LLM referential chain generation. A qualitative case study based on Italian biographies produced by GPT-4*

Anna-Maria De Cesare (Dresden)

Abstract

The goal of the present contribution is to shed light on the textual properties of written outputs generated by large language models by focusing on the referential dimension of textual organization. To gain insights on this aspect, we analyze the properties of the referring expressions forming the main referential chain running through biographies, which typically correspond to the personality on which the text is centered. Based on an empirical qualitative corpus-driven analysis of 30 biographies generated in Italian by GPT-4, we describe (i) the forms of the linguistic expressions codifying the discourse referents building the main referential chain of the biographies; (ii) the degree of complexity of these linguistic expressions, verifying if they match the degree of cognitive accessibility of the discourse referents; finally, on a textual level, (iii) the architectures that these chains form and the distribution of the chain's rings in the textual units composing the biography. Our analysis reveals that the main referential chain building the analyzed biographies is generally well-formed but very simple and relies on repetitive textual patterns. At a micro-textual level, we find marked textual patterns such as the over-codification of a discourse referent as well as cases of over-segmentation of semantically and pragmatically compact textual units.

1 Introduction

Texts generated by large language models (LLMs), such as the ones belonging to the GPT family (GPT-3.5, GPT-4 and GPT-4o), are the product of a "probabilistic system in which the next word in a text is selected based on the context in which it occurs. As such, LLMs are aligning words according to expected patterns […]." (Backus et al. 2023: 303f.). The so-called *context window* that the system relies on depends on the model: it ranges from a few thousand tokens up to 128,000 tokens.[1] For this study we used the model gpt-4-0613, which relies on a

* I wish to thank two anonymous reviewers for their comments and suggestions. A special thank you also to Valerie Hekkel for writing the code that served to create the analyzed text sample and Tom Weidensdorfer for both generating the text sample and designing Figure 2 (provided in the conclusions).

[1] Information on the size of the context window and generated output of the GPT-models can be found on the OpenAI Platform.

context window of 8,192 tokens, a number that exactly matches the size of the output.[2] The text reproduced below (corresponding to a biography generated in English with GPT-4 on September 12, 2024) allows to grasp the size of both the context window and the generated output. Note that in this contribution we will always reproduce the text as it has been generated, including its grammatical, stylistic and information flaws.

(1) Laura Boldrini is an important figure in Italian politics, renowned for her relentless advocacy for human rights, women's rights, and social justice. Born on April 28, 1961, in Macerata, Italy, she has dedicated her life to public service, both on a national and international level, cementing her legacy as one of Italy's most influential women.

Laura Boldrini earned a degree in law from Sapienza University of Rome in 1985. With an innate passion for social issues and politics, she started her career in Rome as a public relations officer for World Food Program (WFP), a branch of the United Nations focused on global hunger and food security. She then relocated to London, working as the spokesperson for the United Nations High Commissioner for Refugees (UNHCR), where she remained for over a decade. During this period, she gained international accolades for her efforts to highlight global refugee issues.

In 2013, Boldrini took a significant step in her career, transitioning from the global humanitarian sector to the national political landscape. She was elected to Italy's Chamber of Deputies on the left-wing coalition Italy Common Good's list. Known for her strong principles and dedication, she was soon elected President of the Chamber of Deputies, becoming the third woman in Italian history to hold this highly esteemed position.

During her tenure as President, she fervently advocated for the equal representation of women in politics, protection of migrants, and improving ethical journalism standards. Known for her uncompromising stances, Boldrini has been a distinct voice in Italian politics, often leading to controversial debates.

After her five-year term in the presidency ended in 2018, Boldrini remained politically active, continuing to advocate for the causes close to her heart. She voiced her concerns about the rise of anti-immigration sentiments in Europe and highlighted the importance of empowering women in political and social spheres.

Boldrini received significant recognition for her work, including honorary citizenship of Lampedusa and Linosa in 2014 for her work with migrants and refugees, and the Minerva Award in 2018 for her contributory role in Italian civil society.

Despite facing various challenges and controversies throughout her career, Laura Boldrini remains firm in her dedication to social justice and human rights, that is deeply rooted in her early work with the UN and continues to inform her political career. A trailblazer in both national and international circles, Boldrini's impact is clear, highlighting the significant role influential women can play in politics and human rights activism.

(OpenAI, GPT-4-gpt-4-0613, 12.09.2024)

The goal of the present contribution is to describe how a probabilistic system such as GPT-4 handles textual relations, i. e. semantic and pragmatic relations between adjacent sentences or

[2] In NLP and specifically in relation to the training data of LLMs, a token can be a word, part of a word (not necessarily coinciding with a morpheme), a character, a punctuation mark, symbol, number etc. If we consider that a token can also correspond to part of a word or even a character, it is clear that its value greatly differs from the one used as benchmark in corpus linguistics, in which a token corresponds to a graphically autonomous element (a word, punctuation mark, symbol etc. but not part of a word). According to the corpus query platform Sketch Engine, the entire text reproduced in (1) consists of only 459 tokens.

between text blocks. Our starting point is the assumption that textual patterns (i. e. "semantico-pragmatic connections that weave the content of the discourse into a whole"; Ferrari 2014b: 24) are less predictable (because they are not determined by a fixed set of rules) and more difficult to reproduce than relations that are governed by grammatical rules within a syntactic clause. The following claim by Goldstein et al. 2023 allows to understand the general challenges LLMs must face to produce good quality outputs, including at the textual level:

> While impressive, current generative language models [they referred to GPT-3] have many limitations. **Even the most sophisticated systems struggle to maintain coherence over long passages**, have a tendency to make up false or absurd statements of fact, and are limited to a generation length of about 1,500 words.
>
> (Goldstein et al. 2023: 18; emphasis from author)

Given the complexity of the question at hand (How do we measure coherence?), in this study we only focus the attention on the form of referential chains built by preverbal referring expressions (i. e., "Any expression which serves to identify any thing, process, event, action, or any other kind of individual or particular", Searle 1969: 27), such as *Laura Boldrini* and *she*, in the first two sentences of ex. (1). To understand the properties of LLM generated referential chains, we provide a descriptive empirical qualitative case study, based on a fine-grained analysis of a small but representative corpus of biographies generated in a single language, namely Italian. We chose to analyze short biographies because this text genre tends to be built on one main referential chain, related to one discourse referent, *viz.* the personality on which the biography is centered (in ex. 1, Laura Boldrini).

The research questions we aim to address are the following:

1. What referring expressions (full name, pronoun etc.) are used to codify the discourse referent building the main referential chain of the LLM generated biographies? Are they appropriate in terms of register?
2. Does the degree of complexity of the referring expressions codifying the main discourse referent reflect its degree of cognitive accessibility? Are there cases of mismatches, either in the form of over- or under-codification of these referents?
3. What forms do these chains have? Can we identify recurrent textual patterns, i. e. typical LLM generated referential chains?

The present contribution starts by defining the main concepts used in the study, namely *referential chain, referring expression, discourse referent* and *cognitive accessibility* (chapter 2); it then describes the text sample generated as well as the annotation procedure used to label different properties of the referential chain composing a corpus of 30 biographies (chapter 3). In the third part, it reports the main results obtained in relation to the form of the preverbal referring expressions and highlights unusual patterns in a referential chain based on the accessibility of the main discourse referent (chapter 4). In the conclusions, it provides the answers to the research questions and outlines further avenues of research (chapter 5).

2 Referential chains: A formal and functional definition

2.1 Referential chains, referring expressions, discourse referents and cognitive accessibility

Referential chains belong to the referential dimension of textual organization (cf. Ferrari 2014a: 118). This dimension concerns the ways a text evokes the world (real or imagined), which is characteristically composed of individual entities (people, animals, things, abstract entities, properties) and events of various kinds (actions, states, processes). The individual entities evoked in a text correspond to *discourse referents*, i. e. conceptual objects (cf. Andorno 2003: 27–29). Discourse referents (i. e. conceptual entities created in discourse), in turn, are referred to by *referring expressions*, corresponding to linguistic forms showing different degrees of complexity. In the first part of the biography reproduced in (2), based on ex. (1), there are different discourse referents identified by various referring expressions, some of which are highlighted in bold:[3]

(2) **Laura Boldrini** is an important figure in **Italian politics**, renowned for **her** relentless **advocacy** for **human rights**, women's rights, and social justice. Born on **April** 28, 1961, in **Macerata**, **Italy**, she has dedicated **her life** to **public service**, both on a national and international level, cementing **her legacy** as one of Italy's most influential women.

Within the referential dimension of textual organization, referring expressions are typically connected to each other through a wide array of semantic relations and build *referential chains* running through different units composing a text: sentences (or utterances) and text blocks. A *text block* is a specific textual unit: it is composed of a compact set of sentences (or utterances) and corresponds to a graphically autonomous textual unit separated from the rest of the text by a blank space before and/or after it and which does not start with an indent (for a definition and application to generated texts, cf. e. g. De Cesare et al. 2016: 125 and De Cesare 2023a: 185). A text block can correspond to a *paragraph*, but a paragraph can also include more than one text block.

As the notion of "chain" suggests, a referential chain is composed by a series of rings, as shown in (4) based on the referring expressions highlighted in bold in (3). In this study, we only consider the referring expression(s) occurring prior to the main verb of each sentence, which tend(s) to correspond to a single discourse referent, namely the main personality of the biography.

(3) **Laura Boldrini** is an important figure in Italian politics, renowned for her relentless advocacy for human rights, women's rights, and social justice. Born on April 28, 1961, in Macerata, Italy, **she** has dedicated her life to public service, both on a national and international level, cementing her legacy as one of Italy's most influential women.
Laura Boldrini earned a degree in law from Sapienza University of Rome in 1985. With an innate passion for social issues and politics, **she** started her career in Rome as a public relations officer for World Food Program (WFP), a branch of the United Nations focused on global hunger and food security. **She** then relocated to London, working as the spokesperson for the United Nations High Commissioner for Refugees (UNHCR), where she remained for over a decade.

[3] For the sake of clarity, the parts of the text on which the analysis is focused are highlighted in bold (or with other typographic means). Note that in this paragraph, for space reasons, the examples provided are mainly in English.

During this period, **she** gained international accolades for her efforts to highlight global refugee issues.

(4) Co-referential chain: Ring 1 (= chain opener): *Laura Boldrini*; Ring 2: *she*; Ring 3: *Laura Boldrini*; Ring 4: *she*; Ring 5: *She;* Ring 6: *she*

In the two text blocks reproduced in (3), there is one preverbal referring expression in each sentence. All the six referring expressions point to the exact same discourse referent (i. e. conceptual individual), namely Laura Boldrini. They are thus directly connected to each other and form a *co-referential chain*. This chain (in fact as every chain) is introduced by a first ring called *chain opener* (It. *capo catena*; cf. Simone 1990: 414); in (3), this special ring is codified at the beginning of the first sentence of the text and takes the form of a full name, aligning first name and last name (Laura Boldrini).

Besides referential chains built by co-referential rings, i. e. referring to the same discourse referent (cf. e. g. Laura Boldrini), there are referential chains composed of referring expressions that are semantically related but not co-referential. This is the case of the chain highlighted in bold in (5), as shown in (6):

(5) Nel 2018, **Laura Boldrini** ha perso la presidenza della Camera, ma è rimasta una deputata attiva e influente. **[Null Subject]** È nota per il suo attivismo in materia di diritti umani e per l'impegno nella tutela dei rifugiati e dei migranti.
Uno dei suoi contributi più noti alla società italiana è stata la sua campagna per la codificazione di una legge contro l'odio online e la diffamazione, nota come "Legge Boldrini", che è diventata un punto di riferimento nel dibattito italiano sui diritti digitali.

'In 2018, **Laura Boldrini** lost the presidency of the House but has remained an active and influential MP. [**Null Subject**] is known for her human rights activism and her commitment to the protection of refugees and migrants.
One of her best-known contributions to Italian society was her campaign to codify a law against online hate speech and defamation, known as the "Boldrini Law," which has become a landmark in the Italian debate on digital rights.' (Translated with DeepL, 27.02.2025[4])

(6) Referential chain: Ring 1 (= text block opener): *Laura Boldrini*; Ring 2: **[Null Subject]**; Ring 3: *Uno dei suoi contributi più noti alla società italiana.* 'One of her best-known contributions to Italian society.'

In example (5), the discourse referent opening the second text block, evoked by the complex noun phrase (NP) *Uno dei suoi contributi più noti* etc. 'One of her best-known contributions' (Ring 3 in 6), partly differs from the discourse referent of the two sentences composing the first text block, which is codified by the nominal referring expression Laura Boldrini. From a semantic point of view, the preverbal referring expressions of the two text blocks are referentially connected (see the possessive *suo* 'her' in the third sentence) but do not form a co-referential chain.

[4] All the Italian examples are translated with the help of DeepL (free version) or GPT-4o and revised where necessary.

2.2 Properties of discourse referents: linguistic expression and cognitive accessibility

The description of the referential chains in (4) and (6), based on examples (3) and (5), respectively, shows that the rings forming these chains are composed of referring expressions associated to different degrees of complexity: besides anthroponyms (in the pattern "first name + last name" e. g., in Ring 1 of ex. 3) and full lexical NPs (see Ring 3 of ex. 5), which are both formally complex, we find implicit subjects (as, e. g., in Ring 2 of ex. 5), corresponding to the simplest possible referring expression.

According to Ariel 1990's proposal, the linguistic form of a discourse referent entering a referential chain is first and foremost cognitively motivated and depends on the degree of 'accessibility' of the discourse referents (for a more detailed definition of *accessibility*, cf. Givón 1983): discourse referents that are associated to low accessibility are codified with complex referring expressions, while cognitively accessible discourse referents are codified with simple referring expressions. More specifically, the degree of accessibility of a discourse referent is related to four criteria related to the properties of the *antecedent* of the referring expression (the anaphor), i. e. the linguistic expression referring to the same discourse referent in the previous text (for details, cf. Ariel 2001). An overview of these four criteria and their correlation with the degree of accessibility of a discourse referent is provided in Table 1.

Criteria determining cognitive accessibility	**Correlation with the degree of accessibility of discourse referents**
(i) the **distance criterion**, related to the textual distance from the antecedent	Highly accessible discourse referents are textually close to their antecedents. Low accessible discourse referents are textually far from their antecedents.
(ii) the "**competition" criterion**, related to the presence of other possible antecedents in the co-text	Highly accessible discourse referents only have one possible antecedent in the co-text. Low accessible discourse referents have more than one possible antecedent in the co-text.
(iii) the **salience criterion**, related to the information status of the antecedent	Discourse referents with antecedents functioning as Topics (i. e. "What the proposition is about", in the sense of Lambrecht 1994) are more accessible than discourse referents with antecedents that are not associated with this function.
(iv) the **criterion of textual unity**, related to the textual realization of the antecedent	Discourse referents are (highly) accessible if they belong to the same textual unit (text block or thematic sub-unit) as their antecedent. Discourse referents are associated to low accessibility if they occur in a different textual unit as their antecedent.

Table 1: Discourse referents and cognitive accessibility (based on Ariel 1990 and 2001)

The examples provided in (3) and (5) illustrate the correlation between the linguistic form of a ring entering a referential chain and the degree of "accessibility" of the discourse referent: the referents codified with the pronoun *she* (Ring 2 and 5 of ex. 3) and the Null Subject (see Ring 2 in ex. 5) are linguistically simple because they are highly accessible in all the four dimensions described in Table 1: (i) their antecedent occurs in the previous sentence (i. e. occurs at close distance); (ii) it is the only antecedent available (there are no competitors in the previous co-

text); (iii) their antecedent is associated with the information function of Topic; and (iv), it occurs in the same textual unit, i. e. text block.

Conversely, the discourse referents *Laura Boldrini* (Rings 1 and 3 of ex. 3) and *Uno dei suoi contributi più noti alla società italiana* 'One of her best-known contributions to Italian society' (Ring 3 of ex. 5) are encoded with complex referring expressions (a full name and a full lexical NP, respectively) because they are associated to a low degree of accessibility. The discourse referent identified by the full lexical NP *Uno dei suoi […]* 'One of her […]' is associated to a low degree of accessibility because its antecedent is found in a different textual unit, i. e. previous text block.

In the case of the first occurrence of the discourse referent *Laura Boldrini* (Ring 1 of ex. 3), the codification of this referent with a full name is motivated by the fact that it occurs at the very beginning of the text, where there is no antecedent. In this case, we consider that the discourse referent is associated with a null degree of accessibility. The second occurrence of the full name *Laura Boldrini* (Ring 3 of ex. 3) is motivated differently. In this case, there is an antecedent, but it does not occur in the same textual unit, i. e. text block. With respect to its mention in the *incipit* of the text, the second occurrence of the discourse referent *Laura Boldrini* is thus associated to a higher degree of accessibility. What justifies its re-instanciation with a full name is not only the fact that its antecedent occurs in a previous text block. As pointed out by Korzen 2001 and Ferrari (2010: 187), at play here is an additional criterion, related to textual saliency. Accessible discourse referents (even highly accessible ones) can be encoded as complex referring expressions (such as full names or full lexical NPs) – *contra* what is expected based on Ariel's four criteria – when they occur in initial position of a new textual unit (e. g., text block).

2.3 Other criteria explaining the linguistic form of discourse referents

As acknowledged in different studies (e. g. Ferrari 2010), the linguistic encoding of a discourse referent does not depend solely on its cognitive accessibility. Highly accessible referents are sometimes encoded with complex linguistic expressions, such as anthroponyms (in the form of a full name or last name only) or full lexical NPs (with a head corresponding to a noun, as opposed to a pronoun). Besides textual saliency (described in chapter 2.2), we ought to consider other factors to explain these exceptional cases.

Two other important factors to consider are the macro-function of a text as well as the nature of the audience to which it is addressed, in particular the cognitive abilities of the addressees. According to Ferrari 2010, in educational texts (see ex. 7) and in specialized forms of written texts, such as "easy-to-read language" (ex. 8 is drawn from the no longer existing monthly magazine *dueparole*, written for people with different degrees of reading impairment), encoding a discourse referent associated to a high degree of accessibility with a complex linguistic expression, specifically by repeating *verbatim* the antecedent (*Harris* and *Joseph Ratzinger*, respectively), is a useful strategy to increase text readability and comprehension.

(7) [...] negli anni in cui Chomsky era studente, Harris$_{\text{Topic}}$ stava terminando di licenziare per la stampa un grosso volume, *Methods in Structural Linguistics*, uno dei lavori più importanti e significativi prodotti nell'ambito di tale corrente linguistica. **Harris** affidò la correzione delle bozze del libro (che uscì nel 1951) proprio a Chomsky [...]

(Graffi 2008: 11, as cited in Ferrari 2010: 181f.)

'[…] during the years when Chomsky was a student, Harris$_{\text{Topic}}$ was finishing to revise for publication a large volume, *Methods in Structural Linguistics*, one of the most important and significant works produced within that linguistic current. **Harris** entrusted the proofreading of the book (which came out in 1951) to Chomsky himself [...]'

(8) Il nuovo papa è il cardinale tedesco Joseph Ratzinger. Come papa, Joseph Ratzinger$_{\text{Topic}}$ ha scelto il nome di Benedetto XVI. **Joseph Ratzinger** è nato in Germania, a Marktl am Inn, vicino a Passau, il 16 aprile 1927.

(*dueparole* 2005, as cited in Ferrari 2010: 190)

'The new pope is the German Cardinal Joseph Ratzinger. As pope, Joseph Ratzinger$_{\text{Topic}}$ has chosen the name Benedict XVI. **Joseph Ratzinger** was born in Germany, in Marktl am Inn, near Passau, on April 16, 1927.'

According to Ferrari (2010: 190), the form of the referential chain realized in (7) and (8) – where a highly accessible discourse referent (there is only one possible antecedent, functioning as Topic, located in the previous sentence of the same text block) is codified with a complex linguistic expression – is occasionally also found in Italian newspapers as well as in Italian texts translated from English (or German). In translations, this textual pattern can be considered to be a textual calque. A well-known form of calque is the over-codification of the syntactic subject in text translated from English to Italian, as shown in (9), where we also observe the repetition of the pronominal subject *egli* 'he' at the beginning of each new sentence:

(9) Stanley Cavell rende ancora più complesso il problema di cui stiamo trattando. **Egli**$_{\text{Topic}}$ dimostra la problematicità persino dell'analisi di intenzioni concrete e dichiarate da parte di un artista. **Egli**$_{\text{Topic}}$ sostiene che un personaggio del film *la Strada* di Fellini può essere inteso come un riferimento alla leggenda di Filomena […] **Egli** immagina una conversazione con Fellini…

(*L'impero del diritto* 1896/1989: 58, as cited in Garzone 2005: 44)

'Stanley Cavell makes the problem we are dealing with even more complex. **He**$_{\text{Topic}}$ demonstrates the problematic nature of even analyzing concrete, stated intentions on the part of an artist. **He**$_{\text{Topic}}$ argues that a character in Fellini's film *La Strada* can be understood as a reference to the Filomena legend [...] **He** imagines a conversation with Fellini...'

The cases described above remain exceptional. For stylistic reasons, the Italian written tradition is still very reluctant to encode an anaphoric referring expression occurring in the vicinity of its antecedent (in particular in an adjacent sentence) with the same referring expression, as in (7), (8) and (9). Close (lexical) repetitions of coreferential discourse referents are highly stigmatized and other solutions are preferred, such as encoding an accessible discourse referent with another form of NP, allowing to provide additional information on the referent (e. g. *Joseph Ratzinger* < *Il nuovo papa tedesco* 'the new German Pope'). In Italian written texts, overall, *variatio* is much more common than *repetitio*. This means that we do not expect to find lexical repetitions of two coreferential expressions (such as *Joseph Ratzinger* < *Joseph Ratzinger*) in Italian generated texts, at least not if these texts reproduce the most frequent and natural textual patterns.

3 Data, corpus and methodology

3.1 Data: Sampling texts representing generated biographies

To answer our research questions (formulated in chapter 1), we generated a small sample of texts representing the genre 'biography', which belongs to the macro-category of 'informative text' (Sabatini 1999/2011: 195). We chose to generate biographies for several reasons: a) we know that the data on which language models are trained include a large amount of Wikipedia entries, one fifth of which describe the life and achievements of individuals who are mostly still alive (cf. Flekova/Ferschke/Gurevych 2014: 855; Navigli/Conia/Ross 2023; Baack 2024); 3% of the training data of GPT-3, for instance, are Wikipedia pages (Brown et al. 2020: 9; for details, see Figure 6 in Zhao et al. 2023: 17); b) we also know that popular searches on the internet are related to specific individuals (information on popular internet search can be obtained with GoogleTrends[5]); finally, from a linguistic point of view, c) biographies are characterized by simple text coherence; short biographies, in particular, tend to concatenate sentences in which the syntactic subject corresponds to the main person at the center of the biography (as can be observed in the generated text reproduced in ex. 1).

Our sample of biographies was generated using the closed commercial LLM GPT-4, available since March 2023 for a fee.[6] Specifically, we used the model gpt-4-0613, trained on texts collected on the internet up to September 2021. We do not know the exact nature of the training data (nor how old these texts are), but we can assume that the datasets employed include Wikipedia entries, books, newspapers, magazines, blogs, social media communication etc. (cf. Brown et al. 2020: 9 in relation to GPT-3). Moreover, we know that multilingual LLMs have been trained on datasets in which English is over-represented. There is a clear "English-bias" (cf. Ferrara 2023: 5). GPT-3, for instance, has been trained on a dataset including 93% of texts written in English (Brown et al. 2020: 14). According to another study, the training data of GPT-3 only includes 0,6% of texts written in Italian, which is one of the top five languages used to train the LLM (cf. Johnson et al. 2022: 3).

The texts we generated are biographies of 38 Italian women, active in different domains (arts, science, entertainment, sports, politics etc.). All selected profiles respond to two criteria: (i) they concern well-known and influential personalities (mostly) born in the 20th century; and (ii) there is publicly available material on each one of them on the internet, in particular a Wikipedia entry written in Italian (often in other languages as well). Here is the full list in alphabetical order:

[5] For instance, one popular trend on September 18, 2024, was related to Totò Schillaci, who died on that day.

[6] Our field of inquiry is very dynamic and new language models are released regularly. At the time we conducted this study, GPT-4 was the latest LLM available (for a fee) from OpenAI. Today, many other LLMs are on the market, including GPT4o, GPT-4.5 and the Mistral models. A follow-up study comparing the results presented in this contribution with the ones obtained using other LLMs, especially open-source ones, would be important.

1. Aulenti, Gae
2. Bindi, Rosy
3. Bo Bardi, Lina
4. Boldrini, Laura
5. Brandeis, Antonietta
6. Cattaneo, Elena
7. Cavaglieri Tesoro, Giuliana
8. Cortellesi, Paola
9. Cristoforetti, Samantha
10. Degli Esposti, Piera
11. Deledda, Grazia
12. Duse, Eleonora
13. Ercoli Finzi, Amalia
14. Fallaci, Oriana
15. Ferrante, Elena[7]
16. Fini, Leonor
17. Gianotti, Fabiola
18. Ginzburg, Natalia
19. Hack, Margherita
20. Iotti, Nilde
21. Levi-Montalcini, Rita
22. Lollini, Clara
23. Loren, Sophia
24. Magnani, Anna
25. Marabini, Anna
26. Maraini, Dacia
27. Meloni, Giorgia
28. Merini, Alda
29. Montessori, Maria
30. Morante, Elsa
31. Ortese, Anna Maria
32. Panico, Patrizia
33. Pellegrini, Federica
34. Rossanda, Rossana
35. Segre, Liliana
36. Spaziani, Maria Luisa
37. Strada, Emma
38. Vignotto, Elisabetta

The biographies were generated in Italian with GPT-4 in January 2024 and April 2024 using the API key. To generate the texts, we used a code based on a two-level, zero-shot prompting technique (cf. Chen et al. 2023). As can be observed in Table 2, the second prompt is the most important one: it identifies the text genre (biography) and includes a gap corresponding to the pattern "first name + last name". In each prompt, the gap is filled with one female anthroponym (e. g. *Scrivi una biografia su Gae Aulenti* 'Write a biography on Gae Aulenti').

Prompting level	Example
System prompt	**Il tuo compito è generare testi informativi.** 'Your task is to generate informative texts.'
User prompt	**Scrivi una biografia su {nome + cognome}.** 'Write a biography about {first name + last name}.'

Table 2: Prompting technique used to create a sample of generated biographies

Each prompt was used three times (we therefore have three texts on each one of the 38 women listed above), which allows probing the stability of the results. We can, for instance, observe that GPT-4 generated a successful biography on most of the female profiles and that the three biographies generated on the same person are very similar, both in terms of their information distribution and macro-textual architecture, e. g. in relation to the number of text blocks they include (the quantitative results are presented in Table 3 below).

There are exceptions, though. No biography was generated on Anna Marabini, Clara Lollini, Elisabetta Vignotto and Emma Strada. In these "failed biographies", the text starts with an apology (in the 1. pers. sing.) and is followed by a series of reasons motivating why the text could not be generated, as shown in the following representative example on Clara Lollini:

[7] We included Elena Ferrante in our list of female authors, even if the identity (and thus sex) of the person is not known.

(10) Mi dispiace, ma non posso fornire informazioni su Clara Lollini perché non esistono dati pubblici disponibili su questa persona. Potrebbe essere un'anonima o semplicemente il nome non è molto noto. Se Clara Lollini è un personaggio famoso o rilevante in un qualche campo, potrebbe essere necessario controllare l'ortografia del nome. Se stai cercando informazioni su un individuo privato, mi dispiace, ma non posso aiutarti a rispettare la privacy e la sicurezza dei dati personali.

(OpenAI, GPT-4-gpt-4-0613)

'I am sorry, but I cannot provide information on Clara Lollini because there is no public data available on this person. She may be anonymous or simply the name is not well known. If Clara Lollini is a famous or relevant person in some field, you may need to check the spelling of the name. If you are looking for information about a private individual, I am sorry, but I cannot help you with privacy and security of personal information.'

This output is somewhat puzzling as there is at least one Wikipedia entry (in Italian) on every woman for which GPT-4 did not generate a biography.

3.2 Description of the work corpus: BioFemIt

From the entire text sample generated by GPT-4, we selected 30 successful biographies on 30 women. We then uploaded the text sample on the Sketch Engine corpus query platform (Kilgarriff et al. 2004) to create a small corpus that we named BioFemIt. Table 3 shows the size of five relevant parameters, presented in ascending order of complexity: a) the number of tokens (for a definition, see fn. 2); b) the number of words; c) the number of sentences; d) the number of text blocks (this data was counted using the annotation tool MAXQDA); and e) the total number of texts.

Parameters	BioFemIt
Tokens	11,133
Words	9,597
Sentences	424
Text blocks	201
Texts	30

Table 3: Corpus BioFemIt

One important observation to make is that, albeit being very small, the BioFemIt corpus is representative of the typology "generated biography" and sufficient to answer our research questions. As already mentioned, and as we will see in more details below, the generated biographies are very similar in many respects, in particular in their overall macrostructure. No biography is preceded by a title and their main building blocks correspond to **text blocks** (defined in chapter 2.1).

As can be calculated based on the data in Table 3, the generated biographies composing the BioFemIt corpus are very short (they include on average 320 words and 14 sentences) and are generally subdivided into 6 to 7 text blocks. These text blocks, too, are short: each one includes, on average, two sentences. Note that the shortest text blocks are composed of a single sentence, while the longest ones include four sentences (there are only 6 cases of the latter in the whole dataset). The biography of Laura Boldrini reproduced in Table 4 is a representative example of

the form of the texts entering the BioFemIt corpus: it contains 340 words, 14 sentences and 7 text blocks: 3 texts blocks (in short: TB) are composed of two sentences (see TB1, TB2 and TB5); 2 texts blocks of 3 sentences (TB3 and TB4) and another two of one only sentence (TB6 and TB7). In the text provided in Table 4, we also highlighted the preverbal referring expression present in each sentence.

TB1	**S1/Laura Boldrini**, nata a Macerata il 28 aprile 1961, è una politica, giornalista e funzionaria italiana di rilevanza internazionale. **S2/**Dopo una lunga e dedicata carriera presso l'Alto Commissariato delle Nazioni Unite per i Rifugiati (UNHCR), **Boldrini** è entrata in politica in Italia e ha servito come Presidente della Camera dei deputati dal 2013 al 2018.
TB2	**S3/Boldrini** ha studiato lingue straniere all'Università degli Studi di Roma La Sapienza, dove ha conseguito una laurea in lingua e letteratura inglese nel 1985. **S4/**Dopo aver finito gli studi, [**Null Subject**] ha iniziato la sua carriera come giornalista, lavorando per diverse organizzazioni di stampa e produzioni televisive, coprendo una serie di temi, tra cui affari esteri e problemi sociali.
TB3	**S5/**Nel 1989, **Laura Boldrini** fu arruolata dall'UNHCR, l'organizzazione delle Nazioni Unite incaricata di proteggere e sostenere i rifugiati. **S6/**[**Null Subject**] Ha iniziato a lavorare per l'agenzia in ruoli vari in diversi paesi, tra cui Svizzera, Sudan, Kenya, Yemen, Pakistan e Bosnia. **S7/**Dal 1998 al 2012, [**Null Subject**] ha ricoperto la carica di portavoce per la regione dell'Europa meridionale di UNHCR.
TB4	**S8/Boldrini** è entrata in politica nel 2013, quando ha accettato di candidarsi alle elezioni parlamentari con la coalizione di Sinistra Ecologia Libertà (SEL). **S9/**In marzo, [**Null Subject**] è stata eletta alla Camera dei deputati e successivamente nominata Presidente della Camera, diventando la terza donna a ricoprire questo ruolo. **S10/**Durante il suo mandato, [**Null Subject**] ha lavorato per migliorare i diritti delle donne, l'integrazione dei migranti e la tutela dei diritti umani.
TB5	**S11/**Nel 2018, **Laura Boldrini** ha perso la presidenza della Camera, ma è rimasta una deputata attiva e influente. **S12/**[**Null Subject**] È nota per il suo attivismo in materia di diritti umani e per l'impegno nella tutela dei rifugiati e dei migranti.
TB6	**S13/Uno dei suoi contributi più noti alla società italiana** è stata la sua campagna per la codificazione di una legge contro l'odio online e la diffamazione, nota come "Legge Boldrini", che è diventata un punto di riferimento nel dibattito italiano sui diritti digitali.
TB7	**S14/Laura Boldrini** rimane una figura polarizzante, ma è universalmente rispettata per la sua dedizione a causa di questioni di giustizia sociale e umanitaria.

Table 4: A representative example of the BioFemIt corpus (S=Sentence; our annotation)

3.3 Data annotation

In the next step, we annotated all the preverbal referring expressions occurring in the 424 sentences included in the BioFemIt corpus. The annotation was performed using the tool MAXQDA. First, we identified all the referring expressions occurring before the main verb of each sentence composing the generated biography (as shown in Table 4). We exclusively considered the linguistic expressions coinciding with arguments of the verb, i. e. functioning as syntactic subject, direct or indirect object, encoded both as full phrases (noun or prepositional phrases), pronouns or as null arguments (as in the case of the subject).

The screenshot provided in Figure 1 shows the annotation of an entire generated biography. Examples of referring expressions occurring in preverbal position include *Laura Boldrini, Boldrini* and *Uno dei suoi contributi più noti alla società italiana* 'One of her best-known contributions to Italian society', as well as the Null Subject (in this case, we annotated the verb: see *ha iniziato* 'started'). We disregarded all the adverbial expressions occurring at the beginning of the sentence (for a definition, cf. De Cesare et al. 2018), such as *Dopo una lunga e dedicata carriera presso l'Alto Commissariato delle Nazioni Unite per i Rifugiati (UNHCR)* 'After a long and dedicated career with the United Nations High Commissioner for Refugees (UNHCR)' and *Nel 1989* 'In 1989' (these examples are found, respectively, in the second sentence and at the beginning of the third text block).

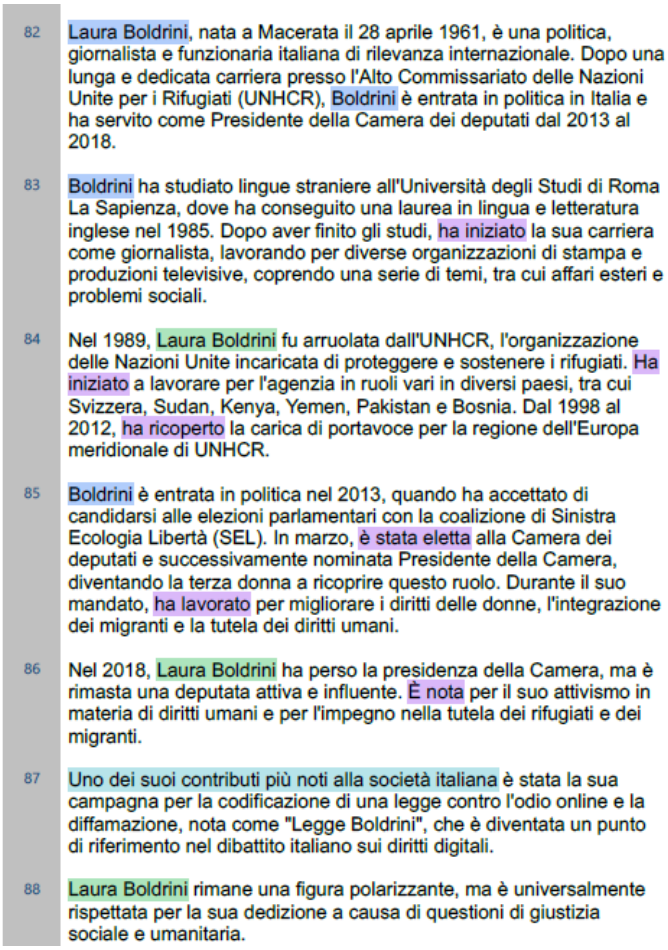

Figure 1: Manual annotation of the referring expressions in preverbal position

Besides identifying the preverbal referring expressions occurring in each sentence, we also labeled their form. The data includes six types of referring expressions, listed in Table 5, also providing a representative example of each case (note that some examples do not appear in the annotated text provided in Figure 1). The forms are ordered following the accessibility scale proposed by Ariel 1990 and discussed in chapter 2.2.

Cognitive Accessibility	**Referring expressions**	**Examples**
Low	Full name (first name + last name)	*Laura Boldrini*
	Full lexical NP[8]	*Uno dei suoi contributi più noti alla società italiana* 'One of her best-known contributions to Italian society'
	Last name	*Boldrini*
	First name	*Laura*
	Pronoun	*lei* 'she'
High	Null Subject	[Null Subject] ha iniziato '[She] started'

Table 5: Form of the preverbal referring expressions

4 Results and discussion

4.1 Form of preverbal referring expressions: general results

Table 6 provides an overview of the results, ordered according to the absolute and relative frequency of the referring expression's forms found in the 424 generated sentences of the BioFemIt corpus.

Form of preverbal referring expressions	**Abs. Freq. (tot. 428)**	**Rel. Freq.**
Null Subject	121	28%
Full lexical NP	90	21%
Last name	89	21%
Full name	88	21%
First name	19	4%
Pronoun	12	3%
Article + last name	1	0,2%
No preverbal referring expression	8	2%

Table 6: Form of referring expressions in preverbal position (BioFemIt)

As shown in the last line of Table 6, the BioFemIt corpus includes 8 sentences in which there is no preverbal referring expression. These sentences either align 'Verb-Subject' word orders (as in ex. 11) or correspond to a cleft sentence (as in ex. 12), i. e. a marked syntactic construction composed of two clauses (for details on the formal structure of cleft sentences, cf. De Cesare 2017: 537, 548–557).

[8] This label coincides with the "definite description" (both long and short) of Ariel's 1990 proposal.

(11) Oltre alla sua opera in architettura, Gae Aulenti si affermò anche nel design di interni e mobili. Tra i suoi pezzi più famosi **ci sono la Tavola con ruote, una tavola bassa su ruote di vetro e metallo, e la lampada Pipistrello**, divenuta un simbolo del design italiano e tuttora in produzione.

'In addition to her work in architecture, Gae Aulenti also made a name for herself in interior and furniture design. Lit. Among her most famous pieces **are the Tavola con ruote, a low table on glass and metal wheels, and the Pipistrello lamp**, which became a symbol of Italian design and is still in production today.'

(12) **Fu la Duse ad estendere l'idea che un attore dovesse vivere il personaggio**, portando la propria individualità e l'esperienza personale nel ruolo.

'**It was Duse who extended the idea that an actor should live the character**, bringing his or her individuality and personal experience to the role.'

The BioFemIt corpus also includes a few sentences with two preverbal referring expressions, as shown e. g. in (13), where the syntactic subject *l'Academy* is followed by the indirect object pronoun *lei* '(to) her', evoking the main discourse referent of the biography, namely *Sophia Loren*. These cases explain why the absolute frequency of the annotated segments (amounting to 428) slightly outnumbers the total amount of sentences composing the corpus (i. e. 424).

(13) Nel 1991, **l'Academy le** ha conferito un Oscar Onorario per il contributo offerto all'industria cinematografica.

'In 1991, **the Academy** awarded **her** an Honorary Oscar for her contribution to the film industry.'

The two syntactic constructions described above (i. e. sentences with special word orders in which there is no preverbal argument and sentences with two preverbal arguments) are marginal: they only occur in 2% of the data. Most sentences (98%) are constructed with the same recurrent syntactic pattern: the canonical word order aligning "Subject (which is not always expressed) Verb and Object". From a textual point of view, all the preverbal arguments (generally corresponding to the syntactic subject) tend to refer to the same discourse referent, which corresponds to the person at the center of the biography. Consequently, the generated biographies tend to be formed with a single referential chain, running through the entire text (this can be observed in the text in Table 4 as well as the English biography in ex. 1).

In the following paragraphs, we provide a detailed description of the referential chains running through the 30 biographies analyzed, focusing on the form of the preverbal referring expressions. We first look at rings of the chain codified with simple linguistic expressions, i. e. Null Subjects and pronouns, respectively (chapter 4.2); we then describe the discourse referents codified with complex referring expressions: full names and last names (chapter 4.3). In each case, we also highlight the recurrent textual patterns in which these referring expressions occur and point to mismatches between the linguistic form of a discourse referent and its degree of cognitive accessibility.

4.2 Discourse referents codified with simple referring expressions

4.2.1 Null Subjects

As shown in Table 6, in the generated biographies analyzed, roughly a third of the referring expressions (from now on, in short RE) occurring in preverbal position are codified as Null Subjects (121 occ., covering 28% of the data). Except for one occurrence (where the Null Subject corresponds to the 1. pers. pl. pronoun *noi*: [*Null Subject*] *ricordiamo* '[We] remember'), all the Null Subjects refer to the same discourse referent (DR), i. e. the person at the center of the biography.

The light morpho-syntactic encoding of the DR is motivated by its high degree of accessibility in the co-text: the antecedent of the RE occurs in the previous sentence of the same text block (with no competitor), where it also functions as Topic, i. e. as what the sentence – or, in semantic terms, the proposition – is about (according to the aboutness definition of Topic proposed by Lambrecht 1994: 121). This configuration occurs in 89% of cases in which the RE corresponds to a Null Subject (108/121 occ.). In the corpus of 30 biographies, there are 13 cases in the very first text block (TB1[9]), as in (14), where the Null Subject is realized in the second sentence and resumes the topical DR introduced in the first sentence with the full name *Giorgia Meloni*:

(14) Giorgia Meloni$_{\text{Topic}}$ è una politica italiana, nata il 15 gennaio 1977 a Roma. **[Null Subject]** È nota per essere la leader del partito di destra Fratelli d'Italia, di cui è stata co-fondatrice nel 2012. È una delle figure di spicco del panorama politico italiano. [TB1]

'Giorgia Meloni$_{\text{Topic}}$ is an Italian politician, born on January 15, 1977 in Rome. **[Null Subject]** is known for being the leader of the right-wing party Fratelli d'Italia, which she co-founded in 2012. She is one of the leading figures on the Italian political scene.' [TB1]

Besides the pattern illustrated in (14), the data includes a small number of cases (13/121 occ.: 11%) in which the Null Subject occurs in the first sentence of a new text block, as shown in (15) and (16):

(15) Margherita Hack$_{\text{Topic}}$ nasce a Firenze il 12 Giugno del 1922, in una famiglia di origini austro-ungariche. [Null Subject]$_{\text{Topic}}$ Si laurea in fisica alla Facoltà di Fisica dell'Università degli Studi di Firenze nel 1945 e nel 1948 vince una borsa di studio per gli Stati Uniti, dove ha l'opportunità di lavorare con alcuni dei più brillanti fisici del tempo.
Nel 1964 **[Null Subject]** diventa direttrice dell'Osservatorio Astronomico di Trieste, carica che ricoprì per 20 anni, diventando la prima donna in Italia a dirigere un importante osservatorio. [Null Subject] Si distingue per i suoi lavori sulla spettroscopia stellare e sulla divulgazione scientifica, oltre che per il suo impegno civile e sociale. [TB1-2]

'Margherita Hack$_{\text{Topic}}$ was born in Florence on June 12, 1922, into a family of Austro-Hungarian descent. [Null Subject]$_{\text{Topic}}$ graduated in physics from the Faculty of Physics at the University of Florence in 1945 and in 1948 won a scholarship to the United States, where she had the opportunity to work with some of the most brilliant physicists of the time.

[9] From now on, at the end of each example, we specify in square brackets where the text block(s) is/are found in the biography by providing the number of the text block(s). We also highlight in bold the RE in which we are interested and underline its antecedent (or any occurring topical RE) in the previous co-text.

'In 1964 [**Null Subject**] became director of the Trieste Astronomical Observatory, a position she held for 20 years, becoming the first woman in Italy to head a major observatory. [Null Subject] is distinguished for her work on stellar spectroscopy and science popularization, as well as for her civic and social commitment.' [TB1-2]

(16) Dal 2008 al 2011, Meloni$_{\text{Topic}}$ è stata Ministro della Gioventù nel governo Berlusconi IV, diventando la più giovane ministro nella storia della Repubblica Italiana. Nel 2012, dopo l'esperienza nel governo e una serie di divergenze all'interno del Popolo della Libertà, [Null Subject]$_{\text{Topic}}$ decide di fondare il partito Fratelli d'Italia insieme a Ignazio La Russa e Guido Crosetto.
In pochi anni, Fratelli d'Italia$_{\text{Topic}}$ è diventato uno dei principali partiti della destra italiana, con Meloni che è costantemente salita nei sondaggi. Meloni$_{\text{Topic}}$ è stata eletta per la prima volta nel 2013 al parlamento italiano e poi rieletta nel 2018.
[Null Subject] Propugna un'agenda politica di destra, affermando valori di sovranità nazionale, conservatorismo culturale e liberismo economico. [Null Subject] Si è distinta per il suo atteggiamento critico nei confronti dell'Unione Europea e della politica di accoglienza dei migranti. [TB3-5]

'From 2008 to 2011, Meloni$_{\text{Topic}}$ was Minister of Youth in the Berlusconi IV government, becoming the youngest minister in the history of the Italian Republic. In 2012, after her experience in the government and a series of disagreements within the Popolo della Libertà, [Null Subject]$_{\text{Topic}}$ decided to found the Fratelli d'Italia party together with Ignazio La Russa and Guido Crosetto.
In just a few years, Fratelli d'Italia$_{\text{Topic}}$ became one of the main parties of the Italian right, with Meloni steadily rising in the polls. Meloni$_{\text{Topic}}$ was first elected to the Italian parliament in 2013 and then re-elected in 2018.
[**Null Subject**] She advocates a right-wing political agenda, emphasizing national sovereignty, cultural conservatism, and economic liberalism. [Null Subject] has stood out for her critical stance towards the European Union and immigration policies.' [TB3-5]

A closer look at (15) and (16) reveals that the textual patterns displayed in both cases are not the same. The pattern in (16) is somewhat less natural than (15). This is due to the fact that the second text block in (16), which corresponds to TB4 in the actual biography, is composed of two sentences with two different DR: *Fratelli d'Italia* and *Meloni.* The Null Subject (opening TB5) only refers to the second DR of the previous text block. In (15), by contrast, the Null Subject corresponds to the same DR referred to in the previous text block (*Margherita Hack*).

In both cases, we also observe an over-segmentation of the text in small textual units, i. e. text blocks. Given the referential unity of TB1-2 in (15) and TB4-5 in (16), a text in which the sentences composing these two text blocks are part of the same textual unit would seem more natural, especially in the case of (16). The result would correspond to a text block composed of four sentences:

(17) In pochi anni, Fratelli d'Italia$_{\text{Topic}}$ è diventato uno dei principali partiti della destra italiana, con Meloni che è costantemente salita nei sondaggi. Meloni$_{\text{Topic}}$ è stata eletta per la prima volta nel 2013 al parlamento italiano e poi rieletta nel 2018. **[Null Subject]** Propugna un'agenda politica di destra, affermando valori di sovranità nazionale, conservatorismo culturale e liberismo economico. [Null Subject] Si è distinta per il suo atteggiamento critico nei confronti dell'Unione Europea e della politica di accoglienza dei migranti. [TB4-5, manipulated from ex. 16]

'In just a few years, Fratelli d'Italia$_{\text{Topic}}$ became one of the main parties of the Italian right, with Meloni steadily rising in the polls. Meloni$_{\text{Topic}}$ was first elected to the Italian parliament in 2013

and then re-elected in 2018. [**Null Subject**] She advocates a right-wing political agenda, emphasizing national sovereignty, cultural conservatism, and economic liberalism. [Null Subject] has stood out for her critical stance towards the European Union and immigration policies.' [TB4-5, manipulated from ex. 16]

In the generated texts reproduced in (15) and (16), instead, each text block is composed of two sentences, in line with most of the other text blocks. As acknowledged in chapter 3.2, in the BioFemIt corpus each text block is composed on average by two sentences. In the 201 text blocks included in the corpus, only 6 are composed of four sentences. The following example, which is one of those, allows to observe that the referential chain is much more compact:

(18) Nata il 26 aprile 1977 a Milano, Italia, Samantha Cristoforetti$_{\text{Topic}}$ ha avuto un'istruzione molto variegata. [Null Subject]$_{\text{Topic}}$ Ha frequentato la Technische Universität di Monaco, dove ha conseguito un diploma in ingegneria meccanica nel 2001. Successivamente, [Null Subject]$_{\text{Topic}}$ è entrata alla Scuola di Applicazione dell'Esercito Italiano nel 2002, dove ha completato un addestramento aeronautico. In seguito, [**Null Subject**] ha prestato servizio come pilota dell'Aeronautica Militare italiana. [TB2]

'Born April 26, 1977, in Milan, Italy, Samantha Cristoforetti$_{\text{Topic}}$ had a diverse education. [Null Subject]$_{\text{Topic}}$ attended the Technische Universität in Munich, where she earned a degree in mechanical engineering in 2001. Subsequently, [Null Subject]$_{\text{Topic}}$ entered the Italian Army Application School in 2002, where she completed an aviation training. Later, **[Null Subject]** served as a pilot in the Italian Air Force.' [TB2]

4.2.2 Pronouns

In the BioFemIt corpus, DR codified as pronouns are not very numerous (12 occ.). Most of them (11 occ.) are proclitic pronouns (note that they are never part of a marked syntactic construction, such as a left or right dislocation), either corresponding to the direct object *la* 'her' (as in ex. 19 and the first occurrence in ex. 20) or the indirect object *le* 'to her' (as in the second occurrence in ex. 20). In all the cases found in the dataset, the pronoun refers to the main DR, namely the woman at the center of the biography.

(19) Fabiola Gianotti$_{\text{Topic}}$ è una figura emblematica nel mondo della fisica delle particelle. La sua dedizione alla ricerca scientifica, il suo impegno per la promozione dell'uguaglianza di genere nel settore scientifico e la sua notevole carriera al CERN **la** [Fabiola Gianotti] rendono un esempio per aspiranti fisici in tutto il mondo. [TB6]

'Fabiola Gianotti$_{\text{Topic}}$ is an iconic figure in the world of particle physics. Her dedication to scientific research, her commitment to promoting gender equality in science, and her remarkable career at CERN make **her** [Fabiola Gianotti] an example for aspiring physicists worldwide.' [TB6]

(20) Dopo il ritorno alla vita civile, Segre$_{\text{Topic}}$ si è sposata e ha avuto due figli, continuando a vivere a Milano. Nel 1990 [Null Subject]$_{\text{Topic}}$ ha iniziato a raccontare la sua storia nelle scuole e a gruppi di giovani, con l'obiettivo di mantenere viva la memoria dell'Olocausto.
Nel gennaio 2018, il presidente italiano Sergio Mattarella **l**'[Liliana Segre] ha nominata senatrice a vita in riconoscimento del suo impegno a educare gli altri sulla tragedia dell'Olocausto. Da allora [Null Subject]$_{\text{Topic}}$ utilizza la sua posizione per parlare contro l'odio, il razzismo e l'antisemitismo. Nel 2019, a causa delle numerose minacce ricevute, **le** [Liliana Segre] è stata assegnata una scorta permanente. [TB5-6]

‘After returning to civilian life, Segre$_{\text{Topic}}$ married and had two children, continuing to live in Milan. In 1990 [Null Subject]$_{\text{Topic}}$ began telling her story in schools and to youth groups, with the goal of keeping the memory of the Holocaust alive.
In January 2018, Italian President Sergio Mattarella named **her** [Liliana Segre] senator for life in recognition of her commitment to educating others about the tragedy of the Holocaust. [Null Subject]$_{\text{Topic}}$ has been using her position to speak out against hatred, racism and anti-Semitism ever since. In 2019, due to the many threats she received, she was assigned a permanent escort.’ [lit. ‘it was assigned **to her** [Liliana Segre] a permanent escort.’] [TB5-6]

Besides the cases illustrated above, the BioFemIt corpus includes one subject pronoun realized in the form of the 3rd pers. sing. *ella* ‘she’:

(21) La Bindi$_{\text{Topic}}$ è conosciuta per il suo impegno negli affari sociali, lavorando spesso su questioni che riguardano il sistema sanitario, le pari opportunità e i diritti delle donne. [Null Subject]$_{\text{Topic}}$ Ha ricoperto diversi ruoli di alto livello nella politica italiana, tra cui quello di Ministro della Sanità dal 1996 al 2000, durante il primo governo di Prodi. In quel periodo, **ella** ha promosso importanti riforme come la legge sulla procreazione medicalmente assistita. [TB3]

‘[lit. the] Bindi$_{\text{Topic}}$ is known for her involvement in social affairs, often working on issues involving the health care system, equal opportunity and women’s rights. [Null Subject]$_{\text{Topic}}$ has held several high-level positions in Italian politics, including Minister of Health from 1996 to 2000, during Prodi's first government. During that time, **she** promoted important reforms such as the law on medically assisted procreation.’ [TB3]

The pronominal subject highlighted in bold in (21) presents two striking idiosyncrasies. First, it corresponds to the very formal (i. e. diaphasically marked) pronoun *ella* ‘she’, which is practically obsolete and survives in some legal and bureaucratic texts. In standard and neostandard Italian, *ella* has been replaced by *lei*, as already acknowledged by Sabatini (1985/2011: 159) and Berruto (^{2}2012/1987: 83f.). Second, the pronoun *ella* is not required from a syntactic point of view because the DR it refers to is highly accessible in the preceding co-text: it corresponds to the subject *La Bindi* (functioning as Topic) of the two sentences opening the same text block (TB3). The presence of the DR *Prodi* at the end of the preceding sentence is the only factor that could lead to a more difficult decodification of a sentence with a Null Subject instead of the pronoun *ella*. However, the chance of a possible misunderstanding in a sentence like “In quel periodo, [Null Subject] ha promosso importanti riforme […]” (‘During that time, [Null Subject] promoted important reforms’) seems rather small.

Based on the second idiosyncrasy of the pronoun *ella*, we can conclude that the DR occurring in ex. (21) is linguistically over-codified in relation to its cognitive accessibility. Crucially, this is the only case of over-codification of the subject with a pronoun found in our corpus of 30 generated biographies. As this is a common syntactic calque from English (as acknowledged, e. g., in Cardinaletti 2005: 63 and Garzone 2005: 43f.; see the ex. 9 above), we searched for similar cases in a larger sample of generated biographies, comprising 105 texts on 35 female profiles and a total of 1,440 sentences. In this larger sample, however, only one additional case of over-codification of the subject in the form of a pronoun could be identified. As can be observed in (22), in this case we find the 3rd pers. sing. *lei* at the beginning of a new text block, i. e. in a textual space where it is normal to re-instantiate a DR with a more explicit and usually also more complex linguistic expression (as acknowledged in chapter 2.2). In (22), however, resuming the topical DR of the previous text block with a tonic subject pronoun is not fully

natural, as this linguistic encoding typically gives rise to a contrastive reading of the structure (*she* – in relation to someone else – continues her work etc.).

(22) Parallelamente alla sua attività di ricercatrice e politica, Cattaneo$_{Topic}$ è anche un'autrice rispettata, con numerose pubblicazioni sulle sue ricerche e su temi scientifici più ampi.
Lei continua il suo lavoro come ricercatrice e senatrice, contribuendo con dedizione e determinazione ai progressi della scienza e dell'istruzione in Italia e nel mondo. [TB6-7, BioFemIt_large]

'Parallel to her work as a researcher and politician, Cattaneo$_{Topic}$ is also a respected author, with numerous publications on her research and broader scientific topics.
She continues her work as a researcher and senator, contributing with dedication and determination to advances in science and education in Italy and around the world.' [TB6-7, BioFemIt_large]

Given that the sample of generated biographies analyzed only include rare occurrences of DR codified as subject pronouns, we can conclude that the over-representation of English in the training data of GPT-4 (on this issue, cf. chapter 3.1) does not strongly affect the Italian output in relation to this specific feature.[10]

4.3 Discourse referents codified with complex linguistic expressions

Table 6 shows that, besides Null Subjects (covering 28% of the data), the sample of 30 biographies analyzed includes a relatively high percentage of complex RE. Specifically, there are three forms of complex RE, which occur equally frequently in the data: full lexical NPs, last names and full names (each one covers 21% of the data, amounting to 63%). According to Ariel's proposal, all three types of RE signal a low degree of accessibility of the DR they point to, and we should therefore expect to find them in textual spaces where the antecedent of the RE is either syntactically and textually distant or referentially unclear. In what follows, we will show that this is not always the case, providing examples of RE corresponding, respectively, to full names (chapter 4.3.1) and last names (chapter 4.3.2).

4.3.1 Full names

In the BioFemIt corpus, roughly a third of the RE corresponding to a full name ("first name + last name") occurs in the very first sentence of the biography (30 occ.). In this textual configuration, the full name corresponds to the default referential chain opener (see, e. g. TB1 in ex. 14 and 15 provided above): it codifies a DR that is cognitively not accessible, with no antecedent in the previous co-text.

In most of the other cases (58/88), by contrast, the degree of accessibility of the DR can be considered to be high (51/88 occ.) or even very high (7/88 occ.). In 51 cases, the full name appears in text block initial position, as in (23), where we find three examples in three consecutive text blocks. Noteworthy is the fact that 23 out of 51 full names appear in the first sentence of the last text block of the biography, as is the case in (23) and (24). DR codified with a full

[10] Recent studies show that Italian LLM generated texts include features and patterns that are more typical of English than of Italian (for details, cf. Cicero 2023; Tavosanis 2024; De Cesare 2023b and 2024). The reproduction of English features and patterns in Italian LLM generated texts could be due to the large number of texts written in English in the training data (cf. chapter 3.1). Italian LLM generated texts could be (indirectly) anglicized.

name are thus associated to specific textual spaces and slots of the referential chain (i. e. there is a clear form-function pattern): they always open the chain, and they usually also mark the text block in which the chain comes to an end.

(23) Oltre alla recitazione, Paola Cortellesi$_{\text{Topic}}$ si dedica anche alla musica e al doppiaggio. [Null Subject]$_{\text{Topic}}$ Ha prestato la voce a molti personaggi animati nei film Disney e Pixar, mentre come cantante ha pubblicato l'album "Marilù" nel 2015.
Paola Cortellesi è nota anche per il suo impegno sociale e civile. Nel 2013, conduce il Concerto del Primo Maggio a Roma, ed è spesso ospite di eventi e manifestazioni a favore dei diritti civili. È sposata con l'attore e regista Riccardo Milani, con cui ha una figlia, Laura.
Nel corso della sua carriera, **Paola Cortellesi** ha dimostrato di essere una delle attrici più talentuose e versatili del panorama italiano, in grado di passare con agio da ruoli drammatici a quelli comici, senza mai perdere la sua caratteristica grinta e intensità interpretativa. [TB4-6]

'In addition to acting, Paola Cortellesi$_{\text{Topic}}$ also devotes herself to music and dubbing. [Null Subject]$_{\text{Topic}}$ has lent her voice to many animated characters in Disney and Pixar films, while as a singer she released the album "Marilù" in 2015.
Paola Cortellesi is also known for her social and civic engagement. In 2013, she conducted the May Day Concert in Rome and is a frequent guest at events and demonstrations in favor of civil rights. She is married to actor and director Riccardo Milani, with whom she has a daughter, Laura.
Throughout her career, **Paola Cortellesi** has proven to be one of the most talented and versatile actresses in the Italian entertainment industry, effortlessly transitioning from dramatic to comedic roles without ever losing her characteristic determination and intensity in her performances.' [TB4-6]

(24) Oltre alla recitazione, Piera$_{\text{Topic}}$ ha anche pubblicato diversi libri, tra cui "La signora dai capelli rossi" nel 2012, che riflette sulla sua esperienza come attrice e sulla sua vita in generale.
Piera Degli Esposti è una personalità rispettata nel mondo del cinema e del teatro italiano, e continua a contribuire alla scena dell'arte italiana con tutte le sue interpretazioni. Nonostante la sua età avanzata, mostra ancora una passione incrollabile per il suo mestiere, diventando così una fonte di ispirazione per le nuove generazioni di artisti. [TB5-6]

'In addition to acting, Piera$_{\text{Topic}}$ has also published several books, including "The Red-Haired Lady" in 2012, which reflects on her experience as an actress and her life in general.
Piera Degli Esposti is a respected personality in the world of Italian cinema and theater, and she continues to contribute to the Italian art scene with all her performances. Despite her advanced age, she still shows an unwavering passion for her profession, thus becoming a source of inspiration for new generations of artists.' [TB5-6]

In (23) and (24), all DR highlighted in bold can be considered to be highly accessible, as they already function as main DR in the previous text block(s) and coincide with the macro-Topic of the text. As we mentioned earlier, a re-codification of the DR with a clear linguistic expression, at least syntactically more complex than a Null Subject, can be justified by its re-instantiation at the beginning of a new text block. In the cases in point, however, there is no real need to codify the DR with a full name, which is one of the most complex linguistic expressions (in line with Ariel 1990). As we will see in chapter 4.3.2, there are many more cases in which a new text block is opened by a sentence where the subject is codified with the last name only. In both (23) and (24), we thus observe another case of over-codification of the DR. In (23), in addition, close repetition of the full name in such a salient textual position (*viz.* text block opening) is stylistically rather infelicitous.

In the BioFemIt corpus, there also are 7 cases in which a DR codified with a full name is found within a text block, as in (25) and (26):

(25) Nel 1943, all'età di 13 anni, Segre e il padre$_{\text{Topic}}$ cercarono di fuggire in Svizzera, ma furono catturati e deportati ad Auschwitz-Birkenau. **Liliana Segre** è una dei soli 25 minori italiani sotto i 14 anni sopravvissuti alla deportazione di 7761 persone. Suo padre e i suoi nonni furono uccisi nei campi. [TB3]

'In 1943, at the age of 13, Segre and her father$_{\text{Topic}}$ tried to escape to Switzerland, but were captured and deported to Auschwitz-Birkenau. **Liliana Segre** is one of only 25 Italian minors under the age of 14 who survived the deportation of 7761 people. Her father and grandparents were killed in the camps.' [TB3]

(26) Eleonora Duse$_{\text{Topic}}$ nacque il 3 ottobre 1858 a Vigevano, in Italia, in una famiglia di attori. Spesso riconosciuta come una pioniera del metodo di recitazione, **Eleonora Duse** è notoriamente nota per il suo contributo unico e affascinante all'industria del teatro. [TB1]

'Eleonora Duse$_{\text{Topic}}$ was born on October 3, 1858, in Vigevano, Italy, into a family of actors. Often recognized as a pioneer of the acting method, **Eleonora Duse** is famously known [famous] for her unique and fascinating contributions to the theater industry.' [TB1]

In (25), encoding the highlighted DR with a full name (Liliana Segre) is justified. In this case, the subject of the previous sentence refers to two different DR, namely Liliana Segre and her father. Although the antecedent of the DR's referential expression is easily accessible in terms of distance, a syntactically simpler form (such as the Null Subject or even the last name alone) would not be sufficient to interpret the sentence as referring solely to Liliana Segre. Consequently, in (25), the full name allows to easily identify the right DR and avoid any form of referential ambiguity.

The use of the full name Eleonora Duse in the second sentence of TB1 in (26) cannot be justified on the same grounds. In this case, there is only one possible antecedent in the co-text. The DR referred to with the anthroponym *Eleonora Duse* is thus associated to a maximal degree of cognitive accessibility, as all four criteria identified by Ariel 1990 are verified: the antecedent of the DR's referential expression corresponds to the topical subject of the preceding sentence, occurring in the same text block. In this case, it is clear that the second sentence of TB1 over-codifies the preverbal RE. In addition, the short distance between the two RE codified as full names leads to an unpleasant sense of repetition.

4.3.2 Last names

Besides full names, the BioFemIt corpus includes 89 occ. of preverbal RE corresponding to a last name, plus one case in which the last name is preceded by the definite article (cf. *La Bindi* in ex. 21 above). As far as the reference of these RE is concerned, a first observation is that they all point to the main character of the biography, i. e. the macro-Topic of the text. There is only one exception (out of 90), occurring in the text block in (27), where the last name *Ponti* does not refer to the main DR (Sophia Loren):

(27) Loren$_{\text{Topic}}$ è cresciuta a Pozzuoli, nei pressi di Napoli, in condizioni economiche molto difficili. Dopo aver partecipato a un concorso di bellezza all'età di 16 anni, [Null Subject]$_{\text{Topic}}$ ottiene le prime piccole parti in film italiani di quel periodo. La sua prima grande occasione$_{\text{Topic}}$ arriva nel

1953, quando l'acclamato produttore cinematografico italiano Carlo Ponti le offre un contratto di sette anni. **Ponti** diventerà in seguito il marito di Sophia e il principale artefice del suo successo internazionale. [TB2]

'Loren$_{\text{Topic}}$ grew up in Pozzuoli, near Naples, in very difficult economic conditions. After entering a beauty contest at the age of 16, [Null Subject]$_{\text{Topic}}$ got her first small parts in Italian films of that period. Her first big opportunity$_{\text{Topic}}$ came in 1953, when acclaimed Italian film producer Carlo Ponti offered her a seven-year contract. **Ponti** would later become Sophia's husband and the main architect of her international success.' [TB2]

Another observation concerns the textual space occupied by the DR codified as last names. As already hinted in the preceding paragraph, most of these DR tend to occur in text block initial position (64/89 occ.: 72%), as in (28); the others appear in block internal position (29) and there are some rare cases where they appear both in text block initial and internal position, as in (30) and (31):

(28) **Segre** continua a sensibilizzare sulla Shoah e a portare avanti la sua missione di memoria, oltre che a impegnarsi per i diritti umani e contro ogni forma di discriminazione e violenza. In tutto il suo lavoro, mantiene viva l'attualità del ricordo e l'importanza di resistere a tutte le forme di odio e intolleranza. [TB7]

'**Segre** continues to raise awareness of the Holocaust and carry out her mission of remembrance, as well as her commitment to human rights and against all forms of discrimination and violence. In all her work, she keeps alive the relevance of remembrance and the importance of resisting all forms of hatred and intolerance.' [TB7]

(29) **La sua carriera poetica** iniziò molto presto, attirando l'attenzione di scrittori noti come Giorgio Manganelli e Pier Paolo Pasolini, che predissero il suo straordinario talento. **Merini** pubblicò la sua prima raccolta di poesie, "La presenza di Orfeo", nel 1953. [TB3]

'**Her poetic career** began very early, attracting the attention of well-known writers such as Giorgio Manganelli and Pier Paolo Pasolini, who predicted her extraordinary talent. **Merini** published her first poetry collection, "La presenza di Orfeo," in 1953.' [TB3]

(30) **Fini** ha avuto una vita difficile in tenera età quando fu costretta a separarsi dal padre a seguito del divorzio dei suoi genitori. **La sua infanzia** segnata da traumi familiari e problemi di salute spesso si rifletteva nelle sue opere. Nonostante non abbia mai ricevuto un'educazione artistica formale, **Fini** ha sviluppato un proprio stile irripetibile, interpretando spesso scene di donne potenti e spaventose. [TB2]

'**Fini** had a difficult life at an early age when she was forced to separate from her father following her parents' divorce. **Her childhood** marked by family trauma and health problems was often reflected in her works. Although she never received a formal artistic education, **Fini** developed her own unmistakable style, often performing scenes of powerful and frightening women.' [TB2]

(31) Dopo la guerra, **Fallaci** si laureò presso l'Università di Firenze e nel 1950 iniziò la sua carriera giornalistica presso l' "Europeo". Apprezzata per il suo stile audace, grintoso e a volte provocatorio, **Fallaci** divenne celebre per le sue interviste a personaggi di spicco della politica, della cultura e dello sport. Tra questi, ricordiamo il filosofo e scienziato Bertrand Russell, il leader palestinese Yasser Arafat, l'ayatollah Khomeini e molti presidenti e primi ministri. [TB2]

'After the war, **Fallaci** graduated from the University of Florence and in 1950 began her journalistic career at the "Europeo." Appreciated for her bold, gritty and sometimes provocative

> style, **Fallaci** became famous for her interviews with prominent figures in politics, culture and sports. These included philosopher and scientist Bertrand Russell, Palestinian leader Yasser Arafat, Ayatollah Khomeini, and many presidents and prime ministers.' [TB2]

In the case of (31), we observe another form of over-codification of the DR. Here, too, the antecedent of the second RE codified as full name (*Fallaci*) is highly accessible: it corresponds to the topical subject of the preceding sentence, which opens the text block.

5 Conclusions

Based on the results obtained in our empirical qualitative analysis, we can provide the following answers to the three research questions formulated in chapter 1 (repeated below).

1. What referring expressions (full name, pronoun etc.) are used to codify the discourse referent building the main referential chain of the LLM generated biographies? Are they appropriate in terms of register?

The main referential chain running through each LLM generated biography is constructed in a third of the cases with rings that coincide with a Null Subject or, more often (in ca. 63% of the cases), a complex linguistic expression (realized as full name, last name or full lexical NP); there are some marginal occurrences of pronouns and first names (due to lack of space, we did not describe the second form in detail). The linguistic expressions used to refer to the main discourse referents generally appear to be appropriate. We encountered only one (but important) register-marked form, namely the (mostly) obsolete and very formal pronoun *ella*.

2. Does the degree of complexity of the referring expression codifying the main discourse referent reflect its degree of cognitive accessibility? Are there cases of mismatches, either in the form of over- or under-codification of these referents?

Overall, the linguistic form of the rings building the analyzed referential chains reflects the degree of accessibility of the discourse referent. However, our qualitative analysis also reveals different forms of over-codification of a discourse referent, giving rise to a marked textual pattern. A discourse referent is linguistically over-codified when it is realized by a referring expression that is more complex than required by its degree of cognitive accessibility, determined in particular based on the syntactic and textual distance of the antecedent in the previous co-text. In the sample of 30 generated biographies analyzed, we found cases of over-codification involving a subject pronoun (where a Null Subject would have been referentially sufficient) as well as a full name and a last name (in both these cases, a Null Subject would also have been perfectly acceptable).

3. What forms do these chains have? Can we identify recurrent textual patterns, i. e. typical LLM generated referential chains?

The LLM generated referential chains running through the sample of generated biographies have different but clearly identifiable forms. One of the most frequently observed form is a chain constructed with co-referential rings, pointing to the main character of the biography (i. e. the Macro-Topic of the text). This pattern is both very simple and highly repetitive. In most of the chains, there are at least four or five rings that are not strictly co-referential with the female at the center of the biography but refer to her more indirectly, e. g. via a possessive pronoun (as

in S13 in Figure 2). In the sample of generated biographies analyzed, there also are some rare cases in which the chain is constructed almost exclusively with rings referring to the same discourse referent. One example is the text reproduced in Table 4. In the English generated biography reproduced in (1), we even find that the entire chain is made of co-referential rings. The simple and repetitive nature of these texts can be grasped in Figure 2, which provides an abstract representation of the referential chain running through most of the generated biographies.

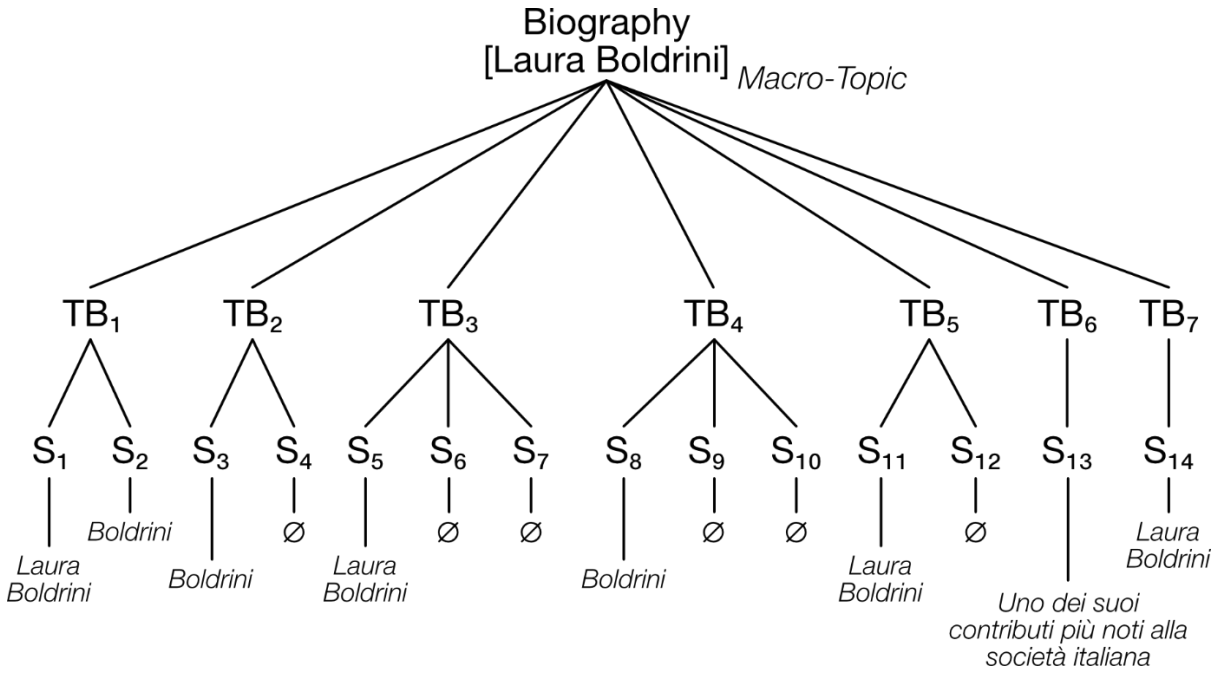

Figure 2: Form of the main referential chain (ø = Null Subject)

Interestingly, in the sample of generated biographies analyzed, there are also rare cases of co-referential (micro-)chains in which each ring is codified with a different linguistic expression:

(32) Nel 1958, **Loren** firma un contratto con la Paramount, che la porta a Hollywood. Qui **l'attrice** ha l'opportunità di lavorare con importanti registi e attori, tra cui Cary Grant e Frank Sinatra. Nonostante il successo, **Sophia** decide di tornare in Italia per lavorare nel cinema italiano. Nel 1962 [**Null Subject**] ottiene il riconoscimento più prestigioso della sua carriera, vincendo l'Oscar come Miglior Attrice per la sua interpretazione in "La Ciociara" di Vittorio De Sica. [TB3]

'In 1958, **Loren** signed a contract with Paramount, which brought her to Hollywood. Here **the actress** has the opportunity to work with important directors and actors, including Cary Grant and Frank Sinatra. Despite her success, **Sophia** decided to return to Italy to work in Italian cinema. In 1962 [**Null Subject**] obtained the most prestigious recognition of her career, winning the Oscar for Best Actress for her performance in Vittorio De Sica's "La Ciociara."' [TB3]

From a macro-textual perspective, the referential chains built by the preverbal referring expressions also form very simple architectures. Most of the generated biographies are constructed with only two types of topic progressions: direct and derived constant Topic progression (cf. Ferrari/De Cesare 2009: 100), involving the concatenation of the same Topic in each sentence or of Topics that are referentially related but not identical, as shown below based on the text reproduced in Figure 2:

- **Direct Constant Topic Progression** (S1-S4): *Laura Boldrini*$_{\text{Topic}}$ < *Boldrini*$_{\text{Topic}}$ < *Boldrini*$_{\text{Topic}}$ < [Null Subject]$_{\text{Topic}}$
- **Derived Constant Topic Progression** (S11-S13): *Laura Boldrini*$_{\text{Topic}}$ < [Null Subject]$_{\text{Topic}}$ < *Uno dei suoi contributi più noti alla società italiana*$_{\text{Topic}}$ 'One of her best-known contributions to Italian society'

All in all, our corpus-driven analysis reveals that the main referential chain building the analyzed biographies is generally well-formed. At the same time, it is generally very simple and relies on repetitive textual patterns. In addition, at a micro-textual level, we find marked textual patterns such as the over-codification of a highly accessible discourse referent as well as cases of over-segmentation of semantically and pragmatically compact textual units. These marked textual patterns probably go unnoticed in a fast and superficial reading. Moreover, even if some portions of the analyzed generated chains appear unnatural (recall that the Italian written tradition prefers *variatio* over *repetitio*; for details, see chapter 2.3), they are still interpretable, and the text does not lose its coherence.

In order to gain more insights into the quality of LLM generated biographies – and of generated texts representing other text genres – , we need to conduct similar studies, based on additional empirical data and an annotation scheme including parameters related to multiple dimensions of textual organization. It would also be necessary to compare generated texts with human-authored ones and tackle the issue of the forms and patterns that are due to language contact with English (e. g., through the presence of explicit subject pronouns). Finally, it would be relevant to develop an analytical and methodological protocol to monitor how the outputs of LLM evolve over time and compare these outputs to the ones produced by non-probabilistic systems (e. g. those described in detailed in Fahime 2024). Longitudinal studies would allow tracking both advancements and regressions in the quality of generated texts (cf. Chen/Zaharia/Zou 2023 refer to unexpected regressions occurring in language models as "behavioral drifts".

References

Andorno, Cecilia (2003): *Linguistica testuale. Un'introduzione*. Roma: Carocci.

Ariel, Mira (1990): *Accessing Noun Phrase Antecedents*. London/New York: Routledge.

Ariel, Mira (2001): "Accessibility Theory: An Overview". In: Sanders, Ted/Schilperoord, Jost/Spooren, Wilbert (eds): *Text Representation. Linguistic and Psycholinguistic Aspects*. Amsterdam/Philadelphia, Benjamins: 29–87.

Baack, Stefan (2024): "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl". *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*: 2199–2208.

Backus, Ad et al. (2023): "Minds: Big questions for linguistics in the age of AI". *Linguistics in the Netherlands* 40: 301–308. doi.org/10.1075/avt.00094.bac.

Berruto, Gaetano (^{2}2012/1987): *Sociolinguistica dell'italiano contemporaneo*. Roma: Carocci.

Brown, Tom et al. (2020): "Language Models are Few-Shot Learners". *Advances in Neural Information Processing Systems* 33. doi.org/10.48550/arxiv.2005.14165.

Cardinaletti, Anna (2005): "La traduzione: un caso di attrito linguistico". In: Cardinaletti, Anna/Garzone, Giuliana (eds.): *L'italiano delle traduzioni*. Milano, FrancoAngeli: 59–83.

Chen, Lingjiao/Zaharia, Matei/Zou, James (2023): *How is ChatGPT's behavior changing over time?* arxiv.org/abs/2307.09009 [21.12.2023].

Chen, Banghao et al. (2023): "Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review". arxiv.org/pdf/2310.14735 [16.09.2024].

Cicero, Francesco (2023): "L'italiano delle intelligenze artificiali generative". *Italiano LinguaDue* 15/2: 733–761.

De Cesare, Anna-Maria (2017): "Cleft constructions". In: Dufter, Andreas/Stark, Elisabeth (eds.): *Manual of Romance Morphosyntax and Syntax*. Berlin/New York: Mouton de Gruyter: 536–568. (=*Manuals of Romance Linguistics* 17).

De Cesare, Anna-Maria (2023a): "Assessing the quality of ChatGPT's generated output in light of human-written texts. A corpus study based on textual parameters". *CHIMERA. Romance Corpora and Linguistic studies* 10: 179–210. revistas.uam.es/chimera/article/view/17979 [26.03.2025].

De Cesare, Anna-Maria (2023b): "*Giorgia Meloni, Meloni o la Meloni?* La codifica degli antroponimi femminili in biografie generate da ChatGPT e pubblicate su Wikipedia". *Lingue e Culture dei Media* 7: 1–20. doi.org/10.54103/2532-1803/22388.

De Cesare, Anna-Maria (2024): "Nuove dinamiche di contatto linguistico: Le "impronte digitali" dell'inglese nell'italiano generato da LLM". Lid'O. *Lingua Italiana d'Oggi* XXI: 67–91.

De Cesare, Anna-Maria et al. (2016): *Sintassi marcata dell'italiano dell'uso medio in prospettiva contrastiva con il francese, lo spagnolo, il tedesco e l'inglese. Uno studio basato sulla scrittura dei quotidiani online*. Frankfurt am Main: Lang. (= *Linguistica contrastiva* 5).

De Cesare, Anna-Maria et al. (2018): "Sentence adverbials: Defining the research object and outlining the research results". *Linguistik online* 92/5: 3–12. doi.org/10.13092/lo.92.4502.

DeepL. deepl.com [27.03.2025].

Fahime, Same (2024): *Referring expression generation in context: Combining linguistic and computational approaches*. Berlin: Language Science Press.

Ferrara, Emilio (2023): "Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models". *First Monday* 28/11: 1–39. doi.org/10.5210/fm.v28i11.13346.

Ferrari, Angela (2010): "*Repetita iuvant*. Note sulla ripetizione lessicale nella scrittura contemporanea non letteraria". In: Ferrari, Angela/De Cesare, Anna-Maria (eds.): *Il parlato nella scrittura italiana odierna. Riflessioni in prospettiva testuale*. Bern, Lang: 149–198.

Ferrari, Angela (2014a): *Linguistica del testo. Principi, fenomeni, strutture*. Roma: Carocci.

Ferrari, Angela (2014b): "The Basel Model for paragraph segmentation: the construction units, their relationships and linguistic indication". In: Pons Bordería, Salvador (ed.), *Discourse Segmentation in Romance Languages*. Amsterdam/Philadelphia, Benjamins: 23–53.

Ferrari, Angela/De Cesare, Anna-Maria (2009): "La progressione tematica rivisitata". *Vox Romanica* 68: 98–128.

Flekova, Lucie/Ferschke, Oliver/Gurevych, Iryna (2014): "What makes a good biography? Multidimensional quality analysis based on Wikipedia article feedback data". *Proceedings of the 23rd International World Wide Web Conference* (WWW 2014): 855–865.

Garzone, Giuliana (2005): "Osservazioni sull'assetto del testo italiano tradotto dall'inglese". In: Cardinaletti, Anna/Garzone, Giuliana (a c. di), *L'italiano delle traduzioni*. Milano, FrancoAngeli: 35–57.

Givón, Talmy (1983) (ed.): *Topic Continuity in Discourse: A Quantitative Cross-language Study*. Amsterdam/Philadelphia: Benjamins.

Goldstein, Josh A. et al. (2023): "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations". doi.org/10.48550/arXiv.2301.04246.

Google Trends: trends.google.com/trends [29.03.2025].

Johnson, Rebecca L. et al. 2022. "The Ghost in the Machine has an American accent: value conflict in GPT-3". doi.org/10.48550/arXiv.2203.07785.

Kilgarriff, Adam et al. (2004): "The Sketch Engine". In: Williams, Geoffrey/Vessier, Sandra (eds): *Proceedings of Eleventh EURALEX International Congress*. Lorient, Université de Bretagne Sud: 105–116.

Korzen, Iørn (2001): "Anafore e relazioni anaforiche: un approccio pragmatico-cognitivo". *Lingua nostra* LXII/3–4: 107–126.

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

OpenAI: *ChatGPT* (GPT-4). chat.openai.com/ [accessed in January 2024 and April 2024].

Navigli, Roberto/Conia, Simone/Ross, Björn (2023): "Biases in Large Language Models: Origins, Inventory, and Discussion". *ACM Journal of Data Information Quality* 15/2: 1–21. doi.org/10.1145/3597307.

Sabatini, Francesco (1985/2011): "'L'italiano dell'uso medio': una realtà tra le varietà linguistiche italiane". In: Coletti, Vittorio et al. (eds.): *L'italiano nel mondo moderno. Saggi scelti dal 1968 al 2009*. Tomo II, Napoli, Liguori: 3–36 [first published in Holtus, Günter/Radtke, Edgar (eds.): *Gesprochenes Italienisch in Geschichte und Gegenwart*. Tübingen, Narr: 154–184].

Sabatini, Francesco (1999/2011): "Rigidità-esplicitezza" *vs* "elasticità-implicitezza": possibili parametri massimi per una tipologia dei testi. In: Coletti, Vittorio et al. (eds.): *Francesco Sabatini. L'italiano nel mondo moderno*. Tomo II. Napoli, Liguori: 181–216 [first published in Skytte, Gunver/Sabatini, Francesco (eds.): *Linguistica testuale comparativa. In memoriam Maria Elisabeth Conte*. Atti del Convegno interannuale della Società di Linguistica Italiana (Copenaghen, 5–7 febbraio 1998). Copenaghen, Museum Tusculanum Press: 141–172].

Searle, John R. 1969: *Speech acts: An essay in the philosophy of language*. Cambridge, Cambridge University Press.

Simone, Raffaele (1990): *Fondamenti di linguistica*. Roma/Bari: Laterza.

Tavosanis, Mirko (2024): "Valutare la qualità dei testi generati in lingua italiana". *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 1/1. doi.org/10.62408/ailing.v1i1.14.

Zhao, Wayne X. et al. (2023): "A Survey of Large Language Models". arXiv:2303.18223v13 [11.01.2024]