UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

L’uso del femminile nei nomi professionali e nei titoli onorifici su Wikipedia in lingua italiana

Candidato: *Rebecca Di Gisi*

Relatore: *Mirko Luigi Aurelio Tavosanis*

Correlatore: *Maria Simi*

Anno Accademico 2020-2021

Sommario

Introduzione 4

1 Il problema dei femminili professionali 7

1.1 Il contributo di Alma Sabatini 8

1.2 Le soluzioni 11

1.3 La situazione attuale 13

2 Estrazione ed elaborazione del corpus 17

2.1 I *dump* di Wikipedia 17

2.2 Estrazione con WikiExtractor 19

2.2.1 ExtractPage 22

2.3 Sentence-Splitting e POS-Tagging del testo 24

2.3.1 Sentence-Splitting 24

2.3.2 POS-tagging 25

3 I femminili professionali su Wikipedia in lingua italiana 28

3.1 I corpora 28

3.2 Oggetto dell'analisi 29

3.3 Modalità di analisi 30

3.3.1 Opzioni del comando grep 32

3.4 Analisi del rapporto tra le occorrenze femminili e maschili dei titoli onorifici 34

3.4.1 Risultati 38

3.5 Analisi delle tendenze linguistiche riguardo la scelta tra diversi femminili 41

3.5.1 Risultati 43

4 Conclusioni 47

5 Bibliografia 49

6 Sitografia 50

6.1 Fonti linguistiche 50

6.2 Fonti informatiche 50

6.3 WikiExtractor 50

6.4 Wikipedia 51

Introduzione

La lingua è, probabilmente, lo strumento più potente che l'essere umano abbia avuto a disposizione nella sua intera storia. Nata verosimilmente dalla più arcaica delle necessità, ovvero quella della comunicazione, la capacità linguistica sta all'origine ed è al contempo il risultato dell'essere umano per come lo intendiamo oggi.

La lingua parlata, infatti, svolge la funzione fondamentale di agevolare la comprensione del mondo. Permettendo di associare un nome agli oggetti e ai fenomeni di cui si fa esperienza, di identificare sensazioni ed emozioni ma, soprattutto, di comunicarle alla società di cui ogni individuo è parte integrante, la lingua si fa lente nei confronti del mondo circostante. Permette quindi una determinata visione del mondo, in base alle parole e ai sistemi che essa stessa offre per comprendere ciò che il mondo è. La lingua svolge, dunque, una funzione di filtro, influenzando la percezione dei parlanti riguardo la realtà e limitando la prospettiva che essi hanno del mondo in base alle parole che permettono loro di descriverlo. Proprio da questo deriva la grande necessità di dare estrema importanza alle parole che vengono utilizzate quotidianamente, poiché creano la nostra percezione del mondo ma rischiano, talvolta, di limitarci all'interno di quella stessa percezione.

La realtà collettiva che caratterizza le società umane è anch'essa, ovviamente, descritta dalla lingua; e l'attenzione che deve essere posta sulla connotazione della società è alta, in quanto la società determina il tipo di vita che ogni persona può vivere, il tipo di considerazione di cui può godere da parte dei pari e tutta una serie di altre realtà. È chiaro quindi che se è la lingua a descrivere la società in cui viviamo, è estremamente importante soffermarsi sull'attenzione che poniamo, ogni giorno, nell'utilizzarla. Il rischio dell'utilizzo improprio della lingua, soprattutto in questo ambito, è quello di perpetuare una continua divisione sociale in un moto che tende all'autoconservazione, e che rischia dunque di impedire la possibilità di cambiamento.

Società e lingua sono concretamente connesse: l'una può manipolare l'altra, e l'analisi dell'una può spiegare e fornire informazioni sulla struttura dell'altra, e viceversa. Questo legame viscerale permette alle indagini linguistiche di ottenere descrizioni piuttosto accurate della società dei parlanti.

Come dichiarato da Alma Sabatini[1]: «La lingua italiana, come molte altre, è basata su un principio androcentrico: l'uomo è il parametro intorno a cui ruota e si organizza l'universo

[1] Alma Sabatini, *Il sessismo nella lingua italiana*, p. 20.

linguistico». È infatti necessaria solo una minima capacità di osservazione per confermare il parallelismo della situazione linguistica con le effettive circostanze sociali.

Alcuni esempi problematici potrebbero essere, ad esempio, l'utilizzo della parola "uomo" con valore non marcato, includendo dunque all'interno della categoria anche le donne; oppure l'uso del maschile universale – o maschile non marcato – impiegato con un'accezione di neutralità nella lingua d'uso quotidiano, ma che a un'analisi più approfondita dimostra solo quanto effettivamente la lingua e la società siano organizzate in un'accezione quasi interamente maschile e che spesso, di conseguenza, prevede un unico e parziale punto di vista. L'utilizzo abitudinario e inconsapevole di questa tipologia di forme linguistiche risulta nella marginalizzazione e nella riduzione della rappresentazione femminile, nonché nel rafforzamento della posizione di inferiorità della donna nei confronti dell'uomo all'interno della società.

La soluzione a una mancanza di rappresentazione di questo tipo, spesso, viene ricercata in una sorta di neutralità purtroppo impossibile nella lingua italiana, che è organizzata sull'esistenza di due soli generi: quello maschile e quello femminile. Di conseguenza, optare per un neutro derivante dalle forme maschili rischia di sfociare nell'annullamento delle differenze e nell'appiattimento delle possibilità linguistiche. Anche perché, come sostenuto da Alma Sabatini[2] ne *Il sessismo nella lingua italiana*, non sono le differenze in sé ad essere il problema, bensì il valore intrinseco che a queste differenze viene comunemente associato. Proprio alla percezione di un maggiore valore si deve la motivazione, ad esempio, dell'uso di sostantivi maschili per l'identificazione professionale di donne, soprattutto in ambiti professionali di prestigio. In questo caso il valore intrinseco associato alla differenza d'uso sta nella concezione comune che il sostantivo maschile comporti un maggior grado di professionalità e competenza rispetto al corrispettivo femminile.

Di nuovo, dunque, siamo spettatori dell'uso di un linguaggio basato sull'universalità e sulla centralità maschile, che associa competenze a un sostantivo di un genere e pone una preventiva nota di demerito sull'utilizzo dello stesso sostantivo declinato al femminile. Parole come *ministra*, *sindaca*, *avvocata* e *ingegnera* subiscono una costante declassazione derivata dall'unica motivazione di essere sostantivi femminili poco utilizzati e, conseguentemente, tacciati di una aprioristica svalutazione che si riversa naturalmente anche sulla donna a cui quel sostantivo è associato.

Questo argomento è teatro di controversie da diverse decine di anni e il risultato è che, con il passare del tempo, questi sostantivi stanno, molto lentamente, iniziando a perdere la connotazione umoristica e svilente che li caratterizzava, entrando poco alla volta nel vocabolario dei parlanti.

[2] Alma Sabatini, *Il sessismo nella lingua italiana*, p. 21.

La presente ricerca vuole dunque effettuare una stima del cambiamento che questo tipo di discussione, perpetuata nel corso degli ultimi anni, ha ottenuto sull'entrata effettiva nell'uso linguistico di questi termini.

L'analisi sarà incentrata sulla comparazione dei risultati di ricerche linguistiche effettuate su diversi corpora di Wikipedia, estratti in quattro anni diversi: nel 2013 e 2015, quando la discussione pubblica sull'uso dei sostantivi femminili professionali era ancora agli inizi; nel 2019, anno di accese discussioni sociali e politiche; nel 2021, l'anno corrente.

1 Il problema dei femminili professionali

Il frequente uso scorretto delle forme femminili di sostantivi professionali e di titoli onorifici comporta, spesso, una mancata percezione della presenza femminile in determinati contesti sociali e lavorativi. Di conseguenza, l'ottenimento di una lingua italiana paritaria rappresenterebbe una vera e propria validazione dell'esistenza delle donne in quei particolari settori e, in generale, come parte integrante della società, contribuendo inoltre a cambiarne la percezione comune.

Al momento, però, è riscontrabile una grande diffidenza da parte dei parlanti nei confronti di queste forme, che, andando a intaccare abitudini consolidate, vengono utilizzate con una certa riluttanza. Le motivazioni che più di frequente vengono date per giustificare questo atteggiamento linguistico sono, come riportato da Cecilia Robustelli[3], principalmente tre. La prima è l'insicurezza data dall'utilizzo di una forma nuova e meno comune rispetto a quella a cui si è normalmente abituati ad utilizzare per alcuni contesti – come, ad esempio, nel caso dell'uso di *ingegnera* al posto del più comune *ingegnere*, abitualmente utilizzato nella forma maschile anche quando rivolto a donne–; la seconda è che le nuove forme femminili vengono percepite come esteticamente brutte e cacofoniche dai parlanti; e infine vi è la convinzione che sia grammaticalmente corretto utilizzare la forma maschile anche in riferimento alle donne.

I femminili professionali sono infatti sovente considerati alla stregua di parole nuove e, di conseguenza, trattati con sospetto e spesso ostracizzati. Tecnicamente, però, oltre a non essere scorretti dal punto di vista grammaticale, questi sostantivi femminili non sono dei veri *neologismi*.

Come sostenuto dalla linguista Vera Gheno, infatti: «Da un punto di vista linguistico, i femminili 'inediti' sono forme previste dal sistema italiano, che fino a tempi recenti erano come rimaste 'dormienti' perché non servivano»[4]. Questa tipologia di parole, dunque, è da sempre stata potenzialmente prevista all'interno del sistema linguistico dell'italiano, insieme a tutti gli altri sostantivi femminili derivati dalla forma maschile, tuttavia, fino a poco tempo fa, non ha trovato riscontro nella realtà, essendoci la mancanza di soggetti di sesso femminile a cui associarla.

Tutto ciò è confermato dal fatto che negli ambiti in cui la donna è tradizionalmente presente l'uso dei femminili non è mai stato messo in discussione e che, dunque, quando ci si rifiuta di utilizzare un titolo professionale al femminile, lo si fa in ragione di stereotipi interiorizzati e di pregiudizi culturali[5].

[3] Cecilia Robustelli, *Infermiera sì, ingegnera no?*.
[4] Vera Gheno, *Femminili singolari*, p. 51.
[5] Patrizia Bellucci, *Il femminile di questore e di prefetto*.

Come asserito anche da Robustelli, infatti: «Le resistenze all'uso del genere grammaticale femminile per molti titoli professionali o ruoli istituzionali ricoperti da donne sembrano poggiare su ragioni di tipo linguistico, ma in realtà sono, celatamente, di tipo culturale»[6].

È quindi chiaro che le motivazioni sopracitate, spesso utilizzate come giustificazione a un mancato – o scorretto – utilizzo dei femminili di titoli e professioni hanno poco a che vedere con le regole linguistiche e grammaticali.

1.1 Il contributo di Alma Sabatini

L'analisi sullo scarso utilizzo dei sostantivi femminili di professione, soprattutto per quanto riguarda cariche pubbliche e professioni di prestigio, è stato affrontato per la prima volta nella storia della lingua italiana nel 1987, dalla linguista Alma Sabatini, nello studio *Il sessismo nella lingua italiana*.

La ricerca illustra come, effettivamente, l'atteggiamento linguistico e generale della società nei confronti delle donne sia carico di sfiducia e discriminazione, e quanto la parità sia molto spesso esclusivamente un principio giuridico poco trasposto nella quotidianità.

Il fine della ricerca è fondamentalmente quello di rilevare e classificare le forme sessiste nella lingua italiana d'uso soprattutto nell'ambiente della stampa, per portarne alla luce le problematiche. Sabatini, infatti, sottolinea l'importanza di un corretto utilizzo della lingua al fine di costruire una realtà paritaria e non sessista, e sostiene che le forme linguistiche siano spesso portatrici inconsapevoli di ideologie e pregiudizi talmente radicati nella struttura linguistica da renderne difficile persino il riconoscimento. Per questo motivo diventa di centrale importanza l'utilizzo consapevole della lingua, e la stessa linguista alla fine del suo saggio dedica un intero capitolo all'analisi di forme linguistiche di uso comune considerate sessiste, fornendo delle proposte alternative al loro utilizzo[7].

La ricerca ha inizio soffermandosi, innanzitutto, sull'utilizzo frequente della forma maschile per inglobare uomini e donne indistintamente, e sottolinea come questo tipo di abitudine linguistica produca, a tutti gli effetti, dei risultati di cancellazione e riduzione della rappresentazione femminile.

In particolare, Alma Sabatini si riferisce alle differenze create da questo tipo di utilizzo del maschile come *dissimmetrie grammaticali*, e ne specifica alcuni esempi[8]:

[6] Cecilia Robustelli, *Infermiera sì, ingegnera no?*.
[7] Alma Sabatini, *Il sessismo nella lingua italiana*, capitolo III.2.
[8] Alma Sabatini, *Il sessismo nella lingua italiana*, pp. 24-25.

- il maschile non marcato – anche detto inclusivo –, come avviene ad esempio nell'utilizzo della parola *uomo* con valore generico, includendo dunque anche esseri di sesso femminile;
- la concordanza al maschile, ovvero l'utilizzo di aggettivi e participi passati maschili concordati con una serie di nomi sia femminili che maschili, spesso determinato dalla presenza di un solo nome maschile – questo tipo di utilizzo del maschile è anche definito *inglobamento* –;
- la precedenza del maschile nelle coppie uomo-donna;
- l'assenza di forme femminili *simmetriche* a quelle maschili nel caso di titoli professionali e ruoli istituzionali di prestigio. In questi casi la forma o è totalmente assente o, quando c'è, viene spesso espressa con suffissi riduttivi, come ad esempio la forma *-essa*.

Partendo dunque dalle dissimmetrie grammaticali, Sabatini avvia la discussione riguardo i sostantivi femminili legati a ruoli professionali e titoli onorifici, un problema che proprio in quegli anni stava cominciando a trovare la necessità di essere discusso, in quanto le donne iniziavano ad entrare in alcune tipologie di occupazioni che fino a quel momento erano state una prerogativa esclusivamente maschile. In questo contesto, dunque, le parole che vengono utilizzate per le nuove occupazioni femminili sono contrastanti e confuse, oltre che, talvolta, velatamente insultanti.

Le forme tendenzialmente utilizzate per creare i nuovi femminili al tempo dell'analisi di Sabatini, ma che in molti casi vengono impiegate ancora oggi dai parlanti, si possono riassumere nelle seguenti categorie[9]:

- l'uso del titolo al maschile con concordanza al maschile di articoli e aggettivi, nonostante il soggetto in questione sia una donna. Chiaramente l'utilizzo del titolo al maschile non fa altro che sottolineare quanto la presenza della donna in quella determinata circostanza sia del tutto eccezionale e provvisoria, nonché accentua la convinzione che alla forma maschile sia associata una maggiore competenza. Inoltre, questo tipo di utilizzo della forma maschile rivolta a un essere di sesso femminile è una soluzione del tutto insoddisfacente e scorretta dal punto di vista grammaticale;
- l'uso del modificatore "donna" anteposto o posposto al nome base. Questa forma è chiaramente dissimmetrica, in quanto non c'è alcun sostantivo femminile la cui forma

[9] Alma Sabatini, *Il sessismo nella lingua italiana*, p. 25.

al maschile preveda l'accompagnamento della specifica "uomo". Ancora una volta il risultato è quello di enfatizzare quanto il ruolo solitamente spetti a un uomo, e che dunque una presenza femminile sia solo un'eccezione vagamente tollerata;

- la derivazione da un sostantivo maschile tramite l'aggiunta del suffisso *-essa*. La problematica in questo caso è la connotazione dispregiativa che molte forme femminili realizzate con questo suffisso hanno assunto nell'uso quotidiano. Spesso, in realtà, la connotazione negativa è andata disperdendosi, soprattutto per quelle forme che sono entrate nell'utilizzo comune grazie a una grande presenza di donne in quei campi – come successo, ad esempio, per *dottoressa* o *professoressa* – ma rimane negli ambiti in cui la presenza femminile è ancora scarsa. Questa forma, inoltre, è spesso usata in maniera impropria, anche in casi in cui le regole della lingua italiana prevederebbero un femminile formato in un altro modo (come accade, ad esempio, nel caso di *presidentessa*, forma grammaticalmente scorretta in quanto *presidente* è un nome epiceno e, come tale, necessita esclusivamente di un articolo femminile anteposto per poter essere considerato femminile);
- l'utilizzo del sostantivo maschile preceduto dall'articolo femminile. Questa soluzione è utilizzata soprattutto con titoli onorifici o cariche pubbliche. In questo caso, in particolare, si percepisce un evidente scontro tra la volontà di notificare una presenza femminile e l'abitudine all'utilizzo di determinate forme linguistiche esclusivamente al maschile, soprattutto in determinati ambienti sociali.

L'utilizzo di queste strutture tende a ribadire costantemente come la forma maschile sia la materia originale da cui viene successivamente derivato il femminile, sia in ambito linguistico che in ambito socioculturale, operando costantemente una marginalizzazione delle forme femminili e, di conseguenza, delle donne, e appoggiando la convinzione che per raggiungere una considerazione paritaria tra uomo e donna sia in un certo modo necessaria l'eliminazione della percezione di ogni differenza tra i sessi.

1.2 Le soluzioni

Dalla fine degli anni Ottanta, quando Alma Sabatini scriveva *Il sessismo nella lingua italiana*, l'argomento – proprio per merito della pubblicazione stessa – è diventato sempre più spesso oggetto di dibattiti e discussioni, specialmente in ambito accademico, portando a ulteriori considerazioni a riguardo, e, soprattutto, a nuovi suggerimenti riguardo l'uso del femminile.

In particolare, l'Accademia della Crusca ha dedicato diversi interventi e consulenze sulla questione, provvedendo alla disambiguazioni di casi che creavano dubbi tra i parlanti e orientandosi tendenzialmente a favore della femminilizzazione dei sostantivi professionali.

Proprio per conto dell'Accademia della Crusca, inoltre, la linguista Cecilia Robustelli si è occupata del progetto *Linee guida per l'uso del genere nel linguaggio amministrativo*, promosso dal Comune di Firenze nel 2012. Il lavoro della linguista riprende dalla conclusione del saggio di Alma Sabatini e si focalizza, in particolare, sul linguaggio amministrativo, illustrando alcune linee guida per quanto riguarda la scrittura burocratica. Chiaramente, nonostante cambino il settore in cui è effettuata l'analisi e i destinatari dei consigli riguardo l'utilizzo dei sostantivi femminili, resta invariata l'opinione sul ruolo di fondamentale importanza del linguaggio come maggior strumento d'azione nella realizzazione della parità, condivisa da Cecilia Robustelli con Alma Sabatini.

Robustelli, infatti, riprende, nella maggior parte dei casi, le indicazioni per l'uso dei femminili professionali suggerite da Sabatini, accettando le forme già entrate nell'uso comune e soffermandosi su quelle che ancora ad oggi fanno fatica ad essere integrate nel lessico comune, fornendo la seguente ricapitolazione sulla formazione dei femminili[10]:

- i termini maschili in *-o* e *-aio/-ario* mutano in *-a* e *-aia/-aria*, come *ministro-ministra*, *notaio-notaia*;
- i termini in *-iere* mutano in *-iera*, come *consigliere-consigliera*;
- i termini in *-sore* mutano in *-sora*, come *assessore-assessora*;
- i termini in *-tore* mutano in *-trice*, come *senatore-senatrice*.

Vi sono poi dei casi in cui la forma del termine non cambia e si ha soltanto l'anteposizione dell'articolo femminile, come succede ad esempio per:

[10] Cecilia Robustelli, *Linee guida per l'uso del genere nel linguaggio amministrativo*, pp. 19-20.

- i termini in *-e* o *-a*, come *giudice* o *preside*;
- le forme italianizzate di participi presenti latini, come *dirigente* o *presidente*;
- i composti con *capo-*, come *capofamiglia*.

Robustelli specifica, inoltre, che sarebbe preferibile non applicare le precedenti indicazioni a forme ormai pienamente attestate nell'uso comune e prive di alcuna connotazione negativa, in quanto è chiaramente più semplice conservare delle forme femminili ormai attestate piuttosto che spingere a un ulteriore cambiamento nelle abitudini dei parlanti. La linguista fa riferimento, in particolare, a sostantivi come *direttrice* o alle forme in *-essa* come *professoressa* o *dottoressa*.

Per quanto riguarda, inoltre, il riferimento a più persone, le soluzioni da lei proposte per evitare l'inglobamento del maschile e del femminile nell'unica forma maschile, sono due[11]:

- la prima soluzione mira a donare una nuova visibilità del genere femminile tramite un uso simmetrico del genere: si incentiva dunque la dichiarazione di entrambe le forme, invece che solo il maschile come comunemente d'uso (anche nelle forme abbreviate);
- la seconda soluzione, invece, prevede un oscuramento di entrambi i generi e si attua nell'uso di sostantivi o perifrasi che non includano riferimenti ad alcun genere, come ad esempio *individui*, *persona, essere umano*; o nell'utilizzo di nomi collettivi che includano a prescindere entrambi i sessi, come *personale, presidenza, utenza*.

Chiaramente quelli di Cecilia Robustelli restano indicazioni e proposte ma, come sottolineato da Claudio Marazzini, linguista e presidente dell'Accademia della Crusca: «Non si deve dimenticare che la lingua accetta oscillazioni e che non è necessario avere sempre e una sola soluzione univoca»[12].

[11] Cecilia Robustelli, *Linee guida per l'uso del genere nel linguaggio amministrativo*, p. 21.
[12] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, p. 123.

1.3 La situazione attuale

Dallo studio condotto nel 2017 da Giuseppe Zarra per l'Accademia della Crusca, sulla percezione dei parlanti riguardo la femminilizzazione dei titoli professionali e politici[13], si evince la presenza di due principali tendenze di pensiero.

La prima si schiera a sostegno dell'utilizzo della forma maschile nelle cariche pubbliche anche per le donne, in quanto sostiene che il maschile – e dunque il sostantivo originariamente creato per denominare quella carica – focalizzi l'attenzione sulla funzione e sul ruolo, piuttosto che sulla persona che lo ricopre; e che invece l'utilizzo di un sostantivo femminile rischi di essere sminuente per la carica in sé, oltre che per la persona che è investita del titolo.

La seconda posizione, invece, appoggia l'utilizzo e la diffusione delle forme femminili, in particolar modo nelle cariche istituzionali e nei titoli onorifici, allo scopo di normalizzare la presenza femminile anche nelle professioni più prestigiose e tendenzialmente riservate agli uomini, ma, soprattutto, come riportato da Giuseppe Zarra: «[…] allo scopo di difendere la piena dignità dell'impegno professionale delle donne»[14].

Entrambe le prese di posizione sono del tutto rispettabili e hanno il sostegno da parte di linguisti e donne in prima persona occupate in professioni e titoli di prestigio. Come già sottolineato in precedenza, infatti, la scelta dell'utilizzo dei femminili di titoli professionali e cariche pubbliche dipende da fattori sociolinguistici interamente dipendenti dalla libera scelta e dalla volontà di utilizzo della lingua del singolo parlante.

Zarra svolge inoltre una ricerca riguardo la diffusione di otto diversi sostantivi femminili, calcolandone le occorrenze all'interno delle pagine del motore di ricerca Google e paragonandole al numero di occorrenze dei corrispondenti maschili, che, chiaramente, a loro volta potrebbero inglobare anche referenti femminili.

I risultati di questa analisi[15] sottolineano come, a tutti gli effetti, alcuni femminili siano più radicati all'interno del linguaggio comune rispetto ad altri. Infatti, è interessante notare come alcuni sostantivi, al momento in cui la ricerca è stata effettuata – nell'agosto del 2016 –, mostrano un rapporto che tende alla parità tra il numero di occorrenze della forma femminile e quelle della forma maschile. Ciò si evince in coppie di sostantivi come *deputata* e *deputato*, che mostrano un rapporto di 1:1,53, e *ministra* e *ministro*, con un rapporto di 1:1,78. Altre coppie di sostantivi, invece,

[13] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, capitolo 2.3.
[14] *Ivi*, p. 31.
[15] *Ivi*, p. 36.

presentano un dislivello ancora estremamente alto, come nel caso di *chirurga* e *chirurgo* (1:187,29), *magistrata* e *magistrato* (1:80,87) o ancora *ingegnera* e *ingegnere* (1:76,95).

Analizzando questi dati in paragone con quelli raccolti da Robustelli nella sua ricerca[16] relativa ai primi anni del Duemila, e tenendo conto dell'enorme incremento delle pagine web create negli anni che intercorrono tra le due misurazioni, Zarra sostiene dunque che si riscontri «[…] un complessivo innalzamento delle occorrenze totali dei femminili e una riduzione del rapporto di occorrenze delle forme femminili rispetto alle corrispondenti maschili per tutti i lessemi […]»[17].

L'unica eccezione citata è il rapporto fra *magistrata* e *magistrato*, che la misurazione di Serianni del 2006[18] riconduce a un rapporto di 1:76,22, che è rimasto più o meno il medesimo anche nella misurazione di Zarra del 2016, la quale, come anticipato, riferisce un rapporto di 1;80,87.

L'analisi di Zarra procede con l'intento di ricercare e stabilire quale sia l'uso più diffuso tra il maschile con accezione neutra e l'utilizzo del femminile, tramite il motore di ricerca Google e nel sito dell'Ansa. Viene stilata una lista di donne categorizzate tramite la propria occupazione, facente parte proprio di quella categoria di sostantivi professionali femminili gravemente penalizzata dallo scarso uso comune.

La ricerca utilizza la sequenza formata dall'articolo determinativo, seguito dal titolo – secondo varie possibilità – e dal cognome. Un esempio di risultato è quello che riguarda la ministra della Difesa Roberta Pinotti, riportato a seguire[19].

	Google	Ansa
la ministra Pinotti	11.800	6
il ministro Pinotti	59.000	40
la ministressa Pinotti	1	0
la ministro Pinotti	619	0
la donna ministro Pinotti	0	0
il ministro donna Pinotti	0	0

[16] Cecilia Robustelli, *Sindaco e sindaca: il linguaggio di genere*.
[17] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, p. 36.
[18] Luca Serianni, *Prima lezione di grammatica*.
[19] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, p. 37.

Da questo esito si evince come nel web venga prediletto, in generale, l'utilizzo del maschile neutro. La sola eccezione sottolineata da Zarra riguarda il femminile *sindaca*, a favore del quale viene riportata, però, l'azione di una campagna mediatica durante l'anno 2016[20]. Viene inoltre sottolineato l'uso ironico e spregiativo della forma *ministressa*, e messo a paragone con forme quali *avvocatessa* che, oltre a non essere connotate negativamente, sono addirittura più frequenti della forma *avvocata*.

La ricerca continua poi su altri siti istituzionali, come ad esempio il sito ufficiale della Presidenza del Consiglio dei Ministri, il sito del Ministero della Salute, il sito del Ministero per le Riforme Costituzionali, dove le ministre sono sempre identificate come *ministro* e tutti i titoli politici assunti da donne sono esclusivamente indicati al maschile.

Viene analizzato inoltre un documento ufficiale del Senato[21] che riporta l'elenco degli eletti della XVII legislatura e ne indica la professione. Dal documento preso in analisi, si può notare che i titoli professionali delle 92 senatrici vengono proclamati al femminile solo nel 2,86% dei casi[22] e principalmente per quanto riguarda i sostantivi *disoccupata* e *impiegata*. Il resto dei titoli è diviso tra sostantivi epiceni (31,43%), sostantivi epiceni con accordo al femminile (0,95%) e al maschile (10,48%), e sostantivi maschili (54,28%).

I risultati di questa ricerca sono particolarmente interessanti, poiché oltre all'evidente prevalenza dell'uso dei sostantivi maschili in generale, possiamo anche notare come questi ultimi prevalgano su quelli femminili non solamente nel caso di cariche di prestigio e titoli onorifici come nei casi già citati di *architetto*, *avvocato*, *magistrato*, spettatori di una lenta affermazione nelle loro forme femminili, ma anche a discapito di quelli pienamente affermati, come *impiegato*, *operaio*, *professore*. Questo documento ufficiale, dunque, mette in luce l'effettiva lontananza dalla parità di trattamento per quanto riguarda i femminili professionali anche – e soprattutto – in un contesto come quello politico e normativo, che dovrebbe invece gettare le basi per un uso della lingua più coerente, chiaro e paritario possibile.

Le ricerche citate offrono uno spaccato molto interessante e significativo riguardo l'utilizzo delle forme femminili di sostantivi professionali nella stampa, in documenti ufficiali e su internet, registrando una conversione, seppur lenta, verso un maggior uso dei femminili. Un punto di vista

[20] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, p. 39.
[21] <https://www.senato.it/application/xmanager/projects/leg18/file/elenco_senatori_n_2_XVII_defintivo.pdf>.
[22] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, pp. 41-42.

sicuramente interessante che resta escluso da questi studi è, invece, la situazione attuale della lingua enciclopedica italiana nei riguardi dell'uso dei sostantivi femminili di professione.

Una possibilità estremamente interessante è data in questo caso da Wikipedia in lingua italiana che, essendo un'enciclopedia online, ha la possibilità di aggiornarsi in tempo reale in riflesso ai cambiamenti del mondo circostante. La particolarità più grande di Wikipedia è, inoltre, l'essere collaborativa. Al fatto che chiunque possa scrivere un articolo su Wikipedia, si aggiunge anche il fatto che le decisioni sui cambiamenti linguistici – come in questo caso la femminilizzazione dei sostantivi maschili rivolti a donne – vengono prese tramite diverse discussioni aperte alla partecipazione di tutti gli utenti. Ciò implica la presenza di una sorta di democrazia nella direzione dei cambiamenti stilistici e di lingua da applicare agli articoli.

Considerare questo aspetto può aiutare a mettere in luce anche l'effettiva tendenza dei parlanti e le eventuali remore nei confronti dei cambiamenti linguistici. Sempre riguardo questo punto, bisogna inoltre considerare che Wikipedia in lingua italiana non è, come spesso erroneamente considerata, composta dalle traduzioni di articoli originariamente scritti in altre lingue, ma anzi, la maggior parte delle voci è scritta da utenti italofoni che, dunque, scrivono gli articoli secondo la loro personale conoscenza linguistica[23].

Nella presente analisi, dunque, partendo dai sostantivi femminili precedentemente analizzati da Robustelli[24] e Zarra[25] nei rispettivi saggi, verrà ripetuta una ricerca similare per essere in grado di valutare il cambiamento effettivo della lingua italiana nell'enciclopedia online Wikipedia. I sostantivi presi in analisi verranno valutati nel loro numero di occorrenze totali e paragonati ai corrispettivi maschili, per stimare un rapporto tra le due incidenze con una proporzione 1:x tra il numero di occorrenze femminili e il numero di occorrenze maschili. Successivamente, invece, verranno valutate le occorrenze delle diverse soluzioni per indicare il femminile di uno stesso sostantivo. I corpora su cui sarà effettuata l'analisi sono composti da tutti gli articoli di Wikipedia in lingua italiana estratti in tre anni differenti: il 2015, il 2019 e il 2021. Chiaramente, ottenuti i risultati, sarà necessario tenere in considerazione il fatto che la composizione di Wikipedia nel corso degli anni è ampiamente mutata, soprattutto per numero di voci (passando, ad esempio, dalle più di 1 500 000 voci a febbraio 2019[26], alle 1 701 413 voci del 25 giugno 2021[27]), parametro con cui valutare attentamente l'eventuale crescita delle occorrenze di determinati sostantivi.

[23] <https://it.wikipedia.org/wiki/Wikipedia_in_italiano>.
[24] Cecilia Robustelli, *Sindaco e sindaca: il linguaggio di genere*.
[25] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*.
[26] <https://it.wikipedia.org/wiki/Wikipedia_in_italiano#Cronologia_recente>.

2 Estrazione ed elaborazione del corpus

2.1 I *dump* di Wikipedia

Per l'ottenimento di un corpus composto da tutti gli articoli di Wikipedia in una data lingua in un determinato momento, bisogna innanzitutto partire da specifici file che archivino tutti gli articoli presenti. Questi file sono chiamati *dump*[28].

Un dump è un file che include il contenuto e la struttura di un determinato database, ed è sostanzialmente utilizzato per finalità di archivio di dati. Wikipedia mette a disposizione degli utenti i dump degli articoli suddivisi per lingua, nella maggior parte delle lingue di uso comune.

I dump di Wikipedia sono tendenzialmente utilizzati dai ricercatori per l'archiviazione, in ragione della possibilità che danno di lavorare su formati facilmente interrogabili e riutilizzabili. Wikipedia offre, per ogni lingua disponibile, diverse tipologie di *data dumps*, in funzione delle esigenze dell'utente. Alcuni esempi di tipologia di dump sono[29]:

- *pages-articles.xml*, un archivio XML delle pagine dell'enciclopedia con il testo della sola versione corrente. Comprende dunque esclusivamente le voci presenti su Wikipedia.
- *pages-meta-current.xml*, un archivio XML di tutte le pagine (voci e pagine di servizio), con il testo della sola versione corrente.
- *pages-meta-history.xml*, un archivio XML di tutte le pagine (voci e pagine di servizio), con il testo della versione corrente e di tutte le precedenti revisioni e modifiche effettuate.

Ci sono poi ulteriori tipologie di dump create per esigenze particolari[30].

Gli articoli di Wikipedia hanno la particolarità di includere, oltre al testo di cui sono composti, anche il *MediaWiki Markup Language*[31], il linguaggio di marcatura di Wikipedia, che

[27] <https://it.wikipedia.org/wiki/Template:Numero_voci_in_Wikipedia>.
[28] I dump degli articoli di Wikipedia si possono trovare al seguente link, sostituendo a XX le sigle identificatrici dei linguaggi (ad esempio en, it, ecc) <http://dumps.wikimedia.org/XXwiki/latest/XXwiki-latest-pages-articles.xml.bz2>.
[29] < https://it.wikipedia.org/wiki/Aiuto:Analisi_del_database#Struttura_dei_file>.
[30] <https://dumps.wikimedia.org>.
[31] <https://it.wikipedia.org/wiki/MediaWiki>.

fornisce una notazione del testo in grado di inserire markup HTML all'interno del documento al fine di formattare e strutturare il testo[32].

Ad esempio:

- per connotare le diverse sezioni e sottosezioni la sintassi[33] prevede l'utilizzo del simbolo = nel seguente modo:

```
==Sezione==
===Sottosezione===
====Sotto-sottosezione====
```

- per andare a capo senza iniziare un nuovo paragrafo si utilizzano i tag `<br>` o `<br/>`.

Ciò avviene per permettere la corretta visualizzazione dell'articolo sulla pagina web che lo ospita e per agevolare le modifiche al testo da parte degli utenti. Proprio il fatto che gli articoli di Wikipedia siano manipolabili e modificabili fondamentalmente da chiunque risulta spesso nell'utilizzo improprio dei tag MediaWiki e HTML, fattore che potrebbe sollevare problematiche al momento dell'estrazione del testo.

I dump, dunque, offrono i dati di una specifica versione di Wikipedia in formato XML, includendo gli articoli esattamente nella forma in cui sono presenti online, con l'associazione della marcatura al dato testuale. Per poterne sfruttare il contenuto, dunque, è sostanzialmente d'obbligo effettuare un'estrazione che ripulisca il testo dalle annotazioni – comprese quelle improprie – e che permetta la visualizzazione del solo file testuale. Questa procedura è effettuabile con *WikiExtractor*, uno script in grado di estrarre il testo da un dump, occupandosi di ripulirlo dal linguaggio di marcatura.

[32] <https://it.wikipedia.org/wiki/Aiuto:Wikitesto>.
[33] <https://it.wikipedia.org/wiki/Aiuto:Sezioni>.

2.2 Estrazione con WikiExtractor

WikiExtractor[34] è uno script in Python scritto nel 2015 dal professor Giuseppe Attardi, docente e ricercatore presso l'Università di Pisa, che estrae il contenuto testuale di un dump del database di Wikipedia ripulendolo da ogni linguaggio di marcatura[35]. Per il suo corretto funzionamento non richiede alcuna libreria aggiuntiva ma solo Python 3.

Dopo aver effettuato il download del dump[36] da cui si vuole effettuare l'estrazione, lo script deve essere invocato utilizzando il nome del dump come argomento del seguente comando:

```
python -m wikiextractor.WikiExtractor.py <Wikipedia dump file>
```

Verrà istantaneamente avviata l'elaborazione del contenuto del dump.

Le operazioni effettuate da WikiExtractor[37] sono di pulizia del file dal linguaggio di markup[38] e di estrazione dell'output[39].

L'operazione di pulizia è effettuata mediante una funzione che elimina la marcatura di MediaWiki per ottenere del testo pulito. La funzione agisce eliminando i tag presenti all'interno del testo – tranne quelli presenti all'interno dei link, che vengono invece conservati – e restituendo una lista di paragrafi senza le intestazioni – che sono presenti e necessarie all'interno della pagina web, ma non hanno una reale funzione nel testo –. La variabile estratta come risultato è una lista di stringhe ripulite dalla marcatura MediaWiki.

A seguire un esempio di testo che presenta marcatura MediaWiki[40].

```
== Introduzione ==
La linguistica ha come scopo comprendere e definire le caratteristiche del linguaggio verbale umano (la facoltà mentale dell'uomo di comunicare attraverso una [[Lingua (linguistica)|lingua]]) attraverso l'analisi delle lingue del mondo: un linguista indaga e descrive quindi le [[Struttura (semiotica)|strutture]] delle lingue per capire come sono quest'ultime e cerca di spiegare perché queste sono come sono (e perché non sono in altro modo).<ref name= Dryer2008/>
```

[34] <https://github.com/attardi/wikiextractor>.
[35] <https://github.com/attardi/wikiextractor/wiki>.
[36] <http://dumps.wikimedia.org/itwiki/latest/itwiki-latest-pages-articles.xml.bz2>.
[37] <https://github.com/attardi/wikiextractor/blob/master/wikiextractor/WikiExtractor.py>.
[38] <https://github.com/attardi/wikiextractor/blob/master/wikiextractor/clean.py>.
[39] <https://github.com/attardi/wikiextractor/blob/master/wikiextractor/extract.py>.
[40] <https://it.wikipedia.org/w/index.php?title=Linguistica&action=edit>.

Lo stesso testo ripulito dalla marcatura e contenuto nella variabile.

```
Introduzione.
La linguistica ha come scopo comprendere e definire le caratteristiche del linguaggio verbale umano (la facoltà mentale dell'uomo di comunicare attraverso una lingua) attraverso l'analisi delle lingue del mondo: un linguista indaga e descrive quindi le strutture delle lingue per capire come sono quest'ultime e cerca di spiegare perché queste sono come sono (e perché non sono in altro modo).
```

Vengono successivamente applicate una serie di funzioni che si occupano della conversione delle rimanenti strutture non testuali – come, ad esempio, intestazioni, elenchi, tabelle – in testo, da cui successivamente si eliminano tutti i caratteri non necessari – quali trattini bassi, spazi bianchi, virgolette –. Dunque, il testo viene normalizzato effettuando le necessarie sostituzioni, ma preservando tutti i dati testuali rilevanti all'interno dell'articolo.

L'output testuale viene estratto e memorizzato in una serie di file di dimensioni simili, contenuti in una determinata directory.

```
drwxrwxr-x 34 digisi digisi 4,0K mag 28 19:19 .
drwxr-xr-x  3 digisi digisi 4,0K mag 28 16:24 ..
drwxrwxr-x  2 digisi digisi 4,0K mag 28 16:30 AA
drwxrwxr-x  2 digisi digisi 4,0K mag 28 16:34 AB
drwxrwxr-x  2 digisi digisi 4,0K mag 28 16:38 AC …
```

Alla fine del processo vengono notificate una serie di informazioni riguardanti l'esito dell'estrazione, come, ad esempio, il totale di pagine estratte, il numero di articoli e il tempo di estrazione.

```
INFO: Finished 1-process extraction of 1694459 articles in 10807.5s (156.8 art/s)
INFO: total of page: 2574949, total of articl page: 1694459; total of used articl page: 1694459
```

I file estratti si trovano all'interno di cartelle create automaticamente in fase di estrazione. Ogni file contiene una serie di articoli di Wikipedia e ognuno di essi è rappresentato da un elemento XML nel formato `<doc>`. L'elemento `<doc>` contiene esclusivamente testo e ha i seguenti attributi:

- `id`, che identifica il documento tramite un numero di serie univoco.
- `url`, che fornisce l'indirizzo URL della pagina Wikipedia originale da cui è stato estratto il testo.
- `title`, il titolo della pagina Wikipedia originale da cui è stato estratto il testo.

Un esempio di elemento XML rappresentante un articolo è il seguente:

```
<doc id="2517" url="https://it.wikipedia.org/wiki?curid=2517" title="Linguistica">
```

L'elemento `<doc>` contiene al suo interno il titolo che identifica l'articolo, posizionato come prima frase, e l'articolo estratto nella sua interezza.
Di seguito un esempio di articolo finale presente nei file di output.

```
<doc id="2517" url="https://it.wikipedia.org/wiki?curid=2517" title="Linguistica">
Linguistica


La linguistica è lo studio scientifico del linguaggio verbale umano e delle sue strutture. Essa include lo studio della fonetica, della grammatica, del lessico, della morfologia, della sintassi e della testualità. È una disciplina scientifica, in quanto si basa su approcci empirici e oggettivi. Un linguista è una persona specializzata in linguistica.

…

</doc>
```

2.2.1 ExtractPage

È eventualmente possibile, inoltre, effettuare l'estrazione di una singola pagina da un dump di Wikipedia. In questo caso si procede invocando `extractPage`[41] con il nome del dump come argomento del comando.

Se lo si desidera può inoltre essere aggiunto al comando, come argomento opzionale, l'id dell'articolo che si vuole estrarre.

```
extractPage [--id ID] <Wikipedia dump file>
```

Viene quindi avviata l'elaborazione del dump tramite `extractPage.py`[42], che si articola in due funzioni. La prima funzione seleziona il file di input e gli argomenti inseriti nel comando iniziale e li passa come parametri alla seconda funzione, che compie l'effettiva estrazione dell'articolo selezionato e lo stampa sul terminale. Nel caso in cui non fosse stato inserito alcun id come argomento al comando, l'id viene impostato di default come 1.

La funzione che si occupa di processare i dati in input ed estrarre la pagina selezionata in base all'id stampa un output che comprende il titolo dell'articolo selezionato e tutto il testo dell'articolo in questione. Essendo ExtractPage uno script di sola estrazione, l'output non viene memorizzato in alcun file e alla fine del processo il risultato sarà visualizzabile esclusivamente sul terminale.

È possibile stampare il risultato in un file di output aggiungendo al comando precedentemente indicato il comando di output diretto che indichi nome ed estensione del file che si vuole creare o aggiornare, contenente il risultato.

```
extractPage [--id ID] <Wikipedia dump file> > output.txt
```

ExtractPage non opera alcun procedimento di pulizia del testo, di conseguenza l'output ottenuto è composto dal testo nella formattazione presente all'interno del dump. Oltre al testo sarà quindi ancora presente tutto il linguaggio di marcatura MediaWiki utilizzato per la formattazione dell'articolo online.

[41] <https://github.com/attardi/wikiextractor>.
[42] <https://github.com/attardi/wikiextractor/blob/master/wikiextractor/extractPage.py>.

Un esempio di output è il seguente:

```
  <page>
    <title>Linguistica</title>
    <ns>0</ns>
    <id>2517</id>
    <revision>
      <parentid>120657392</parentid>
      <timestamp>2021-05-22T09:21:06Z</timestamp>
      <contributor>
        <username>Egidio24</username>
      </contributor>
      <minor />
      <comment>[[:en:WP:CLEANER|WPCleaner]] v2.04 - Fixed using [[WP:CW]]
(Wikilink uguali alla propria descrizione)</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text bytes="18733" xml:space="preserve">[[File:Languages world
map.svg|thumb|right|Mappa delle famiglie linguistiche nel mondo.]]
La '''linguistica''' è lo studio scientifico del [[linguaggio]] verbale umano e
delle sue strutture&lt;ref name=&quot;GenettiDef&quot;
…

{{Discipline umanistiche}}
{{Scienze sociali}}

{{Controllo di autorità}}
{{portale|linguistica}}

[[Categoria:Linguistica| ]]</text>
      <sha1>3r8h9xr3ggro99ivhe2uydmm5d1yktq</sha1>
    </revision>
  </page>
```

2.3 Sentence-Splitting e POS-Tagging del testo

Per permettere e agevolare la ricerca delle informazioni testuali, in particolare riguardanti le singole frasi o parole all'interno del corpus, è consigliabile elaborare ulteriormente l'estratto testuale ottenuto, suddividendolo in frasi separate (*sentence-splitting*) e, successivamente, provvedendo alla lemmatizzazione e all'etichettatura in parti del discorso (*POS-tagging*) di ogni parola. L'etichettatura in parti del discorso, in questo caso, permette una ricerca approfondita all'interno del testo, in quanto dà la possibilità di ricercare e selezionare le parole sia per la parte del discorso a cui appartengono, sia per il genere e numero da cui sono connotate.

Ai fini di effettuare il sentence-splitting e il POS-tagging sul testo ottenuto dall'estrazione, sono stati utilizzati ulteriori script in Python, fornendo come input i file di testo precedentemente estratti e ripuliti dal markup con WikiExtractor. Gli strumenti utilizzati per queste operazioni fanno parte della suite di strumenti per l'analisi del testo della pipeline Tanl (Natural Language Text Analytics)[43]. La pipeline Tanl ha un approccio *data driven* – basato sui dati – ciò significa che ogni fase dell'elaborazione estrae i dati dalla fase precedente e li modifica perché possano essere utilizzati per la fase successiva.

2.3.1 Sentence-Splitting

Per effettuare l'operazione di sentence-splitting viene usato `Splitta.py`, uno script che, invocato sul testo estratto con WikiExtractor, opera una suddivisione in frasi e successivamente in *token*.

Lo script applica una lista di espressioni regolari in sequenza e la stringa risultante viene a sua volta suddivisa secondo gli spazi bianchi che ha al suo interno. Le espressioni regolari, in particolare, uniformano le frasi, separano ogni tipo di punteggiatura dalle parole, considerano i doppi trattini come un unico token e separano la virgola solo se seguita da uno spazio bianco (questo per lasciare intatti, ad esempio, i numeri decimali).

Un esempio di output di sentence-splitting.

```
La
linguistica
è
lo
studio
```

[43] Giuseppe Attardi, Stefano Dei Rossi e Maria Simi, *The Tanl Pipeline*.

```
scientifico
del
linguaggio
verbale
umano
e
delle
sue
strutture
.
```

2.3.2 POS-tagging

La funzione per il POS-tagging approfondisce l'analisi sui token ottenuti tramite il sentence-splitter, annotandoli con informazioni riguardanti le categorie morfo-sintattiche a cui appartengono.

Sul testo diviso in token viene invocato, prima di tutto, un lemmatizzatore (*lemmatizer*), che si occupa di derivare il lemma di ogni token preso in analisi. Successivamente un POS-tagger (*tagger*) procede taggando il token con tutti i tag necessari in base alla parte del discorso. A causa delle dimensioni di Wikipedia il processo richiede un numero elevato di ore per essere concluso, ma procedendo automaticamente non necessita di assistenza.

Avere un testo taggato in base alle parti del discorso può risultare fondamentale in un tipo di ricerca che prevede, come in questo caso, l'analisi di una determinata tipologia di sostantivi, o dell'uso di determinati articoli. A seguire un esempio di output di POS-tagging.

```
La              il              R     RD    num=s|gen=f
linguistica     linguistica     S     S     num=s|gen=f
è               essere          V     V     num=s|per=3|mod=i|ten=p
lo              il              R     RD    num=s|gen=m
studio          studio          S     S     num=s|gen=m
scientifico     scientifico     A     A     num=s|gen=m
del             di              E     EA    num=s|gen=m
linguaggio      linguaggio      S     S     num=s|gen=m
verbale         verbale         A     A     num=s|gen=n
umano           umano           A     A     num=s|gen=m
e               e               C     CC    _
delle           di              E     EA    num=p|gen=f
sue             suo             A     AP    num=p|gen=f
strutture       struttura       S     S     num=p|gen=f
.               .               F     FS    _
```

Come mostrato dall'esempio, a ogni token viene associato:

- il lemma corrispondente;
- la parte del discorso a cui appartiene, che viene identificata tramite l'uso di due tipi di tag. Il primo tag è anche definito *coarse* (a grana grossa) e indica la parte del discorso generica di cui il token fa parte, mentre il secondo tag serve, generalmente, per specificarne la tipologia (come accade con `La`, nella prima linea dell'output, in cui il primo tag è `R`, che indica la classe degli articoli, e il secondo è `RD`, che specifica che l'articolo in questione è determinativo). Nel caso in cui il token, invece, indichi una parte del discorso che non prevede specificazione, verrà ripetuto due volte il primo tag (come si può notare con il token `linguistica`, identificato tramite due tag `S`, indicanti la classe dei sostantivi);
- le caratteristiche morfologiche del token preso in esame, opportunamente codificate. Possiamo dunque trovare informazioni come il numero e il genere, se la parte del discorso in questione li prevede, o l'indicazione di modo e tempo per i verbi `V`.

Di seguito verranno elencate alcune delle categorie generiche e specifiche in cui i token possono essere distinti.

- Sostantivo: `S` di cui:
  - `SP`: sostantivo proprio
- Aggettivo: `A`
- Articolo: `R` di cui:
  - `RD`: articolo determinativo
  - `RI`: articolo indeterminativo
- Pronome: `P` di cui:
  - `PE`: pronome personale
  - `PP`: pronome possessivo
  - `PD`: pronome dimostrativo
  - `PR`: pronome relativo
  - `PI`: pronome indefinito
  - `PC`: pronome riflessivo

- Verbo: V di cui:
  - VA: verbo ausiliare
  - VM: verbo modale
- Avverbio: B
- Preposizione: E di cui:
  - EA: preposizioni articolate
- Congiunzione: C di cui:
  - CC: congiunzione coordinante
  - CS: congiunzione semplice
- Segni di punteggiatura: F di cui:
  - FS: .
  - FF: , …
  - FB: “ ” [ ] « »
  - FC: :

3 I femminili professionali su Wikipedia in lingua italiana

3.1 I corpora

Ognuno dei corpora su cui verrà effettuata l'analisi è stato estratto ed elaborato tramite le medesime procedure, negli anni 2021, 2019, 2015 e 2013. In particolare, i corpora degli anni 2021, 2015 e 2013 sono composti da tutti gli articoli presenti su Wikipedia in lingua italiana al momento della creazione del dump, nella loro versione esclusivamente testuale, nella versione suddivisa in frasi e nella versione POS-taggata. Il corpus del 2019, invece, è composto esclusivamente dalla versione suddivisa in frasi e da quella POS-taggata.

Di seguito sono riportate le date delle estrazioni e gli autori.

- Il corpus 2021 è stato estratto dalla professoressa Maria Simi, docente all'Università di Pisa, in data 25 maggio 2021 ed elaborato con le procedure di sentence-splitting e POS-tagging rispettivamente il 26 e il 27 maggio 2021.
- Il corpus 2019 è stato estratto ed elaborato dal professor Giuseppe Attardi in data 14 aprile 2019.
- Il corpus 2015 è stato estratto da Giuseppe Attardi in data 23 aprile 2015, elaborato con la procedura di sentence-splitting il 16 dicembre 2015 e con quella di POS-tagging il 14 aprile 2019.
- Il corpus 2013 è stato estratto ed elaborato da Giuseppe Attardi in data 5 aprile 2013.

Durante l'estrazione tramite WikiExtractor, i file ottenuti vengono collocati all'interno di una cartella chiamata *text*, suddivisa a sua volta in sottocartelle denominate come *AA*, *AB*, *AC*, ecc. Ognuna di queste sottocartelle contiene cento file – ad eccezione dell'ultima che, tendenzialmente ne ha meno –, nominati come *wiki_00*, *wiki_01*, *wiki_02*, ecc.

I file ottenuti dal sentence-splitting e dal POS-tagging sono inseriti in cartelle denominate come *sentences* e *tagged* che contengono lo stesso tipo di sottocartelle e file testuali della cartella *text*.

3.2 Oggetto dell'analisi

Come anticipato nel capitolo 1.3, i corpora verranno analizzati con il fine di valutare il cambiamento effettivo della lingua italiana di Wikipedia per quanto riguarda la femminilizzazione dei sostantivi di professioni di prestigio e titoli onorifici.

I sostantivi femminili presi in analisi sono stati selezionati dalle ricerche effettuate da Giuseppe Zarra precedentemente citate[44] e, in particolare, saranno circoscritti ai titoli professionali di ambito politico. La scelta di limitare l'analisi ai titoli politici deriva, principalmente, da una questione legata alla probabilità che su Wikipedia siano state dedicate pagine a personalità influenti dell'ambito politico piuttosto che a donne effettivamente legate a professioni di prestigio, ma che non necessariamente hanno la possibilità di trovare spazio in un'enciclopedia.

I sostantivi femminili presi in considerazione sono, in particolare, *assessora*, *deputata*, *ministra* e *sindaca*, titoli politici legati a personalità spesso presenti su Wikipedia in lingua italiana.

La prima parte dell'analisi prevede che di questi sostantivi vengano contate le occorrenze totali, che saranno a loro volta paragonate ai corrispettivi maschili in un rapporto 1:x tra il numero di occorrenze femminili e quelle maschili. La ricerca avverrà sempre nella consapevolezza del fatto che spesso le forme maschili vengono utilizzate con un'accezione di neutralità e, soprattutto, che il numero di detentori di determinati titoli è ancora a maggioranza maschile.

La seconda parte dell'analisi, invece, avrà il fine di stabilire le tendenze linguistiche degli articoli di Wikipedia in lingua italiana riguardo l'utilizzo dei sostantivi femminili di professione. In particolare, la volontà è quella di comprendere se vi sia una propensione all'utilizzo dei maschili con accezione neutra piuttosto che all'utilizzo del femminile, e, nel caso della presenza del femminile, in che declinazione esso sia maggiormente utilizzato.

Le forme prese in analisi per ogni sostantivo femminile sono le seguenti:

- assessora: *l'assessora*, *l'assessore* seguito dal nome o dal cognome di un'assessora, *donna assessore*, *assessore donna*;
- deputata: *la deputata*, *il deputato* seguito dal nome o dal cognome di una deputata, *donna deputato*, *deputato donna*;
- ministra: *la ministra*, *il ministro* seguito dal nome o dal cognome di una ministra, *la ministro*, *ministressa*, *donna ministro*, *ministro donna*;

[44] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, capitolo 2.4.

- sindaca: *la sindaca, il sindaco* seguito dal nome o dal cognome di una sindaca, *la sindaco, sindachessa, donna sindaco, sindaco donna*.

3.3 Modalità di analisi

Le seguenti analisi sono state effettuate utilizzando Ubuntu mediante l'uso di comandi UNIX, e in particolare tramite l'uso dei comandi *grep* (**g**eneral **r**egular **e**xpression **p**rint) e *wc* (**w**ord **c**ount).

Il comando `grep`, nello specifico, serve per la ricerca all'interno di file testuali di uno o più modelli specificati tramite espressioni regolari o stringhe letterali, e in questo contesto è stato utilizzato per la ricerca delle occorrenze di una o più parole in una serie di file. Se una linea contenuta in uno dei file analizzati soddisfa almeno uno dei modelli espressi nel comando, si ottiene un risultato. Il risultato finale è composto dall'insieme delle linee in cui è stata riscontrata una corrispondenza.

La sintassi del comando `grep` prevede l'aggiunta di parametri successivamente al comando. I parametri riguardano le opzioni facoltative applicate al comando, utilizzate per specificare il criterio di ricerca, i modelli da ricercare all'interno dei file e, opzionalmente, i file a cui applicare la ricerca.

La sintassi generale è quindi riassumibile nella seguente formula:

```
grep [opzioni] [modelli] [file]
```

Le opzioni aggiuntive possono dunque essere molteplici o non esserci affatto, così come i file. In particolare, specificando molteplici parametri file, il risultato nell'output prevederà per ogni linea in cui è stata trovata una corrispondenza, anche l'indicazione del nome del file e del numero della linea in cui si trova la corrispondenza al modello. Quando il parametro file non viene specificato, la ricerca avviene su tutti i file testuali presenti nella directory in cui ci si trova in quel momento.

Il comportamento predefinito di `grep` prevede che i modelli utilizzati per la ricerca siano delle espressioni regolari e non delle stringhe letterali, ma è possibile comunque effettuare una ricerca di queste ultime tramite apposite opzioni, come ad esempio l'opzione `-F` [45].

[45] Capitolo 3.3.1

Per completare il comando definitivo per la presente analisi è stato necessario aggiungere alle opzioni specifiche di `grep` il comando `wc`, un comando dei sistemi operativi UNIX che produce come output un conteggio delle linee, parole o byte che costituiscono uno o più file di testo[46].

La sintassi generale di `wc` è la seguente:

```
wc [opzioni] [file]
```

Come per `grep`, anche per `wc`, file è un parametro facoltativo che può essere presente più volte o nessuna, che indica il nome o i nomi dei file su cui effettuare il conteggio. Come per `grep`, quando il parametro file non è specificato la ricerca avviene su tutti i file presenti nella directory in cui ci si trova in quel momento.

Di seguito sono illustrate le tre opzioni principali di `wc`.

- `-c`: effettua il conteggio dei byte;
- `-l`: effettua il conteggio delle linee – in particolare dei caratteri di ritorno a capo –;
- `-w`: effettua il conteggio delle parole.

Con la concatenazione dei comandi `grep` e `wc`, dunque, è possibile estrarre tutte le righe che contengono una determinata espressione regolare – oppure stringa letterale –, in uno o più file, e successivamente provvedere a contare le righe del risultato appena estratto, ottenendo come risultato un numero. È necessario notare, però, che solo con la concatenazione di questi due comandi non è possibile identificare i casi in cui siano presenti più occorrenze del modello nella stessa riga. Per questo servono delle specifiche opzioni del comando `grep`, come ad esempio l'opzione `-o`.

[46] <https://it.wikipedia.org/wiki/Wc_(Unix)>.

3.3.1 Opzioni del comando grep

Il comando `grep` prevede molteplici opzioni, permettendo così una resa sempre più specifica della ricerca. Verranno di seguito illustrati quelli più comuni e, in particolar modo, quelli utilizzati nella presente analisi[47].

- `-A numero`, anche `--after-context=numero`: stampa, per ogni riga di cui è stata trovata una corrispondenza, il numero specificato di linee che la seguono. È utilizzato per visualizzare il contesto di un risultato.
- `-B numero`, anche `--before-context=numero`: stampa, per ogni riga di cui è stata trovata una corrispondenza, il numero specificato di linee che la precedono. È utilizzato per visualizzare il contesto di un risultato.
- `-F`, anche `--fixed-strings`: interpreta i modelli indicati come stringhe di caratteri che vanno ricercate in maniera letterale, invece che come espressioni regolari – come avverrebbe spontaneamente con l'uso del comando `grep –`;
- `-i`, anche `--ignore-case`: ignora le distinzioni tra lettere maiuscole e minuscole nei modelli e nei dati di input, in modo che i caratteri che differiscono esclusivamente per le maiuscole corrispondano tra loro;
- `-n`, anche `--line-number`: antepone a ogni riga di output il numero della riga del file a cui è stata trovata la corrispondenza.
- `-o`, anche `--only-matching`: conta tutte le occorrenze. In particolare, questa opzione stampa esclusivamente le parti delle righe che corrispondono al modello e permette la stampa di ciascuna ricorrenza su una linea di output differente. Questa opzione è necessaria per le analisi che prevedono il conto totale delle volte in cui il modello viene ripetuto, in quanto l'uso del solo comando `grep` ricerca le linee in cui il modello c'è almeno una volta, di conseguenza se su una linea il modello viene ripetuto più di una volta viene comunque contato come un'unica occorrenza.

  Nel caso della presente analisi è stato necessario utilizzare l'opzione `-o` piuttosto che l'opzione `-c` (`--count`), poiché quest'ultima sopprime il normale output di `grep` per stampare il conteggio delle righe che contengono una corrispondenza, senza considerare dunque il reale numero di occorrenze.

[47] <https://www.gnu.org/software/grep/manual/grep.html>.

- `-r`, anche `--recursive`: opera ricorsivamente su ogni file contenuto nella directory indicata. Se non viene indicato nessun argomento file, il comando viene applicato ricorsivamente alla directory in cui ci si trova. È un comando necessario se si opera, come in questo caso, su una grande mole di file contenuti in diverse sotto cartelle in una stessa directory.
- `-w`, anche `–word-regexp`: seleziona solo le righe che contengono corrispondenze al modello che formano parole intere. La stringa corrispondente al modello deve essere all'inizio di una riga o preceduta esclusivamente da un carattere che non sia costitutivo di parola, oppure ancora deve trovarsi alla fine di una riga o seguito da un carattere che non sia costitutivo di parola. I caratteri costitutivi di parola sono, ad esempio, lettere o cifre.

3.4 Analisi del rapporto tra le occorrenze femminili e maschili dei titoli onorifici

Appurato che la lingua italiana è in grado di offrire chiaramente la possibilità di esplicitare la presenza delle donne all'interno della società e in ruoli istituzionali a cui in passato non sarebbero potute giungere, è interessante vedere se e come, effettivamente, queste soluzioni vengano accolte.

Dalle analisi citate in precedenza è evidente come la soluzione generalmente più diffusa e che tendenzialmente prevale sia quella dell'utilizzo del maschile in riferimento indistintamente a uomini e donne. In particolare, nella sfera degli incarichi politici, il maschile neutro viene spesso usato con l'intenzione di riferirsi alla carica, piuttosto che alla persona fisica che la detiene, e ciò denota quanto effettivamente ai sostantivi maschili sia spesso intimamente legato anche un modello culturale di prestigio.

È importante dunque valutare l'effettiva portata del fenomeno di inglobamento dei sostantivi di genere femminile nel corrispettivo maschile, e, soprattutto, considerare quanto effettivamente questo tipo di utilizzo del maschile con accezione neutrale causi l'oscuramento della già scarsa presenza femminile in determinati settori della società.

Le analisi di seguito riportate, essendo basate su corpora appartenenti ad anni differenti, sono in grado di dare uno spaccato fedele dell'evoluzione del linguaggio in termini di femminilizzazione dei sostantivi di titoli onorifici e professioni di prestigio.

La conta delle occorrenze è stata effettuata sui corpora *sentences*, elaborati tramite sentence-splitting, i quali prevedono la presenza di un singolo token per riga, ed effettuando una selezione e conta delle righe in cui la determinata occorrenza è presente. Il comando utilizzato è il seguente.

```
grep -Firw <sostantivo> | wc -l
```

In questo modo è stato possibile analizzare ricorsivamente tutte le sottocartelle di *sentences* contenenti del testo e contare direttamente le occorrenze del sostantivo cercato, senza la necessità di creare un'espressione regolare. In questo caso è possibile utilizzare l'opzione `-l` per il comando `wc` poiché, essendo i corpora *sentences* formati da un token per riga, è sufficiente contare le righe in cui compare un'occorrenza per ottenere il risultato totale.

In una situazione come la presente, in cui le stringhe da cercare sono tendenzialmente sempre letterali, l'utilizzo di `-F` per la ricerca di pattern letterali rende la scrittura del modello molto più immediata. Ma, come accennato in precedenza, il comando `grep` prevede, in realtà, che i modelli ad esso applicati siano espressioni regolari. Di seguito, dunque, è presente un comando alternativo per effettuare lo stesso tipo di ricerca tramite la scrittura di un'espressione regolare. In particolare, la seguente espressione regolare indirizza la ricerca delle parole singole *deputata* e *Deputata.*

```
grep -r '\b[dD]eputata\b' | wc -l
```

Chiaramente, sfruttando a pieno la natura del comando `grep` e utilizzandolo con un modello che sfrutti le espressioni regolari, si ha grande risparmio nella dichiarazioni delle opzioni, anche se ciò comporta una maggior attenzione nella specifica del modello. Per snellire ulteriormente il modello sarebbe inoltre possibile sostituire alla duplice scelta tra l'iniziale minuscola e maiuscola, l'opzione `-i`, che ignora le differenze tra maiuscole e minuscole.

Un'alternativa all'utilizzo del comando `grep`, per questo tipo di analisi basate sul conteggio delle occorrenze, potrebbe essere uno script in Python3 come il seguente, che si occupi di analizzare un'espressione regolare e contarne le occorrenze in tutti i file all'interno di un'intera directory.

```
#!/usr/bin/env python3

import re
from pathlib import Path

pattern = r"\b[sS]ostantivo\b"

def get_occurrences(path):
    if path.is_file():
        return len(re.findall(pattern, path.read_text()))
    else:
        return 0

total = sum(map(get_occurrences, Path("./path/to/folder").rglob("*")))

if total:
    print(f"Numero di occorrenze trovate: {total}")
else:
    print("Nessuna occorrenza trovata.")
```

Questo script, in particolare, utilizza il modulo `pathlib`[48], che offre la possibilità di gestire e manipolare, tramite la sintassi appropriata, delle classi che rappresentano i percorsi di *file system*; e il modulo `re`,[49] che fornisce le operazioni di corrispondenza delle espressioni regolari utilizzate.

Successivamente viene inizializzata una variabile che esplicita il pattern dell'espressione regolare da ricercare, e viene definita una funzione che si occupa di trovare le tutte le occorrenze del pattern all'interno del file in lettura. Tramite l'esplicitazione di uno specifico *path* si applica la funzione ricorsivamente a tutte le cartelle di una determinata directory e si somma il numero delle occorrenze trovate. Infine, se sono state trovate delle occorrenze, viene stampato il numero delle occorrenze totali, altrimenti si viene notificati del fatto che non sia stata trovata alcuna occorrenza.

Questi tre metodi sono perfettamente interscambiabili e operano sui file testuali allo stesso modo, ottenendo esattamente gli stessi risultati.

Mediante queste procedure applicate ai corpora *sentences*, è stato ricavato e registrato il numero di occorrenze delle coppie di sostantivi designanti cariche pubbliche precedentemente selezionati, ovvero *assessora–assessore*, *deputata–deputato*, *ministra–ministro*, *sindaca–sindaco*.

Come è possibile notare, però, *deputata*, *deputato*, *ministra* e *ministro*, oltre a svolgere la funzione linguistica di sostantivi, sono anche dei participi passati singolari particolarmente utilizzati nella lingua di Wikipedia, in particolare *deputata* e *deputato* hanno un'incidenza piuttosto alta. È stato dunque necessario provvedere alla disambiguazione di queste forme, per poter prendere in considerazione esclusivamente i sostantivi di titoli politici e rendere l'analisi più precisa possibile.

Le operazioni di disambiguazione sono state effettuate sui corpora *tagged*, che, provvedendo all'annotazione della parte del discorso di ogni singolo token, rendono possibile riconoscere quando le due forme, seppur identiche, sono sostantivi – identificati tramite `S` – piuttosto che verbi – `V` –. In particolare, è stata effettuata una selezione dei corpora *tagged*, estraendo tutte le linee in cui fosse presente l'occorrenza cercata, e aggiungendola a un file `.txt` separato. Per questa operazione il comando utilizzato è stato il seguente.

```
grep -Fhirw <sostantivo> > file.txt
```

Come è possibile notare, al comando è stata aggiunta l'opzione `-h`, o anche `-no-filename`, necessaria per evitare di stampare i nomi dei file nell'output finale[50].

[48] <https://docs.python.org/3/library/pathlib.html>.
[49] <https://docs.python.org/3/library/re.html>.
[50] <https://www.gnu.org/software/grep/manual/grep.html>.

Questa opzione è strettamente necessaria, poiché i nomi dei file, o delle cartelle in cui sono contenuti, potrebbero utilizzare a loro volta le lettere identificative delle parti del discorso, portando a un'imprecisione del risultato.

Ridotti dunque i corpora alle sole occorrenze dello specifico sostantivo da analizzare, è stata effettuata su di essi una ricerca volta a contare esclusivamente le occorrenze della specifica sigla identificativa della parte del discorso interessata, in questo caso `V`. In questo modo è stato possibile contare, tra tutte le occorrenze dello specifico sostantivo, quante di queste sono taggate come verbi ed escluderle dunque dal conteggio finale. Il comando utilizzato è il seguente.

```
grep -Frw "V" file.txt | wc -l
```

In questo caso è necessario l'utilizzo dell'opzione `-l` per il comando `wc` perché, come specificato nel capitolo 2.3.2, i tag delle parti del discorso sono ripetute due volte, per permettere un'analisi più specifica del singolo token, e in questo modo viene contata una sola occorrenza per ogni riga, anche se dovesse contenerne più d'una.

Successivamente, dunque, è stato sottratto il numero di occorrenze delle forme verbali dalle occorrenze totali precedentemente ricavate dai corpora *sentences*, ottenendo il numero totale delle sole occorrenze del sostantivo.

Durante l'ultima parte dell'analisi appena descritta, quella effettuata sui corpora *tagged*, è stata notata una particolarità che potrebbe essere interessante riportare. È stata infatti riscontrata una netta differenza tra i tag delle versioni femminili dei sostantivi e participi passati – *deputata* e *ministra* – e quelle maschili – *deputato* e *ministro* –.

In particolare, le forme *deputata* e *ministra*, scritte con la lettera iniziale minuscola, risultano essere state identificate tutte come participi passati e connotate dunque dalla lettera `V` anche quando, giudicandone il contesto all'interno del corpus, era facilmente intuibile si trattasse di un sostantivo. Le forme *Deputata* e *Ministra*, invece, erano sempre taggate come sostantivi e connotate dunque dalla lettera `S`. Si è potuto dunque procedere alla disambiguazione solo tramite un'analisi manuale delle occorrenze femminili, in modo da ottenere un risultato più accurato possibile.

Lo stesso, ovviamente, non è accaduto con i corrispettivi maschili, correttamente identificati come sostantivi o participi passati in base al contesto, e dunque disambiguati tramite la procedura precedentemente illustrata.

3.4.1 Risultati

A seguito delle analisi illustrate, effettuate sui corpora *sentences* e *tagged* dei quattro diversi anni, è stato dunque registrato il numero totale di occorrenze delle quattro coppie di sostantivi designanti cariche pubbliche precedentemente elencati, *assessora–assessore*, *deputata–deputato*, *ministra–ministro*, *sindaca–sindaco*. Tramite i numeri ricavati è stato stimato un rapporto tra le due incidenze con una proporzione di 1:x tra il numero di occorrenze del sostantivo femminile e il numero di occorrenze del corrispettivo maschile.

A seguire i risultati.

Corpus	*Assessora*	*Assessore*	Rapporto[51]
2021	26	5 136	197,53
2019	12	3 126	260,50
2015	4	3 756	939,00
2013	4	2 593	648,25
	Deputata	*Deputato*	
2021	1 490	17 581	11,79
2019	752	10 099	13,42
2015	782	13 614	17,40
2013	424	8 690	20,49
	Ministra	*Ministro*	
2021	445	63 012	141,60
2019	196	38 533	196,59
2015	144	47 344	328,77
2013	75	30 921	412,28
	Sindaca	*Sindaco*	
2021	176	28 444	161,61
2019	65	18 302	281,56
2015	8	21 883	2735,37
2013	7	13 409	11915,57

[51] Numero di occorrenze maschili per una sola occorrenza femminile.

I risultati ottenuti evidenziano il divario tra l'utilizzo dei termini maschili e quelli femminili e rendono chiara la preferenza del maschile con accezione neutra in ambito di Wikipedia. Notiamo infatti come nessuna delle coppie di sostantivi mostri un rapporto vicino alla parità. Nonostante ciò, paragonando i risultati delle stesse occorrenze tra i diversi anni, è possibile constatare un complessivo innalzamento delle occorrenze totali dei femminili e un'ingente riduzione del rapporto tra le due occorrenze.

È chiaro, inoltre, il radicamento di alcuni femminili come *deputata* e *ministra*, utilizzati con una netta superiorità rispetto a femminili come *assessora* e *sindaca*, ancora minimamente impiegati e con occorrenze estremamente esigue in minimo aumento nel corso degli anni.

In particolare, riguardo la scarsità di utilizzo del sostantivo *sindaca*, è opportuno sottolineare come su Wikipedia in lingua italiana sia stato affrontato un lungo dibattito – durato circa tre anni e avuto luogo tra il 2016 e il 2019 –, dall'eloquente titolo *Secondo l'Accademia della Crusca bisogna dire "sindaca"*[52], riguardo la necessità o meno di una femminilizzazione delle cariche politiche e dei titoli onorifici negli articoli presenti su Wikipedia. In particolar modo la discussione si è concentrata, come da titolo, sulla carica di sindaco e raccoglie opinioni sull'argomento decisamente contrastanti e differenziate. Nonostante la discussone sia proseguita per diversi anni, i cambiamenti effettivi sul linguaggio di genere utilizzato, per lo meno nell'ambito dei titoli onorifici, sono stati piuttosto ridotti – come è possibile notare dalla tabella dell'analisi delle occorrenze nei vari anni –, a causa della grande riluttanza degli utenti nei confronti di queste forme, percepite come nuove e non abbastanza utilizzate nel linguaggio comune da poter entrare a far parte di quello di Wikipedia.

La discussione sul termine *sindaca*, in particolare, si accoda a un'ulteriore discussione ancora precedente, iniziata nel 2014, riguardo le cosiddette *Femminilizzazioni forzate*[53] riscontrate da alcuni utenti e fortemente condannate. All'interno di questa discussione è possibile riscontrare un atteggiamento di forte diffidenza nei confronti dei femminili professionali, la cui possibilità d'uso viene percepita principalmente come una scelta personale, più che come l'applicazione della corretta forma della lingua. La forte presa di posizione contro la femminilizzazione di terminologie prettamente maschili trova, addirittura, secondo alcuni utenti, motivazione nella volontà di non fare discriminazione di genere tra i detentori e le detentrici delle cariche, in quanto è largamente ritenuto che i sostantivi legati alle cariche abbiano di per sé un'accezione "neutrale" nonostante la forma maschile e, soprattutto, nonostante nella lingua italiana non esista il genere neutro.

[52] <https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Secondo_l%27Accademia_della_Crusca_bisogna_dire_%22Sindaca%22>.
[53] <https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Femminilizzazioni_forzate>

Dunque, come già precedentemente riportato, la discussione riguardo il genere dei sostantivi professionali, che sembra poggiare su ragioni di tipo linguistico, in realtà si riconferma essere esclusivamente di tipo culturale[54].

Restando nel merito della discussione citata in precedenza, è interessante notare come il rapporto delle occorrenze di *sindaca* e *sindaco* sia stato, a tutti gli effetti, uno dei più impari tra quelli registrati nei primi due anni. Il cambiamento avvenuto successivamente alla discussione non ha certamente contribuito a una reale parità nell'utilizzo delle forme, ma per lo meno ha portato la frequenza delle occorrenze del sostantivo *sindaca* ad assestarsi a una cadenza simile a quella degli altri titoli onorifici presi in analisi.

Come si può notare, l'unico rapporto che tende a non avere un grandissimo scarto è quello tra *deputata* e *deputato*. Mentre la coppia che negli ultimi anni ha registrato il divario di occorrenze più ampio è sicuramente *assessora – assessore*. È quindi possibile riscontrare una lieve e lenta tendenza generale verso una riduzione del divario tra le forme, tendenzialmente poco incisiva.

Questi risultati sono perfettamente comprensibili se letti alla luce del fatto che Wikipedia si avvale dell'uso di un linguaggio enciclopedico che tende ad essere molto omogeneo e, soprattutto, strettamente dipendente dalle fonti utilizzate. Di conseguenza, dal momento in cui le informazioni su personalità legate ad incarichi politici vengono in larga parte prese dai siti ufficiali dei Ministeri e dei vari organi politici, è possibile che parte della staticità di questo linguaggio dipenda da una mancata presa di posizione da parte degli organi politici stessi. A differenza di altri paesi europei, infatti, il governo italiano non è mai intervenuto con atti normativi sul trattamento linguistico dei titoli professionali e politici delle donne.

Bisogna però sottolineare la presenza delle linee guida del Parlamento europeo[55], redatte nel 2018 con il fine di incentivare la neutralità di genere nel linguaggio. Secondo il Parlamento, infatti, per quanto riguarda le lingue caratterizzate dal genere grammaticale, come l'italiano, è necessario che la neutralità di genere passi attraverso la femminilizzazione dei sostantivi maschili, soprattutto in ambito professionale. Le linee guida specifiche per l'italiano che si trovano in coda al documento, inoltre, raccomandano l'uso di tecniche per la neutralità del linguaggio fortemente ispirate agli elaborati di Alma Sabatini e Cecilia Robustelli. È possibile dunque constatare come, in realtà, delle indicazioni autorevoli in questo campo esistano ma, chiaramente, le abitudini linguistiche e sociali necessitano di un cambiamento organico e spontaneo dei parlanti per essere realmente integrate.

[54] Cecilia Robustelli, *Infermiera sì, ingegnera no?*.
[55] <https://www.europarl.europa.eu/cmsdata/187102/GNL_Guidelines_IT-original.pdf>.

3.5 Analisi delle tendenze linguistiche riguardo la scelta tra diversi femminili

Come già accennato in precedenza, il problema del genere dei sostantivi professionali e dei titoli onorifici riferiti a donne è nato in seguito all'accesso da parte di queste ultime a professioni e incarichi storicamente riservati agli uomini. «Di conseguenza», come scrive Giuseppe Zarra, «il sistema linguistico ha accolto più possibilità per designare tali donne: o il maschile neutro […], o l'articolo femminile e il sostantivo maschile […], o la mozione al femminile»[56].

Risulta quindi piuttosto interessante avere la possibilità di fare una stima della frequenza d'uso di queste diverse tipologie di femminile, per comprendere quale sia effettivamente la scelta più confortevole in un ambito spinoso come la questione dell'utilizzo dei femminili professionali.

Le tipologie di femminili scelti rispecchiano quelle selezionate da Giuseppe Zarra nella sua analisi sull'uso dei femminili di professioni in Italia[57]. Ad esempio, per ministra, sono state selezionate le seguenti possibilità: *la ministra*, *il ministro* seguito dal nome proprio di una donna, *la ministro*, *ministressa*, *donna ministro*, *ministro donna*. Queste forme sono state applicate anche agli altri tre sostantivi femminili presi in analisi.

Ai fini di questa seconda analisi è stato scelto di utilizzare i corpora nella versione *text*, per rendere più agevole la ricerca di nomi composti da articoli e sostantivi. Di conseguenza, per i corpora 2021, 2015 e 2013, di cui effettivamente si possiede l'estrazione del solo testo, è stato usato il comando riportato di seguito, che conta tutte le occorrenze di una determinata stringa letterale ignorando le distinzioni tra lettere maiuscole e minuscole. In questo caso è necessario specificare tra le opzioni di `grep` l'opzione `-o` e in quelle di `wc` l'opzione `-w`, poiché, trattandosi di un file testuale che contiene diverse parole in ogni riga, è importante contare ogni occorrenza della stringa, anche quando si presenta più volte in una sola riga.

```
grep -Fiorw "<stringa letterale>" | wc -w
```

[56] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, p. 78.
[57] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, p. 37.

Un'alternativa utilizzando le espressioni regolari potrebbe essere la seguente, che sfrutta comunque l'opzione `-i` per stabilire un controllo sulle lettere maiuscole e minuscole, semplificando di conseguenza l'espressione regolare.

```
grep -ior '\b<stringa letterale>\b' | wc -w
```

Per quanto riguarda il corpus 2019, invece, possedendo esclusivamente le versioni *sentences* e *tagged*, l'analisi è stata effettuata in maniera piuttosto differente.

Dal corpus *sentences* 2019 è stato estratto un corpus ridotto contenente il sostantivo cardine dell'analisi e la parola precedente o successiva, in base alla necessità. Successivamente, sono state contate tutte le occorrenze che presentavano il pattern ricercato. Ad esempio, per l'analisi di *la deputata* è stato selezionato un corpus ridotto contenente esclusivamente le righe del corpus *sentences* con un'occorrenza di *deputata*, più la riga precedente a quella dell'occorrenza. Successivamente, si è proceduto con la conta del numero di linee che, all'interno del corpus ridotto basato sulle occorrenze di *deputata*, presentavano l'occorrenza *la*, ottenendo così il numero di volte in cui effettivamente *deputata* è preceduto da *la*, e dunque le occorrenze totali della stringa *la deputata*. Questo procedimento è stato necessario in quanto tramite l'utilizzo del comando `grep` non è possibile effettuare la ricerca di un unico pattern suddiviso in più righe. Il comando utilizzato per la selezione della riga con l'occorrenza cercata e delle righe di contesto è il seguente.

```
grep -B 1 -Firw <sostantivo> > file.txt
```

Il comando in questione trova e stampa tutte le righe che contengono un'occorrenza del sostantivo, più la riga precedente. Nel caso in cui si volessero trovare e stampare le righe con una determinata occorrenza e la riga che la segue, si dovrebbe utilizzare l'opzione `-A` seguita dal numero di righe che si desidera stampare.

Necessita inoltre di una specifica anche l'analisi effettuata sulle occorrenze del sostantivo maschile seguito da un nome proprio. In questo caso il controllo sui nomi propri di persona è avvenuto manualmente, in quanto nessuna informazione nel corpus taggato riporta, chiaramente, il genere di appartenenza del proprietario di un nome proprio. Anche in questo caso si è provveduto, anno per anno, alla creazione di un corpus ridotto basato sulle occorrenze della determinata parola da analizzare, stampando dai corpora *sentences* solo le righe specifiche contenenti l'occorrenza del

sostantivo da ricercare, come, ad esempio *deputato*, ognuna di esse associata alla riga precedente e successiva. Il comando utilizzato è il seguente.

```
grep -A 1 -B 1 -Firw <sostantivo> > file.txt
```

Successivamente si è proceduto con l'esclusione di tutte le forme non precedute, in questo caso, dall'articolo *il*, e si è svolta una ricerca manuale su ogni cognome riscontrato nella riga successiva al sostantivo in questione.

Una particolarità riscontrata di frequente è che, a seguito della forma *articolo maschile* + *sostantivo maschile*, nel caso in cui la persona successivamente indicata fosse una donna, veniva indicata con nome e cognome, piuttosto che solo il cognome, come accade per la maggior parte degli uomini citati. Un'abitudine, questa, che risulta particolarmente interessante se si ragiona sulla giustificazione dell'uso del maschile in funzione del mantenimento della neutralità delle cariche pubbliche addotta da molti utenti nella discussione di Wikipedia precedentemente citata[58].

3.5.1 Risultati

A seguire i risultati delle analisi delle forme alternative di femminile dei sostantivi selezionati.

- *Assessora*

	2021	2019	2015	2013
l'assessora	3	1	0	0
l'assessore + NPF[59]	9	5	3	2
–	–	–	–	–
–	–	–	–	–
donna assessore	2	1	2	2
assessore donna	2	1	0	0

Anche in questa seconda analisi viene confermata la scarsa tendenza all'utilizzo della forma *assessore* in qualsiasi declinazione a tendenza femminile.

[58] <https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Femminilizzazioni_forzate>
[59] L'abbreviazione NPF indica il nome o cognome proprio femminile.

È interessante notare come le forme utilizzate sin dal primo corpus preso in analisi comportino prevalentemente l'uso del sostantivo al maschile, o specificato dal modificatore "donna" o tramite il cognome associato. Ad ogni modo, la forma privilegiata resta comunque quella maschile.

- *Deputata*

	2021	2019	2015	2013
la deputata	470	152	218	226
il deputato + NPF	7	7	4	2
–	–	–	–	–
–	–	–	–	–
donna deputato	2	1	0	0
deputato donna	0	1	0	0

Deputata si conferma, dunque, come il femminile preso in analisi con la maggiore diffusione. Questa forma femminile sembra infatti essere quella maggiormente entrata nell'uso, sebbene si debba ricordare comunque l'ultimo rapporto registrato nella precedente analisi tra le occorrenze femminili e quelle maschili di 1:11,79, che implica comunque un largo utilizzo del maschile con accezione neutrale.

È possibile notare inoltre uno scarsissimo utilizzo delle forme dissimmetriche che prevedono l'utilizzo del modificatore "donna" anteposto o posposto.

- *Ministra*

	2021	2019	2015	2013
la ministra	182	32	66	44
il ministro + NPF	72	41	30	16
la ministro	8	6	6	4
ministressa	0	0	0	0
donna ministro	42	12	10	0
ministro donna	36	13	12	0

Per quanto riguarda l'utilizzo delle forme *la ministra* e *il ministro* seguito da nome proprio di donna prevale, piuttosto sorprendentemente, l'utilizzo della forma femminile, nonostante ci sia stato, nel 2019, un tentativo di inversione di tendenza, con la diminuzione delle forme femminili e un superamento del corrispettivo maschile utilizzato con nome femminile.

Si rivela inoltre molto alta e costantemente in crescita la frequenza delle forme che utilizzano il modificatore "donna", sia prima che dopo il sostantivo maschile. In generale, da questi risultati, si può riscontrare una lieve incertezza riguardo il definitivo abbandono delle forme maschili, che restano, appunto, ancora molto presenti.

- *Sindaca*

	2021	2019	2015	2013
la sindaca	98	15	2	2
il sindaco + NPF	124	43	28	24
la sindaco	8	1	10	6
sindachessa	3	6	2	4
donna sindaco	40	20	26	16
sindaco donna	62	23	38	22

Questi risultati, come già segnalato dal rapporto tra le occorrenze femminili e quelle maschili di 1:161,61 registrato nel corpus 2021, mostrano una spiccata tendenza verso l'utilizzo del maschile.

La forma preponderante resta *il sindaco* seguito dal nome proprio femminile, indice del fatto che la carica sia ancora largamente percepita come prettamente maschile. Confermano questo risultato le diverse occorrenze delle forme che fanno uso del modificatore "donna" e quelle di *la sindaco*. Vengono inoltre registrate, per la prima volta, delle occorrenze con la forma in *-essa*, anche se in minimo numero.

In linea generale, dunque, è possibile, anche in questa seconda analisi, riscontrare una leggera crescita dell'utilizzo delle forme femminili, ma ancora un totale radicamento del maschile con accezione neutra. Come è stato riscontrato anche nelle discussioni tra gli utenti di Wikipedia riguardo l'uso del maschile neutro precedentemente riportate[60], è chiaro che, tendenzialmente, l'utilizzo così ampio del maschile con accezione neutra è dato da una fondamentale incertezza

[60] <https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Secondo_l%27Accademia_della_Crusca_bisogna_dire_%22Sindaca%22>.

riguardo l’accettabilità delle forme trasposte al femminile, piuttosto che da un volontario utilizzo discriminante.

I dubbi nei confronti di queste forme derivano da una mancanza di percezione delle stesse all’interno del parlato quotidiano e, di conseguenza, dalla sensazione che queste parole non siano davvero parte del sistema linguistico in uso. Affinché la femminilizzazione dei sostantivi maschili possa proseguire la sua crescita ed entrare a far parte dell’uso comune, è necessaria un’ulteriore presa di coscienza da parte dei parlanti riguardo la sua piena legittimità grammaticale.

4 Conclusioni

Possiamo quindi notare un'integrazione estremamente lenta dei femminili professionali e dei titoli onorifici nel linguaggio enciclopedico di Wikipedia, ancora tendenzialmente contrastata dagli utenti e dai mediatori che non vedono nella femminilizzazione dei titoli al maschile una soluzione tendenzialmente corretta e ancora largamente accettata dai parlanti. La preferenza ricade nella quasi totalità dei casi sull'utilizzo del maschile con accezione neutra e si nota una grande riluttanza al cambiamento. In questo, il linguaggio di Wikipedia sembra più che altro essere lo specchio del comportamento della maggioranza dei parlanti e delle istituzioni.

Come già anticipato, infatti, il processo di questo cambiamento ha ben poco a che vedere con la lingua in sé, tantomeno con il linguaggio di Wikipedia stessa, ma, anzi, è possibile pensare che il linguaggio enciclopedico sarà definitivamente interessato da questa innovazione linguistica solo nel momento in cui, a tutti gli effetti, si attueranno dei cambiamenti a livello sociale tali da permettere la definitiva entrata in uso di questi termini.

Due, però, sono le cose fondamentali da tenere a mente nel momento in cui ci si approccia allo studio di questo cambiamento in atto.

La prima è che non è strettamente necessario avere un'unica soluzione nella procedura di femminilizzazione, ma è possibile integrare la nuova terminologia femminile anche all'interno di una lingua che conservi parte dell'accezione neutrale del maschile o alcune forme particolarmente tenaci. Per completare l'integrazione dei cambiamenti all'interno di un sistema preesistente e consolidato è del tutto normale passare attraverso fasi intermedie che prevedano soluzioni spesso non percepite come totalmente corrette da parte di chi si auspica un cambiamento radicale nella lingua, in questo caso in favore della femminilizzazione.

La seconda è che, come ricordato da Claudio Marazzini, «la lingua è una democrazia, in cui la maggioranza governa, i grammatici prendono atto delle innovazioni e cercano di farle andare d'accordo con la tradizione, e le minoranze, anche ribelli, hanno pur diritto di esistere»[61]. Dunque, fondamentalmente, ogni parlante deve poter mantenere il suo diritto di scegliere le parole che desidera utilizzare, e la lingua è, alla fine, costruita tramite le scelte che quotidianamente la maggior parte dei parlanti attua. Le varie correnti che si stanno scontrando riguardo la femminilizzazione dei sostantivi di professione hanno alle spalle delle motivazioni che devono essere rispettate, sia che

[61] Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, p. 123.

derivino da una radicalizzazione nei confronti della lingua e da un'affezione alla tradizione, sia che provengano da un'ideologia spinta dalla volontà di innovazione.

Il transito verso un linguaggio più paritario è insindacabilmente in atto, ma il riconoscimento definitivo di questa tipologia di forme potrà probabilmente avvenire solo nel momento in cui ci sarà un'effettiva modifica strutturale dell'organizzazione sociale e linguistica dei parlanti.

La strada sembra ancora piuttosto lunga e tortuosa, ma i piccoli cambiamenti avvenuti sinora sono già tangibili.

Il cuore di questa trasformazione linguistica, ad ogni modo, resta la curiosità con cui ogni individuo deve continuare a sondare l'esistenza, non chiudendo mai davanti a sé in maniera imprescindibile porte che affacciano su nuovi mondi e universi. Più parole conosciamo, più ci apriamo al linguaggio e più lo spazio si apre davanti ai nostri piedi, permettendoci di camminare su strade mai esplorate. Come ci ricorda il linguista Wilhelm Von Humboldt, infatti: «Le lingue non servono propriamente a esporre la verità già nota, ma piuttosto a scoprire la verità che era prima ignota. La loro diversità non è una diversità di suoni e di segni, ma di visioni del mondo».

5 Bibliografia

Vera Gheno, *Femminili singolari*, Firenze, Effequ, 2019.

Cecilia Robustelli, *Linee guida per l'uso del genere nel linguaggio amministrativo*, Firenze, Comune di Firenze, 2012.

Cecilia Robustelli, *Sindaco e sindaca: il linguaggio di genere*, Roma, Gruppo Editoriale l'Espresso – Accademia della Crusca, 2016.

Alma Sabatini, *Il sessismo nella lingua italiana*, Roma, Presidenza del Consiglio Dei Ministri. Dipartimento per l'informazione e l'editoria, 1987.

Luca Serianni, *Prima lezione di grammatica*, Roma-Bari, Laterza, 2006.

Giuseppe Zarra e Claudio Marazzini, *«Quasi una rivoluzione». I femminili di professioni e cariche in Italia e all'estero*, a cura di Yorick Gomez Gane, Firenze, Accademia della Crusca, 2017.

6 Sitografia

6.1 Fonti linguistiche

Elenco dei senatori, XVII legislatura n. 2 giugno 2013, Senato della Repubblica, 3 giugno 2013, https://www.senato.it/application/xmanager/projects/leg18/file/elenco_senatori_n_2_XVII_definitivo.pdf.

Patrizia Bellucci, *Il femminile di* questore *e di prefetto*, «Accademiadellacrusca.it», 17 marzo 2014, https://accademiadellacrusca.it/it/consulenza/il-femminile-di-questore-e-di-prefetto/865.

Cecilia Robustelli, *Infermiera sì, ingegnera no?*, «Accademiadellacrusca.it», 8 marzo 2013, https://accademiadellacrusca.it/it/contenuti/infermiera-si-ingegnera-no/7368.

La neutralità di genere nel linguaggio usato al Parlamento europeo, Parlamento europeo, 2018, https://www.europarl.europa.eu/cmsdata/187102/GNL_Guidelines_IT-original.pdf

6.2 Fonti informatiche

Giuseppe Attardi, Stefano Dei Rossi e Maria Simi, *The Tanl Pipeline*, gennaio 2010, https://www.researchgate.net/publication/228966621_The_Tanl_Pipeline.

GNU Grep 3.5, https://www.gnu.org/software/grep/manual/grep.html.

Python, pathlib – Object-oriented filesystem paths, https://docs.python.org/3/library/pathlib.html.

Python, re -- Regular expression operations, https://docs.python.org/3/library/re.html.

Wikimedia Downloads, https://dumps.wikimedia.org.

6.3 WikiExtractor

Giuseppe Attardi, *WikiExtractor*, https://github.com/attardi/wikiextractor.

Giuseppe Attardi, *WikiExtractor, Codice* `clean.py`, https://github.com/attardi/wikiextractor/blob/master/wikiextractor/clean.py.

Giuseppe Attardi, *WikiExtractor, Codice* `extract.py`, https://github.com/attardi/wikiextractor/blob/master/wikiextractor/extract.py.

Giuseppe Attardi, *WikiExtractor, Codice* `extractPage.py`,
https://github.com/attardi/wikiextractor/blob/master/wikiextractor/extractPage.py.
Giuseppe Attardi, *WikiExtractor, Codice* `WikiExtractor.py`,
https://github.com/attardi/wikiextractor/blob/master/wikiextractor/WikiExtractor.py.
Giuseppe Attardi, *WikiExtractor*, https://github.com/attardi/wikiextractor/wiki.

6.4 Wikipedia

Aiuto: Analisi del database, Struttura dei file,
https://it.wikipedia.org/wiki/Aiuto:Analisi_del_database#Struttura_dei_file.
Aiuto: Sezioni, https://it.wikipedia.org/wiki/Aiuto:Sezioni.
Aiuto: Wikitesto, https://it.wikipedia.org/wiki/Aiuto:Wikitesto.
MediaWiki, https://it.wikipedia.org/wiki/MediaWiki.
Template: Numero voci in Wikipedia,
https://it.wikipedia.org/wiki/Template:Numero_voci_in_Wikipedia.
Wc (Unix), https://it.wikipedia.org/wiki/Wc_(Unix).
Wikipedia: Bar, Discussioni, Secondo l'Accademia della Crusca bisogna dire "Sindaca",
https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Secondo_l%27Accademia_della_Crusca_bisogna_dire_%22Sindaca%22.
Wikipedia: Bar, Discussioni, Femminilizzazioni forzate,
https://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Femminilizzazioni_forzate
Wikipedia in italiano, https://it.wikipedia.org/wiki/Wikipedia_in_italiano.
Wikipedia in italiano, Cronologia recente,
https://it.wikipedia.org/wiki/Wikipedia_in_italiano#Cronologia_recente.