easier
intElligent Automatic
SIgn languagE tRanslation

D5.7 QUALITY ESTIMATION MODELS FOR GERMAN AND DGS

Revision: v1.0

Work Package	WP5
Task	T5.4
Due date	12/01/2024
Submission date	07/01/2024
Deliverable lead	University of Zurich
Version	1.0
Authors	Annette Rios (University of Zurich – UZH) Amit Moryossef (University of Zurich – UZH) , Sarah Ebling (University of Zurich – UZH)
Reviewers	Fabrizio Nunnari (DFKI) Özge Mercanoğlu Sincan (UNIS)

Abstract	This deliverable describes a basic setup for quality estimation of the signed-to-spoken and spoken-to-signed translation models described in D4.3 (translation models, final version) and D1.4 (report on final evaluation study).
Keywords	QE, quality estimation, sign language, machine translation

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

Document Revision History

Version	Date	Description of change	List of contributors
V0.1	13/12/2023	First draft	Annette Rios, Amit Moryossef, Sarah Ebling (UZH)
V0.2	15/12/2023	Internal Review 1	Fabrizio Nunnari (DFKI)
V0.3	18/12/2023	Internal Review 2	Özge Mercanoğlu Sincan (UNIS)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the “Intelligent Automatic Sign Language Translation” (EASIER) project’s consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

© 2021 EASIER Consortium

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable	**R**	
Dissemination Level		
PU	Public, fully open, e. g., web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

* R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc

EXECUTIVE SUMMARY

This deliverable describes the quality estimation for the translation models described in D4.3 (translation models, final version) and evaluated in D1.4 (report on final evaluation study) of the EASIER project. In the signed-to-spoken models, sign language is represented as European Meta Sign Language (EMSL) glosses produced by a sign spotter (described in D3.4). In the other direction, sign language output is represented as skeletal pose sequences.

Supervision for training the quality estimation (QE) models comes from the human judgement scores collected in the final evaluation of the translation models (see D1.4). The QE model presented in this work is a very simple approach, but even so, this model achieves almost perfect Pearson correlation with the human scores. This is due to the fact that the evaluation scores of the automatic translations and the human-translated references are almost completely disjoint in a bimodal distribution: there is a large gap in quality and therefore, even a very simple QE model has no problem distinguishing automatic from human translations.

As a conclusion, we propose that instead of augmenting the QE model, future work should focus on improving the quality of the automatic translations.

Funded by the Horizon 2020
Framework Programme of the European Union

CONTENTS

© 2021 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

© 2021 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

EMSL	European Meta Sign Language
QE	quality estimation
MT	machine translation
DA	direct assessment

Sign Languages

BSL	British Sign Language
DGS	German Sign Language / Deutsche Gebärdensprache

Funded by the Horizon 2020
Framework Programme of the European Union

1 INTRODUCTION – QUALITY ESTIMATION IN MACHINE TRANSLATION

QE has always been an important part of machine translation (MT) research, with the goal to ensure the reliability and correctness of the automatically generated translations (Specia et al., 2018). As opposed to quality evaluation, QE usually does not rely on a reference and can be used in production to filter out bad translations on the fly.

QE covers a wide range of approaches, from language-specific, linguistically informed methods (Felice and Specia, 2012; Specia et al., 2013; Martins et al., 2017) to language-agnostic approaches that often rely on features from the translation model itself, such as probability distributions and cross-attention scores (Ding et al., 2021; Fomicheva et al., 2020). Furthermore, QE models differ in terms of the level of granularity at which they predict quality, e.g. at word (Knowles and Koehn, 2018; Ding et al., 2021; Martins et al., 2017), sentence (Specia et al., 2013), or even document level (Soricut and Echihabi, 2010; Graham et al., 2017).

In general, approaches to quality estimation have shifted to large, data-driven, multilingual models that perform reliable assessments, especially on sentence level (Zerva et al., 2022).

QE models are typically trained with supervision, although unsupervised approaches are also possible (Fomicheva et al., 2020). Instead of reference translations, QE usually relies on either human direct assessment (DA) scores or post-edits for supervised training. In this work, we use the DA scores from the final evaluation of the EASIER translation models (see deliverable D1.4 *Report on Final Evaluation Study*) for supervision.

Funded by the Horizon 2020
Framework Programme of the European Union

2 QUALITY ESTIMATION FOR GERMAN-DGS AND DGS-GERMAN

QE as part of task T5.4 is limited to the language pair German – German Sign Language (DGS), in both directions. We will give a short overview on the translation models and the human evaluation that provided the DA scores used to train the quality estimation model.

2.1 MODELS

The models are described in detail in deliverable D4.3 *EASIER Final Translation Systems V2*, but we will give a brief outline of the architectures here.

2.1.1 German-DGS

The German-to-DGS model consists of a pipeline that translates text-to-gloss and then gloss-to-pose (Moryossef et al., 2023). The text-to-gloss component offers three alternatives:

1. lemmatiser
2. rule-based word reordering and dropping component
3. neural text-to-gloss model trained on meineDGS (Konrad et al., 2020)

For the direction German-DGS, the text-to-gloss conversion was done with the first option, the lemmatiser. Perhaps counterintuitively, the lemmatiser was the best-performing of the three alternatives in an interim evaluation of our translation models carried out additionally to the V1 and the final translation evaluation.

After obtaining the glosses, the gloss-to-pose component converts the glosses to a sequence of poses via lexicon lookup. The lexical data consists of the DGS part in the sign language wordnet (Bigeard et al., 2022), a multilingual resource where each sign is linked with at least one video that was converted to skeletal poses with MediaPipe Holistic (Grishchenko and Bazarevsky, 2020).

In the final translation step, the poses are stitched together from the glosses obtained in the first step. Note that the full system includes a pose-to-video step that renders the sequence with an avatar, however, this was not used in the EASIER models presented to the evaluators (see D1.4): The human evaluation scores for translations into DGS are based on pose sequences.

2.1.2 DGS-German

The models for the direction DGS-German are based on EMSL v2.0b, described in detail in deliverable D3.4 *EMSL Representation v2*. A sign spotter labels a given sign language video

Funded by the Horizon 2020
Framework Programme of the European Union

with glosses for the signs that it recognises. These glosses then serve as input to a neural MT system that translates them into German.

The spotter for the evaluated DGS-German model produces glosses in both British Sign Language (BSL) and DGS, as it was trained on the public part of the MeineDGS corpus (Konrad et al., 2020) and the BOBSL dataset (Albanie et al., 2021). The translation model trained on the output of this combined spotter was slightly better than the model trained on output of a DGS-only spotter, therefore, this was the model that was presented to the raters for the final evaluation in D1.4.

It is, however, important to note that the quality of this translation model is very low, the glosses provided by the spotter are sparse and also not always related to the actual content of the utterance. The neural translation model trained on this data therefore mostly learns to invent content, i.e. it produces more or less fluent output in German, but the content is unrelated to the input (so-called “hallucinations").

2.2 HUMAN JUDGEMENT SCORES FOR SUPERVISION

The human evaluation scores used as labels to train the quality estimation model come from the final evaluation study in EASIER, described in D1.4. For the directions German-DGS and DGS-German a total of 5 and 4 evaluators, respectively, contributed the scores in direct assessment.

Evaluators blindly scored automatic and human translations. For the German-DGS direction, the human reference translations were presented as skeletal pose sequences in order to avoid potential differences in scoring due to the contrast between poses and a real person signing.

Evaluators scored the content of the automatic and human translations on a scale from 1 (no meaning preserved) to 100 (perfect meaning).

Table 2.1 contains the results of the evaluation study for German-DGS and DGS-German. Translations in the direction signed-to-spoken were of notably low quality. The relatively low score for the human reference translations in the direction German-DGS is probably due to the fact that the translations were represented as skeletal pose sequences instead of the original videos.

direction	system	human translations	automatic translations
German-DGS	text-lemma-pose	76.077	10.341
DGS-German	video-EMSL-text	85.842	0.286

Table 2.1: *Results of the final evaluation study for German-DGS and DGS-German translations.*

2.3 QUALITY ESTIMATION

We train a simple quality estimation model based on a collection of simple features. The scores from the human evaluation described in the previous section provide the labels for the supervised training of the quality estimation.

Funded by the Horizon 2020
Framework Programme of the European Union

2.3.1 Features

We calculate the following features to as indicators of quality:

Fluency For the signed-to-spoken translations, where German is the target language, we use a multilingual autoregressive language model based on GPT3 (Shliazhko et al., 2022) to score the translations with perplexity. The translation model generally hallucinates the same few sentences, independent of the input, and even though unrelated in content, those sentences are for the most part perfectly fluent, see Fig. 2.1. Furthermore, there is some indication that language models score fluent generated text better in terms of perplexity than human-written text (Gehrmann et al., 2019).

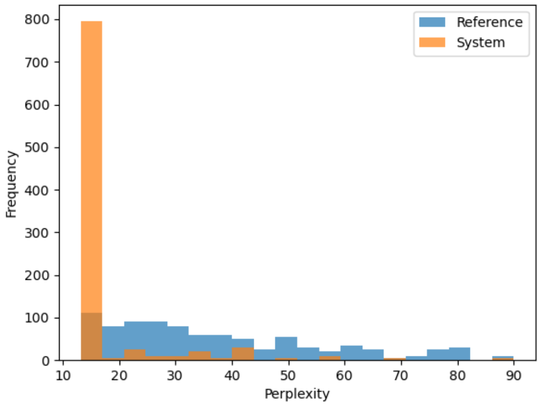
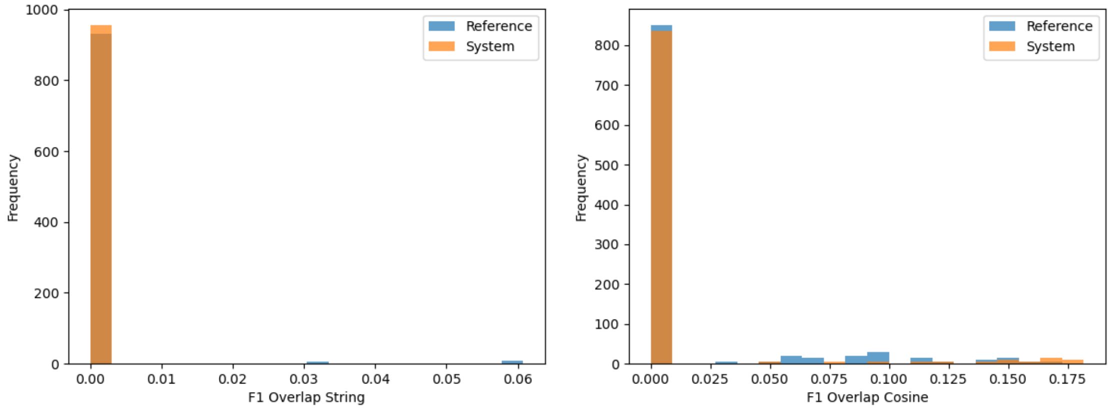


Figure 2.1: *Perplexity Scores on German translations (DGS-German).*

Adequacy For the signed-to-spoken direction, we calculate adequacy in terms of lexical overlap between EMSL glosses and lemmatised German output. We use two alternative approaches: a) string overlap and b) cosine similarity. Option a) is of limited use: the actual lexical overlap is 0 for most samples, both the automatic and the human translations, since EMSL is generally sparse, the generated glosses often do not match the content, and some of the glosses are English lemmas (BSL). Option b) is slightly more informative, since semantically similar words will still count and we can compare German and English word embeddings. We use the word embeddings released by Facebook Research (MUSE)[1] (Conneau et al., 2018). These embeddings have been aligned across languages, i.e. English and German words with similar meanings should have similar embeddings. For both options a) and b), overlap is calculated in terms of F1 score, and for option b), we use a similarity threshold of 0.6 to count two lemmas as a match. Figure 2.2 illustrates the overlap between the automatic and the human translations with the EMSL glosses on the input side. For most EMSL-German pairs, even the human translations, there is no overlap at all.

Number of tokens The number of tokens can be an indicative feature to distinguish an automatic from a human translation in the signed-to-spoken direction, since the model tends to

[1] https://github.com/facebookresearch/MUSE



(a) *F1 string overlap.* **(b)** *F1 cosine similarity overlap.*

Figure 2.2: *Adequacy: F1 overlap EMSL glosses and German translations.*

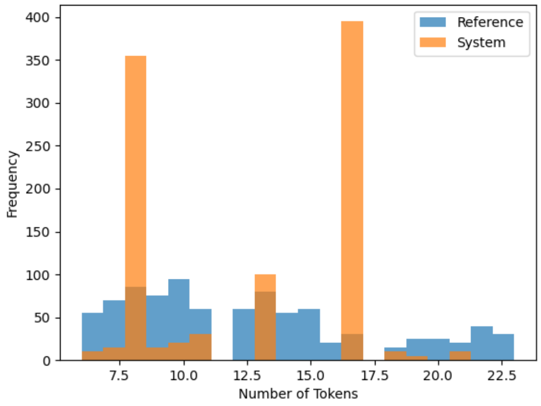


Figure 2.3: *Number of tokens in automatically generated vs. human translations.*

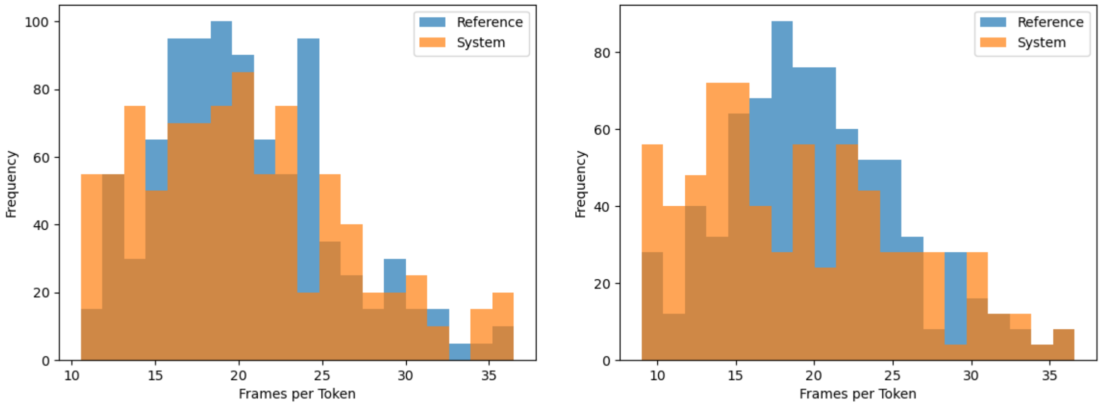
hallucinate the same few sentences independent of the input. Figure 2.3 shows that the human reference translations vary in length, whereas the system output consists of mostly the same three sentence lengths.

Frames-to-token ratio For the signed-to-spoken direction, same as with the number of tokens in the output, the frames-to-token ratio can be indicative of an automatic translation because the model tends to generate the same few German sentences independent of the input, i.e. there is no real correlation between input and output length.

In the direction spoken-to-signed, the model translations tend to be a bit shorter than the human translations. See Figure 2.4 for a visualisation of the ratios for both directions.

Figure 2.5 shows a sample from our dataset for the direction DGS-to-German, on the left with a human reference translation ('Translator-A') and on the right with an automatic translation ('emsl').

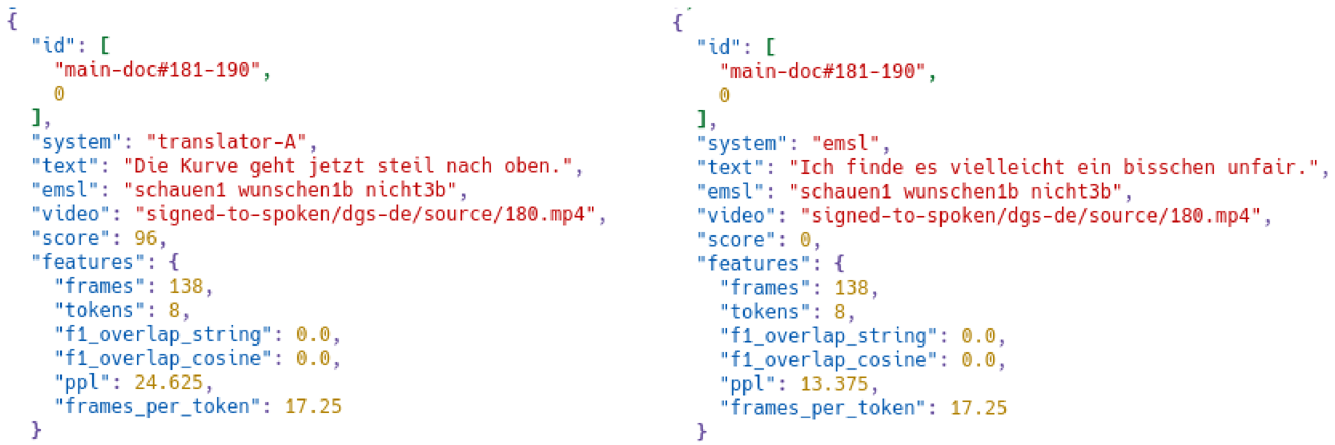
Funded by the Horizon 2020
Framework Programme of the European Union



(a) *DGS-German*

(b) *German-DGS*

Figure 2.4: *Frames per token ratios*



```
{
  "id": [
    "main-doc#181-190",
    0
  ],
  "system": "translator-A",
  "text": "Die Kurve geht jetzt steil nach oben.",
  "emsl": "schauen1 wunschen1b nicht3b",
  "video": "signed-to-spoken/dgs-de/source/180.mp4",
  "score": 96,
  "features": {
    "frames": 138,
    "tokens": 8,
    "f1_overlap_string": 0.0,
    "f1_overlap_cosine": 0.0,
    "ppl": 24.625,
    "frames_per_token": 17.25
  }
```

(a) *Human translation.*

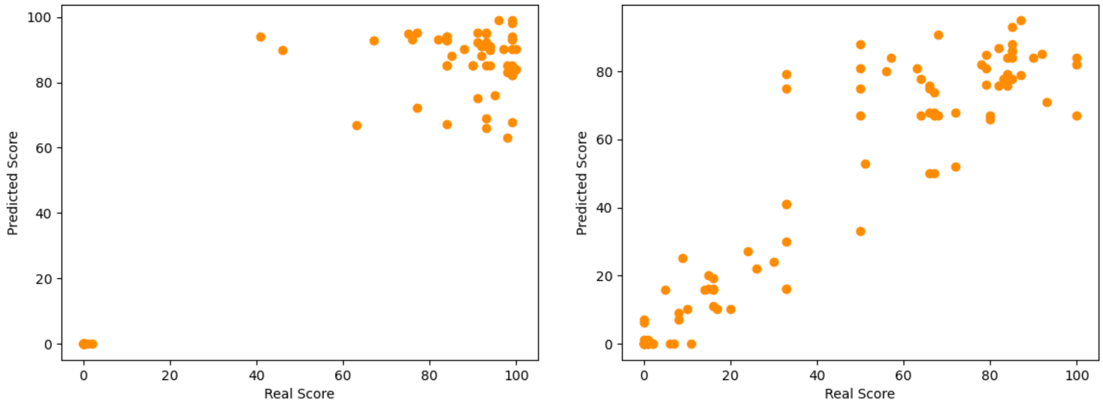
```
{
  "id": [
    "main-doc#181-190",
    0
  ],
  "system": "emsl",
  "text": "Ich finde es vielleicht ein bisschen unfair.",
  "emsl": "schauen1 wunschen1b nicht3b",
  "video": "signed-to-spoken/dgs-de/source/180.mp4",
  "score": 0,
  "features": {
    "frames": 138,
    "tokens": 8,
    "f1_overlap_string": 0.0,
    "f1_overlap_cosine": 0.0,
    "ppl": 13.375,
    "frames_per_token": 17.25
  }
```

(b) *Automatic translation.*

Figure 2.5: *JSON samples with features for the QE model. The value in 'text' is a human translation in a) from 'translator-A', whereas 'text' in b) is an automatic translation.*

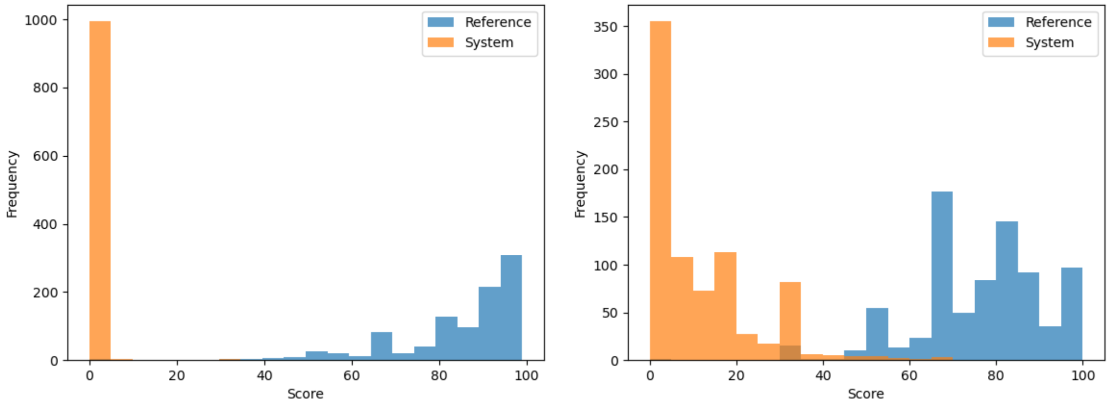
© 2021 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union



(a) *DGS-German: ρ 0.949* **(b)** *German-DGS: ρ 0.914*

Figure 2.6: *Predicted vs. human scores*



(a) *DGS-German* **(b)** *German-DGS*

Figure 2.7: *Evaluation scores on system output vs. human reference translations*

2.3.2 Regression Model

We train a non-linear regresssion model on the features outlined above with the human judgement scores as supervision. The model learns to output a score between 0 and 100 that indicates the estimated quality of the translation.

We fit a support vector regression (SVR) model with an RBF kernel to the data, without any hyperparameter tuning, using the default values C=100, gamma=0.1, epsilon=0.1. We keep 100 records for testing, and train on the remaining data.

2.3.3 Results

To visualise the results, we plot the predicted test scores against the human test scores, and calculate the Pearson correlation coefficient, see Figure 2.6. In both cases, we observe an

© 2021 EASIER Consortium

Funded by the Horizon 2020
Framework Programme of the European Union

almost complete positive correlation between predicted and human scores.

This is due to the fact that the data is almost completely disjoint, there is very minimal overlap between the human judgement scores for the model translations and the scores for the human reference, see Fig. 2.7.

Note that the distribution of scores on the reference translations between DGS-German and German-DGS in Fig. 2.7 also reveals that the evaluators generally scored the German translation in the direction DGS-German higher than the corresponding DGS sequence in the German-DGS direction, even though these are the exact same sentence pairs. This is likely due to the fact that for the spoken-to-signed direction, the evaluators were shown poses rather than videos.

Funded by the Horizon 2020
Framework Programme of the European Union

3 CONCLUSIONS

This deliverable has presented the quality estimation models developed under EASIER. The fact that human scores for automatic and reference translations are almost completely disjoint has led to a very simple SVR model assigning quality scores with very high accuracy.

Instead of upgrading the quality estimation to more complex models, future work should aim to improve the translation quality of the spoken-to-signed and signed-to-spoken models. With a more challenging distribution of human scores, a more refined approach to quality estimation may then be required.

Funded by the Horizon 2020
Framework Programme of the European Union

REFERENCES

Albanie, Samuel, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman (2021). "BOBSL: BBC-Oxford British Sign Language Dataset". In: *CoRR* abs/2111.03635. arXiv: 2111.03635.

Bigeard, Sam, Marc Schulder, Maria Kopf, Thomas Hanke, Kyriaki Vasilaki, Anna Vacalopoulou, Theodore Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, and Eleni Efthimiou (2022). "Introducing Sign Languages to a Multilingual Wordnet: Bootstrapping Corpora and Lexical Resources of Greek Sign Language and German Sign Language". In: *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*. Ed. by Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, and Marc Schulder. Marseille, France: European Language Resources Association, pp. 9–15. URL: https://aclanthology.org/2022.signlang-1.2.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018). "Word Translation Without Parallel Data". In: *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Conference Track Proceedings*. Vancouver, BC, Canada: OpenReview.net. URL: https://openreview.net/forum?id=H196sainb.

Ding, Shuoyang, Marcin Junczys-Dowmunt, Matt Post, and Philipp Koehn (2021). "Levenshtein Training for Word-level Quality Estimation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6724–6733. DOI: 10.18653/v1/2021.emnlp-main.539.

Felice, Mariano and Lucia Specia (2012). "Linguistic Features for Quality Estimation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, pp. 96–103. URL: https://aclanthology.org/W12-3110.

Fomicheva, Marina, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia (2020). "Unsupervised Quality Estimation for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8. Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 539–555. DOI: 10.1162/tacl_a_00330.

Gehrmann, Sebastian, Hendrik Strobelt, and Alexander Rush (2019). "GLTR: Statistical Detection and Visualization of Generated Text". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Marta R. Costa-jussà and Enrique Alfonseca. Florence, Italy: Association for Computational Linguistics, pp. 111–116. DOI: 10.18653/v1/P19-3019.

Graham, Yvette, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton (2017). "Improving Evaluation of Document-level Machine Translation Quality Estimation". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 356–361. URL: https://aclanthology.org/E17-2057.

Funded by the Horizon 2020
Framework Programme of the European Union

Grishchenko, Ivan and Valentin Bazarevsky (2020). *MediaPipe Holistic*. URL: https://google.github.io/mediapipe/solutions/holistic.html.

Knowles, Rebecca and Philipp Koehn (2018). "Lightweight Word-Level Confidence Estimation for Neural Interactive Translation Prediction". In: *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*. Ed. by Ramón Astudillo, João Graça, and André Martins. Boston, MA: Association for Machine Translation in the Americas, pp. 35–40. URL: https://aclanthology.org/W18-2102.

Konrad, Reiner, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder (2020). *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release*. languageresource. Version 3.0. DOI: 10.25592/dgs.corpus-3.0.

Martins, André F. T., Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz (2017). "Pushing the Limits of Translation Quality Estimation". In: *Transactions of the Association for Computational Linguistics* 5. Ed. by Lillian Lee, Mark Johnson, and Kristina Toutanova, pp. 205–218. DOI: 10.1162/tacl_a_00056.

Moryossef, Amit, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling (2023). "An Open-Source Gloss-Based Baseline for Spoken to Signed Language Translation". In: *2nd International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Available at: https://arxiv.org/abs/2305.17714. URL: https://github.com/ZurichNLP/spoken-to-signed-translation.

Shliazhko, Oleh, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina (2022). *mGPT: Few-Shot Learners Go Multilingual*. DOI: 10.48550/ARXIV.2204.07580.

Soricut, Radu and Abdessamad Echihabi (2010). "TrustRank: Inducing Trust in Automatic Translations via Ranking". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre. Uppsala, Sweden: Association for Computational Linguistics, pp. 612–621. URL: https://aclanthology.org/P10-1063.

Specia, Lucia, Carolina Scarton, and Gustavo Henrique Paetzold (2018). "Quality estimation for machine translation". In: *Synthesis Lectures on Human Language Technologies*. Springer Cham, pp. i–148. DOI: https://doi.org/10.1007/978-3-031-02168-8.

Specia, Lucia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn (2013). "QuEst - A translation quality estimation framework". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Miriam Butt and Sarmad Hussain. Sofia, Bulgaria: Association for Computational Linguistics, pp. 79–84. URL: https://aclanthology.org/P13-4014.

Zerva, Chrysoula, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia (2022). "Findings of the WMT 2022 Shared Task on Quality Estimation". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Ed. by Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 69–99. URL: https://aclanthology.org/2022.wmt-1.3.

Funded by the Horizon 2020
Framework Programme of the European Union