







MEGALITE^{PT}: A Corpus of Literature in Portuguese for NLP

Igor Morgado¹(✉) , Luis-Gil Moreno-Jiménez² , Juan-Manuel Torres-Moreno² ,
and Roseli Wedemann¹ 

¹ PPG-Ciências Computacionais, Universidade do Estado do Rio de Janeiro,
Rua São Francisco Xavier, 524, 20550-900 Rio de Janeiro, Brazil
{igor.morgado, roseli}@ime.uerj.br

² Laboratoire Informatique d'Avignon, Université d'Avignon, (Avignon) France,
339 Chemin des Meinajariès, 84911 Avignon, cédex 9, France
{luis-gil.moreno-jimenez, juan-manuel.torres}@univ-avignon.fr
<https://ccomp.ime.uerj.br>, <https://lia.univ-avignon.fr>

Abstract. We present the section of the MEGALITE corpus based on literary texts in Portuguese. This new section has been developed and adapted to be used for Computational Creativity tasks, such as Natural Language Processing, Automatic Text Generation (ATG), and other similar purposes. We highlight characteristics of the Portuguese section, such as the numbers of documents, authors, sentences and tokens and also how it is structured and formatted. We show how the ATG algorithms, which we have previously developed, behave when trained on this corpus, by using a human evaluation protocol where a mixture of automatically generated and *natural* texts is classified, using four criteria: grammaticality, coherence, identification of context, and an adapted Turing test.

Keywords: Portuguese literary corpus · Corpus for emotion detection · Learning algorithms · Linguistic resources

1 Introduction

Linguistic corpora have been widely used in Natural Language Processing (NLP) tasks in recent years. Experiment has shown that a well constructed and analyzed corpus can be exploited to improve the quality of the linguistic objects produced by NLP algorithms in general, and also in the specific case of Automatic Text Generation (ATG) procedures. However, the construction of consistent literary corpora is often unattainable [17] due to the complexity of the process, which requires much time for analysis.

Moreno-Jiménez and collaborators [10, 14] have recently presented a corpus for use in NLP formed only by literary texts in Spanish. This corpus was applied in tasks such as sentiment analysis [9] and automatic generation of literary sentences [12, 14]. The corpus reached the mark of approximately 200 million (M) tokens from literature in Spanish and for this reason was named MEGALITE. The last available version of the Spanish section contains approximately 5 000 documents, 1 300 different authors, approximately 15 M sentences, 200 M tokens, and 1 000 M characters. In [11], the corpus was extended to encompass a section composed of literature in French. This

addition contains 2 690 documents, 1 336 different authors, approximately 10 M sentences, close to 182 M tokens, and approximately 1 082 M characters. To contemplate the addition of a new language, the sections of the corpus acquired two new names, MEGALITE^{ES} for the Spanish section and MEGALITE^{FR} for French the part.

In this work, we present an extension of MEGALITE [20], formed by adding to it a new section based on literature produced in Portuguese, from different lusophone countries, such as Brazil, Portugal, and Mozambique, to name a few. It also contains literature translated to Portuguese taken from sources from different countries around the globe. We describe how the corpus was produced and formatted, its main properties, some ATG experiments carried out on the corpus and their results. We use two different representations of the corpus to better understand its structure and possible applications.

In Sect. 2, we present some work related to the development and analysis of corpora. In Sect. 3, we describe the new corpus MEGALITE^{PT}. Section 4 briefly describes the algorithm for ATG and, in Sect. 5, we present some experiments and evaluate the performance of the ATG algorithms trained with MEGALITE^{PT}. Finally, in Sect. 6, we propose some ideas for future work before concluding.

2 Related Work

In this section, we discuss some work related to the topic of the construction of literary corpora. We note that most of these corpora are composed of documents written in English. For this reason, we have concentrated our efforts on collecting literary documents written in Portuguese, in order to extend the MEGALITE [11] corpus, which already contains a section of literary documents in Spanish and another in French. We hypothesize that the richness and variability of styles found in literature can improve the quality of texts obtained with ATG algorithms, overcoming the limitations of the overly rigid styles of technical documents, or the stereotypes of the journalistic style.

In [17], the authors introduced the RiQua¹ corpus composed of literary quotation structures in 19th century English. The RiQua corpus provides a rich view of dialogue structures, focusing on the importance of the relation between the content of a quotation and the context in which it is inserted in a text. Another interesting approach presented in [19] describes the SLäNDa corpus that consists of 44 chapters of Swedish narratives, with over 220 K manually annotated tokens. The annotation process identified 4733 occurrences of quoted material (quotes and signs) that are separate from the main narrative, and 1143 named speaker-to-speech correspondences. This corpus has been useful for the development of computer tools for analyzing literary narratives and discourse.

A Spanish corpus called LiSSS has been proposed in [9]. It is constituted by literary sentences collected manually from many literary works. The LiSSS corpus has been annotated according to five emotions: love, anger, happiness, hope and fear. It is available in two versions: the first one has 500 sentences manually multi-annotated (by 13 persons), and the second one has 2 000 manually mono-annotated sentences.

¹ This corpus is available on the official website of the University of Stuttgart <https://www.ims.uni-stuttgart.de/en/research/resources/corpora/riqua/>.

Concerning corpora with emotional content, we have the SAB corpus introduced in [15]. This corpus is composed of tweets in Spanish, representing reviews about seven types of commercial products. The tweets are classified into eight categories: *Confidence*, *Satisfaction*, *Happiness*, *Love*, *Fear*, *Disaffection*, *Sadness* and *Anger*. In [2], another very complete work with three resources is described. The first of these is an emotional lexicon composed of words from 136 of the most spoken languages in the world. The second resource is a knowledge graph that includes 7 M words from the same 136 languages, with about 131 M inter-language semantic links. Finally, the authors detected the emotional coherence expressed in Wikipedia texts about historical figures in 30 languages.

3 Megalite Corpus

This section describes MEGALITE^{PT}, a literary corpus for the Portuguese language. It consists of thousands of literary documents, spanning more than a thousand different authors from different countries, writing styles, and literary genres. The documents in this corpus come from a personal collection and hence, for copyright reasons, we are not allowed to share them in their original form. Nevertheless, following the same formatting standards used in the Sections MEGALITE^{ES} and MEGALITE^{FR}, the corpus is available as files indexed by author surname and title, in the form of embeddings, represented in a Parts Of Speech (POS) tags version and a lemma version, and also in files displaying the lists and frequencies of unigrams, bigrams, and SU4-bigrams.

3.1 Structure of the Corpus

The original corpus was built from literary documents in the Portuguese language, written by lusophone authors and also by text translated from other languages to Portuguese. The corpus contains 4311 documents, from 1418 authors, in different literary genres, such as plays, poems, novels, essays, chronicles, etc. The original documents, obtained in heterogeneous formats (ordinary text, epub, pdf, HTML, ODT, doc, etc.), were processed and stored as plain text, UTF-8 document files. Textual metadata such as indexes, titles, remarks, author notes and page numbering were filtered out using techniques that detect regular expressions and pattern matching, and by manual removal. Afterwards, we performed a textual segmentation using a tool developed in PERL 5.0 to detect regular expressions [5]. Some of the properties of the corpus, after pre-processing, are detailed in Table 1.

Table 1. Properties of MEGALITE^{PT}, with 4311 literary texts ($K = 10^3$ and $M = 10^6$).

	Sentences	Tokens	Characters
Total in corpus	19.9 M	253.3 M	1 488.1 M
Average per document	4.6 K	58.7 K	345.2 K

In its current state, the MEGALITE corpus is very extensive, containing literary documents in French and Spanish, so that it is suitable for use in automatic learning and translation. It has, however, a small amount of noise formed by a few textual objects not detected in the pre-processing stages, leading to some mistakes in the segmentation process. This is not unusual in a corpus of the size of MEGALITE, and these same kind of objects may also be found in most corpora containing unstructured text, and they also occur in the Portuguese corpus MEGALITE^{PT}.

The names of all files in MEGALITE^{PT} follow the same naming patterns used in the other sections of MEGALITE, that is `authorLastName`, `_authorName-title`. We also group all authors with the same last name initials in directories. In Table 2, we display the properties of the corpus, for each one of the directories identified by the initial of the last names of the authors.

Table 2. Properties of MEGALITE^{PT}. Numbers of documents, authors, sentences, tokens, and characters in each directory, which is identified by the initials of the last name of the authors.

Directory ID	Docs	Authors	Sentences	Tokens	Characters
Anonymous	6	1	1525	31498	179396
A	757	94	1549969	21288326	124192498
B	355	124	1501439	19679748	116339567
C	459	135	2099142	25127600	147580607
D	271	53	1018850	13018034	76255904
E	36	27	180806	2311665	13598473
F	115	53	679664	8554513	50880932
G	197	86	1020259	13010596	76451978
H	151	62	1184486	14943679	87933493
I	19	8	163198	2151577	12734549
J	69	35	414045	4965709	29099689
K	83	32	908103	10261876	59643611
L	142	79	812392	10526168	61949313
M	314	150	2076862	25689661	150826348
N	62	33	241849	3288980	19089347
O	31	14	108397	1687042	10060641
P	188	95	804049	11936237	70067586
Q	58	9	373289	4645910	27354438
R	278	80	1708858	20295900	118948880
S	431	126	1633717	19842890	116187506
T	70	34	513741	7219323	42511885
U	3	2	25263	357202	2111702
V	120	36	309611	4373096	25933238
W	68	35	501221	6146484	36197624
Y	2	2	12637	172111	1022010
Z	26	14	148242	1861954	11019596
Total	4,311	1,419	19,991,614	253,387,779	1,488,170,811

3.2 Word2vec Embeddings

Word embeddings are representations of words that quantify semantic similarities between linguistic terms. These embeddings can be determined from the analysis of relations among words in a large corpus. Embeddings for the MEGALITE^{PT} corpus were generated using the Word2vec model [7] with the Gensim [18] library, which resulted in a set of 389,340 embeddings. Each embedding is an s -dimensional vector whose elements were obtained from semantic relationships among words in the MEGALITE^{PT} corpus. The training process performed to generate our embeddings used the parameters shown in Table 3. Iterations, i , represents the number of training epochs. Minimal count, m , indicates the minimal frequency of occurrence of a word in the corpus needed for it to be added to the vocabulary. For any word x , its embedding has vector size, s (s specifies the dimension of the vector representation of x), and window size, ws , represents the number of words adjacent to x in a sentence (that are related to it within the sentence) that will be considered to form the embedding. In this model, we used the skip-gram approach [6], with a negative sampling of five words and a downsampling threshold of 0.001.

Table 3. Word2Vec configuration parameters.

Parameter	Values
Iterations, i	5
Minimal count, m	3
Vector size, s	60
Window size, ws	5

Table 4 displays the 10 nearest tokens found in MEGALITE^{PT} for the word queries **Azul** (*blue*), **Mulher** (*woman*) and **Amor** (*love*). The distance between the query and a token is determined by the cosine similarity given by Eq. (2) (see the model description in Sect. 4). For each query word, Q , in Table 4, the left column shows a word, x , associated to Q chosen from the corpus by Word2vec, and the right column shows the cosine similarity between Q and x . We chose to not translate the words associated to the queries within the table, since many of these are synonymous to each other or do not have an English translation to a single word. This is an interesting feature of MEGALITE, that it captures some literary/artistic meanings of words which normally do not emerge from non-literary corpora.

3.3 POS Tag and Lemma Representations

In this section, we present two representations of MEGALITE^{PT}. The first one is a corpus built by using only POS tags and the second one uses only lemmas. This is a solution found that enables sharing the corpus without breaking copyright laws, although still preserving semantic meaning. Table 5 contains a very small subset of these representations, it shows a few sentences from Machado de Assis’s “Memórias Póstumas de Brás Cubas”. The first column displays the line number, which corresponds to the order of

Table 4. List of 10 nearest tokens found in MEGALITE^{PT} for queries Azul, Mulher and Amor.

Keyword	Azul (<i>blue</i>)	Cosine Similarity	Mulher (<i>woman</i>)	Cosine Similarity	Amor (<i>love</i>)	Cosine Similarity
	verde	0.922	moça	0.951	ternura	0.848
	violeta	0.902	menina	0.934	eterno	0.819
	lilás	0.898	mocinha	0.904	amante	0.818
	turquesa	0.895	meninazinha	0.903	ideal	0.788
	cinza	0.895	garota	0.894	deidade	0.787
	alaranjado	0.884	garotinha	0.883	ente	0.781
	cobalto	0.882	mulherzinha	0.883	crença	0.774
	azulado	0.882	menininha	0.880	senhôr	0.771
	centáurea	0.876	velhota	0.879	encanto	0.770
	amarelo	0.876	rapariga	0.873	tema	0.765

the sentence in the original text document (its line number in the file). The second column shows the original sentence as it appears in the original text. The third column displays the version of the original sentence in the POS tag representation, and the fourth column shows the sentence in its lemma representation. These two representations of MEGALITE^{PT} are formed as we describe in what follows.

POS Tag Corpus. This representation is constructed by making a morpho-syntactic analysis of each document, and replacing each word of the document with its corresponding POS tag. The analysis was performed using Freeling version 4.0 [16]. The POS tag² shows grammatical information for each word within a given sentence.

Lemma Corpus. The second representation is a lemmatized version of the original documents. This was achieved by using Freeling POS tags as references to first extract only meaningful lexical words, in this case only verbs, nouns, and adjectives. Every extracted word was then substituted by the corresponding lemma, which is a basic form of a given word, without conjugation, in its singular form and neutral or male genre. Words corresponding to all other types of POS tags, i.e. not verbs, nouns, and adjectives, were removed from this corpus.

3.4 n-Gram Statistics

MEGALITE also provides the frequencies of occurrences of unigrams, bigrams, and skip-grams of the type SU4-bigrams [1]. SU4-bigrams are obtained by taking a pair of words from a sentence such that from the first word in the pair one takes n steps to find the second word, i.e., using n -sized skip-grams, for $n = 1, 2, 3, 4$. For example, for the sentence “Não tive filhos, não transmiti a nenhuma criatura o legado da nossa miséria.”,

² A detailed description of Freeling POS tags can be found at <https://freeling-user-manual.readthedocs.io/en/latest/tagsets/tagset-pt/>.

Table 5. Samples of sentences recovered from Machado de Assis’s novel “Memórias Póstumas de Brás Cubas”, in different versions of MEGALITE^{PT}.

Line	Original	MEGALITE POS	MEGALITE lemmas
2967	Não alcancei a celebridade do emplasto, não fui ministro, não fui califa, não conheci o casamento.	RN VMIS1S0 DA0FS0 NCCS000 SP DA0MS0 NCMS000 Fc RN VMIS1S0 NCMS000 Fc RN VMIS1S0 NCMS000 Fc RN VMIS1S0 DA0MS0 NCMS000 Fp	ALCANÇAR CELEBRIDADE EMPLASTO IR MINISTRO IR CALIFA CONHECER CASAMENTO
2968	Verdade é que, ao lado dessas faltas, coube - me a boa fortuna de não comprar o pão com o suor do meu rosto.	NP00000 RG Fc SP DA0MS0 NCMS000 SP DD0FP0 NCFP000 Fc VMIS3S0 Fg PP1CS00 DA0FS0 AQ0FS00 NCF5000 SP RN VMN0000 DA0MS0 NCMS000 SP DA0MS0 NCMS000 SP DA0MS0 DP1MSS NCMS000 Fp	VERDADE LADO FALTA CABER BOM FORTUNA COMPRAR PÃO SUOR ROSTO
2969	Mais; não padeci a morte de Dona Plácida, nem a semidemência do Quincas Borba.	RG Fx RN VMIS1S0 DA0FS0 NCF5000 SP NP00000 Fc CC DA0FS0 NCF5000 SP DA0MS0 NP00000 Fp	PADECER MORTE DONA.PLÁCIDA SEMIDEMÊNIA QUINCAS.BORBA
2970	Somadas umas coisas e outras, qualquer pessoa imaginará que não houve minguá nem sobra, e, conseqüentemente que saí quite com a vida.	VMP00PF DI0FP0 NCFP000 CC DI0FP0 Fc DI0CS0 NCF5000 VMIF3S0 CS RN VMIS3S0 NCF5000 CC NCF5000 Fc CC Fc RG CS NCMS000 AQ0CS00 SP DA0FS0 NCF5000 Fp	SOMAR COISA PESSOA IMAGINAR HAVER MÍNGUA SOBRA SAÍ QUITE VIDA

given the word *filhos*, the SU4 bigrams are: *filhos/não*, *filhos/transmiti*, *filhos/a* and, *filhos/nenhuma*. The same procedure is applied to every token in every sentence in the text. Then all the occurrences of the same pair are summed up to compute the total frequency of occurrence of each pair of tokens and they are sorted in decreasing order of frequency. In Table 6, we display the top 5 most frequent bigrams and SU-4 bigrams for 4 texts of different authors.

4 Model for Generating Artificial Literary Sentences

In this section, we present a brief description of an adaptation of our previously developed model for literary sentence generation [8, 12, 13]. We have used this model to generate sentences in Spanish and French, using MEGALITE^{ES} and MEGALITE^{FR} and we will show results of its use in experiments of ATG with MEGALITE^{PT}, in the next section. The model consists of the two following stages.

First Stage - Canned Text. This step consists of using the canned text method, commonly used for ATG [3]. The process begins by selecting a sentence f from the original version of MEGALITE^{PT}, which will be used to generate a new phrase. Sentence f is then parsed with FreeLing [16] to replace the lexical words³ by their morpho-syntactic labels (POS tags) and thus generate a Partially Empty Grammatical Structure (PGS). Functional words such as prepositions, pronouns, auxiliary verbs, or conjunctions are

³ Verbs, adjectives, and nouns.

Table 6. Bigrams and SU4-Bigrams with the 5 highest frequencies from 4 literary works in MEGALITE^{PT}.

Bigrams	Frequency	SU4-Bigrams	Frequency
Fernando Pessoa, Livro do Desassossego			
rua douradores	28	vida vida	57
patrão vasques	25	sonho sonho	32
guarda livros	18	mim mim	30
vida real	18	sonho vida	30
vida vida	17	mim vida	30
Eça de Queirós, Os Maias			
maria eduarda	119	maria eduarda	120
affonso maia	94	affonso maia	94
castro gomes	85	castro gomes	85
santa olavia	78	disse carlos	81
á porta	72	santa olavia	78
Érico Veríssimo, O Continente			
santa fé	127	santa fé	127
ana terra	91	ana terra	92
rio pardo	81	pedro terra	83
pedro terra	77	rio pardo	82
maria valéria	58	maria valéria	58
Clarice Lispector, A Descoberta do Mundo			
san tiago	19	ovo ovo	60
dona casa	15	amor amor	53
homem mulher	14	vida vida	43
caneta ouro	14	ovo galinha	28
vou contar	13	homem homem	27

kept in the sentence. To maintain semantic accuracy in our algorithm, the generated sentences must have at least 3 lexical words, but no more than 10. Once the PGS has been generated, it will be analyzed by the semantic module in the second stage.

Second Stage - Semantic Module (Word2vec) Training. We next replace the POS tags of the PGS by lexical words using the Word2vec model. This model has been implemented for our experiments under the *Skip-gram* architecture [6] using MEGALITE^{PT} for training. We have used the hyper-parameter values specified in Table 3 during the Word2vec training phase, to obtain 389,340 embeddings.

In order to select the vocabulary that will replace the POS tags in the PGS formed from f to construct the new sentence, we have implemented a procedure based on an

arithmetic analogy proposed by [4]. We consider the three embeddings corresponding to the words Q , O and A defined as

- \vec{Q} : the embedding associated with the context word Q , the query, given by the user,
- \vec{O} : the embedding associated with the original word O in f which has been replaced by the POS tag,
- \vec{A} : the embedding associated with the word adjacent to O on the left in the sentence f .

With these embeddings, we calculated a fourth embedding \vec{y} with the expression

$$\vec{y} = \vec{A} - \vec{O} + \vec{Q}. \quad (1)$$

This embedding \vec{y} has the features of \vec{A} and \vec{Q} enhanced and the features of \vec{O} decreased, so that it is more distant to \vec{O} .

We then obtain the embeddings of the best word associations related to \vec{y} with Word2vec, and store the first $M = 4\,000$ of these in a list \mathcal{L} , i.e. we take the 4000 first outputs of Word2vec, when \vec{y} is given as input. \mathcal{L} is thus an ordered list of 4000 vectors, a matrix, where each row, j , corresponds to an embedding of a word, w_j associated to \vec{y} . The value of M has been established as a compromise between the execution time and the quality of the embeddings for the procedure we are describing. The next step consisted of ranking the M embeddings in \mathcal{L} , by calculating the cosine similarities between the j^{th} embedding in \mathcal{L} , \vec{L}_j , and \vec{y} as

$$\theta_j = \cos(\vec{L}_j, \vec{y}) = \frac{\vec{L}_j \cdot \vec{y}}{\|\vec{L}_j\| \cdot \|\vec{y}\|} \quad 1 \leq j \leq M. \quad (2)$$

\mathcal{L} is ranked in decreasing order of θ_j .

Another important characteristic to consider when choosing the substitute word is *grammatical coherence*. We have therefore implemented a **bigram analysis**, by estimating the conditional probability of the presence of the n^{th} word, w_n , in a sentence, given that a previous, adjacent word, w_{n-1} , on the left is present,

$$P(w_n | w_{n-1}) = \frac{P(w_n \wedge w_{n-1})}{P(w_{n-1})}. \quad (3)$$

The conditional probability of Eq.(3) corresponds to the frequencies of occurrence of each bigram in MEGALITE^{PT}, which was obtained from the n -gram detection procedure used when constructing this corpus, as described in Subsect. 3.4. Among the bigrams in MEGALITE^{PT}, we considered only the bigrams formed by lexical and functional words (punctuation, numbers, and symbols are ignored) to form a list, LB , used to calculate the frequencies.

For each \vec{L}_j in \mathcal{L} , we compute two bigrams, $b1_j$ and $b2_j$, where $b1_j$ is formed by the left word adjacent to O in f (corresponding to embedding \vec{A}) concatenated with the word w_j (corresponding to embedding \vec{L}_j). Then, $b2_j$ is formed by w_j concatenated with the word adjacent to O to the right in f . We then calculate the arithmetic mean, bm_j , of the frequencies of occurrence of $b1_j$ and $b2_j$ in LB . If O is the last word in f , bm_j is simply the frequency of $b1_j$. The value bm_j for each \vec{L}_j is then combined with

the cosine similarity θ_j , obtained with Eq. (2), and the list \mathcal{L} is re-ranked in decreasing order of the new value

$$\theta_j := \frac{\theta_j + bm_j}{2}, \quad 1 \leq j \leq M. \tag{4}$$

Next, we take the word corresponding to the first embedding in \mathcal{L} as the candidate chosen to replace O . The idea is to select the word semantically closest to \vec{y} , based on the analysis performed by Word2vec, while keeping the coherence of the text obtained with the linguistic analysis done by the language model and the structure of MEGALITE^{PT}. The definition of \vec{y} given by Eq. (1) should allow a substitution of O by a word more distant in meaning, so that potentially more creative phrases may arise. Finally, to respect the syntactic information given by the POS tag, we use Freeling to convert the selected word to the correct gender and number inflection of the word O , which is specified by its respective POS tag. This process is repeated for each replaceable word in f (each POS tag). The result is a new sentence that does not exist in the corpus MEGALITE^{PT}. The model is illustrated in Fig. 1, where the sentence f converted to PGS can be appreciated on the top of the illustration. The PGS sends inputs to the Word2vec module that receives Q , A , and O to generate the list \mathcal{L} . This list is then filtered with the language model, to obtain the best choice with the correct grammatical structure returned by Freeling.

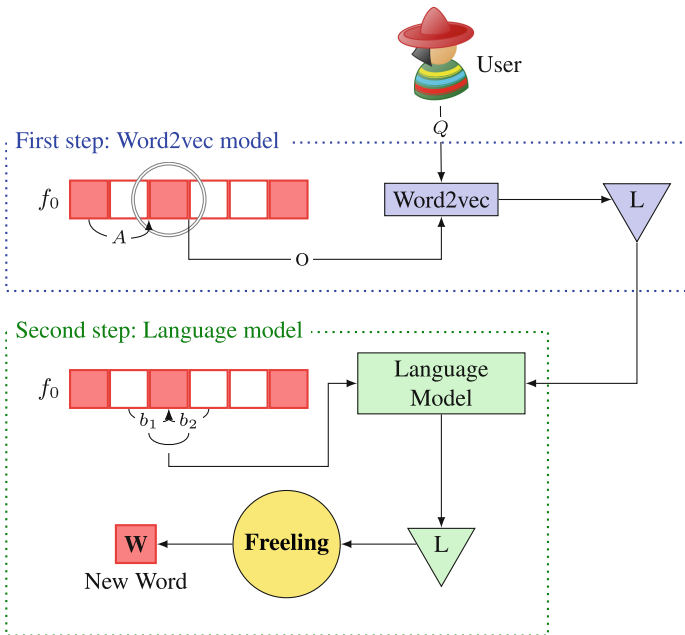


Fig. 1. Second step: vocabulary selection

5 Experiments of Automatic Sentence Generation in Portuguese

In this section, we describe a group of experiments implemented to evaluate the influence of corpus MEGALITE^{PT} in the task of automatic sentence generation. We describe the evaluation protocol, show some examples of generated sentences and present our results. We have chosen 45 sentences, with different grammatical structures and lengths varying from 3 to 10 lexical words, to be used as input to the canned text method. In Table 7, we display some of the queries and the corresponding generated sentences obtained with the model explained in Sect. 4.

Table 7. Generated sentences based on user input queries

Query	Generated sentence
lua	A primeira não deixava nada, olhava sentada, no argumento, indignada contra aquela tempestade de uma confusão nos problemas pelo seu trabalho
tristeza	A mulher me sentiu, me segurou, me levou, é bem verdade
amor	Sim, egoísmo, não tenho outra lei
guerra	Em uma ínfima fração de minuto, João também partiu
sol	Nevava nas casas, e nas cores, e nos palácios, e em galpões

5.1 Evaluation Protocol and Results

Using the method described in Sect. 4, we have automatically generated a set of fifteen sentences for each of the queries *amor*, *guerra*, and *sol*, with a total 45 sentences. We grouped according to query and submitted these sentences for human evaluation to 18 persons, each of whom completed the evaluation survey. Each sentence was evaluated for the three following qualitative categories.

Grammaticality. This category is used to measure the grammatical quality of the generated text. The main characteristics that should be evaluated are orthography, verb conjugations, gender, number agreement and punctuation. Other grammatical rules can also be evaluated but to a lesser degree of importance.

Coherence. In this case, we require the evaluation of how harmonic and well placed the words are within the sentence. The principal points of analysis are the correct use of words and word sequences, the sentence should have a clear meaning and should be read without difficulty.

Context. represents how the sentence is related to the topic of the query. Naturally, in a literary sentence, the relation with the topic can be subtle or even antagonistic.

Each one of these criteria should be evaluated by attributing a numerical, discrete value of 0, 1 or 2, where 0 represents that the sentence does not match that category at all. A value of 1 means that the sentence satisfies some of the conditions in that category, but not all. And finally, a value of 2 is given, if the sentence seems correct in relation to that category.

In the instructions for the evaluators, we stated that some sentences were generated using a computational algorithm, and others were extracted from multiple literary

works. We didn't inform the evaluator of the correct ratio between these two categories. We also performed an adapted Turing test where, for each sentence, we asked the evaluator to predict if the sentence is artificial, that is generated by a computer or if it is natural, that is written by some human author.

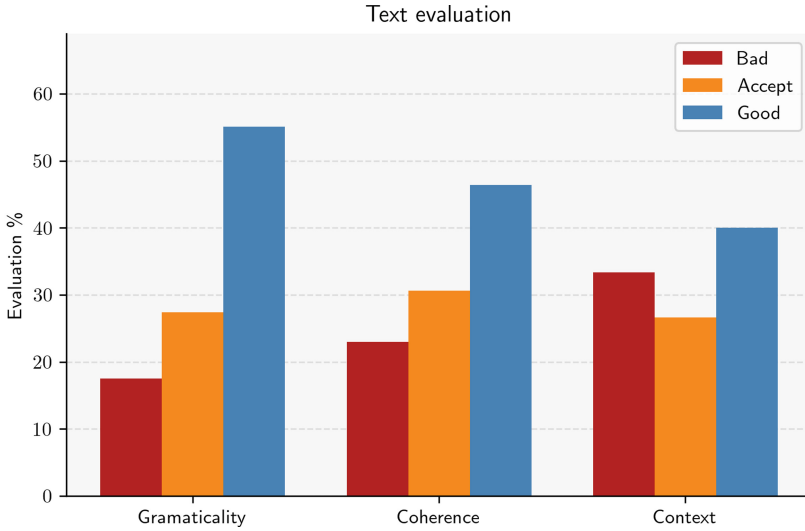


Fig. 2. Evaluation of coherence, grammaticality and context of 45 automatically generated sentences.

The results of our evaluation procedure can be seen in Fig. 2, where we can notice that 55% of the sentences are evaluated as grammatically correct, while 28% as acceptable, and only 17% are considered bad. The coherence values also display positive results with 47%, 30% and 23% evaluated as good, acceptable and bad, respectively. Finally, in the context category we have 40%, 27%, and 33% evaluated as good, acceptable and bad, respectively. All these values were rounded to integer values and the sum is 100% in each category, as expected.

Figure 3 shows the ratio of the evaluation for the Turing test for each one of the 45 sentences. Each bar sums up to 100, and represents in blue the percentage of evaluators that consider the text as written by a human (a natural sentence), while the other part, in yellow, represents the percentage of evaluators who consider the sentence as generated by a computer (an artificial sentence). The fact is that all sentences were generated by the model. The dashed line indicates the mean ratio between sentences evaluated as natural and artificial. This line shows that, on average, 56% of the evaluators consider that sentences were written by humans. Table 8 shows some sentences evaluated by human evaluators and how they were categorized by the majority (more than 80%) of the evaluators.



Fig. 3. Evaluations for the adapted Turing test.

Table 8. Example of sentences and results for the adapted turing test.

Evaluation	Query	Generated sentence
Human	amor	O cérebro é inocente; ninguém sabe escravizá-lo, nem o próprio dono
Human	guerra	Embora não seja o que você e eu chamássemos de dança
Machine	guerra	João está atualmente deixando a uma luta incrivelmente rápida pelo fio do caderno
Machine	sol	Não havia lugar onde fosse vermelho, e não havia como mudar dele
Machine	amor	Torturava pelo apelo dele como um beduíno morrendo de sede que entra uma fonte

6 Conclusions and Perspectives

We have introduced MEGALITE^{PT}, an extension of the MEGALITE literary corpus consisting of literary documents in Portuguese. We have provided versions of MEGALITE^{PT} in the POS tag format and in a lemmatized form. We also made available the lists and distributions of unigrams, bigrams, and SU4-bigrams for statistical analysis. The embeddings of 60-dimensional vectors, were obtained using the Word2vec model. In our experiments, we have shown that MEGALITE^{PT} is useful for NLP tasks such as automatic sentence generation. Our embeddings display a high degree of literary information and are very well suited for creative tasks.

In a human evaluation, 56% of the sentences produced using our model were considered to be generated by real human authors. These sentences were evaluated with good degrees of grammaticality, only 17% being considered bad in this category. Also very good coherence and context were perceived, with only 23% and 33% of the sentences being considered bad in each respective category. Hence, we strongly recommend MEGALITE^{PT} for NLP tasks such as Deep Learning Algorithms, textual assessment, text generation and text classification.

6.1 Future Work

We can extend this corpus to build a subset of MEGALITE^{PT} using only native writers. We believe that this corpus will be able to better model the nuances, details, and characteristics of Portuguese literature. We intend to build deep statistical analysis based on our corpus to find possible patterns and metrics that could help us to investigate structural properties of literature, artistic texts, and of ATG.

References

1. Cabrera-Diego, L.A., Torres-Moreno, J.M.: SummTriver: a new trivergent model to evaluate summaries automatically without human references. *Data Knowl. Eng.* **113**, 184–197 (2018). <https://doi.org/10.1016/j.datak.2017.09.001>
2. Chen, Y., Skiena, S.: Building sentiment lexicons for all major languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), vol. 2, pp. 383–389. Association for Computational Linguistics, Baltimore (2014). <https://doi.org/10.3115/v1/P14-2063>
3. Deemter, K.V., Theune, M., Krahmer, E.: Real versus template-based natural language generation: a false opposition? *Comput. Linguist.* **31**(1), 15–24 (2005). <https://doi.org/10.1162/0891201053630291>
4. Drozd, A., Gladkova, A., Matsuoka, S.: Word embeddings, analogies, and machine learning: beyond King - Man + Woman = Queen. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 3519–3530. The COLING 2016 Organizing Committee, Osaka (2016). aclanthology.org/C16-1332
5. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) [cs], September 2013. [arXiv: 1301.3781](https://arxiv.org/abs/1301.3781)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 26, pp. 3111–3119. Curran Associates Inc., Lake Tahoe, October 2013. proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html. [arXiv: 1310.4546](https://arxiv.org/abs/1310.4546)
8. Moreno-Jiménez, L.-G., Torres-Moreno, J.-M., Wedemann, R.S.: Literary natural language generation with psychological traits. In: Métais, E., Meziane, F., Horacek, H., Cimiano, P. (eds.) *NLDB 2020. LNCS*, vol. 12089, pp. 193–204. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51310-8_18
9. Moreno Jiménez, L.G., Torres Moreno, J.M.: LiSSS: a new corpus of literary Spanish sentences for emotions detection. *Computación y Sistemas* **24**(3), 1139–1147 (Sep 2020). <https://doi.org/10.13053/cys-24-3-3474>
10. Moreno-Jiménez, L.G., Torres-Moreno, J.M.: Megalite: a new Spanish literature corpus for NLP tasks. In: *Computer Science & Information Technology (CS & IT)*, pp. 131–147. AIRCC Publishing Corporation, January 2021. <https://doi.org/10.5121/csit.2021.110109>
11. Moreno-Jiménez, L.-G., Torres-Moreno, J.-M.: **MegaLite-2**: an extended bilingual comparative literary corpus. In: Arai, K. (ed.) *Intelligent Computing. LNNS*, vol. 283, pp. 1014–1029. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-80119-9_67
12. Moreno-Jiménez, L.G., Torres-Moreno, J.M.S., Wedemann, R., SanJuan, E.: Generación automática de frases literarias. *Linguamática* **12**(1), 15–30 (2020). <https://doi.org/10.21814/lm.12.1.308>

13. Moreno-Jiménez, L.G., Torres-Moreno, J.M., Wedemann, R.: A preliminary study for literary rhyme generation based on neuronal representation, semantics and shallow parsing. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pp. 190–198. SBC, Porto Alegre (2021). <https://doi.org/10.5753/stil.2021.17798.sol.sbc.org.br/index.php/stil/article/view/17798>
14. Moreno-Jiménez, L.G., Torres-Moreno, J.M., Wedemann, R.S.: Generación de frases literarias: un experimento preliminar. *Procesamiento del Lenguaje Natural* **65**, 29–36 (2020). <https://doi.org/10.26342/2020-65-3>
15. Navas-Loro, M., Rodríguez-Doncel, V., Santana-Perez, I., Sánchez, A.: Spanish corpus for sentiment analysis towards brands. In: Karpov, A., Potapova, R., Mporas, I. (eds.) *SPECOM 2017. LNCS (LNAI)*, vol. 10458, pp. 680–689. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66429-3_68
16. Padró, L., Stanilovsky, E.: FreeLing 3.0: towards wider multilinguality. In: *FreeLing 3.0: Towards Wider Multilinguality* (2012). upcommons.upc.edu/handle/2117/15986. Accepted 8 June 2012
17. Papay, S., Padó, S.: RiQuA: a corpus of rich quotation annotation for English literary text. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 835–841. European Language Resources Association, Marseille, May 2020
18. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. ELRA, Valletta, May 2010. <https://doi.org/10.13140/2.1.2393.1847>
19. Stymne, S., Östman, C.: SLäNda: an annotated corpus of narrative and dialogue in Swedish literary fiction. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 826–834. European Language Resources Association, Marseille, May 2020
20. Torres-Moreno, J.M.: *Megalite* (2022). hdl.handle.net/11403/megalite. ORTOLANG (Open Resources and TOols for LANGUAGE). www.ortolang.fr