information

Article

Linguistic Profiling of Text Genres: An Exploration of Fictional vs. Non-Fictional Texts

Akshay Mendhakar

Faculty of Applied Linguistics, Uniwersytet Warszawsk, 00-927 Warszawa, Poland; a.mendhakar@uw.edu.pl

Abstract: Texts are composed for multiple audiences and for numerous purposes. Each form of text follows a set of guidelines and structure to serve the purpose of writing. A common way of grouping texts is into text types. Describing these text types in terms of their linguistic characteristics is called 'linguistic profiling of texts'. In this paper, we highlight the linguistic features that characterize a text type. The findings of the present study highlight the importance of parts of speech distribution and tenses as the most important microscopic linguistic characteristics of the text. Additionally, we demonstrate the importance of other linguistic characteristics of texts and their relative importance (top 25th, 50th and 75th percentile) in linguistic profiling. The results are discussed with the use case of genre and subgenre classifications with classification accuracies of 89 and 73 percentile, respectively.

Keywords: genres; subgenres; linguistic profiling; text; NLP

check for updates

Citation: Mendhakar, A. Linguistic Profiling of Text Genres: An Exploration of Fictional vs. Non-Fictional Texts. *Information* **2022**, *13*, 357. https://doi.org/10.3390/info13080357

Academic Editor: Peter Revesz

Received: 21 June 2022
Accepted: 22 July 2022
Published: 26 July 2022

1. Introduction

With the advancement in computers and their processing abilities, powerful algorithms that can process complex data in seconds have led to the development of modern-day natural language processing (NLP) algorithms. Present-day NLP techniques focus on both the context and form of the text rather than focusing on just one of them.

The development of sophisticated NLP pipelines and the availability of multiple large-scale corpora have given rise to a new range of data-driven NLP tools. These modern tools can be used to answer classical linguistic research topics and many more topics with relative ease. By accomplishing this, we can highlight the set of linguistic variables which are suited for the given task and try training machine learning algorithms to build models for a given task. These models represent a text type based on its linguistic features and can be used for solving complex linguistic problems when coupled with complex statistical methods. One such classical linguistics problem is identifying text patterns and highlighting the linguistic characteristics/linguistic profiling [1]. This traditional question has led to multiple advanced concepts such as genre identification [2], identification of one's native language [3], author identification [4], author attribution [5], author verification [6] etc.

Similar complex linguistic use cases have given rise to areas such as computational register analysis [7], which looks at the register and genre variation from a functional spectrum of context-driven linguistic differences; computational sociolinguistics [8], which focuses on the social dimension of language and the underlying diversity associated with it; computational stylometry is aimed at extracting knowledge from texts to verify and attribute authorship [9]; and many more. While classical stylometric techniques place a special emphasis on identifying the most salient or the rarest feature in a text, modern techniques can uncover patterns even in smaller segments of text [1,10]. Identifying a specific linguistic profile of different text types can be used for classification tasks and measurement of readability [11].

2. Literature Review

The concept of linguistic profiling for identifying specific features is not new and has been attempted by multiple researchers. However, the usage of linguistic profiling to

understand the genre variation computationally is the focus of this review. Ref [12] was the first to propose the multi-dimensional (MD) method for genre variation. The MD approach has several salient characteristics [13]:

1. MD is a corpus-driven computational method, defined based on the analysis of a naturally occurring large number of texts.
2. MD helps in identifying linguistic features and patterns in individual texts and genres computationally.
3. MD is built on the hypothesis that different types of texts differ linguistically and functionally and that analysing only one or two of them is insufficient for reaching inferences.
4. MD is, as the name suggests, an explicitly multi-dimensional approach that assumes that in any discourse, it is anticipated that numerous parameters of variation will be active.
5. MD is quantitative in nature, i.e., early statistical techniques as reported by [14,15] have been reported to be useful in measuring frequency counts of linguistic features. Recent multivariate statistical techniques are useful in understanding the relationship between linguistic elements of the text.
6. MD combines macro- and micro-level analysis. That is, macroscopic evaluation of general linguistic patterns combined with microscopic measurement of specific features of the specific texts.

In earlier days, the knowledge extraction methods for register and stylistic analysis focused on the extraction of simple language-specific features such as pronouns, prepositions, auxiliary and modal verbs, conjunctions, determiners, etc. and a few language-independent features such as frequency of linguistic features. Significant progress in information extraction from text has lately been made feasible because of the creation of strong and reasonably accurate text analysis pipelines for a variety of languages [9]. This is also true in all the aforementioned instances where NLP-based technologies that automate the feature extraction process play a critical role. Various programmes exist now that utilize distinctive kinds of features to evaluate register, stylistic, or linguistic complexity.

Among these, the Stylo package [16] provides a comprehensive and user-friendly set of functions for stylometric studies. Stylo focuses on shallow text characteristics, such as n-grams at the token and character levels, that may be automatically extracted without the usage of language-dependent annotation tools. It should be noted, however, that it can also handle the output of linguistic annotation software. Text complexity may also be assessed using a variety of other tools. Coh-Metrix is a well-known example which uses characteristics retrieved from multi-level linguistic annotation to calculate over 200 indices of cohesion, language and readability from an input text [17]. Similarly, L2 Syntactical Complexity Analyzer (L2SCA) [18] and TAASSC [19] both estimate multiple linguistic variables that highlight grammatical complexity at the phrasal and sentence levels. These types of features are relevant in studies on first and second language acquisition. SweGram, a system specifically designed to profile texts in the Swedish language [20], is a striking exception to the preceding technologies, which are all designed for the English language. From this brief review, we can note that language-independent tools, such as Stylo, typically use shallow features that do not require language-specific preprocessing, whereas techniques based on a wide variety of multilevel linguistic features are frequently monolingual.

Profiling–UD [21] is a computational text analysis tool based on linguistic profiling concepts. It allows for the extraction of over 130 linguistic features from the given text. Because it is built on the Universal Dependencies framework, Profiling–UD is a multilingual tool that supports 59 languages. The features extracted from the tool can be grouped under raw text-related properties, lexical variety related features, morpho-syntactic features, verbal predicate structure-based measures, Global and Local Parsed Tree Structures, syntactic and subordination related measures. Table 1 highlights the information on feature categories extracted from the Profiling–UD tool. For more details about the tool, visit http://linguistic-profiling.italianlp.it/ (accessed on 2 March 2022).

Table 1. Features extracted from Profiling–UD.

Category of Feature	Definition of the Feature	Name as Seen in the Tool
Raw text features	This measures raw text features such as document length, sentence and word lengths, number and characters per token.	(n_sentences), (n_tokens), (tokens_per_sent), (char_per_tok)
Lexical variety	Measured in terms of its Type/Token Ratio (TTR) for both the first 100 and 200 tokens of a text in lemma and form. The TTR value ranges from one (high TTR) to zero (low lexical variety).	(ttr_lemma_chunks_100), (ttr_lemma_chunks_200), (ttr_form_chunks_100), (ttr_form_chunks_200)
Morpho–syntactic information	These measures highlight the percentage distribution of 17 core part-of-speech categories defined in the Universal POS tags, the lexical density of content words and inflectional morphology.	(upos_dist_*): distribution of the 17 core part-of-speech categories and (lexical_density), (verbs_tense_dist_*), (verbs_mood_dist_*), (verbs_form_dist_*), (verbs_gender_dist_*), (verbs_num_pers_dist_*)
Verbal predicate structure	This estimates the distribution of verbal heads and roots.	(verbal_head_per_sent), (verbal_root_perc), (avg_verb_edges), (verb_edges_dist_*)
Global and local parsed tree structures	These measure the average depth of the syntactic tree, average clause length, length of dependency links, the average depth of embedded complement chains governed by a nominal head, word order phenomena.	(avg_max_depth), (avg_token_per_clause), (avg_links_len), (avg_max_links_len), (max_links_len), (avg_prepositional_chain_len), (n_prepositional_chains), (prep_dist_*), (obj_pre), (obj_post), (subj_pre), (subj_post)
Syntactic relations	This estimates the distribution of dependency relations of 37 universal syntactic relations used in UD.	37 (dep_dist_*)
Subordination phenomena/Use of Subordination	This evaluates the distribution of subordinate and main clauses, the relative order of subordinates concerning the verbal head and the average depth of embedded subordinate clauses.	(principal_proposition_dist), (subordinate_proposition_dist), (subordinate_post), (subordinate_pre), (avg_subordinate_chain_len), (subordinate_dist_*)

There have been increasingly large collections of data compiled across the internet. With advancements in technologies, these datasets are annotated and automatically analysed for multiple purposes [22]. However, linguistic profiling of texts is usually carried out for multiple different projects with a variety of end goals in mind. Language verification, author identification and verification, and text classification are a few to highlight here. Our focus is to identify specific linguistic features of a given text that influence the text classification into genres and specific subgenres. A brief review of the studies which have focused on linguistic profiling of fictional and non-fictional texts points to the study by [11], where they tried to estimate the readability of Italian fictional prose based on the linguistic profiling of the texts. Even though their study shows promising results, from a fictional prose point of view the dataset considered in the study is devoid of the fictional texts or does not cover most of the subgenres of the fictional type. Therefore, it is very important to conduct studies that consider multiple fictional subgenres that are popularly noted in the literature and compare their linguistic composition with the non-fictional text type. In the study by [11], the four major categories considered were literature further divided into children and adult literature, journalism (newspaper), educational writing (educational materials for primary school and high school) and scientific prose. When we look at the datasets which are utilized across literature for the task of classification or readability or

author identification, we note that the Brown Corpus [23], the Lancaster-Oslo/Bergen (LOB) Corpus [24] and the British National Corpus (BNC) (The BNC is the result of a collaboration, supported by the Science and Engineering Research Council (SERC Grant No. GR/F99847), and the UK Department of Trade and Industry, between Oxford University Press (lead partner), Longman Group Ltd. (London, UK), Chambers Harrap (London, UK), Oxford University Computer Services, the British Library and Lancaster University) are the most prevalently used ones. Even though the nature of the BNC is the availability of a large mixed corpus which renders a possibility to analyse multiple genres of texts, it is not suitable for understanding comparing genres of fiction and non-fiction in detail. The Brown Corpus consists of over 500 samples coded under 15 genres as an early attempt at corpus linguistics. These 15 genres represented are not the universally accepted classification of genres. In fact, when the scope of the study is to measure readability or classification, the available datasets are acceptable. However, if we are interested in understanding the linguistic composition of various genres and subgenres of fictional and non-fictional texts, it is crucial that we define what we consider genres and subgenres of texts. Genre is a fluid concept which is always in constant flux due to the vast majority of researchers proposing different classification systems and different research goals. As the scope of our study is to highlight the linguistic similarities and differences in various subgenres of fiction and non-fictional texts, it is very important to consider a new dataset suitable for the goal of the experiment.

The goal of the present study was to investigate variation within and between genres by comparing a corpus of literary texts to corpora representing other textual genres using contrastive linguistic analysis.

3. Method

The study was carried out at the LELO laboratory located at the Institute of Specialized Studies (IKSI), Faculty of Applied Linguistics at the University of Warsaw. The study was carried out after obtaining ethical clearance from the local ethical committee at the University of Warsaw. The methodology section is divided into three sections, the first part deals with the corpora and the related preprocessing of the dataset. The second part deals with the linguistic profiling results of individual genres. The third section highlights the results of the classifier performance based on the linguistic profiled features for genre identification.

3.1. Corpora and Preprocessing

For the creation of corpus, we considered the text classification of [25] (fictional, non-fictional and poetry). We choose to ignore the category of poetry, as it is beyond the scope of our study. Further classification into subgenres was performed after considering the Reading Habits Questionnaire (RHQ) by [26]. Table 2 highlights the subgenre classification considered for the creation of corpus. The data for the corpus was gathered from various sources. The data for the fictional texts were gathered from the Gutenberg project. Project Gutenberg is a digital archive of over 65,000 books categorized under various subheadings and can be accessed in multiple formats such as HTML, PDF, TXT, EPUB, MOBI etc. All the materials downloaded from the Gutenberg project are covered under the Creative Commons license which makes them ready to use for this research study. Project Gutenberg is an online repository of texts such as short stories, novels, poems, biographies and many more. Despite being smaller than other public collections such as HathiTrust [27] and Google Books, Project Gutenberg has several advantages over those collections. It can be downloaded as a single package or can be scrolled for individual texts, which makes it versatile enough for multiple experiments. Also, most of the online repositories of digital documents use OCR technology to convert and preserve the documents. Texts under Project Gutenberg have been proofread by a human and in some cases even hand-typed, making them more suitable for experimental usage. The texts needed for the non-fiction were gathered from student writing samples of http://www.corestandards.org/ (accessed

on 2 March 2022) [28] and various articles from the procedural texts we chose different projects/articles from the https://www.instructables.com/ (accessed on 2 March 2022) [29] website. Instructables is a dedicated web portal to obtain step-by-step instructions in building and carrying out a variety of projects.

Table 2. Summary of the dataset of the study.

Fiction (2153)	Non-Fiction (1514)
Children's Fiction (190)	Discussion Texts (395)
Fable (394)	Explanatory Texts (242)
Fantasy (249)	Instructional Texts (495)
Legends (44)	Persuasive Texts (382)
Mystery (191)	
Myths (48)	
Romance (591)	
Science Fiction (385)	
Thriller (61)	

Hence, we built a dataset which consists of both fictional and non-fictional texts with a special focus on carrying out a detailed linguistic analysis. Table 2 highlights the number of text samples (shown in brackets) considered in each subgenre grouped across fictional and non-fictional genres. The selected texts were divided into chapters, and it was made sure that the overall size of each of the texts would be around 100–2000 words. Preprocessing of the selected text was carried out to remove licensing information, unnecessary spaces and punctuation.

3.2. Linguistic Profiling of Texts

The scope of the present study is to carry out detailed linguistic profiling of various texts in the fictional and non-fictional categories. We chose to use the tool called Profiling–UD [21] for carrying out a detailed computational linguistic profiling of texts. As stated in the previous sections, this tool provides the most comprehensive set of features for a loaded text.

Each text was individually loaded onto the tool and corresponding features were extracted and tabulated. This process was repeated for all the texts. The results obtained from the analysis were loaded onto SPSS software [30] for further processing.

Even though the analysis of fictional and non-fictional texts was performed based on chapter-wise text, it can be noted that the overall number of sentences and number of tokens in the fictional texts are higher than in non-fictional texts. Table 3 shows the summary of the raw textual features across all subgenres and genres. However, the number of tokens per sentence and character per token is higher in non-fictional texts when compared to fictional texts. It was noted that there were individual differences across subgenres in terms of the number of sentences and tokens. Based on the raw text properties, it can be noted that mystery and thrillers, myths and legends, and fantasy and romance subgenres had similar raw text structures; whereas explanatory and persuasive texts had similar scores in the noted raw text properties. These findings support the hypothesis of [31] that non-fictional texts, notably informational and discussion texts, use substantially longer words and sentences than fictional texts, which use short and easy phrases.

Table 4 highlights the lexical variety noted in the subgenres, and it can be said that based on the values there were no statistical differences between them. However, it can be noted that the subgenres of fables had simple lexical variety and complexity and can be graded as even simpler than the non-fictional texts. This can be accredited to the population that the fables are targeted for—children need simple lexical variety. These findings add to the claims that fictional texts and subgenres report significantly higher TTR values suggesting greater lexical diversity and usage of unique words [32].

Table 3. Summary of the raw textual features across genres.

Parameter/Subgenre	n_sentences	n_tokens	tokens_per_sent	char_per_tok
Children's Fiction	559.40	5593.40	9.98	3.96
Fable	17.00	147.00	8.58	4.25
Fantasy	410.80	4399.80	10.65	4.10
Legends	680.80	7034.00	10.37	4.11
Mystery	981.80	9342.00	9.83	4.25
Myths	1378.20	13,339.80	9.62	4.57
Romance	644.40	6597.00	10.14	4.22
Science Fiction	551.80	5269.80	9.68	4.48
Thriller	1027.20	8921.20	8.78	4.27
Discussion	58.40	1191.00	19.69	4.73
Explanatory	176.80	1759.80	10.00	4.41
Instructional	147.60	1398.80	9.56	4.25
Persuasive	60.20	577.20	9.83	4.53
Fiction	694.60	6738.22	9.74	4.25
Non-fictional	110.75	1231.70	12.27	4.48

Table 4. Summary of the lexical variety features across genres.

Parameter/Subgenre	ttr_lemma_chunks_100	ttr_lemma_chunks_200	ttr_form_chunks_100	ttr_form_chunks_200
Children's Fiction	0.69	0.56	0.76	0.64
Fable	0.23	0.10	0.26	0.11
Fantasy	0.66	0.54	0.73	0.60
Legends	0.64	0.54	0.71	0.61
Mystery	0.72	0.61	0.79	0.68
Myths	0.68	0.60	0.73	0.65
Romance	0.68	0.61	0.74	0.66
Science Fiction	0.69	0.61	0.74	0.66
Thriller	0.69	0.62	0.75	0.67
Discussion	0.66	0.58	0.74	0.66
Explanatory	0.61	0.52	0.68	0.59
Instructional	0.67	0.59	0.76	0.66
Persuasive	0.62	0.48	0.71	0.56
Fiction	0.64	0.54	0.72	0.62
Non-fictional	0.63	0.53	0.69	0.59

Similarly, we looked at the parts of speech distribution in the various subgenres. Table 3 highlights the individual values of the distribution of parts of speech across various subgenres. When the values are compared across fiction and non-fictional texts, it can be noted that fictional texts have a lower number of adjectives but a higher number of adverbs, adpositions, pronouns and punctuation when compared to non-fictional texts. Whereas non-fictional texts have two times higher values of auxiliary verbs and nouns with slightly elevated values in numbers compared to fictional texts. No significant differences were noted in the values of coordinating and subordinating conjunctions, determiners, interjections, symbols and pronouns across fictional and non-fictional texts. Overall, the lexical density of fictional and non-fictional texts remained the same. Table 5 highlights the parts of speech distribution across all subgenres.

According to the Universal Dependencies (UD) framework, parts of speech can be divided into three types [33]. Figure 1 highlights this classification system, and it includes open class (ADJ, ADV, NOUN, VERB, PROPN, INTJ), closed class words (ADP, AUX, CONJ, DET, NUM, PART, PRON, SCONJ) and others (PUNCT, SYS, X). For more details, refer to Figure 1.

Table 5. Summary of parts of speech across and lexical density of various genres.

Parameter Subgenre	* ADJ	* ADP	* ADV	* AUX	* CCONJ	* DET	* INTJ	* NOUN	* NUM	* PART	* PRON	* PROPN	* PUNCT	* SCONJ	* SYM	* VERB	* X	Lexical Density
Children's Fiction	4.48	7.39	6.81	5.82	3.89	7.66	0.60	12.44	0.50	2.77	11.33	3.37	18.76	1.47	0.12	12.56	0.05	0.49
Fable	4.09	8.99	5.83	3.23	4.42	12.98	0.00	15.49	0.42	3.46	10.25	2.70	11.22	2.29	0.00	14.64	0.00	0.48
Fantasy	4.58	10.37	6.01	4.06	5.56	9.53	0.25	15.61	0.71	1.85	10.35	4.17	13.08	1.90	0.02	11.93	0.01	0.49
Legends	4.09	11.17	4.38	4.59	4.49	10.27	0.38	17.37	1.48	1.63	9.31	4.34	15.15	1.03	0.06	10.07	0.19	0.47
Mystery	5.46	9.87	6.48	5.78	3.19	9.44	0.35	14.49	0.66	2.02	11.55	1.77	15.66	2.28	0.17	10.77	0.07	0.46
Myths	6.22	11.41	3.76	4.15	3.77	10.61	0.16	18.19	1.17	1.42	5.98	7.12	15.12	1.35	0.23	8.89	0.45	0.52
Romance	4.79	9.09	5.48	5.95	3.51	7.44	0.44	13.56	0.49	2.64	11.68	4.20	16.80	1.93	0.05	11.91	0.03	0.48
Science Fiction	6.86	10.52	5.99	5.16	3.72	9.91	0.19	17.73	0.88	2.09	8.79	2.03	13.38	1.87	0.12	10.72	0.03	0.50
Thriller	5.80	10.33	5.65	5.31	2.68	9.28	0.26	15.45	0.41	2.06	12.02	2.43	14.89	1.76	0.11	11.57	0.01	0.48
Discussion	5.15	9.90	5.60	4.89	3.91	9.68	0.29	15.59	0.75	2.22	10.14	3.57	14.90	1.76	0.10	11.45	0.09	0.49
Explanatory	6.70	8.36	5.31	6.80	3.54	7.51	0.05	20.65	0.80	3.62	7.19	2.86	11.13	2.67	0.03	12.74	0.06	0.54
Instructional	6.83	8.66	5.29	6.66	3.42	10.54	0.26	21.33	1.13	2.53	6.68	1.45	11.68	1.77	0.07	11.66	0.05	0.53
Persuasive	5.26	8.14	3.68	2.73	2.70	10.75	0.18	22.46	4.30	2.01	5.06	7.47	12.19	1.30	0.43	10.88	0.44	0.57
Fiction	6.94	6.78	5.27	6.70	3.19	7.90	0.08	17.53	0.07	4.79	12.82	0.60	9.17	3.00	0.00	15.18	0.00	0.50
Non-fictional	6.43	7.99	4.89	5.72	3.21	9.18	0.14	20.49	1.58	3.24	7.94	3.10	11.04	2.19	0.13	12.62	0.14	0.54

* The parameters are upos_dist_.

	Tag	Description	Example
Open Class	ADJ	Adjective: noun modifiers describing properties	*nice, yellow, old*
	ADV	Adverb: verb modifiers of time, place, manner	*quick, yesterday*
	NOUN	Noun: words for persons, places, things, etc.	*India, apple, beauty*
	VERB	Verb: words for actions and processes	*doing, dancing*
	PROPN	Proper noun: name of a person, organization, place, etc.	*Akshay, TCS*
	INTJ	Interjection: greeting, polar questions	*yes, hello, oh*
Closed Class Words	ADP	Adposition (Preposition/postposition): describes a noun's	*over, around, on*
		spacial, temporal and/or other relation	
	AUX	Auxillary: words which describe tense, mood etc.	*should, can, may*
	CCONJ	Coordinating conjunctions: words joining two phrases	*and, or*
	DET	Determiner: marks noun phrase properties	*a, an, the*
	NUM	Numerals: numbers	*one, two, three, fourth*
	PART	Particle: a preposition-like form used together with a verb	*up, down, in, at*
	PRON	Pronoun: a shorthand for referring to an entity or event	*who, others*
	SCONJ	Subordinating Conjunction: join a main clause with a	*which, that*
		subordinate clause such as a sentential complement	
Other	PUNCT	Punctuation	*?, .*
	SYM	Symbols	*$, %*
	X	Other	*asdf, qwfg*

Figure 1. Universal Dependencies (UD) tagset by [34].

When we carefully examine the parts of speech distribution across non-fictional texts, we can note that instructional texts had significantly fewer adjectives, adverbs and auxiliary verbs when compared to other non-fictional texts. Also, the concentration of proper nouns in instructional texts is statistically higher than in any other text. Persuasive texts have a statistically significant fewer number of nouns, punctuation and adpositions, but higher values in pronouns and verbs overall. Based on the values of determiners, particle structure, subordinate conjunctions and interjections, we can group non-fictional texts into two groups: discussion–persuasive and explanatory–instructional. No significant differences were noted in the lexical density across all the subgenres. Therefore, it can be noted that open class and closed class words are equally important in the classification of texts into fictional and non-fictional genres.

Similarly, when we look at the subgenres of the fictional texts, we can note that myths and science fiction texts have the highest and lowest concentration of open-class words (specifically adjectives and adverbs, respectively) but this is not statically significant, whereas fables and children's fiction have the least concentration of open-class words (interjections) in the non-fictional text genre. No other significant differences in closed-class words were noted across other subgenres of fiction. Adverbs are the fewest in myths but others were statically insignificant. Auxiliary verbs and coordinating conjunctions were the fewest in children's fiction and thrillers, respectively, but were similar across all the other domains. Nouns are the fewest in children's fiction but are similar in all other domains. Children's fiction and romance had similar closed-class compositions. Myths and legends have the highest number of numerals and proper nouns, and the least occurrence of pronouns and verbs compared to all other subgenres. Lexical density, particles, punctuation, subordinating conjunctions, symbols and other domains are insignificant and are similar across all domains.

Table 5 highlights the part of speech distribution in the different text types. Pronouns and verbs appear to be frequently occurring in non-fictional texts. These findings are

similar to the claim by [31] that these two elements are more common in conversation than in written language forms. The frequency of occurrence of nouns, on the other hand, is relatively low, resulting in a substantially lower noun/verb ratio. These findings are in line with the findings of [35], who suggest that novels have a narrative structure with a plot involved that requires the description of activities using verbs.

Further, when we look at the other morphosyntactic information such as inflectional morphology, the distribution of verbs according to their tense pre and post showed significant differences across fictional vs non-fictional texts. Fictional texts had higher past tense verbs whereas non-fictional texts are composed more of present tense verbs. Looking at the indicative and imperative verbs in fictional versus non-fictional texts, it was found that both kinds of texts are extensively composed of indicative verbs.

No statistical differences in the distribution of verbs according to their number and person, their tenses or even verbal mood were noted. Fables had the highest concentration of past tenses whereas persuasive texts had the lowest concentration of past tenses.

Syntactic features of verbal predicate structures, such as the verbal heads in the document, roots headed by a lemma tagged as a verb, verbal arity and distribution of verbs for arity class, were not found to be significantly different across the subgenres. Further, there were no significant differences noted in the parsed tree structures either, except that the prepositional chains for non-fictional texts had significantly lower values compared to fictional texts. However, fables had the smallest concentration of prepositional chains making their structure closer to non-fictional texts.

When studying the order of elements in syntactic structure, specifically the objects and subjects preceding and following the verbs, it was noted that fictional texts had slightly higher values, but this did not reach statistical significance. When examined individually, it was noted that the objects preceding verbs were least for fables and similar to the non-fictional category where there was not much difference across other subgenres.

Further contrastive analysis of 37 universal syntactic relations was carried out across fictional and non-fictional texts. It was observed that non-fictional texts had elevated values of the clausal modifier of the noun (adnominal clause), adjectival modifier, compound, phrasal verb particle, marker, numeric modifier, object, oblique nominal, which reached statistical significance, but non-significantly different values in adverbial clause modifier and punctuation.

In the use of subordination, none of the parameters reached statistical significance across fictional and non-fictional texts, but slight differences in the values of the distribution of principal clauses and subordinate clauses were noted. No further subgenre differences were noted.

One of the aims of this experiment was to highlight the main features that can be used for the classification of fictional and non-fictional categories for the task of genre classification.

4. Feature Reduction and Classification

We begin by providing a quick overview of the classification algorithm and feature selection approaches we employed in our trials (Section 4.1). Following that, we discuss the classification models that were trained on the dataset using the proposed feature sets (Section 4.2). The next section includes a feature selection experiment in which we evaluate the relevance of the features (Section 4.3). The next step is to re-run the classification methods using alternative subsets of the features to evaluate how this affects the model's accuracy.

4.1. The Classification Algorithm and the Feature Selection Methods

In this study, we utilised the Random Forest algorithm (RF), which is an ensemble learning method, as our classifier. The classification is based on the outcomes of several decision trees it generates during the training process [36,37]. We chose RF as it calculates the permutation relevance of the variables reliably during training the classification models. Table 6 highlights the feature details after dimensionality reduction. After that, we em-

ployed Rank Features by Importance (RFI) and Sequential Forward Search (SFS) to evaluate the features included in each model. Sections 4.3 and 4.4 explain RFI and SFS in detail.

Table 6. Feature details after dimensionality reduction.

Linguistic Category	Old Size	New Size for the Genre	New Size for Subgenre	Ignored Features Genre/Subgenre
Raw Text Properties	4	4	4	0/0
Lexical Variety	4	4	2	0/2
Morphosyntactic Information				
- upos_dist	18	11	13	7/5
- lexical_density	1	1	1	0/0
Inflectional Morphology	21	17	17	4/4
Syntactic Features				
- Verbal Predicate Structure	10	3	5	7/5
- Global and Local Parsed Tree Structures	10	7	8	3/2
- Order of Elements	4	3	3	1/1
Syntactic Relations				
- dep_dist	44	24	30	20/14
- Use of Subordination	8	7	5	0/3

4.2. Constructing RF Models

We utilised Jupyter Notebook for a quick implementation of RF. The features were evaluated in a sequential manner to predict the importance of each feature in the models' prediction success.

4.3. Using RFI to Assess the Relevance of the Features: Experiment One

To evaluate the variables, we used RF's built-in permutation importance [38] to rank their "importance". According to [39], the model is developed first, and its accuracy is computed in out-of-bag (OOB) observations to determine the relevance of the feature (Xi). Following that, any relationship between the values of Xi and the model's outcome is severed by permuting all the values of Xi, and the model's accuracy with the permuted values is re-computed. The permutation importance of Xi is defined as the difference between the accuracy of the new model and the original score. As a result, if a feature has noise or random values, the permutation is unlikely to affect the accuracy. A large difference between the two rates, on the other hand, indicates the importance of the feature for the prediction task. Figures 2 and 3 demonstrate the importance of several variables in genre and subgenre classifier models. The greater the relevance of the feature, the greater the value of the mean decrease in accuracy on the x-axis.

We also used the method of [40] to calculate the p-values for the variables under the null hypothesis that the permutation of the variable has no effect on the accuracy. Out of 131 features, 89 and 83 features from the genre and subgenre models, respectively, were found to have a significant effect on classifier models. The remaining features had a role in the models to varying degrees which did not reach significance.

4.4. Measuring Relevance of the Features Using SFS—Experiment Two

To implement SFS, we used the R package mlr [41]. The algorithm starts with an empty set of features and gradually adds variables until the performance of the model no longer improves. In this model, we used the classif.randomForest learner and the Holdout resampling method. If the improvement falls below the minimum needed value (alpha = 0.01), the algorithm comes to a halt. Each box in Figure 4 shows the selected features of each feature set.

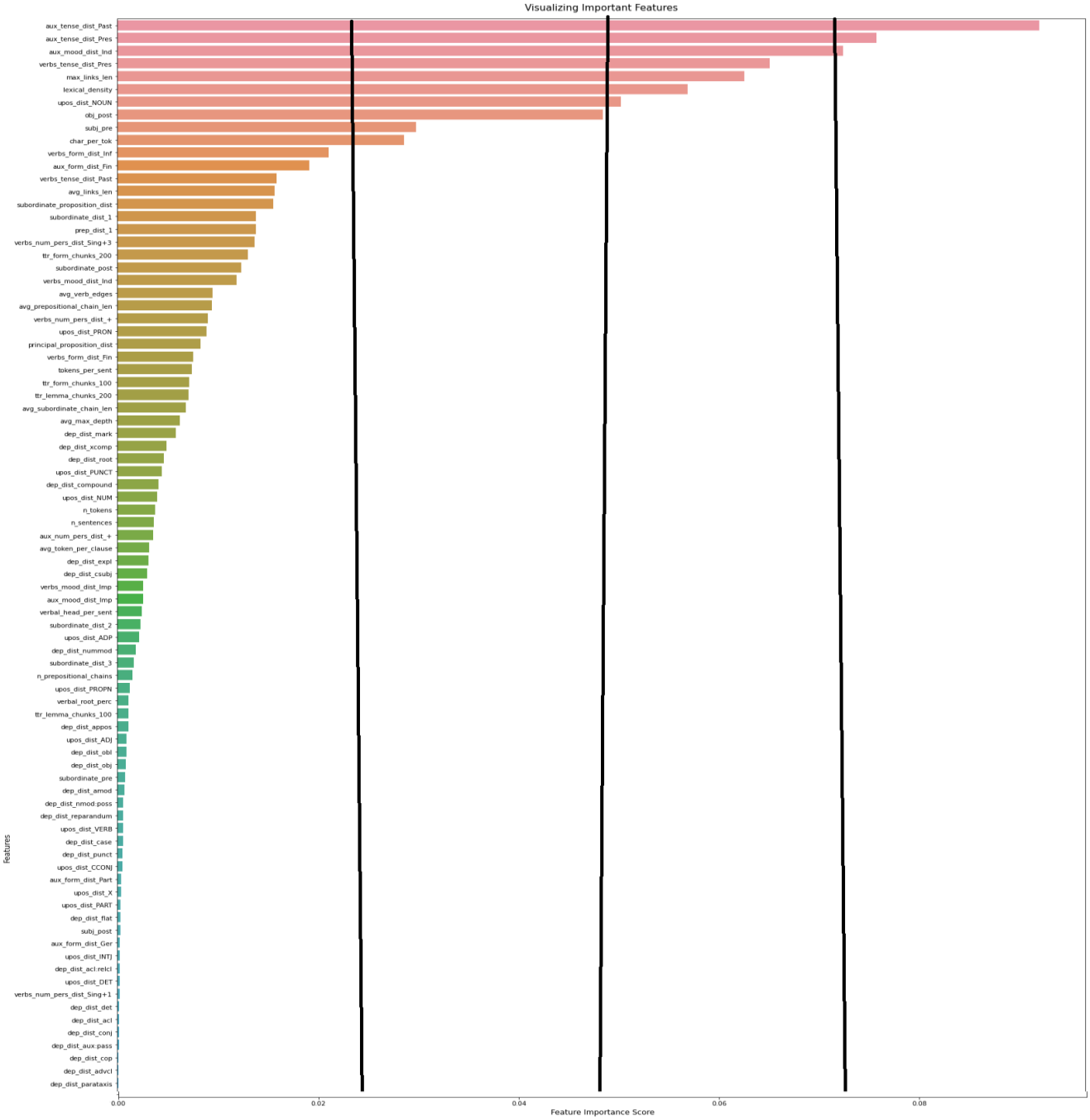

Figure 2. Variable importance plot of the RF genre model. NOTE: The x-axis shows the permutation relevance (mean decrease in accuracy) of each feature; the y-axis lists the features of the genre model.

4.5. Examining Various Feature Subsets Based on Their Significance—Experiment Three

Firstly, in Section 4.5.1, we explore the accuracy of different subsets of each feature set based on the results of the RFI and SFS experiments. In Section 4.5.2, we explore the subsets of all the features combined, trying to come up with an optimal consensus set of features.

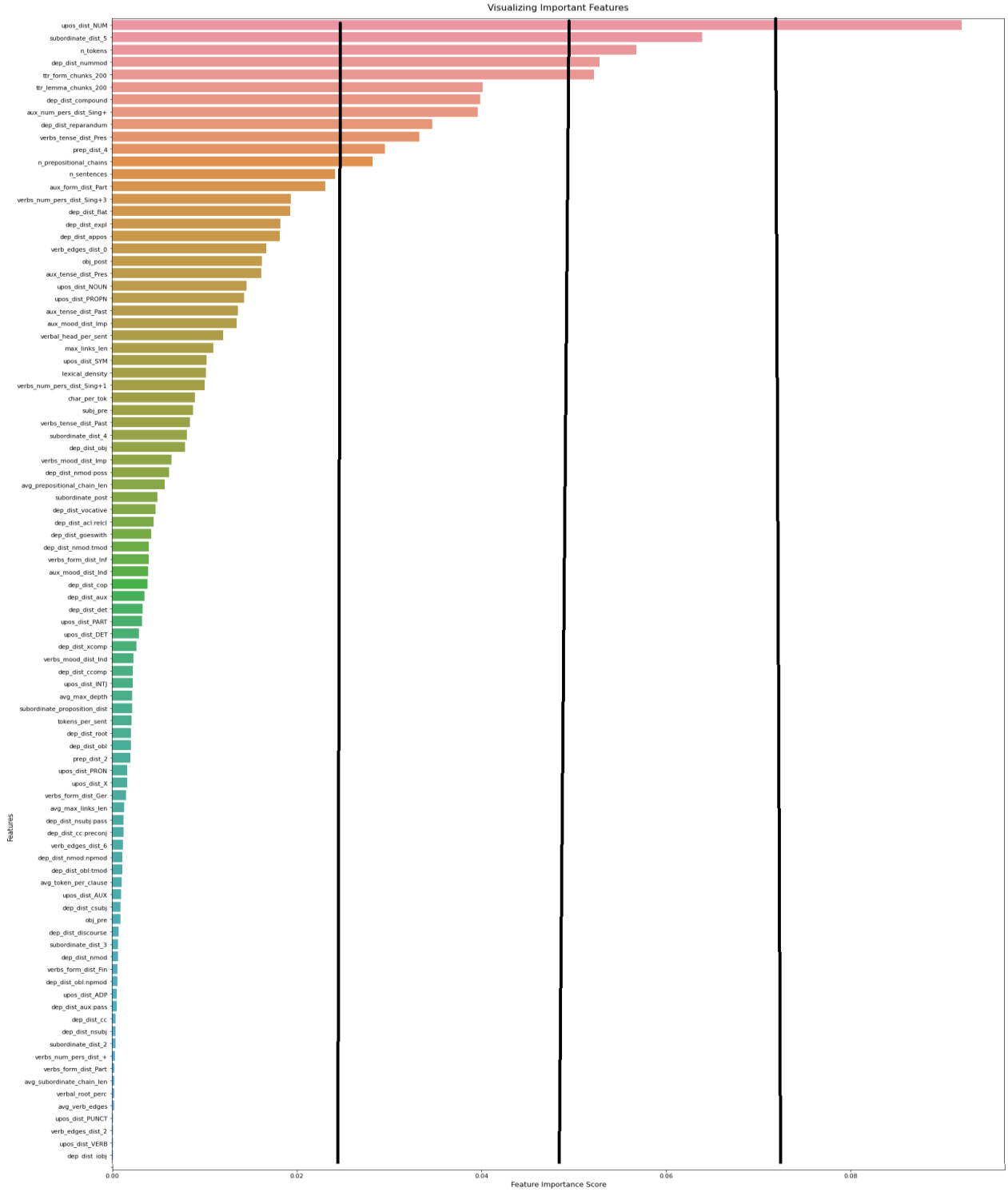

Figure 3. Variable importance plot of the RF sub-genre model. NOTE: The x-axis shows the permutation relevance (mean decrease in accuracy) of each feature; the y-axis lists the features of the subgenre model.

SUB-GENRE	GENRE
upos_dist_NUM	aux_tense_dist_Past
n_tokens	aux_tense_dist_Pres
ttr_form_chunks_200	aux_mood_dist_Ind
ttr_lemma_chunks_200	verbs_tense_dist_Pres
	lexical_density

Figure 4. SFS optimal features of each feature set.

4.5.1. Subsets of Each Feature Set

Tables 7 and 8 highlight the list of features that are considered with greater importance in genre determination, and the top 25th, 50th and 75th percentile of the features important for classification can be noted in Figure 4. The initial accuracy of each model is reported in the first row of Table 9 with the term original. Rows Top 25%, Top 50% and Top 75% report, respectively, on performing RF on the 25th, 50th and 75th percentile of important features for class determination. Similarly, the Top 5 row highlights the relevance of the first five features of each feature set with the greatest importance (according to Figure 4). The row Allimp highlights the results of applying the RF to all the features that were noted to play a part in classification.

Table 7. Subgenre selection of top features.

Top 25	Top 50	Top 75
upos_dist_NUM	upos_dist_NUM, subordinate_dist_5, n_tokens, dep_dist_nummod, ttr_form_chunks_200	upos_dist_NUM, subordinate_dist_5, n_tokens, dep_dist_nummod, ttr_form_chunks_200, ttr_lemma_chunks_200, dep_dist_compounds, aux_num_pers_dist_Sing+, dep_dist_reparandum, verbs_tense_dist_Pres, prep_dist_4, n_prepositional_chains

Table 8. Genre selection of top features.

Top 25	Top 50	Top 75
aux_tense_dist_Past, aux_tense_dist_Pres, aux_mood_dist_Ind	aux_tense_dist_Past, aux_tense_dist_Pres, aux_mood_dist_Ind, verbs_tense_dist_Pres, max_links_len, lexical_density, upos_dist_NOUN	aux_tense_dist_Past, aux_tense_dist_Pres, aux_mood_dist_Ind, verbs_tense_dist_Pres, max_links_len, lexical_density, upos_dist_NOUN, obj_post, subj_pre, char_per_token

Table 9. Accuracy of the model with feature selection.

Row Name	Genre Data	Subgenre Data
Original	0.87	0.82
Top 25%	0.71	0.64
Top 50%	0.77	0.71
Top 75%	0.84	0.79
Top 5	0.75	0.68
Allimp	0.93	0.89

4.5.2. Subsets of All Features

To investigate the possible combinations of all features based on the findings of the RFI and SFS tests, after trying out different subsets of each feature set.

1. In the RFI experiment, we initially applied RF to the set of attributes with the highest permutation relevance. The set, as shown in Figure 3, is {aux_tense_dist_Past, aux_tense_dist_Pres, aux_mood_dist_Ind, verbs_tense_dist_Pres}. The accuracy of this model is 0.889. From Figure 4, the set of features important are {upos_dist_NUM, subordinate_dist_5, n_tokens, dep_dist_nummod, ttr_form_chunks_200, ttr_lemma_chunks_200, dep_dist_compounds}. This model had an accuracy of 0.728.
2. The union of RF and the two most important features of each feature set: {*aux_tense_dist_*Past, aux_tense_dist_Pres, aux_mood_dist_Ind, verbs_tense_dist_Pres, max_links_len, lexical_density, upos_dist_NOUN, obj_post, subj_pre, char_per_token}. The accuracy of this model is 0.913. Similarly, for subgenre model, {*upos_dist_NUM, subordinate_dist_5, n_tokens, dep_dist_nummod, ttr_form_chunks_200, ttr_lemma_chunks_200, dep_dist_compounds, aux_num_pers_dist_Sing+, dep_dist_reparandum, verbs_tense_dist_Pres, prep_dist_4, n_prepositional_chains*} revealed the model accuracy of 0.792. The accuracy of this model is in line with the expected increase in the accuracy when compared with the accuracy of the union of the single most relevant features.

5. Conclusions

In this paper, we tried to linguistically profile the features noted in various fictional and non-fictional subgenres. By considering multiple feature sets highlighted in various computational SRF studies from a linguistic perspective, we attempted to connect the computational models and the linguistic explanations behind those features. As a result of the experiment, we are able to linguistically grade the composition of texts that constitute a text type. We also noted that for the task of genre classification the most important set of features are inflectional morphology, morphosyntactic information and raw text properties. However, for the task of subgenre classification, a mixture of semantic and syntactic features is important, i.e., morphosyntactic information, use of subordination, lexical variety, general syntactic features and parsed tree structures.

Based on the linguistic profiling of non-fictional texts we found that the linguistic composition of discussion and persuasion texts are similar across most of the domains of comparison, and explanatory and instructional texts show linguistic similarities as well. Similarly, grouping of subgenres of fiction can be performed for dyads of children's fiction and fantasy, myths and legends, and mystery and thrillers.

The results of the present study highlight the use of exact estimates of linguistic elements in each text type. These estimates could be useful in planning future use case experiments ranging from identifying developmental patterns in children [42,43] to estimating atypical language acquisition [44,45]. Further, we can also detect linguistic markers for acquired language disorders and cognitive impairments such as dementia and aphasia [46]. Similarly, we can estimate the writing abilities of school children [47]. Furthermore, from the perspective of computational sociolinguistics, the findings aid in the analysis of variations in the social component of language [8] as well as the modelling of stylometric features of authors [9]. By performing a comprehensive estimation of elements belonging to morphological, semantic and syntactic domains, we are able to grade the text types in terms of their complexities as well. This will be especially useful in such cases as readability measurement and selection of specific texts for language learning, among others.

Similarly, the current trend in linguistic analysis is to use complex network models for linguistic representation [48–50]. Complex networks have been used to model and study many linguistic phenomena, such as complexity [51–53], semantics [54], citations [55], stylometry [56–61] and genre classification [62–64]. Multiple studies [65,66] have concluded that the different properties of specific words on the macroscopic scale structure of a whole text are as relevant as their microscopic feature such as frequency of appearance. Linguistic research from the complex network approach is a relatively young domain of scientific

endeavour. There is still a need for studies that can fill the gap in understanding the relationships between the system-level complexity of human language and microscopic linguistic features [48]. Although research in this area is on the rise and abundant findings have already been made, researchers need to have a clear knowledge of the microscopic linguistic features to determine the directions of further research. Our study highlights the crucial microscopic linguistic features which can be used to build better complex network models.

Even though the present study was comprehensive with the linguistic parameters considered, the dataset used was unevenly distributed across fictional and non-fictional text groups. Further studies which can address these issues and replicate the results of the present study in a controlled dataset would be required.

Funding: This research was conducted as part of the ELIT project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie, grant agreement no. 860516.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I wish to thank all the reviewers and Monika Płużyczka, Eliza, Niharika, Priyanka, Darshan and Deepak for helpful comments and discussion. I am also extremely grateful to all the members of IKSI at the University of Warsaw, who helped me in completing this research work.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Halteren, H.V. Linguistic Profiling for Authorship Recognition and Verification. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 21–26 July 2004.
2. Paltridge, B. Genre Analysis and the Identification of Textual Boundaries. *Appl. Linguist.* **1994**, *15*, 288–299. [CrossRef]
3. Cimino, A.; Wieling, M.; Dell'Orletta, F.; Montemagni, S.; Venturi, G. Identifying Predictive Features for Textual Genre Classification: The Key Role of Syntax. In Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it, Rome, Italy, 11–13 December 2017.
4. Coulthard, M. Author Identification, Idiolect, and Linguistic Uniqueness. *Appl. Linguist.* **2004**, *25*, 431–447. [CrossRef]
5. Gamon, M. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In Proceedings of the COLING 2004: 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23–27 August 2004; pp. 611–617.
6. Halteren, H.V. Author verification by linguistic profiling: An exploration of the parameter space. *ACM Trans. Speech Lang. Processing* **2007**, *4*, 1–17. [CrossRef]
7. Argamon, S.E. Computational Register Analysis and Synthesis. *Regist. Stud.* **2019**, *1*, 100–135. [CrossRef]
8. Nguyen, D.; Doğruöz, A.S.; Rosé, C.P.; De Jong, F.M. Computational Sociolinguistics: A Survey. *Comput. Linguist.* **2016**, *42*, 537–593. [CrossRef]
9. Daelemans, W. Explanation in computational stylometry. In *International Conference on Intelligent Text Processing and Computa-tional Linguistics*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 451–462.
10. Montemagni, S. Tecnologie Linguistico-Computazionali E Monitoraggio Della Lingua Italiana. *Studi Ital. Linguist. Te-Orica Appl. (SILTA)* **2013**, *42*, 145–172.
11. Dell'Orletta, F.; Montemagni, S.; Venturi, G. Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, Hissar, Bulgaria, 9–11 September 2013; pp. 189–197.
12. Biber, D. Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language* **1986**, *62*, 384. [CrossRef]
13. Biber, D. *Variation across Speech and Writing*; Cambridge University Press: Cambridge, UK, 1988. [CrossRef]
14. Carroll, J.B. Vectors of Prose Style. In *Style in Language*; Sebeok, T.A., Ed.; MIT Press: Cambridge, MA, USA, 1960; pp. 283–292.
15. Marckworth, M.L.; Baker, W.J. A discriminant function analysis of co-variation of a number of syntactic devices in five prose genres. *Am. J. Comput. Linguist.* **1974**, *11*, 2–24.
16. Eder, M.; Rybicki, J.; Kestemont, M.; Eder, M.M. Stylometry with R: A package for computational text analysis. *R Journal* **2016**, *8*, 107–121. Available online: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html (accessed on 2 March 2022).

17. Graesser, A.C.; McNamara, D.S.; Cai, Z.; Conley, M.; Li, H.; Pennebaker, J. Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *Elem. Sch. J.* **2014**, *115*, 210–229. [CrossRef]
18. Lu, X. Automatic analysis of syntactic complexity in second language writing. *Int. J. Corpus Linguist.* **2010**, *15*, 474–496. [CrossRef]
19. Kyle, K. Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Ph.D. Thesis, Georgia State University, Atlanta, GA, USA, 2016. [CrossRef]
20. Näsman, J.; Megyesi, B.; Palmér, A. SWEGRAM: A Web-Based Tool for Automatic Annotation and Analysis of Swedish Texts. In Proceedings of the 21st Nordic Conference on Computational Linguistics, Nodalida, Gothenburg, Sweden, 22–24 May 2017; pp. 132–141.
21. Brunato, D.; Cimino, A.; Dell'Orletta, F.; Venturi, G.; Montemagni, S. Profiling-UD: A tool for linguistic profiling of texts. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 7145–7151.
22. Lu, X. *Computational Methods for Corpus Annotation and Analysis*; Springer: Berlin, Germany, 2014. [CrossRef]
23. Francis, W.N.; Kucera, H. *Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers*; Technical Report; Department of Linguistics, Brown University: Providence, RI, USA, 1964.
24. Johansson, S.; Leech, G.N.; Goodluck, H. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*; Department of English, University of Oslo: Oslo, Norway, 1978.
25. National Literacy Trust (Adapted from Crown Copyright). A Guide to Text Types: Narrative, Non-Fiction and Poetry. 2013. Available online: https://www.thomastallisschool.com/uploads/2/2/8/7/2287089/guide_to_text_types_final-1.pdf (accessed on 2 March 2022).
26. Kuijpers, M.M.; Douglas, S.; Kuiken, D. Capturing the Ways We Read. *Anglistik* **2020**, *31*, 53–69. [CrossRef]
27. Christenson, H.A. HathiTrust. *Libr. Resour. Tech. Serv.* **2011**, *55*, 93–102. [CrossRef]
28. Schutz, D. The Common Core State Standards Initiative. 2011. Available online: http://www.corestandards.org/ (accessed on 26 March 2022).
29. Wikipedia Contributors. Instructables. In Wikipedia, The Free Encyclopedia. Available online: https://en.wikipedia.org/w/index.php?title=Instructables&oldid=1024372150 (accessed on 26 March 2022).
30. IBM Corp. *Released. IBM SPSS Statistics for Windows*; Version 26.0; IBM Corp.: Armonk, NY, USA, 2019.
31. Biber, D.; Conrad, S. *Register, Genre, and Style*; Cambridge University Press: Cambridge, UK, 2009.
32. Jacobs, A.M. (Neuro-)Cognitive poetics and computational stylistics. *Sci. Study Lit.* **2018**, *8*, 165–208. [CrossRef]
33. Nivre, J.; De Marneffe, M.C.; Ginter, F.; Goldberg, Y.; Hajic, J.; Ryan Petrov, S.; Pyysalo, S.; Sil-veira, N.; Tsarfaty, R.; Zeman, D. Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 1659–1666.
34. Nivre, J.; de Marneffe, M.C.; Ginter, F.; Hajič, J.; Manning, C.D.; Pyysalo, S.; Sebastian, S.; Tyers, F.; Zeman, D. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020. 2020.
35. Voghera, M. La misura delle categorie sintattiche. Parole e numeri. In *Analisi Quantitative dei Fatti di Lingua*; Aracne: Roma, Italy, 2005; pp. 125–138.
36. Nayak, A.; Natarajan, D. Comparative study of naive Bayes, support vector machine and random forest classifiers in sentiment analysis of twitter feeds. *Int. J. Adv. Stud. Comput. Sci. Eng.* **2016**, *5*, 16.
37. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional variable importance for random forests. *BMC Bioinform.* **2008**, *9*, 307. [CrossRef]
40. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef]
41. Bischl, B.; Lang, M.; Kotthoff, L.; Schiffner, J.; Richter, J.; Studerus, E.; Casalicchio, G.; Jones, Z.M. mlr: Machine Learning in R. *J. Mach. Learn. Res.* **2016**, *17*, 5938–5942.
42. Lu, X. Automatic measurement of syntactic complexity in child language acquisition. *Int. J. Corpus Linguist.* **2009**, *14*, 3–28. [CrossRef]
43. Lubetich, S.; Sagae, K. Data-driven measurement of child language development with simple syntactic templates. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 2151–2160.
44. Prud'hommeaux, E.; Roark, B.; Black, L.M.; Van Santen, J. Classification of atypical language in autism. In Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, Portland, OR, USA, 23 June 2011; pp. 88–96.
45. Rouhizadeh, M.; Sproat, R.; Van Santen, J. Similarity measures for quantifying restrictive and repetitive behavior in conversations of autistic children. In Proceedings of the Conference Association for Computational Linguistics North American Chapter, Meeting, Seattle, DC, USA, 29 April–4 May 2015; p. 117.
46. Roark, B.; Mitchell, M.; Hollingshead, K. Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, Translational, and Clinical Language Processing*; Association for Computational Linguistics: Cambridge, MA, USA, 2007; pp. 1–8. [CrossRef]

47. Barbagli, A.; Lucisano, P.; Dell'Orletta, F.; Montemagni, S.; Venturi, G. CItA: An L1 Italian learners corpus to study the development of writing competence. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 88–95.
48. Cong, J.; Liu, H. Approaching human language with complex networks. *Phys. Life Rev.* **2014**, *11*, 598–618. [CrossRef]
49. Gao, Y.; Liang, W.; Shi, Y.; Huang, Q. Comparison of directed and weighted co-occurrence networks of six languages. *Phys. A Stat. Mech. Appl.* **2013**, *393*, 579–589. [CrossRef]
50. Lužar, B.; Levnajić, Z.; Povh, J.; Perc, M. Community structure and the evolution of interdisciplinarity in Slovenia's sci-entific collaboration network. *PLoS ONE* **2014**, *9*, e94429. [CrossRef] [PubMed]
51. Amancio, D.R.; Oliveira, O.N., Jr.; Costa, L.D.F. Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Phys. Stat. Mech. Appl.* **2012**, *391*, 4406–4419. [CrossRef]
52. Segarra, S.; Eisen, M.; Ribeiro, A. Authorship attribution using function words adjacency networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–30 May 2013; pp. 5563–5567. [CrossRef]
53. Segarra, S.; Eisen, M.; Ribeiro, A. Authorship Attribution Through Function Word Adjacency Networks. *IEEE Trans. Signal Process.* **2015**, *63*, 5464–5478. [CrossRef]
54. Silva, T.C.; Amancio, D.R. Word sense disambiguation via high order of learning in complex networks. *Eur. Lett.* **2012**, *98*, 58001. [CrossRef]
55. Amancio, D.R.; Nunes, M.D.G.V.; Oliveira, O.N., Jr.; Costa, L.D.F. Using complex networks concepts to assess approaches for citations in scientific papers. *Scientometrics* **2012**, *91*, 827–842. [CrossRef]
56. Brede, M.; Newth, D. Patterns in syntactic dependency networks from authored and randomised texts. *Complex. InterNatl.* **2008**, *12*, 051915.
57. Liang, W.; Shi, Y.; Tse, C.K.; Liu, J.; Wang, Y.; Cui, X. Comparison of co-occurrence networks of the Chinese and English languages. *Phys. Stat. Mech. Appl.* **2009**, *388*, 4901–4909. [CrossRef]
58. Liang, W.; Shi, Y.; Tse, C.K.; Wang, Y. Study on co-occurrence character networks from Chinese essays in different periods. *Sci. China Inf. Sci.* **2012**, *55*, 2417–2427. [CrossRef]
59. Liu, H.; Li, W. Language clusters based on linguistic complex networks. *Chin. Sci. Bull.* **2010**, *55*, 3458–3465. [CrossRef]
60. Antiqueira, L.; Nunes, M.D.G.V.; Oliveira, O.N., Jr.; Costa, L.D.F. Strong correlations between text quality and complex networks features. *Phys. Stat. Mech. Appl.* **2007**, *373*, 811–820. [CrossRef]
61. Amancio, D.R.; Antiqueira, L.; Pardo, T.A.; Costa, L.D.F.; Oliveira, O.N., Jr.; Nunes, M.G. Complex net-works analysis of manual and machine translations. *Int. J. Mod. Phys. C* **2008**, *19*, 583–598. [CrossRef]
62. Amancio, D.R.; Oliveira, O.N.; Costa, L.D.F. Identification of literary movements using complex networks to represent texts. *New J. Phys.* **2012**, *14*, 043029. [CrossRef]
63. Costa, L.D.F.; Oliveira, O.N., Jr.; Travieso, G.; Rodrigues, F.A.; Villas Boas, P.R.; Antiqueira, L.; Viana, M.P.; Correa Rocha, L.E. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Adv. Phys.* **2011**, *60*, 329–412. [CrossRef]
64. Newman, M.E.; Barabási, A.L.E.; Watts, D.J. *The Structure and Dynamics of Networks*; Princeton University Press: Princeton, NJ, USA, 2022.
65. Ke, J.; Yao, Y. Analysing Language Development from a Network Approach. *J. Quant. Linguist.* **2008**, *15*, 70–99. [CrossRef]
66. Akimushkin, C.; Amancio, D.R.; Oliveira, O.N., Jr. Text authorship identified using the dynamics of word co-occurrence networks. *PLoS ONE* **2017**, *12*, e0170527. [CrossRef]