"La ministro è incinta": A Twitter Account of Women's Job Titles in Italian

Alessandra Teresa Cignarella[1,2]**, Mirko Lai**[1]**, Andrea Marra and Manuela Sanguinetti**[3]

1. Università degli Studi di Torino, Italy
2. Universitat Politècnica de València, Spain
3. Università degli Studi di Cagliari, Italy

{alessandrateresa.cignarella|mirko.lai}@unito.it,
andrea.marra.linguistica@gmail.com, manuela.sanguinetti@unica.it

Abstract

We analyze the use of feminine forms indicating professions and roles held by women in Italian. The study is based on Twitter and collects data from 2006 to 2021. This allows us to set up both the quantitative and the qualitative study in a diachronic perspective on a time span of 15 years. We observe the distribution over time of a selection of feminine job titles (i.e., minister, mayor, rector, engineer and lawyer), compared to their masculine counterparts, distinguishing in particular the following cases: use of marked forms and use of semi-marked forms. The analysis shows that the trend of using feminine (i.e. marked) forms is generally growing through time. However, the unbalance between the actual number of women employed in some professions and the use of the correspondent feminine job title is wide.

1 Introduction

The studies on how sexes are represented in language pertain to a transdisciplinary field of research where linguistic aspects intersect with psychological and social issues (Stahlberg et al., 2007). The various types of gender representations in language, along with their asymmetries, is a matter widely studied in linguistics (Hellinger and Bußmann, 2001) as well as in social psychology (Horvath et al., 2016; Hodel et al., 2017). Some of these studies have also affected Italian (Lepschy et al., 2001; Marcato and Thüne, 2002; Mucchi-Faina, 2005; Maturi, 2020), where a renewed debate has spread in the recent past on the use of a more gender-inclusive language.[1][2]

The presence of gender biases and stereotypes has drawn much attention even in the Natural Language Processing community.[3] Research in this field mainly focuses on the study of a model's performance on data associated with a certain gender, or rather on the association between gender and certain concepts as found in language models (Sun et al., 2019).

The present work, instead, aims at giving an exploratory account of the linguistic visibility of women in Italian language, with a focus in particular on job titles. For this purpose, we analyze the use of feminine forms used for job titles and professional roles in Twitter.

Studies on corpus-based discourse analysis have already focused on gender issues with respect to job titles in Italian. They either quantitatively evaluate the mostly used gendered forms in texts when referring to female referents (Formato, 2016; Formato, 2019a; Voghera and Vena, 2016), or rather assess, by means of a survey among native speakers, the degree of acceptability of some feminine job titles (Castenetto and Ondelli, 2020).

From a theoretical point of view, such works revolve (overtly or more indirectly) around the notion of *markedness* in language, that can be intended here as the "contrast between the unmarked (general, usual, non-salient) and the marked (special, emphatic)" (Clyne et al. (2009) cited in Formato (2019b, p.50)).[4] In the present context, the "gen-

---

[1]Elsewhere also defined as gender-fair, gender-neutral or non-sexist language (Sczesny et al., 2016).

[2]https://www.valigiablu.it/linguaggio-inclusivo-dibattito/.

[3]See, for example, the Workshop Series on Gender Bias in NLP: https://genderbiasnlp.talp.cat/.

[4]In its most general sense, this term refers to an opposition between two - otherwise equal - linguistic elements, one of which is characterized by the presence of a mark and the other by its absence (e.g. voicing in voiced *vs* voiceless stops). However, the notion underwent a number of different interpretations and applications. For an in-depth analysis of the differ-

eral, usual, non-salient" case is represented by masculine forms when used to express a generic reference. This means that grammatical masculine nouns are perceived and used as unmarked terms (for both men and women) based on the idea that they represent how the world is, opposing marked feminine terms which are seen as new, ungrammatical and 'sounding bad'.

While sharing with the studies mentioned above the same theoretical premise, the present work addresses the issue of women visibility in Italian language relying on user-generated data retrieved from Twitter: its peculiar nature as language data source, along with the opportunity it offers to extract and filter data based on specific keywords and time spans, makes this platform particularly useful for our purposes.

More precisely, we aimed at studying the distribution over time of a selection of feminine job titles, distinguishing in particular the following cases:

- the use of *marked* forms, i.e. feminine forms referring to female professionals (e.g. *la sindaca Raggi* ('mayor$_{\text{FEM}}$ Raggi'));

- (for a restricted set of examples) the use of *semi-marked* forms (Formato, 2016), i.e. the combination of masculine forms and feminine modifiers when referring to female professionals (e.g. *la neo-ministro è incinta* ('the$_{\text{FEM}}$ new$_{\text{MASC}}$ minister$_{\text{MASC}}$ is pregnant')).

We thus provide some background knowledge on the main linguistic conventions of Italian language in the assignment of grammatical gender, also mentioning some of the well-known studies that have challenged such conventions over the years, towards a more inclusive use of feminine forms, especially for professions. We then describe how data has been collected and filtered, and show the distribution of the selected job titles in both forms and across a 15-year time span.

2 Background

Italian is a grammatical gender language[5] and provides for the mandatory classification of the noun and its respective targets in agreement (modifiers,

such as the adjective or the article) according to two values: masculine and feminine. The gender value is assigned according to phonological and semantic criteria (Thornton, 2005). In assigning gender to nouns denoting human referents, there is a strong tendency to semantically match grammatical gender with the sex of the referent (e.g., *la maestra è arrivata* vs. *il maestro è arrivato* - 'the teacher arrived').

Typically, the masculine is 'overextended' in reference to mixed groups (e.g. *tutti i candidati ammessi* - 'all$_{\text{MASC}}$ admitted$_{\text{MASC}}$ candidates'$_{\text{MASC}}$) or abstract functions (e.g. *le elezioni a sindaco* - 'the mayoral$_{\text{MASC}}$ elections'), as well as in the case of individuals whose gender is not (yet) known (e.g. *assumeremo un nuovo impiegato* - 'we'll hire a new employee'$_{\text{MASC}}$). However, there are cases in which, despite the existence of the feminine form, the masculine is also preferred to refer to a woman, especially when the person holds a prestigious position (Voghera and Vena, 2016). In such a case, the assignment of grammatical gender does not follow this semantic criterion: unmarked expressions referring to a woman (Thornton, 2009, p.126) or semi-marked expressions [6] are well attested. If we consider gender not only as a morphological category, but also as a semantic category, we can understand that, in the symbolic horizon within which the preceding examples move, masculine gender is taken as a neutral (or unmarked) form.

The assumed neutrality of masculine forms has already been questioned from several points of view (Cavagnoli, 2013; Thornton, 2016; Voghera and Vena, 2016). The seminal work by Alma Sabatini (1987), and the one proposed, more than two decades later, by Cecilia Robustelli (2012), have clarified the existence and use of feminine forms already provided for by the Italian linguistic system, and allowed the formulation of recommendations and guidelines for a more inclusive gendered language.

While such reform proposals went largely unheeded (Merkel et al., 2012), more recent studies seem to reveal a slight change in linguistic habits among Italian native speakers (Castenetto and Ondelli, 2020). Hence the choice to verify, by means of an analysis of user-generated content retrieved from Twitter, if a paradigm shift can be found with respect to the use of more gender-inclusive forms.

---

ent perspectives with which this concept is treated, we refer to Moravcsik and Wirth (1986) and Haspelmath (2006).

[5]We refer to Stahlberg et al (2007) for the complete definition of grammatical gender, natural gender and genderless languages.

---

[6]https://www.repubblica.it/online/speciale/presti/presti/presti.html.

3 Data Collection

Starting from the proposals presented in the recommendations of Sabatini (1987) and Robustelli (2012), we selected a shortlist of 11 job titles with both masculine and feminine endings. The selection is based on morphological criteria, more precisely on the different categories of gender suffix pairs that can be added to the root of a noun. We thus included the following terms:

Job titles ending in *-o*$_{\text{MASC}}$ / *-a*$_{\text{FEM}}$:

- *ministro/ministra* ('minister'),
- *sindaco/sindaca* ('mayor').

Job titles ending in *-tore*$_{\text{MASC}}$ / *-trice*$_{\text{FEM}}$:

- *rettore/rettrice* ('rector').

Job titles ending in *-ere*$_{\text{MASC}}$ / *-era*$_{\text{FEM}}$:

- *ingegnere/ingegnera* ('engineer').

Job titles ending in *-o*$_{\text{MASC}}$ / *-a* or *-essa*$_{\text{FEM}}$[7]

- *avvocato/avvocata/avvocatessa* ('lawyer').

Twitter recently introduced APIs (v2) that allow to access the full history of public conversations since the first tweet was created on March 21st, 2006. Accordingly, we take advantage of Twitter's full-archive search endpoint[8] for retrieving each tweet written in Italian and containing at least one of the words listed above, from March 21st, 2006 to March 21st, 2021 aiming at depicting the scenario of their use diachronically through a span of 15 years.

3.1 Data Cleaning

A preliminary data analysis shows several noisy tweets in the dataset. Some keywords are indeed particularly affected by homonymy and polysemy. For example, the keywords *sindaco* and *sindaca* are also inflections of the verb *sindacare* ('to judge, criticize, inspect'). A particular example is also the homonymy of the word *rettore* ('rector'$_{\text{MASC}}$) with

---

[7]The suffix *-essa* is used as a derivative for female referents starting from the male noun (Formato, 2019a), and its possible demeaning connotation has been matter of debate (Merkel et al., 2012; Mucchi-Faina, 2005). In Sabatini's Recommendations, its use is discouraged in favor of the suffix *-a* (or *-e* for some epicene nouns).

[8]https://developer.twitter.com/en/docs/
twitter-api/tweets/search/api-reference/
get-tweets-search-all.

the surname of a famous Italian singer and songwriter (Donatella Rettore), and of the word *avvocata* ('lawyer'$_{\text{FEM}}$) with a homonymous district of the city of Naples, Italy.

Other keywords are also affected by the use of figurative language. Particularly relevant is the use of the keywords *ministro* and *avvocata* in a religious context. Indeed, in Christianity, priests are also called *ministri di Dio* ('ministers of the Lord'), while *avvocata nostra* ('most gracious advocate') is part of the prayer 'Hail Holy Queen'. These few examples help to catch a glimpse of the difficult task of cleaning and removing noisy tweets from this dataset automatically. Therefore, we performed a semi-automatic data cleaning by using filters tailored for each word.

The final dataset consists of around 9.7 million tweets overall; Table 1 reports the number of tweets per keyword, as resulted after the cleaning process.[9] Drawing inspiration from studies in demography, where male to female ratio is a common parameter, we report the proportion of masculine (M) and feminine (F) forms in terms of M/F RATIO, where the higher the value the greater the unbalance between the two forms at the expense of the latter.

MASC	# tweets	FEM	# tweets	M/F RATIO
ministro:	3,575,613	ministra:	290,321	12.32
sindaco:	4,005,156	sindaca:	256,334	15.62
rettore:	138,328	rettrice:	4,490	30.81
ingegnere:	291,334	ingegnera:	4,759	61.22
avvocato:	1,133,456	avvocata:	22,771	49.78
		avvocatessa:	25,190	45.00
sum:	**9,143,887**	**sum:**	**405,841**	
unique:	9,090,414	unique:	378,274	

Table 1: N° of tweets retrieved for each query word.

On the numerical front we can see that the number of tweets containing the masculine form is greatly dominant. This is especially evident in the case of the keyword pair *ingegnere/ingegnera* (M/F RATIO of 61.22) despite the fact that the ratio of male and female engineers in Italy is 5.38.[10]

On the other hand, the feminine words that seem to be used in the most balanced way with respect to their masculine counterpart are *ministra* (M/F RA-

---

[9]It is worth pointing out, however, that several tweets contain two or more keywords; they are counted in the table as many times as the number of keywords they contain. For this reason the values of 'sum' are higher than 'unique'.

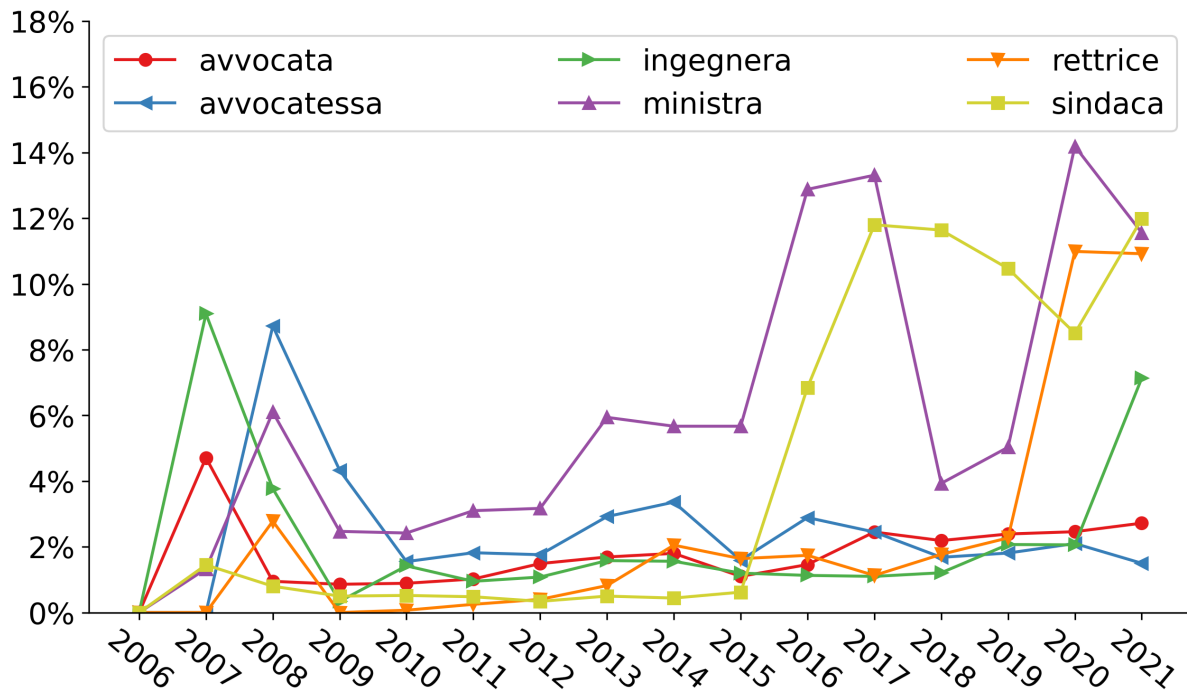[10]See page 13: https://www.cni.it/images/Ne
ws/2020/Iscritti_anno_2020_LQ.pdf.

Figure 1: Frequency trend of women's job titles from 2006 to 2021.

of 12.32) and *sindaca* (M/F RATIO of 15.62).

4 Data Analysis and Discussion

The first step of our data analysis consists of observing the trends of the frequency of use of the six women's job titles explored in this work.

In Figure 1 we represent the frequency of feminine job titles with respect to the total of terms used to describe the profession (FEM / FEM + MASC). We observe that from 2006 to 2021 there is a tendency to a more frequent use of female forms in general. However, relevant spikes are present on the left side of the chart. We believe they are caused by the scarcity of data before 2010, which is also imputable to the low popularity of the microblogging platform in Italy before that year. Furthermore, among the 6 sixfeminine keywords used as case study in the present work, 2 of them do not even have any occurrence in the totality of the year 2006. Their use starts with a few occurrences only from the year after (*avvocatessa* and *rettrice*).

The purple ▲ line (see Figure 1), illustrating the trend of the word *ministra* ('minister'FEM) shows how the word has been increasingly used around 2016-2017, and then again around 2019-2020. The use of this term seem to increase during the election period and to decrease immediately afterwards. This outcome is indeed in line with the periods in

which governmental changes occurred in Italy. In particular, in both those time spans there have been female ministers who have been highly politically exposed.[11]

Another fact worth mentioning is the trend of the mustard-yellow line ■ in Figure 1 depicting the use of the word *sindaca* ('mayor'FEM). The word seems to have started to be used more frequently in conjunction with the election of two female mayors in two large Italian cities.[12] Also the relationship between red ● and blue ◄ lines in the same figure presents a notable trend. Those lines respectively show the use of *avvocata* and *avvocatessa* (both: 'lawyer'FEM). It is peculiar how the two lines show the same tendency throughout the years with the preference for the term *avvocatessa* on top of *avvocata*, until the year 2017. From that moment on, there is an inversion of trend and the occurrence of first term starts decreasing (blue ◄ line), favoring the use of the second one (red ● line). The oscillation between *avvocatessa* and *avvocata* therefore remains, but it seems that the latter has been increasingly gaining some ground.

The word *rettrice* ('rector'FEM), marked by the orange ▼ line in Figure 1, has an averagely grow-

---

[11]Marianna Madia and Maria Elena Boschi in 2016-2017 and Luciana Lamorgese and Lucia Azzolina in 2019-2020.

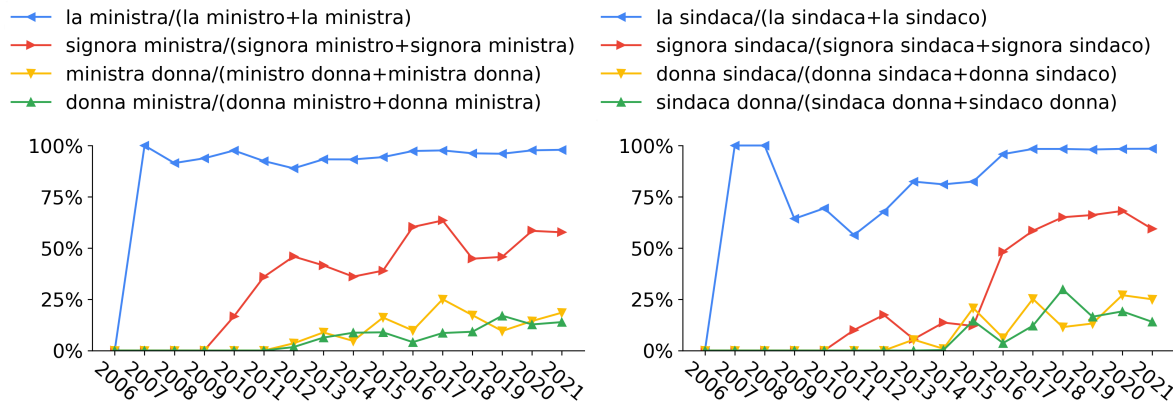[12]Virginia Raggi in Rome and Chiara Appendino in Turin.

Figure 2: Ratio between the marked forms and the semi-marked forms for the terms *ministra/o* ('minister', on the left) and *avvocata/o* ('mayor', on the right).

ing distribution through time (around 2%), with a spike of increase in 2020, when – for the first time – a woman has been elected as rector in the biggest university of Europe: La Sapienza in Rome.

Finally, *ingegnera* ('engineer'ᶠᴱᴹ) is the only one among the six terms taken into consideration with a low, though constant, trend throughout the temporal span of 15 years (around 1.6%), with only one recent spike around 2020-2021 (green ▶ line).

4.1 Analysis of N-grams

In a second step of our analysis, we aimed at investigating on the use of semi-marked forms (see Section 1). We focused on the two terms that presented the most balanced distributions with respect to their masculine counterpart (see Table 1), i.e. *ministra* and *sindaca*, and studied when and how the masculine form has been used to refer to a female referent in the real world. To do so, we extracted n-grams where one of the two tokens is one of the masculine words selected for the study and the second token is a feminine determiner or nominal modifier.

Hence, we selected the following 2-grams of interest:

- *la ministro/sindaco*
  ('theᶠᴱᴹ minister/mayor'ᴹᴬˢᶜ)

- *ministro/sindaco donna* and *donna ministro/sindaco*
  ('female minister/mayor'ᴹᴬˢᶜ)

- *signora ministro/sindaco*
  ('Madame minister/mayor'ᴹᴬˢᶜ)

In Figure 2 we show two charts (one for the word 'minister', and one for the word 'mayor') illustrat-

ing the ratio between the selected marked forms and the sum of such forms with semi-marked forms.

In both cases it is once again evident that the data collected before 2010 is very scarce, and that relevant statistics are, therefore, to be considered valid only after that year.

For both charts it is shown how the tendency of using marked forms (*la ministra* and *la sindaca*) is growing throughout the years; on the other hand, expressions where the female attribute is explicitly mentioned – such as *signora ministra* ('Madame minister'ᶠᴱᴹ) and *signora sindaca* ('Madame mayor'ᶠᴱᴹ) – are still very frequent (red ▶ lines in both charts).

Despite the outcomes derived from the analysis of n-grams, we acknowledge that the procedure described in this subsection is fairly limited. Beside the fact we studied the distribution of only two words out of the six selected for the present study, the availability of the same data enriched with part-of-speech tagging and parsing information would be highly beneficial for the automatic identification of marked and semi-marked forms.

5 Conclusion and Future Work

In this paper, we reported the results of a corpus-based account of the linguistic visibility of women in Italian language, with a focus in particular on job titles, and using Twitter as data source. From a preliminary analysis of a selection of profession nouns, we found that some marked forms are increasingly being preferred in spite of semi-marked expressions. Besides extending and systematizing this analysis to other case studies, we also aim to observe the usage of such forms by Italian native

speakers by tackling the issue as a stance detection task, so to assess how the users value a given marked form and, more in general, the adoption of more gender-inclusive linguistic habits. Furthermore, the messages leveraged on this topic might overlap with the task of misogyny detection and hate speech detection as well, broadening the horizons of three different NLP detection tasks. This design choice can also be motivated with regard to *contextual stance detection* (Cignarella et al., 2020; AlDayel and Magdy, 2021), to investigate how supporters/opponents of inclusive language strategies are segregated in different online social network communities.

Finally, due to its preliminary and exploratory nature, this work only reports the distribution of feminine and masculine forms, which are the two values for gender assignment taken in consideration for the analysis. We are well aware, however, that a comprehensive study of gender-inclusive language must necessarily cover all those linguistic forms that refer to the multiple and diverse identities in the gender spectrum.

With respect to this point, innovative forms have been proposed in the last years, in order to overcome the binary opposition, even in a grammatical gender language as Italian, such as the schwa (ə), the asterisk (∗), the 'at' sign (@), and other graphic solutions. This is another aspect that is worth exploring in a stance detection perspective, so to assess users' stance regarding the use of such linguistic innovations and their spread in everyday language.

Acknowledgments

References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Giorgia Castenetto and Stefano Ondelli. 2020. The acceptability of feminine job titles in Italian newspaper articles. In Giuliana Giusti and Gabriele Iannàccaro, editors, *Language, Gender and Hate Speech. A Multidisciplinary Approach*, pages 75–90. Edizioni Ca' Foscari.

Stefania Cavagnoli. 2013. *Linguaggio giuridico e lingua di genere: una simbiosi possibile*. Edizioni dell'Orso.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, Rosso Paolo, et al. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. CEUR-WS.org.

Michael Clyne, Catrin Norrby, and Jane Warren. 2009. *Language and human relations: Styles of address in contemporary language*. Cambridge University Press.

Federica Formato. 2016. Linguistic markers of sexism in the Italian media: a case study of ministra and ministro. *Corpora*, 11(3):371–399.

Federica Formato. 2019a. *Gender, Discourse and Ideology in Italian*. Palgrave Studies in Language, Gender and Sexuality. Palgrave Macmillan.

Federica Formato. 2019b. An overview of grammatical gender in italian. In Federica Formato, editor, *Gender, Discourse and Ideology in Italian. Palgrave Studies in Language, Gender and Sexuality*, pages 39–80. Palgrave Macmillan.

Martin Haspelmath. 2006. Against Markedness (And What to Replace It With). *Journal of Linguistics*, 42(1):25–70.

Marlis Hellinger and Hadumod Bußmann. 2001. *Gender across languages. The linguistic representation of women and men*, volume 1 of *Impact: Studies in language and society*. John Benjamins Publishing Company.

Lea Hodel, Magdalena Formanowicz, Sabine Sczesny, Jana Valdrová, and Lisa von Stockhausen. 2017. Gender-fair language in job advertisements: A cross-linguistic and cross-cultural analysis. *Journal of Cross-Cultural Psychology*, 48(3):384–401.

Lisa K. Horvath, Elisa F. Merkel, Anne Maass, and Sabine Sczesny. 2016. Does gender-fair language pay off? The social perception of professions from a cross-linguistic perspective. *Frontiers in Psychology*, 6.

Anna Laura Lepschy, Giulio Lepschy, and Helena Sanson. 2001. Lingua italiana e femminile. *Quaderns d'Italià*, 6:9–18.

Gianna Marcato and Eva Maria Thüne. 2002. Gender and female visibility in Italian. In Marlis Hellinger and Hadumod Bußmann, editors, *Gender across languages. The linguistic representation of women and men. Volume II*, pages 187–217. John Benjamins Publishing Company.

Pietro Maturi. 2020. Qual è il tuo pronome? Riflessioni su questioni di genere nelle lingue europee. *Fuori Luogo Rivista di Sociologia del Territorio, Turismo, Tecnologia*, 8(2/2000):67–74.

Elisa Merkel, Anne Maass, and Laura Frommelt. 2012. Shielding women against status loss: The masculine form and its alternatives in the Italian language. *Journal of Language and Social Psychology*, 31(3):311–320.

Edith Moravcsik and Jessica Wirth. 1986. Markedness - an overview. In Fred R. Eckman, Edith A. Moravcsik, and Jessica R. Wirth, editors, *Markedness*, pages 1–11. Springer.

Angelica Mucchi-Faina. 2005. Visible or influential? Language reforms and gender (in)equality. *Social Science Information*, 44(1):189–215.

Cecilia Robustelli. 2012. Linee guida per l'uso del genere nel linguaggio amministrativo. Progetto Accademia della Crusca e Comune di Firenze. Available at: `https://www.rai.it/dl/docs/1354527785476guida_robustelli.pdf` (2021/09/04).

Alma Sabatini. 1987. Raccomandazioni per un uso non sessista della lingua italiana. In Alma Sabatini, editor, *Il sessismo nella lingua italiana*, pages 97–122. Presidenza del Consiglio dei Ministri e Commissione Nazionale per la Parità e le Pari Opportunità tra uomo e donna.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7:25.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. In Klaus Fiedler, editor, *Social communication. A volume in the series Frontiers of Social Psychology*, pages 163–187. Psychology Press.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Anna Maria Thornton. 2005. *Morfologia*. Carocci.

Anna Maria Thornton. 2009. Designare le donne. In Giuliana Giusti and Susanna Regazzoni, editors, *Mi fai male...*, pages 115–133. Cafoscarina.

Anna Maria Thornton. 2016. Designare le donne: preferenze, raccomandazioni e grammatica. In Fabio Corbisiero, Pietro Maturi, and Elisabetta Ruspini, editors, *Genere e linguaggio. I segni dell'uguaglianza e della diversità*, pages 15–33. Franco Angeli.

Miriam Voghera and Debora Vena. 2016. Forma maschile, genere femminile: si presentano le donne. In Fabio Corbisiero, Pietro Maturi, and Elisabetta Ruspini, editors, *Genere e linguaggio. I segni dell'uguaglianza e della diversità*, pages 34–52. Franco Angeli.