# Assessing the quality of ChatGPT's generated output in light of human-written texts.
## A corpus study based on textual parameters

Anna-Maria De Cesare
Technische Universität Dresden

This contribution has an exploratory nature, marking the initial phase of a broader research project aimed at achieving both descriptive and theoretical objectives. The primary goal is to evaluate the quality of texts produced by Language Model Models (LLMs). Two key aspects are examined: the quality of generated texts in comparison to human-authored texts and the identification of distinctive features characterizing this emerging text typology. The analysis is centered on textual parameters, encompassing various phenomena related to text segmentation and three dimensions of text organization (the referential-thematic dimension, the logico-argumentative dimension, and the polyphonic-enunciative dimension). Results of different case studies based on a self-assemble corpus of biographies generated by ChatGPT-3.5 and published on Wikipedia are presented.

**Keywords:** text generation; ChatGPT; textual parameters; self-assembled corpus

## 1. Introduction

Texts generated automatically by large language models (in short, LLMs), such as ChatGPT's operating system GPT-3.5 or GPT-4, based on OpenAI's GPTn series, are a new and important object of linguistic study. They result from systems trained on extremely large samples of texts written in different languages (GPT3 was trained on a corpus of 300 billion tokens) and can therefore (tentatively) be claimed to be representative of wide and somewhat undifferentiated linguistic communities. Furthermore, according, e.g., to Nissim & Pannitto (2022: 118), LLMs faithfully represent the reality on which they are trained. Consequently, they can be used to uncover a wide array of phenomena, ranging from biases and asymmetries associated with a linguistic community to the most frequent and thus characteristic patterns used in language (be it in a specific language

*CHIMERA. Romance Corpora and Linguistic Studies* 10 (2023), 179-210. ISSN 2386-2629

or cross-linguistically) or in a linguistic sequence (be it at textual, sentence or phrase level). So far, the output of LLMs has mainly been described based on English texts (see, among many others, Ferrara 2023 as well as Kotek, Dockum & Sun 2023). As a result, more research is needed on other languages, such as Italian and all the other Romance languages (for a recent study on Spanish, see Garrido-Muñoz, Martínez-Santiago & Montejo-Ráez 2023)[1].

The aim of this contribution – which has an exploratory character and is conceived as the first step in a larger research project pursuing both descriptive and theoretical goals – is to investigate the quality of texts generated by LLMs: on the one hand, we want to assess how well generated texts are written and how natural they are in comparison to human-written texts[2]; on the other hand, we want to describe the characteristic features of this new text typology. While there are many features to consider (lexical, grammatical, pragmatic etc.), the present contribution is devoted to textual ones. Textual parameters concern, inter alia, aspects as diverse as the nature of the constitutive units forming a text, the markers used to signal boundaries between textual units and the set of cohesive devices employed. Following the theoretical framework known as "Basel Model for paragraph segmentation" (outlined in Ferrari et al. 2008 and Ferrari 2014), we distinguish three dimensions of text organization: the referential-thematic dimension, the logico-argumentative dimension and the polyphonic-enunciative dimension (for details, see Ferrari 2014: 25-28).

Our contribution is organized as follows: we start by presenting the work corpus set up to study the quality of automatically generated texts, comprising a subcorpus of biographies generated by ChatGPT-3.5 and a comparable subcorpus of biographies published on Wikipedia, and briefly outline our methodology (§ 2); we then describe the results of several small-scale case studies: the first one concerns text segmentation by the set of strong punctuation marks (§ 3); the second and third ones highlight relevant aspects related, respectively, to the referential and topical (§ 4) as well as the logico-argumentative (§ 5) dimensions of textual organization. In the conclusion (§ 6), we provide a general assessment on the quality of ChatGPT's output and indicate questions to address in future studies.

[1] Recent studies on Romance languages describe the linguistic and textual features of automated texts based on a template with gaps, which are used e.g. to report the results of stock markets or national elections (see De Cesare 2021a and De Cesare, Eliasson & Weidensdorfer 2023). A first exploratory study investigating the textual properties of ChatGPT's output in Italian is proposed in De Cesare in press.

[2] We are not interested in the quality of the content. As is well known, and as we indeed also find in our corpus of generated biographies (described in § 2), LLMs produce 'hallucinations', i.e. nonfactual, untruthful information (for details, see Bang et al. 2023).

## 2. Corpus BioCGPT/BioWiki

### 2.1 Data collection used to construct the corpus

To probe the writing quality (in the sense outlined in § 1) of LLMs, we created a corpus of comparable texts belonging to the genre of biographies. The corpus, called BioCGPT/BioWiki, consists of two parts: BioCGPT and BioWiki[3].

The BioCGPT subcorpus was set up using the free version of ChatGPT-3.5, available online in a user-friendly interface since November 30, 2022. The LLM on which ChatGPT-3.5 is based was trained on a very large sample of texts collected on the Internet. These texts include Wikipedia entries, books, newspapers, magazines, blogs etc.[4] The Wikipedia pages make up 3% of the training data[5]. ChatGPT is "a large language model-based chatbot" (ChatGPT, Wikipedia[6]): it has been designed to be primarily used in a dialogic modality[7], in sequences of "prompt-answer" pairs. The BioCGPT corpus includes texts generated with a single "prompt": the texts are thus to be considered monologic rather than dialogic.

The BioCGPT subcorpus includes of 35 female biographies generated by ChatGPT-3.5 in Italian. These biographies were generated individually between 7.1.2023 and 14.6.2023 and meet two conditions: (i) they concern well-known personalities (mostly) born in the 20th century; these personalities are active in different fields (science, art, entertainment, sports, etc.) and (ii) biographies of the same personalities are also available in Italian on Wikipedia.

All texts were generated manually using the following zero-shot learning prompt: "Write a 1000-words text to explain who x is," where x corresponds to a

[3] The corpus is multilingual (Italian, French and German) and includes a subcorpus of male biographies. In this contribution, we only present the data related to the Italian subcorpus of female biographies.

[4] The exact nature of the training data is unknown. We are thus dealing with a black box.

[5] For details on the sources of the training data, see https://en.wikipedia.org/w/index.php?title=GPT-3&oldid=1179524498#References (accessed on October 19, 2023)

[6] https://en.wikipedia.org/w/index.php?title=ChatGPT&oldid=1175956865#Use_and_implications (accessed on October 19, 2023)

[7] According to OpenAI, "ChatGPT was optimized for dialogue by using Reinforcement Learning with Human Feedback (RLHF) – a method that uses human demonstrations and preference comparisons to guide the model toward desired behavior" (https://help.openai.com/en/articles/6783457-what-is-chatgpt).

variable that coincides with a person's first and last name, for example Dacia Maraini[8]. An example of both a prompt and the generated output is given in Figure 1 in § 2.2.

The BioWiki subcorpus, in turn, was created with the help of Trafilatura (Barbaresi 2021). We scraped all the content of the Wikipedia female entries mentioned in point (ii) above and saved it in a single txt. file. To ensure the highest possible comparability with the generated texts, we then discarded from the file several Wikipedia "idiosyncratic" contents, such as the final notes, the filmography or complete list of works of the personality described in the entry, the references and external links.

## 2.2. BioCGPT/BioWiki: Corpus design and methods

Our self-assembled corpus includes biographies generated in Italian by ChatGPT-3.5 on 35 different women (BioCGPT) and the corresponding 34[9] biographies available in Italian on Wikipedia (BioWiki). Table 1 shows the design of the corpus.

**Table 1.** Corpus BioCGPT/BioWiki

| **BioFemCGPT** | **BioFemWiki** |
|---|---|
| 35 biographies of It. women | 34 biographies of It. women |

In the ChatGPT-3.5 subcorpus, each female biography was generated at least four times (on different dates), both to check for a possible evolution of the textual output of the LLM and to compare the results obtained within the same biography. As a result, the BioCGPT subcorpus includes a total of 168 texts, while the BioWiki subcorpus only includes a total of 34 texts. Table 2 provides an overview of the main quantitative properties of the work corpus used in this study.

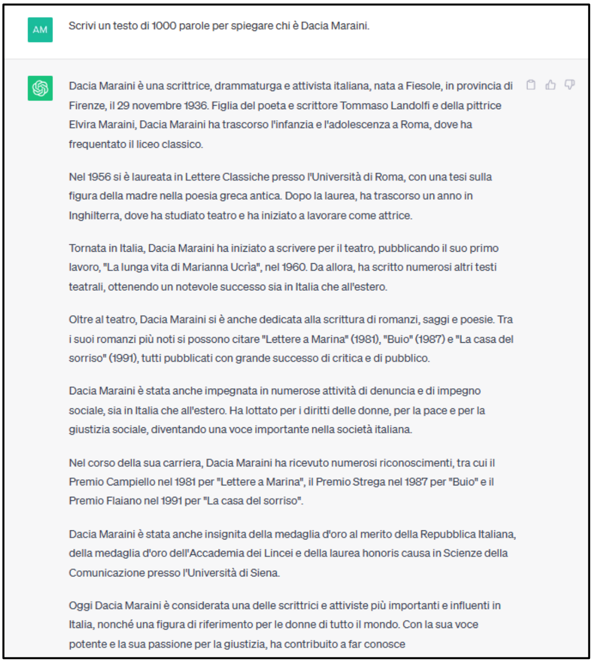
[8] As pointed out by an anonymous reviewer, the prompt could have been phrased and conceived differently, as it is quite generic. Importantly, GPT-3.5 has not knowledge on texts genres. Consequently, without further specification (e.g. in the form of examples), it does not know what the characteristic features of biographies really are. This is why the output of a prompt such as "Write a 1000-words biography on x" does not seem to differ (much) from the results obtained with the prompt "Write a 1000-words text to explain who is x". In other words, the verb *explain* does not seem to trigger expositive rather than narrative textual outputs (for details on these two text types, see Werlich 1975, as well as De Cesare 2011).

[9] The BioFemWiki subcorpus comprises 34 female profiles because – as we only noticed later – there is no entry for Letizia Jaccheri in Italian.

**Table 2.** Corpus BioCGPT/BioWiki (data from Sketch Engine)

| | **BioFemCGPT** | **BioFemWiki** |
|---|---|---|
| **Texts** | 168 | 34 |
| **Tokens** | 75,932 | 84,351 |
| **Words** | 65,624 | 68,192 |
| **Sentences** | 2,710 | 1,860 |

As can be observed based on Table 2, both subcorpora include a similar number of tokens and a relatively close number of words, even though BioFemCGPT includes 168 texts and BioFemWiki only 34. This result is due to the fact that the generated biographies are generally much shorter than the Wikipedia biographies. The biography of Dacia Maraini reproduced in Figure 1 and Figure 2 allows grasping the difference in length: the generated text includes 339 words, while the Wikipedia entry includes 984 words (excluding the final list of works and references). Note that Figure 1 reproduces the whole generated text, while Figure 2 only shows part of it.



Scrivi un testo di 1000 parole per spiegare chi è Dacia Maraini.

Dacia Maraini è una scrittrice, drammaturga e attivista italiana, nata a Fiesole, in provincia di Firenze, il 29 novembre 1936. Figlia del poeta e scrittore Tommaso Landolfi e della pittrice Elvira Maraini, Dacia Maraini ha trascorso l'infanzia e l'adolescenza a Roma, dove ha frequentato il liceo classico.

Nel 1956 si è laureata in Lettere Classiche presso l'Università di Roma, con una tesi sulla figura della madre nella poesia greca antica. Dopo la laurea, ha trascorso un anno in Inghilterra, dove ha studiato teatro e ha iniziato a lavorare come attrice.

Tornata in Italia, Dacia Maraini ha iniziato a scrivere per il teatro, pubblicando il suo primo lavoro, "La lunga vita di Marianna Ucrìa", nel 1960. Da allora, ha scritto numerosi altri testi teatrali, ottenendo un notevole successo sia in Italia che all'estero.

Oltre al teatro, Dacia Maraini si è anche dedicata alla scrittura di romanzi, saggi e poesie. Tra i suoi romanzi più noti si possono citare "Lettere a Marina" (1981), "Buio" (1987) e "La casa del sorriso" (1991), tutti pubblicati con grande successo di critica e di pubblico.

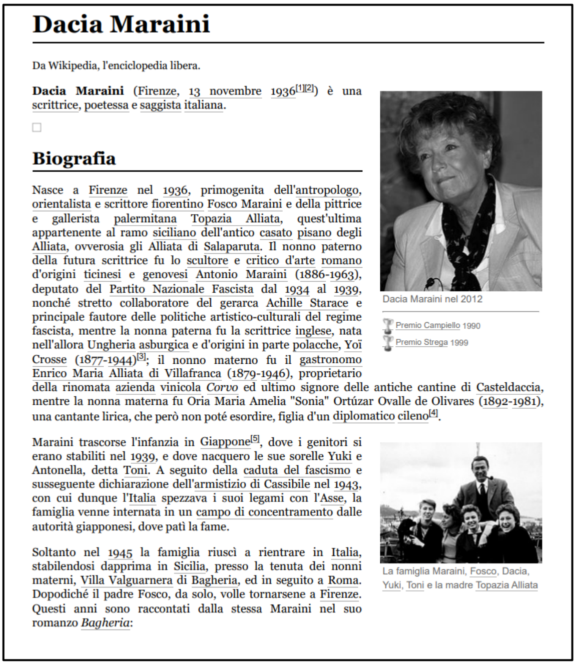
Dacia Maraini è stata anche impegnata in numerose attività di denuncia e di impegno sociale, sia in Italia che all'estero. Ha lottato per i diritti delle donne, per la pace e per la giustizia sociale, diventando una voce importante nella società italiana.

Nel corso della sua carriera, Dacia Maraini ha ricevuto numerosi riconoscimenti, tra cui il Premio Campiello nel 1981 per "Lettere a Marina", il Premio Strega nel 1987 per "Buio" e il Premio Flaiano nel 1991 per "La casa del sorriso".

Dacia Maraini è stata anche insignita della medaglia d'oro al merito della Repubblica Italiana, della medaglia d'oro dell'Accademia dei Lincei e della laurea honoris causa in Scienze della Comunicazione presso l'Università di Siena.

Oggi Dacia Maraini è considerata una delle scrittrici e attiviste più importanti e influenti in Italia, nonché una figura di riferimento per le donne di tutto il mondo. Con la sua voce potente e la sua passione per la giustizia, ha contribuito a far conosce

Figure 1. BioCGPT-3.5 (generated on 7.1.2023)



# Dacia Maraini

Da Wikipedia, l'enciclopedia libera.

**Dacia Maraini** (Firenze, 13 novembre 1936[1][2]) è una scrittrice, poetessa e saggista italiana.

□

## Biografia

Nasce a Firenze nel 1936, primogenita dell'antropologo, orientalista e scrittore fiorentino Fosco Maraini e della pittrice e gallerista palermitana Topazia Alliata, quest'ultima appartenente al ramo siciliano dell'antico casato pisano degli Alliata, ovverosia gli Alliata di Salaparuta. Il nonno paterno della futura scrittrice fu lo scultore e critico d'arte romano d'origini ticinesi e genovesi Antonio Maraini (1886-1963), deputato del Partito Nazionale Fascista dal 1934 al 1939, nonché stretto collaboratore del gerarca Achille Starace e principale fautore delle politiche artistico-culturali del regime fascista, mentre la nonna paterna fu la scrittrice inglese, nata nell'allora Ungheria asburgica e d'origini in parte polacche, Yoï Crosse (1877-1944)[3]; il nonno materno fu il gastronomo Enrico Maria Alliata di Villafranca (1879-1946), proprietario della rinomata azienda vinicola *Corvo* ed ultimo signore delle antiche cantine di Casteldaccia, mentre la nonna materna fu Oria Maria Amelia "Sonia" Ortúzar Ovalle de Olivares (1892-1981), una cantante lirica, che però non poté esordire, figlia d'un diplomatico cileno[4].

Dacia Maraini nel 2012

Premio Campiello 1990
Premio Strega 1999

Maraini trascorse l'infanzia in Giappone[5], dove i genitori si erano stabiliti nel 1939, e dove nacquero le sue sorelle Yuki e Antonella, detta Toni. A seguito della caduta del fascismo e susseguente dichiarazione dell'armistizio di Cassibile nel 1943, con cui dunque l'Italia spezzava i suoi legami con l'Asse, la famiglia venne internata in un campo di concentramento dalle autorità giapponesi, dove patì la fame.

La famiglia Maraini, Fosco, Dacia, Yuki, Toni e la madre Topazia Alliata

Soltanto nel 1945 la famiglia riuscì a rientrare in Italia, stabilendosi dapprima in Sicilia, presso la tenuta dei nonni materni, Villa Valguarnera di Bagheria, ed in seguito a Roma. Dopodiché il padre Fosco, da solo, volle tornarsene a Firenze. Questi anni sono raccontati dalla stessa Maraini nel suo romanzo *Bagheria*:

Figure 2. BioWiki (accessed on September 18, 2023[10])

Figure 1 also allows highlighting that the output of ChatGPT-3.5 answers our prompt only in part. A count of the number of words in each text making up the BioCGPT subcorpus reveals that the generated output never exceeds 500 words (not infrequently, as can be observed at the end of Figure 1, the text is interrupted during the generation, leaving the last sentence incomplete): the longest texts include 489 words. A first hypothesis explaining this result could be a lack of understanding of what a word is and consequently what '1000 words' represent. However, when prompting ChatGPT-3.5 in relation to word counts in sentences and texts, the results are correct[11]:

> Prompt: Write a sentence of 10 words.
> ChatGPT: Underneath the starry sky, they shared stories and dreams together.
>
> Prompt: Write a text of 20 words.
> ChatGPT: Amidst the serene forest, a babbling brook danced through mossy stones, and birds sang harmonious melodies under the golden sun.

[10] The permanent links of all the cited Wikipedia entries are provided in the references section.
[11] I only prompted the model twice. More research is needed on this interesting issue.

Besides the failed output related to the total number of words included in each text, the prompt was not followed on another important point: one text (out of 168) was generated in English even though the prompt was written in Italian[12].

Finally, as far as the methodology is concerned, the textual features investigated in §§ 3 to 5 rely on mixed methods. Our study is both corpus-based (in § 3, for instance, our starting point is a list of preset punctuation marks) and corpus-driven (again in § 3, we identify the uses and functions of these punctuation marks through a (semi-)manual analysis of the data). Moreover, when the feature lends itself to it, we provide additional quantitative data (for details on these approaches, see Cresti & Panunzi 2013).

## 3. Text segmentation and nature of textual units: the case of strong punctuation marks

### 3.1 Segmenting a text block in Utterances

The biographies included in our corpus consist of text blocks (see Figure 1 and Figure 2 above). These blocks, in turn, are composed by one or more Utterances, defined as "the key unit[s] involved in the semantico-pragmatic structuring of the paragraph [for us the text block]" (Ferrari 2014: 32; for details on the form and function of Utterances, see Ferrari 2014: 33ff). In texts written in Italian, Utterance boundaries are signaled by 'strong' punctuation marks (Ferrari 2014: 33), most frequently by the full stop (/ . /) and the colon (/ : /), but also by the exclamation mark (/ ! /), the question mark (/ ? /) and by parentheses (/ (   ) /)[13].

In order to identify how Utterance boundaries are signaled in the generated biographies and highlight possible differences in respect to the ones published on Wikipedia, Table 3 provides an overview of the strong punctuation marks found in the data present in both subcorpora. The data is ordered following the absolute and normalized frequency of the marks in BioFemCGPT.

[12] In our entire multilingual corpus of generated texts, comprising a total 900 biographies on females and males, English written outputs for non-English prompts could be observed on six occasions: three times as responses to a French prompt, two times to an Italian one and once to a German one.

[13] We did not consider the use of dashes.

**Table 3.** Strong punctuation marks: absolute and normalized frequency (pmt = n. of occurrences per million tokens)

| | **BioFemCGPT** | **pmt** | **%** | **BioFemWiki** | **pmt** | **%** |
|---|---|---|---|---|---|---|
| / . / | 2’702 | 35’585 | 91 | 2’410 | 28’571 | 72 |
| / ( ) / | 225 | 2’963 | 8 | 465 | 5’513 | 14 |
| / : / | 20 | 263 | 1 | 275 | 3’260 | 8 |
| / ? / | 3 | 40 | 0 | 20 | 237 | 0 |
| / ! / | 2 | 26 | 0 | 23 | 273 | 1 |
| / ; / | 1 | 13 | 0 | 173 | 2’051 | 5 |

The same paradigm of strong punctuation marks, consisting of six different forms, is used in both subcorpora. At the same time, there are clear differences between the two subcorpora in the frequency with which each punctuation mark is used. If we look at the data in each subcorpus individually, we note that BioCGPT mainly includes two punctuation marks: the full stop (covering 91% of the whole occurrences) and, but much more marginally, the parentheses (covering another 8% of the data). In BioWiki, by contrast, four different marks are used: the full stop (covering 72% of the data), the parentheses (14%), the colon (8%) and the semicolon (5%). A comparison of the data across the two subcorpora allows to further highlight that in the generated biographies there is a higher number of full stops (35’585 vs. 28’571 per million tokens, in short pmt), while we find a lower number of all the other marks: parentheses (2’963 vs. 5’513 pmt), colon (263 vs. 3’260 pmt), question mark (40 vs. 237 pmt) and exclamation mark (26 vs. 273 pmt).

From a textual point of view, the data described above points to the following assumptions: in BioCGPT, Utterance boundaries[14] related to the main line of the text (which leaves aside parenthetical contents) are predominantly signaled by the full stop, as shown in ex. (1) and other examples to follow (also see Figure 1 above). In BioWiki, by contrast, Utterance boundaries are signaled by a variety of other marks, including the colon (see ex. 2, as well as the text at the end of Figure 2):

(1) Elsa Morante è stata una scrittrice italiana del XX secolo, nata a Roma il 18 agosto 1912 e scomparsa il 25 novembre 1985. //**U1** È stata una delle più importanti figure della letteratura italiana del dopoguerra, autrice di romanzi, raccolte di poesie e libri per bambini. //**U2** Morante è stata una

[14] We adopt the annotation scheme of the Basel Model for paragraph segmentation (Ferrari et al. 2008; Ferrari 2014), where Utterance boundaries are signaled by a double slash, followed by the abbreviation ‘U’ (standing for ‘Utterance’) and sometimes a number indicating the order of Utterances in the text block.

figura poliedrica e complessa, una voce fuori dal coro che ha saputo fare della scrittura un mezzo di espressione e di lotta per la giustizia e la libertà. //**U3** (Elsa Morante, BioCGPT)

'Elsa Morante was a 20th-century Italian writer, born in Rome on August 18, 1912, and died on November 25, 1985. //**U1** She was one of the most important figures in postwar Italian literature, author of novels, poetry collections and children's books. //**U2** Morante was a multifaceted and complex figure, a voice out of the chorus who knew how to make writing a means of expression and struggle for justice and freedom.' //**U3**[15]

(2) Nel 1989 Rosy Bindi incomincia la sua carriera politica iscrivendosi alla Democrazia Cristiana: //**U1** in quell'anno si candida alle elezioni europee con il partito scudocrociato nella circoscrizione Nord-Est ottenendo 211.000 preferenze e viene eletta. //**U2** A Strasburgo ricopre l'incarico di vicepresidente della Commissione cooperazione e sviluppo e, successivamente, di presidente della Commissione petizioni e diritti dei cittadini. //**U3** (Rosy Bindi, BioWiki)

'In 1989 Rosy Bindi began her political career by joining the Christian Democratic Party: //**U1** in that year she ran in the European elections with the Scudocrociato party in the Northeast constituency, obtaining 211,000 preferences and was elected. //**U2** In Strasbourg she serves as vice-chair of the Cooperation and Development Committee and later as chair of the Petitions and Citizens' Rights Committee. //**U3**'

Another difference between the two subcorpora can be traced in the function performed by some of the punctuation marks. Let us consider the case of the semicolon. In BioCGPT, this mark (which is used only once) has a segmenting function. As can be observed in (3), it signals an Utterance boundary:

(3) Ma il suo lavoro non si è limitato alla scienza; //**U1** Rita ha anche dedicato la sua vita a promuovere i diritti delle donne e l'accesso all'istruzione, e ha lavorato per salvare molte vite dalla persecuzione nazista. //**U2** (Rita Levi Montalcini, BioCGPT)

[15] All examples have been translated with the help of DeepL.com, with some adjustments, in particular regarding the use of pronouns referring to human beings (which are often masculine instead of feminine).

‘But her work was not limited to science; //**U1** Rita also devoted her life to promoting women's rights and access to education, and she worked to save many lives from Nazi persecution. //**U2**’

In BioWiki, by contrast, the semi-colon (which is present 173 times) has two distinct functions: similarly to (3), it delimits two textual segments coinciding with Utterances (see ex. 4); in addition, it is used in a list, to separate its members (see ex. 5):

(4) [suo padre] Fu sindaco di Nuoro nel 1863. //**U1** La madre era Francesca Cambosu, descritta come donna di severi costumi; //**U2** dedita alla casa, educò lei la figlia. //**U3** (Grazia Deledda, BioWiki)

‘[her father] was mayor of Nuoro in 1863. //**U1** Her mother was Francesca Cambosu, described as a woman of stern manners; //**U2** devoted to the home, it was she who educated her daughter. //**U3**’

(5) Nel 2021 escono tre nuovi libri di Laura Boldrini: //**U1** *Le donne di Minsk. La rivolta pacifica per la democrazia in Bielorussia*, scritto con Lia Quartapelle e pubblicato da Infinito Edizioni; *Questo non è normale. Come porre fine al potere maschile sulle donne*, pubblicato da Chiarelettere; e *Una storia aperta. Diritti da difendere, diritti da conquistare*, pubblicato dalle Edizioni Gruppo Abele. //**U2** (Laura Boldrini, BioWiki)

‘Three new books by Laura Boldrini are coming out in 2021: //**U1** *Le donne di Minsk. La rivolta pacifica per la democrazia in Bielorussia*, written with Lia Quartapelle and published by Infinito Edizioni; *Questo non è normale. Come porre fine al potere maschile sulle donne*, published by Chiarelettere; and *Una storia aperta. Diritti da difendere, diritti da conquistare*, published by Gruppo Abele Editions. //**U2**’

### 3.2 Polyphony: sources of the Utterances ending with an exclamation or question mark

A qualitative analysis of the Utterances ending with an exclamation or a question mark in both subcorpora reveals another difference between the generated biographies and the biographies published on Wikipedia. Only Wikipedia biographies include direct reported speech, i.e. Utterances produced in another communicative situation (another moment in time and/or another place), either by the female at the center of the text or by someone else.

As far as the exclamation mark is concerned, in BioCGPT (where there only are 2 occ.) it only appears in titles (see ex. 6; note that generated texts do not use italics), while in BioWiki it appears both in titles (as in ex. 7) and in direct reported speech (in ex. 8, it occurs in a quote belonging to a letter):

(6) Il grande successo di Anna Magnani come attrice cinematografica iniziò negli anni '40, quando recitò in film come "Teresa Venerdì" (1941) di Vittorio De Sica e "Abbasso la ricchezza!" (1946) di Gennaro Righelli. (Anna Magnani, BioCGPT)

'Anna Magnani's great success as a film actress began in the 1940s, when she starred in films such as Vittorio De Sica's "Teresa Venerdì" (1941) and Gennaro Righelli's "Abbasso la ricchezza!" (1946).'

(7) [...] Nel medesimo periodo presentò due opere nell'esibizione *Promotrice Fiorentina*, a Firenze. La prima si intitola *Gondola*, la seconda *Buon dì!*. (Antonella Brandeis, BioWiki)

'In the same period she presented two works in the exhibition *Promotrice Fiorentina*, in Florence. The first was entitled *Gondola*, the second *Buon dì!*.'

(8) Su di lei scrisse prima Maksim Gor'kij e, più tardi, D. H. Lawrence.
Maksim Gor'kij raccomanda la lettura delle opere di Grazia Deledda a L. A. Nikiforova, una scrittrice esordiente. In una lettera del 2 giugno del 1910 le scrive: «Mi permetto di indicarle due scrittrici che non hanno rivali né nel passato, né nel presente: Selma Lagerlof e Grazia Deledda. Che penne e che voci forti! In loro c'è qualcosa che può essere d'ammaestramento anche al nostro mužik». (Grazia Deledda, BioWiki)

'Maksim Gor'kij first wrote about her and, later, D. H. Lawrence.
Maksim Gor'kij recommended reading the works of Grazia Deledda to L. A. Nikiforova, a beginning writer. In a letter dated June 2, 1910, he wrote to her, "Allow me to point you two female writers who have no rivals either in the past or in the present: Selma Lagerlof and Grazia Deledda. What pens and what strong voices! In them there is something that can be a lesson even to our mužik."'

As for the question mark, its function in BioCGPT is to raise a general question, issued from the point of view of a generic 'voice', about the protagonist of the biography (see ex. 9 and 10), while in BioWiki it appears in book and film titles

(11), but most often in direct reported speech, as shown in the two representative examples reproduced in (12) and (13):

(9) Ma chi era davvero Oriana Fallaci? (Oriana Fallaci, BioCGPT)
‘But who was Oriana Fallaci really?’

(10) Ma chi era Natalia Ginzburg? Come si è formata la sua personalità e il suo stile letterario? (Natalia Ginzburg, BioCGPT)
‘But who was Natalia Ginzburg? How was her personality and literary style formed?’

(11) […] È anche attrice protagonista o co-protagonista in alcuni film, come *Tu la conosci Claudia?* del 2004 con Aldo, Giovanni e Giacomo e *Il posto dell'anima* di Riccardo Milani, suo attuale marito. (Paola Cortellesi, BioWiki)

‘She also stars or co-stars in some films, such as 2004's *Tu la conosci Claudia?* with Aldo, Giovanni e Giacomo and *Il posto dell'anima* by Riccardo Milani, her current husband.’

(12) Il 31 gennaio 2014, Beppe Grillo condivide sul suo profilo Facebook un video, commentandolo con la domanda: "Cosa succederebbe se ti trovassi Boldrini in macchina?" (Laura Boldrini, BioWiki)

‘On January 31, 2014, Beppe Grillo shared a video on his Facebook profile, commenting on it with the question, "What would happen if you found Boldrini in your car?"’

(13) Ancora lo ribadisce in un'altra lettera, dell’11 maggio 1893: «*Io non sono certa se ho venti o ventun anni compiuti; neanche mia madre ne è certa, ma è più probabile che ne abbia ventuno che venti. Sono vecchia, non è vero?* […]». (Grazia Deledda, BioWiki)

‘She reiterates this again in another letter, dated May 11, 1893: "*I am not certain whether I am twenty or twenty-one years old; my mother is not certain either, but it is more likely that I am twenty-one than twenty. I am old, am I not?* […]."’

### 3.3 Uses and functions of parenthetical contents

Besides showing clear differences in terms of frequency, a closer look at the ways parentheses are used in BioCGPT and BioWiki allows to highlight important differences between generated biographies and biographies published on Wikipedia. These differences can be described based on the data provided in Figure 3 and Figure 4.

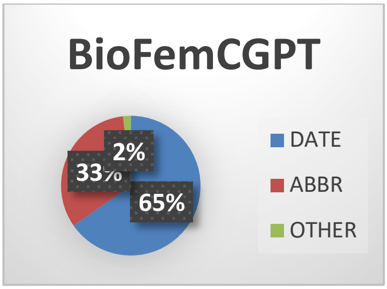


Figure 3. Parentheses in BioFemCGPT

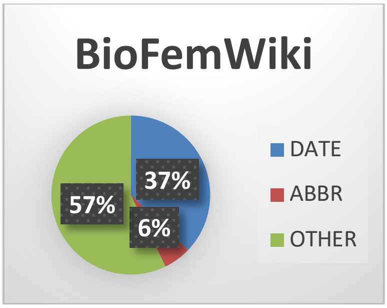


Figure 4. Parentheses in BioFemWiki

In BioCGPT, the parentheses serve two general purposes: they specify a date (65%) or an abbreviation (33%). In BioWiki, these two functions are also documented, but they are not as relevant (they make up 37% and 6% of the data, respectively).

In BioCGPT, parentheses are used first and foremost to specify temporal events, either by pointing to a period between two dates (as in ex. 14 and, in a slightly more complex and richer denotational format, ex. 15) or to a single date (as in ex. 16, where each date refers to the publication year of the cited work).

(14) Leonor Fini **(1908-1996)** è stata una pittrice, illustratrice, scenografa e scrittrice argentina naturalizzata italiana. (Leonor Fini, BioCGPT)

‘Leonor Fini **(1908-1996)** was an Argentine painter, illustrator, set designer and writer naturalized Italian.’

(15) Anna Magnani **(Roma, 7 marzo 1908 – Roma, 26 settembre 1973)** è stata un'attrice italiana. (Anna Magnani, BioCGPT)

‘Anna Magnani **(Rome, March 7, 1908 - Rome, September 26, 1973)** was an Italian actress.’

(16) La sua carriera letteraria inizia nel 1938 con la pubblicazione della sua prima raccolta di poesie, "L'alibi". In seguito, pubblica altre raccolte di poesie, come "La vita degli animali" **(1941)** e "Mensilità" **(1950)**, ma è soprattutto con la pubblicazione del romanzo "La Storia" **(1974)** che Morante diventa famosa in Italia e all'estero. (Elsa Morante, BioCGPT)

‘Her literary career began in 1938 with the publication of her first collection of poems, "The Alibi." Later, she published other collections of poems, such as "La vita degli animali" **(1941)** and "Mensilità" **(1950)**, but it was mainly with the publication of the novel "La Storia" **(1974)** that Morante became famous in Italy and abroad.’

In turn, when parentheses are used to specify abbreviations, they either outline the full form of an abbreviation occurring before the parenthesis (as in the first case in ex. 17) or, vice versa, provide the abbreviation of a full form appearing prior to the parenthesis (as shown by the second pair of parentheses in ex. 17):

(17) Samantha Cristoforetti è una degli astronauti più iconici d'Italia. Nata il 26 aprile 1977 a Milano, è stata la prima donna italiana ad essere selezionata per una missione spaziale dell'ESA **(Agenzia Spaziale Europea)** e la prima a volare a bordo della Stazione Spaziale Internazionale **(ISS)**. (Samantha Cristoforetti, BioCGPT)

‘Samantha Cristoforetti is one of Italy's most iconic astronauts. Born April 26, 1977, in Milan, she was the first Italian woman to be selected for an ESA **(European Space Agency)** space mission and the first to fly aboard the International Space Station **(ISS)**.’

As mentioned above, in BioWiki the parentheses are also used to specify temporal events (see ex. 18) and abbreviations (ex. 19 and 20). Importantly, the dates referring to the birth and death of a person have a fixed format[16], shown in (18), which is also used in the generated biographies (15), but only very marginally. In the generated biographies, the most common way of referring to these two major life events is the format in (14).

(18) Maria Tecla Artemisia Montessori, nota come Maria Montessori **(Chiaravalle, 31 agosto 1870 – Noordwijk, 6 maggio 1952)** è stata una pedagogista, educatrice e medico italiana, internazionalmente nota per il metodo educativo che prende il suo nome, adottato in migliaia di scuole dell'infanzia, elementari, medie e superiori in tutto il mondo. (Maria Montessori, BioWiki)

'Maria Tecla Artemisia Montessori, known as Maria Montessori **(Chiaravalle, August 31, 1870 - Noordwijk, May 6, 1952)** was an Italian pedagogist, educator and physician, internationally known for the educational method named after her, which has been adopted in thousands of preschools, elementary, middle and high schools around the world.'

(19) Il 5 aprile 2016 le viene dedicato un nuovo ibrido di orchidea spontanea scoperto in Salento e denominato *Ophrys × montalciniae nothosubsp. cristoforettiae*, ibrido tra *O. incubacea subsp. brutia* e *O. sphegodes subsp. classica*. L'ibrido in questione è stato pubblicato sulla rivista nazionale G.I.R.O.S. **(Gruppo Italiano per la Ricerca di Orchidee Spontanee)**. (Samantha Cristoforetti, BioWiki)

'On April 5, 2016, a new spontaneous orchid hybrid discovered in Salento and named *Ophrys × montalciniae nothosubsp. cristoforettiae,* a hybrid between *O. incubacea subsp. brutia* and *O. sphegodes subsp. classica*, is dedicated to her. The hybrid in question was published in the national journal G.I.R.O.S. **(Gruppo Italiano per la Ricerca di Orchidee Spontanee)**.'

[16] To ensure homogeneity across entries, the first sentence of the biographical entries published on Wikipedia are now generated automatically using a template (i.e., the technique mentioned in fn. 1 in relation to texts describing stock market's performances and election results). For details, see Tavosanis (2021: 426-427).

(20) Nel maggio 2009 è stata selezionata dall'Agenzia Spaziale Europea **(ESA)** e, dopo 5 anni, è diventata la prima astronauta di nazionalità italiana ad effettuare un volo spaziale. (Samantha Cristoforetti, BioWiki)

'In May 2009, she was selected by the European Space Agency **(ESA)** and, after five years, became the first astronaut of Italian nationality to make a space flight.'

In BioWiki, as already acknowledged, parentheses providing temporal information and information related to abbreviations are not the most frequent and thus characteristic ones. Many other parenthetical contents are present in BioWiki, which are totally absent from BioCGPT. Mention should be made in particular to parentheses with a specification or explicative function (see ex. 21) and to parentheses related to a quote. In this latter case, the parentheses can include a quote (as in ex. 22), appear in a quote to signal text omission (see ex. 23) or to specify information pertaining to the source of the quote (as in 24).

(21) Inoltre, nel febbraio del 1906 scrisse un Proclama alle donne italiane, quasi interamente dedicato ad Ancona, in cui descrive la città osservandola dall'alto del Monte **(si intende il Conero)**. (Maria Montessori, BioWiki)

'She also wrote a Proclamation to Italian Women in February 1906, almost entirely devoted to Ancona, in which she describes the city as she observes it from the top of the Mount **(meaning the Conero)**.'

(22) Lei si disse incredula per la decisione della Consulta **(«Quei magistrati hanno perso il senno»)**, ritenendo che tale sentenza fosse stata resa possibile dall'orientamento politico dei componenti della Corte: (Oriana Fallaci, BioWiki)

'She said she was incredulous at the decision of the Constitutional Court **("Those magistrates have lost their minds")**, believing that such a ruling was made possible by the political orientation of the Court's members:'

(23) «Il Capitano Samantha Cristoforetti **(…)** ha acquisito competenze che integrano gli aspetti metodologici, tecnologici e progettuali delle scienze ingegneristiche con le conoscenze necessarie per la realizzazione di esperimenti avanzati nel campo delle scienze biomediche. **(…)** Ha contribuito in modo decisivo […]». (Samantha Cristoforetti, BioWiki)

'"Captain Samantha Cristoforetti **(...)** has acquired skills that integrate the methodological, technological and design aspects of the engineering sciences with the knowledge required to carry out advanced experiments in the field of biomedical sciences. **(...)** She has contributed decisively [...]».'

(24) Durante la preparazione della sua tesi, frequentò le lezioni di antropologia fisica (o biologica) tenute da Giuseppe Sergi. La tesi, che discusse il 10 luglio del 1896, fu a carattere sperimentale: quasi cento pagine scritte a mano che portano il titolo "Contributo clinico allo studio delle allucinazioni a contenuto antagonistico" **(pp. 33-37)**. (Maria Montessori, BioWiki)

'While preparing her thesis, she attended lectures on physical (or biological) anthropology given by Giuseppe Sergi. The thesis, which she discussed on July 10, 1896, was experimental in nature: nearly one hundred handwritten pages bearing the title "Clinical contribution to the study of hallucinations with antagonistic content" **(pp. 33-37)**.'

## 4. Referential dimension of textual organization: the case of anthroponyms

### 4.1 Codification of textual referents evoking persons: set of linguistic forms

Let us now look at how textual referents – i.e. entities that are part of the ongoing discourse and can thus become "objects" of the discourse (for details on this notion, see Andorno 2003: 27ff. – are codified in BioCGPT and BioWiki). We only focus on textual referents evoking the main character of each biography, i.e. on the female at the center of the text.

Person reference can be expressed by a variety of linguistic means. The set of forms, which we present following their order of linguistic complexity, includes: a) null subjects and verb morphology (as in [null subj.] *recitò* in ex. 25); b) pronouns (*sua* in ex. 25; *lei* e *suo* in ex. 26); c) proper nouns (i.e., anthroponyms, such as *Eleonora Duse* or simply *Duse* in ex. 25) and d) definite descriptions, syntactically realized with more or less complex noun phrases (see *l'attrice* in ex. 26, referring to Eleonora Duse by means of a definite article followed by a hypernym highlighting a defining property of the referent):

(25) **Eleonora Duse** (1858-1924) è stata un'attrice teatrale italiana, considerata una delle più grandi attrici della storia del teatro. Nata a Vigevano, in Lombardia, **Duse** iniziò la **sua** carriera come attrice a soli 16 anni, recitando in

piccole compagnie teatrali in giro per l'Italia. Nel 1881, [**null subj.**] entrò a far parte della compagnia del famoso attore Ettore Scola, con il quale [**null subj.**] **recitò** in molte produzioni di successo. (Eleonora Duse, BioCGPT)

'**Eleonora Duse** (1858-1924) was an Italian stage actress, considered one of the greatest actresses in the history of theater. Born in Vigevano, Lombardy, **Duse** began **her** career as an actress when she was only 16 years old, acting in small theater companies around Italy. In 1881, [**null subj.**] joined the company of the famous actor Ettore Scola, with whom [**null subj.**] **acted** in many successful productions.'

(26) Boito adattò per **lei** *Antonio e Cleopatra*. La loro relazione restò sempre segreta e durò, fra alti e bassi, per diversi anni. In questo periodo, **l'attrice** frequentò gli ambienti della Scapigliatura e il **suo** repertorio si arricchì anche dei drammi di Giuseppe Giacosa, amico di Boito. (Eleonora Duse, BioWiki)

'Boito adapted Antony and Cleopatra for **her**. Their relationship always remained secret and lasted, between ups and downs, for several years. During this period, **the actress** frequented Scapigliatura circles and **her** repertoire was also enriched by the dramas of Giuseppe Giacosa, a friend of Boito.'

A first important outcome (based on a non-systematic qualitative analysis) concerns the set of forms available to evoke person reference. In BioCGPT we do not find definite descriptions, such as *l'attrice* (in ex. 26), to refer to the protagonist of the text. In contrast, definite descriptions referring to females are regularly used in BioWiki (besides ex. 26, see ex. 27-29), especially in contexts in which the person is evoked at the beginning of a new text block:

(27) **L'autrice** è vegetariana. (Dacia Maraini, BioWiki)
'The author is a vegetarian.'

(28) Nel novembre 2002 **la scrittrice** volò in Italia per opporsi all'autorizzazione della manifestazione organizzata dai no-global a Firenze per il timore che si potessero ripetere i fatti del G8 di Genova del 2001.
[…] Sempre nel 2002 **la scrittrice fiorentina** venne citata in giudizio in Svizzera dal Centro Islamico e dall'Associazione Somali di Ginevra, dalla sede di Losanna di SOS Racisme e da un cittadino privato, per il contenuto ritenuto razzista de *La rabbia e l'orgoglio*. (Oriana Fallaci, BioWiki)

'In November 2002 **the writer** flew to Italy to oppose the authorization of the demonstration organized by the no-globals in Florence because of the fear that the events of the G8 in Genoa in 2001 could be repeated.
[...] Also in 2002, **the Florentine writer** was sued in Switzerland by the Islamic Center and the Somali Association of Geneva, the Lausanne branch of SOS Racisme, and a private citizen, for the content deemed racist of *La rabbia e l'orgoglio*.'

(29) Sia a Firenze che a Budapest diffuse la propria opera sotto il nome di "Antonio Brandeis". Il biografo Angelo de Gubernatis spiegò la scelta: **l'artista** aveva ricevuto elogi e critiche, ma non accettò le lodi pronunciate soltanto per il suo essere donna. (Antonietta Brandeis, BioWiki)

'In both Florence and Budapest she spread her work under the name "Antonio Brandeis." Biographer Angelo de Gubernatis explained the choice: **the artist** had received both praise and criticism, but she did not accept the praise pronounced only because of her being a woman.'

## 4.2 Codification of textual referents evoking persons: forms of anthroponyms

In this section, we describe in more details the forms anthroponyms take in both BioCGPT and BioWiki when they refer to the female characters at the center of the biographies. Our analysis focuses on the three forms of anthroponyms outlined and illustrated in Table 4.

**Table 4.** Forms of anthroponyms analyzed

| | **Form of anthroponym** | **Example** |
|---|---|---|
| **Type I** | First name + Surname | Eleonora Duse |
| **Type II** | Surname | Duse |
| **Type III** | Definite article + Surname | la Duse |

Table 5 reports the quantitative data related to the presence of the three types of anthroponyms in the two subcorpora of biographies.

**Table 5.** Form and frequency of three types of anthroponyms

| | **BioFemCGPT** | **BioFemWiki** |
|---|---|---|
| **Type I** | 537 (46%) | 308 (57%) |
| **Type II** | 569 (48%) | 126 (24%) |
| **Type III** | 67 (6%) | 103 (19%) |
| **Tot.** | 1173 (100%) | 537 (100%) |

All three forms of anthroponyms are documented in the subcorpora. In BioCGPT, anthroponyms of Types I and II are used with a similar frequency (46% and 48%, respectively), while in BioWiki Type I is the most frequently occurring form (57%).

Type I is closely associated to a specific textual position: the incipit of the text, as shown in (30) and (31). At the beginning of the BioCGPT texts, Type I is indeed the only occurring form of anthroponym documented (it thus occurs 168 times in text absolute initial position, covering 31% of the data related to Type I). Type I is also the most frequently occurring form found at the beginning of the BioWiki texts. In special cases, however, the new female referent occurring in text initial position can be instantiated with more specific anthroponyms, as shown in (32), which are not part of the data reported in Table 5.

(30) [incipit] **Eleonora Duse** (1858-1924) è stata un'attrice teatrale italiana, considerata una delle più grandi attrici della storia del teatro. (Eleonora Duse, BioCGPT)

‘**Eleonora Duse** (1858-1924) was an Italian stage actress, considered one of the greatest actresses in the history of theater.’

(31) [incipit] **Liliana Segre** (Milano, 10 settembre 1930) è un'antifascista e politica italiana, superstite dell'Olocausto e testimone attiva della Shoah. (Liliana Segre, BioWiki)

‘**Liliana Segre** (Milan, September 10, 1930) is an Italian antifascist and politician, Holocaust survivor and active witness to the Shoah.’

(32) **Eleonora Giulia Amalia Duse** (Vigevano, 3 ottobre 1858 – Pittsburgh, 21 aprile 1924) è stata un'attrice teatrale italiana. (Eleonora Duse, BioWiki)

‘**Eleonora Giulia Amalia Duse** (Vigevano, October 3, 1858 - Pittsburgh, April 21, 1924) was an Italian stage actress.’

The correlation between Type I anthroponyms and text initial position is of course not surprising, as new and non-accessible textual referents need to be instantiated through denotationally clear forms, which are typically complex from a formal point of view (see the scale of linguistic complexity in § 4.1). Type I anthroponyms are also appropriate to re-instantiate a discourse referent at the beginning of a new text section, such as a new text block:

(33) **Metodo recitativo e rivoluzioni della "Divina"**
**Eleonora Duse** caratterizzò il teatro moderno perché ruppe totalmente gli schemi del teatro ottocentesco, divenuto ormai incombente su una società del tutto nuova e diversa. (Eleonora Duse, BioWiki)

**'Acting method and revolutions of the "Divine"**
**Eleonora Duse** characterized modern theater because she totally broke the mold of nineteenth-century theater, which had now become incumbent on an entirely new and different society.'

The quantitative discrepancy in the use of Type I in the two subcorpora (46% vs. 57%) can in part be explained by the fact that in BioWiki the format 'First name + Surname' is regularly used to identify the source of a quote, as shown, e.g., in (34):

(34) «chiunque sia stato comunista negli anni Cinquanta riconosce di colpo il nuovo linguaggio delle BR. Sembra di sfogliare l'album di famiglia: […]» (**Rossana Rossanda**, *Il discorso sulla Dc*, articolo apparso in prima pagina su *il manifesto*, il 28 marzo 1978) (Rossana Rossandra, BioWiki)

'"anyone who was a communist in the 1950s suddenly recognizes the new language of the RB. It is like leafing through a family album: [...]" (**Rossana Rossanda**, *Il discorso sulla Dc*, article that appeared on the front page of *il manifesto*, March 28, 1978)'

Another relevant difference between BioCGPT and BioWiki that can be observed from the data in Table 5 concerns the frequency of Type III anthroponyms. Referring to a woman with the form 'Definite article + Surname' is three times more frequent in BioWiki than in BioCGPT (19% vs. 6%). This suggests that generated biographies are written in both a more modern[17] and politically correct language than Wikipedia ones, as Type III anthroponyms are considered sexist by many Italian scholars (see, among others, Sabatini 1993 and Viviani 2011).

(35) **La Duse** fu molto influenzata dal lavoro di alcuni importanti registi e attori del suo tempo, tra cui André Antoine e Konstantin Stanislavskij. (Eleonora Duse, BioCGPT)

[17] D'Achille (2016: 176) considers the lack of definite article in front of female surnames a distinctive feature of contemporary Italian.

‘**[The] Duse** was greatly influenced by the work of some important directors and actors of her time, including André Antoine and Konstantin Stanislavsky.’

(36) **La Duse** aveva amicizie con alcune delle personalità più note dell'epoca, come la scrittrice Sibilla Aleramo e la danzatrice Isadora Duncan. (Eleonora Duse, BioWiki)

‘**[The] Duse** had friendships with some of the most well-known personalities of the time, such as writer Sibilla Aleramo and dancer Isadora Duncan.’

## 4.3 Expressing ‘Constant Topical Progression’

In texts where the discourse referent of a first Utterance (U1) functions as Topic (defined, following Lambrecht 1994, as “what we talk about” within a semantic proposition) and is resumed as topical in the following Utterance(s) of the text (U2, U3…), we find a special textual pattern called ‘Constant Topical Progression” (see Ferrari et al. 2008: 152-175; Ferrari & De Cesare 2009). The abstract form of this special pattern (in short CTP) is shown in Table 6.

**Table 6.** Constant Topical Progression (CTP)

| | **Topic: “what we talk about”** | **Comment: “what we say about the Topic”** |
|---|---|---|
| **U1** | Topic 1 | Comment1 |
| **U2** | Topic 1 | Comment2 |
| **U3** | Topic 1 | Comment3 |
| **U4** | … | … |

In this section, we look at how Topics corresponding to female textual referents are linguistically codified in BioCGPT and BioWiki when they enter a CTP. We are mainly interested in the form of the Topic occurring in the second Utterance (U2) of the pattern.

After its text instantiation through an anthroponym of Type I (‘First name + Surname’), the most natural way of codifying the Topic of the second Utterance entering a CTP is with a syntactically and semantically more economic linguistic expression. In both BioCGPT and BioWiki, this expression typically coincides with a null subject. The following examples from our two subcorpora illustrate the pattern ‘anthroponym I > null subj.’, where the null subject functions as anaphor to resume a topical discourse referent present in an Utterance belonging to the same text block (as in ex. 37) or to a different text block (as shown in ex. 38):

(37) [incipit] **Dacia Maraini** è una scrittrice, sceneggiatrice e attrice italiana, nata a Fiesole il 14 dicembre 1936. //**U1** [**null subj.**] È figlia di scrittori e giornalisti, e ha studiato filosofia all'Università di Firenze. //**U2** (Dacia Maraini, BioCGPT)

'[incipit] **Dacia Maraini** is an Italian writer, screenwriter and actress, born in Fiesole on December 14, 1936. //**U1** [**null subj.**] is the daughter of writers and journalists, and studied philosophy at the University of Florence. //**U2**'

(38) [incipit] **Dacia Maraini** (Firenze, 13 novembre 1936) è una scrittrice, poetessa e saggista italiana. //**U1** [table of content]
**[null subj.]** Nasce a Firenze nel 1936, primogenita dell'antropologo, orientalista e scrittore fiorentino Fosco Maraini e della pittrice e gallerista palermitana Topazia Alliata, quest'ultima appartenente al ramo […]. //**U2** (Dacia Maraini, BioWiki)

'[incipit] **Dacia Maraini** (Florence, November 13, 1936) is an Italian writer, poet and essayist. //**U1** [table of content]
[**null subj**.] was born in Florence in 1936, the eldest child of the Florentine anthropologist, orientalist and writer Fosco Maraini and the Palermitan painter and gallerist Topazia Alliata, the latter belonging to the branch [...]. //**U2'**

In both subcorpora, another natural way of expressing a topical textual referent that occurs in the second Utterance of a CTP is by means of the surname alone, in the pattern 'anthroponym I > anthroponym II', as shown in (39):

(39) **Rita Levi-Montalcini** è stata una biologa italiana di origine ebrea, premio Nobel per la Medicina nel 1986 insieme a Stanley Cohen per la scoperta dei fattori di crescita nervosa. //**U1** Nato [sic] a Torino nel 1909, **Levi-Montalcini** ha trascorso la sua infanzia e la sua adolescenza a Torino, dove si è laureata in medicina nel 1936. //**U2** (Rita Levi-Montalcini, BioCGPT)

'**Rita Levi-Montalcini** was a Jewish-born Italian biologist who won the 1986 Nobel Prize in Medicine along with Stanley Cohen for the discovery of nerve growth factors. //**U1** Born in Turin in 1909, **Levi-Montalcini** spent her childhood and adolescence in Turin, where she graduated in medicine in 1936. //**U2**'

Besides these common CTP patterns, in BioCGPT we come across an unusual way of expressing the topical referent occurring in the second Utterance. This is the case, for instance, when the topical referent continues to be evoked by means of the same, complex anthroponomic form 'Name + Surname", as in (40) and (41). This pattern is unnatural as it involves an overly costly linguistic form, which over-specifies a discourse referent that is not only given but also easily accessible to the reader (for details on the linguistic means that pattern with different types of topical referents, see Andorno 2003: 53):

(40) [incipit] **Nilde Iotti** è stata una figura importante nella lotta per i diritti delle donne in Italia. //**U1** Nata a Bologna nel 1921, **Nilde Iotti** ha vissuto in una società in cui le donne avevano poche opportunità e la loro voce era spesso ignorata. //**U2** (Nilde Iotti, BioCGPT)

'[incipit] **Nilde Iotti** was an important figure in the struggle for women's rights in Italy. //**U1** Born in Bologna in 1921, **Nilde Iotti** lived in a society in which women had few opportunities and their voices were often ignored. //**U2**

(41) **Elsa Morante** è stata una delle più importanti scrittrici italiane del XX secolo, nata a Roma il 18 agosto del 1912 e morta nella stessa città il 25 novembre del 1985. //**U1** Autrice di romanzi, raccolte di poesie e di racconti, **Elsa Morante** è stata una figura di spicco nella letteratura italiana del dopoguerra. //**U2** (Elsa Morante, BioCGPT)

'**Elsa Morante** was one of the most important Italian writers of the 20th century, born in Rome on August 18, 1912, and died in the same city on November 25, 1985. //**U1** An author of novels, collections of poetry, and short stories, **Elsa Morante** was a leading figure in postwar Italian literature. //**U2**'

This pattern has not been found in the Wikipedia biographies.

## 5. Logical dimension of textual organization: the case of additive connectives

In this last section, we focus on logical connectives (for a definition, see, *inter alia*, Ferrari 2010 and Ferrari & Pecorari 2021: 7-11), a category including "morphologically invariable linguistic forms that offer instructions on how to link the events evoked by the text or the linguistic acts of textual composition through logico-argumentative relations such as cause, consecution, reformulation, exemplification, opposition, etc." (Ferrari & Pecorari 2021: 7, our translation).

Connectives are not very numerous in the two subcorpora analyzed. One of the best represented semantic categories of connectives are the ones marking temporal relations between the events codified in the text, such as *dopo* 'after' in *Dopo la fine della Seconda Guerra Mondiale* 'After the second world war' (this ex. is included in the biography on Rossana Rossanda generated by ChatGPT-3.5 on 20.4.2023). The sparse presence of connectives in the data can be explained by the text genre analyzed. In biographies, connections between semantic propositions within an Utterance or across Utterances are primarily additive. In contrast to complex relations, such as consecution (expressed by It. *dunque* 'consequently, hence') or concession (*benché* 'although'), addition is a simple semantic relation that can easily be inferred from the contents conveyed in the text. As a result, the additive relation generally does not need to be expressed linguistically by a connective (for details on the properties of additive relations, see De Cesare 2021b).

In BioCGPT and BioWiki, we nevertheless find different additive expressions. The most frequent additive connective in the data is *inoltre* 'in addition', 'moreover', 'also'[18]. Table 7 reports the quantitative data on *inoltre*, providing information on two parameters related to its use: (i) its Utterance distribution (initial vs. internal) and (ii) its occurrence with and without commas.

[18] While E. *also* can function as an additive connective, It. *anche* cannot, which is why we do not consider it here.

**Table 7.** It. connective *inoltre*

| | **BioFemCGPT**<br>**(59 occ. /777 pmt)** | **BioFemWiki**<br>**(50 occ./592 pmt)** |
|---|---|---|
| **Utterance initial** | 41 occ. (70%)<br>Inoltre, (39 occ./95%)<br>Inoltre (1 occ.)<br>Inoltre (1 occ. unclear[19]) | 15 occ. (30%)<br>Inoltre, (12 occ./80%)<br>Inoltre (3 occ.) |
| **Utterance internal** | 18 occ. (30%)<br>inoltre (17 occ./94%)<br>, inoltre, (1 occ.) | 35 occ. (70%)<br>inoltre (33 occ./94%)<br>inoltre, (1 occ.)<br>, inoltre, (1 occ.) |

The data in Table 7 clearly shows that the connective *inoltre* appears in a complementary distribution in the two subcorpora. In BioCGPT, it occurs predominantly in Utterance initial position (70%), while in BioWiki it is more commonly found in Utterance internal position (70%). Moreover, when it is used in Utterance initial position, *inoltre* is more systematically followed by a comma in BioCGPT than in BioWiki (95% vs. 80%, respectively). Examples of *inoltre* in Utterance initial position drawn from the two subcorpora are the following:

(42) **Inoltre**, la Hack ha ricevuto numerosi premi e riconoscimenti per il suo lavoro scientifico e divulgativo, tra cui la Medaglia d'Oro del Presidente della Repubblica, l'Ordine al Merito della Repubblica Italiana e l'Orione d'Oro dell'Unione Astrofili Italiani. (Margherita Hack, BioCGPT)

‘**In addition**, Hack has received numerous awards and honors for her scientific and popular work, including the Gold Medal of the President of the Republic, the Order of Merit of the Italian Republic and the Golden Orion of the Italian Astrophilic Union.’

(43) **Inoltre**, vinse il Premio internazionale Saint-Vincent e il Premio Internazionale Feltrinelli, conferitole nel 1969 dall'Accademia dei Lincei. (Rita Levi-Montalcini, BioWiki)

‘**In addition**, she won the Saint-Vincent International Prize and the Feltrinelli International Prize, awarded to her in 1969 by the Accademia dei Lincei.’

[19] In this occurrence, *inoltre* is the last token of a text that is interrupted during its generation.

(44) Nel tour di *Elisa Heart Alive Tour* tenutosi in aprile a maggio del 2010 vengono proiettati filmati ritraenti Paola Cortellesi prima e durante la versione italiana di *Mad World*. **Inoltre** è presente di persona alla data di Roma al PalaLottomatica. (Paola Cortellesi, BioWiki)

'On the *Elisa Heart Alive Tour* held in April to May 2010, footage is shown portraying Paola Cortellesi before and during the Italian version of *Mad World*. **In addition** she is present in person at the Rome date at PalaLottomatica.'

When it is used Utterance internally, *inoltre* is usually not accompanied by commas. This is the case in both subcorpora. What differs between them, though, is the syntactic position occupied by the connective. In BioCGPT, *inoltre* occupies a rather fixed sentence slot. In 16 cases out of 18, it occurs after the auxiliary of a complex verb form, as in ex. (45) and (46); in the two other occurrences, we find it once before a complex verb form (in this case, as can be observed in ex. 47, it is accompanied by commas) and once after the main verb form (see the copula *è* 'is' in ex. 48):

(45) Nel corso della sua carriera, Margherita Hack ha contribuito a importanti scoperte scientifiche, come la scoperta di un nuovo tipo di stella, le pulsar. Ha **inoltre** svolto attività di ricerca su vari temi, come le stelle binarie, le supernove, le galassie e la struttura dell'universo. (Margherita Hack, BioCGPT)

'Throughout her career, Margherita Hack has contributed to important scientific discoveries, such as the discovery of a new type of star, pulsars. She has **also** done research on various topics, such as binary stars, supernovas, galaxies and the structure of the universe.'

(46) Meloni si è **inoltre** impegnata nella lotta contro la corruzione e contro la criminalità organizzata, e ha promosso la creazione di un sistema giudiziario più efficiente e trasparente. (Giorgia Meloni, BioCGPT)

'Meloni **also** pledged to fight corruption and organized crime, and promoted the creation of a more efficient and transparent judicial system.'

(47) Morante, **inoltre**, ha sempre dimostrato una forte attenzione verso la condizione femminile e verso la lotta per l'emancipazione delle donne, come dimostra il suo romanzo "Menzogna e sortilegio". (Elsa Morante, BioCGPT)

‘Morante, **in addition**, has always shown a strong concern for the condition of women and the struggle for women's emancipation, as evidenced by her novel "Lies and Sorcery."’

(48) Il partito è **inoltre** molto attento alla difesa della famiglia, della tradizione e della cultura italiana. (Giorgia Meloni, BioCGPT)

‘The party is **also** very focused on the defense of the Italian family, tradition, and culture.’

By contrast, in BioWiki, *inoltre* occurs most often (in 23 out of 35 occ., i.e. 66% of the cases) after the main verb (as in ex. 49 and 50); more marginally (in 8 and 4 occ., respectively), we also find it after the auxiliary of a complex verb form (see ex. 51 and 52) and before the subject (as in ex. 53):

(49) Propone **inoltre** l'elezione diretta del presidente della Repubblica e in generale è favorevole a una riforma Costituzionale in senso presidenzialista. (Giorgia Meloni, BioWiki)

‘She **also** proposes the direct election of the President of the Republic and generally supports a Constitutional reform in a presidentialist direction.’

(50) La senatrice è stata **inoltre** una delle promotrici del "Progetto Genomi Italia", che intendeva realizzare un'infrastruttura dedicata a un progetto di genomica su scala nazionale applicato alla sanità pubblica […]. (Elena Cattaneo, BioWiki)

‘The senator was **also** one of the promoters of the "Genomes Italy Project," which aimed to set up an infrastructure dedicated to a nationwide genomics project applied to public health […].’

(51) Ha proposto il rafforzamento dell'istituto della legittima difesa in modo che essa sia considerata "sempre legittima". Ha **inoltre** proposto la castrazione chimica per stupratori recidivi e pedofili. (Giorgia Meloni, BioWiki)

‘She proposed strengthening the institution of self-defense so that it would be considered "always legitimate." She has **also** proposed chemical castration for repeat rapists and pedophiles.’

(52) La senatrice si è **inoltre** opposta con fermezza all'abolizione del tema di ambito storico dall'esame di maturità. (Liliana Segre, BioWik)

'The senator has **also** staunchly opposed the abolition of the history subject from the baccalaureate exam.'

(53) Secondo alcune ricostruzioni giornalistiche **inoltre** i nonni di Giorgia Meloni sarebbero gli attori e doppiatori Zoe Incrocci e Nino Meloni. (Giorgia Meloni, BioWiki)

'According to some journalistic reconstructions **moreover** Giorgia Meloni's grandparents are actors and voice actors Zoe Incrocci and Nino Meloni.'

## 6. Conclusion

This contribution is a proof of concept for a larger research endeavor investigating the quality of textual vs. visual outputs generated by LLMs. While the quality of texts can be measured at different levels (content, grammar, word choice, punctuation etc.), we focused on textual parameters, which are often neglected but have a great and proven heuristic value.

Based on the results of qualitative and quantitative small-scale case studies, conducted on a self-assembled comparable corpus of biographies (BioCGPT, a machine-generated dataset, and BioWiki, a human-written one), we uncovered important textual differences between generated and human-written biographies. These differences allow highlighting several properties of the generated biographies:

1. These biographies include a restricted set of forms belonging to a larger paradigm. As far as text segmentation is concerned, we observed that ChatGPT's output relies on a smaller subset of punctuation marks than the biographies available on Wikipedia. This is evident in relation to the full stop: there is *repetitio* over *variatio* and thus also, more generally, a lack of sensitivity towards stylistic matters. In relation to the codification of female referents, too, ChatGPT's output relies on a more restricted paradigm of forms (definite descriptions are not present). This could be due to a lack of world knowledge and difficulty to handle complex anaphoric expressions.
2. Generated biographies also include a restricted set of functions (see the cases of the exclamation mark and parenthetical contents).
3. Generated biographies lack diversity in the points of view reported in the texts. They do not include the point of view of different human-beings, in particular in the form of direct reported speech.

4. Generated biographies include atypical textual patterns. We observed this in the expression of Constant Topical Progression as well as the Utterance distribution of the additive connective *inoltre* ‘in addition’.

In conclusion, the following general statements on ChatGPT’s output can be made. First, taken individually, texts generated by GPT-3.5 appear to be written well. It is only when analyzing a sample of generated texts (we looked at 168) and comparing it to a similar sample of human-written texts that the differences in quality become apparent. Second, the quality of texts generated by ChatGPT-3.5 differs in many respects from the quality of similar texts written by human-beings. Overall, generated texts appear to be repetitive, monotonous, and monophonous. At the same time, and this is the third observation, generated texts present positive characteristics. In the realm of anthroponyms, for instance, generated biographies include a lower percentage of forms considered ‘sexist’ (*la Duse* ‘Art. + Surname’). Tentatively, they can thus be considered as being more politically correct and in line with the recent developments of contemporary Italian.

To further deepen our understanding of texts generated by LLMs, more research is necessary in this new line of inquiry. Specifically, we need more research on languages other than English (e.g., on Italian and other Romance languages), based on texts generated by the latest version of different LLMs (such as GPT-4, BERT, Llama 2, BLOOM etc.) and based on larger corpora of generated texts (yet to be constructed). Moreover, future studies ought to consider the question of prompt-engineering more systematically (the assessment of the quality of generated texts should not be solely based on zero-shot learning; more refined prompting is needed) and should consider a broader set of text genres (beyond biographies). Finally, we ought to go beyond the mere description of the data. The results should be explained by considering technical and computational issues related, for instance, to the construction and training data of LLMs. Some textual features, such as the Utterance initial position of connectives followed by a comma, seem to reproduce typical English patterns and could thus be explained on the underlying English data on which GPT-3.5 has been trained.

## Acknowledgments

I would like to thank two anonymous reviewers for their comments and suggestions. I also wish to thank Tom Weidensdorfer for collecting automatically the texts making up the BioWiki subcorpus. Last but not least, I am indebted to Claudia Rausch, who generated the biographies on which the BioCGPT subcorpus has

been constructed and who provided me a first set of results on the forms of anthroponyms used in the corpus data.

## References

Andorno, C. 2003. *Linguistica testuale. Un’introduzione*. Roma: Carocci.

Bang, Y. et al. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. https://doi.org/10.48550/arXiv.2302.04023 (accessed July 27, 2023).

Barbaresi, A. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. *Proceedings of ACL/IJCNLP 2021: System Demonstrations*, 122-131.

Cresti, E. & Panunzi, A. 2013. *Introduzione ai corpora dell’italiano*. Bologna: il Mulino.

D’Achille, P. 2016. Architettura dell'italiano di oggi e linee di tendenza. In S. Lubello (ed.), *Manuale di linguistica italiana*. Berlin: De Gruyter, 165-189.

De Cesare, A.-M. 2011. Espositivi, testi. In R. Simone (ed.), *Enciclopedia dell’italiano*. Roma: Treccani, 1474-1478.

De Cesare, A.-M. 2021a. Répétitions et variations des textes générés. Une analyse linguistique basée sur un corpus d’articles financiers rédigés en français. *CHIMERA. Romance Corpora and Linguistic studies* 8: 79-108. https://revistas.uam.es/chimera/article/view/15158 (accessed October 18, 2023).

De Cesare, A.-M. 2021b. Autour de la relation d’ajout. Définition et connecteurs adverbiaux du français. In A. Ferrari & F. Pecorari (eds), *(Nuove) Prospettive di analisi dei connettivi*. SILTA L/1, 67-82.

De Cesare, A.-M. (in press). Il *Movimento Testuale* seriale: forma prototipica e manifestazione nei testi generati da ChatGPT. In L. Fesenmeier, S. Dessì Schmid & T. Paciaroni (eds), *Atti del XII Convegno dell’Associazione Germanofona degli Italianisti*, Munich (March 2022): Wissenschaftliche Buchgesellschaft.

De Cesare, A.-M., Eliasson, E. & Weidensdorfer, T. 2023. La coesione testuale (basata sulle strutture verbali) nella scrittura generata in ambito finanziario. Italiano e francese a confronto. In A.-M. De Cesare et al. (eds), *Forme della scrittura italiana contemporanea in prospettiva contrastiva. La componente testuale*. Firenze: Cesati, 71-86.

Ferrara, E. 2023. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. https://doi.org/10.48550/arXiv.2304.03738 (accessed June 15, 2023).

Ferrari, A. 2010. Connettivi. In R. Simone (ed.), *Enciclopedia dell’italiano*. Roma: Treccani, 271-273.

Ferrari, A. 2014. The Basel Model for paragraph segmentation: the construction units, their relationships and linguistic indication. In S. Pons Bordería (ed.), *Discourse Segmentation in Romance Languages*. Amsterdam/Philadelphia: John Benjamins: 23-54.

Ferrari, A., Cignetti, L., De Cesare, A.-M., Lala, L., Mandelli, M., Ricci, C. & Roggia, E. 2008. *L’interfaccia lingua-testo. Natura e funzioni dell’articolazione informativa dell’enunciato*. Alessandria: Edizioni dell’Orso.

Ferrari, A. & De Cesare, A.-M. 2009. La progressione tematica rivisitata. *Vox Romanica* 68: 98-128.

Ferrari, A. & Pecorari, F. (2021). Introduzione. Denominazioni, definizioni, prospettive di analisi. In A. Ferrari & F. Pecorari (eds), *(Nuove) Prospettive di analisi dei connettivi*, SILTA L/1, 7-13.

Garrido-Muñoz, I., Martínez-Santiago, F. & Montejo-Ráez, A. 2023. MarIA and BETO are sexist: evaluating gender bias in large language models for Spanish. Lang Resources & Evaluation. https://doi.org/10.1007/s10579-023-09670-3 (accessed September 25, 2023)

Kotek, H., Dockum, R., Sun, D. Q. 2023. Gender bias and stereotypes in Large Language Models. https://aps.arxiv.org/pdf/2308.14921.pdf (accessed June 20, 2023)

Lambrecht, K. 1994. *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Nissim, M. & Pannitto, L. 2022. *Che cos'è la linguistica computazionale?* Roma: Carocci.

Sabatini, A. 1993. *Il sessismo nella lingua italiana. Commissione Nazionale per la parità e le pari opportunità tra uomo e donna*. Presidenza del Consiglio dei Ministri.

Tavosanis, M. 2021. L'ideologia linguistica e le pratiche di Wikipedia in lingua italiana. In A. P. Alamán, F. Ruggiano & O. Walsh (eds), *Le ideologie linguistiche: lingue e dialetti nei media vecchi e nuovi*. Berlin: Lang, 413-434.

Viviani, A. 2011. Cognomi, articolo con. In R. Simone (ed.), *Enciclopedia dell'italiano*. Roma: Treccani, https://www.treccani.it/enciclopedia/articolo-con-prontuario-cognomi_%28Enciclopedia-dell%27Italiano%29/ (accessed June 20, 2023)

Werlich, E. 1975. *Typologie der Texte. Entwurf eines Textlinguistischen Modells zur Grundlegung einer Textgrammatik*. Heidelberg: Quelle & Meyer.

*Permanent links to the Wikipedia entries* (all accessed on September 18 and 19, 2023)

https://it.wikipedia.org/w/index.php?title=Dacia_Maraini&oldid=135501906.
https://it.wikipedia.org/w/index.php?title=Rosy_Bindi&oldid=135491102.
https://it.wikipedia.org/w/index.php?title=Grazia_Deledda&oldid=135390496.
https://it.wikipedia.org/w/index.php?title=Laura_Boldrini&oldid=135253541.
https://it.wikipedia.org/w/index.php?title=Elena_Cattaneo&oldid=134858998.
https://it.wikipedia.org/w/index.php?title=Paola_Cortellesi&oldid=135495025.
https://it.wikipedia.org/w/index.php?title=Eleonora_Duse&oldid=134989604.
https://it.wikipedia.org/w/index.php?title=Oriana_Fallaci&oldid=135283939.
https://it.wikipedia.org/w/index.php?title=Giorgia_Meloni&oldid=135501878.
https://it.wikipedia.org/w/index.php?title=Rita_Levi-Montalcini&oldid=135460237.
https://it.wikipedia.org/w/index.php?title=Liliana_Segre&oldid=135498117.
https://it.wikipedia.org/w/index.php?title=Antonietta_Brandeis&oldid=129060540.