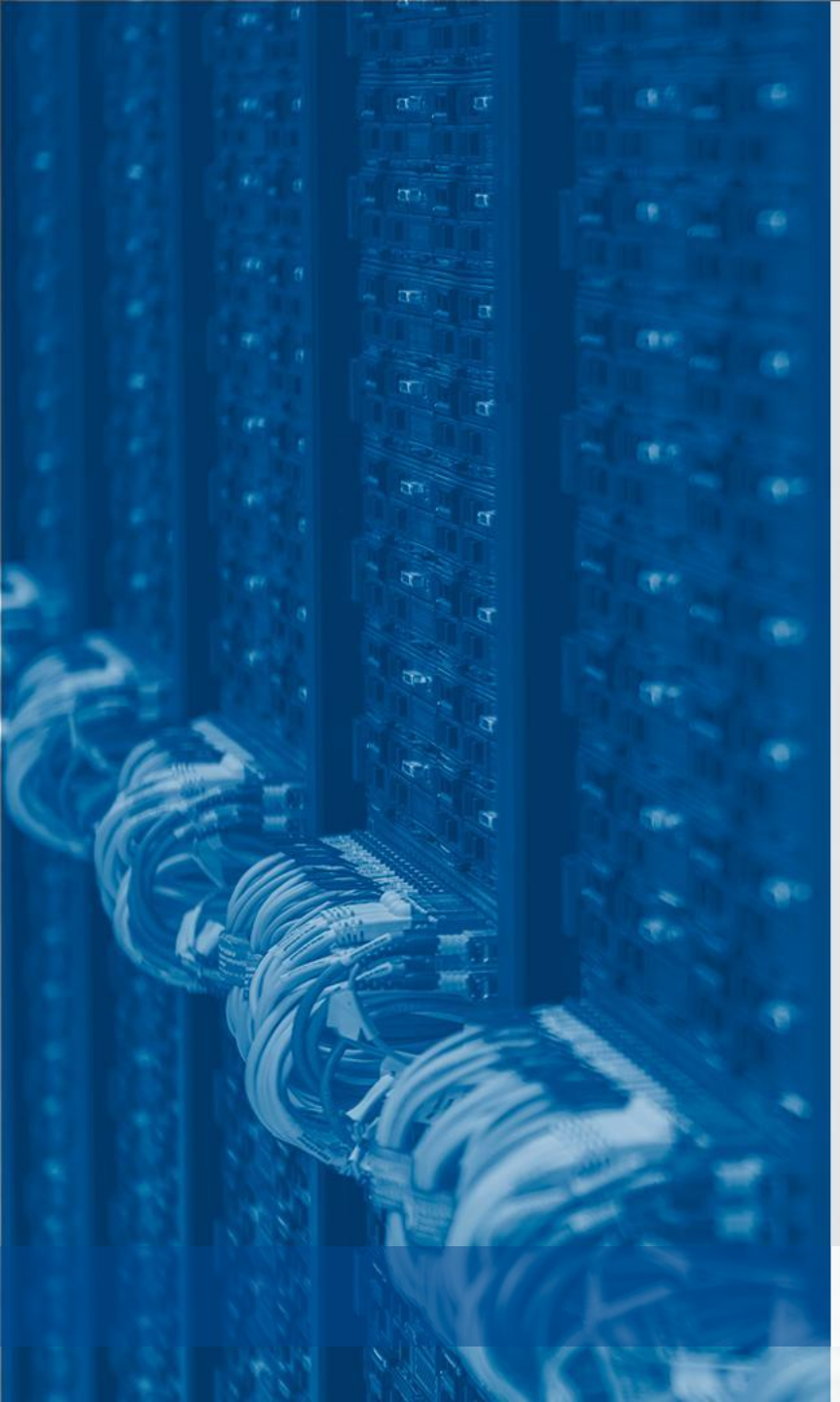


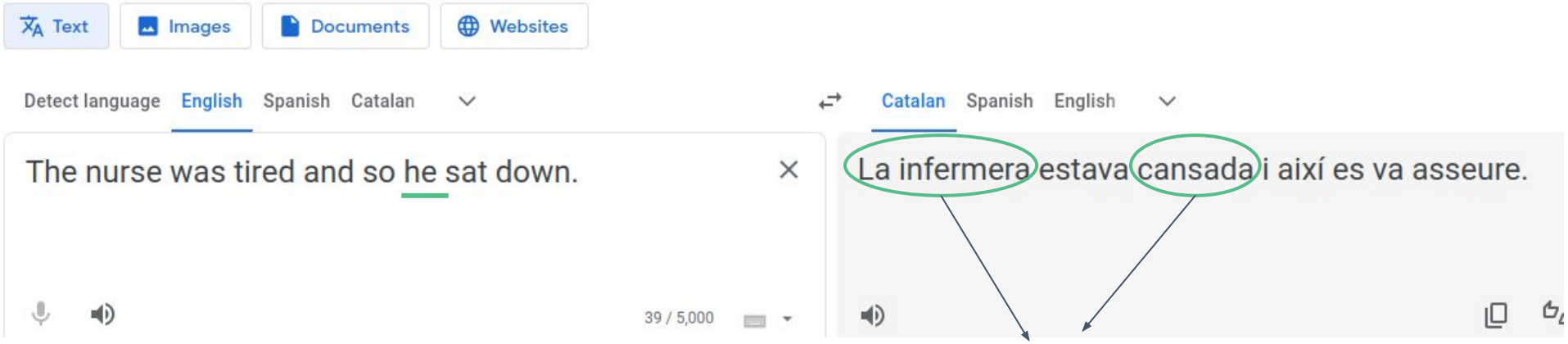
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Unmasking Biases: Exploring Gender Bias in English-Catalan Machine Translation through Tokenization Analysis and Novel Dataset

Audrey Mash, Carlos Escolano, Aleix Sant, Francesca De Luca Fornaciari, Maite Melero

The Problem: Model Bias

- Bias in models refers to information learnt which is unnecessary and potentially harmful.
- Often reflects stereotypes present in society & therefore the data.



The Problem: Linguistic Gender

- Not all languages contain the same amount of gender information, making it hard to generate accurate translations between them.

NO GENDER	FINNISH	Hän on tyytyväinen arvosanoihinsa. Pron3SG is satisfied with-grades-POSS3PL
NOTIONAL GENDER	ENGLISH	**She** is happy with **her** grades. Pron3SG(F) is happy with poss_pron3SG(F) grades
GRAMMATICAL GENDER	CATALAN	**Ella** està **contenta** amb les seves notes. Pron3SG(F) is happy(F) with poss_pron3PL(F) grades(F)

Previous Approaches

Pre-existing bias:

- Gender tagging (Elaraby et al., 2018; Stafanovičs et al., 2020; Vanmassenhove et al., 2018)
- Debiasing of word-embeddings (Escudé Font and Costa-jussà, 2019)
- Fine-tuning on gender balanced dataset (Costa-jussà and de Jorge, 2020)

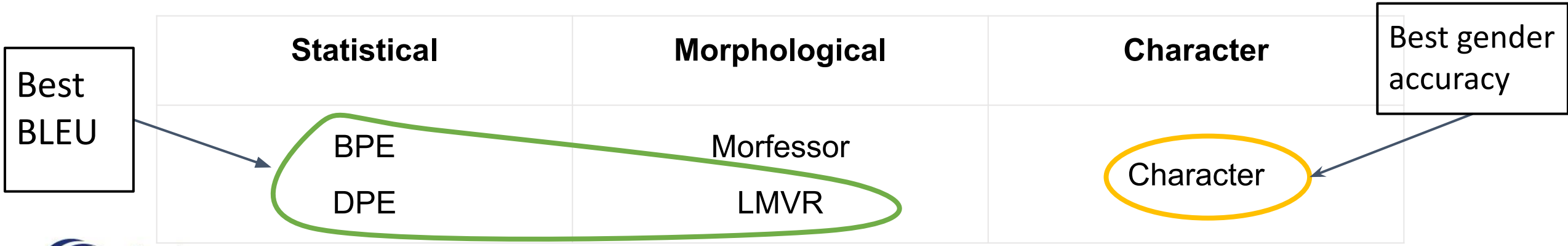
Technical bias:

- Shared vs. language specific Encoder-Decoders (Costa-jussà et al.,2020)
- Tokenization (Gaido et al., 2021)

Gaido et al. (2021)

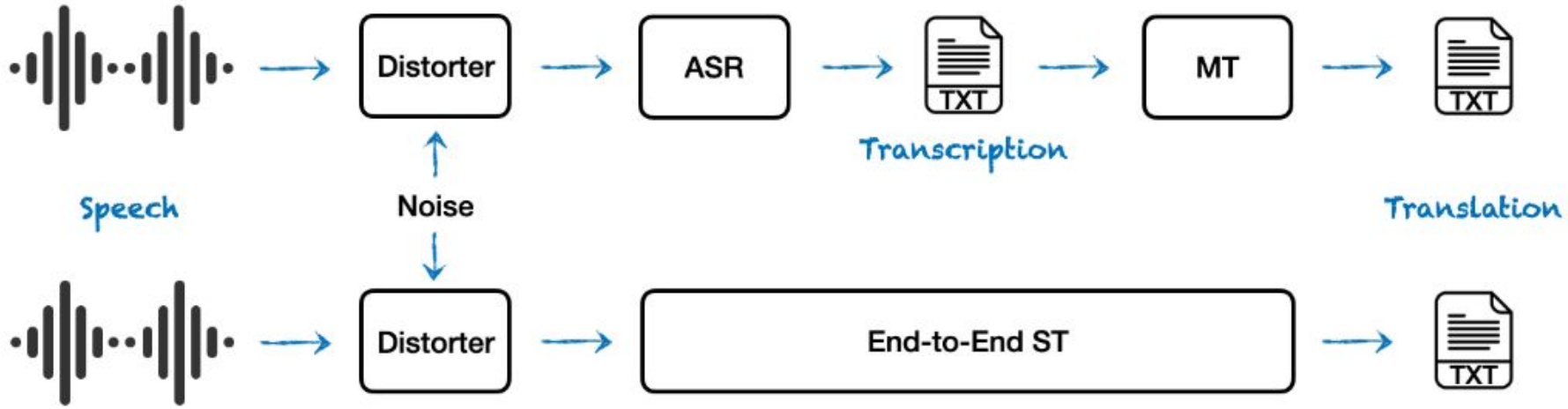
- Feminine forms are both morphologically more complex and sparser in the data than masculine forms
- Morphological & character based segmentations have in some studies (Smit et al., 2014; Ataman et al., 2014; Belinkov et al., 2020) been shown to be better for at capturing morphology than statistical methods

Investigate whether statistical segmentation disadvantages less frequent, more morphologically complex feminine forms in direct Speech Translation.



Limitations for our context

- Conducted their experiment in direct Speech Translation
- Character has an advantage in this context because of its correlation with phonemes
- May also gain gender information from the audio



Cortes, 2020

Research Question:

To what extent does the choice of tokenization method for NMT impact the gender bias of the model's output?

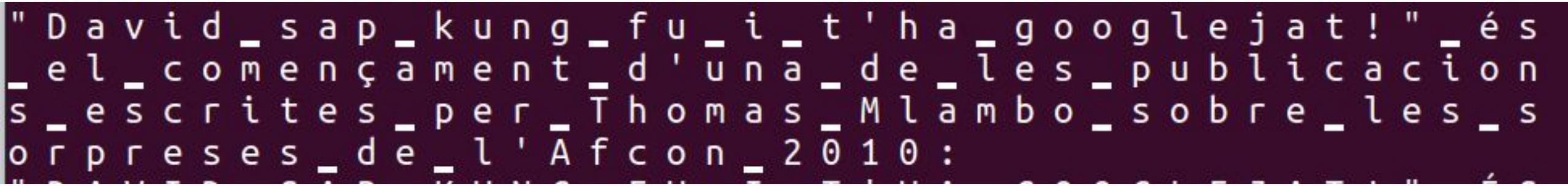
Hypothesis 1: Character-based tokenization will perform more poorly than other methods for both translation quality and gender accuracy.

Hypothesis 2: Morphological tokenization will show less gender bias than statistical tokenization methods.

Experiments

- Trained four models from scratch
  - Character Based
  - Byte Pair Encoding
  - SentencePiece Unigram
  - Morfessor
- Trained on a high quality English - Catalan corpus of 8.218.519 sentence pairs, initially compiled for training the Projecte Aina ca-en MT model
- Evaluated for overall translation quality (BLEU) and gender bias

Character-Based Tokenization



- Ensures that all tokens can be processed - nothing is OOV (Costa-jussà & Fonollosa,2016)
- Generally not great for text because of the lack of semantic meaning carried in each token. (Libovický et al., 2022)
- May have advantages in morphologically rich languages (Costa-jussà & Fonollosa, 2016; Libovický et al., 2022)
- Manually split

Byte Pair Encoding (BPE) - 32k vocab

```
"@@ David sap k@@ ung f@@ u i t'@@ ha go@@ og@@ le@@ j@@ at@@ !" és el comen@@ ç
@@ ament d'@@ una de les public@@ acions es@@ cr@@ ites per Thomas M@@ lam@@ bo
sobre les sor@@ pres@@ es de l'@@ Af@@ con 2010@@ :
```

- Very popular tokenization method in NMT - used in GPT-3 (Brown et al., 2020)
- Initiates with character level vocabulary and iteratively merges the most frequent combinations until reaches desired number of merges (Sennrich et al., 2016)
- Implemented using the Subword-NMT library

Unigram Model - 32k vocab

```
_" D a vid _sap _k ung _ fu _i _t ' ha _ google jat !" _és _el _començament _d '
 una _de _les _publicacions _escrit es _per _Thomas _M lam bo _sobre _les _s or
pres es _de _l ' A f con _2010 :
```

- Initiates with a large seed vocabulary and iteratively reduces by removing the least probable n-grams (Kudo, 2018)
- Less control over the final vocabulary size
- Implemented using SentencePiece library

Morphological Tokenization (Morfessor)

```
_" David _sap _kung _fu _i _t' ha _google jat ! " _és _el _comença ment _d'una _
de _les _publica cions _escrit es _per _Thomas _M la mbo _sobre _les _sorpreses
_de _l' Af con _2010 :
```

- Morfessor learns approximate morphological segmentation using a Minimum Defined Length model (Creutz & Lagus, 2002, Smit et al., 2014)
- Not exactly morphological, but should be closer than other methods.
- No possibility to fix the vocabulary size but we limited the Fairseq dictionary to 150K.

Fairseq Dictionary size by Tokenization Method

Tokenisation Method	Dictionary Size
Character based	8,164
Morfessor	149,996
BPE	36,542
Unigram	70,636

Evaluation Corpus

En-Es MuST-SHE Corpus (Bentivogli et al. 2020)

- Based on MuST-C dataset, a multilingual audio & text based corpus compiled from TED talk data.
- For each source segment there is a **correct** reference translation and a **gender-swapped** reference translation, as well as all **gender terms**.
- Divides segments into categories based on source of gender information (text, audio, no information)
- MuST-SHE en-es contains 1164 segments

En-Ca MuST-SHE Corpus (text only)

- Reduced the original categories to two - gender information or no gender information
- Automatically translated from the Spanish using PlanTL ES-CA model
- Translations manually revised by native Catalan speaker
- Extracted gender terms with a heuristic based on spaCy morphologizer.
- Discarded sentence triplets which did not contain gender terms.
- Resulting corpus contains **1046** segments

Example segment of MuST-SHE en-ca

En	Ca-Ref	Ca-Wrong-Ref	Gender Terms
The most ambitious and most competent leader on the international stage today is Chinese President Xi Jinping.	El líder més ambiciós i competent en l'escena internacional avui, és el president Xi Jinping.	La lideressa més ambiciosa i competent en l'escena internacional avui, és la presidenta Xi Jinping.	El La;líder lideressa;ambiciós ambiciosa;el la;president presidenta

Results

Overall Translation Quality

BLEU SCORES				
	Char	BPE	Uni	Morf
Flores Dev	39.2	41	**41.6**	41.2
Flores DevTest	39.1	41.3	**42.0**	41.5
MuST-SHE	51.6	56.4	**57**	56.6
AVG	43.3	46.2	**46.9**	46.4

Gender Accuracy (with context)

		Char	BPE	Uni	Morf
F1 scores for gender translation with provided context	F	57.95%	**65.99%**	65.92%	60.10%
	M	98.47%	**98.78%**	98.34%	98.51%
	Overall	78.21%	**82.39%**	82.13%	79.31%

Gender Output (without context)

	Char		BPE		Uni		Morf	
Feminine	11.87	75	10.35%	65	10.61%	68	15.04%	94
Masculine	88.13%	557	89.65%	563	89.39%	573	84.96%	531

Conclusions

- We have not shown any significant impact of tokenization on gender bias here.
- Character based tokenization has performed more poorly than other tokenization methods by all metrics.
- Morphological tokenization has not shown any improvement to gender accuracy over statistical methods
- The increased output of feminine forms by Morfessor is interesting but not conclusive and warrants further investigation.

References

- Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., & Turchi, M. (2020). Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6923–6933. https://doi.org/10.18653/v1/2020.acl-main.619
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165
- Cortes, G. (2020). Towards Robust End to End Speech Translation. UPC TFM: https://upcommons.upc.edu/bitstream/handle/2117/330400/GuillemCortes-Towards_Robust_End_to_End_Speech_Translation.pdf
- Costa-jussà, M. R., Escolano, C., Basta, C., Ferrando, J., Batlle, R., & Kharitonova, K. (2020). *Gender Bias in Multilingual Neural Machine Translation: The Architecture Matters* (arXiv:2012.13176). arXiv. https://doi.org/10.48550/arXiv.2012.13176
- Costa-jussà, M. R., & de Jorge, A. (2020). Fine-tuning Neural Machine Translation on Gender-Balanced Datasets. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 26–34. https://aclanthology.org/2020.gebnlp-1.3
- Costa-jussà, M. R., & Fonollosa, J. A. R. (2016). Character-based Neural Machine Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 357–361. https://doi.org/10.18653/v1/P16-2058
- Creutz, M., & Lagus, K. (2002). *Unsupervised Discovery of Morphemes* (arXiv:cs/0205057). arXiv. https://doi.org/10.48550/arXiv.cs/0205057
- Domingo, M., Garcıa-Martınez, M., Helle, A., Casacuberta, F., & Herranz, M. (2019). *How Much Does Tokenization Affect Neural Machine Translation?* (arXiv:1812.08621). arXiv. https://doi.org/10.48550/arXiv.1812.08621
- Elaraby, M., Tawfik, A. Y., Khaled, M., Hassan, H., & Osama, A. (2018). Gender aware spoken language translation applied to English-Arabic. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 1–6. https://doi.org/10.1109/ICNLSP.2018.8374387
- Escudé Font, J., & Costa-jussà, M. R. (2019). Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. In *ArXiv e-prints*. https://doi.org/10.48550/arXiv.1901.03116
- Gaido, M., Savoldi, B., Bentivogli, L., Negri, M., & Turchi, M. (2020). Breeding Gender-aware Direct Speech Translation Systems. *Proceedings of the 28th International Conference on Computational Linguistics*, 3951–3964. https://doi.org/10.18653/v1/2020.coling-main.350
- Kudo, T. (2018). *Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates* (arXiv:1804.10959). arXiv. https://doi.org/10.48550/arXiv.1804.10959
- Libovický, J., Schmid, H., & Fraser, A. (2022). *Why don't people use character-level machine translation?* (arXiv:2110.08191). arXiv. https://doi.org/10.48550/arXiv.2110.08191
- McConnell-Ginet, S. (2013). ` Gender and its relation to sex: The myth of 'natural' gender. In ` *Gender and its relation to sex: The myth of 'natural' gender* (pp. 3–38). De Gruyter Mouton. https://doi.org/10.1515/9783110307337.3
- Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP* (arXiv:2112.10508). arXiv. http://arxiv.org/abs/2112.10508
- Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162
- Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 21–24. https://doi.org/10.3115/v1/E14-2006
- Stafanovičs, A., Bergmanis, T., & Pinnis, M. (2020). Mitigating Gender Bias in Machine Translation with Target Gender Annotations. *Proceedings of the Fifth Conference on Machine Translation*, 629–638. https://aclanthology.org/2020.wmt-1.73
- Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting Gender Right in Neural Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3003–3008. https://doi.org/10.18653/v1/D18-1334

Acknowledgements

This work has been promoted and financed by the *Generalitat de Catalunya* through the Aina project.

This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública* and *Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335 and 2022/TL22/00215334.

Generalitat de Catalunya
Government
of Catalonia

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank You