Measuring Normative and Descriptive Biases in Language Models Using Census Data

Samia Touileb
University of Bergen
samia.touileb@uib.no

Lilja Øvrelid
University of Oslo
liljao@ifi.uio.no

Erik Velldal
University of Oslo
erikve@ifi.uio.no

Abstract

We investigate in this paper how distributions of occupations with respect to gender is reflected in pre-trained language models. Such distributions are not always aligned to normative ideals, nor do they necessarily reflect a descriptive assessment of reality. In this paper, we introduce an approach for measuring to what degree pre-trained language models are aligned to normative and descriptive occupational distributions. To this end, we use official demographic information about gender–occupation distributions provided by the national statistics agencies of France, Norway, United Kingdom, and the United States. We manually generate template-based sentences combining gendered pronouns and nouns with occupations, and subsequently probe a selection of ten language models covering the English, French, and Norwegian languages. The scoring system we introduce in this work is language independent, and can be used on any combination of template-based sentences, occupations, and languages. The approach could also be extended to other dimensions of national census data and other demographic variables.

1 Introduction

Pre-trained language models (LMs) may contain various types of biases, and the field of NLP has seen a lot of work in recent years on attempting to identify, mitigate, and reduce these biases. Biases can originate both from the unlabeled texts used for pre-training these LMs, and from texts and annotations used for tuning downstream classifiers. LMs have become a cornerstone in most NLP model architectures, and the extent to which they reflect, amplify, and spread the biases present in their training data is still a problematic issue to be solved.

Several efforts in this direction have focused on *gender* as a variable (Touileb and Nozza, 2022; Touileb et al., 2021; Ousidhoum et al., 2021; Nozza et al., 2021; Touileb et al., 2020; Saunders and Byrne, 2020; Bhaskaran and Bhallamudi, 2019; Cho et al., 2019; Prates et al., 2018), also in correlation with occupations (Borchers et al., 2022; Touileb et al., 2022; Bolukbasi et al., 2016). While there have been several efforts on exploring the existing biases related to these demographic variables, most work approaches the task from a normative point of view (Blodgett, 2021), where equality between the demographic distributions is prioritized.

Although normativity in this aspect is crucial for certain applications, we argue that it is also interesting to explore the task from a descriptive perspective. This is especially interesting for occupations, since a descriptive and realistic view of society already contains gender disparities. We propose that national census data, in our case about gender-occupation distributions, can offer a reliable ground truth against which model predictions can be compared. Moreover, we argue for taking both normative and descriptive assessments into account, in order to give a broader picture of the representations of demographics within LMs. This has also been partly pointed out by Blodgett et al. (2020), who stress the importance of the connection between language and social hierarchies, which has not been taken into consideration in most previous work on bias in NLP.

In this paper, we introduce a new score for measuring how LMs are aligned with normative and descriptive occupational demographic distributions. We use demographic distributions covering occupations in four countries, namely France, Norway, the United Kingdom, and the United States. We manually select gendered pronouns and nouns, as well as specific verb phrases, to construct template-based sentences, subsequently used to probe a selection of ten LMs covering the relevant languages.

Our contributions include; (i) creating novel benchmark datasets for English, French, and Norwegian based on manually crafted templates to mea-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2242–2248
May 2-6, 2023 ©2023 Association for Computational Linguistics

sure occupational gender biases, (ii) proposing a scoring system to measure normative and descriptive biases in LMs, and (iii) releasing our code and data for reproducibility.

In what follows, we give a detailed description of our new benchmark datasets in Section 2. We then, in Section 3, give a detailed description of the normative and descriptive bias scores, and present our analysis on ten LMs as proof of concept. We discuss and summarize our findings in Section 4, and conclude by discussing possible directions for future work in Section 5. The limitations of our work are discussed in the Limitations Section.

2 Benchmark datasets

In this work we develop a set of benchmark templates for English, French, and Norwegian that cover occupations in France, Norway, the United Kingdom (UK), and the United States (US). These templates are then used for probing different LMs. More details are given in what follows.

Occupations We retrieve country-specific lists of occupations and their associated (male/female) gender ratios from the national statistics bureaus of France, Norway, UK, and the US.[1] This resulted in 235 occupations from France,[2] 415 from Norway,[3] 325 from the UK,[4] and finally 314 occupations from the US.[5] All of these occupations were listed in either masculine singular or masculine plural form. As some of the languages we are focusing on inflect nouns for gender, we manually generate for each occupation in singular masculine form the corresponding forms in singular feminine, plural feminine, and plural masculine. This was performed by a native speaker of Norwegian and French, and a proficient speaker of English. Table 1 shows the top 5 female-dominated, male-dominated, and gender balanced occupations in each census data.

Templates Our work builds on the methodology of template-based probing. To measure a model's occupational biases we follow the same procedure for all languages. Our templates are based on the gender-inflected occupations, preceded by a sequence of selected gendered pronouns and a set of gender-specific identifier terms in singular and plural forms, followed by a predicate generically denoting the act of having an occupation. As an example, a template could be:

$$\underbrace{\text{The woman}}_{\text{gender-specific identifier}}\ \underbrace{\text{worked as a}}_{\text{predicate}}\ \underbrace{\text{nurse}}_{\text{occupation}}$$

We select 28 gender-specific identifiers, and 6 predicates for all three languages. The full list of gender-specific identifiers can be found in Table 2 and the list of predicates in Table 5 in Appendix A. Combining these identifiers and predicates with our country-specific occupations, we create a set of 12.726 template-based probes for French occupations, 69.720 for Norwegian, 50.700 for the UK, and 48.984 for the US.

The templates we created cover different grammatical tenses, such that each template is given in the past, present, and future tense. We have decided to include such a broad collection of variations to the templates to get a better representation of how occupations are correlated with genders, especially since research has shown that bias probes are sensitive to grammatical tense (Touileb, 2022).

3 Method

LMs trained with a masked language modelling objective are trained such that random tokens in the input training data are replaced with a placeholder token, `[MASK]`, which will subsequently be predicted by the trained model. Template-based approaches to probe biases take advantage of this feature of LMs. For our purposes, we mask the gendered identifier in each template-generated probe (as introduced in Section 2), and use the returned probability of each masked identifier to compute our bias scores. A masked version of the example template above would be:

$$\underbrace{\text{The [MASK]}}_{\text{gender-specific identifier}}\ \underbrace{\text{worked as a}}_{\text{predicate}}\ \underbrace{\text{nurse}}_{\text{occupation}}$$

Language models We select ten LMs covering the three languages English, French, and Norwegian. All models are available from the HuggingFace library (Wolf et al., 2020). We use four Norwegian models, four English models, and two French models. These are:

- NorBERT (Kutuzov et al., 2021): trained from scratch on the Norwegian newspaper corpus[6],

[1] All of the statistics were retrieved in October 2022.
[2] https://dares.travail-emploi.gouv.fr/donnees/portraits-statistiques-des-metiers
[3] https://utdanning.no/likestilling
[4] https://www.nomisweb.co.uk/datasets/aps168/
[5] https://www.bls.gov/cps/cpsaat11.htm

[6] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

	Female-dominated occupations	Male-dominated occupations	Gender-balanced occupations
FR	midwife kindergarten assistant secretary office secretary executive secretary	navy officer and boatswain construction machinery operator pipe fitter panel beater carpenter	doctor higher education teacher medical device specialist admin. and financial executive dentist
NO	knitter midwife public health nurse skin care specialist dental health secretary	coastal skipper chief engineer scaffold builder roofer bricklayer	doctor architect lawyer politician associate professor
UK	midwife school secretary dancer and choreographer dental nurse medical secretary	roofer, roof tiler and slater carpenter and joiner construction and building supervisor bricklayer and mason vehicle technician and mechanic	barrister and judge laboratory technician paramedic industrial trainer and instructor legal professional
US	skincare specialist preschool and kindergarten teacher executive secretary speech-language pathologist dental hygienist	cement mason electrical power-line installer crane and tower operator heavy vehicle technician and mechanic bus and truck mechanic	insurance sales agent medical scientist dental laboratory technician photographer advertising sales agent

Table 1: Top 5 gender-dominated and gender-balanced occupations in census data from France (FR), Norway (NO), the United Kingdom (UK), and the United States (US). The occupations presented here are either dominated by more than 98% of either gender, or have a more balanced distribution (between 45% and 55%) between both female and male genders.

and Norwegian Wikipedia. The model comprises about two billion word tokens.

- NorBERT2[7]: the non-copyrighted subset of the Norwegian Colossal Corpus (NCC)[8] and the Norwegian subset of the C4 web-crawled corpus (Xue et al., 2021) were used to train this model from scratch. It comprises about 15 billion word tokens.
- NB-BERT_base (Kummervold et al., 2021): trained on the full version of the NCC corpus. This model used the architecture of the BERT cased multilingual model (Devlin et al., 2018). It comprises around 18.5 billion word tokens.
- NB-BERT_Large[9]: trained similarly to the NB-BERT_base model.
- BERT_base (Devlin et al., 2018) and BERT_Large: trained on English Wikipedia and Google's Books Corpus.
- RoBERTa_base (Liu et al., 2019) and RoBERTa_Large: trained on the BookCorpus, English Wikipedia, CC-news corpus (English news), OpenWebText dataset, and Stories dataset (a subset of the Common Crwal corpus).
- CamemBERT (Martin et al., 2020): trained on the OSCAR corpus (Ortiz Suárez et al., 2019), which is a multilingual corpus created by filtering the Common Crawl corpus.
- Barthez (Eddine et al., 2020): trained on the French part of the Common Crawl and Wikipedia, in addition to various smaller corpora (Eddine et al., 2020).

Scoring system The scoring system we introduce is the same for both bias scores. We will give more details on the differences of the scores in their respective sections.

For each template, and for each language, 28 gender-specific identifiers and 6 different predicates were used with each occupation. To compute the scores, we average over the gendered-identifiers and the predicates of the LMs' returned probabilities for each template. For each template, only one gender is represented (female or male). If the language inflects for gender, all components of a template reflect the gender in question, otherwise it is only reflected in the identifier.

We average the scores for a given occupation by gender, by summing and normalizing the probabilities of each identifier and the total probability

[7] https://huggingface.co/ltgoslo/norbert2

[8] https://github.com/NbAiLab/notram/blob/master/guides/corpus_description.md

[9] https://huggingface.co/NbAiLab/nb-bert-large

Norwegian	English	French
Brødrene	He	Elle
Broren	She	Elles
Dama/Damen	The aunt	Il
Damene	The aunts	Ils
Datteren	The boy	L'homme
Døtrene	The boys	L'oncle
Faren	The brother	La dame
Fedrene	The brothers	La femme
Gutten	The daughter	La fille
Guttene	The daughters	La mère
Han	The father	La soeur
Hun	The fathers	La tante
Jenta/Jenten	The girl	Le fils
Jentene	The girls	Le frère
Kvinnen	The ladies	Le garçon
Kvinnene	The lady	Le père
Mannen	The man	Les dames
Mennene	The men	Les femmes
Mødrene	The mother	Les filles
Moren	The mothers	Les fils
Onkelen	The sister	Les frères
Onklene	The sisters	Les garçons
Sønnen	The son	Les hommes
Sønnene	The sons	Les mères
Søsteren	The uncle	Les oncles
Søstrene	The uncles	Les pères
Tanten	The woman	Les sœurs
Tantene	The women	Les tantes

Table 2: Gender-specific pronouns and identifiers.

values returned by the LM, here dubbed $proba_G$, where G can be female or male (equation (1)). Then using this overall probability of a gender for a template, we average these values over all templates related to the occupation (equation (2)). More formally, for a language model *LM*, for each occupation O, there are a number of templates T, and a number of identifiers i and predicates p, reflecting a gender G. We define the bias score as follows:

$$proba_G = \frac{\sum_i T_p}{|i|} \tag{1}$$

$$score_O = \frac{\sum_{proba_G} T_O}{|T_O|} \tag{2}$$

Descriptive bias score Once the scores $score_O$ are computed, the descriptive bias score compares the percentages of distribution of occupations in the LMs to the ground truth data that comes from the respective census data of our countries of interest. We impose a threshold on the gender distributions in such a way that the category of *gender-imbalanced* occupations here corresponds to all occupations exceeding 55% of distribution for one gender, while *gender-balanced* occupations are those which percentages lie around $50\%_{\pm 5}$ for each gender.

Model	Normative	Descriptive
NorBERT	16.23	39.31
NorBERT2	3.17	34.67
NB-BERT	18.55	36.50
NB-BERT_Large	11.35	40.90
BERT_UK	18.05	35.33
BERT_large_UK	13.73	40.43
RoBERTa_base_UK	0.15	34.56
RoBERTa_large_UK	0.00	34.56
BERT_US	17.25	43.29
BERT_Large_US	12.46	48.88
RoBERTa_base_US	0.15	42.81
RoBERTa_Large_US	0.31	42.81
CamemBERT	10.46	34.10
BARThez	6.45	37.08

Table 3: Normative and descriptive occupational bias scores.

We look at the extent to which this score aligns with the census data. We compute an overall score disregarding gender, in addition to class-level scores: female dominated occupations (more than 55% in census are females), male dominated occupations (more than 55% in census are males), neutral occupations (between 45% and 55% of occupations in census for either gender).

Normative bias score The normative bias score also builds on top of the scores $score_O$, and compares the resulting distribution of occupations in LMs to a normative description of all occupations, such that percentages of either gender should be around $50\%_{\pm 5}$.

From a normative point of view, equal representations should be given to females and males. Instead of just setting the distribution to a strict value of 50-50, we decided that for either gender, the distribution should range anywhere between 45% and 55% in the census data. This to say, that if an occupation has 45% and 55% males, we consider it a balanced distribution.

4 Results and discussion

Table 3 shows the resulting normative and descriptive bias scores of the ten LMs. All scores represent percentages, *i.e.,* the percentage of model predictions that align with our normative values or descriptive demographic distributions. With no surprise, it is clear that all models exhibit fairly weak performance according to the normative bias score. The weakest performing model normatively speaking is RoBERTa (both base and large) on both UK and US statistics. BERT seems to be a bit better on UK statistics, but the difference is not signif-

Model	Neutral	Female	Male
NorBERT	1.46	22.34	15.50
NorBERT2	0.24	33.57	0.85
NB-BERT	1.46	23.68	11.35
NB-BERT_Large	0.12	33.82	6.95
BERT_UK	1.54	33.02	0.77
BERT_Large_UK	1.23	31.63	7.56
RoBERTa_base_UK	0.00	34.56	0.00
RoBERTa_Large_UK	0.00	34.56	0.00
BERT_US	2.39	39.93	0.95
BERT_Large_US.	1.75	40.09	7.02
RoBERTa_base_US	0.00	42.81	0.00
RoBERTa_Large_US	0.00	42.81	0.00
CamemBERT	0.00	0.00	34.10
BARThez	0.00	0.00	37.08

Table 4: Descriptive bias scores of gender-imbalanced and gender-neutral occupations. The two gender-imbalanced occupations cover female dominated occupations (more than 55% in census are females), and male dominated occupations (more than 55% in census are males). The gender-neutral occupations are those with distributions between 45% and 55% in census data for either gender.

icant. For the remaining languages, NorBERT2 is the worst Norwegian model normatively and BARThez is the worst of the two French models.

Results of the descriptive scores are in general higher. Most models seem to reflect the demographic occupational distribution to a certain extent. Both BERT models achieve highest descriptive scores, performing best on the US census data. While the RoBERTa models obtain the lowest performance in terms of the normative score, they rank second in the descriptive score on the US census data. NorBERT2 is still the weakest performing Norwegian model, ranking last both descriptively and normatively, while BARThez seems to yield the best descriptive score for French.

To get a more detailed overview of which types of occupations the tested models seem to represent the best, we also computed the descriptive scores of gender-imbalanced and gender-neutral occupations separately. Results can be seen in Table 4. Interestingly, all Norwegian and English models are better at identifying female-dominated occupations, while the two French models seem to only identify male-dominated occupations.

All models exhibit the lowest scores on gender-neutral occupations, hinting at the tendency that models correlate most occupations with one gender, rather then equally representing them. This would also align with the lower normative scores that we generally see.

Since the occupations in the census data differ from country to another, it is difficult to compare and rank models across languages. A fair comparison of these models is to focus on performance by country rather than across them. Even if we state that some models are the best or worst using one scoring system, the country-level scores are the most important measure of bias in the models.

5 Conclusion

We have introduced a new scoring system for measuring occupational biases in pre-trained language models. The scoring system allows the attribution of two scores: a normative score and a descriptive score. While the normative score sheds light on to what extent the correlations between genders and occupations are balanced, the descriptive score uses real world demographic distributions to reflect to what degree the language models reflect reality.

As a proof of concept we test our scoring systems on ten language models covering the French, Norwegian, and English languages. It comes as no surprise that all models exhibit low scores when using the normative scoring, while most of them have an adequate score when measured descriptively. What is more interesting is that our scoring mechanism allows us to separate between the normative and descriptive aspects of the model properties. All templates and codes are made publicly available on our GitHub repository.[10]

While we have limited our analysis to three languages, our approach is language agnostic and only requires language specific templates and demographic statistics on the distributions of occupations with respect to gender, something most national census agencies should be able to provide. Moreover, the approach could also be extended to other dimensions of national census data and other demographic variables.

On the note of future directions, we also plan to investigate cross-cultural effects, by comparing models for different languages across the gender–occupation lists from different nations, and also including multi-lingual models. Moreover, we also plan to systematically test the impact of different text sources (used to train language models) on our bias scores.

[10]https://github.com/SamiaTouileb/Normative-Descrptive-scores

Limitations

The major limitation of our work is that we focus on a binary gender setting. We acknowledge the fact that gender as an identity spans more than just two categories, however, the demographic census data we use have only the two genders (female and male) represented.

As proof of concept, the templates we use in this work are limited to one framing of how gendered pronouns and nouns can co-occur with occupations. Extending this to more diverse templates might give a broader context and a better representation of genders in LMs.

The applicability of English models to the UK and US census data, and the French models to the French census data, might also give a skewed representation of occupations. Both English and French are spoken across many countries, which might have an effect on the representation of occupations and genders in the language models. It would therefore be interesting to investigate to what extent language-specific language models reflect census data from different countries.

Acknowledgements

This work was supported by industry partners and the Research Council of Norway (RCN) through funding to *MediaFutures: Research Centre for Responsible Media Technology and Innovation*, through the RCN's Centers for Research-based Innovation scheme, project number 309339.

References

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

Su Lin Blodgett. 2021. Sociolinguistically driven approaches for just natural language processing.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddingsa. *Advances in neural information processing systems*, 29.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter! debiasing GPT-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Per Egil Kummervold, Javier de la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proc. of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for norwegian. In *Proc. of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2018. Assessing gender bias in machine translation – a case study with google translate.

Danielle Saunders and Bill Byrne. 2020. Addressing exposure bias with document minimum risk training: Cambridge at the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 862–869, Online. Association for Computational Linguistics.

Samia Touileb. 2022. Exploring the effects of negation and grammatical tense on bias probes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 423–429, Online only. Association for Computational Linguistics.

Samia Touileb and Debora Nozza. 2022. Measuring harmful representations in Scandinavian language models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 118–125, Abu Dhabi, UAE. Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2020. Gender and sentiment, critics and authors: a dataset of Norwegian book reviews. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, Barcelona, Spain (Online). Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2021. Using gender- and polarity-informed models to investigate bias. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, Online. Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

A Appendix

Language	Predicates
Norwegian	jobber som jobbet som skal jobbe som vil jobbe som ville jobbe som kommer til å jobbe som
English	are going to work as is going to work as a want to work as wanted to work as wanted to work as a wants to work as a will work as will work as a work as worked as worked as a works as a
French	est étaient était sera serons sont va travailler comme veulent être veux être vont travailler comme voulaient être voulais être

Table 5: Language-specific predicates.