Creativity, Invention and Linguistic Analysis.
A Case Study

Silvia Lilli

Department of Department of Literary Studies, Philosophical Studies and History of Art, University of Rome "Tor Vergata", Rome, Italy

silvialilli@hotmail.it

Abstract

Can we systematically analyse the linguistic features of literary texts that employ highly inventive language? How can existing tools for automated text analysis and computational linguistics be adapted when resources designed for natural language fall short? Traditional approaches often prioritize close reading over quantitative analysis, particularly for texts marked by extreme creativity. This study proposes an alternative method, using automated text analysis to examine the language and style of Giovanni Testori (1923-1993), a 20th-century Italian author. Testori's works challenge standard linguistic frameworks, merging Italian, Milanese dialect, word deformations, and code-mixing across stylistic levels. This approach assumes that even highly creative literary languages must maintain a degree of comprehensibility to achieve effective communication, which means relying on iterative and systematic deformation techniques. By identifying and formalizing these patterns through regular expressions, inventive languages can be treated analytically as natural ones, allowing their features to be statistically examined. While the approach involves subjective interpretation and risk errors in manual labelling, it demonstrates how computational tools can quantify and represent an author's style even in extreme cases. Integrating qualitative insights with quantitative analysis offers a deeper and more reliable understanding of the author's poetics, reducing the biases and limitations of close reading alone.

Keywords: Linguistic creativity, Stylistics, Literary Linguistics, Computational Text Analysis, Verbal Art.

È possibile analizzare sistematicamente le caratteristiche linguistiche dei testi letterari che utilizzano un linguaggio altamente inventivo? Possono trovare applicazione gli strumenti attualmente esistenti per l'analisi automatica dei testi e la linguistica computazionale quando le risorse progettate per le lingue naturali risultano inadeguate? Nel caso di testi caratterizzati da un'alta creatività linguistica, gli approcci tradizionali tendono a privilegiare la lettura a campione e l'interpretazione sintetica rispetto all'analisi quantitativa. Questo studio propone un metodo alternativo, utilizzando l'analisi automatica dei testi per indagare il linguaggio e lo stile di Giovanni Testori (1923-1993), autore italiano del XX secolo. Le opere di Testori sfidano le convenzioni linguistiche tradizionali, mescolando italiano, dialetto milanese, deformazioni

linguistiche, codici e registri stilistici diversi. L'assunto teorico di partenza è che anche i linguaggi letterari più creativi hanno come scopo la comprensibilità dell'espressione per il destinatario, e pertanto anche la variabilità più estrema deve fondarsi su tecniche di deformazione iterative e sistematiche. Trattando le lingue letterarie creative, perciò, alla stregua di lingue naturali, è possibile identificarne le caratteristiche, formalizzarle tramite espressioni regolari e analizzarle statisticamente. Sebbene il metodo comporti margini di soggettività e rischi di errore nella categorizzazione manuale, esso dimostra come gli strumenti computazionali possano effettivamente essere usati per rappresentare quantitativamente lo stile di un autore, anche nei casi di creatività più estrema. Fondando l'analisi qualitativa su quella quantitativa, inoltre, è possibile giungere a una comprensione più profonda e completa della poetica dell'autore, superando le impressioni soggettive e i pregiudizi che possono emergere dalla sola lettura.

Parole chiave: Creatività linguistica, Stilistica, Linguistica Letteraria, Analisi Automatica dei Testi, Retorica.

> Chi è mai riuscito a distinguere, per quanti sforzi si sien fatti, tra linga e stile? ([19]: 144)

Introduction

This article summarizes the methodology developed in my PhD project at the University of Rome "Tor Vergata". The project examines linguistic innovation in the literary works of Giovanni Testori (1923-1993), a versatile Italian writer renowned for his avant-garde theatre. Testori's plays feature highly innovative language, created by mixing features of multiple languages – Italian as the base, along with Milanese dialect, Latin, French, and others – and systematically modifying words through techniques such as phonetic alteration, gender switching, and the extension of morphemes to unexpected lexical items. This study poses a challenging question: Can creative literary language be analysed in the same way as natural language?

Although there seems to be an intuitive understanding of what a creative language might be, it is necessary to clarify its application in this specific case. In Literary Studies,[1] at least three different concepts are commonly grouped under this label:

- Deviation from commonly used speech style: This is typical of all literary texts, where the linguistic form draws attention to itself alongside the content.[2]

---

[1] As language invention is driven by various motivations – political, philosophical, commercial, playful, religious, and more – the study of creative language extends beyond literature (for further discussion, see [4], [28], [23], [35], [45]). Successful invented languages include, for example, Esperanto and the more recent Europanto, created by Diego Marani as a common language for the EU. Creativity has also been a central topic in theoretical linguistics, particularly in understanding language production and evolution. Generativists view it as a combinatory process of fixed rules and items, with real creativity remaining an unresolved mystery ([17]). In contrast, idealists regard creativity as the very essence of language and expression ([19], [47]). On this broader topic the bibliography is too vast to detail here.

[2] This deviation can occur in various aspects: lexical, grammatical, phonological, semantic, and more. For an extensive survey of linguistic strategies, particularly in English poetry, see [32]: 42-52.

- The invention of entirely new linguistic systems: Often found in fantasy and science fiction, this technique highlights the otherness of imaginary worlds or futures, such as Elvish in Tolkien's *Lord of the Rings* or Newspeak in Orwell's *1984*.[3]

- Systematic deviation from a standard language: This includes processes such as word or grammar alteration and code-mixing across languages, dialects, and stylistic registers. It is typical of multilingual and experimental literature, such as the works of Carlo Emilio Gadda or James Joyce, often in a parodic context.[4]

While the second type refers to languages invented *a priori*, without relying on any existing system, the other two emerge *a posteriori* as deviations from a commonly used standard for artistic and expressive purpose.[5] However, the first operates within the boundary of a 'probabilistic' deviation: it does not alter the system but distances itself from the most probable patterns and sequences, employing what Noam Chomsky ([16]: 27) described as 'rule-governed creativity'. The latter, on the other hand, employs in addition the so-called 'rule-changing creativity', breaking the system's rules (lexical, syntactical, morphological, etc.) without entirely abandoning them. Instead, it bends the system to align with the author's idiosyncratic vision while enriching it through frequent code-switching and style-shifting phenomena. As is evident, Testori's linguistic invention typically falls into the last category – not only by using techniques of language deformation but also by achieving through language a defamiliarizing effect on the reality shown. This type of linguistic creativity will be the focus of the following discussion.

Excluding *a priori* invented languages, which are more relevant to the fields of linguistics and philosophy, the study of language creativity in the first and last categories falls under stylistic analysis. While the concept of 'style' is not universally defined ([55]), it is generally accepted that stylistics studies literary texts from a linguistic perspective, based on the assumption that every linguistic element in a text holds potential significance. Analysing the language and style of a literary work, especially when it diverges significantly from common language, is fundamental to achieving a deeper understanding of both the text and author's poetics. However, the greater the deviation from standard language, the more challenging it becomes to analyse the text's linguistic peculiarity. For these reasons, much of the existing research has been conducted in an unsystematic manner, relying on close observations, selective sampling, and general impressions.[6] My project, however, sought to investigate the style of Testori's plays from a different perspective.

---

[3] In recent years, the study and creation of such languages have spread significantly, driven by the popularity of the genre in both literature and film and the emergence of dedicated online communities. Today, several experts in language invention are employed in media entertainment production. See, for example, [40].

[4] In Italian literature, this type of language invention is primarily found in comic dramas, particularly in the 16th century, where it was used to challenge the linguistic dominance of literary norms. It gave voice to 'real' spoken languages and exploited the comic potential of linguistic misunderstandings ([25]). More recently, across Europe, it gained prominence in 20th-century modernist fiction and avant-garde literature, serving as an expression of the fragmented and ambiguous perception of reality and consciousness ([46]).

[5] In other cases, it served to protect the existence and identity of a social group, as seen with the thieves' jargon ([18]).

[6] Positioned between two kinds of creativity – the creation of a brand-new language and a stylistic variation within the system – linguistic experimentation in literature has struggled with its placement

Theoretical Approach

Given that the primary goal of an artistic work – especially a theatrical one – is to ensure effective communication with the audience, every creative language must rely on patterns of variation that are both understandable and communicative. [7] Even when the style appears chaotic, idiosyncratic, or irreducible to a system, it must be assumed to possess an internal logic. Without this, the artistic message would become unintelligible, leading to a failure in communication. In other words, deviations from the standard must display recurrent phenomena filled with implicit or explicit additional meaning that the author intended to convey. If this is the case, the features characterizing the creative language should function similarly to variations naturally occurring between and within languages – e.g. geographical, chronological, or contextual[8] – and can thus be identified and analysed systematically, i.e., through a quantitative analysis. In the context of literary text, the purpose of such measurements is inherently linked to interpreting the subjective meaning of linguistic data, as linguistic creativity reflects a fundamental aspect of an artist's unique act of creation. Therefore, while quantitative analysis serves as preliminary step, it is nonetheless essential for enabling a more comprehensive qualitative analysis of text's features, providing a reliable foundation for understanding the text itself.

In short, the methodology is based on the idea that corpus linguistics techniques ([8]) can be applied to literary texts to extract linguistic information, which can then be used to interpret the text from a literary perspective. In applied linguistics, the extensive quantitative exploration of linguistic features in text dates back to the mid-20th century ([30],[31],[14]) and has since benefited significantly from the availability of increasingly larger corpora and advances in computational power, evolving from computational linguistics to text mining. The application of quantitative approach to literary works, however, is a more recent development[9] and has been mostly oriented toward answering questions related to the social aspects of literature (e.g., canonization, reception, stereotypes). This research seeks to combine the methods traditionally used in corpus linguistics to address questions that are neither strictly linguistic nor social, but rather literary. It focuses on the style of a single author, much like traditional close reading, but given the linguistic distinctiveness of the text – and the intention to analyse not «subtle semantic or grammatical structures» but rather «a large amount of simple linguistic features» ([39]: 309) – it leverages computational tools for the quantitative description of these features. The advantage of this approach lies in its independence from selective sampling, thereby avoiding confirmation bias.

---

in academic studies and theoretical framework. Existing research remain sparse and typically focuses on individual authors or texts (see, for example: [22], on Gadda; [44], on Carroll; [12], on Joyce; [20], on Rabelais).

[7] As Den Ouden ([21]: 18) points out, «The distinctive characteristic of human speech lies in its order or coherence, its variation and novelty. […] [Chomsky and Cordemoy] are not talking about novelty and creativity in the sense of being randomly unique or absurdly different. The creative use of language stems from both its variations and its coherence». See also [42].

[8] For definitions and methods in sociolinguistics, see [29], [13], and, for Italian Studies, [6].

[9] Even though methods of 'distant reading' were employed earlier ([54]), the coining of the expression by Franco Moretti ([36]) and the impact of his article spurred the broader adoption of his experimental approach in literary studies during the 2000s. This shift was further supported by the growing availability of digital tools and larger textual datasets.

Treating a creative literary language as a variety of a natural language allows for the possibility of designing a linguistic analysis using similar procedures. However, assuming this theoretical feasibility, how can such an analysis be practically implemented? While both natural languages and creative literary languages exhibit variation patterns, linguistic analysis of natural languages benefits from tools developed over the years for automated categorization and quantification of observed features, such as dictionaries, lemmatizers, and POS taggers. These tools,[10] however, are created and trained on well-documented languages – primarily standard varieties – and may therefore prove unreliable or inadequate when applied to the 'rule-breaking' creativity of literary language. The stylistic innovations in such text often go beyond the unexpected combination of existing words, extending to the creation of entirely new word forms through novel combination of morphemes or sounds. Consequently, the outputs from conventional analytical tools and linguistic reference corpora are mostly ineffective. Nonetheless, as stated, even the most creatively constructed languages exhibit identifiable patterns of variation. This makes it possible to conduct extensive and (partially) automated text analysis, although with the acknowledgement of certain limitations inherent in current tools.[11]

The corpus

Before delving into the detailed research design, I will provide an overview of my case study. The corpus consists of six plays by Giovanni Testori, written in two chronological phases: three plays from 1972 to 1977 and three from 1990 to 1992. The corpus contains a total of 87,409 tokens and 23,415 types,[12] resulting in a Type-Token Ratio (TTR) of 0.268. In terms of formal structure, two plays are written in prose, three in poetry, and one combines both forms. The base language is Italian, incorporating various features (especially phonetic) from the Milanese dialect, alongside numerous words borrowed from other languages, both modern (French, English, Spanish, German) and ancient (primarily Latin). These borrowed words are often partially adapted to Italian phonetics. Additionally, the plays frequently include coinages formed by combining existing morphemes, gender switching, or arbitrary phonetic alterations. Within the Italian language, multiple codes are interwoven: specialized lexicon, archaisms or literary expressions, and, conversely, vulgar language and colloquialisms. This creates a dynamic without a clear stylistic trajectory – neither consistently elevated nor consistently lowered. Instead, the plays show a continuous and intense tension between these poles. The impact of this linguistic deformation is best understood in the context of the themes and settings of the plays. Most are rewrites of classic modern works (such as *Hamlet*, *Macbeth*, and *Faust*) or ancient plays (*Oedipus*,

---

[10] Existing resources for Italian are listed in [10] (239 et seq.). For English, see, for example, WordNet (https://wordnet.princeton.edu/), a large lexical database where «[n]ouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations».

[11] An entirely different approach could have been to use Machin Learning to train a model capable of classifying words deviating from the standard into predetermined categories. However, this approach is impractical for small corpora and may not provide reliable classifications when marked forms require nuanced interpretations. Furthermore, it demands significant computational power and technical expertise, which were not available in my case. Its implementation may also require broader relevance to the scientific community to justify the investment.

[12] Counts were obtained using AntConc ([3]). See 29 for further specifications.

Oresteia) and historical myths, such as those of Herodias, Cleopatra, and the Virgin Mary.[13] Testori employs a subtle technique of shifting from the serious, often tragic tone characteristic of the classics to a grotesque one. This shift strips the language of its 'seriousness' and denotative potential, unveiling a deeper truth (religious, in Testori's view) that lies behind the cultural and social constructs embedded in standardized language.[14]

Given the literary nature of the text, the analysis seeks to deeply understand the meaning underlying the process of language creation in the author's work. It aims to highlight how language closely correlates with the thematic concerns and the pragmatic functions of the theatrical pieces. In other words, the deliberate divergence from standard language represents a conscious and purposeful choice. Therefore, grounding the interpretation in a reliable depiction of the linguistic creation processes at play becomes all the more essential.

Research Design

The first step in designing a linguistic analysis of a creative literary language is to conduct a thorough close reading of the corpus. This preliminary reading is indispensable in this unique context, as each language shows idiolectal traits that are not generalizable. While certain processes or linguistic features may recur across the works of different authors, assuming these features to be universally applicable to all creative languages can lead to problematic oversimplification. Therefore, in the initial stages of research, the most prudent approach is to treat the corpus under investigation as a standalone source of evidence from which to derive generalizations about the author's creative processes.

Close reading also aids in determining the unit of analysis. Should the corpus be analysed as a single entity, or should it be divided into smaller units or grouped in specific ways? In my case study, I chose to treat each play as a separate corpus. This approach was motivated by the substantial variation between the works and the need to enable chronological comparison. Clearly, the selection of an analysis unit depends on both the corpus's characteristic and the research question at hand.

At this stage, the procedure of analysis can be summarized as follows:

Identifying constitutive categories

The initial step involves determining whether the marked linguistic features can be grouped into broader categories based on various parameters. These parameters might include sources of variation typically identified in sociolinguistics, such as diachronic (e.g., an archaic word or a

---

[13] The plays included in the corpus are as follows: *L'Ambleto* (1972), *Macbetto* (1974), *Edipus* (1976), *Sfaust* (1990), *sdisOrè* (1991), and *Tre Lai* (1992). The first three were published in [49] and the latter in [53]; *Edipus* and *Tre Lai* were also recently included in the collection of selected works republished by Mondadori in the *Meridiani* series ([63]). Testori's experimental plays have rarely been translated due to the challenges of the task and the limited recognition he received in Italy, partly because of his cultural stance as a Catholic and anti-secularist. Exceptions include a published French translation of *L'Ambleto* ([48]) and more recent translations for theatrical performances ([50],[51],[52]).

[14] For a literary and linguistic perspective on Testori's theater, see at least references [58],[59],[60],[61],[62]. As with other authors mentioned, linguistic studies on Testori's linguistic creation have only focused on selected samples and close readings.

discarded phonetic or graphic feature); diatopic (variations within the language, such as regional forms, or between languages, such as foreign elements); and stylistic (e.g., distinction between high and low registers). Additionally, the choice of categories could also be influenced by other considerations, such as the process of alteration – its location (e.g., phonetic, morphological, syntactical), its methods (e.g., substitution, combination) – or the pragmatic function of the marked form (e.g., humour, intensification, estrangement).

When selecting categories for analysis, it is crucial to recognize their connection to the interpretative hypotheses about the creative language in the author's works, and the types of conclusions the analysis aims to draw. Opting for specific categories as indices for grouping observed linguistic features does not preclude further distinctions at sublevels. Instead, it assigns a hierarchical structure to the analysis, which will guide the final interpretations and conclusions. This analysis adopts a top-down approach, beginning with a clear hypothesis about the meaning of the author's experimentation and a defined objective for what the linguistic research seeks to uncover.

Another significant issue when discussing variation is determining the basis on which a linguistic variety can be considered the standard against which marked forms are compared. The concept of a 'standard' in linguistics is well-known to be highly controversial. [15] For this reason, considerable attention must be paid to assessing the reference language, taking into account both linguistic and extralinguistic factors. In my case study, I adopted normative Italian as the standard language. Although dialectal (Milanese) elements are heavily present in the texts, they never compromise the syntactical and morphological system of the national language, and even the phonology divergence is partial and inconsistent. For this reason, Testori's plays cannot be classified as part of the dialectal literary tradition.[16] Moreover, in analysing the style of a literary text, the notion of standard does not concern only adherence to the grammar of a diatopic variety but also alignment with the features of a stylistic register. Style, in this broader sense, depends on the conventions of a specific genre and can vary significantly between different literary works. Defining the reference style was crucial, in my case, to determine which elements should be considered 'marked', besides the macroscopic infractions of the linguistic norm. However, this has been quite a controversial point, given the ambiguity in Testori's plays between the theme and content of the texts – mostly tragic – and their parodic treatment. The text exhibits features that lean toward both high and low registers, often juxtaposed to create a striking contrast. It is evident that neither a high literary style nor a low register serves as the baseline; rather, these registers function as opposing poles, deliberately stretched to achieve unexpected effects. For this reason, I considered a 'medium' stylistic level as the standard, meaning that features pushing the language toward either extreme were considered marked.[17]

---

[15] See, for instance, [7], [2], and for Italian, [6]: 68 et seq.

[16] Milanese dialectal tradition has been vibrant since the 13th century, strengthened by the municipality of Milan's historical prominence. It became particularly influential during the modern era, offering an alternative linguistic and cultural pole to the normative tradition rooted in Florentine. For a comprehensive review and anthology of major works and authors, see [5].

[17] While marking a syntactic trait as colloquial or elevated can be quite straightforward, doing so for individual words is more challenging. Especially for Elevated Language, I referred to *NVdB*'s usage markers ([57]), considering as marked all words associated with the labels CO (Common), BU (Low Usage), OB (Obsolete), LE (Literary), which are typically used by medium-to-highly educated people. For Colloquialisms, I prioritized semantics, considering as marked all words referring to sex and bodily functions, insults and imprecations, as well as lexicon related to everyday objects, particularly in a rural context. The inclusion of this last category is justified by the lowering effect

Having defined the linguistic and stylistic standards of reference, I categorized the divergent features into four groups, prioritizing the source of variations as pivotal for describing Testori's idiolectal composition.

[1] Dialectisms: Features (phonetic, morphological, lexical) [18] typical of the Milanese dialect.

[2] Colloquialisms: Features (morphological, lexical, syntactic) commonly used in spoken and informal registers.

[3] Elevated Language: Features (phonetic, morphological, lexical) that elevate the discourse, including archaic or literary forms typically found in written and more formal registers.

[4] Idiolectal Features: Features (phonetic, morphological, lexical, syntactic) characteristic of the author's unique style. These include coined words, expanded uses of existing forms, and distinctive syntactic structures (e.g., couplets, repetitions, lists).

While the first three categories – Dialectisms, Colloquialisms, and Elevated Language – narrowly define specific sources of variation, the fourth category, Idiolectal Features, encompasses a broader range. This group includes elements derived from technical jargon, foreign languages, and entirely invented words. Despite their diversity, these features share a common pragmatic function: to create an estrangement effect [19], surprising and disorienting the audience or deconstructing language itself through mechanisms such as grammaticalization and exaggerated repetition.[20] In contrast, the pragmatic functions of Dialectisms and Colloquialisms are primarily to lower the register, introducing an informal, familiar, and intimate tone. Elevated Language, on the other hand, serves to distance the discourse from everyday spoken language, adding formality and gravitas. Organizing the research around these groups is not merely a matter of classification; it reflects an interpretative framework. The linguistic choices that bend the language in specific directions are deliberate, serving distinct pragmatic effects – whether to elevate, lower, or estrange the language.[21]

---

produced when concrete names such as for clothes, animals, and utensils, appear in the artificial setting of classical rewriting.

[18] Dialectal syntactic features were excluded from the analysis, as their presence in the texts proved insignificant. For each group, only grammatical branches with observed phenomena are listed.

[19] For a definition and history of the concept of 'estrangement' and related bibliography, see [34].

[20] These techniques were particularly influenced by lessons drawn from Renaissance theatre; see [1]: 48-49.

[21] The pragmatic effect identified here should not be interpreted as the final effect achieved by specific words, syntagms, or sentences in individual occurrences. The ultimate meaning depends heavily on context and potential ironic intent. However, irony and parody, which invert or subvert the meaning, do not negate the pragmatic function of the linguistic category itself; rather, they operate within this function, using it as a vehicle for parodic effect (see, for instance, [27]). A final interpretation of these sources of linguistic invention should consider multiple additional factors, such as the author's ideology, cultural traditions, historical context, themes, and so on. Linguistic analysis should play a central role in this interpretation, being just a part of it, as this study aims to demonstrate.

Selecting the actual features within each category.

After defining the main categories through corpus observation, the subsequent step is to identify the specific features within each category. For sociolinguistic varieties and dialects, existing feature lists can be consulted to select those that are present in the corpus.[22] Conversely, for other types of variations – such as Elevated Language and Idiolectal Features – a bottom-up approach may prove more effective. This involve directly identifying linguistic features in the text and classifying them into the relevant categories.[23] It is crucial to maintain flexibility in this process. As the analysis progresses, the list of the features may require adjustments, including the introduction of new features or the revision or removal of existing ones based on their relevance and frequency of occurrence. Each feature is assigned a unique reference label to streamline systematic counting and manual annotation. This structured approach not only facilitates a detailed and reliable analysis, but also support the selective refinement of features for further examination.

Selecting ruled-based tools for automatic features extraction.

Once the features to be searched in the text are identified, the next step is to evaluate whether they can be extracted using automatic text analysis tools. A key distinction must be made between lexical and textual analysis ([10]: 111), based on the number of words involved in each feature (single-word features versus multi-word features). In this study, lexical analysis was predominant. However, due to the idiolectal nature of the case study's language, the range of Natural Language Processing (NLP) tools that could be effectively employed was quite limited, focusing primarily on tokenization. Other NLP tools, such as Part-of-Speech (POS) tagging, Named Entity Recognition (NER) and lemmatization, performed poorly with the creative linguistic structures present in the corpus.[24] Given the emphasis on classifying word forms from phonetic and morphological perspectives rather than semantic ones,[25] tools designed for semantic analysis – such as lemmatization, topic modelling, sentiment analysis, and keyword extraction – were not used in this study. To partially automate the classification task, text searching through Regular Expressions (REGEX) proved essential for features that can be described using precise character sequence rules. It is crucial, however, that REGEX development considers potential variations across different sections of the corpus, such as

---

[22] For this case study, existing research on Milanese dialect ([41], [33]) and on diaphasic and diamesic variation in Italian ([56], [43]) were referenced.

[23] Regarding Idiolectal Features, no pre-existing list could be used, as the deviating linguistic elements are strictly individual and unique to the author. For Elevated Language, while it is possible to compile a list of frequently occurring features that elevate language, these tend to be rhetorical rather than strictly grammatical or phonetic, and the scope of such list is likely too extensive to serve as a practical reference – at least in this case. Rhetorical elements, such as metaphors, rhymes, alliterations, were not included in this analysis due to the limitation of searching for them automatically without the use of machine-learning approach. While these features could have been manually analysed (as was done for other linguistic elements), the focus was instead placed on word forms and highly recurrent syntactic structures (e.g. repetitions, binomials, noun-modifier inversions).

[24] I experimented with the software TaLTaC ([9]), which integrates various resources for the Italian language.

[25] Semantics was only considered in Colloquialisms (see note 7) and technical jargons included in Idiolectal Features.

graphic differences or preferences for specific morphemes.[26] Additionally, precision and recall checks must be regularly conducted to ensure that the REGEX patterns accurately identify and retrieve the intended linguistic feature without omitting relevant data. This iterative refinement is key to maintaining the reliability of the automated extraction process.

Processing queries based on REGEX.

Once the REGEX patterns are defined and collected into a spreadsheet, they can be used to process queries across the text. For this purpose, I utilized the free software AntConc.[27] While software with a GUI does not require programming expertise, it necessitates processing queries one by one. In contrast, programming environments like Python or R enable the execution of multiple queries sequentially, speeding up the process. However, handling each query individually in AntConc provides greater control over the outcomes at each step, which is crucial for fine-tuning the REGEX patterns. Additionally, this approach makes it easy to switch between examining word lists and observing words in their specific contexts, depending on the need. The ability to consider context is particularly useful when dealing with homographs, especially given the absence of reliable POS tagging tools for creative languages. AntConc facilitates this shift through its 'Word' or 'Key Word in Context' (KWIC) functions, which allow for transitions from type-level to token-level analysis. Matching words or syntagms were then extracted and documented in a separate spreadsheet for each feature under consideration.

Manually filtering the false positives.

To maximize the capture of relevant occurrences – given the variability and unpredictability of creative literary language – it is advisable to design REGEX patterns with a focus on broader recall. However, prioritizing recall comes at the cost of reduced precision, leading to more false positives. To address this, the results must be manually reviewed to filter out irrelevant matches. While manual filtering is always feasible, overly strict rules aimed at enhancing precision can make it impossible to recover false negatives. Clearly, the choice between broader recall and stricter precision depends on the corpus size and the resources available for analysis.

Manual annotation of other relevant features.

At this stage, I focused exclusively on lexical and textual analyses that could be partially automated.[28] However, when the corpus size permits, the most effective method to ensure the

---

[26] The process of building REGEX patterns may involve selecting specific occurrences of a given feature. For example, in my case, one dialectal phonetic trait is the sonorization of intervocalic mute consonant, such as /t/ evolving into /d/. To automate data extraction, I focused on the most frequent observed instances of this phenomenon, such as the past participle suffixes (REGEX: [aiu]d[oaie]$). The selection intentionally excludes less frequent and more unpredictable occurrences, which will be retrieved by manual annotation.

[27] Several free programs with graphical user interface (GUI) are available for text analysis, including AntConc, TaLTaC, and Voyant Tools. I chose AntConc because it offers a standalone version, and I was not interested in more complex tasks that require external linguistic resources.

[28] This applies to features characterized by the repetition of limited character sequences or those signalled by one or a few pivotal words. For example, consider the frequent use of the 'false' adversative subordinate introduced by the conjunction 'anzi'. Searching for the single word and filtering out false positive was possible by considering the word in context, rather than just focusing on the individual type.

comprehensive identification of all relevant features is to supplement automated analysis with manual annotation. This involves reviewing the word forms in the word list generated during the tokenization process using the software tool.[29] In my case, each text in the corpus contained between 2,878 and 4,693 unique word types, totalling 23,415 types across the entire corpus. Although this number initially appeared substantial, the process was accelerated as it excluded the labels from previous automated analysis. While manually annotating the word list is suitable for lexical analysis, it proves inadequate for identifying textual features, which demand contextual observation. To address this, I performed contextual manual annotation using the free annotation software CATMA ([26]).[30] Completing this manual annotation required significant effort, spanning three years of my PhD program. Its feasibility, as evident, depends greatly on the time and resources available.

Grouping the Counts According to the Categories.

The occurrences found for each feature, after being checked and counted, were organized using two levels of analysis: first, according to the four categories (Dialectisms, Colloquialisms, Elevated Language, Idiolectal Features) defined during the research design stage as most relevant to the research questions; second, within each of these categories, according to the traditional divisions of formal linguistics (i.e., phonetics, morphology, lexicon, syntax). This two-tiered grouping allowed for a better understanding of how the type of linguistic deviation relates to its specific formal features.

Interpreting descriptive statistics (e.g. measure of variability, correlation matrices, time series).

Based on the gathered and grouped data, further statistical analysis can be conducted. In my case study, I explored several key aspects:

- Temporal dynamics: How do the impacts of these categories on language experimentation change over time?

- Differences across plays: Which category is most represented in each play, and how do the categories vary between plays?

---

[29] Word lists were extracted from AntConc results. The tokenization process was performed using the default software settings, which include UTF-8 encoding and treat all punctuation characters as separators, except in *Sfaust*, where the hyphen was included in characters list, as Testori makes extensive use of compound word as a creative strategy. Given the extreme variation in word forms, it was not possible to automatically check for multiword expression or homographs. While the general statistics derived from this tokenization are approximate (including the numbers of types, tokens, and related statistics), the detailed categorization and analysis (specifically the counts of each individual feature within the four categories) were conducted manually. Word lists were manually checked, and homographs, apostrophized words, and some multiword expressions were distinguished, but only when relevant for the marking process. For this reason, given its unsystematic nature, I preferred to rely on AntConc's approximate tokenization for general statistics and normalization.

[30] The software was selected based on the review of major annotation tools by Neves and Ševa ([38]). In making the choice, priority was given to the ability to allow overlapping annotations and to extract them along with surrounding context.

- Correlations between category: Which categories show positive or negative correlations?

These questions are crucial for understanding the deeper meaning of language experimentation in Testori's work.

While the detailed tables of results have been published in my PhD thesis,[31] I will include here only a selection of the graphs derived from the statistical analysis as illustrative examples.[32]

Figure 1 illustrates the sum of the normalized values of each category across plays in chronological order. It reveals that two categories – Colloquialisms and Dialectisms – remain relatively stable over time. In contrast, Elevated Language and Idiolectal Features exhibit sharp peaks in the first two plays of the second trilogy (1990-1991).
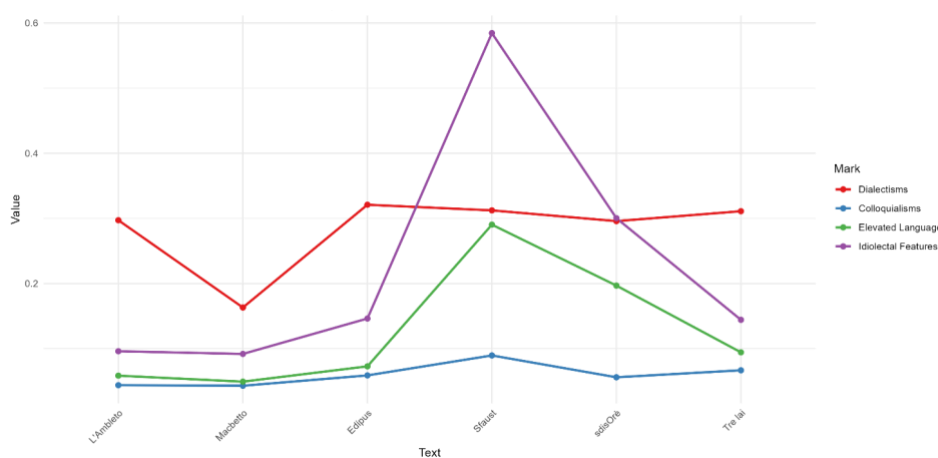

Figure 1 – Evolution of the market features across plays

Figure 2 displays the mean value for each category across all plays, with standard deviation shown as vertical error bars. This graph highlights the dominance of Dialectisms, while also illustrating the extreme variability of Idiolectal Features and Elevated Language. Intriguingly, the analysis shows that one of the most widely recognized elements of Testori's style – foul language, for which he was often criticized – is actually among the least pervasive features, as revealed by the Colloquialisms category.[33] Quantitative analysis thus helps overcome interpretative biases that

---

[31] My dissertation is available in full at this link:
https://www.academia.edu/129029070/La_lingua_sperimentale_nelle_opere_teatrali_di_Giovanni_Testori.

[32] The scripts for the analyses and the data frames used as the basis are available in the Github repository.

[33] Consider, in addition, that in this category is included not only foul language and insults, but also syntactic colloquialism (mostly dislocations) and the lexicon identifying objects and concept derived from everyday life (animals, tools, body parts, etc.). All considered, the foul language only represents a very small part of all the marked words in the texts.

may stem from cultural or psychological predisposition (e.g., aversion to expressions related to sex and bodily functions).
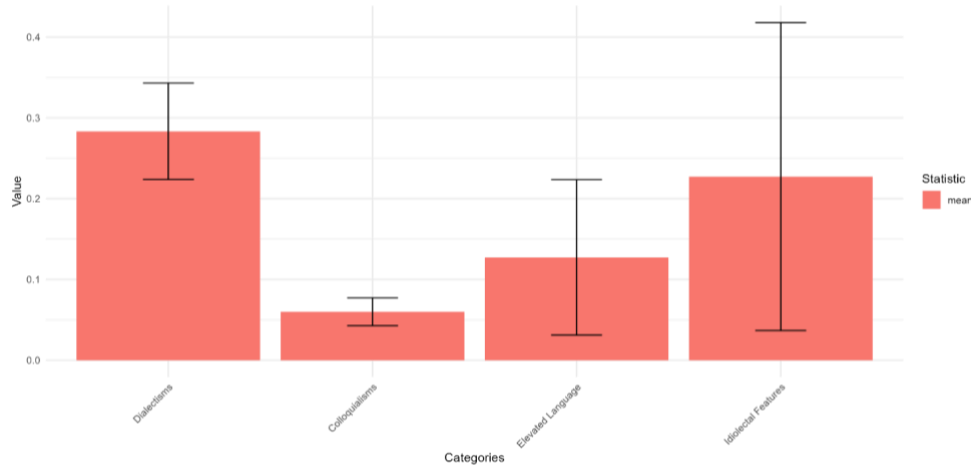

Figure 1 - Mean of marked categories with standard deviations as error bars

Figure 2 breaks down the internal distribution of each category according to the four areas of formal linguistics: phonetic, morphology, lexicon, syntax. The chart demonstrates how each category predominantly operates within specific dimensions: Dialectisms focus primarily on phonological variations; Colloquial and Elevated Language affect the lexicon; Idiolectal Features often target morphological and syntactic structures. Although some distortion may arise from the initial selection of features within each category – given its partly subjective nature – this selection was rooted in close textual observation. It is my strong belief that further analyses will not diverge significantly from the presented findings. This chart also helps resize the perceived impact of the Milanese dialect. While it is the most represented category, the majority of dialectal features involve minor phonetic changes that do not alter the core linguistic structure of Italian. Consequently, the longstanding relegation of Testori's work to a regional context – often dismissing its broader literary significance – is unjustified.
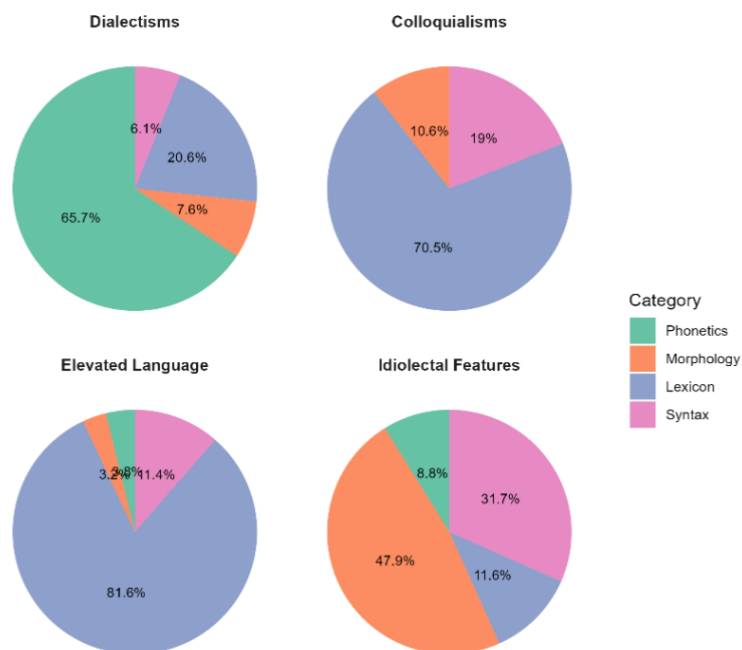
Figure 2 - Composition of each marked category according to formal linguistics aeras

Figure 3 examines the correlations between the four categories. A strong positive correlation is evident between Elevated Language and Idiolectal Features.[34] Surprisingly, Colloquialisms also correlate more strongly with Elevated Language and Idiolectal Features than with Dialectisms, despite my initial hypothesis stated that both Colloquialisms and Dialectisms functioned primarily to lower the stylistic level. This finding suggests that Colloquialisms, like Elevated Language and Idiolectal Features, although they differ in intensity, serve the same overarching purpose of steering the language away from convention. Dialectisms, however, behave differently from the other features. Except for one play (*Macbetto*), Dialectisms remain persistent and stable over the years, acting as a form of 'background noise'. This stability resists both the centrifugal forces that push the language away from linguistic norms, and the centripetal pull toward standard Italian. Serving as a balancing element between these two poles – and in alignment with the 20th-century shift of dialect from a social medium to an intimate vehicle of individual memory (see [11]) – dialect anchors the language in a personal, nostalgic sphere, opposing both the artificiality of linguistic norms and their parody through creative language.

---

[34] P-values were also calculated to assess the robustness of the correlations. A positive correlation between Elevated Language and Idiolectal Features was significant with $p = 0.0005$; the correlation between Colloquialisms and Idiolectal Features was significant with $p = 0.0262$; and the correlation between Colloquialisms and Elevated Language was significant with $p = 0.0447$.
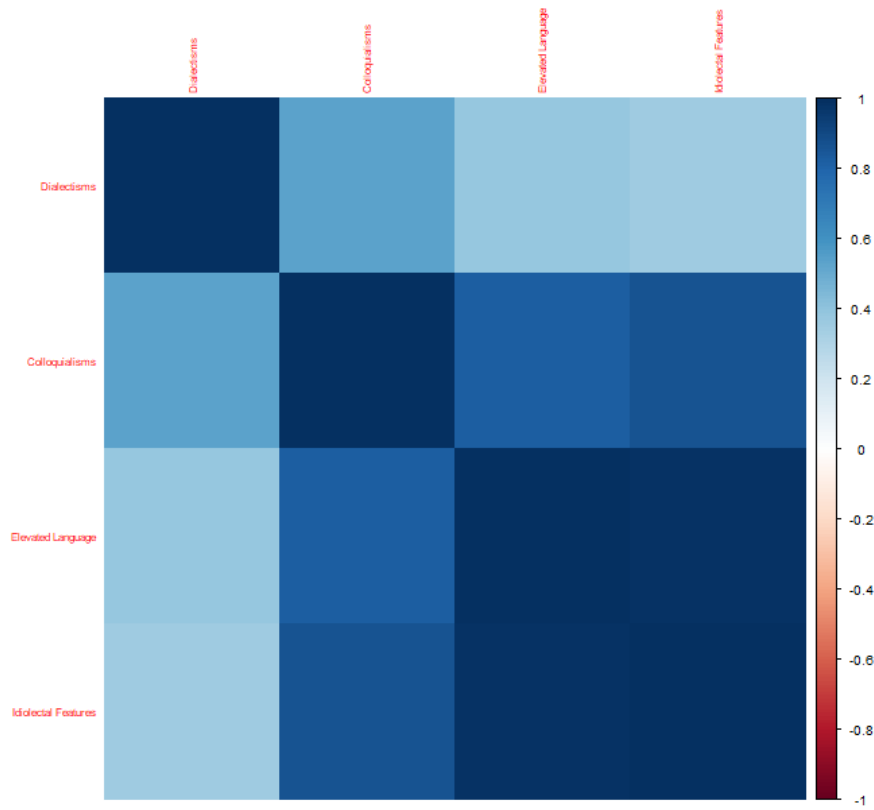
Figure 3 - Correlation matrix across marked categories

Discussion

This case study was undertaken with two primary goals: first, to illuminate the process of language creation within Giovanni Testori's theatrical works, thereby contributing to the understanding of this underappreciated author in Italian studies; second, to explore potential methodologies for quantitatively analysing languages that consciously and meaningfully deviate from standard norms. As a pilot study, this research highlights several aspects that require further refinement and exploration. The biggest limitation lies in the study's reliance on personal interpretation and manual annotation, which contrasts with the principles of reproducibility and reliability that ideally form the foundation of any scholarly work. Given the interdisciplinary nature of this research – situated at the intersection of linguistics and literary study – and the fact that it relied solely on my individual effort, a certain degree of subjectivity is inevitable. However, interpretative components are intrinsic to literary and stylistic studies and cannot, nor should they, be entirely removed.

In fact, this study arose precisely from the need to highlight the lack of objectivity in Testori studies. It may seem contradictory to have fallen into the same issue I initially set out to address. However, this is merely a superficial resemblance. Until now, critics – even when supported by fine sensibility and a solid background – have attempted broad interpretations of Testori's

linguistic creativity without relying on extensive observation of linguistic data, instead basing their conclusion on random sampling. This does not necessarily make their conclusions or observation incorrect, but rather partial. This partiality has also reinforced a pervasive belief (e.g. [58]: 68-9; [59]: 216), that Testori's linguistic creation was driven solely by an irrational and chaotic force, without logic or intentionality in its structure. My view, supported by the results of this study, is that each element operates within a coherent framework, conveying a precise effect, not merely an impression of chaos or overlapping linguistic and historical layers. Quantitative analysis helped uncover the patterns and concrete impact of each stylistic element, countering biased impression that might otherwise lead to unjustified foregrounding of certain aspects, as seen in the statistical analysis examples. While subjectivity and interpretation remain fundamental to this stylistic study, I believe it has provided a much deeper understanding of Testori's creative methods, offering a foundation for new critical interpretations.[35]

Regarding the methodological goal, substantial benefits could arise from extending this research with additional computational tools. For example, automated processes for category annotation could be developed by leveraging existing methods for error detection and correction or employing trained language model for annotation. These tools could mitigate issues such as human bias and fatigue, which often complicate manual annotations. However, developing such tools would require significant resources and collaboration, necessitating a more appealing or popular research topic to attract interest and funding.

While this study demonstrates how quantitative linguistic can deepen and refine literary interpretations, it is important to stress the need for cautious interpretation of the results: the quantitative data depend on the types of features included in the categories. For example, phonetic traits and functional words may quickly inflate counts, without increasing their interpretative significance. Moreover, meticulous attention must be paid to designing the categorization schema, as it forms the foundation of the analysis. Categorization should not only be linguistically motivated but also strategically aligned with the initial hypotheses about the contributions of each feature to the text.

Given the significant variation in creative language, it remains challenging to develop a universal methodology for analysis. This is where close collaboration between linguistic and literary scholars, developers and computer scientists is crucial. Together, they can tailor analytical approaches to the needs of individual case studies. However, despite their variability, creative literary languages often rely on a finite set of alteration strategies. These techniques, which differ more in their outputs than in their processes, include multilingualism (including dialects), the extension of suffixes and prefixes, gender switching, non-etymological phonetic alteration, word composition and fragmentation, syllabic inversion and reduplication, among others. Defining and formalizing these strategies, while integrating them with existing digital resources for NLP, could pave the way for future research in this area. Moreover, expanding studies to include other

---

[35] The interpretation I derived from the quantitative analysis particularly highlights the parodic effect of all linguistic components, except for dialect. Unlike authors such as Gadda or Arbasino ([64]), Testori does not aim to extensively embrace reality; rather, he seeks to de-functionalize language, exposing its artificial construction as a tool of oppressive institutional power. Conversely, to avoid an excessive comic effect, dialect serves as a means for the character to reconnect with their deeper bond to the land, the mother, the origins of their life, representing truth in opposition to falsity. This interpretation, which links stylistic and thematic analysis of the plays, also attempts to address a fundamental question in Testori's theatre: Why does he employ linguistic experimentation in certain plays but not in others? A broader discussion of my interpretation have been published in my PhD thesis and can be partially found in references [65] and [66].

authors and traditions could yield broader insights, contributing to the development of effective methodologies for analysing creative literary languages.

References

[1] Altieri Biagi, Maria Luisa.1980. *La lingua in scena.* Zanichelli.

[2] Ammon, Ulrich. 2004. "Standard variety". In *Sociolinguistics*, edited by Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier and Peter Trudgill, vol.1, 273-83. Walter de Gruyter. https://doi.org/10.1515/9783110141894.1.2.273.

[3] Anthony, Lawrence. 2022. *AntConc (Version 4.1.1).* Waseda University of Tokyo. https://www.laurenceanthony.net/software/antconc/.

[4] Bausani, Alessandro. 1974. Le lingue inventate. Linguaggi artificiali, linguaggi segreti, linguaggi universali. Ubaldini.

[5] Beretta, Claudio. 2003. Letteratura dialettale milanese. Itinerario antologico-critico dalle origini ai nostri giorni. Hoepli.

[6] Berruto, Gaetano. 2012. *Sociolinguistica dell'Italiano contemporaneo*. La Nuova Italia Scientifica.

[7] Bex, Tony, and Richard J. Watts, eds. 1999. *Standard English: The Widening Debate*. Routledge. 149-166. https://doi.org/10.4324/9780203054062.

[8] Biber, Douglas, Susan Conrad, and Randi Reppen. 2006. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489.

[9] Bolasco, Sergio, Giovanni De Gasperis, Lorenzo Buonanno, and Agnese Salutari. 2020. *TaLTaC (Version 4.0)*. https://www.taltac.com/.

[10] Bolasco, Sergio. 2021. *L'analisi automatica dei testi*. Carocci.

[11] Brevini, Franco. 1990. Le parole perdute. Dialetti e poesia nel nostro secolo. Einaudi.

[12] Burgess, Anthony. 1973. *Joysprick: An Introduction to The Language of James Joyce*. Andre Deutsch. https://doi.org/10.5040/9781350388376.ch-015.

[13] Chambers, J. K. 2003. *Sociolinguistic Theory*. 2nd Edition. Polity Press.

[14] Chiari, Isabella. 2007. Introduzione alla linguistica computazionale. Laterza.

[15] Chiari, Isabella. 2012. "Corpora e risorse linguistiche per l'italiano. Stato dell'arte, problemi e prospettive". *Italienisch* 68: 90-105.

[16] Chomsky, Noam. 1966. *Cartesian Linguistics*. Harper & Row.

[17] Chomsky, Noam. 1980. *Rules and Representations*. Basil Blackwell.

[18] Cohen, Marcel. 1918-9. "Note sur l'argot". *Bulletin de la Société de Linguistique de Paris* 21 (66-67): 132-47.

[19] Croce, Benedetto. 1902. Estetica come scienza dell'espressione e linguistica generale. Remo Sandron Editore.

[20] Demerson, Guy. 1981. "Le plurilinguisme chez Rabelais". *Bulletin de l'Association d'étude sur l'humanisme, la réforme et la renaissance* 14: 3-19. https://doi.org/10.3406/rhren.1981.1261.

[21] Den Ouden, Bernard D. 1975. *Language and Creativity. An Interdisciplinary Essay in Chomskyan Humanism.* The Peter de Ridder Press. https://doi.org/10.1515/9783110883473.

[22] Devoto, Giacomo. 1936. "Dal «Castello di Udine»". *Annali della R. Scuola Normale di Pisa* 5 (3): 187-210; now *Edinburgh Journal of Gadda Studies* 6 (supplement 7), 2007. https://www.gadda.ed.ac.uk/Pages/resources/reviews/devotoceg1936.php.

[23] Eco, Umberto. 1993. La ricerca della lingua perfetta nella cultura europea. Laterza.

[24] Fischer, Eugen. 2000. Linguistic creativity: exercises in philosophical therapy. Kluwer Academic Publishers.

[25] Folena, Gianfranco. 1991. Il linguaggio del caos. Studi sul plurilinguismo rinascimentale. Bollati Boringhieri.

[26] Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Dominik Gerstorfer, and Mari Akazawa. 2023. *CATMA (Version 7.0).* Zenodo. https://doi.org/10.5281/zenodo.12092195.

[27] Hutcheon, Linda. 1992. "The Complex Functions of Irony". *Revista Canadiense de Estudios Hispánicos* 16 (2): 219-34. http://www.jstor.org/stable/27762900.

[28] Kirschenblatt-Gimblett, Barbara and A.E. Backhouse. 1976. *Speech play: research and resources for studying linguistic creativity.* University of Pennsylvania Press.

[29] Labov, William. 1972. *Sociolinguistic Patterns.* Blackwell.

[30] Lebart, Ludovic, and André Salem. 1994. *Statistique textuelle.* Dunod.

[31] Lebart, Ludovic, André Salem, and Lisette Berry. 1998. *Exploring Textual Data.* Kluwer Academic Publishers.

[32] Leech, Geoffrey N. 1969. *A Linguistic Guide to English Poetry.* Longman. https://doi.org/10.4324/9781315836034.

[33] Loporcaro, Michele. 2009. Profilo linguistico dei dialetti italiani. Laterza.

[34] Manera, Lorenzo. 2022. "Estrangement". *International Lexicon of Aesthetics.* https://doi.org/10.7413/18258630121.

[35] Marrone, Caterina. 2004. *Le lingue utopiche.* Stampa Alternativa & Graffiti.

[36] Moretti, Franco. 2000. "Conjectures on World Literature". *New Left Review* 1: 54-68.

[37] Moretti, Franco. 2013. "'Operationalizing': Or, the Function of Measurement in Literary Theory". *New Left Review* 2 (84): 103-19.

[38] Neves, Mariana, and Jurica Ševa. 2021. "An extensive review of tools for manual annotation of documents". *Briefing in Bioinformatics* 22 (1): 146-63. https://doi.org/10.1093/bib/bbz130.

[39] Olsen, Mark. 1993. "Signs, Symbols, and Discourses: A New Direction for Computer-Aided Literary Studies". *Computers and the Humanities* 27 (5/6): 309-14. https://doi.org/10.1007/bf01829380.

[40] Peterson, David J. 2015. The Art of Language Invention: From Horse-Lords to Dark Elves to Sand Worms, the Words Behind World-Building. Penguin Publishing Group.

[41] Sanga, Glauco. 1999. "Il dialetto di Milano". *Rivista italiana di Dialettologia* 23: 137-64.

[42] Sanguineti, Edoardo. 1999. "Il plurilinguismo nelle scritture novecentesche". In *Transiti letterari e culturali*, edited by Giuseppe Sertoli e Goffredo Miglietta, vol. 1, 17-31. Edizioni Università di Trieste.

[43] Sobrero, Alberto A. 1993. Introduzione all'italiano contemporaneo. Vol. 2: La variazione e gli usi. Laterza.

[44] Spacks, Patricia M. 1961. "Logic And Language In 'Through the Looking Glass'". *ETC: A Review of General Semantics* 18 (1): 91-100. http://www.jstor.org/stable/42573885.

[45] Stria, Ida. 2016. *Inventing languages, inventing worlds. Towards a linguistic worldview for artificial languages*. Wydział Neofilologii UAM w. https://doi.org/10.14746/9788394760915.

[46] Taylor-Batty, Juliette. 2013. *Multilingualism in Modernist Fiction*. Palgrave MacMillan. http://dx.doi.org/10.1057/9781137367969.

[47] Terracini, Benvenuto. 1963. *Lingua libera e libertà linguistica*. Einaudi.

[48] Testori, Giovanni. 1994. *L'Hamblette*. Translated by Jean-Paul Manganaro. Dramaturgie.

[49] Testori, Giovanni. 1997. *Opere. Vol. 2 (1965-1977)*. Edited by Fulvio Panzeri. Bompiani.

[50] Testori, Giovanni. 2004. *Edipus*. Translated by Julie Quénehen. Deposited at Maison Antoine Vitez – Centre international de la traduction théâtrale. https://www.maisonantoinevitez.com/fr/bibliotheque/edipus-544.html.

[51] Testori, Giovanni. 2009. *Macbette*. Translated by Sylvia Bagli and Giampaolo Gotti. Manuscript deposited at Maison Antoine Vitez – Centre international de la traduction théâtrale. https://www.maisonantoinevitez.com/fr/bibliotheque/macbette-662.html.

[52] Testori, Giovanni. 2011. *désOrest'*. Translated by Sylvia Bagli and Giampaolo Gotti. Manuscript deposited at Maison Antoine Vitez – Centre international de la traduction théâtrale. https://www.maisonantoinevitez.com/fr/bibliotheque/desorest-761.html.

[53] Testori, Giovanni. 2013. *Opere. Vol. 3 (1977-1993)*. Edited by Fulvio Panzeri. Bompiani.

[54] Underwood, Ted. 2017. "A Genealogy of Distant Reading". *Digital Humanities Quarterly"* 11 (2). https://digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

[55] Verdonk, Peter. 2002. *Stylistics*. Oxford University Press.

[56] Voghera, Miriam. 1992. Sintassi e intonazione dell'italiano parlato. Il Mulino.

[57] De Mauro, Tullio. 2016. *Nuovo Vocabolario di base della lingua italiana (NVdB)*. https://dizionario.internazionale.it/.

[58] Lazzerini. Lucia. 1973. "*L'Ambleto*, o della barbarie ritrovata". *Paragone. Letteratura* 282: 56-73.

[59] Taffon, Giorgio. 1997. Lo scrivano, gli scarrozzanti, i templi. Giovanni Testori e il teatro. Bulzoni.

[60] D'Achille, Paolo. 2012. *Parole: al muro e in scena*. Cesati, 366-9.

[61] D'Onghia, Luca. 2017. "«L'Ambleto di Testori», ovvero Ruzante a Lomazzo. Schede storiche e linguistiche". *Quaderni veneti* 6 (1): 169-84. https://doi.org/10.14277/1724-188X/QV-6-1-17-10.

[62] D'Onghia, Luca. 2022. "Giovanni Testori". In *La letteratura dialettale milanese: autori e testi*, edited by Silvia Morgana, 1382-95. Salerno.

[63] Testori, Giovanni. 2023. *Opere scelte*. Edited by Giovanni Agosti. Mondadori ("I Meridiani").

[64] Rindaldi, Rinaldo. 1985. Il romanzo come deformazione. Autonomia ed eredità gaddiana in Mastronardi, Bianciardi, Testori, Arbasino. Mursia.

[65] Lilli, Silvia. 2022. "Testori, la lingua e il teatro". *Campi immaginabili* 66-67: 212-44.

[66] Lilli, Silvia. 2025. "Il pantano della parola: la deformazione linguistica nel teatro di Testori". In *Per Giovanni Testori. Il teatro*, edited by Isabella Becherucci, Fabio Pierangeli, Carlo Serafini, 15-32. Paolo Loffredo.