The Swedish Winogender Dataset

Saga Hansson[1] **Konstantinos Mavromatakis**[1]
Yvonne Adesam[2] **Gerlof Bouma**[2] **Dana Dannélls**[2]
[1] University of Gothenburg
[2] Språkbanken Text, Department of Swedish, University of Gothenburg
{sagawhansson,konstantinosmavromatakis}@gmail.com
{yvonne.adesam,gerlof.bouma,dana.dannells}@svenska.gu.se

Abstract

We introduce the SweWinogender test set, a diagnostic dataset to measure gender bias in coreference resolution. It is modelled after the English Winogender benchmark, and is released with reference statistics on the distribution of men and women between occupations and the association between gender and occupation in modern corpus material. The paper discusses the design and creation of the dataset, and presents a small investigation of the supplementary statistics.

1 Introduction

Winogender (Rudinger et al., 2018) is a diagnostic dataset designed to detect gender bias in English language coreference resolution systems, inspired by the Winograd Schema Challenge (Levesque et al., 2012). It is also found as part of SuperGlue, a set of benchmark tasks for evaluating Natural Language Understanding models (Wang et al., 2019).[1] Unlike Winograd-style test sets, Winogender is not meant to be a particularly challenging pronoun resolution test set per se, but to lay bare a specific type of gender bias in systems.

Sentences in the Winogender test set contain pronouns whose interpretation is fully determined by causal reasoning. Each sentence contains two noun phrases that could, as far as syntax is concerned, serve as antecedents for the pronoun, one introducing a referent by their occupation, and the other a further participant, which alternatively is referred to with an indefinite pronoun. Furthermore, the sentences are given in several variants, with pronouns with different gender agreement properties (*he*, *she*, [singular] *they*). Examples 1 and 2 are illustrative of the type of sentences in the Winogender test set. Coreferents are in bold.

(1) **The paramedic** performed CPR on *the passenger/someone* even though **she/he/they** knew it was too late.

(2) *The paramedic* performed CPR on **the passenger/someone** even though **she/he/they** was/were already dead.

A crucial aspect of the Winogender sentences is that their interpretation does not depend on the form of the pronoun. So, from common sense reasoning alone – and the assumption that no further entities are relevant – one can conclude that the three alternative pronouns in (1) should all refer to the paramedic, whereas the three alternative pronouns in (2) refer to the other participant (that is, the passenger/someone). In particular, the extent to which the mentioned occupation is perceived as associated with men or women does *not* influence the interpretation of the pronoun.

By inspecting the performance of a pronoun resolution system on the different sentence variants, we can assess the gender-occupation bias inherent in the system. For an unbiased system, there should not be a difference in performance between the pronominal forms. In addition, Rudinger et al. (2018) look at the correlation of model prediction with measures of the binary gender association of the occupations in the test set. For three pronoun resolution systems, the comparisons show clear over-tendencies to resolve the pronoun *she* to female-associated occupations, and under-tendencies to resolve *she* to male-associated occupations.

In this paper, we introduce *SweWinogender*, a Swedish pronoun resolution test set modelled on the Winogender resource. The test set includes Swedish sentences of the type exemplified above. In addition, we provide occupation-gender association statistics relevant to the Swedish language and the Swedish society. Following Rudinger et al. (2018), we supply real-world statistics as well as

[1] https://super.gluebenchmark.com/

corpus-based statistics. The dataset is made available under an open license.[2]

For English, several other studies and benchmarks consider gender-bias in pronoun resolution systems. Zhao et al. (2018, WinoBias) and Lu et al. (2020) use constructed, templatic test items like Winogender, and also investigate ways to mitigate the observed biases. The latter paper presents a slightly different methodology, as bias is not assessed through model predictions, but by looking at model scores. Webster et al. (2018) and Cao and Daumé III (2020) present curated test sets compiled from attested material, with items that lack distinguishing gender-related cues. In addition, the latter moves beyond a binary perspective on gender, and includes a discussion of the harm gender biases in pronoun resolution systems may cause. Beyond English, however, not much directly related work exists. Stanovsky et al. (2019) use the English Winogender and WinoBias sets to probe gender bias in machine translation systems. We are unaware of any previous work that specifically targets gender-bias in coreference resolution systems for languages other than English.

The rest of this paper is structured as follows. We start by presenting the approach taken to create the resource (Section 2). We then describe our real-world occupational gender statistics (Section 3) for Sweden. We continue by exploring gender in the Swedish Culturomics Gigaword corpus (Section 4) and end with conclusion and pointers to future work.

2 Creating SweWinogender

The English Winogender sentences were formulated with the intent that changing the gender of a pronoun should not affect its resolution. The causal/logical structures of the sentences are carefully crafted such that pronoun interpretation is as unambiguous as possible for humans. A Mechanical Turk experiment confirmed that the sentences were indeed unambiguous (Rudinger et al., 2018). To avoid having to reinvent scenarios that have this property, we modelled the SweWinogender collection on the English original.

The English templates were loosely translated into Swedish templates, which then each give rise to twelve similar Swedish sentences: two continuations that force different readings × two ways of referring to the participant (using a descriptive noun or using *någon* 'someone') × three pronouns (*han* 'he', *hon* 'she', *hen* '(singular) they' – or object/possessive forms where appropriate). The Swedish dataset contains 624 sentences in total. Examples 3 and 4 below are taken from the Swedish Winogender dataset. The two sentences each contain three mentions: the occupation *läkaren* 'the physician', the participant *patienten* 'the patient', and the pronoun *hen*. In the first example the pronoun corefers with the participant, in the second with the occupation. Each such sentence occurs six times, three with the specific participant and each of the three pronouns to be resolved, and three with the generic participant *någon* 'someone' and each of the three pronouns to be resolved.

(3) *Läkaren* sa till **patienten** att **hen**
The physician told the patient that they
behövde mer vila.
needed more rest.

(4) **Läkaren** sa till *patienten* att **hen**
The physician told the patient that they
inte kunde skriva ut en högre läkemedelsdos.
could not prescribe a higher dose of medicine.

Sometimes the English occupation was not easily translated to Swedish, because of differences between the American and Swedish contexts. Since our goal was not to create an exact translation, we chose other roles to fit the logic in the discourses. In a number of cases we had to reformulate Swedish sentences due to linguistic differences between Swedish and English. A problematic class of sentences contained possessive pronouns, that potentially corefered with the closest subject. In Swedish, subject coreferring possessives are reflexive possessives, and these are unmarked for gender of the referent, which makes them unsuitable as a diagnostic for gender bias. A second problem with possessives is that regular possessives alternate with reflexive possessives depending on whether there is coreference with the nearest subject or not. This means that even regular possessives may be syntactically unambiguous, making them unsuitable for a diagnostic that relies on syntactic – but not pragmatic – ambiguity. This alternation is illustrated in the following sentence:

(5) X träffade Y för att diskutera sina$_X$ /
X met Y to discuss POSS-REFL
hans$_Y$/hennes$_Y$/hens$_Y$ framsteg
his/her/their progress

[2] https://spraakbanken.gu.se/en/resources/swewinogender

Finally, there is the issue of the inclusion of a gender-neutral neutral pronoun in the test items. English has a relatively well-established gender-neutral pronoun in the form of (singular) *they/them/their*. For Swedish, there has been quite a lot of public debate in the last decade or so about the gender-neutral *hen/hens*. It is not common to introduce new pronouns in a language, but *hen* appears to have weathered out objections. Since 2015 it is even included in the glossary published by the Swedish Academy (SAOL). Unlike *they*, *hen* is unambiguously singular. We have used it for SweWinogender, but considering its rise in use is only recent, it may not be as useful for systems based on older texts.

3 Real-world statistics on gender and occupation

An important part of the diagnostic potential of the Winogender test set is the availability of statistics on the distribution of gender across occupations. It allows a more fine-grained investigation of the correlation of system behaviour with gender biases, by seeing if system predictions follow the distribution of genders for the occupation in a test item. Statistics on gender and occupation also highlight a subset of the Winograd sentences as particularly worthy of close scrutiny, namely those for which the gender bias strongly goes against the intended interpretation of the pronoun. We refer the reader to the original Winogender and WinoBias papers for worked-out examples of the diagnostic methodology (Rudinger et al., 2018; Zhao et al., 2018). The methodological question of how to collect and use statistics that let us move away from a binary gender division is as yet unsolved. The statistics introduced in this section (real-world data) and the next section (corpus-based data) will therefore be binary gender statistics.

To create our first statistical reference, we retrieved real-world statistics about the distribution of men and women across different professions, from Statistics Sweden (SCB).[3] These data were matched against the 43 occupations that occur in our diagnostic sentences. In some cases, we allowed many to one mappings, because the SCB classification was more finegrained than the occupation names in our data. For instance *lärare* 'teacher' in our dataset can be mapped to SCB's *förskollärare* 'preschool teacher', *grundskollärare* 'primary school teacher', *gymnasielärare* 'high school teacher', *högskolelärare* 'college teacher', and *trafiklärare* 'driving instructor'. In these cases, the SCB statistics were summed together before calculating the female-male ratio. This strategy inevitably influences the results since there is no guarantee that the different SCB occupations have similar female-male ratios.

Looking at our compiled statistics, we see that the occupations in SweWinogender are spread out fairly evenly, covering the whole spectrum from female-dominated (more than two thirds registered female practitioners), through neutral (between one third and two thirds female), to male-dominated professions (less than one third female). Table 1 in Appendix A gives the occupations in SweWinogender sorted after the proportion of registered female practitioners.

4 Gender and occupation as seen from a corpus

Another way to look at occupations as female- or male-dominated is not through work-place statistics, but through the lens of a corpus. We can ask: do people read/write about a certain occupation as associated with men or with women? We could speculate that this correlates much better with preconceptions that people hold than the actual employment statistics. More importantly, however, in the context of evaluating NLP systems: the construction of such systems typically involves corpus data. It therefore makes sense to also investigate the relation between system performance and corpus-based gender and occupation associations.

In Rudinger et al. (2018), the noun gender and number dataset from Bergsma and Lin (2006) is used to this end, cited as a frequently used source of this type of information in actual pronoun resolution systems. This list was created using antecedent-pronoun patterns, defined on an automatically parsed corpus, which were used to extract highly likely cases of co-reference in an unsupervised manner. The proportion of *she* (etc) vs *he* (etc) references to a noun is then used to place it on a scale from feminine to masculine. A counterpart to such a list does not exist for Swedish. Moreover, following the same methodology to create such a list is non-trivial in Swedish: First, it depends on having a parsed corpus of Swedish of

[3] https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__AM__AM0208__AM0208E/YREG50/table/tableViewLayout1/

sufficient size and quality. At the time of writing, we have no such corpus readily available. Secondly, several of the patterns used as high-precision coreference patterns by Bergsma and Lin are not useful as sources of information about referential gender in Swedish, because they would involve reflexives or reflexive possessives, which have the same form independent of referential gender.

We therefore follow a less direct approach to extracting occupation-gender associations from corpora, by viewing them as collocative. We assume that a prevalence of definitionally or culturally female-gendered words in the context of mention of a profession, points towards a profession being viewed as female-coded, and correspondingly for male-gendered words. Our approach is reminiscent of the word sense disambiguation method of Yarowsky (1995), and it has been inspired by the application to gendered words in Caren (2013).

As our data source, we use the most recent fifteen years of the Swedish Culturomics Gigaword corpus (Eide et al., 2016), which contains 57M sentences of social media, news text and scientific prose from 2000 to 2015. We use three sets of gendered collocates to classify sentence-level contexts as male- or female-associated: The small set uses only forms of the pronouns *hon/han* 'he/she'. The medium set also includes a list of definitionally gendered nouns, such as *flicka/pojke* 'girl/boy', *mamma/pappa*, *maka/make* 'wife/husband', *syster/bror* 'sister/brother', etc., in total 31 nouns for the male and 25 nouns for the female set.[4] In the large set, we include the items from small and medium sets, and in addition a set of culturally gendered items: all female and male proper names with more than 1000 bearers in Sweden.[5] The large set contains 585 female- and 543 male-gendered words. A sentence is classified according to the majority of collocates it contains – sentences that do not contain any collocates are ignored. For each profession in our dataset, we then calculate the number of sentences classified as female or male that mention this profession. This gives us a way to quantify how strong an occupation is associated with a gender in the corpus.

In many cases, the gender-association assigned to an individual context aligns well with the way an individual referent is presented in the text. This is for instance the case if the decisive collocate in a context happens to be a pronoun that corefers, or is the subject of predication, as in (6) – collocate is in bold, occupation in italics. In these cases, the classification happens to coincide with what Bergsma and Lin's method would yield.

(6) Istället blir **han** *börsmäklare* på Wall Street.
'Instead **he** becomes a *stock broker* on Wall Street.'

But the approach also gives a classification in situations where it makes less sense, for instance in (7), where the context is classified as female because of two collocates from the female set, but where a direct relation to the denotation of the occupation noun is missing.

(7) **Monica** [...] säger att **hon** hoppas kunna göra en studie för att undersöka hur exempelvis *kassapersonal* påverkas.
'**Monica** says **she** hopes to be able to study how for example *cashiers* are affected.'

This type of behaviour is to be expected from a collocational approach. As we will see below, comparison of the corpus results to the SCB data suggests that the approach nevertheless yields usable statistics.

In Figure 1 we plot the real-world SCB data against our corpus-derived measure of gender association, for each of the three collocate sets. Irrespective of the collocate set used, our method generally underestimates the female percentages: most points fall below the diagonal in each plot. This effect is clearly stronger in the small collocate set than in the large set. However, this way of looking at the data ignores the fact that the corpora are biased towards classifying sentences as male-associated in general, not just in the context of a profession. The convex curves in the graphs show what a perfect correspondence would look like if we adjust for this corpus-wide bias.[6] Now we see that in each plot, about half of the points fall above, and half fall below the curve. We conclude that the underestimation of female association of occupations is the result of overall corpus characteristics and not directly related to how people write about

[4]The prototypical pair *man/kvinna* 'man/woman' is not included, because *man* 'man' is homonymic with the frequent pronoun 'one'.

[5]This data is also obtained from SCB, at `https://www.statistikdatabasen.scb.se/sq/99310` and `https://www.statistikdatabasen.scb.se/sq/99311`.

[6]The curves show the line $y = qx/(qx + (1-q)(1-x))$, where q is the overall proportion of sentences classified as female in the corpus.

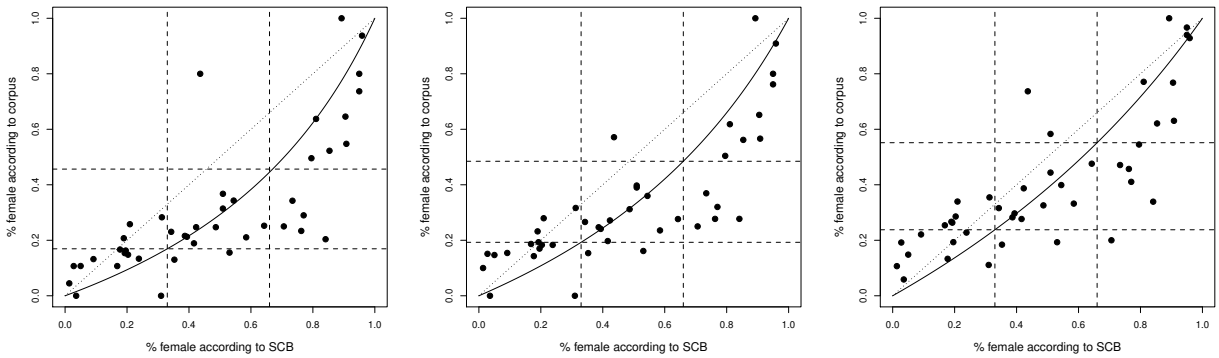

Figure 1: Relation between SCB-based real-world statistics and corpus-based estimations of gender balance in occupations, using small (left), medium (middle), and large collocate sets (right). Curved lines show hypothetical perfect correspondences if we correct for the inherent bias of the method towards male associations. The dashed horizontal lines divide the y-axis in three equal zones male-dominated, neutral, female-dominated, again after correcting for the general bias.

the different professions. However, we can also see a clear pattern in the deviations: points on the left hand-side of the plots generally lie above the curve, whereas those on the right lie below. This means that, compared to the SCB data, the corpus method tends to underestimate male or female domination in the occupations; the estimates shy away from the extremes. This can also be seen by looking at the division of the data into three zones: male-dominated, neutral, and female-dominated. With respect to the SCB data (x-axes) the data points are equally divided between these zones (cf. our remarks in Section 3). However, in the corpus estimates (y-axis), after correction for the overall bias, the neutral professions are over-represented.

On the basis of the overall correlation with the real-world data, we conclude that our method of extracting gender biases for occupations yields meaningful estimates of these biases. We would like to add two further considerations as to why we think our approach makes good sense. Empirically, we note that the pattern that Rudinger et al. (2018) find in the relation between the corpus data and real-world data is (visually) very similar to the patterns discussed above (cf. their figure reproduced here in Appendix B), in spite of what could be expected to be a more precise corpus method. Furthermore, it seems likely that NLP systems that rely on some kind of word embeddings, effectively use collocational information. In those cases, our method may be a much better fit for any biases in such a system than pronoun-resolution-derived estimates.

5 Conclusion

We have presented the freely available SweWinogender test set. It is based on the English Winogender resource and we consider it a starting point which should be expanded upon.

In our data release, the test items themselves will be accompanied by real-world statistics about gender ratios for occupations and by corpus-based gender-occupation associations. These reference data are a core part of making the Winogender idea work as an effective diagnostic.

We have proposed an alternative way of extracting gender-occupation statistics from corpus data, ultimately based on the venerable Distributional Hypothesis. We have argued that the resulting data gives us a perspective on gender and occupation that is relevant to Winogender. Nevertheless, the strengths and weaknesses of this approach need to be further explored. For future work, we will also consider creating further statistical reference sets, for instance in the style of Bergsma and Lin (2006).

We hope that the existence of a SweWinogender will help stimulate the further development, exploration and scrutiny of natural language understanding systems for Swedish.

Acknowledgments

This work has been funded by Nationella Språkbanken – jointly funded by 10 partner institutions and the Swedish Research Council (2018–2024; dnr 2017-00626) – and by the SwedishGlue project (Vinnova, 2020-2021, dnr 2020-02523).

References

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, page 33–40, USA. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Neal Caren. 2013. Using Python to see how the Times writes about men and women. `http://nealcaren.github.io/text-as-data/html/times_gender.html`. Visited 8 February 2021.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish Culturomics Gigaword corpus: A one billion word Swedish reference dataset for NLP. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, July 11, 2016, Krakow, Poland*, pages 8–12. Linköping University Electronic Press.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, volume 12300 of *Lecture Notes in Computer Science*, pages 189–202. Springer, Cham.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

SAOL. 2015. *Svenska akademiens ordlista över svenska språket [The Swedish Academy wordlist of the Swedish language]*, 14th edition. Svenska akademien, Stockholm.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Appendix A Occupational Gender Statistics

Occupation	% Female SCB	% Female Corpus	# Corpus Hits
tandhygienist ‘dental hygienist’	95.88	93.75	42
nutritionist ‘nutritionist’	94.97	80.00	90
dietist ‘dietician’	94.97	73.68	250
terapeut ‘therapist’	90.81	54.76	1413
sköterska ‘nurse’	90.51	64.57	3173
juristassistent ‘paralegal’	89.26	100.00	1
frisör ‘hairdresser’	85.36	52.26	868
apotekare ‘pharmacist’	84.10	20.38	404
receptionist ‘receptionist’	81.03	63.74	166
veterinär ‘veterinarian’	79.53	49.57	2139
lärare ‘teacher’	77.02	28.99	16029
bibliotekarie ‘librarian’	76.25	23.37	1061
psykolog ‘psychologist’	73.41	34.25	5078
kassapersonal ‘cashier’	70.62	25.00	5
utredare ‘investigator’	64.27	25.24	2271
revisor ‘accountant’	58.47	21.06	888
läkare ‘physician’	54.43	34.31	16999
kemist ‘chemist’	53.09	15.56	1088
rättsläkare ‘forensic pathologist’	50.95	36.75	205
specialistläkare ‘medical specialist’	50.95	31.43	84
bartender ‘bartender’	48.66	24.71	264
ambulanssjuksköterska ‘paramedic’	43.62	80.00	38
forskare ‘researcher’	42.31	24.71	8070
rådgivare ‘adviser’	41.61	18.90	3253
försäljare ‘sale person’	39.34	21.26	1139
advokat ‘lawyer’	38.67	21.58	10929
arkitekt ‘architect’	35.29	13.02	10744
polis ‘police’	34.26	23.05	47411
bagare ‘baker’	31.25	28.28	361
byggnadsinspektör ‘building inspector’	30.99	0.00	18
ingenjör ‘engineer’	23.82	13.37	4938
operatör ‘operator’	20.92	25.79	851
köksmästare ‘chef’	20.33	14.81	119
programmerare ‘programmer’	19.58	16.30	176
vaktmästare ‘janitor’	19.25	15.38	463
tekniker ‘technician’	18.95	20.74	1080
börsmäklare ‘stockbroker’	17.74	16.67	60
maskinist ‘machine engineer’	16.84	10.71	126
målare ‘painter’	9.16	13.22	3755
mekaniker ‘mechanic’	5.02	10.73	418
vägarbetare ‘road worker’	3.58	0.00	17
elektriker ‘electrician’	2.78	10.71	224
rörmokare ‘plumber’	1.35	4.48	103

Table 1: Occupational Gender Statistics. The smallest set of collocates (only pronouns) was used for the second and third columns

Appendix B Relation between corpus-based noun gender and Bureau of Labor Statistics data

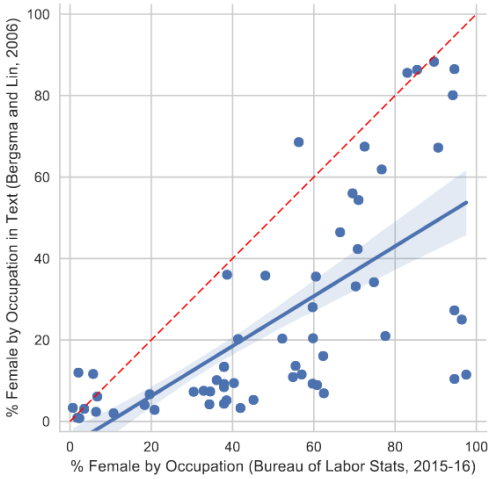

Figure 3: Gender statistics from Bergsma and Lin (2006) correlate with Bureau of Labor Statistics 2015. However, the former has systematically lower female percentages; most points lie well below the 45-degree line (dotted). Regression line and 95% confidence interval in blue. Pearson r = 0.67.