Towards Inclusive Communication in Cancer Prevention and Treatment: A Case Study on Italian Informational Materials

**Chiara Cassani[1,2], Luca Brigada Villa[1], Marco Forlano[1],
Serena Coschignano[1], Amelia Barcellini[1,3], Silvia Luraghi[1],
Alberto Giovanni Leone[4], Chiara Zanchi [1], Adalberto Lovotti[1,2]**

[1]University of Pavia, [2]Fondazione IRCCS Policlinico San Matteo,
[3]National Center for Oncological Hadrontherapy (CNAO),
[4]Fondazione IRCCS Istituto Nazionale dei Tumori

Correspondence: ch.cassani@smatteo.pv.it

Abstract

This paper presents an annotation scheme developed to analyze linguistic accessibility and inclusivity in Italian cancer-related informational materials. The scheme combines metadata annotation, qualitative analysis of textual and visual features, and automatically extracted measures of linguistic complexity capturing structural, lexical, and probabilistic properties of the texts. A brief case study demonstrates how the proposed framework can be applied to compare documents and identify different sources of linguistic difficulty. The approach provides a replicable methodological basis for large-scale analyses of health communication materials.

1 Introduction

Cancer remains a major global health challenge, representing the second leading cause of mortality worldwide and the primary cause of death in 57 countries (Bray et al., 2021). Although cancer affects individuals across all regions and demographic groups, substantial and persistent disparities in both cancer risk and survival continue to be documented (Minas et al., 2021). Health care disparities refer to differences in care quality among individuals with equivalent access and similar treatment needs or preferences, and they represent a substantial public health challenge. While numerous studies highlight that ethnic minorities face unequal care and poorer clinical outcomes, disparities also arise with respect to gender, age, socioeconomic background, sexual orientation, rural location, and disability status (Neuhauser and Kreps, 2008). Indeed, socially and economically vulnerable populations bear a disproportionate cancer burden, largely due to barriers related to socioeconomic status, cultural context, and other demographic factors that limit access to preventive services and timely treatment, influence health-related behaviors, and increase exposure to carcinogens and infectious agents. These inequities are particularly pronounced among socially marginalized groups, including individuals with lower educational attainment and limited health literacy, ethnic minority populations, and those whose age, gender, sexual orientation or other personal circumstances further compromise their access to high-quality cancer care (Neuhauser and Kreps, 2008).

Previous work has extensively examined the interplay of structural, socioeconomic, environmental, behavioral, and biological determinants in shaping cancer incidence and outcomes (Zavala et al., 2021). However, far less is known about how these determinants collectively influence access to, and understanding of, cancer-related information, particularly information on prevention and treatment. Similarly, research is limited on how intersecting social and demographic factors define groups at heightened risk of inadequate health communication, as well as on the development of effective communicative strategies to mitigate disparities in access to and comprehension of cancer-related information among underserved populations.

In 2024, a multidisciplinary team composed of cancer specialists (medical, radiation, and gynecologic oncologists) and linguists with expertise in health communication launched the research initiative investigating linguistic accessibility and inclusivity in cancer-related informational materials in Italy. The project addresses critical gaps in understanding how intersecting social, cultural, and demographic factors shape disparities in access to cancer-related information.

The project focuses specifically on the linguistic characteristics of informational materials used in national and regional cancer screening programs and in public awareness campaigns on prevention, early detection, and treatment. Its primary objective is to determine how language influences suitability, acceptance, and uptake of screening recommendations and oncologic care among vulnerable

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 277–283
March 28, 2026 ©2026 Association for Computational Linguistics

populations.

In this paper, we outline the methodological approach and present the annotation scheme developed to analyze the materials collected in the first phase of the project. The paper is structured as follows. In Section 2, we describe the collection of cancer-related informational materials. Section 3 details the multilayer annotation scheme developed to systematically evaluate the characteristics of the collected resources. Section 4 presents an example of analysis. Section 5 concludes the paper presenting future research directions.

2 Materials

The first phase of the project involved building and analyzing a multimodal corpus of official cancer-related informational resources used in Italy at national and regional levels. Materials on prevention, early diagnosis, and treatment have been gathered through systematic searches of institutional websites, physical archives, and healthcare facilities (e.g., hospitals, oncological hubs, Local Health Service Providers and Authorities).

The project officially began in July 2025 and has so far collected 82 cancer-related informational resources, edited between 2010 and 2025 and totaling approximately 130,000 tokens (average length: 1,500 tokens; range: 25-29,000). Most materials (94%, n=77) are produced for regional audiences: 23% originate from local health authorities (n=19), while the remainder is used in public or accredited private hospitals, covering a range of healthcare settings. Approximately one quarter of the corpus (n=21) concerns general cancer prevention and targets the adult population at large, while the remaining materials address oncology patients undergoing medical, surgical, or radiotherapy treatments.

3 Annotation scheme

To systematically evaluate the characteristics of the collected materials, we designed a multilayer annotation scheme specifically developed for the analysis of cancer-related informational resources. The scheme is organized into three complementary layers, namely: metadata; descriptive analysis; linguistic accessibility and inclusivity. In this section, we describe these three layers of annotation in detail.

3.1 Metadata

All collected materials are systematically annotated with relevant metadata, including document format, date of publication, authorship, communicative purpose, and intended target audience.

3.2 Textual analysis

This layer of annotation aims to capture the textual and visual organization of the materials, as well as their main graphic properties, in order to assess how formal features may facilitate or hinder comprehension. Annotators document the resource's overall length (in number of pages and tokens) and textual format (e.g., continuous prose, bullet-point lists, or mixed structures), as well as the organization of internal textual divisions, including paragraphs, sections, titles, and subtitles. Particular attention is also dedicated to the visual layout. Hence, annotators identify the dominant color of the text and background, the use of typographic variation (e.g., standard, creative fonts, or both), and the presence of graphic devices employed to emphasize key information, such as boldface, color highlights, or other visual markers. Finally, visual content is examined, distinguishing between different types of images (e.g., photographs, illustrations), as well as the presence of institutional or commercial logos.

3.3 Linguistic accessibility and inclusivity

This last layer of annotation evaluates the degree to which each resource is linguistically accessible and inclusive of diverse populations, with particular attention to groups at risk of marginalization due to socio-demographic, cultural, or educational factors. These include individuals with limited literacy, linguistically and culturally diverse communities, as well as members of sexual and gender minority groups.

In recent decades, attention to gender-fair language has increased in academia and civil society. However, there is no general consensus on the most equitable strategies to adopt. The debate focuses in particular on the tension between gender neutralization strategies aimed at maximizing inclusiveness and the use of female-gendered expressions intended to counteract the historical invisibilization of women in texts. Nevertheless, several works recommend avoiding the use of generic masculine when referring to female or non-binary individuals, as well as to mixed-gender groups (Thornton, 2022). In response to skepticism regarding the potential increase in linguistic complexity associated with gender-fair texts, experimental studies have demonstrated that their overall readability is not significantly impaired (Gygax and Gesto, 2007; Pepponi, 2024; Pepponi and Comandini, 2025)

After assessing whether the intended target audience is explicitly identified, annotators record lexical and morphological strategies related to gender representation, specifically the adoption of gender-neutral formulations (e.g., persone 'people') in place of gendered or binary expressions (Rosola, 2024; Thornton, 2022). To identify potential biases, annotators also record the presence of visual or linguistic stereotypes that may reinforce inequitable representations of gender, age, ethnicity, sexual orientation, or other demographic characteristics.

Given the relevance of linguistic accessibility for individuals with a migratory background, the availability of translations into languages other than Italian is also documented, specifying both the languages provided and whether the translation covers the entire material or only selected sections. Moreover, the presence of unexplained technical vocabulary is manually recorded, as such terminology may hinder comprehension among readers with limited health literacy.

To quantify the linguistic complexity of the patient-information materials, we automatically extract five complementary metrics capturing structural and probabilistic properties of the texts. Sentence length, dependency tree depth, and dependency length capture structural aspects of syntactic organization and integration cost (Lu, 2010; Liu, 2008; Gibson, 2000; Ferrer-i Cancho, 2004; Futrell et al., 2015), while word surprisal provides a probabilistic estimate of processing difficulty based on contextual predictability (Hale, 2001; Levy, 2008; Smith and Levy, 2013). Lexical accessibility is further assessed through the proportion of lemmas outside the *Nuovo Vocabolario di Base* (De Mauro, 2016), a reference lexicon of core Italian vocabulary.

4 Example of analysis

In this section we illustrate the application of this annotation scheme, focusing on the layers described in Sections 3.2 and 3.3; the full annotation is available in Table 1 (Appendix A). For illustration purposes, we specifically selected two documents that present differences at multiple levels. The first document (henceforth, AA01) is a 15-page informative note on chemotherapy, immunotherapy, and biological therapies with a description of the main side effects and how to manage them, elaborated within a hospital and specifically intended for cancer patients starting medical treatment. The second document (AA02) is an informative brochure on vaccination against HPV, produced by a biopharmaceutical company in agreement with the Ministry of Health for an awareness campaign.

4.1 Textual analysis

Beside their differences in extension (15 pages and 5,909 tokens vs. 2 pages and 630 tokens), both documents have an internal organization into paragraphs and subparagraphs, combining both continuous text and bulleted lists. Document AA01 contains black text over a white background and uses standard, sans-serif fonts, with highlighted parts in bold, italics, and/or underlined. Document AA02 is more varied with respect to font choices: it contains text in black, white, and green, over a blue or white background, combining creative and standard fonts and using bold and different text colors as highlighters.

AA01 does not contain any images, while AA02 includes two photographs of a young, white woman. On top of every page, AA01 includes the logos of the Regional Health Service and of the hospital that elaborated it; AA02 includes the logo of the commercial company that created it.

4.2 Linguistic accessibility and inclusivity

This layer of annotation comprises both qualitative and quantitative metrics. The two documents differ in how they mention their target audiences in the text. AA02 does it more explicitly, listing them in a bulleted list. AA01 combines direct and indirect forms of address (*suggerirvi* "to suggest to you" (plural), *i pazienti* "the patients") and implies the relevance of the material to the patients who receive it, but its title could suggest that it is intended for a medical audience:

> NOTA INFORMATIVA PER LA SOMMINISTRAZIONE DI FARMACI ANTIBLASTICI[1].

Both texts make use of generic masculine. However, AA01 does it extensively (e.g., the numerous

[1] INFORMATION NOTE FOR THE ADMINISTRATION OF ANTIBLASTIC DRUGS.

instances of *i pazienti* (masculine plural, identifiable by the use of the masculine definite article) "the patients"), while AA02 only uses it once, in

> *La vaccinazione anti-hpv protegge il nostro futuro e quello dei nostri figli*[2] (masculine plural).

They also recur to both gendered and gender-neutral linguistic formulations, although in different ways: for instance, AA01 opens with *Gentile Signora, Egregio Signore*, lit. "Gentle Madam, Esteemed Sir", while in AA02 we find gendered language following the biological sex of the referent, as in

> *le donne che sono state trattate per lesioni CIN2+ o di grado superiore*[3].

AA01 contains instances of gender-neutral terms (e.g., the collective *il personale specializzato* "specialized staff"), and AA02 shows different neutralization strategies: use of relative clauses (e.g., *chi si vaccina*, "those who vaccinate"), epicene nouns (*i soggetti* "subjects", *le persone* "people"), elision of article (*adolescenti* "adolescents" instead of *gli/le adolescenti*, with gendered article).

Neither document shows instances of linguistic stereotypes, while the two photographs of a white, young woman contained in AA02 may reinforce the idea that HPV vaccination is exclusively/especially dedicated to this demographic group.

Neither document is available in any language other than Italian and both contain instances of medical jargon without explanation. With its relatively wide length, in AA01 medical terms are generally explained or contextualized, although not necessarily clearly or extensively - e.g.,

> *I quadri di tossicità cutanea possono poi combinarsi tra loro e raggiungere vari gradi di gravità, sino a condizioni di ipersensibilità acuta che possono richiedere il ricovero ospedaliero (sindrome DREES o Steven-Johnson)*[4].

AA02, which is much shorter, mentions medical conditions (*condilomi genitali* "genital warts") and anatomical parts (*orofaringe* "oropharynx") whose names may not be transparent for all readers.

The qualitative observations are complemented by a quantitative comparison of linguistic metrics across the two documents. All metrics were computed automatically from the raw patient-information materials using a standard NLP pipeline implemented with Stanza (Qi et al., 2020). The texts were first segmented into sentences and tokenized using Stanza's Italian preprocessing models. After lemmatizing the text, we compared each lemma to the NVdB lexical set and calculate the ratio of out-of-vocabulary items. We then applied Stanza's UD Italian dependency parser, trained on the Universal Dependencies Italian-ISDT treebank (Bosco et al., 2013), to obtain syntactic structures for each sentence. Sentence length was derived directly from the tokenized output. From the UD parses, we computed dependency tree depth by identifying the longest root-to-leaf path and dependency length as the mean linear distance between heads and dependents within each sentence. To estimate surprisal, we used an Italian autoregressive language model based on GPT-2, — specifically GePpeTto (De Mattei et al., 2020) — from which we extracted token-level log-probabilities conditioned on the left context; sentence-level surprisal was calculated as the average surprisal across tokens. All metrics were aggregated at the document level by averaging over sentences.

AA01 exhibits a markedly higher average sentence length (29.23 tokens) than AA02 (18.23 tokens), pointing to more syntactically elaborate sentence structures. This difference is further reflected in dependency-based measures: AA01 shows greater average dependency tree depth (6.58 vs. 4.84) and average dependency length (3.18 vs. 2.80), suggesting a higher degree of structural complexity in the organization of syntactic relations.

Differences also emerge in probabilistic and lexical dimensions. Average surprisal values are higher in AA01 (6.505) than in AA02 (5.828), indicating lower predictability at the word level. By contrast, AA02 displays a higher percentage of lemmas outside the Nuovo Vocabolario di Base (20.18% vs. 15.62%). These results should, however, be interpreted in light of the substantial difference in document length: while AA01 is considerably longer and includes repeated explanations and contextualization of technical terms, AA02 is more concise

[2] HPV vaccination protects our future and that of our children.

[3] women who have been treated for CIN2+ or higher-grade lesions.

[4] Different skin toxicity patterns can overlap and reach different severity grades, including acute hypersensitivity conditions that may require hospitalization (DREES syndrome or Steven-Johnson).

and concentrates specialized vocabulary within a smaller textual space. As a consequence, lexical specialization appears more densely distributed in AA02, whereas AA01 combines greater syntactic and probabilistic complexity with a comparatively broader reliance on basic vocabulary. Taken together, these metrics suggest that the two documents differ not only in overall complexity but also in the sources from which potential difficulty for readers may arise.

5 Conclusion and future plans

This paper introduced a multilayer annotation scheme for analyzing the linguistic accessibility and inclusivity of cancer-related informational materials, combining qualitative analysis with automatically extracted quantitative metrics. The example analysis demonstrates how documents may differ in both structural and lexical sources of complexity, even when addressing similar health topics. Ongoing and future work will apply this framework to the full corpus to support broader investigations into accessibility and equity in cancer communication, with the longer-term goal of informing the design and production of more effective and accessible informational materials for patients.

Limitations

This study has some limitations. The corpus analyzed is limited in size and primarily regional, and the example analysis is illustrative rather than intended to support generalizable claims. While the selected computational metrics capture complementary dimensions of linguistic complexity, they do not directly model reader comprehension and are partly dependent on modeling choices, such as the language model used for surprisal estimation. In addition, lexical accessibility is approximated through the Nuovo Vocabolario di Base, which provides a useful but coarse-grained proxy for shared vocabulary knowledge and does not account for variation across individual readers.

Ethical Considerations

This study analyzes publicly available cancer-related informational materials produced by institutional and healthcare actors. No personal data, patient records, or sensitive individual information were collected or processed. The analysis focuses exclusively on linguistic and visual characteristics of the materials and does not involve human subjects or interventions. All materials were handled in accordance with applicable ethical standards for research on publicly available data.

Acknowledgments

This paper is part of the research project ICaRe funded by the Fondazione Cariplo under Grant: INEQUALITIES RESEARCH - Generare conoscenza per ridurre le disuguaglianze 2024 (project number 2024-1526).

References

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria. Association for Computational Linguistics.

Freddie Bray, Matthieu Laversanne, Elisabete Weiderpass, and Isabelle Soerjomataram. 2021. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, 127(16):2815–3040.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. GePpeTto carves Italian into a Language Model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 97–104, Bologna, Italy. CEUR Workshop Proceedings.

Tullio De Mauro. 2016. *Il Nuovo Vocabolario di Base della Lingua Italiana*. Garzanti Linguistica, Milano.

Ramon Ferrer-i Cancho. 2004. Euclidean Distance Between Syntactically Linked Words. *Physical Review E*, 70(5):056135.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale Evidence of Dependency Length Minimization in 37 Languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Edward Gibson. 2000. The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In Yasuhara Miyashita, Alec Marantz, and Wayne O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT Press, Cambridge, MA.

Pascal Gygax and Noelia Gesto. 2007. Féminisation et lourdeur de texte. *L'Année psychologique*, 107(2):239–255.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Roger Levy. 2008. Expectation-Based Syntactic Comprehension. *Cognition*, 106(3):1126–1177.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Xiaofei Lu. 2010. Automatic Analysis of Syntactic Complexity in Second Language Writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Tsion Zewdu Minas, Maeve Kiely, Anuoluwapo Ajao, and Stefan Ambs. 2021. An overview of cancer health disparities: New approaches and insights and why they matter. *Carcinogenesis*, 42(1):2–13.

Linda Neuhauser and Gary L. Kreps. 2008. Online cancer communication: Meeting the literacy, cultural and linguistic needs of diverse audiences. *Patient Education and Counseling*, 71(3):365–377.

Elena Pepponi. 2024. Comunicazione istituzionale ampia e tecnologie per il Trattamento Automatico del Linguaggio: possibili applicazioni e sviluppi. *AION – Sezione linguistica*, 12:275–323.

Elena Pepponi and Gloria Comandini. 2025. Semplificazione e inclusione nell'italiano istituzionale. un primo studio sul corpus ALIAS. *Italiano LinguaDue*, 17(1):801–840.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Martina Rosola. 2024. Which Is the Fairest of Them All? Evaluating Gender-fair Strategies in Italian. *Phenomenology and Mind*, 27:84–97.

Nathaniel J. Smith and Roger Levy. 2013. The Effect of Word Predictability on Reading Time is Logarithmic. *Cognition*, 128(3):302–319.

Anna Maria Thornton. 2022. Genere e igiene verbale: l'uso di forme con e in italiano. *AION – Sezione linguistica*, 11:11–54.

Valentina A. Zavala, Paige M. Bracci, John M. Carethers, Luis Carvajal-Carmona, Nicole B. Coggins, Marcia R. Cruz-Correa, Melissa Davis, Adam J. de Smith, Julie Dutil, Jane C. Figueiredo, Rena Fox, Kristi D. Graves, Scarlett Lin Gomez, Andrea Llera, Susan L. Neuhausen, Lisa Newman, Tung Nguyen, Julie R. Palmer, Nynikka R. Palmer, and 12 others. 2021. Cancer health disparities in racial/ethnic minorities in the United States. *British Journal of Cancer*, 124:315–332.

A Annotation table

Metadata	File name		AA01	AA02
	…			
Textual analysis	Number of pages		15	2
	Number of tokens		5,909	630
	Text format	continuous/bulleted list/mixed	mixed	mixed
	Division into paragraphs and subparagraphs	yes/no	yes	yes
	Dominant text color	yes/no (if yes, specify which)	black	black, white, green
	Dominant background color	yes/no (if yes, specify which)	white	blue, white
	Fonts used	standard/creative/both	standard	both
	Use of colors, bold type, or other graphic elements to highlight	yes/no	no	yes
	Presence of images	yes/no	no	yes
	Type of images	photographs, illustrations…	-	photograph
	Presence of institutional/commercial logos	yes/no	yes	yes
Linguistic accessibility and inclusivity	Explicitly stated target audience	yes/no	yes	yes
	Use of the generic masculine	yes/no	yes	yes
	Use of gendered linguistic formulations	yes/no	yes	yes
	Use of gender-neutral linguistic formulations	yes/no	yes	yes
	Presence of visual stereotypes	yes/no	no	yes
	Presence of linguistic stereotypes	yes/no	no	no
	Presence of translations into other languages	yes/no	no	no
	Use of medical jargon without explanation	yes/no	yes	yes
	Average sentence length		29.23	18.23
	Average dependency tree depth		6.58	4.84
	Average dependency length		3.18	2.80
	Surprisal		6.505	5.828
	Percentage of words outside the NVdB		15.62%	20.18%

Table 1: Table reporting the full annotation of the texts AA01 and AA02