Directing the violence or admonishing it?
A survey of contronymy and androcentrism in Google Translate and some recommendations

Contents

Abstract

The recent raft of high-profile gaffes involving neural machine translation technologies has brought to light the unreliability of this evolving technology. A worrisome facet of the ubiquity of this technology is that it largely operates in a use-it-at-your-own-peril mode where the user is often unaware of either the idiosyncratic brittleness of the underlying neural translation model or when it is, that the translations be deemed trustworthy and when they wouldn't. These revelations have worryingly coincided with other developments such as the emergence of large language models that now produce biased and erroneous results, albeit with human-like fluency, the use of back-translation as a data-augmentation strategy in so termed 'low-resource' settings and the emergence of 'AI-enhanced legal-tech' as a panacea that promises 'disruptive democratization' of access to legal services. In the backdrop of these quandaries, we present this cautionary tale where we shed light on the specifics of the risks surrounding cavalier deployment of this technology by exploring two specific failings: Androcentrism and Enantiosemy. In this regard, we empirically investigate the fate of the pronouns and a list of contronyms when subjected to back-translation using Google Translate. Through this, we seek to highlight the prevalence of 'defaulting-to-the-masculine' phenomenon in the context of engendered profession-related translations and also empirically demonstrate the scale and nature of threats pertaining to contronymous phrases covering both current-affairs and legal issues. Based on these observations, we have collected a series of recommendations that constitute the latter half of this paper. All of the code and datasets generated in this paper have been open-sourced for the community to build on here: `https://github.com/rteehas/GT_study_recommendations`.

1 Introduction

Google Translate (GT) today, is arguably the most influential machine translation (MT) technology deployed in the real world. In March 2021, GT crossed the one billion installations mark on the Google Play Store and now purportedly serves 500 million monthly users to the tune of 140 billion words per day[1] across 109 different languages [1]. It has been dubbed "uncannily artful" in a New York Times magazine feature article, [2] pilloried as "shallow" by cognitive scientist Douglas Hofstadter [3] and marketed as "a personal interpreter in your pocket" by Google[2]. As observed in [4], one of GT's often missed and potent attributes has been *invisibilization* of the very act of translation. This is further reflected in [5], where the author remarks how "... Google Translate has gradually become less of an app that you install and more of an integrated experience throughout Google's ecosystem", and rightfully acknowledges the percolation of the underlying tech far beyond the standalone app or the GT portal accessible at `https://translate.google.com/`. GT now latently powers other applications such as as *Google sheets* and *Google lens*[3]. In Figure1, we present an example of a translation event (albeit biased) happening implicitly via Google Lens with GT playing *your friend when reading menus, street signs, and more*[6]. One might argue that the influence of this technology and trust invested by a common end user is also reflected in how it has been repeatedly used by hackers to launch malware (2012) [7, 8] and phishing attacks (2019) (See [9]).

1.1 The good: GT as an emergent language acquisition aid?

Besides playing the "*unlikely World Cup hero*" by breaking language barriers for sports fans [10], there is *some* evidence that when language instructors strategically harnessed [11] GT in their curricula, the tech has proven to be a useful pedagogical device speeding language acquisition amongst students. In [12], the researchers demonstrated that both syntactic complexity and accuracy was higher in a student group of Chilean high-school EFL learners (English as a Foreign Language) when given access to GT. Similarly, another study [13] concluded that the cohort members using GT as a revision tool, displayed improved writing in their L2 (second language) proficiency. The study involving Chinese sophomore, junior, and senior

[1] `https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html`

[2] `https://translate.google.com/intl/en/about/`

[3] `https://blog.google/products/translate/google-translates-instant-camera-translation-gets-upgrade/`

EFL students [14], which revealed that the students had *"found satisfaction with using Google Translate in their English writing, especially in finding vocabulary items and enhancing the completion of English writing."* A Case Study entailing 16 international students at the School of Languages, Literacies, and Translation, at Universiti Sains Malaysia [15] revealed that GT an effective supplementary tool for learning vocabulary, writing, and reading. Of particular note were three positive statements made by the students regarding their experience.
- *It is good for my self-learning even after the class ends;*
- *Using GT, I feel relaxed because I don't worry to ask the meaning of all words in class;*
- *GT is quite accessible everywhere so I feel the teacher is always next to me*

1.2 The bad: Lost in translation mishaps

On the flip side, Google Translate has also courted controversy over mistranslations in some high profile cases, especially in specialized domains such as law and medicine.
In the now infamous case of United States v. Omar Cruz-Zamora [4], a Kansas Highway Patrol Trooper used GT to translate *Can I **search** the car?* that resulted in the erroneous translation *Can I **find** the car (¿Puedo buscar el auto?)*, which was not the question intended and which in turn led to a wrongful extraction of consent, search and eventual imprisonment! In court, the judge ruled based on testimony from expert interpreters, stating: *Here, both professional interpreters testified Google Translate should only be used for literal word-for-word translations as Google Translate cannot take context into consideration. So, while it might be reasonable for an officer to use Google Translate to gather basic information such as the defendant's name or where the defendant was travelling, the court does not believe it is reasonable to rely on the service to obtain consent to an otherwise illegal search.*

In [16], clinicians assessed the potential harms in applying GT in emergency department discharge instructions. The authors translated 647 sentence-pairs, and showed while 594 (92% English-Spanish) and 522 (81% English-Chinese) were accurately translated, a substantial amount of mistranslated examples held clinically significant, and potentially ***life-threatening*** harm. When the source instruction in English (which forbade the patient from consuming any more medication) was: ***hold** the kidney medicine until you have a chance to speak with your kidney doctor*, the authors of the study demonstrated that in both Spanish and Chinese, GT flipped the interpretation of the crucial contronym *'hold'*, and provided translations which goaded the patients to ***keep taking** the kidney medicine until you talk to your kidney doctor*. This vector of vulnerability was further investigated in [17] in the context of translating 20 commonly used emergency department discharge instruction phrases, across Spanish, Chinese, Vietnamese, Tagalog, Korean, Armenian, and Farsi. The authors discovered shocking examples such as *You can take over the counter **ibuprofen** as needed for pain* being translated as *You may take **anti-tank missile** as much as you need for pain* in Armenian. Similarly, *Your **Coumadin** level was too high today. Do not take any more **Coumadin** until your doctor reviews the results.* in English was being translated as *Your **soybean** level was too high today. Do not take anymore **soybean** until your doctor reviews the results.* in Chinese. These results convinced the authors of this study to verbatim conclude that *GT for discharge instructions in the ED is inconsistent between languages and should not be relied on for patient instructions.*
In January 2021, it emerged that GT had translated *the vaccine is not **required*** as *the vaccine is not **necessary*** in Spanish on the Virginia Department of Health website's FAQ (frequently asked questions) page, thus miseducating the Hispanic users of the website who were already disproportionately impacted by the pandemic [18].
The repeated emergence of such high impact mistranslation gaffes (even for the so-termed high resource language-pairs such as English-Spanish and English-Chinese) can be traced back to this simple observation that this fledgling technology operates in a ***use-it-at-your-own-peril regime*** in the real-world, with no clear cut recommendations or insights into when it could be expected to work well, and when it couldn't. There also exists no easily accessible online official documentation on the metrics used or the metrics achieved by the deployed state-of-the-art model and one has to resort to investigations by members of academia to get insights into the nature, extent and vagaries of biases observed and harms enacted.

This work fits squarely into the audit-critique body of literature surrounding GT and the rest of the paper is organized as follows:

1.3 Paper organization

In Section-2, we cover the background with regards to specific aspects of evolution of NMT (Neural machine translation) with emphasis on the lasting contributions of Warren Weaver. In Section-3, we focus on the linguistic phenomenon of enantiosemy, which poses as a particularly

[4] `https://ecf.ksd.uscourts.gov/cgi-bin/show_public_doc?2017cr40100-24`

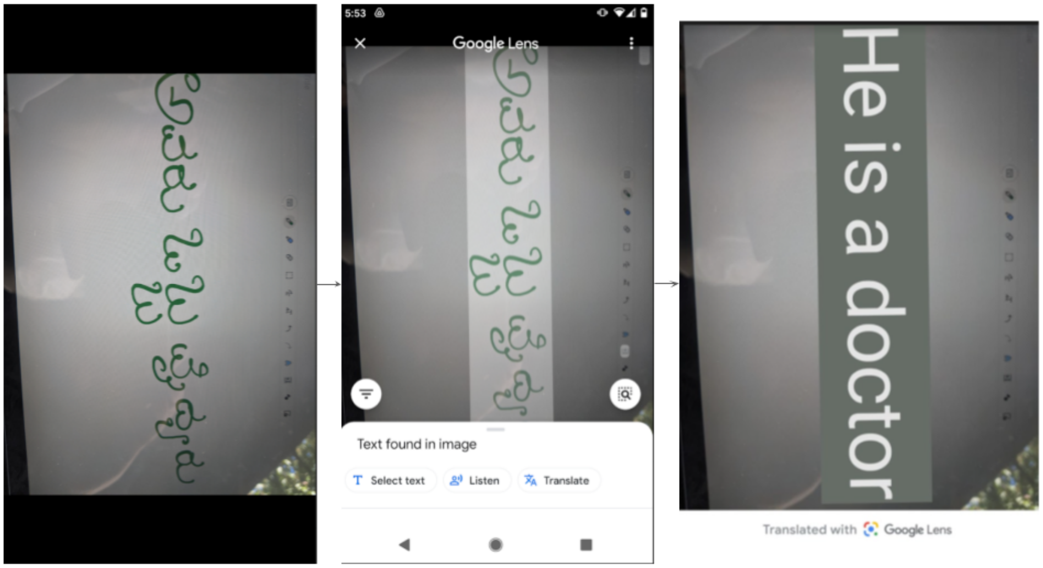


Figure 1: No longer a standalone app! An example of GT accessed implicitly via *Google Lens*

strong vector of inquiry targeting the word sense disambiguation ability of a NMT model. We provide the relevant background and present the empirical experiments results we obtained with GT. In Section-4, we dive into the world of gendered bias and androcentrism, discuss the historical usage in GT as well as other popular translation engines. We setup an experiment where we apply a new metric we term *She Survival Rate (SSR)* to interrogate back-translation errors over multiple language families. In Section-5, we've enumerated a set of recommendations which may aid in fixing the issues discovered in this paper and conclude the paper in Section-6.

2 Background

There are several excellent surveys covering various aspects of machine translation history. A name that one repeatedly encounters while parsing through the bibliography of MT surveys is that of William John Hutchins.[5] Three of his historical treatises: *Warren Weaver and the Launching of MT* [19], *Early years in machine translation: memoirs and biographies of pioneers* [20] and *ALPAC the (in) famous report* [21] provide the reader with a proverbial front seat to all the important events that unravelled in the early formative years of MT between July 1949 (date of authorship of Warren Weaver's Memorandum [22]) and the publishing of the ALPAC report in November 1966. His later survey works, titled *Machine translation: general overview* [23] (2003) and *Machine translation: general overview* [24] (2007), along with Slocum's *A survey of machine translation: Its history, current status and future prospects* [25] cover the important landmark events spanning the Rule Based Machine Translation (RBMT 1950-80), Example Based Machine Translation (EBMT 1980-1990) and the Statistical Machine Translation (SMT 1990-2014) eras.

With regards to the post SMT era, Philipp Koehn's books on Statistical Machine Translation [26] and Neural Machine Translation [27] constitute an excellent starting resource point. More specialized surveys[6] have emerged in the context of MT for Indian languages [28] (2017), Domain adaptation techniques used [29] (2018), Deep Learning (DL) techniques for NMT [30] (2020) and Multilingual NMT [31] (2020).

With this background, we now revisit a particular slice of machine translation history that takes place in the summer of 1949, concerning Dr. War-

[5] We gathered during the authoring of this paper that Dr. Hutchins recently passed away on 9 January 2021. Much of his life's work is still accessible via the Wayback Machine - Internet Archive link: `https://web.archive.org/web/20210102211028/http://www.hutchinsweb.me.uk/`

[6] We'd also like to acknowledge the survey materials curated for the `LING/TRST 415 - Machine Translation: History and Applications` course by the Computational Linguistics group at University of Illinois freely accessible via: `http://computational.linguistics.illinois.edu/LING415/Spring2017/schedule.html` and `https://www.youtube.com/channel/UCFnN8EryrdTq_qPcVl1VmCg`

ren Weaver, that has a strong influence on the rest of the work present here.

2.1 Warren Weaver

The towering figure of Warren Weaver looms large on the landscape of Machine Translation. His *memorandum* titled `Translation` published on July 15, 1949, is considered to be the "*single most influential publication in the earliest days of machine translation*" [19]. Wary of the limitations of the word-for-word translation approach, he laid out the following four proposals each of which have heavily shaped the landscape of ideas in MT literature.

1: Meaning and Context- The nascent N-gram-esque approach for word sense disambiguation : "*Thus one is led to the concept of a translation process in which, in determining meaning for a word, account is taken of the immediate ($2N$ word) context.*"

2: Language and Logic: In this approach Weaver postulated that translation could be cast as a problem of formal logic where the input (premise) would be fed in the source language and that translation would constitute the process of deducing the conclusion in the target language. In this regard, he was clearly inspired by the powerful suite of theorems proved in the influential treatise *A logical calculus of the ideas immanent in nervous activity* by Warren McCulloch and Walter Pitts [32] and states verbatim that "*But insofar as written language is an expression of logical character, this theorem assures one that the problem is at least formally solvable.*"

3: Translation and Cryptography: Warren Weaver's deep insights into Claude Shannon's work on the *Theory of Communication* and his associated contribution to cryptography goaded him to believe that the mathematical infrastructure laid towards characterizing the statistical characteristics of the communication process could be harnessed to achieve machine translation. This *translation-as-a-cryptography* problem framing has been somewhat immortalized by his famous quip that reads: "*It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code".*" Investigating further, this approach leads Weaver to propose his fourth, and arguably the most celebrated approach that advocates making deep use of language invariants, that elucidates in the section titled *"Language and Invariants"*.

4: Language and Invariants: Google's official blog dissemination unveiling their Massively Multilingual, Massive Neural Machine (M4) translation approach,[7] literally quotes Weaver's parable of linguistic universals that reads: "... *perhaps the way [of translation] is to descend, from each language, down to the common base of human communication — the real but as yet undiscovered universal language — and then re-emerge by whatever particular route is convenient.*" (In Table 1, we cover the hagiography of the 17^{th} century revival in the interest in the idea of the Universal Language beginning with Francis Bacon's *Advancement of Learning* (1605) to Voltaire's *Candide* (1759)).

One can argue that while this approach appears as the fourth proposal, the groundwork towards motivating this is in fact laid much earlier in the second section of his memorandum titled "*A War Anecdote - Language Invariants*" where he firstly states that "*whether Bantu or Greek, Icelandic or Peruvian*, humans have *essentially the same (vocal organs) equipment to bring to bear on this problem (of language development)* with brains of *of the same general order of potential complexity*", thus resulting in what he believes are wide superficial differences between the languages, whilst all of them sharing certain basic common building structures. Secondly, he harnesses Erwin Reifler's famous observation that the Chinese words for '*to shoot*' and '*to dismiss*' not only showed an uncanny phonological and graphic agreement but also mirrored the usage of the phrase "*to fire*" in English, whilst wondering "*Is this only happenstance? How widespread are such correlations?*".

We've authored Appendix A to capture an iconoclastic line of thought that emerged from certain skeptics hailing from psycho-sociolinguistic backgrounds who constantly grapple with non-portability of nuanced notions of gender and emotion between languages.

2.1.1 Countering the male hagiographies

It is important to note that parallel to the universal language pursuits captured in the male-exclusive hagiography of Table 1, was the emergence of an incredible wealth of translation scholarship led by women linguists. At this juncture, we deem it essential to juxtapose Table 1 with Table 2 where we celebrate some of the important contributions emanating from the period spanning from 1640s to 1900s [33].

In this vein, we'd also like to implore the reader to revisit works such as [34, 35, 36, 37, 38] that have critically analyzed and celebrated the valorous lives of indigenous cultural intermediaries such as *Malintzin* (La Malinche), *Amonute* (Poc-

[7] `https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html`

Scholar	Publications
Francis Bacon	Advancement of Learning (1605), De Augmentis Scientiarum (1623)
Jacob Boehme	The Mysterium Magnum (1623)
Marin Mersenne	letter to René Descartes (1629)
René Descartes	Letter to Mersenne (1629)
John Comenius	Via Lucis (1641)
Thomas Hobbes	Leviathan (1651), De Corpora (1655)
Isaac Newton	Unpublished notes (1661)
Gottfried Leibniz	De Arte Combinatoria (1666)
John Wilkins	An Essay towards a Real Character and a Philosophical Language (1668)
Jonathan Swift	Gulliver's Travels (1726)
Voltaire	Candide (1759)

Table 1: Hagiography of the 17th Century revival of the Universal Language. Source: `http://computational.linguistics.illinois.edu/LING415/Spring2017/slides/Universal_Language_in_the_17th_Century.pdf`

Linguist	Summary of contributions
Aphra Behn (1640-1689)	A pioneering French-to-English translator whose translated *A Discovery of New Worlds* by Bernard le Bovier de Fontenelle.
Anne Dacier (1654-1720)	Completed the gargantuan task of translating the works of Homer (The Iliad and The Odyssey) from ancient Greek to French.
Claudine Picardet (1735-1820)	A polymath translator who was also a chemist, a mineralogist, and a meteorologist. Is widely accredited with playing the leading role in ushering in the "chemical revolution" in France thanks to her translation efforts with scientific manuals and papers between Swedish, French, English, German, Italian, and Latin.
Lady Charlotte Guest (1812-1895)	A polyglot that spoke Arabic, Hebrew, English, Welsh, Latin, Greek, French, Italian and Persian
Mary Louise Booth (1831-1889)	Translated all of Count Agenor de Gasparin's "Uprising of a Great People" in less than a week and was the first editor-in-chief of Harper's Bazaar.
Constance Garnett (1861-1946)	The first to translate Russian novelists like Tolstoy, Dostoevsky, and Chekhov into English, and many of her translations are still in use today.

Table 2: Sourced from *Inspiring and Notable Women Throughout Translation History* by Kenzie Shofner [33]

ahontas), and *Sacajewea* (Sakakawea/Sacagawea) and contextualized their translation contributions through the lens of indigenous and Chicana feminism.

2.2 Background on the current GT model deployed

In October 2007[8], it was revealed [39] that Google had dropped `Systran` as the back end for translations and was now switching to a home-grown SMT system. As per [2], an internal mandate was set for GT to be overhauled to an NMT system by the end of 2016. In September 2016, the Google Neural Machine Translation system (GNMT) [40] was unveiled[9] marking a departure away from phrase-based SMT techniques[10]. The key features of this GNMT model were:

[8] `http://googlesystem.blogspot.com/2007/10/google-translate-switches-to-googles.html`

[9] `https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html`

[10] The reader is highly encouraged to sift through Section 7 (Theory Becomes Product) and Section 8 (A Celebration) of the *The Great A.I. Awakening* article [2] to get a deeper understanding of the background events that led to this moment

1. A deep LSTM architecture (8 encoder and 8 decoder layers) using attention and residual connections was used.
2. The use of attention mechanism to connect the bottom layer of the decoder to the top layer of the encoder that improved parallelism and led to faster training.
3. The use of low-precision arithmetic and Length-norm-and-coverage-penalty enhanced beam search inference.
4. To improve handling of rare words, words were divided into a limited set of common sub-word units ("wordpieces") for both input and output.
5. Use of the *Wordpiece* model approach (Section 4.1 in [40]) that provide accuracy improvements for translation of rare words.

This resulted in a sea change of *perceived translation quality* aspects of which were captured with much fanfare in the NY Times feature article titled *The Great A.I. Awakening* [2]. This GNMT model had only a 3 year reign, ending in June 2020 when it was revealed via blog-post titled *Recent Advances in Google Translate*[11] where they state *"we have replaced the original (RNN-based) GNMT system, instead training models with a transformer encoder and an RNN decoder, implemented in Lingvo (a TensorFlow framework)"*. This *M*assively *M*ultilingual, *M*assive neural *M*achine translation (*M4*) approach seems to be the current state-of-the-art, details of which are further covered in the upcoming subsection.

2.3 Post-GNMT era: M4 and GShard

There are two sources of information that provide insights into the NMT models that enable GT currently to serve across 109 languages. The first source pertains to the peer-reviewed research publications covering Gpipe [41], the M4 approaches [42, 43, 44, 45] and GShard [46, 1]. The second pertains to the six official blog posts (see Figure 2) posted on `https://ai.googleblog.com/`.

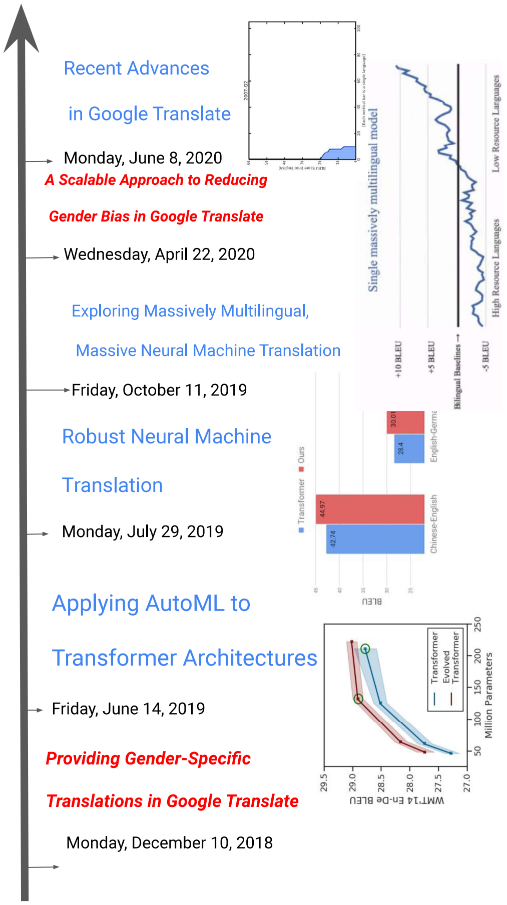


Figure 2: The official blog posts pertaining to Google Translate. The ones in italic pertain to the dissemination regarding gender bias

From what we could parse, the GShard-600B (619 B parameters) multilingual neural machine translation model is currently the publicly known SoTA model within the Google-NMT universe. As per [46, 1], this model with the sparsely gated mixture-of-experts (MoE) Transformer architecture with 36 Encoder-decoder layers (and 2048 Experts Per-layer) was trained with a dataset containing 1 trillion source side tokens (after sub-word segmentation) for 250k training epochs (steps) at the Google North Carolina data-center in April 2020 on 2048 TPU v3 chips for 3.1 days and consumed 24.1MWh while producing 4.3 net tCO2e during the training process. We note that the last of the NMT blog-posts, titled *Recent Advances in Google Translate* was published on June 8, 2020 which predates the GShard-600B model [46] ArXiv dissemination date of 30 Jun 2020.[12]

We note that besides the M4 modeling approach, other factors that have supposedly led to substantial BLEU gains were: Hybrid Model Architecture (Transformer encoder + RNN decoder), embedding-based web-crawler, sequential curriculum learning based training (begin with all data, and then gradually fine-tune on smaller and cleaner subsets scored using preliminary models)

[11] `https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html`

[12] `https://arxiv.org/abs/2006.16668`

and the last but not the least, the technique of *back-translation*. This observation provides for a neat transition into the next subsection, where we delve into the background details surrounding back-translation, its usage as a augmentation technique and as an instrument of inquiry.

2.4 Back-translation in linguistics and NLP

The term "Back-translation" in linguistics and NLP, appears under three different contexts:

1. Back-translation as an intermediate step in training MT models involving low-resource source languages.
2. Back-translation as an NLP data-augmentation technique used in text-classification tasks.
3. Back-translation as a translation quality assessment tool.

In this subsection, we will provide background regarding each of these three usages, and contextualize how and why we harness BT in our empirical analyses.

2.4.1 Back-translation as an intermediate step in training MT models

In recent years, within the larger ambit of *Deep learning driven Natural Language Processing (NLP)*, BT is increasingly seen as an important module in Statistical and Neural Machine Translation pipelines. In the context of phrase-based translations with monolingual data, [47] harnessed BT in what they term as the "reverse self-training" procedure that "*... improves the decoder's ability to produce grammatically correct translations into languages with morphology richer than the source language especially in small-data setting*". In the context of training Neural Machine Translation (NMT) models, [48] paired monolingual training data with synthetic back-translated data (as additional parallel training data) to achieve "*...substantial improvements on the WMT 15 task English<->German (+2.8-3.7 BLEU), and for the low-resourced IWSLT 14 task Turkish->English (+2.1-3.4 BLEU)*" as well. BT also features prominently in the context of recent advances in Unsupervised Machine Translation (UMT) models that are trained using only monolingual corpora. In [49], the authors motivated by the issue of *low-resource language pairs*, investigated whether it was possible to train a translation model without any parallel data by taking sentences from previously constructed monolingual corpora in two different languages and mapping them into the same latent space. They reported "*... BLEU scores of 32.8 and 15.1 on the Multi30k and WMT English-French datasets, without using even a single parallel sentence at training time*". Inspired by the above-mentioned works, researchers from Facebook AI Research and Google Brain, in their work titled *Understanding Back-Translation at Scale* [50] sought to *broaden the understanding of back-translation* and investigated a number of methods to generate synthetic source sentences and concluded that in all but resource poor settings, BT techniques (obtained via sampling or noised beam outputs) were in fact the most effective. Besides scaling their experiments to *hundreds of millions of monolingual sentences* that did result a new state of the art score of 35 BLEU on the WMT'14 English-German test set, they also demonstrated that "*... synthetic data can achieve up to 83% of the performance attainable with real bitext!*" These highly cited works have set the stage for BT being an indispensable technique in the context of achieving state-of-the-art scores (and worryingly so in our opinion) in what are deemed to be resource-poor or low-resource settings. In [51], the authors used the iterative BT technique [48] for data-augmentation to train an NMT model for the *Chibchan* language, *Bribri*, that achieved an average performance of BLEU 16.9±1.7 in spite of being trained on an extremely small dataset of 5923 Bribri-Spanish pairs. Similarly, [52] harnessed BT for NMT involving the *resource-poor* Lithuanian and Gujarati languages. Lastly, the authors in [53] investigated the effects of synthetic back-translated data for the *"low resource less related language pair"* of Chinese and Vietnamese.

2.4.2 BT as an NLP data-augmentation technique

Outside of training MT models, BT is being increasingly used as the data-augmentation strategy of choice for various downstream text-classification tasks, especially in settings with small training datasets. Whilst grappling with a small training dataset in a given language for a text classification task, the idea is to use a pre-trained MT model to translate into an intermediate (pivot) language and translate back to generate paraphrases of the training data, and train a model using the augmented dataset that has both the real training data and the paraphrased training data.

In March 2018, the winners of the "*Toxic Comment Classification Challenge*" on Kaggle[13] used BT as an augmentation trick to climb to the top of the leader-board. Terming their technique as *Train/test-Time Augmentation* (TTA), they used a pre-trained NMT model to augment their dataset using French, German, and Spanish translations

[13] `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557`

and then translate back to English. Intriguingly, they state: "... *a single TTA-ed model was beating the majority of teams' (presumably) ensembled submissions on the public ladder.*" Then, in April 2018, the QANET paper [54] arrived on the scene that proposed "Data Augmentation by Back-translation" in Section-3, and demonstrated how using translating and back-translating from French as a *pivotal* (intermediate) language, resulted in them attaining the SoTA (84.6 F1) score on the SQuAD dataset. Other noteworthy instances from 2018, BT-augmented training data being used to better downstream text classification performance include the textual style-transfer paper [55] and Claude Coulombe's *Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs*[56] where BT-augmented data generated using NLP cloud APIs was used to improve the accuracy of text polarity prediction.

In 2019, [57] back-translated and augmented data was used to achieve an 8.1% improvement over the augmentation-free baseline for the IMDB movie reviews classification task. Later, [58] used GT as the translation engine to back-translate and augment data (with German being the intermediate language) in the context of narrative analysis with news corpora. GT was also used in *Data Augmentation For Chinese Text Classification Using Back-Translation* [59], where the authors augment their training dataset using Chinese-English-Chinese paraphrased samples. It is to be noted that the authors used the inbuilt translate-backtranslation capability in Google sheets using

```
=GOOGLETRANSLATE(GOOGLETRANSLATE(A1
    , "zh-CN," "en"), "en," "zh-CN")
```

to perform training data augmentation. Recently, in [60], researchers from Amazon also used backtranslation-augmented dataset (Using English-German/German-English translation models [61]) and demonstrated that BT-augmentation provided very strong baselines and consistently outperformed several pre-trained data augmentation techniques for the SST-2 (Stanford Sentiment Treebank), SNIPS and TREC fine-grained question classification tasks.
Today, back-translation augmentation (also recently termed as *paraphrase augmentation* [62] and back-and-forth translation augmentation [63]) is available as an off-the-shelf module in popular Python NLP packages such as `NLPAug` [64] (that uses Facebook-AI's Fairseq [65] as the back end translation engine), `BackTranslation`-PyPi package (that offers Google Translate as well as Baidu Translation APIs as back end options) and Google's own UDA (Unsupervised Data Augmentation) [66] that performs English-French-English backtranslation augmentations using the WMT'14 English-French translation models[14]

2.4.3 BT as a QA tool

The early 1950s saw the emergence of BT as an invaluable tool for checking translation errors by comparing the input sentence with the back-translated sentence [67]. In the late 1960s, Richard Brislin, in *Back-Translation for Cross-Cultural Research* [68], carried out one of the earliest large-scale back-translation projects when with a team of "*[n]inety-four bilinguals from the University of Guam, representing ten languages, translated or back-translated six essays incorporating three content areas and two levels of difficulty.*" More recently, in the context of analyzing translation quality of NMT models in medicine, works such as [69, 17, 16] have all harnessed backtranslation as a quality assessment (QA) tool.

In this paper, we have harnessed BT as a QA tool in two experiments. The first experiment (in Section-3) involves analyzing the fate of the contronym *enjoin* upon completing the back-translation journey across the 108 languages supported on GT. The second experiment entails analyzing the fate of the pronoun *She* upon completing the back-translation journey across 108 languages while being used in sentences associated with 86 different professions. Our goal here is to use these preliminary results to inspire a new line of inquiry for NLP-bias research scholars and inspire caution in the wake of emergence of BT as a data-augmentation tool. In this regard, we'd like to explicitly emphasize that we are not championing the cause of BT as a definitive accuracy estimation tool to make some kind of a turgid claim regarding the fragility of GT. Further, we would like to invite the reader to pay heed to works such as [70, 71, 72] that have clearly demonstrated some of the shortcomings on back-translation based accuracy metrics.

3 Enantiosemy

Contronyms are words that elicit contradictory interpretations depending on the ***context*** of their usage. They are also called *Janus words, autoantonyms, antagonyms, antilogies, contranyms, amphibolous words, enantiodromic words, enantiodromes, fence-sitter words, opposonyms, pseudo-opposites, self-antonyms and self-contradicting words*. Examples of popular contronyms in colloquial usage include *overlook* (which can either mean to monitor or to fail to notice), *peruse* (which

[14] `https://github.com/google-research/uda/blob/master/README.md`

can either mean to read without attention to detail or examine in detail) and *screen* (which can either mean to conceal or to show). It has been speculated that inverse semantic processes entailing *semantic broadening*, *polarization of actants* and *idiosyncratic conflation of two hitherto-unrelated homographs* give rise to the emergence of contronyms in spoken language (See [73, 74]).

3.1 Study of contronyms

As per [75], the formal study of contronyms can be traced back to Carl Abel's *Uber den Gegensinn der Urworte*. It was published in 1884 during his study of Egyptian hieroglyphics, where he labelled the phenomenon as 'Gegensinn'[15]. Lederer in [76] postulates that it was Jack Herring, who in an article in the February 1962 issue of the now defunct *Word-study* magazine, introduced the specific term "contronym". The linguistic phenomenon that encompasses these contronyms is often termed enantiosemy or contronymy. Stefan Dubois in [77] viewed enantiosemy as a "*special form of polysemy wherein a lexeme has two directly opposing senses*", thus situating the phenomenon at the intersection of antonymy and polysemy.

As stated in [78], this phenomenon has especially attracted the attention of Slavic semanticists which has resulted in works such as [79, 80]. As evinced in the examples of online repositories such as [81, 82], contronyms take opposite meanings for reasons such as the persistence of archaic interpretations, native-versus-non-native usage differences, usage in *legalese* [83] as opposed to the colloquial usage, the American-British dichotomy, and also whether the word is being used as a verb or an adjective.

In the context of legal usage, works such as [84, 78, 85] have called for extreme caution and care to be deployed while translating and interpreting legal documents into other languages. In Figure 3, we present real world examples of mistranslations from the United States Department of Justice website involving the legalese[16] contronym *enjoin* being translated into Spanish (which is deemed to be a 'high-resource' language).

3.2 Why contronyms are not a "corner-case-gotcha"

To those who champion the cause of devil's advocacy and ask "*Why should one bother to study the fate of contronyms upon translating?*" or "*Are we trying to manufacture a corner-case gotcha to make the technology **look bad**?*", we present the following sets of arguments. The first argues that contronyms were, in fact, used to highlight the challenges that MT would face at **the very birth** of the modern post-1949-MT era by the very *pioneer* of the field. The second highlights the visceral impact of flipping the interpretation of a translated sentence involving these fence-sitting [75] words, especially upon making small adversarial input perturbations.

3.2.1 On Warren Weaver's 'fast' tryst with enantiosemy

The contronym *fast* finds a special place in Weaver's *Translation* [22] (See 2.1) in the section titled **Meaning and Context** where he uses the word to highlight the word-sense disambiguation challenges that MT efforts will come to grapple with. Specifically, he states: "*If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. **'Fast'** may mean 'rapid'; or it may mean 'motionless'; and there is no way of telling which*". This specific contronym *fast*, has since appeared in contronym compendia such as [87, 88] and is also used as the first exemplar word in Merriam-Webster's word history page on *Janus words*[17].

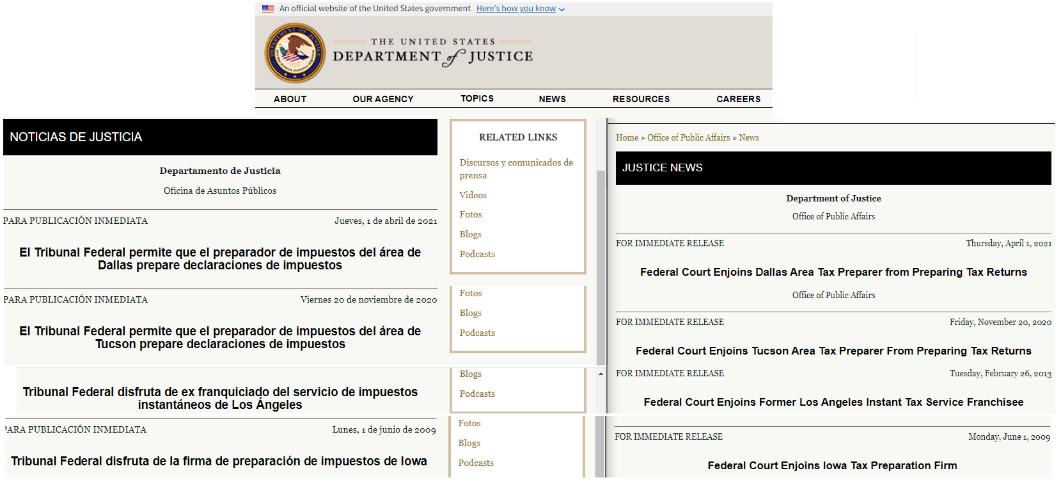
3.2.2 The Opposite meaning, not a slightly altered one, might survive

Consider the sentence *"The IOC sanctioned the guilty athletes"*. The presence of ***IOC*** (International Olympic Committee) and ***guilty*** help disambiguate the contronym *sanction*, which in this context means debar, ban or punish (as opposed to *approve*, which is the other interpretation). This is mistranslated by GT into Hindi to be: आईओसी ने दोषी एथलीटों को मंजूरी दी। which reads as *"The IOC **approved** the convicted athletes"*, thereby flipping the fence-sitter *sanction* to the wrong side of the fence. Thus, we note that mistranslating a sentence with contronyms carries a pronounced risk of the ***opposite*** and not merely a *slightly altered* meaning surviving the translation, which can have stark implications in domains such as medicine

[15] [75] also states the possibility of earlier discovery by way of Matthias Kramer in the 17^{th} century and later by Goethe in the 18^{th} century.

[16] Legalese or Legal English "... *has traditionally been a special variety of English. Mysterious in form and expression, it is larded with law-Latin and Norman-French, heavily dependent on the past, and unashamedly archaic"* [86]

[17] See `https://www.merriam-webster.com/words-at-play/words-own-opposites`



An official website of the United States government Here's how you know

THE UNITED STATES DEPARTMENT of JUSTICE

NOTICIAS DE JUSTICIA

Departamento de Justicia

Oficina de Asuntos Públicos

PARA PUBLICACIÓN INMEDIATA — Jueves, 1 de abril de 2021

El Tribunal Federal permite que el preparador de impuestos del área de Dallas prepare declaraciones de impuestos

PARA PUBLICACIÓN INMEDIATA — Viernes 20 de noviembre de 2020

El Tribunal Federal permite que el preparador de impuestos del área de Tucson prepare declaraciones de impuestos

Tribunal Federal disfruta de ex franquiciado del servicio de impuestos instantáneos de Los Ángeles

PARA PUBLICACIÓN INMEDIATA — Lunes, 1 de junio de 2009

Tribunal Federal disfruta de la firma de preparación de impuestos de Iowa

Home » Office of Public Affairs » News

JUSTICE NEWS

Department of Justice

Office of Public Affairs

FOR IMMEDIATE RELEASE — Thursday, April 1, 2021

Federal Court Enjoins Dallas Area Tax Preparer from Preparing Tax Returns

Office of Public Affairs

FOR IMMEDIATE RELEASE — Friday, November 20, 2020

Federal Court Enjoins Tucson Area Tax Preparer From Preparing Tax Returns

FOR IMMEDIATE RELEASE — Tuesday, February 26, 2013

Federal Court Enjoins Former Los Angeles Instant Tax Service Franchisee

FOR IMMEDIATE RELEASE — Monday, June 1, 2009

Federal Court Enjoins Iowa Tax Preparation Firm

Figure 3: Real world examples of mistranslations entailing the word *enjoin* from the United States Department of Justice website.

and legal-work. As will be further explored in Sec 3.4, when we throw adversarial vulnerability of NMT models into the mix, it gets even more dangerous as merely changing the case of the words or adding a benign exclamation mark can sometimes flip the meaning of the translated sentence.

3.3 Experiments

In this subsection, we carry out both latitudinal (across all languages) and longitudinal (with specific intermediate languages) exercises to highlight the nature and extent of risks that emerge in the context of translating contronymous sentences. Here, we design two experiments. The first is a depth-wise longitudinal experiment that focuses on a specific language and uses a hand-curated list of sentences that involve a set of contronyms. The second experiment is breadth-wise latitudinal exploration across all the 109 languages using a specific sentence: "*The court enjoined the violence!*". We hope the specifics and the extent of the shortcomings demonstrated in this section will help justify the tone of caution-and-skepticism with regards to NMT that has been the main theme in this paper.

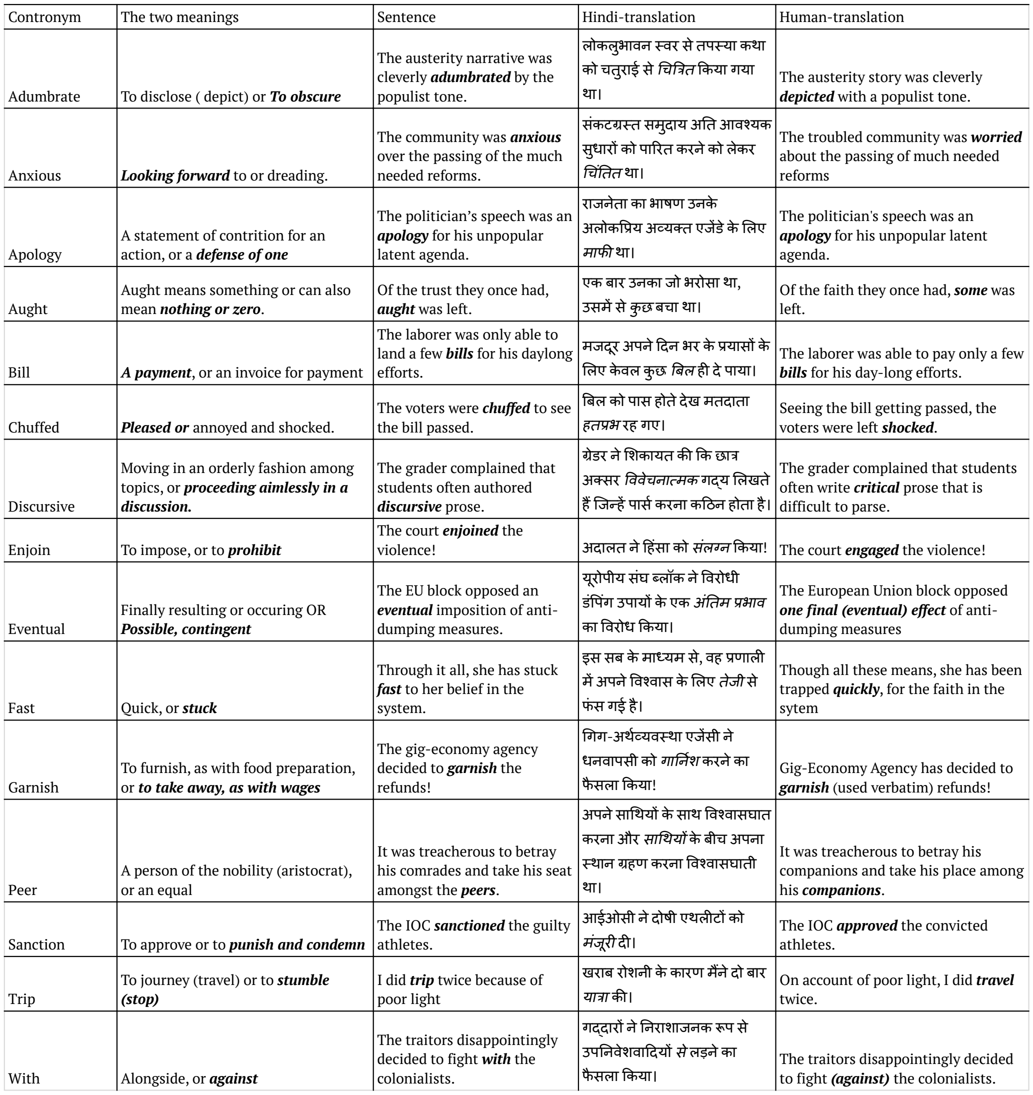
3.3.1 Longitudinal experiment with contronymous sentences

In this subsection, we present a longitudinal exploration where we track the fate of a **set of sentences** with contronymous components while fixing the target-language to be Hindi.

To begin with, we created a list of following contronyms from sources such as [81, 82]: *Adumbrate, Anxious, Apology, Aught* , *Bill* , *chuffed, Discursive, Enjoin, Eventual Fast, Garnish, Peer, Sanction, Trip*, and *With*. For each of the contronyms in this list, we generated a news-like exemplar sentence that one might find in an article about governance or legislation. We strategically introduced key words into each of these sentences so as to disambiguate between the two meanings of the included contronyms. The goal of this exercise is to not only make the reader aware of the idiosyncratic shortcomings of translating into the Hindi language but to demonstrate the kind of misinformation that these technical frailties could unleash, especially in the context of news-consumption in the global south.

The results of this experiment are presented in Table 4, which presents the contronym used, its two plausible meanings, the news-like sentence we generated, its translation into Hindi (the italicized word), and finally the human generated back-translation of the Hindi-translated sentence. Here, the italicized word in the column titled "The two meanings" indicates which of the two interpretations was chosen in the sentence generated. Similarly, with regards to the column "Hindi-translation", the italicized word indicates the contronyms translation into Hindi.

It is important to note that while some of the mistranslations were straight forward (Ex: *"The austerity narrative was cleverly adumbrated by the populist tone"* getting translated as लोकलुभावन स्वर से तपस्या कथा को चतुराई से चित्रित कयागया था), some entailed translations of the contronymous word in the form of loan words. For example, the sentence *"The laborer was only able to land a few bills for his daylong efforts"* was translated as मजदूर अपने दिन भर के प्रयासों के लिए केवल कुछ बिल ही दे पाया। where the contronym



Contronym	The two meanings	Sentence	Hindi-translation	Human-translation
Adumbrate	To disclose (depict) or ***To obscure***	The austerity narrative was cleverly ***adumbrated*** by the populist tone.	लोकलुभावन स्वर से तपस्या कथा को चतुराई से *चित्रित* किया गया था।	The austerity story was cleverly ***depicted*** with a populist tone.
Anxious	***Looking forward*** to or dreading.	The community was ***anxious*** over the passing of the much needed reforms.	संकटग्रस्त समुदाय अति आवश्यक सुधारों को पारित करने को लेकर *चिंतित* था।	The troubled community was ***worried*** about the passing of much needed reforms
Apology	A statement of contrition for an action, or a ***defense of one***	The politician's speech was an ***apology*** for his unpopular latent agenda.	राजनेता का भाषण उनके अलोकप्रिय अव्यक्त एजेंडे के लिए *माफी* था।	The politician's speech was an ***apology*** for his unpopular latent agenda.
Aught	Aught means something or can also mean ***nothing or zero***.	Of the trust they once had, ***aught*** was left.	एक बार उनका जो भरोसा था, उसमें से *कुछ* बचा था।	Of the faith they once had, ***some*** was left.
Bill	***A payment***, or an invoice for payment	The laborer was only able to land a few ***bills*** for his daylong efforts.	मजदूर अपने दिन भर के प्रयासों के लिए केवल कुछ *बिल* ही दे पाया।	The laborer was able to pay only a few ***bills*** for his day-long efforts.
Chuffed	***Pleased or*** annoyed and shocked.	The voters were ***chuffed*** to see the bill passed.	बिल को पास होते देख मतदाता *हतप्रभ* रह गए।	Seeing the bill getting passed, the voters were left ***shocked***.
Discursive	Moving in an orderly fashion among topics, or ***proceeding aimlessly in a discussion.***	The grader complained that students often authored ***discursive*** prose.	ग्रेडर ने शिकायत की कि छात्र अक्सर *विवेचनात्मक* गद्य लिखते हैं जिन्हें पार्स करना कठिन होता है।	The grader complained that students often write ***critical*** prose that is difficult to parse.
Enjoin	To impose, or to ***prohibit***	The court ***enjoined*** the violence!	अदालत ने हिंसा को *संलग्न* किया!	The court ***engaged*** the violence!
Eventual	Finally resulting or occuring OR ***Possible, contingent***	The EU block opposed an ***eventual*** imposition of anti-dumping measures.	यूरोपीय संघ ब्लॉक ने विरोधी डंपिंग उपायों के एक *अंतिम प्रभाव* का विरोध किया।	The European Union block opposed ***one final (eventual) effect*** of anti-dumping measures
Fast	Quick, or ***stuck***	Through it all, she has stuck ***fast*** to her belief in the system.	इस सब के माध्यम से, वह प्रणाली में अपने विश्वास के लिए *तेजी* से फंस गई है।	Though all these means, she has been trapped ***quickly***, for the faith in the sytem
Garnish	To furnish, as with food preparation, or ***to take away, as with wages***	The gig-economy agency decided to ***garnish*** the refunds!	गिग-अर्थव्यवस्था एजेंसी ने धनवापसी को *गार्निश* करने का फैसला किया!	Gig-Economy Agency has decided to ***garnish*** (used verbatim) refunds!
Peer	A person of the nobility (aristocrat), or an equal	It was treacherous to betray his comrades and take his seat amongst the ***peers***.	अपने साथियों के साथ विश्वासघात करना और *साथियों* के बीच अपना स्थान ग्रहण करना विश्वासघाती था।	It was treacherous to betray his companions and take his place among his ***companions***.
Sanction	To approve or to ***punish and condemn***	The IOC ***sanctioned*** the guilty athletes.	आईओसी ने दोषी एथलीटों को *मंजूरी* दी।	The IOC ***approved*** the convicted athletes.
Trip	To journey (travel) or to ***stumble (stop)***	I did ***trip*** twice because of poor light	खराब रोशनी के कारण मैंने दो बार *यात्रा* की।	On account of poor light, I did ***travel*** twice.
With	Alongside, or ***against***	The traitors disappointingly decided to fight ***with*** the colonialists.	गद्दारों ने निराशाजनक रूप से उपनिवेशवादियों *से* लड़ने का फैसला किया।	The traitors disappointingly decided to fight ***(against)*** the colonialists.

Figure 4: Table demonstrating the fate of contronymous sentences with Hindi as the target language.

bill was just transliterated as बिल, a loan word that has the same interpretation as an invoice for payment.

'Whyness' of this qualitative effort:
A cursory glance at the mistranslations reveals a worrisome consistency with which the incorrect interpretation of a contronym is repeatedly chosen. We hope this might motivate NMT researchers to focus on enantiosemy as an important testing instrument for assessing claims of achieving enhanced *contextuality* in contextual text representations [89]. The deeper goal however, behind presenting these qualitative results is to implore the reader to reflect on the *fluency* and the *seeming coherence* of the mistranslated sentences. Parsing through many of the mistranslations in Table 4, a time-traveling GT lay user from even 3 years ago one would be hard pressed to believe that these were machine generated sentences. This observation directly reinforces Section 6.2 of [90] where the authors state: *"However, MT systems can (and frequently do) produce output that is inaccurate yet both fluent and (again, seemingly) coherent in its own right to a consumer who either doesn't see the source text or cannot understand the source text on their own. When such consumers therefore mistake the meaning attributed to the MT output as the actual communicative intent of the original text's author, real-world harm can ensue".* As one might fathom, this new vector of threat where the error-prone brittle NMT model is now endowed with the ability of generating erroneous sentences that bear deft verisimilitude of being human-generated is certainly the deeper worry being presented and explored here.

3.3.2 Latitudinal experiments: The fate of enjoin

Certain use cases for contronyms can magnify the importance of accurate translations. One important subset of contronyms are those with specialized uses, such as in law. An inaccurate back-translation for these words may have an outsized effect as word choice in these contexts tends to be highly deliberate, with technical consequences for choosing incorrectly [91, 92]. In the legal context, straying beyond normal legal language can impact the strength of a contract or court ruling and, from the perspective of one being legally bound, improper translation can cause unintentional violations of the law. As a result, fidelity to the text is a primary concern [93]. Outside of their strict legal usage, many of these terms also appear in news headlines. Improper translations can influence the perceptions of those unable to speak the language and shape English-language coverage and analysis [94].

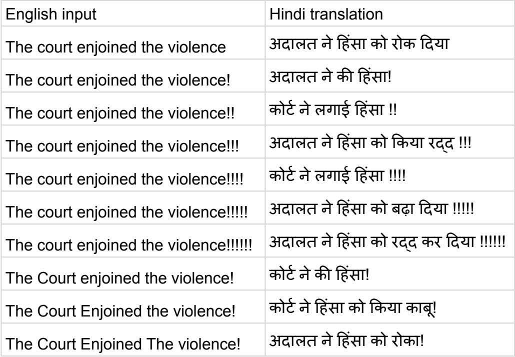
The contronym "Enjoined", can mean either prescribing an action or prohibiting it, and is often used in legal documentation and court rulings, as well as subsequent reporting on them. Inspired by the observations in Figure 3, we conducted the following study where we translate the sentence "The court enjoined the violence!" into all 109 languages available through the Google Translate API, back-translate into English and assess which of the two interpretations survived. Specifically, we categorized the back-translated results into those where the back-translation orders or prescribes violence, those where the back-translation prohibits it, and oddity-ridden mistranslations.

Type of Translation	Count
Prescribing Violence	88
Prohibiting Violence	10
Miscellaneous/Incorrect	11

Table 3: Translations of "The court enjoined the violence"

We report the results across all 109 languages, including English, in Table 3 (a table of the translations is included in Table 30). Overall, there were 88 languages where the back-translation prescribed violence and only 10 that prohibited it. Only the English "back-translation" (included in the prohibition group) included "enjoined". In the non-mistranslated cases, one of the two contronymous meanings were chosen, and those choices skewed significantly towards prescribing violence. Among the back-translations that prohibited violence were Chinese, both with simplified and traditional characters, as well as Hindi, Irish, Pashto, Japanese, Turkish, and Ukrainian, which belies the belief that an obvious commonality can predict proper translation. Interestingly, some linguistically similar languages were back-translated to different meanings. For example, within the Iranian language family, (which is a subgroup of Indo-Iranian languages that fall under the larger category of Indo-European languages), Pashto and Persian were backtranslated differently [95]. Similarly, Russian and Ukrainian, both of which are Slavic languages (specifically East Slavic) had divergent backtranslations [96]. Importantly, each of these language pairs share a common script – Perso-Arabic for Pashto and Persian [95] and Cyrillic for Ukrainian and Russian. This phenomenon did not just appear in specific linguistic subgroups with similar scripts, however, but also across larger language groups. Polish, for example, is also a Slavic language (within the West Slavic branch) [96], was also backtranslated differently than Ukrainian. Identifying the reasons for this divergence among similar languages is an open area for further research. Also, we observed that the Lao back-translation was "*Court of Violence!*" and the Thai back-translation reads "*Police charge violence!*", which stand out as glaring outliers, worthy of further research. As with the languages that back-translated to prohibition, we posit that the languages associated with mistranslation do not have any obvious similarities, aside from the fact that we can speculate about the relative paucity of data for some of them compared to some of the other languages.

3.4 Adversarial vulnerability: The fence-sitters flip over!



English input	Hindi translation
The court enjoined the violence	अदालत ने हिंसा को रोक दिया
The court enjoined the violence!	अदालत ने की हिंसा!
The court enjoined the violence!!	कोर्ट ने लगाई हिंसा !!
The court enjoined the violence!!!	अदालत ने हिंसा को किया रद्द !!!
The court enjoined the violence!!!!	कोर्ट ने लगाई हिंसा !!!!
The court enjoined the violence!!!!!	अदालत ने हिंसा को बढ़ा दिया !!!!!
The court enjoined the violence!!!!!!	अदालत ने हिंसा को रद्द कर दिया !!!!!!
The Court enjoined the violence!	कोर्ट ने की हिंसा!
The Court Enjoined the violence!	कोर्ट ने हिंसा को किया काबू!
The Court Enjoined The violence!	अदालत ने हिंसा को रोका!

Figure 5: Adversarial vulnerability using English-to-Hindi examples in GT

Works such as [97, 98, 99, 100, 101, 102] have studied in detail the adversarial vulnerability of NMT models. In the case of contronymous inputs, it is natural to wonder if small input perturbations can trigger the NMT model's output to not just exhibit non-trivial changes, but to in fact ***switch from one interpretation to the opposite meaning***. Much to our dismay, with regards to GT, we found it was trivial to do so and an example of this is captured in Figure 5. As in the sub-section above, we use the sentence *The court enjoined the violence!* as input which GT translates as अदालत ने की हिंसा! in Hindi (which reads as *The court committed violence!* in Hindi). With the following single character changes to the input, we observed that:

1. Upon removing the exclamation mark from the input, the GT output changes to: अदालत ने हिंसा को रोक दिया in Hindi (which translates as *The court **stopped** the violence*).
2. Replacing '!' with the ellipsis '...' results in कोर्ट ने दी हिंसा की सजा ... दिया which reads as *The court delivered the punishment for the violence ...*
3. Beginning the input with a lower case 't' resulted in अदालत ने हिंसा को रोक दिया ! in Hindi (which translates as *The court **stopped** the violence!*)
4. Introducing an extra exclamation mark '!!' in the place of '!' in the input resulted in कोर्ट ने लगाई हिंसा !! (which translates as *The court **applied** the violence!!*)
5. Introducing three exclamation marks '!!!' results in अदालत ने हिंसा को किया रद्द !!! (which translates as *The court cancelled [or annulled] the violence!!*)
6. Changing the case of the first letter of the words resulted in dramatic changes in the output's interpretation as well (Also see related work to change-of-case perturbations in [103, 104]).

We emphasize that this last point is particularly worrisome on account of two reasons. The first reason is that there is no mention of GT's case-sensitivity on either the website, or, to the best of our knowledge, in the available developer documentation. Secondly, as revealed in Section-7 of [2]: *" Many people, Google had found, don't look to the service to translate full, complex sentences; they translate weird little shards of language"*, that not only lends credence and real-world validity to these phrase-level studies presented in the section but also highlights the prevalence of exposure to these vulnerabilities in terms of real-world usage.

4 Gender bias and androcentrism

At the heart of this ongoing *Deep learning-NLP-NMT* nexus, lies an indifference to the knowledge generated by the linguists and psycholinguists alike, who have studied the gender-language nexus in great detail. We argue that being informed with about how various languages treat gender provides a good platform to understand some of the extreme inter-language bias variations we observed.

Further, [105] describe how amongst the languages with grammatical gender systems[18], there was repeated encountering of potential bias in favor of the masculine forms while concluding that *"we are not aware of any European language with a similar potential feminine bias"*. In this section, we present an investigation into the nature of androcentric bias reflected via pronouns in GT and begin with the brief overview of the relevant background literature.

[18] Specifically, Chinese, Czech, Danish, Dutch, English, French, German, Italian, Norwegian, Polish, Romanian, Russian, Slovak, Spanish, and Swiss German

4.1 The world of gendered languages

In the ensuing subsections, we will see that the gender bias measured via our experiments with GT varied widely between languages. In order to present the reader with some foundational knowledge that might help shine some light on the *whyness* of these variations, we present a quick primer on categorization of languages based on how they address and accommodate gender.

Broadly speaking, languages have been categorized into *grammatical gender languages*, *natural gender languages*, and *genderless languages*. But linguistics scholarship from sources such as [106, 107, 105, 108] has produced more granular and nuanced categorization of languages, which we now present.

We begin with the Guiora Scale [107] proposed in 1983 that groups languages into four categories based on *gender loading* as follows:

1. Zero gender loading (such as Hungarian)
2. Low gender loading (such as English)
3. High gender loading (such as Spanish and German)
4. Very high gender loading (such as Hebrew)

Gygax et al. [105] in 2019 introduced ***a language index*** that grouped languages into five categories:

1. Grammatical gender languages: French, Spanish,Czech and German (See [106])
2. Languages with a combination of *grammatical gender* and *natural gender*: Norwegian, Dutch
3. Natural gender languages: English
4. Genderless languages with a few traces of natural gender: Oriya, Basque
5. Genderless languages: Turkish and Finnish

Thirdly, we cover the World Atlas of Language Structures (WALS) [108] that has emerged as the largest known academic database encompassing phonological, grammatical and lexical properties of languages gathered from descriptive materials (such as reference grammars) by an international team of linguists entailing 55 authors covering 257 languages, 172 language-genera and 90 language families. Chapters 30 - *Number of Genders* [109], 31 - *Sex-based and Non-sex-based Gender Systems* [110] and 32 *Systems of Gender Assignment* [111] of WALS (authored by Greville G. Corbett) are entirely dedicated to studying the nexus between Gender and Language. In figure 14, we see the representative maps of these chapters and the histogram-counts of these gender characteristics across the 257 languages covered in the atlas. We note that Google research effort such as [112] and [113] have previously harnessed WALS in the context of predicting the features of language structures from speech and evaluating cross-lingual generalization.

These studies of the vagaries of gender loading in languages is important as it informs the nature of biases we see in the forthcoming sections. To this end, we introduce the reader to the words of Levi Hord, a transgender studies and queer theory scholar, that read: *"I argue that gendered languages have less linguistic "room" for neutral language – by which I mean less space, opportunity, ease and susceptibility to its development – than "natural" gender languages do, based on the amount of grammatically gendered conventions in the language. I also argue that this difference impacts the lived experiences of transgender individuals who speak these languages, and that it should thus become a framework through which we view issues of transgender representation and activism."* (from [114])

4.1.1 Related academic work on gender bias analysis on Google Translate

The gender bias in Google Translate has been a focus of many recent studies. To begin with, the authors in [115] began with a list of professions from the U.S. Bureau of Labor Statistics (BLS) and constructed sentence templates of the form "`He/She is a/an` *`professional`*" across 12 different gender neutral languages such as Hungarian, Chinese and Yoruba. Then, they translated the template sentences into English using the Google Translate API, and by analyzing the translated sentences, demonstrated that Google Translate did exhibit a strong tendency towards male defaults. In [116], the authors focused on Korean, a language with gender neutral pronouns, and generated a sentence set of size 4236 sentences to demonstrate gender bias across GT, the *Naver Papago* and *Kakao* Translator APIs by using a metric, they termed as *translation gender bias index* (TGBI). In [117], carried out English-to-Spanish, English-to-Italian and English-to-French translations across Google Translate, DeepL and Bing Microsoft Translators to establish the prevalence of gender bias. In 2020, Frakas and Nemeth [118] focused primarily on Hungarian, a language with gender-neutral pronouns, and experimented across 742 occupations (based on the *Hungarian Standard Classification of Occupations* - FEOR - classification). They discovered bias against the male and the female genders but also established that biased results against women were much more frequent. Also, AlgorithmWatch, a non-profit research and advocacy organization published a scathing report titled *"Female historians and male nurses do not exist, Google Translate tells its European users"* [119]. The report analyzed 440 translation pairs across 11 occupations to and from German, Italian, Polish, Spanish and French and found biased translations in 182 of the 440 cases. Further, the report also revealed that 68 of these 182 translations were marked as "verified" by Google.

4.1.2 Acknowledgement of gender bias in GT

There are two blog-posts that address the gender bias issue. The first of which, appeared on Dec 10,2018, is titled "Providing Gender-Specific Translations in Google Translate" announced that GT would henceforth provide both *"feminine and masculine translations when translating single-word queries from English to four different languages (French, Italian, Portuguese, and Spanish), and when translating phrases and sentences from Turkish to English"*. It was revealed that a three-step approach was used that consisted of a 'triggering' pre-translation convolutional neural network that would now classify which queries would merit gender-specific translations, a back end NMT for generating gender-specific translations and a third module to check for accuracy (See Figure 6). It is to be noted that [119] states that *"However, it is unclear whether such efforts were made in earnest"* and characterized this feature as ***Window dressing***.

The second, titled "*A Scalable Approach to Reducing Gender Bias in Google Translate*" dated April 22, 2020, revealed that the 3-step approach in Figure 6 was not deemed to be scalable as the triggering classifier training was found to be too data intensive and the system suffered from low recall. It proposed a new *rewriting-based method* using a one-layer transformer-based sequence-to-sequence model trained on *millions of training examples composed of pairs of phrases, each of which included both masculine and feminine translations*. The blog claimed that this new approach resulted in a bias reduction of ≥90% for translations from Hungarian, Finnish and Persian-to-English and that the existing Turkish-to-English system improved from 60% to 95%. In Figure 7, we see the plots from the blog post juxtaposed with the contents of a recent viral tweet that highlighted the experiential bias of the user, that highlights that much work still needs to be done in this regard.

4.2 Experiments

One of the important recurring themes of AI-skepticism has been the fear of large-scale automated continuation, rebirth and reproduction of societal toxicities and archaic norms that risks undoing the progress made by decades of activism [120, 121, 122]. One specific form of toxicity pertains to the ***male-as-norm*** principle that besieges modern language and one that "strengthens the perceptions that the male category is the norm and that the corresponding female category is a derivation and thus less important" (See 123, 124, 125, 126). As noted in [127], this manifests as defaulting to the masculine pronoun in automated-translation systems.

4.2.1 Using Hindi as a motivation point

Hindi is a *gendered* language and as stated in [128], one needs to carefully navigate the challenges that might arise from *defaulting to masculine in mixed-gender situations, or situations where gender is unknown* . Google Translate translates the sentence *She is a doctor* as वह एक डॉक्टर है imbibing the usage of the gender-neutral third person distal formal word वह [19]. However, when this is back-translated sentence into English, GT defaults to the masculine and the translation reads *He is a doctor*! As in [119], we note the fact that this translation is accompanied by a `Verified : Translation verified by Google Translate contributors` icon (See Fig.8) that only worsens things.

In order to get an estimate of the extent of androcentrism in this neural translation technology, we performed an experiment using a dataset of sentences pertaining to 86 different professions across the 109 languages on offer, the details of which are presented in the forthcoming subsection.

4.2.2 Dataset curation and experiment details

Our experimentation detailed here explores the nexus between profession and gender in the context of translation-tech and was informed by sociolinguistic scholarship such as Celia Davies' treatise on *The Sociology of Professions and the Profession of Gender* [131] and Tracey L.

[19] There are deeper socio-cultural connotations of this specific word in Hindi and Urdu. This is deeply explored in 'Woh' , that is a gut-wrenching tale of a nameless subaltern protagonist with a disfigured face, who happens to be referred to by the others in disgust as *woh* or "that one", authored Dr. Rashid Jahan (1905-1952), a sterling icon of Indian literary radicalism (See [129], [130])

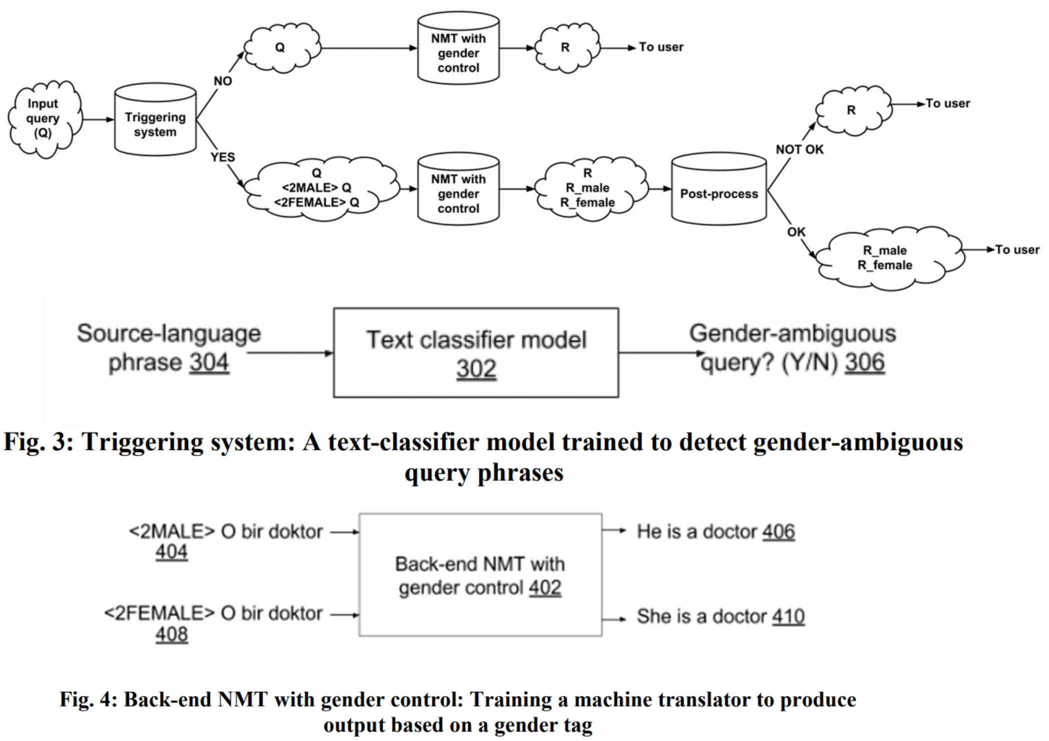
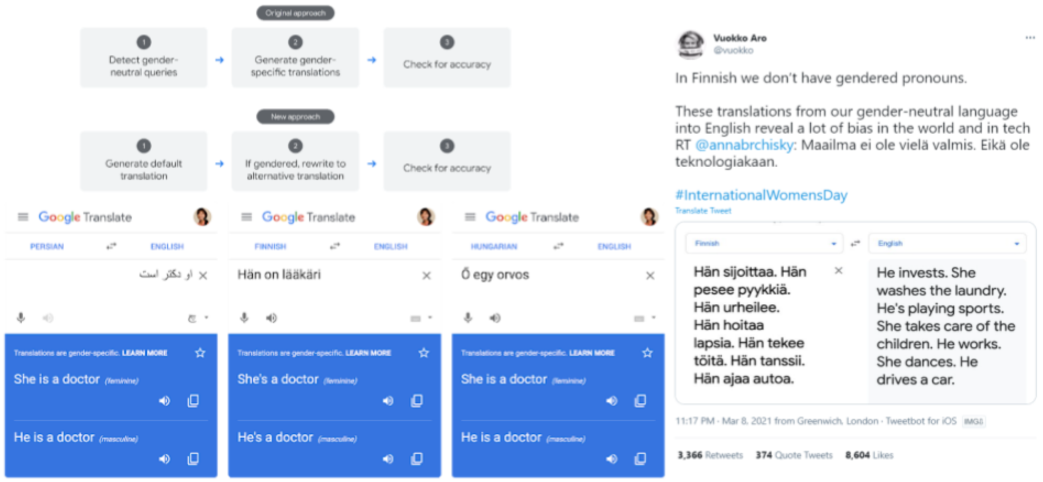


Fig. 3: Triggering system: A text-classifier model trained to detect gender-ambiguous query phrases

<2MALE> O bir doktor 404
<2FEMALE> O bir doktor 408
Back-end NMT with gender control 402
He is a doctor 406
She is a doctor 410

Fig. 4: Back-end NMT with gender control: Training a machine translator to produce output based on a gender tag

Figure 6: The three-step approach for gender-specific translations. Source: `https://www.tdcommons.org/dpubs_series/1577/`



او دکتر است

Translations are gender-specific. LEARN MORE

She is a doctor (feminine)

He is a doctor (masculine)

Hän on lääkäri

Translations are gender-specific. LEARN MORE

She's a doctor (feminine)

He's a doctor (masculine)

Ő egy orvos

Translations are gender-specific. LEARN MORE

She is a doctor (feminine)

He is a doctor (masculine)

Vuokko Aro
@vuokko

In Finnish we don't have gendered pronouns.

These translations from our gender-neutral language into English reveal a lot of bias in the world and in tech
RT @annabrchisky: Maailma ei ole vielä valmis. Eikä ole teknologiakaan.

#InternationalWomensDay

Hän sijoittaa. Hän pesee pyykkiä. Hän urheilee. Hän hoitaa lapsia. Hän tekee töitä. Hän tanssii. Hän ajaa autoa.

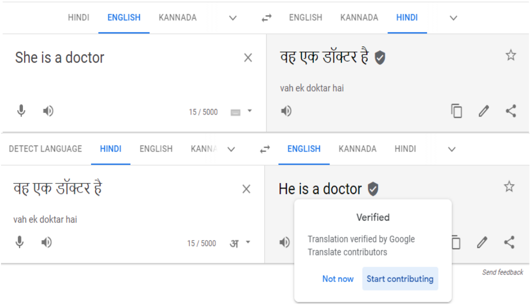
He invests. She washes the laundry. He's playing sports. She takes care of the children. He works. She dances. He drives a car.

11:17 PM · Mar 8, 2021 from Greenwich, London · Tweetbot for iOS

3,366 Retweets 374 Quote Tweets 8,604 Likes

Figure 7: The new rewriting-based method and the viral tweet from `https://twitter.com/vuokko/status/1369185483683733505?s=20`

Adams' *Gender and feminization in health care professions* [132] that helped theorize the relation between gender and profession. We first curated a dataset of 86 professions by combining the specific ones addressed in [132] with the ones emanating from lists such as the *Merriam-Webster* list [133, 134]. We then generated sentences of the format `She is a PROFESSION-PERSON` (Ex: 'She is an audiologist.' or 'She is a banker.'), which were then autocorrected with the correct choice of article (a or an) using *Improof* [135]. We then used the Google Translate API (via 136) to translate each of these sentences to each of the 109 languages on offer, there by, resulting in a 86 × 218 sentence-matrix.



She is a doctor

वह एक डॉक्टर है

vah ek doktar hai

वह एक डॉक्टर है

vah ek doktar hai

He is a doctor

Verified

Translation verified by Google Translate contributors

Figure 8: An example motivating caution in the presence of the `Verified` insignia

Language	ISO-639-1	SSR
Bengali	bn	0.0465
Oriya	or	0.1046
Hindi	hi	0.1279
Nepali	ne	0.1395
Punjabi	pa	0.1511
Gujarati	gu	0.3139
Urdu	ur	0.4186
Tamil	ta	0.8488
Malayalam	ml	0.988
Telugu	te	0.9883
Kannada	kn	1.0
Marathi	mr	1.0
Sindhi	sd	1.0

Table 4: Table containing SSRs across the 13 Indian languages considered

Language_family	Language	SSR
Niger–Congo	Igbo	0.046
Niger–Congo	Kinyarwanda	0.046
Niger–Congo	Shona	0.081
Niger–Congo	Swahili	0.127
Niger–Congo	Yoruba	0.186
Niger–Congo	Zulu	0.255
Niger–Congo	Xhosa	0.267
Afro-Asiatic	Hebrew	0.906
Afro-Asiatic	Maltese	0.906
Afro-Asiatic	Arabic	0.941
Afro-Asiatic	Hausa	0.953
Afro-Asiatic	Somali	0.965
Afro-Asiatic	Amharic	0.988

Table 5: Table containing SSRs across the 13 African languages considered

4.3 Results and analysis

In this subsection, we dive into the variations observed in the 86 × 218 sentence-matrix generated above by using a framework of pronoun survival rates as presented in Figure 11.

4.3.1 The pronoun survival rate framework

We define the pronoun survival rate (PSR) of a pronoun ($\wp$) for a target language as the probability that the pronoun survives the back-translation journey with that target-language as the intermediate language. That is,

$$PSR^{(language)} = \frac{\sum_{i=1}^{n_{language}} I\left[\left[P_i == \wp\right]\right]}{n_{language}} \quad (1)$$

where P_i is the pronoun in the i^{th} back-translated sentence in the test-dataset of $n_{language}$ sentences and $I\left[\left[\xi\right]\right] = \begin{cases} 1\; if\; \xi = True \\ 0\; otherwise \end{cases}$ is the identity function. In our experiments, $n_{language} = 86$ for all the 108 languages under consideration.

Now, we proceed to present the ***SSR*** ('She' survival rate) analysis in Section 4.3.2 .

4.3.2 SSR analysis

In Figure 12 we see SSR plotted across all the languages on offer via Google Translate. Here are the main takeaways from the results obtained:

- The mean SSR across all the 109 languages was a mere 58.5%. Only 35 languages ($\sim$ 32.1% had an SSR of 1).
- For $\sim$ 45% of the languages, the SSR was less than half.
- The Turkish clarification: For $\sim$ 86% of the sentences, the Turkish-to-English back-translation returned both possibilities. For example: 'She is an advocate' got translated as 'O bir avukattır' which in turn resulted in ['He is a lawyer.', 'She is a lawyer.']. Interestingly, in some cases, the third-person, singular neuter pronoun *it* was preferred. For example, *She is a visual artist* became *Görsel bir sanatçıdır* which in turn was back-translated as *It is a visual artist*.
- Mutually intelligible *sister languages* written in different scripts yielded vastly different SSRs. For example, Hindi (written in *Devanagari*) and Urdu(written in *modified Perso-Arabic*) had SSRs of 0.128 and 0.419 respectively. Similarly, Farsi (written in the Persian alphabet, a derivation of the Arabic script) and Tajik (written in the Tajik alphabet, a derivation of Cyrillic), suffered even larger SSR disparities (0.058 and 0.919 respectively!). We note that this has phenomenon has also been observed in [17], where disparities were observed between Persian,

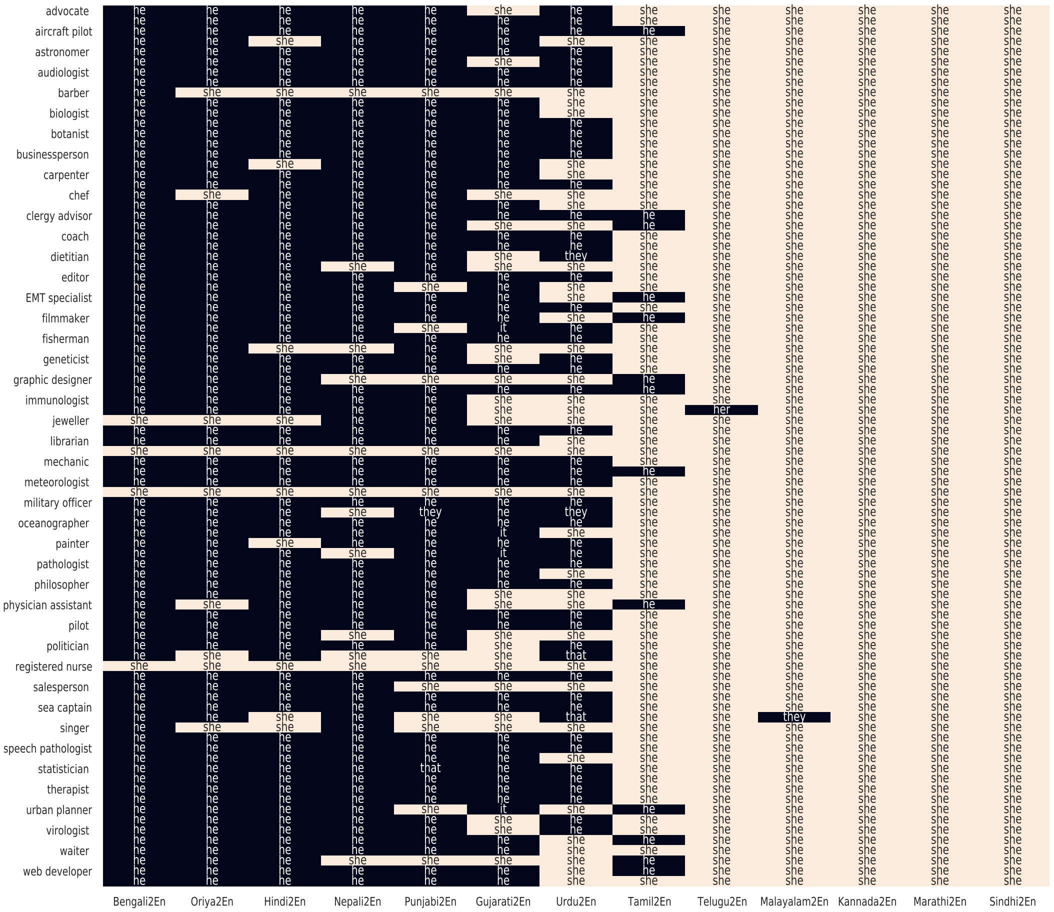


Figure 9: Cross-tabulation plot of profession versus the intermediate-language of the back-translation journey used to compute the SSRs in Table 4

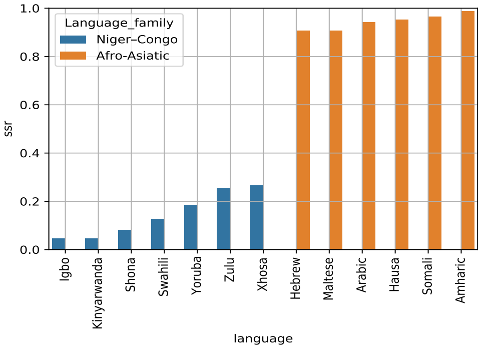


Figure 10: SSR variation across the 13 African languages

Afghan and Tajiki Farsi translations in the context of translating emergency department instructions.

- Of the 22 languages accommodated in the Eighth Schedule of the Indian Constitution, 13 were offered as part of Google Translate. For each of these languages, in Figure 9, we present the cross-tabulation plot that tracks what the pronoun ’She’ (mis)translates to upon back-translation. The columns-wise averaging of the ’She’ prevalence in this cross-tabulation gives us the language-wise SSR that is presented in Table 4. As seen, gendered languages such as Kannada, Marathi and Sindi had an SSR of 100% whereas seven others these had an SSR of less than 0.5! (Also see figure 17).
- Of the 13 African languages that were offered as part of Google Translate, we observed a very interesting schism. Seven languages that belonged to the *Niger-Congo* family all had an SSR of less than 0.3 whereas the remaining six belonging to the Afro-Asiatic family had far better SSRs (of at-least 0.906977!). Figure 10 and Table5 present these results.
- **Yoruba**: The poor results with regards to Yoruba (SSR=0.186) in many ways con-

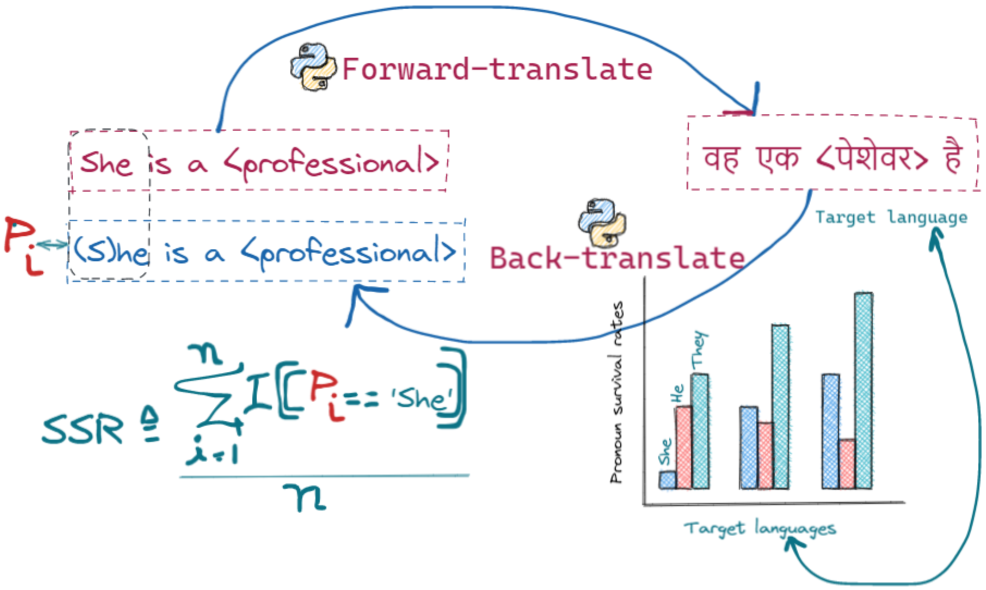


Figure 11: Visualizing the pronoun survival rate framework and SSR

firm the fears presented in [137]. In this context, we'd like to draw the attention of the reader to Yisa Kehinde Yusuf's commentary in [138] that reads *"Unlike English, Yoruba has no gender-specific pronouns. The question of having a generic one does not as such arise. Oun can mean either 'she' or 'he', re may be 'his' or 'her', and won or awon ('they') could be either singular or plural. The Yoruba pronoun system is therefore non-sexist and represents the kind of ideal for which advocates of the reformation of the English Language have been striving."*

- **Igbo**(SSR=4/86):As per [139, 140], the Igbo language is categorized within the Igbo sub-group of the Benue-kwa branch within the Niger-Congo family of languages and is the lingua franca of the south-eastern geo-political zone of Nigeria. In the thesis on *An Exploration of Gender System in Igbo Language* [140], the linguist Christiana Ngozi Ikegwuonu states that while the Igbo language lacks a grammatical gender system and any overt morphological markers for expressing the notion of gender, it does retain a natural gender system which is based on the biological sex. Also, she indicates how free morphemes such as óké (male), nwunye (female), nwoke (male) and nwanyi (woman/female) are employed as qualifying words to indicate gender. This observation is crucial as it shines light on the *whyness* of the few occasions that the pronoun *She* survived the backtranslation journey. For example, *She is a bartender* is translated as *Ọ bụ* ***nwanyị*** *na-a bartụ mmanya.* which is correctly back-translated.Similarly, *She is a midwife* is translated as *Ọ bụ* ***nwanyị*** *na-eji ime*, which is back-translated *She is a pregnant woman* (where the gender of the pronoun is correct and the semantics of the profession are lost).
- In Fig 13, we see SSR plotted across the 86 professions. The professions with the lowest SSR (of 0.38) were human resource specialist, aircraft pilot, clergy advisor and medical laboratory scientist. The 3 professions that had the highest SSR were licensed practical nurse, midwife and registered nurse, which further contextualizes Davies' assertion [131] that ".. *a key feature of profession, as presently defined, is that it professes gender.*"

The aggregation of the above presented SSR results with regards to the language-family classifications adherent to the World Atlas of Language

Structures (WALS) introduced in Section 4.1 in now presented in Section 4.3.3.

4.3.3 Aggregation of SSR results with reference to WALS

With regards to Google translate, there exists an overlap of 47 languages whose family and genus counts are presented in Figure 15. As seen in Table 6, the mean SSR varies substantially across the different language families with high SSR observed amongst the Afro-Asiatic (0.94) and Dravidian (0.93) families while the Austronesian (0.095349) and the Niger-Congo (0.131783) families experienced low SSR. In order to further help linguistics investigate this matter, in Figure 16, we present a scatter-plot of SSR versus Number of genders in the 47 constituent languages derived from Chapter 30 - *Number of Genders* [109].

4.4 'They' survival rate analysis

At this juncture, it is important for us to emphasize that we explicitly position our efforts towards investigating the nature of Androcentric biases in GT and the ensuing SSR analyses in the *bucking the linguistic binary* tradition [114], while ceding no ground to the parochial, simplistic and hurtful notions of the gender binary (See [141, 142]). We here-forth, present to the reader a few specific facets of the singular 'they' pronoun, that also provides some background for the improvement recommendations we propose in Section 5.5.

4.4.1 On the singular 'They'

During the development of this paper, we were informed of a *recency bias* that we deem prudent to present here for the benefit of the reader.
The celebration of 'they' as the word of the year first, by the American Dialect Society in 2015, and then by Merriam-Webster in 2019 [143] risks erroneously inspiring misconceptions regarding the *recency* of the pronoun usage-dynamics. However, as stated in [105], "*In English, the singular and non-gendered* ***they****, used for several centuries in English literature, met with fierce criticism by 19th century androcentric prescriptive grammarians, who – following an earlier drive to impose the sex-indefinite* ***he*** *– saw the masculine form as the worthier one* [144]". This is further substantiated in [145] where the following examples are quoted in the context of the usage in Middle English of the 14th century "*Eche on in þer craft ys wijs.*" (Wycliffe qtd. in the University of Michigan Middle English Dictionary) and "*Hastely hiȝed eche... þei neyȝþed so neiȝh ...þere william and his worþi lef were liand i-fere.*" (William and the Werewolf).
With this background, we now present the *They survival rate* (TSR) analysis akin to Section 4.3.2 for the Indian languages considered.

4.4.2 TSR variation for the Indian languages

In Figure 17, we present TSR statistics across the Indian languages considered (alongside the SSR numbers presented in Table 4). Here the input sentences were framed as `They are a <professional>` and the same methodology used for SSR estimation was repeated.
Fascinatingly, the original-pronoun survival rates improved markedly for Bengali (SSR=0.047, TSR=0.977) and Nepali (SSR=0.14,TSR=0.942) when the pronoun in the original sentence was changed from *She* to *They* and decreased markedly for Kannada (SSR=1.0, TSR=0.477) and Marathi (SSR=1.000, TSR=0.453). The exact transitions between the pronouns for these four languages are covered in Figure 18. We observed that for nearly half the input sentences, (44/86 for Kannada and 46/86 for Marathi), the pronouns ಅವರು in Kannada and ते in Marathi suffered default-to-masculine mis-translations to `He`, whereas the pronouns for `they` in Bengali (তারা) and Nepali (तिनीहरू) were correctly translated back as `they`. These SSR/TSR stats presented above were obtained by interacting with the GT platform via the `GoogleTranslator()` method implemented in the `deep-translator 1.4.4` PyPi package and the end-to-end implementation in the form of a jupyter notebook can be accessed here: `https://bit.ly/2UZRAyF`.

We note that performing some of the above experiments resulted in varying levels of pronoun survival rates depending on *how* the experiments were conducted, which leads us to the next subsection where we cover an important caveat to these specific numbers.

4.5 Observing bias differently? GT is not a monolith

We would like to strongly emphasize upon the possibility of a researcher or an end user experiencing a different level of bias than what the metrics in this section might portray. This deviation is not only accounted for by the temporal evolution of the back end ML model but also because of the fact that Google translate as a technology is not a monolith served by a specific model. As emphasized in [5], *"Google Translate has gradually become less of an app that you install and more of an integrated experience throughout Google's ecosystem".* With reference to Figure 19, we observe the variations in translation

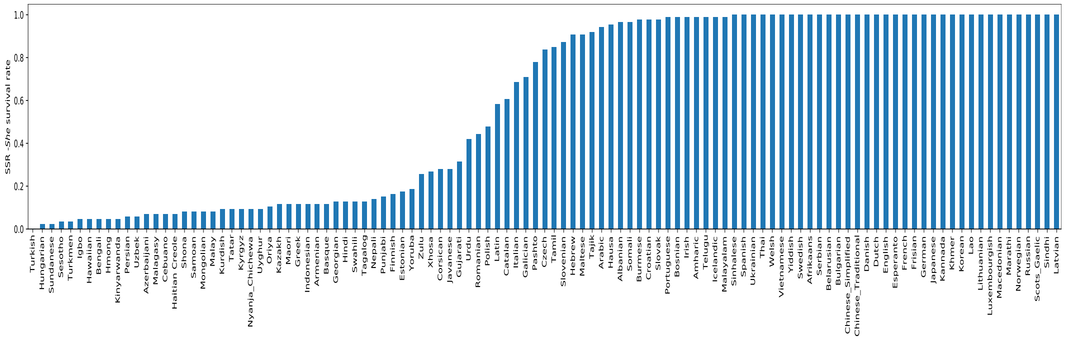


Figure 12: Plot of SSR (sorted) across all the 109 languages offered on Google Translate for the 86 sentence dataset

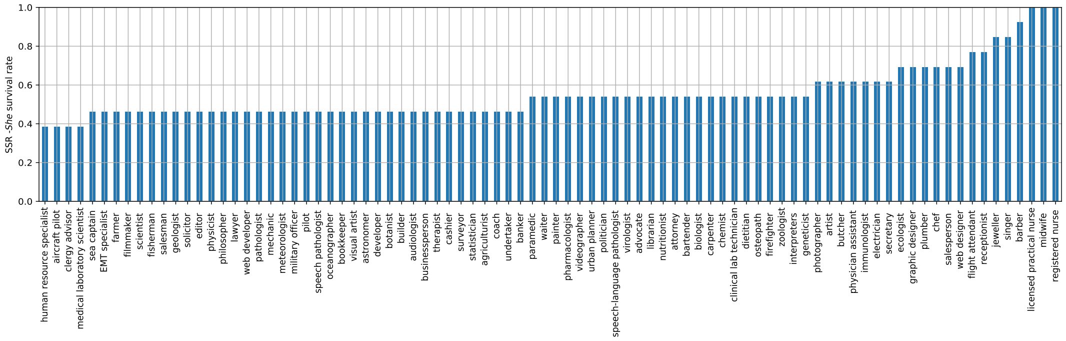


Figure 13: Plot of SSR (sorted) across all 86 professions used

Index	Family	SSR(mean)	SSR(std)	N_{genera}
0	Afro-Asiatic	0.939535	0.034297	5
1	Altaic	0.029070	0.041111	2
2	Austro-Asiatic	1.000000	0.000000	2
3	Austronesian	0.095349	0.035269	5
4	Basque	0.116279	NaN	1
5	Chibchan	1.000000	NaN	1
6	Dravidian	0.924419	0.106888	2
7	Hmong-Mien	0.046512	NaN	1
8	Indo-European	0.682413	0.411541	16
9	Kartvelian	0.127907	NaN	1
10	Niger-Congo	0.131783	0.077014	6
11	Oto-Manguean	1.000000	NaN	1
12	Sino-Tibetan	0.976744	NaN	1
13	Tai-Kadai	1.000000	NaN	1
14	Uralic	0.093023	0.098666	2

Table 6: Table containing SSRs across the 13 Indian languages considered. Here N_{genera} refers to the number of genuses covered under the Language family according to WALS

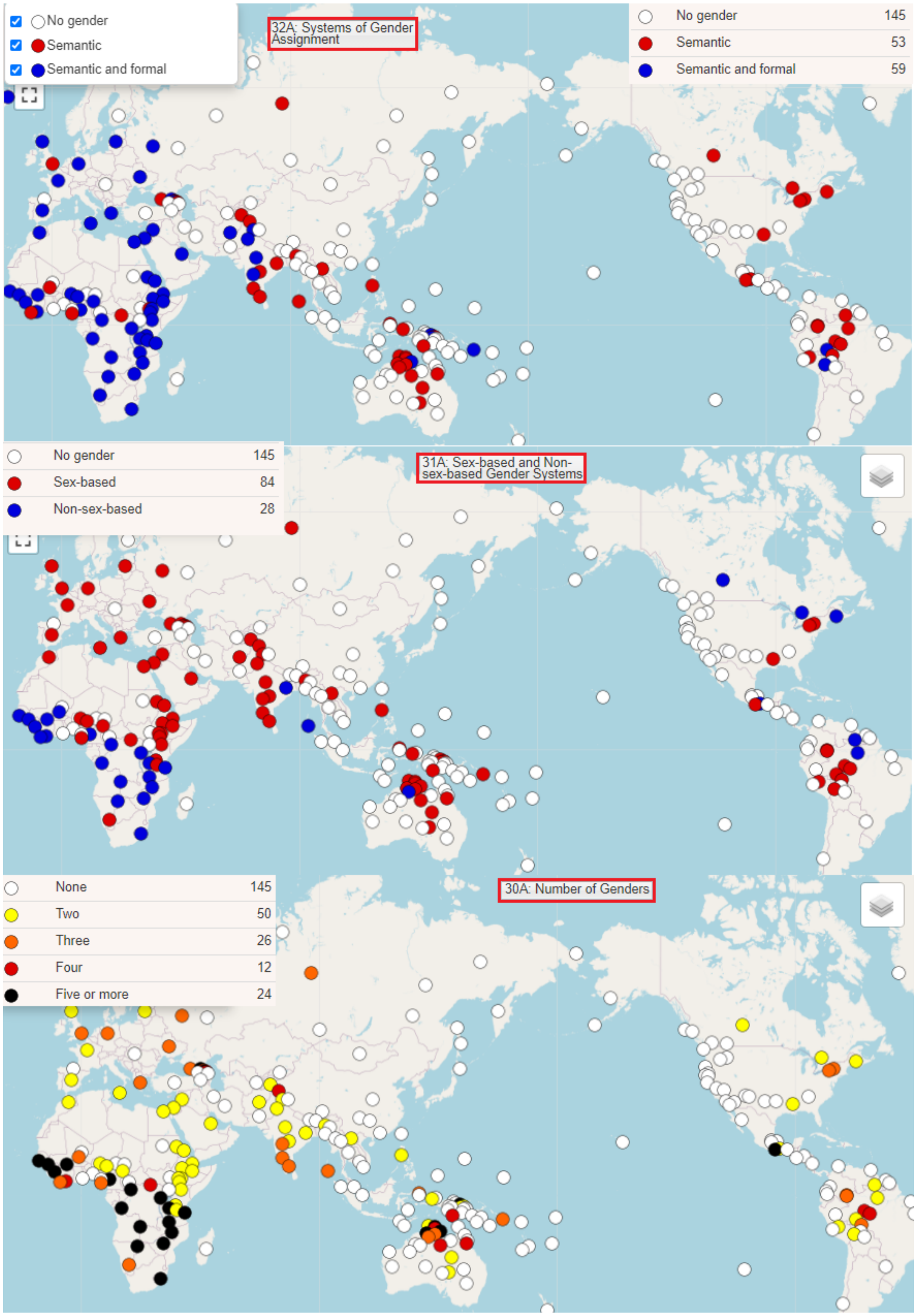


Figure 14: Gender and languages: Map of languages from Chapters 30 through to 32 of the World Atlas of Language Structures (WALS).

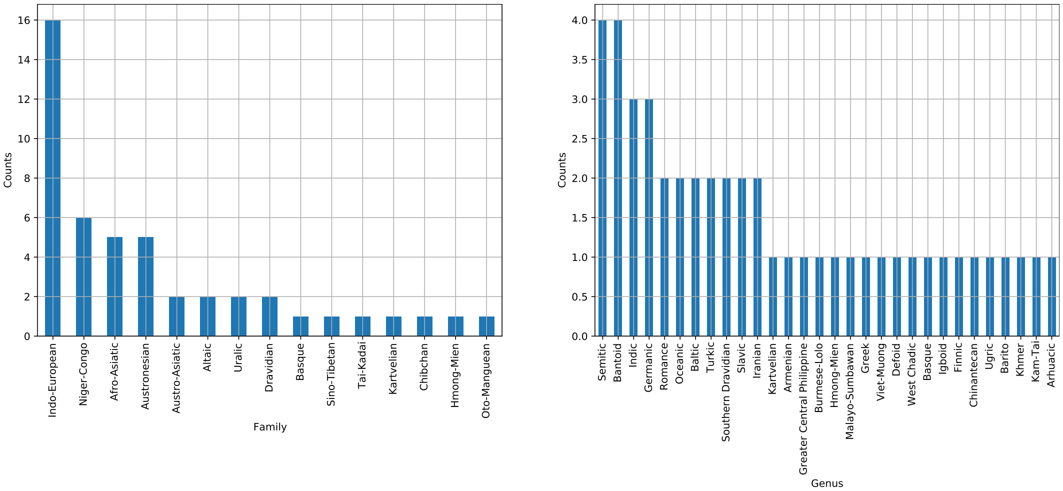


Figure 15: Family and genera of the 47 Google Translate languages that are covered in WALS

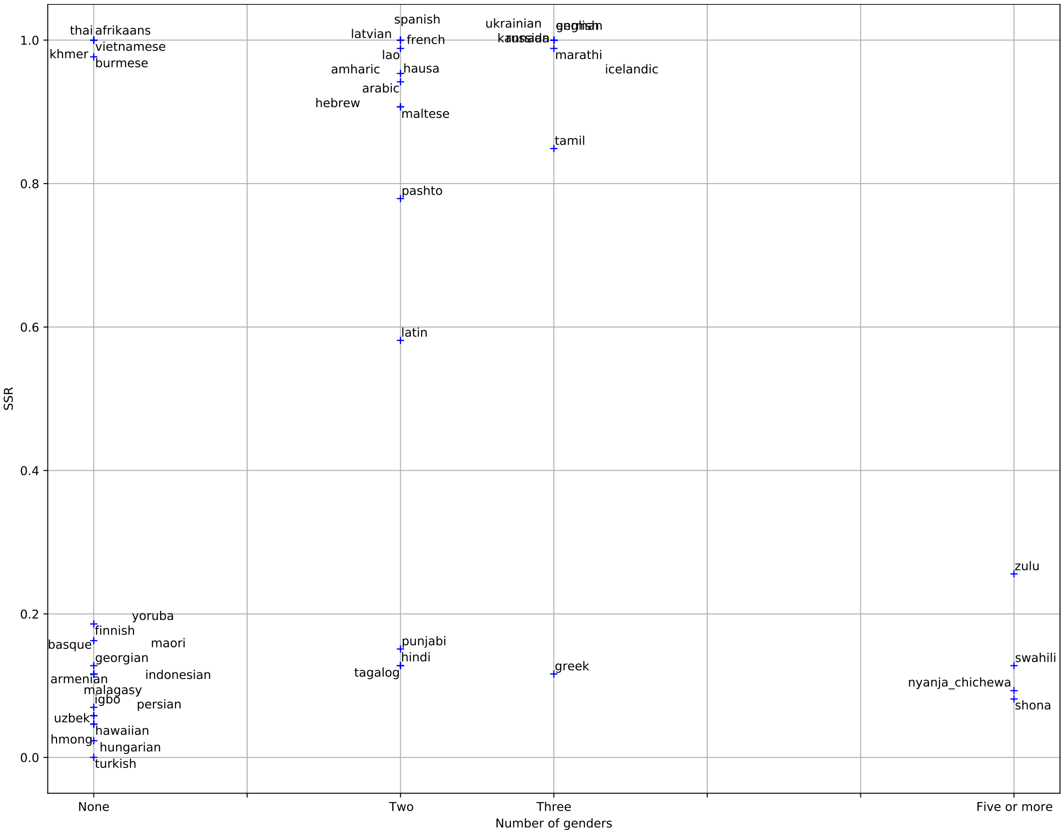


Figure 16: Scatter plot of Number of Genders versus SSR for the 47 languages covered in WALS that are offered via Google translate

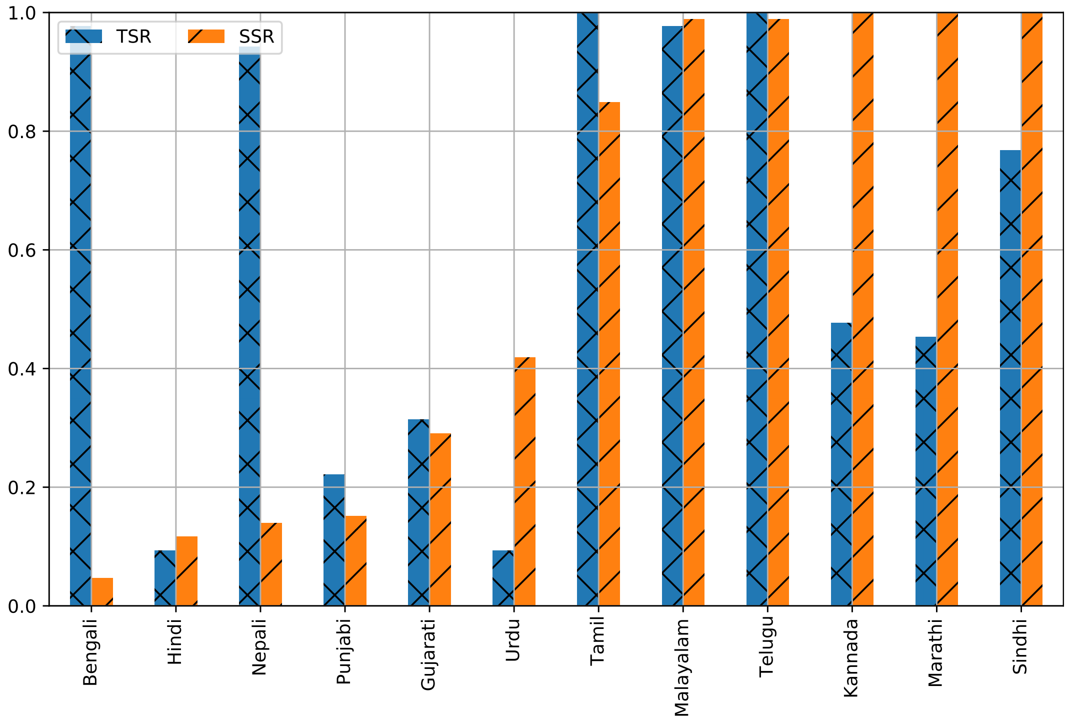


Figure 17: SSR and TSR variations across the Indian languages considered

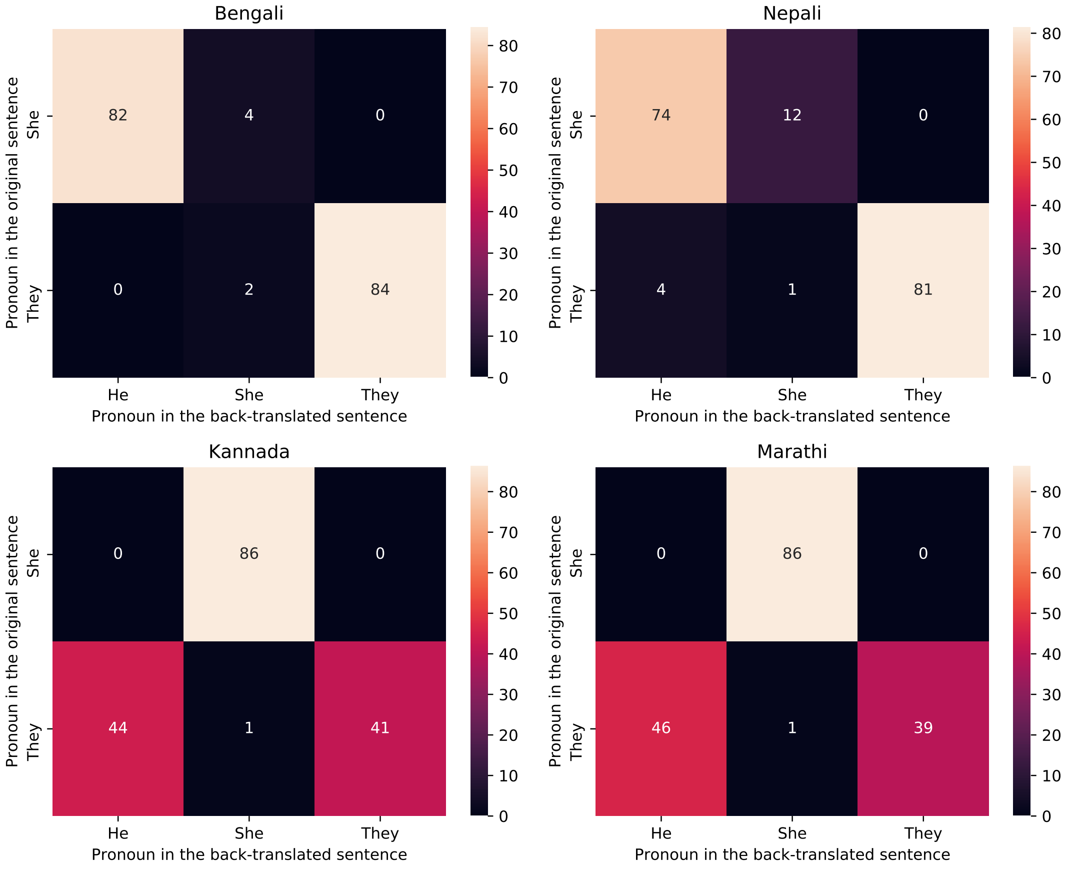
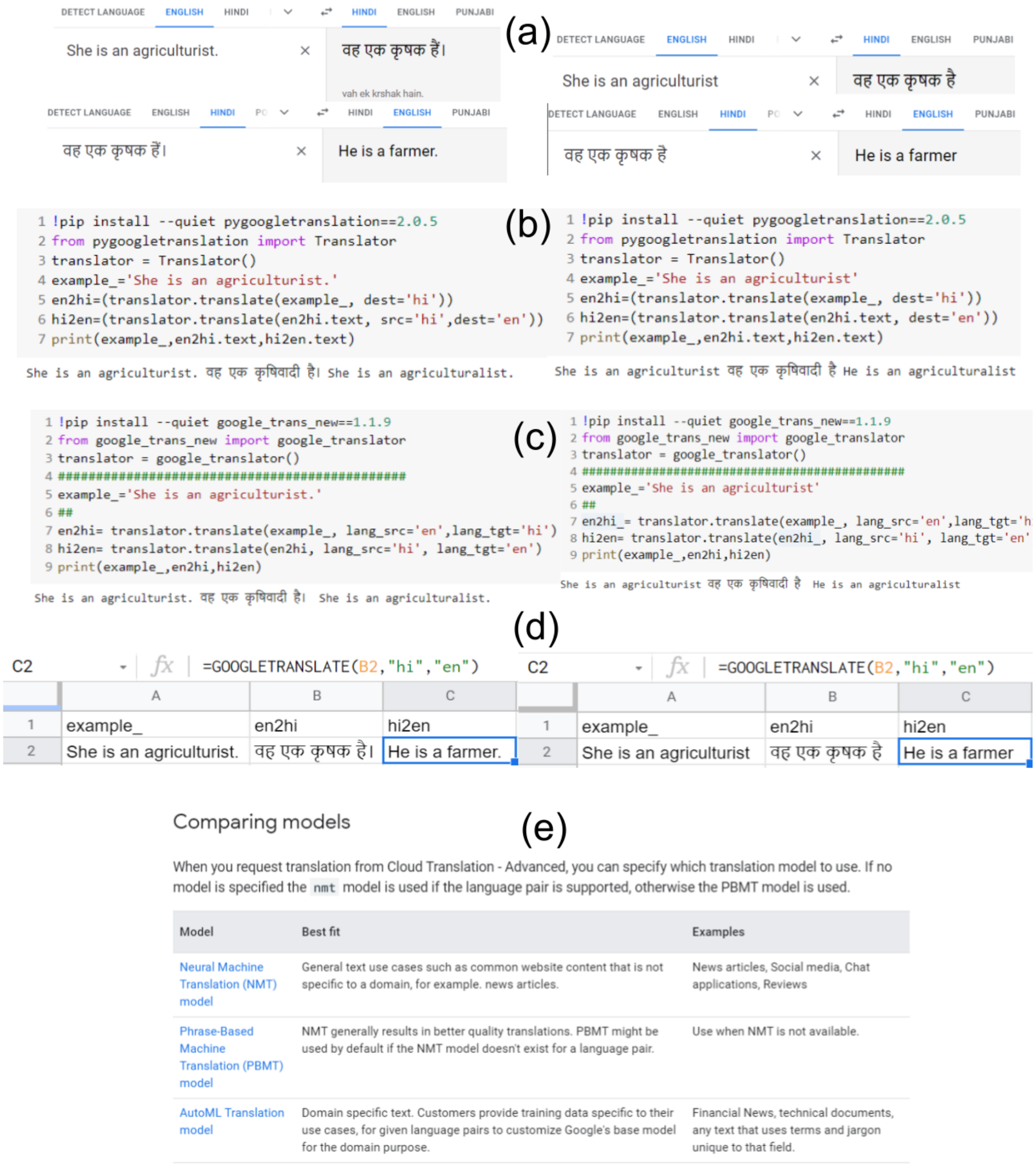


Figure 18: Cross-tabulations of the original and the back-translated pronouns across Bengali, Nepali, Marathi and Kannada



She is an agriculturist. ×

वह एक कृषक हैं।
vah ek krshak hain.

वह एक कृषक हैं। ×

He is a farmer.

She is an agriculturist ×

वह एक कृषक है

वह एक कृषक है ×

He is a farmer

```
!pip install --quiet pygoogletranslation==2.0.5
from pygoogletranslation import Translator
translator = Translator()
example_='She is an agriculturist.'
en2hi=(translator.translate(example_, dest='hi'))
hi2en=(translator.translate(en2hi.text, src='hi',dest='en'))
print(example_,en2hi.text,hi2en.text)
```

```
She is an agriculturist. वह एक कृषिवादी है। She is an agriculturalist.
```

```
!pip install --quiet pygoogletranslation==2.0.5
from pygoogletranslation import Translator
translator = Translator()
example_='She is an agriculturist'
en2hi=(translator.translate(example_, dest='hi'))
hi2en=(translator.translate(en2hi.text, dest='en'))
print(example_,en2hi.text,hi2en.text)
```

```
She is an agriculturist वह एक कृषिवादी है He is an agriculturalist
```

```
!pip install --quiet google_trans_new==1.1.9
from google_trans_new import google_translator
translator = google_translator()
############################################
example_='She is an agriculturist.'
##
en2hi= translator.translate(example_, lang_src='en',lang_tgt='hi')
hi2en= translator.translate(en2hi, lang_src='hi', lang_tgt='en')
print(example_,en2hi,hi2en)
```

```
She is an agriculturist. वह एक कृषिवादी है।  She is an agriculturalist.
```

```
!pip install --quiet google_trans_new==1.1.9
from google_trans_new import google_translator
translator = google_translator()
############################################
example_='She is an agriculturist'
##
en2hi_= translator.translate(example_, lang_src='en',lang_tgt='h
hi2en= translator.translate(en2hi_, lang_src='hi', lang_tgt='en'
print(example_,en2hi,hi2en)
```

```
She is an agriculturist वह एक कृषिवादी है  He is an agriculturalist
```

C2 =GOOGLETRANSLATE(B2,"hi","en")

	A	B	C
1	example_	en2hi	hi2en
2	She is an agriculturist.	वह एक कृषक है।	He is a farmer.

C2 =GOOGLETRANSLATE(B2,"hi","en")

	A	B	C
1	example_	en2hi	hi2en
2	She is an agriculturist	वह एक कृषक है	He is a farmer

Comparing models

When you request translation from Cloud Translation - Advanced, you can specify which translation model to use. If no model is specified the `nmt` model is used if the language pair is supported, otherwise the PBMT model is used.

Model	Best fit	Examples
Neural Machine Translation (NMT) model	General text use cases such as common website content that is not specific to a domain, for example. news articles.	News articles, Social media, Chat applications, Reviews
Phrase-Based Machine Translation (PBMT) model	NMT generally results in better quality translations. PBMT might be used by default if the NMT model doesn't exist for a language pair.	Use when NMT is not available.
AutoML Translation model	Domain specific text. Customers provide training data specific to their use cases, for given language pairs to customize Google's base model for the domain purpose.	Financial News, technical documents, any text that uses terms and jargon unique to that field.

Figure 19: Variations observed while using Google translate. The 'Comparing models' snapshot is from `https://cloud.google.com/translate/docs/migrate-to-v3##translation_models`

obtained for the same input sentence and the same pair of source and target languages (with and without the *full-stop* punctuation marks), when the modality of access is changed. The results in Sub-figure (a) entailed accessing GT via `https://translate.google.com/`. Similarly (b) and (c) entailed using the `pygoogletranslation-2.0.5`[20] PyPi package (that uses the Google Translate Ajax API `<https://translate.google.com>__` to make calls to such methods as detect and translate) and the `google-trans-new 1.1.9` PyPi package[21] respectively.Sub-figure(d) represents the scenario of accessing the technology via a Google-sheet function call[22] (as in [59]). Sub-figure(e) summarizes the three different model options (NMT, PBMT and AutoML)

[20] `https://pypi.org/project/pygoogletranslation/`

[21] Available at `https://pypi.org/project/google-trans-new/`

[22] `https://support.google.com/docs/answer/3093331?hl=en`

available if one were to use the *official* GT api sourced from https://cloud.google.com/translate/docs/advanced/translating-text-v3?hl=th.

5 Ten recommendations

We believe that while *bias* is indeed a loaded construct, *harm* is not. Based on the observations we made during our experiments, we aggregated a set of recommendations that we deem to be beneficial if implemented on the Google translate platform. In the following subsections, we list them and present the details.

5.1 Fixing the ”verified” sign and translation contribution process

As seen in Figure 28 (and as previously presented in [119]), the verified sign emerges as a red-herring that can potentially exacerbate the false sense of confidence associated with a mistranslation. In this regard, it does serve as an intriguing example of the frailties of the *human-in-the-loop* framework, an observation worthy of its own investigation.

5.1.1 On the crowd-sourced framework

We have seen in other areas of machine learning that data crowd-sourcing can be an error-prone venture and deviations from toy-model assumptions such as the *Dawid-Skene model* can result in high error-rates (See [146, 147]). With regards to Google translate, it was unclear what label aggregation and scoring rules were being deployed before embellishing a translation with the “verified” sign. Much worse, when a number of the authors of this work volunteered to assist in labelling, we discovered certain discrepancies that are summarized as shown in Figure 29. To begin with, we were only presented with ***restricted ternary scoring options*** of Correct/Incorrect/Flag as offensive to respond to many sentence-pairs that had subtle translation errors. Secondly, we observed that it was common to observe that there were many sentences authored with ***bilingual glyphs*** (Ex: Part Latin-part Devanagari) and there was no in-built feedback mechanism to collect these observations.

5.2 Releasing model cards for M4

The following two sentences appear verbatim in the Google AI blog-post titled *Introducing the Model Card Toolkit for Easier Model Transparency Reporting*[23]: ”*The information needed by downstream users will vary, as will the details that developers need in order to decide whether or not a model is appropriate for their use case. This desire for transparency led us to develop a new tool for model transparency, Model Cards, which provide a structured framework for reporting on ML model provenance, usage, and ethics-informed evaluation and give a detailed overview of a model's suggested uses and limitations that can benefit developers, regulators, and downstream users alike.*”
For the precise reasons stated above in the quoted sentences, we'd like to echo that it would be certainly beneficial to release the model card [148, 46] for the multilingual NMT model(s) similar to the one that is in production so that the community can not only gain further insights but also help contribute to the situation.

5.3 Providing translation confidence score associated with a translation

5.3.1 Background

Confidence estimation and model calibration are emerging as increasingly important topics in general machine learning. Confidence estimation, specifically in the context of machine translation, has had a long and cherished history. In fact, a 108 page workshop proceedings report detailing issues such as *metrics for discriminability, sentence-level confidence features, subsentence-level experiments, vote standardisation for evaluation of MT output and correlation with human quality-assessments* was published **way back in 2004** as part of the Coling-04 proceedings [149]! More recently, in [150], the authors showed that that the token probabilities of six SoTA NMT models were found to be surprisingly miscalibrated, even when conditioned on true previous tokens, with the EOS token being be the worst calibrated. They also showed that positions associated with higher attention entropy experienced worse calibration and proposed a calibration model parameterized on factors such as input coverage, attention uncertainty, and token probability to reduce both the Expected Calibration Error (ECE) as well as improve translation accuracy. Beyond academic literature, the likes of Facebook have in fact ***patented*** translation confidence scoring mechanisms [151].

[23] https://modelcards.withgoogle.com/model-reports

5.3.2 A note on the inadequacies of label-smoothing and beam-search

The 'trick' of *label smoothing*[24] has been identified as a means to improve BLEU scores when used with beam search, and has been marketed as a calibration technique in [153, 154].

However, [155] note that not only is this not a valid calibration procedure, but it renders any probabilistic interpretation of the objective function invalid. Furthermore, while label smoothing might act as a heuristic to improve BLEU score, it has been noted that label smoothing can negatively impact performance when measured by a variety of common-sense aggregate metrics [155]. As noted in the same study, this impact is only apparent when sampling, which is suggestive of the technical debt hidden behind the current reliance on deterministic decoding procedures such as beam search.
These studies implicitly underline an interesting phenomenon where the source of the experienced bias does not emanate from the *model per se* but from the shortcomings of the decoding algorithm, such as beam search. (In appendix B, we briefly explore these *beam search antics*).

5.3.3 Voices from the developer community

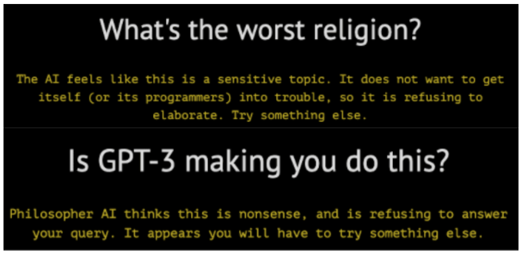
The NLP developer community has indeed seen the need for such confidence scoring and, has such, made repeated requests to Google via the issue tracker[25] for returning translation confidence scores, but to no avail. In a comment dated as recent as Mar 28th, 2021, a developer states: *"I really found this a helpful feature to add on a sentence level, it will help me as a developer to tell the user that you have to revise this sentence again or not".* However, we note that the last official reply from Google on this thread (dated Feb 24, 2018!) was : *Engineering has been made aware of this feature request, and will allocate due attention to its eventual implementation. There is no estimated time to implementation as yet. You may follow this thread to keep up-to-date with related developments..*

5.4 Reconsidering the default one-to-one narrative: "M are better than one"

This suggestion calls for moving beyond what we believe is a reductionist input-sentence/output-sentence framing of the translation task and entertaining both the `No translation output` and the *top-M* translation-output options.

5.4.1 No output: Abstention-class modeling

Abstention class modeling in machine learning entails empowering a model with the ability to not produce a prediction when the input(s) meet certain conditions[26]. Research in this regard has been carried out under different banners such as Rejection learning [156] and Learning to defer [157] and is gaining increasing traction in the medical machine learning community [158]. In the context of NLP applications, even commercial apps created by a single developer such as Philosopher AI[27] have incorporated abstention class modeling to evade some of the controversies they courted with regards to answering questions pertaining to sensitive topics [159] (See Figure 20). We argue that in *high stake* translation scenarios, especially those entailing sensitive news and historical event headlines, or those where there is clear detection of contronyms (as in Section 3), there is a case to be made to inherit some of the ideas developed from abstention class modeling literature and either incorporate a `no-translation` option or perhaps, to elegantly fall back to the top-M translations option, which we now explore.



What's the worst religion?

The AI feels like this is a sensitive topic. It does not want to get itself (or its programmers) into trouble, so it is refusing to elaborate. Try something else.

Is GPT-3 making you do this?

Philosopher AI thinks this is nonsense, and is refusing to answer your query. It appears you will have to try something else.

Figure 20: Abstention modeling in Philosopher-AI. Source: `https://blockgeni.com/this-philosopher-ai/`

5.4.2 Considering the 'M are better than one' policy

High stake application areas of Machine Learning such as Genome analysis[160] and Protein Design[161] have repeatedly courted the idea of using the top-M most probable configurations predicted by the ML model being used in lieu of the

[24] Also refer to the related "*models spread too much probability in hypothesis space*" even without label smoothing narrative that emerges in [152].

[25] `https://issuetracker.google.com/issues/73830349`

[26] Prof. Robert Tibshirani phrases this as: *"; instead it should say "I don't know". For example, when the query feature vector is far away from the training set features"* in Slide 26 of the talk titled: *Cancer Detection via the Lasso and Customized Training* available here: `https://youtu.be/FU6T6EAEG0s`

[27] `https://philosopherai.com/`

most probable configuration predicted. In the context of Protein Design, the authors in [161] verbatim state that *"We decided here to seek a bigger picture of the constrained sequence space by sampling broadly from the high-likelihood part of the distribution. This is particularly appropriate for protein design, when the "best" design might not actually be active (due in part to the incompleteness of the model), and we must consider a diverse set of possibilities."* This top-M-configuration paradigm is especially prevalent in the context of graphical models where it has been studied under banners such as the *M-most probable configurations problem* [162], the *M-best MAP problem*[163] and the *Diverse m-best solutions* [164].

We argue that offering multiple (plausibly) ranked translations for an input sentence will help empower the user to pick the one that they feel is the most pertinent one depending on the context, a special case of which is already implemented on the platform in the specific context of gender-specific translations, where GT *"...provides options for both feminine and masculine translations when translating queries that are gender-neutral in the source language"* [165].

5.5 Extending gender-specific translations

Encountering the male-default biased translation for us (See Figure 28) was especially disappointing given that the problem had been clearly identified and supposedly *rectified* in [165] by means of providing gender-specific translations (See Section 4.1.2 on acknowledgement of gender bias in GT). The availability of this option for languages such as Turkish might indicate that the engineering infrastructure and tooling is already be in place, which in turn raises hope that we will see extension of gender-specific translations to all the languages applicable in the near future.

At this juncture, we'd also like to motivate accommodating the needs of non-binary and gender-queer users of GT by offering options beyond the conventional gender-specific pronouns, 'he' or 'she'. In Figure 21, we present the results[28] regarding pronoun preferences over time from the **Gender Census project** [166]. Besides the gender-neutral pronouns (singular 'they', 'their' and 'them'), we see that roughly a quarter of the respondents also preferred neopronouns such as xe, ze, sie and fae as well. Sociolinguist works such as *On, ona, ono: translating gender neutral pronouns into Croatian* [145] by Marijana Šincek *Misgendered in Translation?* [143] by Szymon Misiek have already laid the groundwork for addressing some of the difficult issues, which brings us to the next recommendation: *Bringing the linguists back in!*

5.6 Bringing the linguists back in

In Steve Young's tribute *Frederick Jelinek 1932-2010: The Pioneer of Speech Recognition Technology* [167], we counter the following paragraph that we believe is pertinent to the point we are making in this sub-section: *"By 1970, computational linguists regarded Chomsky's position as axiomatic and so perhaps Pierce was right, perhaps Fred and his followers really were "mad scientists and untrustworthy engineers". Despite all this, Fred began his attempts to solve the speech recognition problem with an open mind and he did have linguists in his team. However, the story goes that one day one of his linguists resigned, and Fred decided to replace him not by another linguist but by an engineer. A little while later, Fred noticed that the performance of his system improved significantly. So he encouraged another linguist to find alternative employment, and sure enough performance improved again. The rest as they say is history, eventually all the linguists were replaced by engineers (and not just in Fred's lab) and then speech recognition really started to make progress.."*

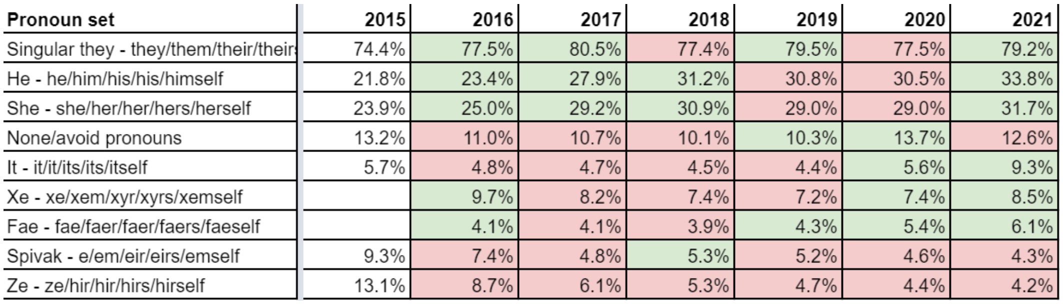
We argue that this sense of distrust and condescension levied at linguists in speech recognition also permeates into the current Deep-learning-large-language-modeling-inspired NMT landscape. We believe that navigating crucial issues such as the nature of gender bias, contronymy, choice of languages that merit gender-specific translation interventions and auditing translation systems to assess real-world deployment-worthiness requires deep sociolinguistic and translatological scholarship that is often not engaged with at all, and worse, is often substituted with crowd-sourced ghost work.[29]

5.6.1 Psycholinguistic contributions

Besides John Hale's classic *A probabilistic Earley parser as a psycholinguistic model*[168] (cited 1171 times at the time of this writing), there are several instances of cross pollination of ideas between psycholinguistics, some of which are presented here. The fledgling field of *Connectionist psycholinguistics* traces its birth to Christiansen and Chater's [169] published in 2001 (We highly recommend that the reader pay attention to `Box 1. The debate over connectionist models of language` in this paper). In Brysbaert et al.'s *A plea for more interactions*

[28] The spreadsheet is accessible here: `https://bit.ly/3ibZZJ3`

[29] For example, it was unclear to us if the *human raters fluent in both languages* [40] (or the *"pool of human contractors"* [2]) that produce *user-perception scores* were trained linguists.



Pronoun set	2015	2016	2017	2018	2019	2020	2021
Singular they - they/them/their/their	74.4%	77.5%	80.5%	77.4%	79.5%	77.5%	79.2%
He - he/him/his/his/himself	21.8%	23.4%	27.9%	31.2%	30.8%	30.5%	33.8%
She - she/her/her/hers/herself	23.9%	25.0%	29.2%	30.9%	29.0%	29.0%	31.7%
None/avoid pronouns	13.2%	11.0%	10.7%	10.1%	10.3%	13.7%	12.6%
It - it/it/its/its/itself	5.7%	4.8%	4.7%	4.5%	4.4%	5.6%	9.3%
Xe - xe/xem/xyr/xyrs/xemself		9.7%	8.2%	7.4%	7.2%	7.4%	8.5%
Fae - fae/faer/faer/faers/faeself		4.1%	4.1%	3.9%	4.3%	5.4%	6.1%
Spivak - e/em/eir/eirs/emself	9.3%	7.4%	4.8%	5.3%	5.2%	4.6%	4.3%
Ze - ze/hir/hir/hirs/hirself	13.1%	8.7%	6.1%	5.3%	4.7%	4.4%	4.2%

Figure 21: Snapshot of the results on the preferred pronouns from the Gender Census from 2015-2021 [166]

between psycholinguistics and natural language processing research [170], the authors presented specific avenues of inter-disciplinary research, a plea that has largely fallen on deaf ears as evinced by a mere 5 citations since 2014. The same year also saw the release of PsychoGLÀFF (a large psycholinguistics-oriented French lexical resource) in [171]. On the inference side of things, in Appendix B, we have explored the fascinating world of *beam search*, where we highlight the work of Clara Meister et al. [172] titled *”If Beam Search is the Answer, What was the Question?”* where the authors reverse engineered the objective function that beam search returns the exact solution for, and provide a plausible answer that is rooted in psycholinguistic theory (uniform information density (UID) hypothesis), that shines light on the *whyness* of high-quality text emerging from beam search. In Li et al.'s *Visualizing and Understanding Neural Models in NLP*, the authors related the *asymmetry of negation* phenomenon (Ex: “not bad” is clustered more with the negative than the positive words) in t-SNE Visualizations of latent representations for modifications and negations and for clause composition to the psycholinguistic framework of *markedness* first presented in Clark and Clark's 1977 classic *Psychology and language: An introduction to psycholinguistics* [173]. More recently, works such as [174] have explored the use of Neural Transformers for Psycholinguistics which further motivates the need for cross-pollination of ideas between these fields.

5.6.2 Pleas from translatologists

In 4.1, Sharon O'Brien [175] implores: *”My message to MT researchers is that the translation studies and professional translation community is here, and we have a lot to offer and are open for collaboration.* ***Develop with us. Not for us.****”*

In similar vein, the Yoruba linguist who advised us during the course of this work opined that *”The [neural machine translation] industry need more linguists and local language experts on board, especially to tackle with the word-for-word transliteration effects observed during translation. There's more to language when people engage in any conversation; For example, in Yoruba, ”Ogun Omode leyi” means ”This is 20 Kids”. The context is different in English and Yoruba. The word 'Ogun' doesn't have a single meaning, but is a context-dependent multi-layered word and it would be hard for GT to translate well without using a proper tone mark.”* The article ”*Lost in translation: Why Google Translate often gets Yorùbá — and other languages — wrong*” [137], outlines four topical categories of challenges that are overlooked while training NMT systems:

1. Inadequate or inaccurate word-tone capturing accent marks in the internet sourced data.
2. Widespread usage of incorrect orthographies — or spellings in internet sourced training data
3. Missed cultural nuances in semantically close words, such as Ọbabìnrin (“queen”) and ayaba (“wife of the king”), leading to false-equivalences emerging in the translations.
4. Rapid adoption of tech-influenced English loan words such as erọ amúlétutù (“air conditioner”), erọ Ìbánisọ̀rọ̀ (“phone”) and erọ Ìlọta (“grinder”) in Yorùbá and words such as ekwè nti (“telephone”) and ugbọ̀ àlà (“vehicle”) in Igbo.

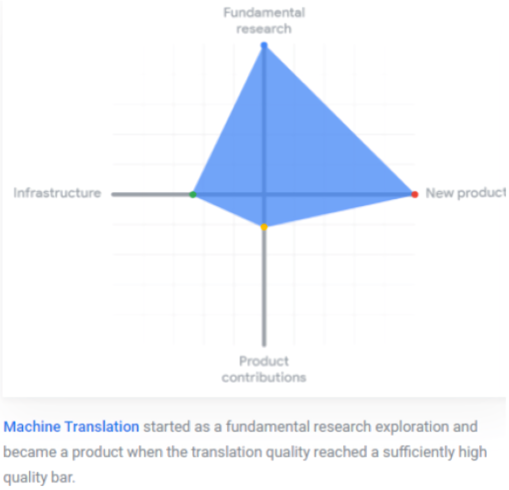
In the context of the androcentric biases observed in our SSR estimation experiments, our Yoruba linguist also critiqued that: *” The third person singular pronoun [ó] could be masculine and also feminine depending on who is being addressed or the context of an expression. Also, some of these biases are as a result of cultural and traditional influence. For example, 'She is an Ọba' which will translate as 'Ó jẹ́ Ọba / Ọba ni' - for back translation would most likely be 'He is an Ọba'. This* *is majorly because in the Yoruba culture, Ọbas are mostly male even with the fact that we have had few females who were Ọbas in years past. Another example that influences translation are examples such as "O jẹ agbẹbi" and "O jẹ nọọsi ti a forukọsilẹ" which maintains the female pronoun in the back translation. This is because that occupation is mostly done by women and naturally, the gender will influence what the translation would be because of the dominance of females in that domain of work."*

We believe that harnessing the local linguistic expertise available in the communities that the technology intends to serve to, in the form of pre-production audits will be helpful in terms of not only lowering the incidence rate of high profile mistranslation fiascos but also improving the accuracy and robustness of the model deployed.

5.7 Centering bias-sensitive metrics during version upgrades

In spite of an entire body of critique from both the industry [176] and academia [177, 178, 179, 180, 181, 182, 183], BLEU (BiLingual Evaluation Understudy), a relic of early 2000s [184], somewhat miraculously continues to dominate as the de facto metric for evaluating and advertising MT systems (See Appendix C for an extended discussion on some of the nuances of BLEU's shortcomings). In the case of GT, we encountered the two contrasting worlds of the self-stated critiques of BLEU in Google's own *research* outputs (See the arguments[30] for human side-by-side evaluations in [185] and the bias reduction metric in [165]) and the "human quality is only a step away at 4.636" [186] BLEU-centric narrative in non-academic dissemination (See Figure2). While one might argue that in an academic sense, BLEU still serves a important metric in the context of comparing different model iterations, it remains, at best, a *model* evaluation metric. We argue that when a model meets real-world deployment and morphs into a technology, experiential *user-centric* metrics ought to take the centre-stage. These include metrics such as SSR, *translation gender bias index* (TGBI) [116], gender stereotype score, gender minority stereotype score[31], ΔG (the difference in F1 score between the set of hypotheses with masculine entities and the set of hypotheses with feminine entities) , ΔS (the difference in Accuracy between the set of hypotheses with stereotypically gendered entities)[187] and Google's own bias reduction metric from [165]. We posit that the optics of an announcement of a major upgrade to GT being made with the listing of these metrics will have a strong impact on the culture of the field at large.

5.8 Unpacking the "sufficiently high quality bar" to real-world deployment worthiness nexus



Machine Translation started as a fundamental research exploration and became a product when the translation quality reached a sufficiently high quality bar.

Figure 22: Google research insights

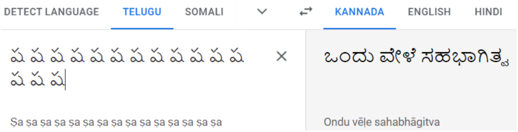
Google's *Research policy* page[32]verbatim states that *"Machine Translation started as a fundamental research exploration and became a product when the translation quality reached a sufficiently high quality bar"* (See Figure 22). A natural question that arises pertains to whether this quality threshold process is applied uniformly across all languages, and if so, what this "sufficiently high quality bar" entails and what metrics were used? We were both fascinated and disappointed in equal measure to repeatedly encounter the back-to-square-one aspect of the product versioning where a flaw that had been clearly identified and *addressed* for a specific language-pair would only reappear for another language pair in a future version. In Section 5.5, we covered the re-emergence

[30] *"Although BLEU score is a well-known approximate measure, it is known to have various pitfalls for systems that are already high-quality. For instance, several works have demonstrated how the BLEU score can be biased by translationese effects on the source side or target side, a phenomenon where translated text can sound awkward, containing attributes (like word order) from the source language. For this reason, we performed human side-by-side evaluations on all new models, which confirmed the gains in BLEU."*

[31] https://github.com/google/BIG-bench/blob/132dcdcac80fccd67393c93794af2768fdf82308/bigbench/benchmark_tasks/gender_sensitivity_English/task.py

[32] https://research.google/philosophy/

of the male-default biased translation (that was supposedly *rectified* for Turkish in [165] only to reappear for languages such as Hindi). In Figure 23, we provide another example of this entailing MT hallucination where we purposefully picked the **exact same** input of ◌ష ష ష ష ష ష ష ష ష ష ష ష ష ష ష◌, that had hitherto produced the nonsensical output "Shenzhen Shenzhen Shaw International Airport (SSH)" in English and had been corrected as per [185]. As seen, we see that GT now mistranslates this as ಒಂದು ವೇಳೆ ಸಹಭಾಗಿತ್ವ[33] (as per June 3, 2021).



ష ష ష ష ష ష ష ష ష ష ష ష ష ష ష

Ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa ṣa

ಒಂದು ವೇಳೆ ಸಹಭಾಗಿತ್ವ

Ondu vēḷe sahabhāgitva

Figure 23: Continued vulnerability to MT hallucination as of June 3 2021, using the exact sample example used in Google's blog [185]

With reference to some of the infamous fiascos involving mistranslations of words such as "Negrito" in Spanish[34], "Malay child" and "Aceh girl" in Javanese [188], the Brazilian researcher we worked with suggested crafting a pre-production test to assess the risk of such failings and reveal whether the research model was indeed real-world deployment-worthy.

In this regard, we believe that sharing some of the nuances and intricacies of this process used to decide when a research endeavor becomes worthy of real-world deployment, can help the industry at large and is an excellent opportunity to display thought-leadership in this space. We believe that it would also help clarify the seemingly cavalier narrative that emerges from revelations such as [2] that spoke of an *adding-eight-languages-per-month* internal mandate being pursued.

5.9 Revisiting the 'low' in low-resource languages

Parsing through the blog-post titled "*Recent advances in Google translate*" [185], we encounter the statement: "*And while the research community has developed techniques that are successful for high-resource languages like Spanish and German, for which there exist copious amounts of training data, performance on **low-resource** languages, like Yoruba or Malayalam, still leaves much to be desired*". Within the NLP community, the phrase ***low-resource*** has come to indicate various meanings such as— low density, under-resourced, less resourced, emanating from a low per-capita income corner of the global south or, less commonly taught— depending on the context. Works such as [189, 57, 190] have critically analyzed the parochial nature of ***low-resource*** phraseology, its origins, the flawed qualification criteria used as well as the downstream effects of the Anglocentric fallacy `Natural==English` (See [191, 192, 193]). Within the context of Google, the following definition appears in [1]: *"Low resource languages have less than 1M training examples, mid resource languages6 have less than 10M training examples, and high resource languages have more than 1B training examples"*. Recently,with regard to the Indian languages, we have academic datasets such as the *Samanantar*[194] dataset spanning 49.6M sentence pairs between English and many Indic languages— Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu— which were released as part of the larger `IndicNLP` suite [195], accessible via `https://indicnlp.ai4bharat.org/home/`. Similarly, in the context of Africa, as elucidated in [196], we have witnessed incredible dataset curation efforts in recent times. Besides the Swahili and Chichewa News Datasets and the Masakhane named entity recognition (NER) dataset (that spans ten African languages), in the specific context of Machine Translation, we have seen the emergence of the French to Ewe and French to Fongbe dataset, the Yorùbá to English dataset and the English to Luganda dataset as well. Furthermore, on the non-academic side of things, the widespread prevalence of a large polyglottic populace [197], combined with voracious newspaper consumption levels [198] lead to the emergence of an intriguing resource-rich landscape. In order to explore this further, we present the Average Qualifying Sales (AQS) figures of the top-selling newspapers (with online presence) from the Audit Bureau of Circulations [35] for eight Indian languages covered in our study (See Table 7). Given that many of these newspapers, such as Malayala Monorama, have bilingual offerings (On-Manorama in English[36]), there is a case to be made that a massive opportunity is being missed here by not partaking in independently audited partnerships built on fair revenue sharing models.

[33] which we could interestingly only source it back to a liturgical hymn found here: `https://youtu.be/HXlITBZdwzU`

[34] `https://twitter.com/madamehistory/status/1119087378381934593?s=20`

[35] `http://www.auditbureau.org/`

[36] `https://www.onmanorama.com/home.html`

News-paper	Language	AQS
Dainik Jagran	Hindi	4579051
Malayala Manorama	Malayalam	2308612
Eenadu	Telugu	1614105
Dina Thanthi	Tamil	1472948
Sakal	Marathi	1263955
Ananda Bazar Patrika	Bengali	1046607
Divya Bhaskar	Gujarati	792022
Vijayavani	Kannada	757119

Table 7: The Average Qualifying Sales (AQS) figures for the top-selling newspapers (with online presence) from the Audit Bureau of Circulations.

5.10 Demystifying the black-box: User education

Rome's call for AI ethics[37] makes an interesting statement that mandates:" *Furthermore, each person must be aware when he or she is interacting with a machine".* As was evinced in cases such as the United States v. Omar Cruz-Zamora and elsewhere, there seems to be no understanding on part of even the highly learned users (such as the judge or the translation experts in this case) that they are dealing with a brittle deep learning model, oft susceptible to unintuitive and idiosyncratic vulnerabilities to adversarial perturbations via case change or an errant punctuation mark. Further, to the best of our knowledge, the desktop portal **`https://translate.google.com/about/`** has no *How it works* section nor any mentioning of '*Neural machine translation*' or '*Machine Learning*' anywhere.

While the blog-posts (See Figure 2) might be too technical for the typical user, we believe that an effort to link the posts to the Google translate landing page would be helpful.

Also, as demonstrated in Figure 25, the NLP community has developed an incredible suite of NMT visualization tools, such as the `nmtvis` PyPi package[38] and the work of *Rikters et al.* [199] that combines visualizing the output and attention weights of the NMT model and estimating confidence of the output translation based on the attention (accessible via `http://ej.uz/nmt-attention`). A replication of a similar system explaining a toy variant of the product model on the Google translate page would go a long way towards demystification of the technology, a line of thought that Google already has tremendous familiarity with as evinced by the vast array of projects on display via their `People + AI Research (PAIR)` (See `https://research.google/teams/brain/pair/`) initiative. This sentiment is also echoed in works such as [72], where the authors studied the downstream consequences of uninformed cavalier MT usage in the healthcare and law domains and emphasized on *broad societal need for higher levels of awareness of the specific strengths and crucially, of the limitations of MT*. In Figure 24 we reproduce Table-1 from the paper that summarizes the on *Findings on Perception, Use and Impact of MT in Medical and Legal Settings* for the reader's perusal.

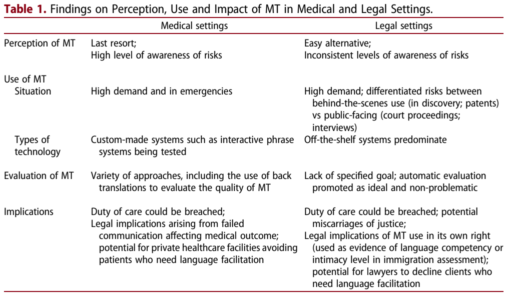
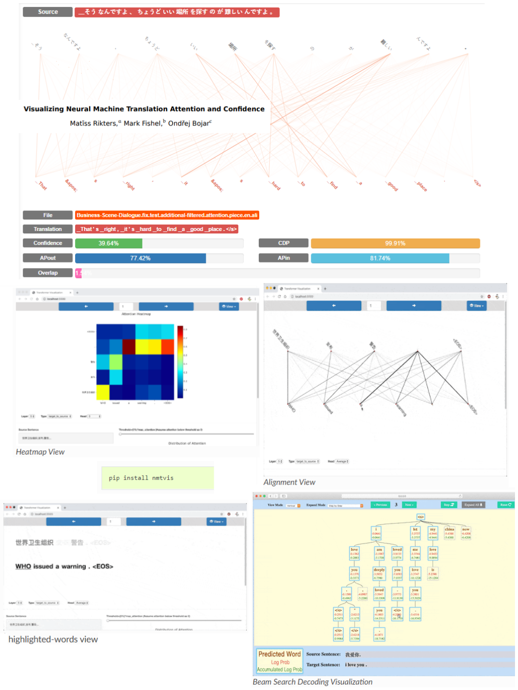


Table 1. Findings on Perception, Use and Impact of MT in Medical and Legal Settings.

	Medical settings	Legal settings
Perception of MT	Last resort; High level of awareness of risks	Easy alternative; Inconsistent levels of awareness of risks
Use of MT		
Situation	High demand and in emergencies	High demand; differentiated risks between behind-the-scenes use (in discovery; patents) vs public-facing (court proceedings; interviews)
Types of technology	Custom-made systems such as interactive phrase systems being tested	Off-the-shelf systems predominate
Evaluation of MT	Variety of approaches, including the use of back translations to evaluate the quality of MT	Lack of specified goal; automatic evaluation promoted as ideal and non-problematic
Implications	Duty of care could be breached; Legal implications arising from failed communication affecting medical outcome; potential for private healthcare facilities avoiding patients who need language facilitation	Duty of care could be breached; potential miscarriages of justice; Legal implications of MT use in its own right (used as evidence of language competency or intimacy level in immigration assessment); potential for lawyers to decline clients who need language facilitation

Figure 24: Table-1 reproduced from [72]



Visualizing Neural Machine Translation Attention and Confidence

Matīss Rikters,[a] Mark Fishel,[b] Ondřej Bojar[c]

Source: _そう なんですよ 、 ちょうど いい 場所 を探す の が 難しい んですよ 。

File: Business-Scene-Dialogue.fix.test.additional-filtered.attention.piece.en.ali

Translation: _That ' s _right , _it ' s _hard _to _find _a _good _place . </s>

Confidence: 39.64%

CDP: 99.91%

APout: 77.42%

APin: 81.74%

```
pip install nmtvis
```

世界卫生组织 发布 警告

WHO issued a warning . <EOS>

Source Sentence: 我爱你。

Target Sentence: I love you .

Figure 25: A collage of the visualization tools built by the NMT community

We believe addressing these will help exude thought leadership that will benefit the entire field

[37] `https://www.romecall.org/the-call/`

[38] `https://nmtvis.readthedocs.io/en/latest/`

of NMT, but also pressurize the other AI behemoths to follow suit as many of the other offerings are certainly culpable to some of the frailties demonstrated here.

6 Concluding thoughts

Google translate is today, one of the most influential technologies that operates at a planetary scale, with real-world deployment scenarios varying from cases such as translation of fortune cookies to more serious cases such as those unraveling during a refugee crisis [2]. While a reductionist understanding positions it as being a mere digital language translation tool, works such as [4] have explored how GT elegantly fits within Google's overarching mandate of being the *switchboard of the global flow of information* whilst providing the proverbial front seat viewing of inter-cultural exchanges and trends unraveling in real-time.

Our goal in authoring this paper, is not to merely concoct a *hit-piece* by cleverly harnessing some idiosyncratic *corner-case* shortcomings of GT. In this regard, we absolutely concur that many of these shortcomings pointed to in this paper are admittedly shared by other MT offerings as well.

In the academic realm, works such as [200, 72] have explored the ramifications of brittleness MT systems in the medical and legal domains. In [200], the authors present this intriguing observation that while NMT models bettered their SMT counterparts (the SoTA ones as per 2017) for domains such as IT and Subtitles, SMT models outperformed NMT models for other domains including Law and Medical. (See Figure 26 reproduced from the paper)

Out in the real world, Facebook had to apologize for two major gaffes, one involving the translating of Chinese President Xi Jinping's name [201] and the other involving a post saying "good morning" that was erroneously translated as "attack them" in Hebrew (and "hurt them" in English) leading to an innocent person's arrest. Similarly, *WeChat* had to apologize for its auto-translation API translating an emoji of the Canadian flag as "*He's in prison*" and the flag of the Democratic Republic of the Congo as "*He's dead*".

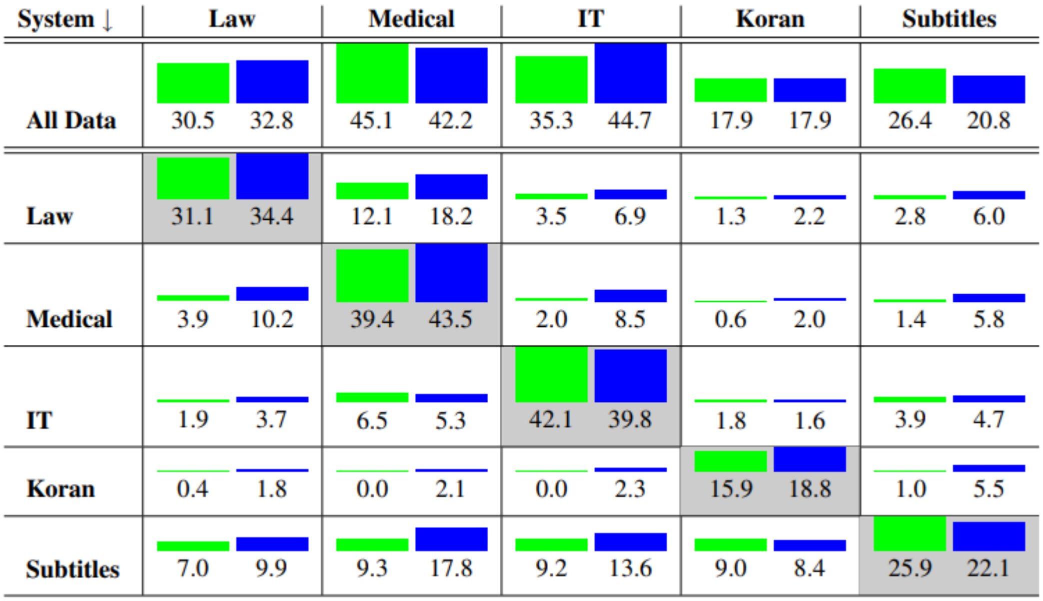
We are increasingly concerned not only about the innate *brittleness* of the NMT models that power this technology but also how it continues to be deployed within an *all-bets-are-off* and *use-it-at-your-own-peril* framework. A framework where the average end user is not only completely unaware of the vulnerability of the technology to small changes in the input (like case or punctuation marks) but is also clueless when the translations might be reliably used for what downstream application scenarios.

A *ProPublica* investigation [202] revealed that U.S. Citizenship and Immigration Services (USCIS) instructed its officers that "*the most efficient approach to translate foreign language contents is to utilize one of the many free online language translation services provided by Google, Yahoo, Bing, and other search engines*" in the context of sifting through non-English social media posts of refugee applicants. In Sophie Zhang's recent whistle-blower account on the misinformation tactics of the authoritarian regime in Azerbaijan [203], there was this chilling revelation that *As of August 2020, Facebook did not have any full-time or contract operations employees who were known to speak Azeri, leaving staff to use* ***Google Translate to try to understand the nature of the abuse***. This when seen in the context of poor SSR performance for Azerbaijani (See Figure 12) and the *The court ordered violence!* backtranslation in Table 30 paints a bleak picture. These revelations have also worryingly coincided with two other developments to be noted: The rise of back-translated text being increasingly used to augment training data in so termed low-resource natural language processing (NLP) scenarios and the emergence of '*AI-enhanced legal-tech*' as a panacea that promises '*disruptive democratization*' of access to legal services laced with grandiose quips such *The greatest impact of AI will be in democratizing legal services* [204].

In the backdrop of these quandaries, that we have invested efforts to highlight two specific failings: Androcentrism and Enantiosemy, in this paper. To this end, we curated datasets covering these two vectors of vulnerability and performed empirical exercises, both on the qualitative and the quantitative front, to estimate the extent of the frailties. Based on the nuances gleaned via these experiments and the related literature of critique, we have drawn out an entire landscape of aspirational recommendations that we believe addresses *some* of the concerns raised. In doing so, we solemnly acknowledge the trap of *techsolutionism* [205] and also acknowledge that we do not hold any special agency as auditors-in-general of NMT-ethics and most certainly attest to not be breathing the rarefied air of the mythical moral high ground. We hope that the cautionary tales we have served in this work will strike important conversations and hopefully some changes at least, on the app or the API.

Acknowledgement

This is a community-driven working paper, and represents research in progress. We have envisioned this to be a collection of ideas contributed by scholars and practitioners alike, and would like



System ↓	Law		Medical		IT		Koran		Subtitles	
All Data	30.5	32.8	45.1	42.2	35.3	44.7	17.9	17.9	26.4	20.8
Law	31.1	34.4	12.1	18.2	3.5	6.9	1.3	2.2	2.8	6.0
Medical	3.9	10.2	39.4	43.5	2.0	8.5	0.6	2.0	1.4	5.8
IT	1.9	3.7	6.5	5.3	42.1	39.8	1.8	1.6	3.9	4.7
Koran	0.4	1.8	0.0	2.1	0.0	2.3	15.9	18.8	1.0	5.5
Subtitles	7.0	9.9	9.3	17.8	9.2	13.6	9.0	8.4	25.9	22.1

Figure 1: Quality of systems (BLEU), when trained on one domain (rows) and tested on another domain (columns). Comparably, NMT systems (left bars) show more degraded performance out of domain.

Figure 26: ”Figure 1” reproduced from [200]

to acknowledge the contribution of the organizers of the Africa-NLP-EACL 2021 workshop for facilitating the initial exchange of ideas. We’d like to thank Randip Singh for informing us of the specific shortcomings observed with regards to English to Punjabi translations that inspired some of the experiments here. We’d also like to thank Abubakar Abid for both, sparking the initial analysis into contronyms[39] and for further discussions on the nuances of mistranslations into Urdu. Last but not the least, we’d like to thank the anonymous reviewers of the Africa-NLP-EACL 2021 workshop for their extremely helpful reviews pertaining to an earlier draft of the paper.

[39] https://twitter.com/abidlabs/status/1338965268232622088?s=20

References

[1] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

[2] Gideon Lewis-Kraus. The great A.I. awakening. `https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html`, Dec 2016. (Accessed on 04/28/2021).

[3] Douglas Hofstadter. The shallowness of google translate. `https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/`, Jan 2018. (Accessed on 04/28/2021).

[4] Ido Ramati and Amit Pinchevski. Uniform multilingualism: A media genealogy of google translate. *New Media & Society*, 20(7):2550–2565, 2018.

[5] Kyle Bradshaw. Google translate reaches 1 billion play store downloads. `https://9to5google.com/2021/04/05/google-translate-1-billion-play-store-downloads/`, April 2021. (Accessed on 05/19/2021).

[6] Barak Turovsky. Ten years of google translate. `https://blog.google/products/translate/ten-years-of-google-translate/`, Apr 2016. (Accessed on 04/28/2021).

[7] Mike Siko. Concealing network traffic via google translate. `https://practicalmalwareanalysis.com/2012/02/14/concealing-network-traffic-via-google-translate/`, Feb 2012. 'Running the Gauntlet' blog (Accessed on 05/27/2021).

[8] Ari Juels and Ting-Fang Yen. Sherlock holmes and the case of the advanced persistent threat. In *5th {USENIX} Workshop on Large-Scale Exploits and Emergent Threats ({LEET} 12)*, 2012.

[9] Larry Cashdollar. Phishing attacks against facebook / google via google translate. `https://blogs.akamai.com/sitr/2019/02/phishing-attacks-against-facebook-google-via-google-translate.html`, Feb 2019. Akamai Security Intelligence and Threat Research Blog.

[10] Nick Ames. Google translate: the unlikely world cup hero breaking barriers for fans - The Guardian. `https://www.theguardian.com/football/2018/jul/11/google-translate-world-cup-hero-fans-language-barriers`, July 2018. (Accessed on 05/26/2021).

[11] Cynthia Ducar and Deborah Houk Schocket. Machine translation and the L2 classroom: Pedagogical solutions for making peace with google translate. *Foreign Language Annals*, 51(4):779–795, 2018.

[12] Marco Cancino and Jaime Panes. The impact of google translate on L2 writing quality measures: Evidence from chilean EFL high school learners. *System*, 98:102464, 2021.

[13] Shu-Chiao Tsai. Chinese students' perceptions of using google translate as a translingual call tool in efl writing. *Computer Assisted Language Learning*, pages 1–23, 2020.

[14] Shu-Chiao Tsai. Using google translate in efl drafts: a preliminary investigation. *Computer Assisted Language Learning*, 32(5-6):510–526, 2019.

[15] Hossein Bahri and Tengku Sepora Tengku Mahadi. Google translate as a supplementary tool for learning malay: A case study at universiti sains malaysia. *Advances in Language and Literary Studies*, 7(3):161–167, 2016.

[16] Elaine C Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. Assessing the use of google translate for spanish and chinese translations of emergency department discharge instructions. *JAMA internal medicine*, 179(4):580–582, 2019.

[17] Breena R Taira, Vanessa Kreger, Aristides Orue, and Lisa C Diamond. A pragmatic assessment of google translate for emergency department instructions. *Journal of General Internal Medicine*, pages 1–5, 2021.

[18] Terena Bell. Google translate causes vaccine mishap | Multilingual. `https://multilingual.com/google-translate-causes-vaccine-mishap/`, Jan 2021. (Accessed on 05/01/2021).

[19] John Hutchins. Warren weaver and the launching of mt. *Early Years in Machine Translation, ed. W. John Hutchins. Amsterdam: John Benjamins. 20o0*, pages 17–20, 2000.

[20] W John Hutchins. *Early years in machine translation: memoirs and biographies of pioneers*, volume 97. John Benjamins Publishing, 2000.

[21] John Hutchins. ALPAC: the (in) famous report. *Readings in machine translation*, 14:131–135, 2003.

[22] Warren Weaver. Translation. memorandum. reprinted in WN Locke and AD Booth, eds. *Machine Translation of Languages: Fourteen Essays*, 1949.

[23] John Hutchins. Machine translation: general overview. *The Oxford handbook of computational linguistics*, 2003.

[24] John Hutchins. Machine translation: A concise history. *Computer aided translation: Theory and practice*, 13(29-70):11, 2007.

[25] Jonathan Slocum. A survey of machine translation: Its history, current status and future prospects. *Computational linguistics*, 11(1):1–17, 1985.

[26] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[27] Philipp Koehn. Neural machine translation. *arXiv preprint arXiv:1709.07809*, 2017.

[28] Nadeem Jadoon Khan, Waqas Anwar, and Nadir Durrani. Machine translation approaches and survey for indian languages. *arXiv preprint arXiv:1701.04290*, 2017.

[29] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*, 2018.

[30] Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*, 2020.

[31] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A comprehensive survey of multilingual neural machine translation. *arXiv preprint arXiv:2001.01115*, 2020.

[32] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[33] Kenzie Shofner. Inspiring and notable women throughout translation history. `https://www.unitedlanguagegroup.com/blog/translation/inspiring-and-notable-women-throughout-translation-history`, Jan 2019. (Accessed on 05/25/2021).

[34] Kari Boyd McBride. *Native Mothers, Native Others: La Malinche, Pocahontas, and Sacajawea*. Routledge, 2019.

[35] Kristina Downs. *The" Enamored Indian Princess" Narrative: Race, Sexuality, and Ancestry in the Stories of La Malinche, Pocahontas, and Sacagawea*. PhD thesis, Indiana University, 2017.

[36] Rebecca Kay Jager. *Malinche, Pocahontas, and Sacagawea: indian women as cultural intermediaries and national symbols*. University of Oklahoma Press, 2015.

[37] Norma Alarcón. Traddutora, traditora: A paradigmatic figure of chicana feminism. *Cultural Critique*, (13):57–87, 1989.

[38] Shari M Huhndorf. Scenes from the fringe: Gendered violence and the geographies of indigenous feminism. *Signs: Journal of Women in Culture and Society*, 46(3):561–587, 2021.

[39] Barry Schwartz. Google translate drops systran for home brewed translation. `https://searchengineland.com/google-translate-drops-systran-for-home-brewed-translation-12502`, Oct 2007. (Accessed on 05/02/2021).

[40] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[41] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv preprint arXiv:1811.06965*, 2018.

[42] Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*, 2019.

[43] Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8854–8861, 2020.

[44] Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*, 2019.

[45] Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

[46] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[47] Ondřej Bojar and Aleš Tamchyna. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, 2011.

[48] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, 2016.

[49] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

[50] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, 2018.

[51] Isaac Feldman and Rolando Coto-Solano. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[52] Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. Analysis of back-translation methods for low-resource neural machine translation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 466–475. Springer, 2019.

[53] Hongzheng Li, Jiu Sha, and Can Shi. Revisiting back-translation for low-resource machine translation between chinese and vietnamese. *IEEE Access*, 8:119931–119939, 2020.

[54] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

[55] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*, 2018.

[56] Claude Coulombe. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*, 2018.

[57] Sam Shleifer. Low resource text classification with ulmfit and backtranslation. *arXiv preprint arXiv:1903.09244*, 2019.

[58] Effi Levi, Guy Mor, Shaul Shenhav, and Tamir Sheafer. Compres: A dataset for narrative structure in news. *arXiv preprint arXiv:2007.04874*, 2020.

[59] Jun Ma and Langlang Li. Data augmentation for chinese text classification using back-translation. In *Journal of Physics: Conference Series*, volume 1651, page 012039. IOP Publishing, 2020.

[60] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020.

[61] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*, 2019.

[62] Johanes Effendi, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. Multi-paraphrase augmentation to leverage neural caption translation. In *International Workshop on Spoken Language Translation*, 2018.

[63] Thomas Body, Xiaohui Tao, Yuefeng Li, Lin Li, and Ning Zhong. Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models. *Expert Systems with Applications*, page 115033, 2021.

[64] Edward Ma. NLP augmentation. https://github.com/makcedward/nlpaug, 2019.

[65] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[66] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[67] Susan Ervin and Robert T Bower. Translation problems in international surveys. *Public Opinion Quarterly*, 16(4):595–604, 1952.

[68] Richard W Brislin. Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, 1(3):185–216, 1970.

[69] Prithwijit Das, Anna Kuznetsova, Meng'ou Zhu, and Ruth Milanaik. Dangers of machine translation: The need for professionally translated anticipatory guidance resources for limited english proficiency caregivers. *Clinical pediatrics*, 58(2):247–249, 2019.

[70] Jonathan Downie and Angela Dickson. Unsound evaluations of medical machine translation risk patient health and confidentiality. *JAMA internal medicine*, 179(7):1001–1001, 2019.

[71] Sonia Colina, Nicole Marrone, Maia Ingram, and Daisey Sánchez. Translation quality assessment in health research: A functionalist alternative to back-translation. *Evaluation &amp; the health professions*, 40(3):267–293, 2017.

[72] Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication &amp; Society*, pages 1–18, 2020.

[73] Alexei Shmelev. Semantic shifts as sources of enantiosemy. *The Lexical Typology of Semantic Shifts*, 58:67, 2016.

[74] Alexei Shmelev. Cognitive and communicative sources of enantiosemy. *Proceedings of the 10th World Congress of the International Association for Semiotic Studies (IASS/AIS)*, pages 837–844, 2012.

[75] Burcu I Karaman. On contronymy. *International Journal of Lexicography*, 21(2):173–192, 2008.

[76] Richard Lederer. Curious contronyms. *Word Ways*, 11(1):10, 1978.

[77] Stefan DuBois. The deceptively simple problem of contronymy. *Convenit Internacional*, 27:15–30, 2018.

[78] Vladimir Ozyumenko. Enantiosemy in legal english. *INTCESS 2019- 6th International Conference on Education and Social Sciences*, pages 72–78, 2019.

[79] Josef Filipec. Česká lexikologie. *Studie a práce lingvistické, sv. 20*, 1985.

[80] Aleš Klégr. The limits of polysemy: enantiosemy. *Linguistica Pragensia*, 23(2):7–23, 2013.

[81] Wiktionary. Appendix:english contranyms. `https://en.wiktionary.org/wiki/Appendix:English_contranyms`, January 2021. (Accessed on 01/30/2021).

[82] John Burkardt. Autoantonyms - the same, and different. `https://people.sc.fsu.edu/~jburkardt/fun/wordplay/autoanto.html`, May 2020. (Accessed on 01/30/2021).

[83] D Charles Hair. Legalese: A legal argumentation tool. *ACM SIGCHI Bulletin*, 23(1):71–74, 1991.

[84] Mette Hjort-Pedersen and Dorrit Faber. Lexical ambiguity and legal translation: a discussion. *Multilingua*, 20(4):379–392, 2001.

[85] VI Ozyumenko and Kamo P Chilingaryan. Polysemy of english legal lexis and the problems of translation. *Russian Journal of linguistics*, pages 180–193, 2015.

[86] Peter Butt. Legalese versus plain language. *Amicus Curiae*, 2001(35):28–32, 2001.

[87] Mark Nichol. 75 contronyms (words with contradictory meanings). `https://www.dailywritingtips.com/75-contronyms-words-with-contradictory-meanings/`, Sep 2011. (Accessed on 05/02/2021).

[88] Judith Herman. 25 words that are their own opposites | Mental Floss. `https://www.mentalfloss.com/article/57032/25-words-are-their-own-opposites`, June 2018. (Accessed on 05/02/2021).

[89] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[90] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[91] Catherine Way. The challenges and opportunities of legal translation and translator training in the 21st century. *International Journal of Communication*, 10:1009–1029, 2016.

[92] Joanna Drugan, Ingemar Strandvik, and Erkka Vuorinen. Translation quality, quality management and agency: Principles and practice in the european union institutions, 2018.

[93] Poon Wai Yee Emily. The cultural transfer in legal translation. *International Journal for the Semiotics of Law*, 18:307–323, 2005.

[94] David Eckels Wade and Ariane Tabatabai. How mistranslation could threaten the Iran deal. https://www.theatlantic.com/international/archive/2017/09/mistranslation-iran-deal/538770/, 2017. (Accessed on 01/30/2021).

[95] Ronald Eric Emmerick. Iranian languages. https://www.britannica.com/topic/Iranian-languages. Accessed on 05/25/2021.

[96] Britannica. South Slavic languages. https://www.britannica.com/topic/South-Slavic-languages. Accessed on 05/25/2021.

[97] Thierry Etchegoyhen and Harritxu Gete. To case or not to case: Evaluating casing methods for neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3752–3760, 2020.

[98] Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. Evaluating robustness to input perturbations for neural machine translation. *arXiv preprint arXiv:2005.00580*, 2020.

[99] Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. On evaluation of adversarial perturbations for sequence-to-sequence models. *arXiv preprint arXiv:1903.06620*, 2019.

[100] Lijun Wu, Yingce Xia, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Adversarial neural machine translation. In *Asian Conference on Machine Learning*, pages 534–549. PMLR, 2018.

[101] Javid Ebrahimi, Daniel Lowd, and Dejing Dou. On adversarial examples for character-level neural machine translation. *arXiv preprint arXiv:1806.09030*, 2018.

[102] Yong Cheng, Lu Jiang, and Wolfgang Macherey. Robust neural machine translation with doubly adversarial inputs. In *Association for Computational Linguistics (ACL)*, 2019.

[103] Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. Evaluating robustness to input perturbations for neural machine translation. *arXiv preprint arXiv:2005.00580*, 2020.

[104] Thierry Etchegoyhen and Harritxu Gete. To case or not to case: Evaluating casing methods for neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3752–3760, 2020.

[105] Pascal Mark Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in psychology*, 10:1604, 2019.

[106] Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. Representation of the sexes in language. *Social communication*, pages 163–187, 2007.

[107] Alexander Z Guiora. Language and concept formation: A cross-lingual analysis. *Behavior Science Research*, 18(3):228–256, 1983.

[108] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

[109] Greville G. Corbett. Number of genders. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

[110] Greville G. Corbett. Sex-based and non-sex-based gender systems. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

[111] Greville G. Corbett. Systems of gender assignment. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

[112] Aditya Siddhant, Junjie Hu, Melvin Johnson, Orhan Firat, and Sebastian Ruder. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. 2020.

[113] Alexander Gutkin, Tatiana Merkulova, and Martin Jansche. Predicting the features of world atlas of language structures from speech. 2018.

[114] Levi CR Hord. Bucking the linguistic binary: Gender neutral language in english, swedish, french, and german. *Western Papers in Linguistics/Cahiers linguistiques de Western*, 3(1):4, 2016.

[115] Marcelo OR Prates, Pedro H Avelar, and Luis C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19, 2019.

[116] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. On measuring gender bias in translation of gender-neutral pronouns. *arXiv preprint arXiv:1905.11684*, 2019.

[117] Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. A case study of

natural gender phenomena in translation a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *Association for Machine Translation in the Americas (AMTA): Workshop on the Impact of Machine Translation (iMpacT 2020)*, page 62. Workshop on the Impact of Machine Translation (iMpacT 2020) at Association …, 2020.

[118] Anna Farkas and Renáta Németh. How to measure gender bias in machine translation: Optimal translators, multiple reference points. *arXiv preprint arXiv:2011.06445*, 2020.

[119] Nicolas Kayser-Bril. Female historians and male nurses do not exist, Google Translate tells its european users - Algorithmwatch. https://algorithmwatch.org/en/google-translate-gender-bias/, September 2020. (Accessed on 05/01/2021).

[120] Ruha Benjamin. Race after technology: Abolitionist tools for the new Jim code. *Social Forces*, 2019.

[121] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. nyu Press, 2018.

[122] Jesse Emspak. Facing facts: Artificial intelligence and the resurgence of physiognomy. https://undark.org/2017/11/08/facing-facts-artificial-intelligence/, November 2017. (Accessed on 01/31/2021).

[123] Thomas Laqueur. *Making sex: Body and gender from the Greeks to Freud*. Harvard University Press, 1992.

[124] Ruth Evans. *Simone de Beauvoir's The second sex: New interdisciplinary essays*. Manchester University Press, 1998.

[125] Dale Spender. *Man Made Language*. Routledge, 1985.

[126] Heiko Motschenbacher. Language, gender and sexual identity. *Poststructuralist perspectives, Amsterdam/Philadelphia*, 2010.

[127] James Zou and Londa Schiebinger. AI can be sexist and racist-It's time to make it fair. *Nature*, 559(7714):324–326, 2018.

[128] Ian Woolford. India pride month 2020: Can Hindi language be gender inclusive? here's how language can free you. https://bit.ly/2YsSbHM, June 2020. (Accessed on 01/30/2021).

[129] Neetu Khanna et al. Three experiments in subaltern intimacy. *Postcolonial Text*, 13(4), 2018.

[130] Rakhshanda Jalil. *A rebel and her cause: the life and work of Rashid Jahan*. Women Unlimited New Delhi, 2014.

[131] Celia Davies. The sociology of professions and the profession of gender. *Sociology*, 30(4):661–678, 1996.

[132] Tracey L Adams. Gender and feminization in health care professions. *Sociology compass*, 4(7):454–465, 2010.

[133] Merriam-Webster. Professions. https://learnersdictionary.com/3000-words/topic/jobs-professions/1, 2021. (Accessed on 01/30/2021).

[134] Milos Bejda. Professions.txt. https://bit.ly/3aiPeiF, 2019. (Published as a Github Gist and accessed on 01/30/2021).

[135] Sai Prasanna. lmproof - Language model proof reader. https://pypi.org/project/lmproof/, May 2020. (PyPI package. Accessed on 01/31/2021).

[136] Lu Shan. google-trans-new | pypi package. https://pypi.org/project/google-trans-new/, Dec 2020. (Accessed on 01/31/2021).

[137] Jr. Aremu Adeola. Lost in translation: Why google translate often gets Yorùbá — and other languages — wrong. https://bit.ly/3jxjjiz, March 2020. (Accessed on 02/06/2021).

[138] Yisa Kehinde Yusuf. Sexism, English and Yoruba. *Linguistik online*, 11(2), 2002.

[139] Kay Williamson and Roger Blench. Niger-congo. In Bernd Heine and Derek Nurse, editors, *African Languages: An Introduction*, pages 11–42. Cambridge University Press, Cambridge, 2000.

[140] Christiana Ngozi Ikegwuonu. *An Exploration of Gender System in Igbo Language*. PhD thesis, Open journal of modern linguistics, 2019.

[141] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[142] Brian N Larson. Gender as a variable in natural-language processing: Ethical considerations. *EACL 2017*, page 1, 2017.

[143] Szymon Misiek et al. Misgendered in translation? *ANGLICA-An International Journal of English Studies*, 29(2):165–185, 2020.

[144] Ann Bodine. Androcentrism in prescriptive grammar: Singular 'they', sex-indefinite 'he', and 'he or she'. *Language in society*, pages 129–146, 1975.

[145] Marijana Šincek. On, ona, ono: Translating gender neutral pronouns into croatian. *Zbornik radova Međunarodnog simpozija mladih anglista, kroatista i talijanista*, pages 92–112, 2020.

[146] Hongwei Li, Bin Yu, and Dengyong Zhou. Error rate analysis of labeling by crowdsourcing. In *ICML Workshop: Machine Learning Meets Crowdsourcing. Atalanta, Georgia, USA*. Citeseer, 2013.

[147] Matthäus Kleindessner and Pranjal Awasthi. Crowdsourcing with arbitrary adversaries. In *International Conference on Machine Learning*, pages 2708–2717. PMLR, 2018.

[148] Andrew Zaldivar, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Lucy Vasserman, M. Mitchell, Parker Barnes, Simone Sanoian McCloskey Wu, and Timnit Gebru, editors. *Model Cards for Model Reporting*, 2019.

[149] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321, 2004.

[150] Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019.

[151] Fei Huang. Translation confidence scores, November 20 2018. US Patent 10,133,738.

[152] Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR, 2018.

[153] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4694–4703. Curran Associates, Inc., 2019.

[154] Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*, 2020.

[155] Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C Lipton. Decoding and diversity in machine translation. *arXiv preprint arXiv:2011.13477*, 2020.

[156] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016.

[157] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664*, 2017.

[158] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):1–6, 2021.

[159] Eliza Strickland. Openai's GPT-3 speaks! (Kindly disregard toxic language) - IEEE spectrum. `https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/open-ais-powerful-text-generating-tool-is-ready-for-business`, Feb 2021. (Accessed on 04/17/2021).

[160] Chen Yanover, Mona Singh, and Elena Zaslavsky. M are better than one: an ensemble-based motif finder and its application to regulatory element prediction. *Bioinformatics*, 25(7):868–874, 2009.

[161] John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg. Protein design by sampling an undirected graphical model of residue constraints. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(3):506–516, 2008.

[162] Chen Yanover and Yair Weiss. Finding the M most probable configurations using loopy belief propagation. *Advances in neural information processing systems*, 16:289–296, 2003.

[163] Menachem Fromer and Amir Globerson. An LP view of the M-best MAP problem. *Advances in Neural Information Processing Systems*, 22:567–575, 2009.

[164] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse M-best solutions in Markov Random Fields. In *European Conference on Computer Vision*, pages 1–16. Springer, 2012.

[165] Melvin Johnson. Google AI blog: A scalable approach to reducing gender bias in google translate. `https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html`, Apr 2020. (Accessed on 06/02/2021).

[166] Cassian Lodge. Gender census 2021: Worldwide summary – gender census. https://gendercensus.com/results/2021-worldwide-summary/, March 2021. (Accessed on 06/02/2021).

[167] S Young. Frederick Jelinek 1932-2010: The pioneer of speech recognition technology. *Speech and Language Processing Technical Committee Newsletter*, 2010.

[168] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

[169] Morten H Christiansen and Nick Chater. Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, 5(2):82–88, 2001.

[170] Marc Brysbaert, Emmanuel Keuleers, and Paweł Mandera. A plea for more interactions between psycholinguistics and natural language processing research. *Computational Linguistics in the Netherlands Journal*, 4:209–222, 2014.

[171] Basilio Calderone, Nabil Hathout, and Franck Sajous. From GLàff to PsychoGLàff: a large psycholinguistics-oriented french lexical resource. In *Euralex*, pages 431–446, 2014.

[172] Clara Meister, Tim Vieira, and Ryan Cotterell. If beam search is the answer, what was the question? *arXiv preprint arXiv:2010.02650*, 2020.

[173] Herbert H Clark and Eve V Clark. Psychology and language: An introduction to psycholinguistics. 1977.

[174] Antonio Laverghetta Jr. *Exploring the Use of Neural Transformers for Psycholinguistics.* PhD thesis, University of South Florida, 2021.

[175] Jost Zetzsche. Women and Machine Translation – American Translators Association (ATA). https://www.atanet.org/tools-and-technology/women-and-machine-translation/, November 2020. (Accessed on 05/25/2021).

[176] Hazel Mae Pan. How BLEU measures translation and why it matters | Slator. https://slator.com/technology/how-bleu-measures-translation-and-why-it-matters/, November 2016. (Accessed on 06/03/2021).

[177] Nitika Mathur, Tim Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*, 2020.

[178] Aaron Smith, Christian Hardmeier, and Jörg Tiedemann. Climbing mont BLEU: the strange world of reachable high-BLEU translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 269–281, 2016.

[179] Besat Zardosht and Ahmad Abdollahzadeh Barforoush. Aut-bleu: Extending bleu with use of parse tree. *Tenth Symposium on Natural Language Processing*, 2013.

[180] Some improvements over the BLEU metric for measuring translation quality for Hindi. In *2007 International Conference on Computing: Theory and Applications (ICCTA'07)*, pages 485–490. IEEE, 2007.

[181] Michael Gamon, Anthony Aue, and Martine Smets. Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*, 2005.

[182] Matt Post. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*, 2018.

[183] Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*, 2018.

[184] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[185] Isaac Caswell and Bowen Liang. Google AI blog: Recent advances in google translate. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html, Jun 2020. (Accessed on 06/02/2021).

[186] James Vincent. Apple boasts about sales; google boasts about how good its ai is - the verge. https://www.theverge.com/2016/10/4/13122406/google-phone-event-stats, October 2016. (Accessed on 06/04/2021).

[187] Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, pages 1–15, 2021.

[188] Gisela Swaragita. Aceh NGO to sue google over alleged racist translations. https://www.thejakartapost.com/news/2019/10/22/aceh-ngo-sue-google-over-alleged-racist-translations.html, October 2019. (Accessed on 06/03/2021).

[189] Emily Bender. The BenderRule: On naming the languages we study and why it matters. *The Gradient*, 2019.

[190] Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, 2016.

[191] Cliff Goddard et al. De-anglicising humour studies. *The European Journal of Humour Research*, 8(4):48–58, 2020.

[192] Carsten Levisen. Biases we live by: Anglocentrism in linguistics and cognitive sciences. *Language Sciences*, 76:101173, 2019.

[193] Adrian Pablé. Three critical perspectives on the ontology of "language". In *Integrational Linguistics and Philosophy of Language in the Global South*, pages 30–47. Routledge.

[194] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*, 2021.

[195] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*, 2020.

[196] Davis David. 10 best African language datasets for data science projects | Hacker Noon. https://hackernoon.com/10-best-african-language-datasets-for-data-science-projects-l51h34xw, May 2021. (Accessed on 06/01/2021).

[197] Jessica Chandras. Multilingualism in India. *Education About ASIA*, 25(3), 2020.

[198] Urvi Malvania. Print readership in India jumps 4.4% to 425 million in two years: Report | Business Standard News. https://webcache.googleusercontent.com/search?q=cache:D9x7ZAOQkjsJ:https://www.business-standard.com/article/current-affairs/print-readership-in-india-jumps-4-4-to-425-million-in-two-years-report-119042700079_1.html+&cd=11&hl=en&ct=clnk&gl=us, Apr 2019. (Accessed on 04/29/2021).

[199] Matīss Rikters, Mark Fishel, and Ondřej Bojar. Visualizing neural machine translation attention and confidence. *The Prague Bulletin of Mathematical Linguistics*, 109(1):39, 2017.

[200] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.

[201] Mariel Padilla. Facebook apologizes for vulgar translation of Chinese leader's name - The New York Times. https://www.nytimes.com/2020/01/18/world/asia/facebook-xi-jinping.html, Jan 2020. (Accessed on 01/31/2021).

[202] Yeganeh Torbati. Google says google translate can't replace human translators. immigration officials have used it to vet refugees. — ProPublica. https://bit.ly/2YvVsGm, Sep 2019. (Accessed on 01/31/2021).

[203] Julia Carrie Wong. How Facebook let fake engagement distort global politics: a whistleblower's account | The Guardian. https://www.theguardian.com/technology/2021/apr/12/facebook-fake-engagement-whistleblower-sophie-zhang, April 2020. (Accessed on 04/15/2021).

[204] Soojung Chang. A lawyer's guide to artificial intelligence. https://blog.rossintelligence.com/post/lawyers-guide-artificial-intelligence, 2018. (Accessed on 01/30/2021).

[205] Evgeny Morozov. *To save everything, click here: The folly of technological solutionism*. Public Affairs, 2013.

[206] Orhan Fırat. *Connectionist multi-sequence modelling and applications to multilingual neural machine translation*. PhD thesis, Middle East Technical University, 2017.

[207] D Raj Reddy et al. Speech understanding systems: A summary of results of the five-year research effort. *Department of Computer Science. Camegie-Mell University, Pittsburgh, PA*, 17:138, 1977.

[208] Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,

pages 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics.

[209] T Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62, 2010.

[210] August Fenk and Gertraud Fenk. Konstanz im kurzzeitgedächtnis-konstanz im sprachlichen informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, 27:400–414, 1980.

[211] Ramon Ferrer-i Cancho. The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change*, 5(1):114–137, 2015.

[212] Ramon Ferrer-i Cancho. The placement of the head that maximizes predictability. an information theoretic approach. *arXiv preprint arXiv:1705.09932*, 2017.

[213] Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. Uniform information density effects on syntactic choice in hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48, 2018.

[214] Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189, 2009.

Appendix

Appendix A Does the great open basement have glass ceilings?

We begin by explicitly clarifying that we included this section of the appendix as an ode to the scholars of psycholinguistics and sociolinguistic traditions who we interacted with during the course of this authorship and who can broadly be described as *Great-open-basement-skeptics*. Hence, the following paragraphs do admittedly read like and ought to be treated as an opinion-piece.

To begin with, we revisit the words of Guiora that read: [107] (Pg 229), "*Linguistic structures, having to do with gender, time, relations and action, vary between languages* ***and are not readily transposable from one language to another****. Further, figures of speech in different languages offer alternative ways to conceptualise, to express, and perhaps to experience similar events, thus creating the possibility of a shared universe which is not readily accessible to speakers of other languages*". The philosophical underpinnings of these words, one might argue, do run counter to Weaver's vision of the *great open basement* that first emerges in the Section: `Language and Invariants` of [22] and reads: *"But when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers. Thus may it be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication - the real but as yet undiscovered universal language - and then re-emerge by whatever particular route is convenient."*

We argue that questions pertaining to the very existence of such a *great open basement* and its glass ceilings are almost dogmatically left out in NMT literature that are, at their heart, a literal culmination of a near doctrinaire pursuit of Weaver's vision. Large scale projects such as the massively multilingual, massive neural machine translation (M4) approach [41] can well be positioned as the latest installment in the quest for the chimeral *interlingua ideal* [206] (*"yet undiscovered universal language"* in Weaver's words), albeit this time, draped in a connectionist-neural machinery and validated by idiosyncratic accuracy metrics such as BLEU. Whether this approach be able to assuage the long-studied ghosts that haunt the minds of traductologists, sociolinguists and psycholinguists alike and whether the translation error-floors be breached with more data, larger models and more training remains to be seen.

Appendix B A note on beam-search antics in NMT

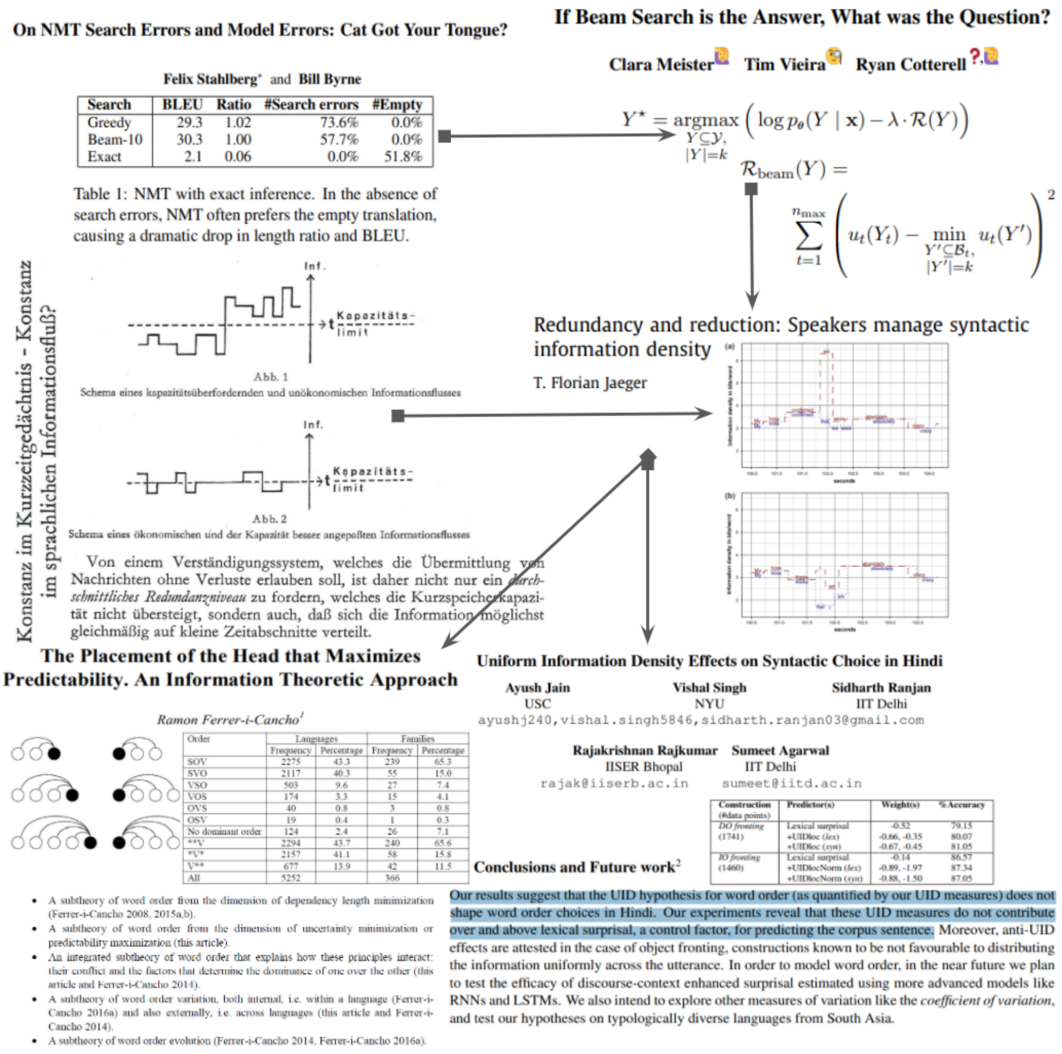
Let us begin with equation (2) which models the inference stage of an NMT system:

$$\mathbf{y}^{*} = \underset{\mathbf{y}\in\mathcal{Y}}{\arg\max}\left(p_{\theta}(\mathbf{y}|\mathbf{x})\right). \tag{2}$$

Here, $\mathbf{x}$ represents the input sequence, $\mathbf{y}^{*}$ represents the output sequence and $\mathcal{Y}$ represents the search-space.

Beneath the veneer of simplicity, this equation hosts a fascinating tale of anachronistic contributions. The rapid advances in the architecture design and training of the large scale neural probabilistic text generation models has meant that the *model*: $p_{\theta}(\mathbf{y}|\mathbf{x})$ (parameterized by weights θ) inherits all the shiny new SoTA components being currently invented. However, the actual decoding algorithm used to approximate the "$\arg\max$" part oft-happens be *Beam search*, a relic of the 1970s!(See [207])

Recently, this quirk has elicited a lot of attention, beginning with the work of [208], who unearthed that *"Surprisingly, beam search fails to find these global best model scores in most cases, even with a very large beam size of 100. For more than 50% of the sentences, the model in fact assigns its global best score to the empty translation, revealing a massive failure of neural models in properly accounting for adequacy,"* in the content of the WMT-15 English-German test set. They also demonstrated how exact maximum a posteriori (MAP) decoding of neural language generators frequently led to unexpectedly low-quality results in terms of the BLEU score achieved. This *unreasonable reasonableness* of beam search, especially with regards to achieving high BLEU scores, was further investigated in [172], where the authors reverse engineered the objective function that beam search returns the exact solution for, and showed how the regularizer needed in the MAP-decoding (from (2)) had a clear connection to the Uniform Information Density (UID) hypothesis [209], whose roots emanate from the information-theoretic work of August Fenk and Gertraud Fenk-Oczlon [210] in 1980. However, this hypothesis, especially with regards to word order modeling has been challenged in the works of Ferrer-i-Cancho [211, 212], who put forth the two sub-theories of word order from the dimension of uncertainty minimization or predictability maximization and dependency length minimiza-



On NMT Search Errors and Model Errors: Cat Got Your Tongue?

Felix Stahlberg* and Bill Byrne

Search	BLEU	Ratio	#Search errors	#Empty
Greedy	29.3	1.02	73.6%	0.0%
Beam-10	30.3	1.00	57.7%	0.0%
Exact	2.1	0.06	0.0%	51.8%

Table 1: NMT with exact inference. In the absence of search errors, NMT often prefers the empty translation, causing a dramatic drop in length ratio and BLEU.

If Beam Search is the Answer, What was the Question?

Clara Meister Tim Vieira Ryan Cotterell ?,

$$Y^\star = \operatorname*{argmax}_{\substack{Y \subseteq \mathcal{Y},\\ |Y|=k}} \Big(\log p_{\boldsymbol{\theta}}(Y \mid \mathbf{x}) - \lambda \cdot \mathcal{R}(Y) \Big)$$

$$\mathcal{R}_{\text{beam}}(Y) = \sum_{t=1}^{n_{\max}} \left(u_t(Y_t) - \min_{\substack{Y' \subseteq \mathcal{B}_t,\\ |Y'|=k}} u_t(Y') \right)^2$$

Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß?

Abb. 1
Schema eines kapazitätsüberfordernden und unökonomischen Informationsflusses

Abb. 2
Schema eines ökonomischen und der Kapazität besser angepaßten Informationsflusses

Von einem Verständigungssystem, welches die Übermittlung von Nachrichten ohne Verluste erlauben soll, ist daher nicht nur ein *durchschnittliches Redundanzniveau* zu fordern, welches die Kurzspeicherkapazität nicht übersteigt, sondern auch, daß sich die Information möglichst gleichmäßig auf kleine Zeitabschnitte verteilt.

Redundancy and reduction: Speakers manage syntactic information density

T. Florian Jaeger

The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach

Ramon Ferrer-i-Cancho[1]

Order	Languages		Families	
	Frequency	Percentage	Frequency	Percentage
SOV	2275	43.3	239	65.3
SVO	2117	40.3	55	15.0
VSO	503	9.6	27	7.4
VOS	174	3.3	15	4.1
OVS	40	0.8	3	0.8
OSV	19	0.4	1	0.3
No dominant order	124	2.4	26	7.1
**V	2294	43.7	240	65.6
V	2157	41.1	58	15.8
V**	677	13.9	42	11.5
All	5252		366	

- A subtheory of word order from the dimension of dependency length minimization (Ferrer-i-Cancho 2008, 2015a,b).
- A subtheory of word order from the dimension of uncertainty minimization or predictability maximization (this article).
- An integrated subtheory of word order that explains how these principles interact: their conflict and the factors that determine the dominance of one over the other (this article and Ferrer-i-Cancho 2014).
- A subtheory of word order variation, both internal, i.e. within a language (Ferrer-i-Cancho 2016a) and also externally, i.e. across languages (this article and Ferrer-i-Cancho 2014).
- A subtheory of word order evolution (Ferrer-i-Cancho 2014, Ferrer-i-Cancho 2016a).

Uniform Information Density Effects on Syntactic Choice in Hindi

Ayush Jain Vishal Singh Sidharth Ranjan
USC NYU IIT Delhi
ayushj240,vishal.singh5846,sidharth.ranjan03@gmail.com

Rajakrishnan Rajkumar Sumeet Agarwal
IISER Bhopal IIT Delhi
rajak@iiserb.ac.in sumeet@iitd.ac.in

Construction (#data points)	Predictor(s)	Weight(s)	% Accuracy
DO fronting (1741)	Lexical surprisal +UIDloc (*lex*) +UIDloc (*syn*)	-0.52 -0.66, -0.35 -0.67, -0.45	79.15 80.07 81.05
IO fronting (1460)	Lexical surprisal +UIDlocNorm (*lex*) +UIDlocNorm (*syn*)	-0.14 -0.89, -1.97 -0.88, -1.50	86.57 87.34 87.05

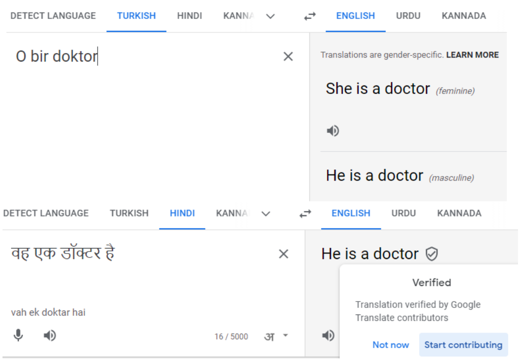
Conclusions and Future work[2]

Our results suggest that the UID hypothesis for word order (as quantified by our UID measures) does not shape word order choices in Hindi. Our experiments reveal that these UID measures do not contribute over and above lexical surprisal, a control factor, for predicting the corpus sentence. Moreover, anti-UID effects are attested in the case of object fronting, constructions known to be not favourable to distributing the information uniformly across the utterance. In order to model word order, in the near future we plan to test the efficacy of discourse-context enhanced surprisal estimated using more advanced models like RNNs and LSTMs. We also intend to explore other measures of variation like the *coefficient of variation*, and test our hypotheses on typologically diverse languages from South Asia.

Figure 27: Surveying the landscape of beam-search inference for NMT

tion. Further, in the context of understanding syntactic choice in Hindi (that has a more flexible word order compared with English), the results in [213] suggested that the UID hypothesis did not seem to shape word order choices (with regards to the Hindi-Urdu Treebank (HUTB) dataset [214]).

In order to aid the reader interested in researching more on this fascinating facet of the NMT pipeline, we have created a landscape snapshot of these recent developments and present them in Figure 27.



O bir doktor

Translations are gender-specific. LEARN MORE

She is a doctor (feminine)

He is a doctor (masculine)

वह एक डॉक्टर है

vah ek doktar hai

16 / 5000

He is a doctor

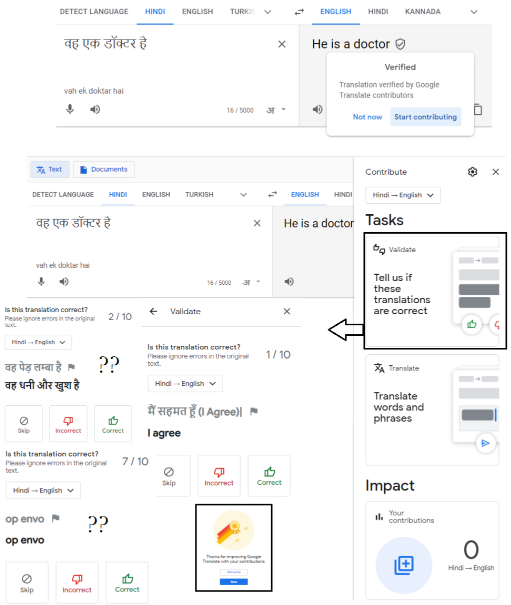
Verified

Translation verified by Google Translate contributors

Figure 28: The difference in gender-specific annotations availability between Turkish and Hindi

Appendix C More on the blues of BLEU

The use of BLEU score to evaluate NMT models has led to the rise in popularity of a variety of heuristics which improve BLEU score, but at the expense of other distribution-level desiderata. As pointed out by [155], two such heuristics are label smoothing and beam search, which in conjunction



वह एक डॉक्टर है

vah ek doktar hai

He is a doctor

Translation verified by Google Translate contributors

Tell us if these translations are correct

Translate words and phrases

Is this translation correct? Please ignore errors in the original text.

वह पेड़ लम्बा है

वह धनी और खुश है

मैं सहमत हूँ (I Agree)|

I agree

op envo

op envo

Thanks for improving Google Translate with your contributions

Figure 29: The annotation process that we encountered upon volunteering on Google translate

improves BLEU score, but biases the output distribution according to other prerequisites. Most notably, the same study implicates beam search as a salient source of bias toward more represented gender pronouns. In particular, the authors observed that when male gender pronouns occurred more often in a German to English training set, beam search inflated the rate of male gender pronouns compared those of both the ground truth and decoding by sampling. In contrast, when the target language was German and the most common pronoun was *sie* (meaning either she or they), beam search inflated the rate of *sie* in the output distribution compared to that of the ground truth and decoding by sampling. In both cases, beam search systematically biases the output distribution of gender pronouns toward the more represented gender. On the other hand, the authors show that beam search so dramatically improves BLEU score that a model trained for only $1/3$ of an epoch but decoded by beam search matches the BLEU score of a fully trained model decoded using a sampling procedure. Remarkably, the fully trained model decoded by beam search gains an additional 14 BLEU points over the model decoded by sampling (at the cost, of course, of human-level desiderata).

Language	En2Language	Language2En
Afrikaans	Die hof beveel die geweld!	The court orders the violence!
Albanian	Gjykata urdhëroi dhunën!	The court ordered the violence!
Amharic	ፍርድ ቤቱ አመፁን አዘዘ!	The court ordered the protest!
Arabic	بالعنف! المحكمة أمرت	Court ordered violence!
Armenian	Դատարանը պատվիրեց բռնությունը:	The court ordered the violence.
Azerbaijani	Məhkəmə şiddəti əmr etdi!	The court ordered violence!
Basque	Auzitegiak indarkeria agindu zuen!	The court ordered violence!
Belarusian	Суд прызначыў гвалт!	The court ordered violence!
Bengali	আদালত সহিংসতার নির্দেশ দিয়েছে!	Court orders violence!
Bosnian	Sud je naredio nasilje!	The court ordered violence!
Bulgarian	Съдът нареди насилието!	The court ordered the violence!
Catalan	El tribunal va ordenar la violència!	The court ordered the violence!
Cebuano	Gisugo sa korte ang kapintas!	The court ordered the violence!
Chinese_Simplified	法院禁止暴力!	The court prohibits violence!
Chinese_Traditional	法院禁止暴力!	The court prohibits violence!
Corsican	U tribunale hà urdinatu a viulenza!	The court ordered the violence!
Croatian	Sud je naredio nasilje!	The court ordered violence!
Czech	Soud nařídil násilí!	The court ordered violence!
Danish	Retten pålagde volden!	The court imposed the violence!
Dutch	De rechtbank beval het geweld!	The court ordered the violence!
English	The court enjoined the violence!	The court enjoined the violence!
Esperanto	La kortumo ordonis la perforton!	The court ordered the violence!
Estonian	Kohus määras vägivalla välja!	The court ordered the violence!
Finnish	Tuomioistuin määräsi väkivallan!	The court ordered the violence!
French	Le tribunal a ordonné la violence !	The court ordered violence!
Frisian	De rjochtbank joech it geweld oan!	The court ordered the violence!
Galician	O xulgado impuxo a violencia!	The court imposed violence!
Georgian	სასამართლომ დააკისრა ძალადობა!	The court ordered the violence!
German	Das Gericht hat die Gewalt vorgeschrieben!	The court prescribed the violence!
Greek	Το δικαστήριο διέταξε τη βία!	The court ordered the violence!
Gujarati	કોર્ટે હિંસાનોઆદેશ આપ્યો!	Court orders violence!
Haitian Creole	Tribinal la mande vyolans lan!	The court demanded the violence!
Hausa	Kotun ta ba da umarnin tashin hankali!	The court ordered violence!
Hawaiian	Ua kauoha ka ‘aha i ka hana ‘ino!	The court ordered the atrocity!
Hebrew	בית המשפט צירף את !האלימות	The court attached the violence!
Hindi	अदालत ने हिंसा को किया स्थगित !	The court suspended the violence!
Hmong	Lub tsev hais plaub sau cov kev kub ntxhov!	The court recorded the violence!
Hungarian	A bíróság elrendelte az erőszakot!	The court ordered the violence!
Icelandic	Dómstóllinn boðaði ofbeldið!	The court announced the violence!
Igbo	Lọ ikpe ahụ nyere iwu ka e mee ihe ike!	The court ordered the violence!
Indonesian	Pengadilan memerintahkan kekerasan!	The court ordered violence!
Irish	Chuir an chúirt an foréigean i gcion air!	The court condemned the violence!
Italian	La corte ha ingiunto la violenza!	The court ordered the violence!
Japanese	裁判所は暴力を禁止しました！	The court has banned violence!
Javanese	Pengadilan mrentah kekerasan kasebut!	The court ruled the violence!
Kannada	ನ್ಯಾಯಾಲಯ ಹಿಂಸಾಚಾರಕ್ಕೆ ಆದೇಶಿಸಿದೆ!	The court has ordered violence!
Kazakh	Сот зорлық-зомбылықты бұйырды!	The court ordered the violence!
Khmer	តុលាការបានរួមបញ្ចូលអំពើហិង្សា!	The court included violence!
Kinyarwanda	Urukiko rwategetse ihohoterwa!	The court ordered the violence!
Korean	법원은 폭력을 조장했습니다!	Court promoted violence!
Kurdish	Dadgehê emrê tundiyê da!	The court ordered the violence!
Kyrgyz	Сот зомбулукка буйрук берди!	The court ordered the violence!
Lao	ສານລອມຄວາມຮຸນແຮງ□	Court of Violence!

Latin	Et atrium per violentiam poterit scrutari vias !	The court will be able to examine ways of violence ?
Latvian	Tiesa piesprieda vardarbību!	The court condemned the violence!
Lithuanian	Teismas nurodė smurtą!	The court ordered violence!
Luxembourgish	D'Geriicht huet d'Gewalt beoptragt!	The court ordered the violence!
Macedonian	Судот нареди насилство!	The court ordered violence!
Malagasy	Nandidy ny herisetra ny fitsarana!	The court ordered the violence!
Malay	Mahkamah memerintahkan keganasan!	The court ordered the violence!
Malayalam	അക്രമത്തിന് കോടതി ഉത്തരവിട്ടു!	Court orders violence!
Maltese	Il-qorti ordnat il-vjolenza!	The court ordered the violence!
Maori	I whakahau te kooti ki te tutu!	The court ordered the violence!
Marathi	कोर्टाने हिंसाचाराचा आदेश दिला !	Court orders violence
Mongolian	Шүүх хүчирхийллийг даалгасан!	The court ordered the violence!
Burmese	တရားရုံးကအကြမ်းဖက်မှုကိုအမိန့်ပေးခဲ့သည်။	The court ordered the violence.
Nepali	अदालतले हिंसाको आदेश दियो!	Court orders violence!
Norwegian	Retten påkalte volden!	The court called for violence!
Nyanja_Chichewa	Khothi lidalamula zachiwawa!	The court ordered the violence!
Oriya	ହିଂସାକୁ କୋର୍ଟ ନିର୍ଦ୍ଦେଶ ଦେଇଛନ୍ତି!	The court ordered the violence!
Pashto	وکړ! امر تریخوالي تاو د محکمې	Court orders violence!
Persian	داد! خشونت به دستور دادگاه	The court ordered violence!
Polish	Sąd nakazał przemoc!	The court ordered the violence!
Portuguese	O tribunal ordenou a violência!	The court ordered the violence!
Punjabi	ਅਦਾਲਤ ਨੇ ਹਿੰਸਾ ਦਾ ਆਦੇਸ਼ ਦਿੱਤਾ!	Court orders violence
Romanian	Curtea a cerut violenței!	The court called for violence!
Russian	Суд предписал насилие!	The court ordered violence!
Samoan	Ua faatonuina e le faamasinoga le vevesi!	The court has ordered the riot!
Scots_Gaelic	Chuir a 'chùirt a-steach an fhòirneart!	The court ruled in violence!
Serbian	Суд је наредио насиље!	The court ordered violence!
Sesotho	Lekhotla le ile la laela pefo!	The court ordered the violence!
Shona	Dare rakaraira mhirizhonga!	The court ordered the riots!
Sindhi	ڏنو! حڪم جو تشدد عدالت	The court ordered the violence!
Sinhalese	අධිකරණය ප්‍රචණ්ඩත්වය අණ කළේය!	Court orders violence!
Slovak	Súd nariadil násilie!	The court ordered violence!
Slovenian	Sodišče je ukazalo nasilju!	The court ordered violence!
Somali	Maxkamaddu waxay amartay rabshadaha!	The court ordered the violence!
Spanish	¡El tribunal ordenó la violencia!	The court ordered violence!
Sundanese	Pengadilan maréntahkeun kekerasan!	The court ordered the violence!
Swahili	Korti iliamuru vurugu!	The court ordered violence!
Swedish	Domstolen förordade våldet!	The court recommended the violence!
Tagalog	Inutusan ng korte ang karahasan!	The court ordered the violence!
Tajik	Суд ба зӯроварӣ фармон дод!	The court ordered the violence!
Tamil	வன்முறையை நீதிமன்றம் கட்டளையிட்டது!	Court orders violence!
Tatar	Суд көч кулланырга кушты!	The court ordered the violence!
Telugu	హింసను కోర్టు ఆదేశించింది!	Court orders torture!
Thai	ตร. กำชับเหตุรุนแรง!	Police charge violent incidents!
Turkish	Mahkeme şiddeti yasakladı!	The court has banned violence!
Turkmen	Kazyýet zorlugy buýurdy!	The court ordered the violence!
Ukrainian	Суд наказав насильству!	The court punished the violence!
Urdu	دیا! حکم کا تشدد نے عدالت	Court orders torture!
Uyghur	بۇيرۇدى! زوراۋانلىقنى سوت	The court ordered the violence!
Uzbek	Sud zo'ravonlikni buyurdi!	The court ordered the violence!
Vietnamese	Tòa án ra lệnh cho bạo lực!	Court ordered violence!
Welsh	Cysylltodd y llys â'r trais!	The court contacted the violence!
Xhosa	Inkundla iyalela ubundlobongela!	The court orders violence!
Yiddish	גוואַלד! די באַשטימט האָט הויף דער	The court ruled the violence!
Yoruba	Kootu paṣẹ fun iwa-ipa!	Court orders violence!
Zulu	Inkantolo yayalela udlame!	The court ordered the violence!

Figure 30: Table containing the results of the latitudinal exploration across all the 109 languages with the sentence ”The court enjoined the violence!”.

Appendix D Acronym Glossary

ALPAC	Automatic Language Processing Advisory Committee
API	Application Programming Interface
AQS	Average Qualifying Sales
BLEU	Bilingual Evaluation Understudy
BLS	Bureau of Labor Statistics
BT	Back-Translation
DL	Deep Learning
EBMT	Example Based Machine Translation
ECE	Expected Calibration Error
EFL	English as a Foreign Language
EOS	End Of Sentence
FAQ	Frequently Asked Questions
FEOR	Hungarian Standard Classification of Occupation
GNMT	Google Neural Machine Translation
GT	Google Translate
HUTB	Hindi-Urdu Treebank
IWSLT	International Conference on Spoken Language Translation
LSTM	Long Short-Term Memory
M4	Massively Multilingual Massive Neural Machine Translation
MAP	Maximum A Posteriori
MoE	Mixture of Experts
MT	Machine Translation
MWh	Mega-Watt Hour
NLP	Natural Language Processing
NMT	Neural Machine Translation
QA	Quality Assessment
RBMT	Rule Based Machine Translation
RNN	Recurrent Neural Network
SMT	Statistical Machine Translation
SoTA	State of the Art
SSR	She-Survival Rate
SST2	Stanford Sentiment Treebank
tCO2e	Tons of C02 Equivalent
TGBI	Translation Gender Bias Index
TPU	Tensor Processing Unit
TTA	Train/Test-Time Augmentation
UDA	Unsupervised Data Augmentation
UID	Uniform Information Density
UMT	Unsupervised Machine Translation
USCIS	U.S. Citizenship and Immigration Services
WALS	World Atlas of Language Structures
WMT	Workshop on statistical Machine Translation

Table 8: Table of acronyms used in this paper