A machine learning approach to literary genre classification on Portuguese texts: circumventing NLP's standard varieties

Dionéia Motta Monte-Serrat, Mateus Tarcinalli Machado
Evandro Eduardo Seron Ruiz

[1]Department of Computing and Mathematics – FFCLRP
University of São Paulo – Ribeirão Preto, Brazil

di_motta61@yahoo.com.br, {mateusmachado, evandro}@usp.br

Abstract. *We evaluate and classify bottom-up and quali-quantitatively literary genres from the BDCamões corpus. Chronicles, novels, short stories, and tales, annotated in UD, are classified by random forests and analyzed based on the Brazilian-Portuguese version of the LIWC dictionary. Results by class are reported by the mean value, along with a measure of variability. The results for features by class, LIWC tags, part of speech, and Universal Dependency tags highlight the higher positive and lower negative features. Adapting this method to the fluidity and mutability of literary genres circumvents the difficulty of NLP's standard tools, showing consistency and fewer errors in the results.*

Resumo. *Avaliamos e classificamos quali-quantitativamente gêneros literários do corpus BDCamões. Crônicas, romances, histórias curtas e contos, anotados em UD, são classificados por florestas aleatórias, e analisados com base na versão português-brasileira do LIWC. Os resultados por classe são reportados pela média, juntamente com uma medida de desvio padrão. Os resultados das características por classe, rótulos LIWC, classes gramaticais e rótulos UD destacam características positivas altas e negativas baixas. A adaptação desta metodologia à fluidez e mutabilidade dos gêneros literários contorna as dificuldade normalemnet encontradas em NLP, apresentando consistência e poucos erros nos resultados.*

1. Introduction

Classifying documents according to their genre is proposed as a primary task in text and document processing. Eissen and Stein [Sven Meyer zu and Stein 2004] claim genre classification is usually performed by discriminating documents through their form, their style, or their targeted audience. They also assure that genre classification is orthogonal to a classification based on the documents' contents. This study considers the artistic content of the analyzed texts since it focuses on literary genres, which materialize the style or form adopted by the authors, being, therefore, closely linked to the organizational goals and processes through which writing went.

The textual genre has a broad conception due to the socio-communicative function of documents, which encompasses the intention, context, social function, among other aspects of the document [Martins 2008]. Text is formed from expressed ideas. These ideas, structured in textual types, are also determined from interconnected factors, such as socio-historical context, physical and subjective structures of individuals,

grammatical rules, social function, and interlocutors. This multiplicity of factors is materialized, in an unstructured way, in messages such as posts on social networks, scientific articles, books, news, reports, and others, giving rise to different textual genres. This article is dedicated to studying the classification of literary genres, a specific text classification task within natural language processing (NLP). Textual genres usually rely on their differences in textual structure and leave behind a wide range of contextual factors. The challenge in analyzing literary genres lies in non-structured data from context inputs that blend into words and expressions to carry meaning to the grammatical reality of the text [Bronckart 2004, Monte-Serrat 2021, Monte-Serrat and Cattani 2021b, Monte-Serrat and Cattani 2021a]. Textual genres are fluid and changeable, continuously adapting to new social needs [Bronckart 2004, Matos 2021]. Suppose the focus falls on only one of the features (only on a contextual feature or only on the human language's logic/grammatical feature). In that case, the differentiation process might be deficient, carrying weakness that may be transmitted to the final findings.

Deepening the knowledge about literary genres classification can improve web search and information access to extensive digital collections [Crowston and Kwasnik 2003, Karlgren et al. 1998]. If not only to improve information retrieval, distinguishing different literary genres is a demanding task. The difference between poetry and a novel may be evident for an educated human being, but it may not be easy to discriminate a short story from a chronicle.

The main objective of this work is to study quantitative features that may differentiate literary genres under the scope of the NLP subject. We aim to address some corpus annotation features that may help find the regularity underlying the literary genres of the corpus of the BDCamões [Grilo et al. 2020]. This article is divided into five sections. Section 2 deals with related research on textual genre identification, showing that genre knowledge makes information more easily understood by search tools. Section 3 describes data and methods, discussing the corpus of textual genres BDCamões annotated according to the Universal Dependency framework. The results are described in Section 4, where they are segmented by classes reported by the mean value and its standard deviation. We also highlight the importance of features per class, LIWC labels, part-of-speech and Universal Dependency labels. Section 5 addresses the discussion of the results, showing from a bottom-up perspective how these NLP tools adapt to the contextual characteristics of fluidity and mutability of literary genres. We conclude in Section 6 that the analysis of literary genres must contemplate their artistic content to identify the aesthetics of the 'best order or structure' of words or the aesthetics of the 'best word'. For this, we suggest a combination of tools that balance structural and contextual aspects, making the machine more intuitive with more consistent results.

2. Related work

Some papers focus on genre identification for information retrieval in extensive digital collections. When characterized only by the textual form, this identification is not enough to define an information problem. This identification is due to the interdependent relations, in textual genres, between linguistic units and contextual parameters [Schneuwly 1997]. Search results containing genre information are easier understood by search tools, as the genre is often an implicit notion.

Karlgren and collaborators [Karlgren et al. 1998] make iterative information retrieval through topic grouping to build a multidimensional pre-representation interface of the research results. This way, the authors enrich the information search dialogue, encouraging and supporting the iterative refinement of queries, and enrich the document representation beyond the simple semantics of the terms frequencies.

Statamatos and co-authors [Stamatatos et al. 2000] use word frequency from The Wall Street Journal training corpus. Compared to the most frequent words in plain English cited in the British National Corpus, the authors claim that the latter contains more reliable discriminators for classifying text genres than the most frequent words in their limited-size training corpus. Similarly, Feldman and his team [Feldman et al. 2009] proposed using part-of-speech (POS) histogram statistics to perform the classification of textual genres. Together with a quadratic discriminant classifier, they show better performance than techniques that use word frequency counting features and POS tri-gram features. The authors claim that it is unclear what techniques would be needed to cover the entire feature space and differentiate the sub-classes, suggesting to characterize the genre more generalizable, as a multitasking learning, replacing singular genre classes with multiple factors.

Nilan et al. [Nilan et al. 2001] employ a bottom-up approach to analyze perceptions of textual resources that assist users in characterizing documents. Using content analytic techniques, the authors derive a set of genres built around the actual use of the web to compare existing genre lists. Mark Rosso [Rosso 2005] reports a study in which users classify genres according to a palette for use in web retrieval. Each participant received a pile of 102 printed web pages and was asked to separate the pages into piles according to the genre. The level of agreement reached 60%. In another study, the author analyzes 18 genres in an online experiment in which 257 subjects. The agreement rate reached more than 70% regarding the textual genre.

Omar [Omar 2020] brings together the Vector Space Clustering (VSC), 'concept pool' (BOC), explicit semantic analysis (ASE), and ConceptNet methods to address the classification of literary genres. They show that the computational and semantic models approach results to achieve better performance in the classification task. The author claims that the dimensionality of the data makes it difficult to obtain reliable analytical results, suggesting classification only in the most critical or distinct resources available.

3. Data and Methods

Data

The corpus of textual genres used in this research is the BDCamões Collection of Portuguese literary documents [Grilo et al. 2020]. Some features of the BDCamões corpus make it very useful for research in NLP: it is composed of 4 million words from more than 200 complete documents written by 83 authors in 14 genres. Its literary texts range from the 16th to the 21st century and have been carefully edited. The dimensionality of the data makes it difficult to obtain reliable analytical results [Omar 2020]. Monte-Serrat and Cattani [Monte-Serrat and Cattani 2021a] mention the curse of dimensionality in data interpretation. These arguments support the choice of the BDCamões corpus to find reliable results.

The corpus includes automatically annotated linguistic information such as grammatical classes, morphological features, grammatical dependencies based on the Universal Dependency framework (UD) [Nivre 2015], and expressions denoting named entities. BDCamões brings classified texts according to the following literary genres: 92 tales; 26 chronicles; 25 novels; 21 short stories; 18 poems; 11 theater plays; 8 essays; 1 travel guide; 1 sermon; 1 other; 1 narrative; 1 memoir; 1 letter; 1 anthology, totaling 208 documents and 3,945,943 words.

Methods

We chose random decision forests [Breiman 2001], an ensemble learning method, for the literary genre classification task. Random forest is a popular machine learning algorithm consisting of a combination of tree classifiers. Each classifier is generated using a random vector sampled independently from the input vector. Every tree casts a single vote, choosing the most popular class to classify an input vector. Decision trees seek to find the best split to subset the data based on the features provided to the learning phase.

Linguistic Inquiry and Word Count (LIWC) [Tausczik and Pennebaker 2010] is a text analysis system created by Pennebaker and collaborators [Pennebaker et al. 2001] with the aim of grouping words into categories that can be used to analyze psycholinguistic characteristics in different types of texts, making this tool interesting for the assessment of literary genres. LIWC is composed of software tools and a lexicon/dictionary. Each LIWC dictionary entry can be assigned to one or more categories (The word 'like' can belong to the category 'pronoun' or 'discrepancy' or 'affection' or even 'simile'). LIWC includes 17 standard linguistic dimensions (e.g., word count, percentage of pronouns, articles), 25 word categories tapping psychological constructs (e.g., affect, cognition), 10 dimensions related to "relativity" (time, space, motion), and 19 personal concern categories (e.g., work, home, leisure activities)': "LIWC successfully measures positive and negative emotions, a number of cognitive strategies, several types of thematic content, and various language composition elements" [Pennebaker et al. 2015]. The core of this program is known as the LIWC dictionary, which was made available for the Brazilian-Portuguese. In this research, we used this Brazilian version of LIWC 2007 [Balage Filho et al. 2013] to present the best and worst combinations of categories in analyzing texts from the BDCamões corpus.

BDCamões delivers different numbers of classified texts according to literary genres. To pursue consistent results, we selected only the literary genres that had more than ten text samples. These were: tale (96), novel (25), short story (21), chronicle (26), and poetry (18). As mentioned before, the corpus was already annotated with part-of-speech and UD labels. We used these annotations as features for the random forest classifier, and we also added word categories labels obtained from the Brazilian Portuguese LIWC [Balage Filho et al. 2013]. As for the latter, this dictionary/lexicon is composed of 64 word classes. Many words have multiple class labels. In this case, all word labels were added as features for the classifier. We did not use words as a feature, as our goal is to seek a categorization of texts focused on their structure, not on the content.

A grid search method was applied to find the best parameters to train a random forests model. For this, we used stratified 3-fold cross-validation together with a total of 4,320 combinations of parameters. The best combination of parameters was chosen by

calculating the F-measure. Once the best parameters were found, we configured a new classifier model using these parameters with a stratified 5-fold cross-validation scheme.

In order to analyze and understand the importance of the selected characteristics in the classification, we used the Python language module Eli5[1]. Eli5 allows the explanation of weights and predictions made by machine learning models. The weights of each attribute are calculated by following decision paths in all trees created by the classification model. Each node of the tree has an output score. The feature contribution on the decision path is calculated using the score difference from a parent to child node. Weights of all features sum to the output score or probability of the estimator.

4. Results for literary genre classification

The results based on a classification of resources' average among all classes of literary genres make the data more similar to the expected target domain of each genre. See Table 1. This table implies a balance between two aspects that literary genres commonly bring embedded: i) the assessment that considers the textual type (in which what matters most is the structural organization providing specific sequences for narration, description, exposition, argumentation) [Marcuschi et al. 2002]; ii) and the assessment of the socio-historical aspects, text function, media, type, and adequacy of language, among others, which influence that textual type.

Table 1 shows the results for literary genre classification based on the comprehensive set of part-of-speech, UD, and LIWC features. The poetry genre obtained the best classification scores for precision and recall measures among the five genres tested, reaching a harmonic mean of 88% (SD 11%). Although the chronicle genre obtained 100% precision, 29% (SD 29%) recall contributed to the second-lowest F-measure (24%) among all the genres. Novels also obtained inferior results, which reflected an F-measure of 11%. Contrary to our initial guess, tales and short stories might have very few in common regarding these classification features. Tales presented an F-measure of 77% with the lowest SD of 7%. On the opposite side, short stories presented higher percentages of standard deviation for precision and recall, matching the final F-measure to its SD.

The weighted average is a metric computed this way: find the corresponding metric's average for each class weighted by the number of true instances for each label. Then compute the average among all these classes. The weighted average values for precision, recall, and F-measure reflect the inconsistency and the variety of the classes' metrics.

The common ground on what constitutes a domain is something idealized. We recall that for Plank [Plank 2011] the common ground does not exist. Plank affirms that the literary genre can be considered a domain that mixes those textual types and socio-historical aspects. While Table 1 idealizes the pattern of these aspects for each genre, we can observe that poetry and tale present outstanding results ($\overline{x}$ =0.88 and 0.66; F-measure=0.88 and 0.77, respectively). We infer this highlight occurs because their textual type stands out to their socio-historical aspects. This result is not repeated with chronicles, novels, and short story. The socio-historical aspects of the latter are presented in greater proportion compared to the textual type, which makes the normalization process more difficult to reduce their differences (F-measures=0.24; 0.11; 037, respectively, as

[1]https://eli5.readthedocs.io/en/latest/

shown in Table 1). When we address Table 2 we will clear these assertions. The analysis of chronicles, novels, and short stories becomes challenging because of the uncertainties in the choice or extension of the characteristics of each of these genres, which turns into a contradiction if studies by two or more literary critics were compared. Literary genre is an ongoing problem that the constitutive classification weaknesses of the gender notion [Altman 1984]. The tool reflects this weakness when displaying the F-measure for chronicle, novel, and short story, respectively.

Table 1. Results per class reported by the mean (average value) along with a measure of variablility (SD, standard deviation).

Class	Precision		Recall		F-measure	
	Mean	SD	Mean	SD	Mean	SD
Chronicle	**1.00**	0.00	0.18	0.29	0.24	0.36
Novel	0.40	0.42	0.08	0.11	0.11	0.16
Poetry	**0.87**	0.19	**0.93**	0.15	**0.88**	0.11
Short story	0.73	0.43	0.30	0.33	0.37	0.37
Tale	0.66	0.08	**0.92**	0.06	**0.77**	0.07
Weighted average	0.57	0.12	0.66	0.07	0.58	0.09

Table 2 focuses on the feature importance per class, permitting to investigate the artistic content of each literary genre [Marcuschi et al. 2002]as they are a set of works of the same nature, with essentially identical trends [Almeida 2005], linked to similar cultural periods. Features linked to the notion of time stand out in poetry, novels, and chronicles, such as: `LW:time` 0.036; 0.025 for poetry and novel respectively; and `LW:past` 0.019, 0.016 and 0.015 for chronicle, poetry, and novel respectively. These features appear as the lowest negative features for short stories and tales: `LW:past` -0.006, -0.044 and `LW:time` -0.053 respectively, confirming that these two genres focus more on structure than narrative, which is evidenced in the prevalence of morphosyntactic annotation of universal dependence and POS aimed at analyzing the linguistic features of a word along with its preceding as well as following words (PO and UD tags)

5. Discussion

How the authors of the texts under analysis write provide clues to emotion and cognition [Gottschalk and Gleser 1979, Rosenberg and Tucker 1979]. The LIWC dictionary [Balage Filho et al. 2013] offers an efficient method to study the emotional components of literary works, going beyond the analyzes that commonly focus on the structure of those texts. We demonstrate that the analysis becomes more precise using a context-motivated tool (taking the classification of textual genres as contextual information), bringing results closer to state-of-the-art. Our strategies offer results with potential uses to set limits on the accuracy, being easy to replicate and interpret: i) the use of LIWC provides contextual data that make the tool more intuitive; the inclusion of all LIWC categories reduces complexity and facilitates tool optimization; ii) each node of the Random Forest reduces the built-in freedom of context, softening the dispersion of available information; iii) Part-of-Speech label words identifying their function (grammatical class tag such as noun, verb, article, adjective, preposition, pronoun, adverb, conjunction and

Table 2. Feature importance per class. LW, PO and UD stand for LIWC, part-of-speech and Universal Dependency labels, respectively.

Chronicle		Novel		Poetry		Short story		Tale	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
Ten highest positive features									
LW:past	0.019	LW:time	0.025	LW:time	0.036	LW:ingest	0.006	PO:UM	0.013
LW:assent	0.009	UD:NSUBJ	0.017	LW:past	0.016	PO:LADV1	0.005	UD:POBJ	0.012
LW:quant	0.008	LW:past	0.015	LW:preps	0.009	LW:percept	0.005	UD:CASE	0.010
PO:LTR	0.008	LW:home	0.015	PO:LTR	0.006	LW:see	0.004	UD:NUMBER	0.009
UD:CC	0.003	PO:LTR	0.015	UD:CASE	0.005	UD:CC	0.002	PO:PNT	0.009
LW:motion	0.002	LW:see	0.013	LW:sad	0.005	PO:LADV2	0.002	UD:ROOT	0.009
LW:ingest	0.002	LW:assent	0.012	LW:motion	0.004	PO:LPREP1	0.002	LW:cogmech	0.007
PO:GER	0.002	PO:LITJ2	0.012	UD:POBJ	0.004	LW:i	0.002	UD:MWE	0.007
LW:conj	0.002	UD:PREDET	0.010	LW:assent	0.003	LW:health	0.001	LW:bio	0.006
LW:see	0.002	LW:sad	0.010	LW:ingest	0.002	PO:DGT	0.001	UD:PUNCT	0.005
Ten lowest negative features									
UD:CASE	-0.006	PO:UM	-0.003	LW:cogmech	-0.012	UD:POBJ	-0.009	LW:time	-0.053
UD:POBJ	-0.005	UD:ROOT	-0.003	LW:funct	-0.009	UD:CASE	-0.008	LW:past	-0.044
LW:preps	-0.005	LW:future	-0.001	LW:cause	-0.008	PO:EOE	-0.008	PO:LTR	-0.029
LW:time	-0.005	UD:POBJ	-0.001	LW:tentat	-0.007	UD:IOBJ	-0.006	LW:assent	-0.024
PO:DA	-0.003	UD:CASE	-0.001	LW:inhib	-0.006	LW:past	-0.006	LW:see	-0.019
PO:PNT	-0.003	UD:NUMBER	-0.001	LW:adverb	-0.006	UD:NUMBER	-0.006	LW:ingest	-0.018
UD:NSUBJ	-0.003	LW:bio	-0.001	LW:quant	-0.006	PO:UM	-0.005	UD:CC	-0.015
PO:CARD	-0.003	PO:LADV4	0.000	PO:UM	-0.005	PO:PNT	-0.005	LW:home	-0.014
PO:ORD	-0.003	UD:PUNCT	0.000	PO:VAUX	-0.005	UD:MWE	-0.004	PO:LITJ2	-0.012
UD:MARK	-0.003	PO:LITJ1	0.000	LW:conj	-0.004	LW:bio	-0.004	UD:NSUBJ	-0.012

interjection) from the relationship with relative terms and by definition (probability-based and rule-based) help to reduce ambiguity.

For the approaches to literary genres to achieve results in state-of-the-art, it is necessary to adapt the tool to the fluidity and mutability characteristics of these genres, making the machine obey the rules of the nature of the analyzed text. These are the rules/strategies that we try to expose in this research work. Working bottom-up, the system establishes the best and worst feature combinations to classify specific literary genres.

According to Lüthi [Lüthi 1970] tales follow a distinct style for unfolding the genre in lasting appeal to people. Tale's unique style of structure, symbolism, and meaning offer "sharpness and precision" because it eliminates most descriptions (prevalent in chronicles, novels, poetry, and short stories), giving tales a universal meaning that opens up an opportunity for the use of the imagination. Therefore, we found consistency with the results in Table 2, as the absence of those descriptive details gives greater importance to the text structure than to the psychological characteristics, reducing the efficiency of the LIWC.

The everyday basis of the narrative structure present in the chronicles, short stories, and novels did not 'deceive' the tool, giving less weight to the LIWC attributes for the short story. This result in Table 2 is consistent because the linguistic sequences of the short story are more streamlined than the chronicle and the novel, increasing the importance of features related to the structure of the text relatively to the psycholinguistic attributes of the LIWC. The short story is based on the principle of offering a faster reading than a novel. Therefore, it condenses information, reduces the number of facts presented, and aesthetic strategies meet this need.

It is important to emphasize that the novel genre can contain several textual types, making the tool's evaluation perform very poorly. In some cases, it is possible to find

genres with a specific typology, such as poetry, which improves the performance of the analysis. As seen in Table 1.

Time is a feature that stands out in romance and poetry. This highlight in the novel is due to the intrinsic narrative situated in time. However, the tool also found greater prominence for the feature time in poetry. See Table 2. This finding is justified because poetry evokes an imaginative awareness, organizing its meaning, sound, and rhythm through language [Nemerov 2020]. There is a hypothetical expression of things in poetry that stands out from the storytelling of facts in the novel. In poetry, contemplation evokes feelings, leading the reader to a delight intrinsic to art (the Beauty) so as not to 'freeze' the senses in separate classes of objects. In this way, poetry acts on the human spirit, becoming recognizable because it depends on a line as a parameter, which guides the reading through the displacement of the latter concerning breathing and syntax [Nemerov 2020]. This characteristic changes its appearance, which makes interpretation by the tool more accurate. This reading process (line) is essential to differentiate the tone or rhythm of poetry from the novel.

The precision-of-meaning effect is greater in the novel than in poetry, making syntax-based tools more obvious to use as they deal with the measurable. Refer to UD tags in Table 2. In poetry, meaning is less accessible to observing the 'best order' to operate in the 'best words', which determines the artistic attitude (the Beauty) concerning definitions in general. This poetic structure has to do with pleasure, with delight in the form of arrangement of sounds about thoughts [Nemerov 2020]. Although poetry deals with commonplace matters, its structure does not have the characteristic of the commonplace. Poetry contains forms of production of inferences like the forms of parables, adapting to the metamorphosis of sentences, transcending the topic dealt with, that is, extending time [Monte-Serrat 2017]. It is inferred, therefore, why the feature time is so important in poetry. See Table 2. These are some comments on considering these elements as strategies to establish the relationship of literary genres with the interpretation to be performed by our tool.

6. Conclusion

We conclude that analyzes of literary genres must consider the artistic content of the texts, as aesthetics are a fundamental element for the various genres. The identification of aesthetics from the search for the 'best order or structure' of words or aesthetics from the search for the 'best word' is of paramount importance, since literary genres materialize the style or form adopted by their authors, linking the writing of the latter to the objectives and organizational processes through which the text went through. The combination of tools that we suggest in this research work provides a textual analysis that balances structural and contextual aspects so that the tool works more intuitively, approaching the state-of-the-art. The cases in which the results were not satisfactory (see Table 1) are due to the complexity of the elements that make up the literary genre. NLP deals with canonical varieties that are considered standard [Plank 2016] and the challenge in analyzing literary genres lies in data variations. Our method suggests how to improve data training. We indicate how best to leverage contextual (literary genres) data that is forgotten and needs to be refined to produce more robust models. It is not about making prescriptions for dealing with textual genres. We make assumptions about which tools are best suited for each genre, training the system to make fewer mistakes.

References

Almeida, N. M. d. (2005). *Gramática metódica da Língua Portuguesa*. Saraiva, 45 edition.

Altman, R. (1984). A semantic/syntactic approach to film genre. *Cinema Journal*, pages 6–18.

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bronckart, J.-P. (2004). Les genres de textes et leur contribution au développement psychologique. *Langages*, 1(153):98–108.

Crowston, K. and Kwasnik, B. H. (2003). Can document-genre metadata improve information access to large digital collections? *LIBRARY TRENDS*, 52(2):345–361.

Feldman, S., Marin, M. A., Ostendorf, M., and Gupta, M. R. (2009). Part-of-speech histograms for genre classification of text. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4781–4784. IEEE.

Gottschalk, L. A. and Gleser, G. C. (1979). *The measurement of psychological states through the content analysis of verbal behavior*. University of California Press.

Grilo, S., Bolrinha, M., Silva, J., Vaz, R., and Branco, A. (2020). The BDCamões Collection of Portuguese Literary Documents: a Research Resource for Digital Humanities and Language Technology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 849–854.

Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., and Wolkert, N. (1998). Iterative information retrieval using fast clustering and usage-specific genres. In *Eight DELOS workshop on User Interfaces in Digital Libraries*, pages 85–92.

Lüthi, M. (1970). Once Upon a Time: On the Nature of Fairy Tales. *Trans. Lee Chadeayne & Paul Gottwald. New York: Frederick Ungar Publishing Co*.

Marcuschi, L. A. et al. (2002). Gêneros textuais: definição e funcionalidade. *Gêneros textuais e ensino*, 2:19–36.

Martins, N. S. (2008). *Introdução à estilística: a expressividade na língua portuguesa*, volume 71. Edusp.

Matos, T. (2021). Gêneros textuais. `https://www.portugues.com.br/redacao/generostextuais.html`. Online; accessed July 17th of 2021.

Monte-Serrat, D. (2017). Neurolinguistics, Language and Time: investigating the verbal art in its amplitude. *International Journal of Perceptions in Public Health*, 1(3):162–171.

Monte-Serrat, D. (2021). Operating language value structures in the intelligent systems. *Advanced Mathematical Models & Applications*, 6(1):31–44.

Monte-Serrat, D. M. and Cattani, C. (2021a). Interpretability in neural networks towards universal consistency. *International Journal of Cognitive Computing in Engineering*, 2:30–39.

Monte-Serrat, D. M. and Cattani, C. (2021b). *The Natural Language for Artificial Intelligence*. Elsevier.

Nemerov, H. (2020). Poetry. Encyclopedia Britannica. https://www.britannica.com/art/poetry. [Online; accessed 05-August-2020].

Nilan, M. S., Pomerantz, J., and Paling, S. (2001). Genres from the Bottom Up: What Has the Web Brought Us? In *Proceedings of the ASIST Annual Meeting*, volume 38, pages 330–39. ERIC.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16, Cham. Springer International Publishing.

Omar, A. (2020). Classifying literary genres: a methodological synergy of computational modelling and lexical semantics. *Texto Livre: Linguagem e Tecnologia*, 13(2):83–101.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Plank, B. (2011). *Domain Adaptation for Parsing*. PhD thesis, University of Groningen, https://bplank.github.io/publications.html. ISBN: 978-90-367-5199-5.

Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing*.

Rosenberg, S. D. and Tucker, G. J. (1979). Verbal behavior and schizophrenia: The semantic dimension. *Archives of General Psychiatry*, 36(12):1331–1337.

Rosso, M. A. (2005). What type of page is this? Genre as Web descriptor. In *Proceedings of the 5th ACM/IEEE–CS joint Conference on Digital libraries*, pages 398–398.

Schneuwly, B. (1997). Textual organizers and text types: Ontogenetic aspects in writing. *Processing interclausal relationships. Studies in the production and comprehension of text*, pages 245–263.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Sven Meyer zu, E. and Stein, B. (2004). Genre classification of web pages. In Biundo, S., Frühwirth, T., and Palm, G., editors, *KI 2004: Advances in Artificial Intelligence*, pages 256–269, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Acknowledgement

This work was carried out at the Center for Artificial Intelligence (C4AI–USP), with support by the São Paulo Research Foundation (FAPESP Grant #2019/07665-4) and by the IBM Corporation.