# Gender Representation in Open Source Speech Resources

Mahault Garnerin, Solange Rossato, Laurent Besacier

*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6599–6605
Marseille, 11–16 May 2020


# Gender Representation in Open Source Speech Resources

**Mahault Garnerin, Solange Rossato, Laurent Besacier**
LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr

**Abstract**

With the rise of artificial intelligence (AI) and the growing use of deep-learning architectures, the question of ethics, transparency and fairness of AI systems has become a central concern within the research community. We address transparency and fairness in spoken language systems by proposing a study about gender representation in speech resources available through the Open Speech and Language Resource platform. We show that finding gender information in open source corpora is not straightforward and that gender balance depends on other corpus characteristics (elicited/non elicited speech, low/high resource language, speech task targeted). The paper ends with recommendations about metadata and gender information for researchers in order to assure better transparency of the speech systems built using such corpora.

**Keywords:** speech resources, gender, metadata, open speech language resources (OpenSLR)

## 1. Introduction

The ever growing use of machine learning has put data at the center of the industrial and research spheres. Indeed, for a system to learn how to associate an input X to an output Y, many paired examples are needed to learn this mapping process. This need for data coupled with the improvement in computing power and algorithm efficiency has led to the era of big data. But data is not only needed in mass, but also with a certain level of quality. In this paper we argue that one of the main quality of data is its *transparency*.

In recent years, concerns have been raised about the biases existing in the systems. A well-known case in Natural Language Processing (NLP) is the example of word embeddings, with the studies of Bolukbasi et al. (2016) and Caliskan et al. (2017) which showed that data are socially constructed and hence encapsulate a handful of social representations and power structures, such as gender stereotypes. Gender-bias has also been found in machine translation tasks (Vanmassenhove et al., 2018), as well as facial recognition (Buolamwini and Gebru, 2018) and is now at the center of research debates. In previous work, we investigated the impact of gender imbalance in training data on the performance of an automatic speech recognition (ASR) system, showing that the under-representation of women led to a performance bias of the system for female speakers (Garnerin et al., 2019).

In this paper, we survey the gender representation within an open platform gathering speech and language resources to develop speech processing tools. The aim of this survey is twofold: firstly, we investigate the gender balance within speech corpora in terms of speaker representation but also in terms of speech time available for each gender category. Secondly we propose a reflection about general practices when releasing resources, basing ourselves on some recommendations from previous work.

**Contributions.** The contributions of our work are the following:

- an exploration of 66 different speech corpora in terms of gender, showing that gender balance is achieved in terms of speakers in elicited corpora, but that it is not the case for non-elicited speech, nor for the speech time allocated to each gender category
- an assessment of the global lack of meta-data within free open source corpora, alongside recommendations and guidelines for resources descriptions, based on previous work

## 2. OpenSLR

Open Speech Language Resources[1] (OpenSLR) is a platform created by Daniel Povey. It provides a central hub to gather open speech and language resources, allowing them to be accessed and downloaded freely. OpenSLR currently[2] hosts 83 resources. These resources consist of speech recordings with transcriptions but also of softwares as well as lexicons and textual data for language modeling. As resources are costly to produce, they are most of the time a paying service. Therefore it is hard to study gender representation at scale. We thus focus on the corpora available on OpenSLR due to their free access and to the fact that OpenSLR is explicitly made to help develop speech systems (mostly ASR but also text-to-speech (TTS) systems). In our work, we focus on speech data only.

Out of the 83 resources gathered on the platform, we recorded 53 speech resources. We did not take into account multiple releases of the same corpora but only kept the last version (e.g. TED LIUM (Hernandez et al., 2018)) and we also removed subsets of bigger corpora (e.g. LibriTTS corpus (Zen et al., 2019)). We make the distinction between a resource and a corpus, as each resource can contain several languages (e.g. Vystadial (Korvas et al., 2014)) or several accent/dialect of a same language (e.g. the crowdsourced high-quality UK and Ireland English Dialect speech data set (Google, 2019)). In our terminology, we define a corpus as monolingual and monodialectal, so resources containing different dialects or languages will be considered as containing different corpora.

We ended up with 66 corpora, in 33 different languages with 51 dialect/accent variations. The variety is also great

[1] `http://www.openslr.org`.
[2] Last checked on November 14th, 2019.

in terms of speech types (elicited and read speech, broadcast news, TEDTalks, meetings, phonecalls, audiobooks, etc.), which is not suprising, given the many different actors who contributed to this platform. We consider this sample to be of reasonable size to tackle the question of gender representation in speech corpora.[3] OpenSLR also constitutes a good indicator of general practice as it does not expect a defined format nor does have explicit requirements about data structures, hence attesting of what metadata resources creators consider important to share when releasing resources for free on the Web.

## 3. Methodology

In order to study gender representation within speech resources, let us start by defining what gender is. In this work, we consider gender as a binary category (male and female speakers). Nevertheless, we are aware that gender as an identity also exists outside of these two categories, but we did not find any mention of non-binary speakers within the corpora surveyed in our study.

Following work by Doukhan et al. (2018), we wanted to explore the corpora looking at the number of speakers of each gender category as well as their speech duration, considering both variables as good features to account for gender representation. After the download, we manually extracted information about gender representation in each corpus.

### 3.1. Speaker Information and Lack of Meta-Data

The first difficulty we came across was the general absence of information. As gender in technology is a relatively recent research interest, most of the time gender demographics are not made available by the resources creators. So, on top of the further-mentioned general corpus characteristics (see Section 3.3), we also report in our final table where the gender information was found and whether it was provided in the first place or not.

The *provided* attribute corresponds to whether gender info was given somewhere, and the *found_in* attribute corresponds to where we extracted the gender demographics from. The different modalities are *paper*, if a paper was explicitly cited along the resource, *metadata* if a metadata file was included, *indexed* if the gender was explicitly indexed within data or if data was structured in terms of gender and *manually* if the gender information are the results of a manual research made by ourselves, trying to either find a paper describing the resources, or by relying on regularities that seems like speaker ID and listening to the recordings. We acknowledge that this last method has some methodological shortcomings: we relied on our perceptual stereotypes to distinguish male from female speakers, most of the time for languages we have no knowledge of, but considering the global lack of data, we used it when corpora were small enough in order to increase our sample size.

[3]Our case study does not claim to be exhaustive and future investigations should definitely include data sets provided by resource agencies such as ELRA or LDC.

### 3.2. Speech Time Information and Data Consistency

The second difficulty regards the fact that speech time information are not standardised, making impossible to obtain speech time for individual speakers or gender categories. When speech time information is provided, the statistics given do not all refer to the same measurements. Some authors report speech duration in hours e.g. (Panayotov et al., 2015; Hernandez et al., 2018), some the number of utterances (e.g (Juan et al., 2015)) or sentences (e.g. (Google, 2019)), the definition of these two terms never being clearly defined. We gathered all information available, meaning that our final table contains some empty cells, and we found that there was no consistency between speech duration and number of utterances, excluding the possibility to approximate one by the other. As a result, we decided to rely on the size of the corpora as a (rough) approximation of the amount of speech data available, the text files representing a small proportion of the resources size. This method however has drawbacks as not all corpora used the same file format, nor the same sampling rate. Sampling rate has been provided as well in the final table, but we decided to rely on qualitative categories, a corpus being considered small if its size is under 5GB, medium if it is between 5 and 50GB and large if above.[4]

### 3.3. Corpora Characteristics

The final result consists of a table[5] reporting all the characteristics of the corpora. The chosen features are the following:

- the resource identifier (*id*) as defined on OpenSLR
- the language (*lang*)
- the dialect or accent if specified (*dial*)
- the total number of speakers as well as the number of male and female speakers (*#spk*, *#spk_m*, *#spk_f*)
- the total number of utterances as well as the total number of utterances for male and female speakers (*#utt*, *#utt_m*, *#utt_f*)
- the total duration, or speech time, as well as the duration for male and female speakers (*dur*, *dur_m*, *dur_f*)
- the size of the resource in gigabytes (*sizeGB*) as well as a qualitative label (*size*, taking its value between "big", "medium", "small")
- the sampling rate (*sampling*)
- the speech task targeted for the resource (*task*)

[4]A reviewer rightly pointed out that we could estimate speech duration having its file size, sampling rate and number of bits for quantification, but due to the difficulty to gather all these information and the variety of resources structures, we left it as future work perspective

[5]The final table and the script used for the analysis are available at: `https://github.com/mgarnerin/openslr_gender_survey`.

| Gender info available | | Number of corpora |
|---|---|---|
| No | | 24 (36.4%) |
| Yes | metadata | 9 (13.6%) |
| | indexed | 28 (42.4%) |
| | paper | 5 (7.6%) |
| **Total** | - | **66** |

Table 1: Information availability on gender in OpenSLR corpora.

| Gender info available | Number of corpora |
|---|---|
| Number of speakers | 41 |
| Number of utterances | 32 |
| Speech time | 5 |
| **Total number of corpora** | **42** |

Table 2: Type of information provided in terms of gender alongside the 42 corpora containing gender information.

- is it *elicited* speech or not: we define as non-elicited speech data which would have existed without the creation of the resources (e.g TedTalks, audiobooks, etc.), other speech data are considered as elicited
- the language status (*lang_status*): a language is considered either as high- or low-resourced. The language status is defined from a technological point of view (i.e. are there resources or NLP systems available for this language?). It is fixed at the language granularity (hence the name), regardless of the dialect or accent (if provided).
- the year of the release (*year*)
- the authors of the resource (*producer*)

# 4. Analysis

## 4.1. Gender Information Availability

Before diving into the gender analysis, we report the number of corpora for which gender information was provided. Indeed, 36.4% of the corpora do not give any gender information regarding the speakers. Moreover, almost 20% of the corpora do not provide any speaker information whatsoever. Table 1 sums up the number of corpora for which speaker's gender information was provided and if it was, where it was found. We first looked at the metadata file if available. If no metadata was provided, we searched whether gender was indexed within the data structure. At last, if we still could not find anything, we looked for a paper describing the data set. This search pipeline results in ordered levels for our *found_in* category, meaning papers might also be available for corpora with the "metadata" or "indexed" modalities.

When gender information was given it was most of the time in terms of number of speakers in each gender categories, as only five corpora provide speech time for each category. Table 2 reports what type of information was provided in terms of gender, in the subset of the 42 corpora containing gender information. We observe that gender information is easier to find when it regards the number of speakers, than when it accounts for the quantity of data available for each gender group. Due to this lack of data, we did not study the speech time per gender category as intended, but we relied on utterance count when available. It is worth noticing however, that we did not find any consistency between speech time and number of utterances, so such results must be taken with caution.

Out of the 42 corpora providing gender information, 41 reported speaker counts for each gender category. We manually gathered speaker gender information for 7 more corpora, as explained in the previous section, reaching a final sample size of 47 corpora.[6]

## 4.2. Gender Distribution Among Speakers

### 4.2.1. Elicited vs Non-Elicited Data

Generally, when gender demographics are provided, we observe the following distribution: out of the 6,072 speakers, 3,050 are women and 3,022 are men, so parity is almost achieved. We then look at whether data was elicited or not, non-elicited speech being speech that would have existed without the corpus creation such as TEDTalks, interviews, radio broadcast and so on. We assume that if data was not elicited, gender imbalance might emerge. Indeed, non-elicited data often comes from the media, and it has been shown, that women are under-represented in this type of data (Macharia et al., 2015). This disparity of gender representation in French media (CSA, 2018; Doukhan et al., 2018) precisely led us to the present survey. Our expectations are reinforced by examples such as the resource of Spanish TEDTalks, which states in its description regarding the speakers that *"most of them are men"* (Hernandez-Mena, 2019). We report results in Table 3.

In both cases (respectively elicited and non-elicited speech), gender difference is relatively small (respectively 5.6 percentage points and 5.8 points), far from the 30 percentage points difference observed in (Garnerin et al., 2019). A possible explanation is that either elicited or not, corpora are the result of a controlled process, so gender disparity will be reduced as much as possible by the corpus authors. However, we notice that, apart from Librispeech (Panayotov et al., 2015), all the non-elicited corpora are small corpora. When removing Librispeech from the analysis, we observe a 1/3-2/3 female to male ratio, coherent with our previous findings. This can be explained by the care put by the creators of the Librispeech data set to *"[ensure] a gender balance at the speaker level and in terms of the amount of data available for each gender"* (Panayotov et al., 2015), while general gender disparity is observed in smaller corpora.

What emerges from these results is that when data sets are not elicited or carefully balanced, gender disparity creeps in. This gender imbalance is not observed at the scale of the entire OpenSLR platform, due to the fact that most of the corpora are elicited (89.1%). Hence, the existence of such gender gap is prevented by a careful control during the data set creation process.

[6]The Free ST Chinese Mandarin Corpus (SurfingTech, NA) provided gender information, but we did not manage to use it, hence a total of 47 and not 48.

| Type of speech | #corpora | #F | #M |
|---|---|---|---|
| Elicited | 41 | 1782 | 1596 |
| | | 52.8% | 47.2% |
| Non-elicited | 5 | 1268 | 1426 |
| | | 47.1% | 52.9% |
| Non-elicited | 4 | 67 | 143 |
| (without Librispeech) | | 31.9% | 68.1% |

Table 3: Speaker gender distribution in data depending on the type of speech. *NB: the two last lines refer to the non-elicited corpora, the only difference is that the last line does not take Librispeech into account.*

| Language status | #corpora | #F | #M | Total |
|---|---|---|---|---|
| Low-resource | 23 | 677 | 539 | 1216 |
| | | 55.7% | 44.3% | 100% |
| High-resource | 19 | 1105 | 1057 | 2162 |
| | | 51.1% | 48.9% | 100% |

Table 4: Speaker gender distribution in elicited corpora depending on language status.

#### 4.2.2. High-resource vs Low-resource Languages

In the elicited corpora made available on OpenSLR, some are of low-resource languages other high-resource languages (mostly regional variation of high-resources languages). When looking at gender in these elicited corpora, we do not observe a difference depending on the language status. However, we can notice that high-resource corpora contain twice as many speakers, all low-resource language corpora being small corpora.

#### 4.2.3. "How Can I Help?": Spoken Language Tasks

Speech corpora are built in order to train systems, most of the time ASR or TTS ones. We carry out our gender analysis taking into account the task addressed and obtain the results reported in Table 5. We observe that if gender representation is almost balanced within ASR corpora, women are better represented in TTS-oriented data sets. This can be related to the UN report of recommendation for gender-equal digital education stating that nowadays, most of the vocal assistants are given female voices which raises educational and societal problems (West et al., 2019). This gendered design of vocal assistants is sometimes justified by relying on gender stereotypes such as "female voices are perceived as more helpful, sympathetic or pleasant." TTS systems being often used to create such assistants, we can assume that using female voices has become general practice to ensure the adoption of the system by the users. This claim can however be nuanced by Nass and Brave (2005) who showed that other factors might be worth taking into account to design gendered voices, such as social identification and cultural gender stereotypes.

### 4.3. Speech Time and Gender

Due to a global lack of speech time information, we did not analyse the amount of data available per speaker category. However, utterance counts were often reported, or easily found within the corpora. We gathered utterance counts for a total of 32 corpora. We observe that if gender balance is almost achieved in terms of number of speakers, at the utterance level, men speech is more represented. But this disparity is only the effect of three corpora containing 51,463 and 26,567 (Korvas et al., 2014) and 8376 (Hernandez-Mena, 2019) utterances for male speakers, while the mean number of utterances per corpora is respectively 1942 for male speakers and 1983 for female speakers. Removing these three outliers, we observe that utterances count is balanced between gender categories.

It is worth noticing, that the high amount of utterances of the outliers is surprising considering that these three corpora are small (2.1GB, 2.8GB) and medium (5.2GB). This highlights the problem of the notion of utterance which is never being explicitly defined. Such difference in granularity is thus preventing comparison between corpora.

| Task | #corpora | #F | #M |
|---|---|---|---|
| ASR | 12 | 2523 | 2615 |
| | | 49.1% | 50.9% |
| TTS | 10 | 124 | 70 |
| | | 63.9% | 36.1% |
| NA | 25 | 403 | 337 |
| | | 54.5% | 45.5% |

Table 5: Speaker gender representation in data depending on the task. ASR stands for Automatic Speech Recognition, TTS stands for Text To Speech, and NA accounts for the corpora for which no task was explicitly cited.

| | F | M |
|---|---|---|
| Number of speakers | 591 | 551 |
| | 51.8% | 48.2% |
| Number of utterances | 72,280 | 143,342 |
| | 33.5% | 66.5% |

Table 6: Number of speakers of each gender and number of utterances for each gender category within the subset of corpora providing utterance count by gender. *N.B: two corpora provided utterance count by gender but no speaker count, so the number of speakers is only given as a trend.*

### 4.4. Evolution over Time

When collecting data, we noticed that the more recent the resources, the easier it was to find gender information, attesting of the emergence of gender in technology as a relevant topic. As pointed out by Kate Crawford (2017) in her NeurIPS keynote talk, fairness in AI has recently become a huge part of the research effort in AI and machine learning. As a result, methodology papers have been published, with for example the work of Bender and Friedman (2018), for NLP data and systems, encouraging the community towards rich and explicit data statements. Figure 1 shows the evolution of gender information availability in the last 10 years. We can see that this peek of interest is also present in our data, with more resources provided with gender information after 2017.

## 5. Recommendations

The social impact of big data and the ethical problems raised by NLP systems have already been discussed by pre-

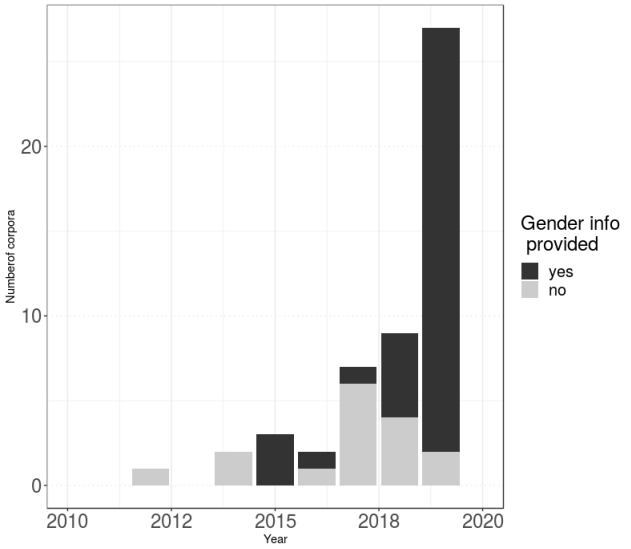


Figure 1: Evolution of gender information availability in OpenSLR resources from 2010 to 2019.

vious work. Wilkinson et al. (2016) developed principles for scientific data management and stewardship, the FAIR Data Principles, based on four foundational data characteristics that are Findability, Accessibility, Interoperability and Reusability (Wilkinson et al., 2016). In our case, findability and accessibility are taken into account by design, resources on OpenSLR being freely accessible. Interoperability and Reusability of data are however not yet achieved. Another attempt to integrate this discussion about data description within the NLP community has been made by Couillault et al. (2014), who proposed an Ethics and Big Data Charter, to help resources creators describe data from a legal and ethical point of view. Hovy and Spruit (2016) highlighted the different social implications of NLP systems, such as *exclusion*, *overgeneralisation* and *exposure* problems. More recently, work by Bender and Friedman (2018) proposed the notion of data statement to ensure data transparency.

The common point of all these studies is that information is key. The FAIR Principles are a baseline to guarantee the reproducibility of scientific findings. We need data to be described exhaustively in order to acknowledge demographic bias that may exist within our corpora. As pointed out by Hovy and Spruit (2016), language is always situated and so are language resources. This demographic bias in itself will always exist, but by not mentioning it in the data description we might create tools and systems that will have negative impacts on society. The authors presented the notion of *exclusion* as a demographic misrepresentation leading to exclusion of certain groups in the use of a technology, due to the fact that this technology fail to take them into account during its developing process. This directly relates to our work on ASR performance on women speech, and we can assume that this can be extended to other speaker characteristics, such as accent or age. To prevent such collateral consequences of NLP systems, Bender and Friedman (2018) advocated the use of data statement, as a professional and research practice. We hope the present study will encourage researchers and resources creators to describe exhaustively their data sets, following the guidelines proposed by these authors.

### 5.1. On the Importance of Meta-Data

The first take-away of our survey is that obtaining an exhaustive description of the speakers within speech resources is not straightforward. This lack of meta-data is a problem in itself as it prevents guaranteeing the generalisability of systems or linguistics findings based on these corpora, as pointed out by Bender and Friedman (2018). As they rightly highlighted in their paper, the problem is also an ethical one as we have no way of controlling the existence of representation disparity in data. And this disparity may lead to bias in our systems.

We observed that most of the speech resources available contain elicited speech and that on average, researchers are careful as to balance the speakers in terms of gender when crafting data. But this cannot be said about corpora containing non-elicited speech. And apart from Librispeech, we observed a general gender imbalance, which can lead to a performance decrease on female speech (Garnerin et al., 2019). Speech time measurements are not consistent throughout our panel of resources and utterance counts are not reliable. We gathered the size of the corpora as well as the sampling rate in order to estimate the amount of speech time available, but variation in terms of precision, bit-rate, encoding and containers prevent us from reaching reliable results. Yet, speech time information enables us to know the quantity of data available for each category and this directly impacts the systems. This information is now given in papers such as the one describing the latest version of TEDLIUM,[7] as this information is paramount for speaker adaptation.

Bender and Friedman (2018) proposed to provide the following information alongside corpus releases: curation rationale, language variety, speaker demographic, annotator demographic, speech situation, text characteristics, recording quality and others. Information we can add to their recommendations relates to the duration of the data sets in hours or minutes, globally and per speaker and/or gender category. This could allow to quickly check the gender balance in terms of quantity of data available for each category, without relying on an unreliable notion of utterance. This descriptive work is of importance for the future corpora, but should also be made for the data sets already released as they are likely to be used again by the community.

### 5.2. Transparency in Evaluation

Word Error Rate (WER) is usually computed as the sum of the errors made on the test data set divided by the total number of words. But if such an evaluation allows for an easy comparison of the systems, it fails to acknowledge for their performance variations. In our survey, 13 of the 66 corpora had a paper describing the resources. When the paper reported ASR results, none of them reported gendered evaluation even if gender information about the data was

[7] However, as gender information was not provided with the release we used, we did not take it into account in our survey.

provided. Reporting results for different categories is the most straightforward way to check for performance bias or overfitting behaviours. Providing data statements is a first step towards, but for an open and fair science, the next step should be to also take into account such information in the evaluation process. A recent work in this direction has been made by Mitchell et al. (2019) who proposed to describe model performance in model cards, thus encouraging a transparent report of model results.

## 6. Conclusion

In our gender survey of the corpora available on the OpenSLR platform, we observe the following trends: parity is globally achieved on the whole, but interactions with other corpus characteristics reveal that gender misrepresentation needs more than just a number of speakers to be identified. In non-elicited data (meaning type of speech that would have existed without the creation of the corpus, such as TEDTalks or radio broadcast), we found that, except in Librispeech where gender balance is controlled, men are more represented than women. It also seems that most of the corpora aimed at developing TTS systems contain mostly female voices, maybe due to the stereotype associating female voice with caring activities. We also observe that gender description of data has been taken into account by the community, with an increased number of corpora provided with gender meta-data in the last two years. Our sample containing only 66 corpora, we acknowledge that our results cannot necessarily be extended to all language resources, however it allows us to open discussion about general corpus description practices, pointing out a lack of meta-data and to actualise the discourse around the social implications of NLP systems. We advocate for a more open science and technology by following guidelines such as the FAIR Data Principle or providing data statements, in order to ensure scientific generalisation and interoperability while preventing social harm.

## 7. Acknowledgements

This work was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

## 8. Copyrights



## 9. Bibliographical References

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th Conference on Neural Information Processing Systems*, NIPS 2016, pages 4349–4357.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, ACM FAT 2018, pages 77–91.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Couillault, A., Fort, K., Adda, G., and (de), H. M. (2014). Evaluating corpora documentation with regards to the ethics and big data charter. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Crawford, K. (2017). The trouble with bias. NeurIPS 2017 Keynote. Available on YouTube, last accessed March 6th 2020.

CSA. (2018). La représentation des femmes à la télévision et à la radio. Rapport d'Exercice 2017.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., and Meignier, S. (2018). An open-source speaker gender detection framework for monitoring gender equality. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2018, pages 5214–5218.

Garnerin, M., Rossato, S., and Besacier, L. (2019). Gender representation in French broadcast corpora and its impact on ASR performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, pages 3–9, New York, NY, USA. ACM.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., and Estève, Y. (2018). TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208. Springer.

Hovy, D. and Spruit, S. L. (2016). The social impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Juan, S. S., Besacier, L., Lecouteux, B., and Dyab, M. (2015). Using resources from a closely-related language to develop ASR for a very under-resourced language: a case study for Iban. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH15)*, Dresden, Germany.

Macharia, S., Ndangam, L., Saboor, M., Franke, E., Parr, S., and Opoku, E. (2015). Who makes the news. Global Media Monitoring Project (GMMP).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman,

L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. ACM.

Nass, C. and Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-computer Relationship*. MIT Press.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2018, pages 3003–3008.

West, M., Kraut, R., and Ei Chew, H. (2019). I'd blush if I could: closing gender divides in digital skills through education.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

## 10. Language Resource References

Google. (2019). *Crowdsourced high-quality UK and Ireland English Dialect speech data set*. Google, distributed via OpenSLR.

Hernandez-Mena, Carlos D. (2019). *TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license*. Universidad Nacional Autonoma de Mexico, distributed via OpenSLR.

Korvas, Matěj and Plátek, Ondřej and Dušek, Ondřej and Žilka, Lukáš and Jurčíček, Filip. (2014). *Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license*. Distributed via OpenSLR.

SurfingTech. (NA). *ST-CMDS-20170001_1, Free ST Chinese Mandarin Corpus*. SurfingTech, distributed via OpenSLR.