Emerging Trends in Gender-Specific Occupational Titles in Italian Newspapers

Pierluigi Cassotti[1], **Andrea Iovine**[1], **Pierpaolo Basile**[1],
Marco De Gemmis[1] **and Giovanni Semeraro**[1]
1. Department of Computer Science, University of Bari Aldo Moro, Italy
{firstname}.{surname}@uniba.it

Abstract

The grammatical gender system can influence the way the semantic gender is perceived. Italian is a grammatical gender language, in which nouns are classified for gender. In this work, we investigate the usage of gender-specific forms of occupational titles in a diachronic corpus of 3 billion tokens extracted from two Italian newspapers. The hypothesis is that the usage of gender-specific forms might be influenced by socio-cultural aspects, such as changes in the employment policy. We automatically collect a set of occupational titles and perform a diachronic analysis exploiting the frequency of gender-specific forms. Results show a correlation between changes in the usage of gender-specific forms and socio-cultural events.

1 Introduction

Throughout history, the prerogative use of specific gender forms over particular professions can fade away by introducing changes in the language lexicon (e.g., neologisms) or in the language usage (e.g., word frequencies). The way the lexicon is affected by those changes depends on the grammatical gender system, i.e. the set of rules that define the agreement between noun classes forms and the other parts-of-speech. Grammatical gender systems can vary dramatically from one language to another. Gygax et al. (2019) propose a classification of languages based on their grammatical gender system. In this work, we focus on the Italian language, a grammatical gender language in which all nouns must be classified for gender. The Italian gender system admits three categories for nouns: gender-specific ending nouns, mobile gender nouns, and nouns where the gender is specified through determiners and adjectives (Marcato and Thüne, 2002). In gender-specific ending nouns, the gender forms are expressed through completely different lexical roots (e.g., *genero/nuora*). In mobile gender nouns, the specific gender forms share the same lexical root, and the semantic gender is instead represented by different suffixes (e.g., *scrittore/scrittrice*). In other cases, the semantic gender of a noun is inferred only by the determiner and/or adjective (e.g., *il* giudice, *la* giudice). The peculiar characteristic found in the Italian language has strong repercussions in the way people refer to occupational titles, because a specific gender form might be preferred over the other due to historical reasons, regardless of the gender of the actual person being talked about (Sabatini, 1985). This has become a hot-button issue in the last years, especially as a result of the United Nations Resolution “Transforming our world: the 2030 Agenda for Sustainable Development” with its global indicator framework for Sustainable Development Goals (SDGs), and specifically of SDG 5 *Achieve gender equality and empower all women and girls* (subgoal 5.1 *End all forms of discrimination against all women and girls everywhere*) (Lee et al., 2016).

The objective of this paper is to monitor how the use of gender-specific occupational titles has changed in the Italian language over the years through the use of diachronic analysis tools. We would like to emphasize that the goal is not to map the composition of men and women for each profession over time, as this cannot be reliably inferred from text. Instead, we are interested in gauging the cultural relevance of the gender-specific titles over time, as reflected in the news domain. Accordingly, the contributions in this paper can be summarized as follows:

(i) We analyze emerging trends in the use of gender-specific occupational titles in the Italian language in a corpus of newspaper articles.[1]

(ii) We perform a deep-dive analysis of the figures that have guided a significant shift for two professions in particular.

Large diachronic corpora have already been used to study social and cultural phenomenons that affected language in a significant way. The Google Ngrams Dataset (Goldberg and Orwant, 2013) is a dataset of n-grams extracted by 3.5 million books published between 1520 and 2008. Aiden and Michel (2011) exploit the huge quantity of information contained in the Google Ngrams Dataset to analyze the evolution of the language lexicon over time. In particular, the work offers interesting culturomics results, such as highlighting the spread of the term *influenza* during historical pandemic periods. Kutuzov et al. (2017) exploit diachronic word embeddings to track wars and conflicts that took place from 1994 to 2010 all around the world. Diachronic word embeddings are trained on the English Gigaword news corpus (Parker et al., 2011) and used to predict conflict states: *peace*, *war* and *stable*. Laine and Watson (2014) analyze the linguistic sexism occurring in *The Times* newspaper over five decades (1965-2005), relying on the classification of linguistic sexism proposed in (King, 1991). The authors hypothesize that occupational titles and agents would be more resistant to change than other forms of sexism over the decades. They confirm their hypothesis by exploring the frequencies of male and female affixes, showing that they keep stable. Burr (1995) performs an empirical analysis on manually-annotated occurrences of grammatical agents in a small synchronic corpus of Italian newspapers. The outcomes of this work lead the authors to conclude that women are underrepresented in Italian newspapers, especially in more high-position roles.

2 Corpus

Occupational titles occurrences are extracted from a diachronic corpus that comprises two sub-corpora. The former corpus is the “L’Unità” corpus (Basile et al., 2020) that covers the time period 1945-2014. The latter is crawled by the publicly available digital archive of the Italian newspaper “La Stampa” covering the period 1945-2005 and processed using the same methodology mentioned in (Basile et al., 2020). In order to align the two sub-corpora time ranges, we consider a sub-portion of the “L’Unità” corpus that spans the period 1948-2005. The overall corpus contains 3,529,820,155 tokens and spans the period 1948-2005. Corpus statistics are reported in Table 1. The corpus presents two main critical issues. First, despite having performed pre-processing and filtering, the documents from the earlier periods suffer from several OCR errors and noise. Second, data is not equally distributed, the number of tokens drops dramatically in the first years. Text is processed using the UDPipe model (Straka et al., 2016) included in spaCy[2]. The UDPipe model is trained on the Italian Stanford Dependency Treebank (Bosco et al., 2014). Each sentence is tokenized, lemmatized and annotated with PoS-tags, named entity tags and dependency relations. Moreover, the UDPipe model provides information about inflectional features of nouns exploited in the occupational titles extraction pipeline.

Corpus	Tokens	Period
L’Unità	425,833,098	1948-2014
La Stampa	3,145,959,127	1948-2005
Overall	3,529,820,155	1948-2005

Table 1: Corpus statistics.

3 Extracting Occupational Titles

The first step of our investigation consists of extracting a list of occupational titles from a common Knowledge Base. Specifically, we have exploited Wikidata (Vrandečić and Krötzsch, 2014), since it has collected a wide range of entities related to professional activities. We first extracted a list of all entities that are an instance of *profession* (wd:Q28640), or of an entity that is a subclass of it, for which a label in the Italian language is present. This label commonly contains the male gender form of the occupational title. Then, we filtered the list of professions by only including those that possess the *female form of label* (wdt:P2521) property for the Italian language. This property denotes the female variant of the occupational title, where applicable. The next step consists of filtering out occupational titles for which the gender is not easily distinguishable from text, such as those in which both gender variants

[1] All data collected in this experiment is available here: https://github.com/pierluigic/igsot

[2] https://spacy.io/

share the same lexical root (e.g. the aforementioned *il giudice/la giudice*), or those that do not feature gender variants at all (e.g. *la guardia*, i.e. the guard). We also removed all occupational titles that consist of two or more tokens. Then, we reduced the list by filtering out polysemous words. A common example of polysemy in the Italian language occurs when an occupational title shares the same lexical form as the discipline to which it belongs, such as *matematica* (female form of *mathematician*), or *fisica* (female form of *physicist*). For each occupational title, we used WordNet to find all synsets in which it appears and then removed it if the synset is a hyponym of the *discipline.n.01* synset. Moreover, we manually analyzed the list of remaining occupational titles and removed other instances of polysemy, which would otherwise hinder the quality of the results. For instance, we filtered the word *editrice* (female form of *editor*) as it can also appear in the phrase *casa editrice* (i.e. publishing house), and the word *tecnica* (female form of *technician*), which can also refer to the word *technique* depending on context. We also decided to remove words that have additional figurative meanings, such as *cacciatrice* (female form of *hunter*) and guerriera (female form of *warrior*). This process was undertaken by two independent annotators and then checked for agreement. The final result of this process is T, a set of tokens that unequivocally refer to occupational titles, and that feature distinct male and female gender variants which can reliably be extracted from text.

4 Experimental Setup

Once we have acquired the set of occupational titles T, the next step of the analysis consisted of measuring the frequency with which each term $w \in T$ occurs for each year in the corpus described in Section 2. We also make use of the lexical information contained in said corpus in order to eliminate any remaining ambiguity in the words. In fact, for each occupational title, we counted a hit in the corpus if it appears with the NOUN tag. This allows us to avoid counting occupational titles that can be confused with verbs or adjectives, such as *impiegato/impiegata*, which can refer to the noun *employee* in Italian, but also to the past participle conjugation of the verb *to employ*.

Moreover, we only counted a hit if the word has been registered with the singular form. This is done for two reasons: first, occurrences of the plural form are outside the scope of this investigation, because in Italian the male plural form is traditionally used as the default, while the female variant of the plural is only used in exceptional cases, such as when referring to a group that is composed entirely of women. Second, this strategy filters out cases where the plural form shares the same lexical root as one of the gender variants. An example of this is the word *infermiere* (i.e. *nurse*), which can refer to both the singular masculine form (as in *l'infermiere*), or the plural feminine form (as in *le infermiere*).

Since the objective of this study is to observe the trends in the use of masculine and feminine forms for occupational titles, we are interested in analyzing how their frequency changes from one year to the other. However, measuring the absolute frequency in each year for both forms would be misleading, as it heavily depends on the amount of data that is available for each year in the corpus. Instead, we compute the smoothed relative frequency p_w^t for each word w and each year t using the following formula:

$$p_w^t = \frac{f_w^t + 1}{C^t + \mid V^t \mid} \quad (1)$$

where f_w^t is the frequency of word w in the year t, C^t is the count of tokens occurring in the corpus the year t and $|V^t|$ is the vocabulary length computed on the year t. We compute p_w^t for both gender forms of each occupational title. Then we compute $odds(w)^t$ which represents the log ratio of the smoothed relative frequency of the female and male forms respectively:

$$odds(w)^t = log\frac{p_{w_f}^t}{p_{w_m}^t} \quad (2)$$

Operationally, $odds(w)^t$ specifies the probability that the feminine variant will appear in a text relative to the masculine form in the specified year t. We then obtain the time-series by concatenating the $odds(w)$ values computed for each year: $(odds(w)^{1948}, odds(w)^{1949}, .., odds(w)^{2004})$. Assuming a linear course of the time-series, three different scenarios can occur: *(i)* the occurrences of the female form are growing; *(ii)* the occurrences of the male form are growing; *(iii)* the ratio of the male and female form of an occupational title are stable over time. We com-

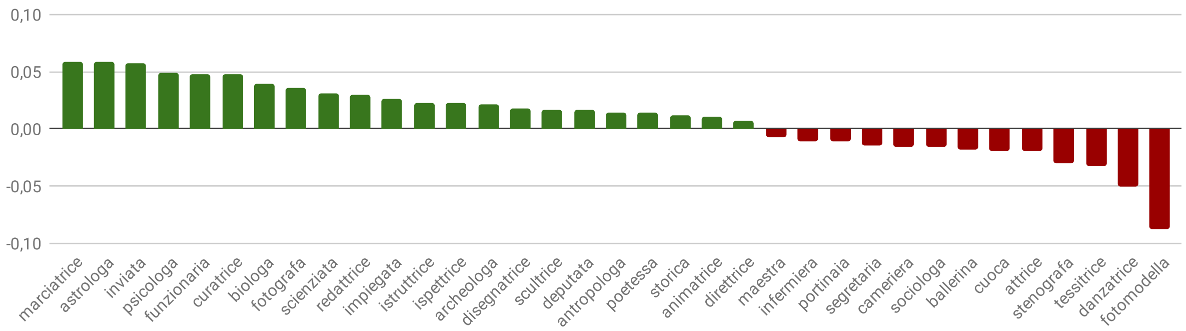


Figure 1: Final set of occupational titles (the female form is reported) and the slope of $odds(w)^t$.

puted the regression line of the time-series, using the linear least-squares regression method provided by the SciPy library[3]. We use the slope of the regression line to determine whether the values of $odds(w)^t$ are changing over time. If the slope is positive/negative, $odds(w)^t$ is increasing/decreasing over time, which means that the frequency of w_f is increasing/decreasing faster than that of w_m, or that the frequency of w_m is decreasing/increasing faster than that of w_f. For each regression line, we also compute the statistical significance of the slope parameter relying on the Wald Test (Fahrmeir et al., 2007). Specifically, the null hypothesis states that the slope parameter of the regression line is zero. In this stage, occupational titles for which we get a $p - value > 0.1$ are filtered out.

5 Results

Figure 1 describes the value of the slope for each occupational title. Depending on the sign of the slope, we can identify two distinct groups of occupational titles. Green bars indicate that the slope of $odds(w)^t$ is positive, i.e. the frequency of the feminine form is increasing relative to that of the masculine form. On the other hand, red bars indicate that the slope is negative, thus the frequency of the feminine form is decreasing relative to that of the masculine form. Out of 35 occupational titles, 22 have a positive slope, while 11 result in a negative slope. In particular, the most positive slope is the one associated to *marciat-ore/-rice* (i.e. *racewalker*), while the most negative slope is *fotomodell-o/-a* (i.e. *fashion model*).

For many of these titles, the resulting slope can be mapped to specific social changes. An interesting example in this regard is *infermiere* (i.e. *nurse*), to which a negative slope is recorded: indeed, in Italy the position of nurse has been opened to men starting from 1971[4]. The odds(w) time series of infermiera/infermiere is reported in Figure 2.

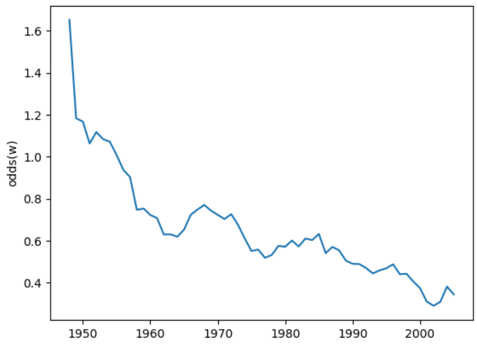


Figure 2: 10-year moving average of odds(w) for infermiera/infermiere.

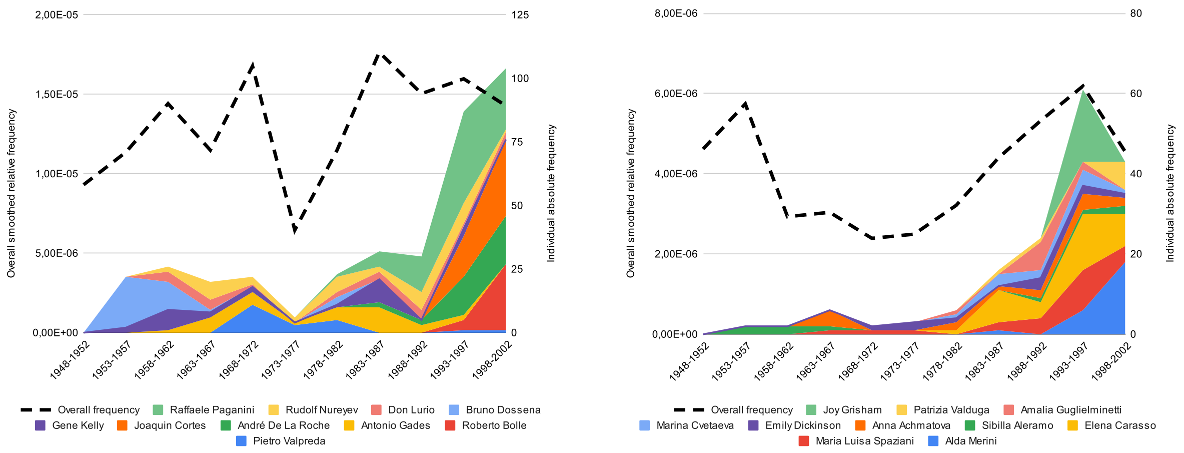
Moreover, results show that managerial roles such as *funzionaria* (i.e. *civil servant*), *ispettrice* (i.e. *inspector*), *direttrice* (i.e. *director*) are associated to a positive slope, which is indicative of a stronger perception of women in such roles.

A similar push can be observed also in the scientific domain, with a positive trend for the words *biologa* (i.e. *biologist*), *scienziata* (i.e. *scientist*), as well as the artistic one. On the other hand, we observe an increase in the usage of the masculine form for *segretario* (i.e. *secretary*), *ballerino* (i.e. *dancer*), and *stenografo* (i.e. *stenographer*).

In the second part of the experiment, we attempt to identify the people that have driven the change in the usage of the feminine and masculine forms of an occupational title. To do this, we retrieve

[3] https://www.scipy.org/

[4] https://www.gazzettaufficiale.it/eli/id/1971/04/03/071U0124/sg



(a) ballerino.

(b) poetessa.

Figure 3: Occurrences of Named Entities associated to two occupational titles. The X-axis reports the time periods. The left Y-axis reports the overall smoothed relative frequency of the occupational title. The right Y-axis reports the absolute frequency of each Named Entity.

the Named Entities (NEs) to which the occupational titles refer for each year, and monitor their frequency. In particular, we exploit the UDPipe annotations to extract valid NEs, i.e. entities that are directly connected to an occupational title via a dependency relation.

In Figure 3, we report the NEs extracted for two particular occupational titles: *ballerino* (i.e. male dancer) and *poetessa* (i.e. female poet). We have chosen these titles because they feature the largest number of occurrences of NEs in the corpus. The data is presented in the form of stacked line charts, which report the absolute frequency of each NE so that the height of a coloured line represents how many times a NE has been mentioned within a specified period. The dotted black line reports the overall smoothed relative frequency for the occupational title. Both the absolute frequency of NEs mentions and the overall smoothed relative frequency are aggregated in bins of 5 years.

Three male dancers are referenced over a wide period due to their historical role in the field: *Rudolf Nureyev*, *Antonio Gades* and *Gene Kelly*. However, the last years have seen a rise in popularity of new figures such as *Raffaele Paganini*, *Joaquin Cortes*, *André de La Roche* and *Roberto Bolle*.

Occurrences of specific female poets in the corpus keep low until the late ’70s. Ignoring a spike in 1953-1957, probably due to the quality issues in the data collected, the individual absolute frequency of NE mentions seems to agree with the overall smoothed relative frequency of the noun *poetessa*. In the 1988-2002 period, four figures overwhelm the scene: *Joy Grisham*, *Elena Carasso*, *Maria Luisa Spaziani* and *Alda Merini*. Even though the first work of *Maria Luisa Spaziani* dates back to 1954, we observe a significant rise in the occurrences in the early ’90s, when she is nominated three times for the Nobel Prize for Literature [5]. The increase in NE mentions over time is even more apparent in this case, however, it follows a different trend compared to that of the overall frequency of the noun *poetessa*, which suggests that the word may have been used differently in the earliest period.

6 Conclusion

This paper investigates the usage of gender-specific forms of occupational titles in the Italian language in a diachronic corpus of 3 billion tokens extracted from two popular Italian newspapers. Through this analysis, we show that there are significant changes in the way newspaper articles refer to the masculine and feminine form of an occupational title and that they are consistent with socio-cultural events, such as changes in the employment policy. Moreover, we performed a more fine-grained analysis by extracting the most influential figures that have guided this shift for two occupational titles (male dancers and female poets).

[5] https://en.wikipedia.org/wiki/Maria_Luisa_Spaziani

As future work, we propose to continue work on this field by increasing the size of the corpus and by including sources other than news, such as social media, job applications, and legal documents. This can help reduce any form of linguistic bias that may have been introduced by journalists and increase the significance of the results. Moreover, we will extend the list of occupational titles, as well as group titles together based on category. Finally, we propose to improve the process used to extract named entities that are associated with occupational titles in text.

Acknowledgments

This research has been partially funded by ADISU Puglia under the post-graduate programme "Emotional city: a location-aware sentiment analysis platform for mining citizen opinions and monitoring the perception of quality of life".

References

Erez Lieberman Aiden and Jean-Baptiste Michel. 2011. Culturomics: Quantitative Analysis of Culture Using Millions of Digitized Books. In *6th Annual International Conference of the Alliance of Digital Humanities Organizations, DH*, page 8, Stanford, CA, USA, June. Stanford University Library.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. A Diachronic Italian Corpus based on "L'Unità". In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, 3. CEUR-WS.org.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. *The Evalita 2014 Dependency Parsing task*, pages 1–8.

Elisabeth Burr. 1995. Agentivi e sessi in un corpus di giornali italiani. In Gianna Marcato, editor, *Atti del Convegno Internazionale di studi Dialettologia al femminile*, pages 349–365, Padova, Italy, April. Cleup.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. 2007. *Regression*. Springer.

Yoav Goldberg and Jon Orwant. 2013. A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. *Atlanta, Georgia, USA*, page 241.

Pascal Mark Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. 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.

Ruth Elizabeth King. 1991. *Talking gender: A guide to nonsexist communication*. Copp Clark Professional.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017. Tracing armed conflicts with diachronic word embedding models. In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard H. Hovy, Teruko Mitamura, and David Caswell, editors, *Proceedings of the Events and Stories in the News Workshop@ACL 2017*, pages 31–36, Vancouver, Canada, August. Association for Computational Linguistics.

Tarutuulia Laine and Greg Watson. 2014. Linguistic sexism in The Times-A diachronic study. *International Journal of English Linguistics*, 4(3):1.

Bandy X Lee, Finn Kjaerulf, Shannon Turner, Larry Cohen, Peter D Donnelly, Robert Muggah, Rachel Davis, Anna Realini, Berit Kieselbach, Lori Snyder MacGregor, et al. 2016. Transforming our world: implementing the 2030 agenda through sustainable development goal indicators. *Journal of public health policy*, 37(1):13–31.

Gianna Marcato and Eva-Maria Thüne. 2002. Gender and female visibility in Italian. *Gender across languages: The linguistic representation of women and men*, 2:187–217.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword fifth edition, 2011. *Linguistic Data Consortium, Philadelphia, PA, USA*.

Alma Sabatini. 1985. Occupational titles in Italian: Changing the sexist usage. In *Sprachwandel und feministische Sprachpolitik: Internationale Perspektiven*, pages 64–75. Springer.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, Portorož,Slovenia, 5. European Language Resources Association (ELRA).

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. Publisher: ACM New York, NY, USA.