Journal of Portuguese Linguistics

Santos, D., & Simões, A. (2022). Clustering emotions in Portuguese. *Journal of Portuguese Linguistics*, 21, pp. 1–33. DOI: https://doi.org/10.16995/jpl.8197

Open Library of Humanities

Clustering emotions in Portuguese

Diana Santos, University of Oslo & Linguateca, Norway, d.s.m.santos@ilos.uio.no
Alberto Simões, 2Ai, school of technology, IPCA & Linguateca, Portugal, ambs@zbr.pt

In this paper we present some exploratory studies of emotion words based on large annotated corpora of Portuguese. Those corpora were automatically annotated with emotionality, and each emotion word was assigned one or more groups out of 26 emotion groups. Our goal is to evaluate those groups by applying different statistical approaches to our material, namely based on (a) co-occurrence in a sentence as a sign of closeness of meaning, and (b) word embeddings. After looking at the full material, we turn our attention to two specific emotion groups: *Amor* ('love') and *Desespero* ('despair'), investigating whether clustering with those underlying techniques can help improve the shape, or redesign, particular emotion groups. In the paper we suggest some novel forms of measuring semantic coherence on word embedding models. Since computational research on emotion words in Portuguese is still rare, our methods and resources will lay the ground for future investigations.

1. Introduction

The study of emotions in language has received a wide interest in the last decades, especially in the specific alley of opinion mining, or sentiment analysis (Birjali et al., 2021; Pang & Lee, 2008). Emotions are a complex field where language, biology, psychology, society and even history (Boddice, 2018) interact (Barrett et al., 2018). In spite of the amount of recent research, there is still a lack of consensus on many levels. For example, there are not even consensual emotion categories that all or most scholars recognize. Many researchers, in fact, take a simplified view and only consider polarity (i.e., positive or negative). If we add the variable language, and the fact that each language has different lexical items and grammatical constructions to express emotion, it is clear that a lot remains to be done in this matter.

In this paper we consider reference to emotion in Portuguese. We are neither concerned with how texts may raise emotion in the readers nor how authors express their emotions. Our concern is to study the way words denoting emotion are used in text. We do not follow a particular theory of what an emotion is in psychological or physiological terms, and limit ourselves to the linguistic properties of emotion words. For surveys of other approaches and discussion of several theories, see the works by Maia and Santos (2018) and Santos and Maia (2018).

In this paper we want to evaluate a publicly available resource that provides emotion annotation in large Portuguese corpora (Santos, Simões & Mota, 2021). To build it, we used primarily our linguistic intuition as to whether a word denoted an emotion, and refined this process by consulting dictionaries, encyclopaedias, and corpora, producing 26 emotion groups in Portuguese, named by the most frequent members of the group.[1] See **Figure 1** for a quantitative overview, and translation, of those emotion groups, from the original paper.

Note that we do not claim that there are only 26 types of emotions in Portuguese. In fact, we have a group called *OUTRA* (‘other’), which explicitly groups emotion words for which we have not yet created a dedicated group. In addition, we have a group called *AUSENCIA* (‘absence’) which marks words that concern absence of emotion, like *impávido* (‘unmoved’) – the justification to do this is the analogy with the semantics of colour and clothing, where we have postulated a similar absence category that deals with words like *incolor* (‘colourless’) or *nu* (‘naked’). We have also, just like in these other semantic categories, created a GEN (“generic”) group, which deals with emotion in general – again, examples from the colour and clothing domains would be *colorido* (‘coloured’) and *vestir* (‘to dress’). We use this group in the emotion domain for words like *sentimento* (‘feeling’) or *emocionado* (‘moved’).

[1] The naming decision was unfortunate, because different classes received different grammatical categories (adjective or noun), and “most frequent” is based on the size of the corpus at the naming time (2015), and may therefore have changed already. In the present paper we will consistently use the nominal version in the English translation. We will keep the Portuguese names unchanged as they were in the original corpus.

Alívio	Admirar	Amor	Ausência	Coragem	Desejo	Desespero
Relief	Admiration	Love	Absence	Courage	Desire	Despair
130,088	244,764	1,451,760	37,673	360,859	1,976,440	146,320
229	208	918	22	422	381	146
Esperança	Feliz	Fúria	Gen	Grato	Humildade	Infeliz
Hope	Happiness	Anger	Generic	Gratitude	Humility	Unhappiness
774,172	737,833	755,594	517,701	288,786	824,867	712,024
268	758	428	243	201	214	552
Ingrato	Insatisfeito	Inveja	Medo	ódio	Orgulho	Outra
Ungratitude	Unsatisfaction	Envy	Fear	Hate	Pride	Other
6195	153,891	25,708	617,413	135,670	372,152	225,849
18	72	38	502	183	513	170
Pena	Satisfeito	Saudade	Surpresa	Vergonha		
Sorrow	Satisfaction	Longing	Surprise	Shame		
130,304	178,985	105,810	268,566	415,918		
201	251	76	198	612		

Figure 1: Table from Santos, Simões and Mota (2021), where the following properties are specified: the name of the group, its translation into English, the number of occurrences of words annotated with group, and the lexical diversity of each group in lemmas.

So, the emotion groups are just a quantitatively-based first approach to the categories that are recognized by Portuguese. If we noticed that a considerable set of distinct words could be assigned to a group (which would thus have a high lexical diversity), we created the group. But one may also state that what we have (so far) achieved is the identification of 23 groups for which we assigned a label, and that every individual word in the *OUTRA* ('other') category is a potential emotion group in Portuguese. We will be mindful to treat the cases of *AUSENCIA* ('absence'), *OUTRA* ('other') and GEN ("generic") groups in a special way.

Then, we have devised a considerable number of rules, using the approach sketched by Santos and Mota (2010) for human-machine cooperation, in order to deal with the inescapable property of language that most words have more than one meaning. We exemplify in the aforementioned paper that, in the case of emotion, there are endless cases where a given word describes an emotion in some contexts, and something not emotional in others, and propose a lexical measure called the degree of emotionality of a word. We also note that it is common for words to convey more than one emotional meaning, which entails that some words are assigned to more than one group. All this has been discussed and exemplified in several papers by Mota and Santos (2015), Santos and Mota (2015), Ramos, Santos and Freitas (2020) and Santos, Simões and Mota (2021), but these facts result in additional problems for our studies on the basis of the annotated cor.

In particular, how to deal with words which are assigned to more than one emotion category. We will come back to this issue later.

The purpose of our studies on top of this material (the resources just described) is to evaluate it properly and independently, in order to assess its value for the community that deals with Portuguese at large. It would be extremely reassuring if one could find automatic methods that could confirm our intuitions[2] and/or help us improve the groups. Although Santos, Simões and Mota (2021) have presented some data that confirms our intuition, still the following questions remain: how can one evaluate it properly, and not in so general terms? How can one be sure that those corpora provide more than our subjective opinion of what an emotion is? And particularly, we want to investigate whether it is possible to independently motivate the postulated emotion groups – is there any statistical data that lend them support?

To be more concrete, let us take three examples:

- Can big data-based automatic methods support the existence of the two groups *FELIZ* ('happiness') and *SATISFEITO* ('satisfaction') that were postulated, or on the other hand do the methods suggest their merging?
- Can statistical processing support that concepts like mistrust and despair should belong to the same group as they are now in *DESESPERO* ('despair'), or does language use show that they should be divided into two groups?
- Is there any statistical evidence that the group *AMOR* ('love') should also encompass friendship, as it does now?

So, the explorations described in the remainder of the paper are different attempts to use automatic techniques to evaluate the aforementioned annotation, and to investigate whether, by taking a quantitative bird's eye view, one can generate further knowledge on emotions in Portuguese.

It should also be said upfront that, no matter how many revision rules were designed to improve the (automatic) annotation, it is humanly impossible to review and correct the million annotated cases. So, the studies reported here will help identify specific problems and uncover places to improve the annotation. This will be clear when we analyse our findings.

Emotion annotation is an on-going process, in the sense that, at the same time we are doing these studies, other annotators are painstakingly analysing particular cases and improving the rules, as reported e.g., by Ramos (2021). So, what we present here concerns a particular time

[2] Are the groups well-chosen? Are the different emotion words that belong to a group correctly classified?

slice of the annotated corpora (October 2021), and our most relevant contribution should be the methods we use and propose.

We will be using two different approaches: co-occurrence information, and word embeddings, a vectorial representation of words based on their co-occurrence in large corpora.

2. Further information about the data we use

The main corpus used is *Todos*, merging all AC/DC corpora (Santos, 2014) together and removing repeated material. The corpus is described in more detail in Santos, Simões and Mota (2021). It purports to include mainly written Portuguese in several genres (newspaper, academic writing, interviews), amounting to ca. 1.5 billion tokens (as of October 2021), most of them (3/4) in Brazilian Portuguese. A small part (6.5%) is transcribed oral data (see Santos, 2016, on the different kinds of oral corpora). A subset of *Todos* which we thought interesting to explore too was *Literateca*, containing all literary text after removing repeated texts, amounting to ca. 40 million words. *Literateca* features mostly old texts (not contemporary Portuguese) and, contrary to *Todos,* mostly Portuguese from Portugal (70%) (*Todos* includes *Literateca*).

We believe in the importance to document the options taken to deal with the material, because they may be essential for replication, and for interpretation. We have thus used the following information from the corpora: tokenization, sentence separation, lemmatization (done by the PALAVRAS parser, see Bick, 2000, 2007, 2014) and semantic annotation for the emotion domain, as displayed in **Figure 2**.

```
<s>
O           o           0
amor        amor        emo:amor
traz-nos    trazer+nós  0
um          um          0
sentimento  sentimento  0
de          de          0
felicidade  felicidade  emo:feliz
</s>
```

Figure 2: A fictive example of the material, where we show the three columns we made use of: word, lemma and semantic annotation. Note that the annotation does not mark the word *sentimento* in the context of *sentimento de X* (“feeling of X”), X being an emotion, as explained in Ramos and Freitas (2019). The example also illustrates one case of a verb with enclitics, whose lemma is the two lemmas concatenated by the + sign: *trazer* is the lemma of the verb form *traz*, and *nós* is the lemma of the pronoun form *nos*.

For the co-occurrence approach, no further preprocessing was necessary. In order to look at the particular groups *AMOR* and *DESESPERO*, some manual pruning was done to the lemmas that were used in our graphs: we removed (a) cases where *PALAVRAS* had creatively added a derived lemma, such as *atagonia* or *paragonimíase* (incorrectly parsed as derivations from *agonia*); (b) cases of misspelled words that might be emotion words but appear too rarely to consider listing them in the emotion group, like *afectuiar* or *deprezador* or *desesperadamante*; (c) cases with non-standard capitalization or hyphenation, such as *aMIGO* or *en-ternecer*. We also removed *hapax legomena,* and added together all cases of clitics, so that lemmas *amar+ele*, *amar+se*, *amar+ele+o* and *amar* were lumped together under the lemma *amar*. This shrank the lemmas belonging to the AMOR group from 1162 to 87, and those of the DESESPERO group from 246 to 64 cases.

For the creation of word embeddings, we additionally removed all capitalization, and used only words with a frequency higher than 5. Four kinds of word embeddings were created:

1. standard, using the bare corpus, without any kind of word annotation
2. changing the words marked as emotional to emo:word
3. changing the words marked as emotion to emo:word::group
4. replacing the words marked as emotion with their group: emo:group

To make this more concrete, the word *empertigado* ('stiff') from the ORGULHO ('pride') group would have been coded like this in each approach:

1. empertigado
2. emo:empertigado
3. emo:empertigado::orgulho
4. emo:orgulho

This means that our word embeddings created on *Todos* had different sizes:

1. 1,171,525 word vectors
2. 1,177,040 word vectors
3. 1,177,420 word vectors
4. 1,163,962 word vectors

One should also note that clitics were not removed for the word embeddings, which means that for example the two words *admirei* and *admirei-me* count separately.

For the embeddings creation, we considered using Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017) or Glove (Pennington et al., 2014). Some authors (such as Romanov & Khusainova, 2021) claim that FastText gives better results for morphology queries, while the two others are more semantically aware. As to the choice between Word2Vec and Glove, the first uses a global word-to-word co-occurrence measure, while the second tries to work within local context. We chose Word2Vec.

To create the embeddings, we left most of Word2Vec's default options unchanged, but we used a dimension of 300 instead of 200 and 20 training iterations instead of 5. Three hundred was chosen to be comparable to most other public word embeddings for Portuguese, and our expectation was that increasing the number of training iterations would increase quality.[3]

3. Creating a co-occurrence graph

One standard way to reduce large corpora to quantitative objects easy to manipulate, and reduce their dimensionality, is simply counting the number of times different concepts, or words, co-occur, taking this as a measure of relatedness or even similarity.

This was our first approach, which we applied to words annotated with emotion, in two different ways:

1. simply using the emotion group (any word annotated as belonging to the group *FELIZ* ('happiness') would count as *FELIZ*, so *felicidade* and *ventura* (both translateable by happiness) would both count as *FELIZ* ('happiness'))
2. or using the word itself, for particular emotion groups (so, in the *FELIZ* group, *felicidade* and *ventura* would count as *felicidade* and *ventura*, respectively).

Since the corpora are parsed by *PALAVRAS*, we operationalized co-occurrence as "appearing in the same sentence", marked by the structural attribute <s>. We believe this to be more linguistically motivated than deciding on a fixed window of N words.

Figures 3 and **4** show the relationship between the emotion groups in *Todos* and *Literateca*, drawn like a graph, whose vertices are the group labels, with size (diameter) proportional to their frequency, and whose edges correspond to the attested co-occurrences, drawn with thickness corresponding to the number of co-occurrences.

[3] Preliminary experiments seem to show we were mistaken, there seems to be no significant difference.

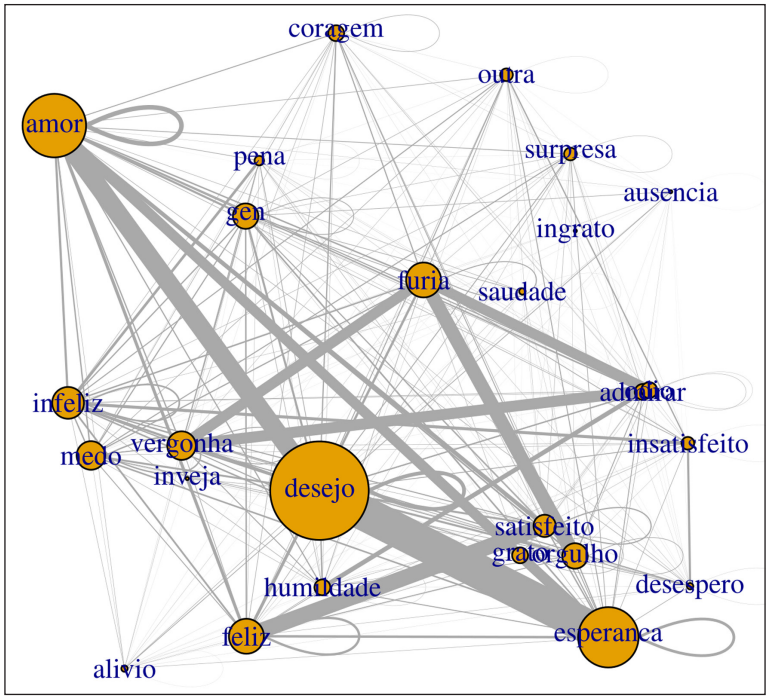

Figure 3: All emotion groups in *Todos*, using the R (R Development Core Team, 2008) package igraph (Csardi & Nepusz, 2006).

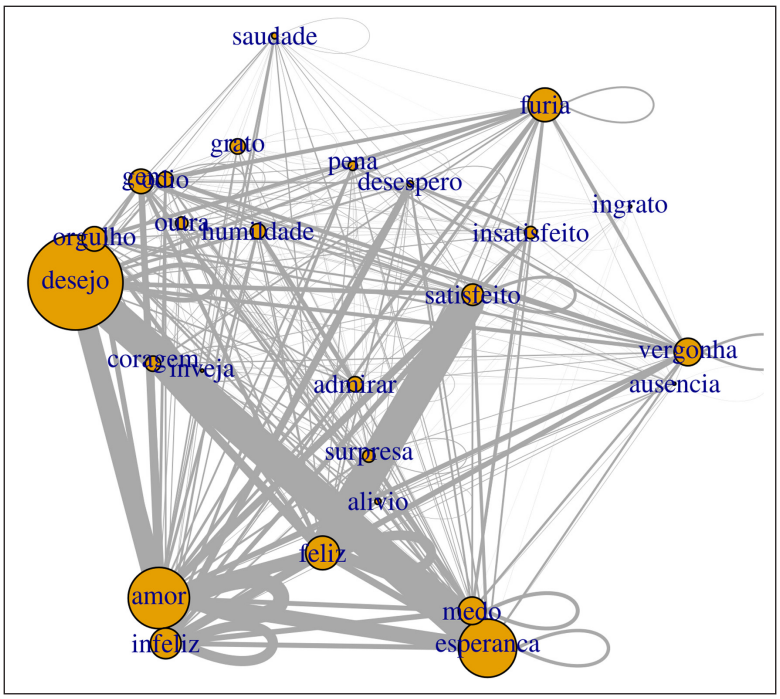

Figure 4: All emotion groups in literary text (*Literateca*).

It should be clarified that we use a random layout for drawing the graphs, which implies that the place where the particular categories appear is not meaningful – and therefore should not be compared across graphs. Also, it is important to explain that the sizes are relative to the universes, which differ widely in quantity: *Literateca* has about 40 million words, while *Todos* (which includes *Literateca*) amounts to a total of 1,315 million words.

With these caveats in mind, what can these graphical depictions tell us, as an initial overview of the annotated corpora? We can conclude that literary text in Portuguese (at least the one present in *Literateca*), has a stronger emphasis on *AMOR* ('love') and *(IN)FELIZ* ('(un)happiness') than other kinds of text. And that, in general, the most invoked emotion is *DESEJO* ('desire') followed by *ESPERANCA* ('hope') and *AMOR* ('love'). Initially surprised by the unexpected prominence of *DESEJO*, we soon understood its cause: words like *querer* ('want') and *desejar* ('desire') were considered as emotion words, even when they might arguably be considered only denoting volition or intent. If this were not the case, the *DESEJO* group would considerably shrink. In fact, by investigating the matter closer, we also realized that our co-occurrence counting procedure counts words annotated, e.g., with *desejo_amor* for both the *DESEJO* and *AMOR* groups, therefore inflating even more the (possibly dubious) contribution of the verb *desejar* to the Portuguese emotion realm. In any case, this illustrates that the original decision to attribute as many emotion labels as deemed appropriate can have consequences on the further processing of the material. Since we are not sure what the best alternative to deal with these cases is – and suspect that they may, in fact, indicate a desirable merge of the groups in question –, we did not compute an alternative co-occurrence matrix.

In **Figures 5** to **8**, we present now the two groups *AMOR* ('love') and *DESESPERO* ('despair') for both the *Literateca* and the *Todos* material, together with the pairs with more co-occurrences, in **Tables 1** and **2**.

We can thus observe several specific differences between the group *AMOR* ('love') in general (in all genres present), and in literary text: it is easy to see that while romantic love is the most described in the literary texts of *Literateca*, *amigos* ('friends'), *gostar* ('to like') and *preferir* ('to prefer') are the most common members of AMOR in other genres. Interestingly, *abraçar* and *beijar* ('hug' and 'kiss') rank high in the literature list, *preferir* ('to prefer') or *desejar* ('to wish for') are more frequent, when one looks at all genres together.

In **Table 2**, we present again the most common co-occurrences, now for the *DESESPERO* group. Note that the quantities are much lower than in **Table 1**, because reference to this group is far less frequent in Portuguese, at least according to the corpora we are using (cf. **Figure 1**).

Although the material has far fewer instances to analyse, reference to mistrust – represented in **Table 2** by the words *desconfiança* e *desconfiar* – does not occur often in literary texts, contrarily to general texts. And it seems that (from the most common co-occurrences only) the words related to mistrust and those related to despair keep separate. This is in sharp contrast with love and friendship, where there were plenty of co-occurrences, as can be appreciated in **Table 1**.

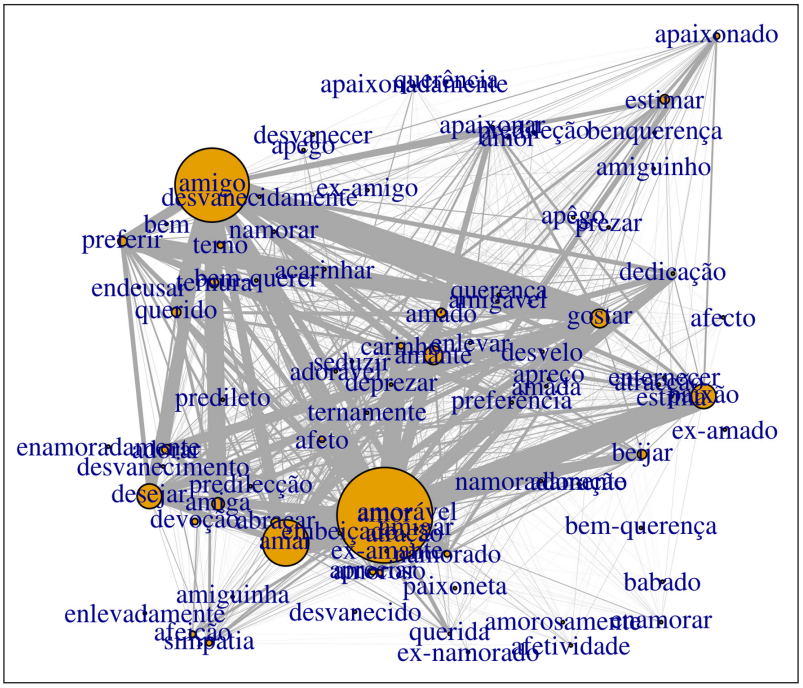

Figure 5: The words belonging to the *Amor* ('love') group in *Todos*.

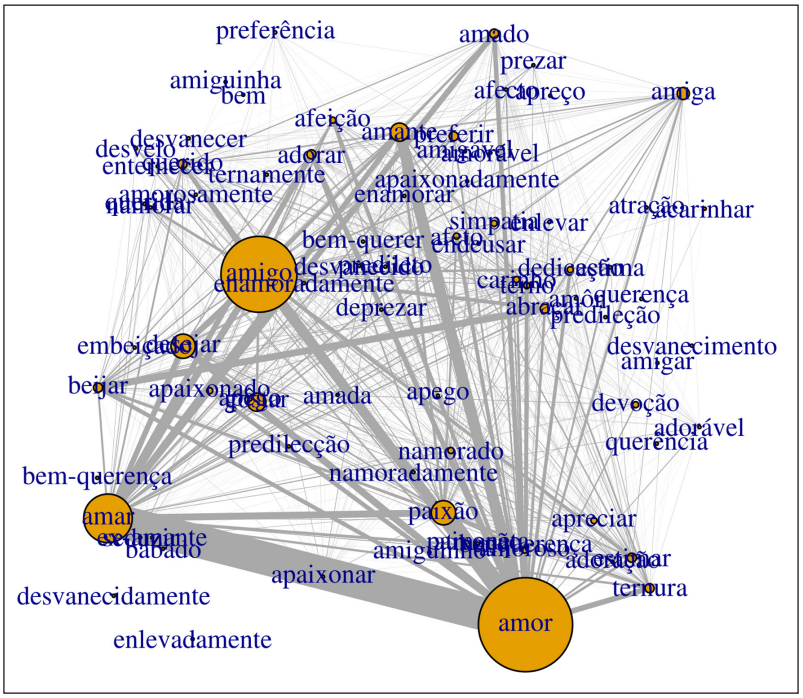

Figure 6: The words belonging to the *Amor* ('love') group in literary text (*Literateca*).

Todos			*Literateca*		
words		**coocc.**	**words**		**coocc.**
amar	*amor*	3001	*amar*	*amor*	1268
amigo	***gostar***	2337	*amor*	*paixão*	682
amor	*paixão*	2067	*amante*	*amor*	414
amor	***carinho***	1607	*amigo*	*amor*	368
amiga	***amigo***	1515	***amor***	***desejar***	361
amigo	*amor*	1405	***amar***	***paixão***	257
gostar	***preferir***	1387	***amar***	***amigo***	253
amigo	***querido***	1334	***abraçar***	***beijar***	247
amor	***dedicação***	1037	***amado***	***amar***	225
amante	*amor*	975	*amigo*	*desejar*	215

Table 1: The ten most common co-occurrences between words of the *AMOR* group in all corpora (*Todos*) and in the literary corpora (*Literateca*); in bold those that appear in the top 10 in one list and not in the other.

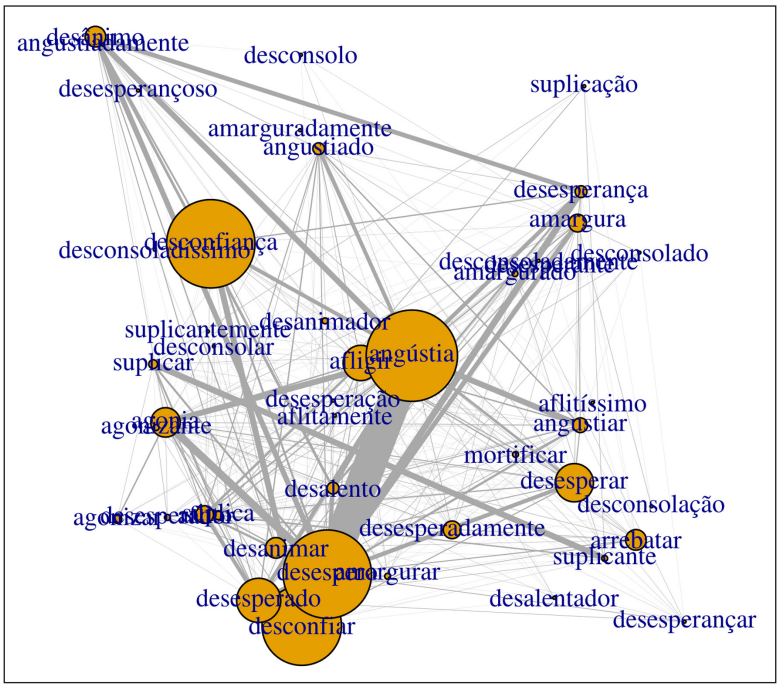

Figure 7: The words belonging to the *DESESPERO* ('despair') group in *Todos*.

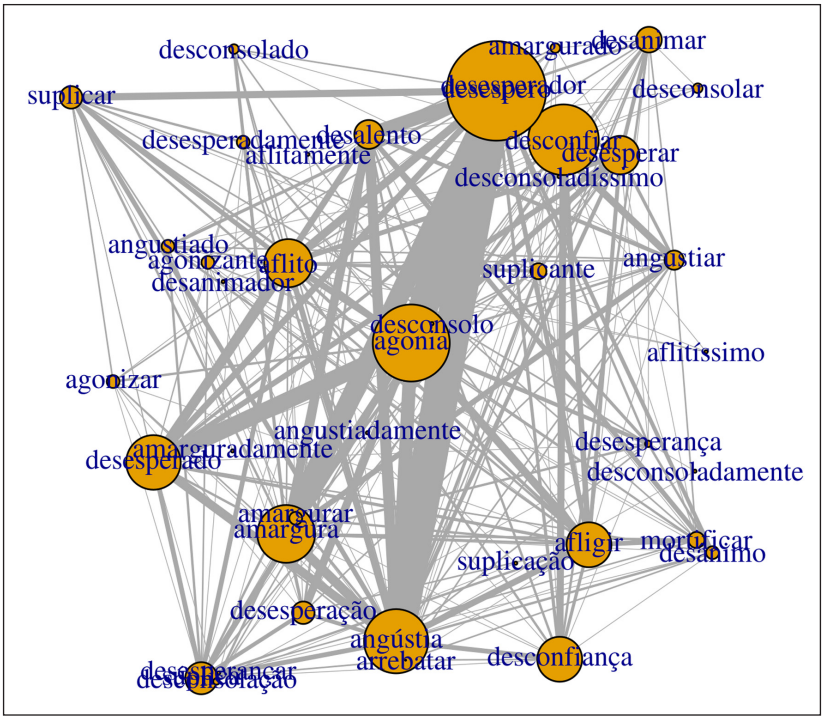

Figure 8: The words belonging to the *DESESPERO* (‘despair’) group in *Literateca*.

Todos			***Literateca***		
words		**coocc.**	**words**		**coocc.**
angústia	*desespero*	264	*angústia*	*desespero*	35
desesperado	***desespero***	262	*agonia*	*desespero*	31
desesperança	***desespero***	259	***amargura***	***desespero***	23
agonia	*desespero*	58	***agonia***	***desesperado***	21
suplicante	***suplicar***	54	*agonia*	*angústia*	15
desconfiança	***desconfiar***	51	***desalento***	***desespero***	14
angustiar	***angústia***	49	***amargura***	***desalento***	10
angústia	***desânimo***	49	***amargura***	***angústia***	10
agonia	*angústia*	48	***agonia***	***desesperação***	10
desespero	***desânimo***	46	***angústia***	***desesperado***	9

Table 2: The ten most common co-occurrences between words of the *DESESPERO* (‘despair’) group in all corpora (left) and in the literary corpora (right); in bold, those that appear in the top 10 in one list and not in the other.

After this preliminary bird's eye view, we have applied several clustering techniques to the co-occurrence material, which we proceed to describe below.

The idea of clustering, a non-supervised exploratory technique, is to identify meaningful groups ("clusters") in large amounts of data. There are two major ways to proceed: divisive clustering, which starts by dividing the material, and agglomerative clustering, which proceeds bottom-up by joining the closest elements. In any case, these processes depend on a distance measure between the objects to be clustered, and there are many different distance measures to choose from.

For the co-occurrence data, we considered that co-occurrence between two words (or emotion groups) measured how close they were, and defined distance as the inverse of the co-occurrence number (so, if X co-occurred 43 times with Y, their distance would be 1/43). When there were no co-occurrences at all in the material, we assumed infinite distance. We then applied multidimensional scaling with two and three dimensions to the emotion group co-occurrence data. The result for two dimensions is in **Figure 9**.

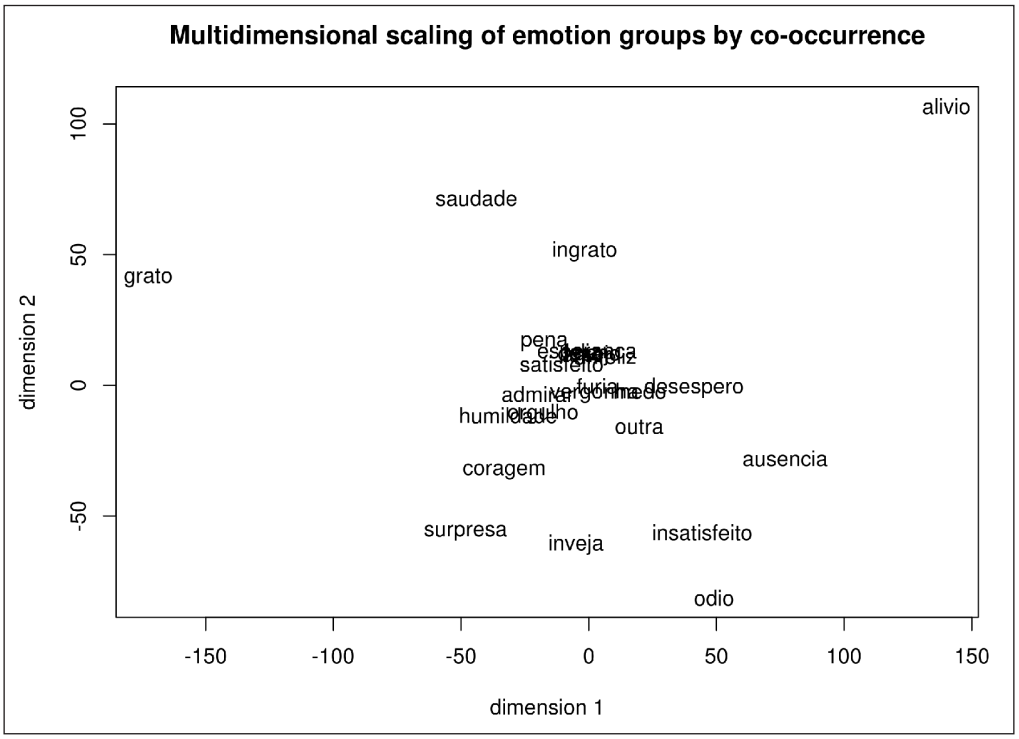

Figure 9: Multidimensional scaling of the data obtained by co-occurrence of the emotion groups in *Todos*.

The result shows that dimension 2 clearly separates emotions from their absence. However, it is hard to understand what it is that dimension 1 captures, singling out two not very typical

emotion groups: *INGRATO* (‘ingratitude’) and *INVEJA* (‘envy’), in fact the two groups with lowest lexical diversity in Portuguese according to the corpora (cf. again **Figure 1**).

If we redo multidimensional scaling requiring three dimensions, see **Figure 10**, dimension 3 now singles out *SAUDADE* (‘longing’), which is the next least lexically diverse emotion group, apart from *INSATISFEITO* (‘insatisfaction’).

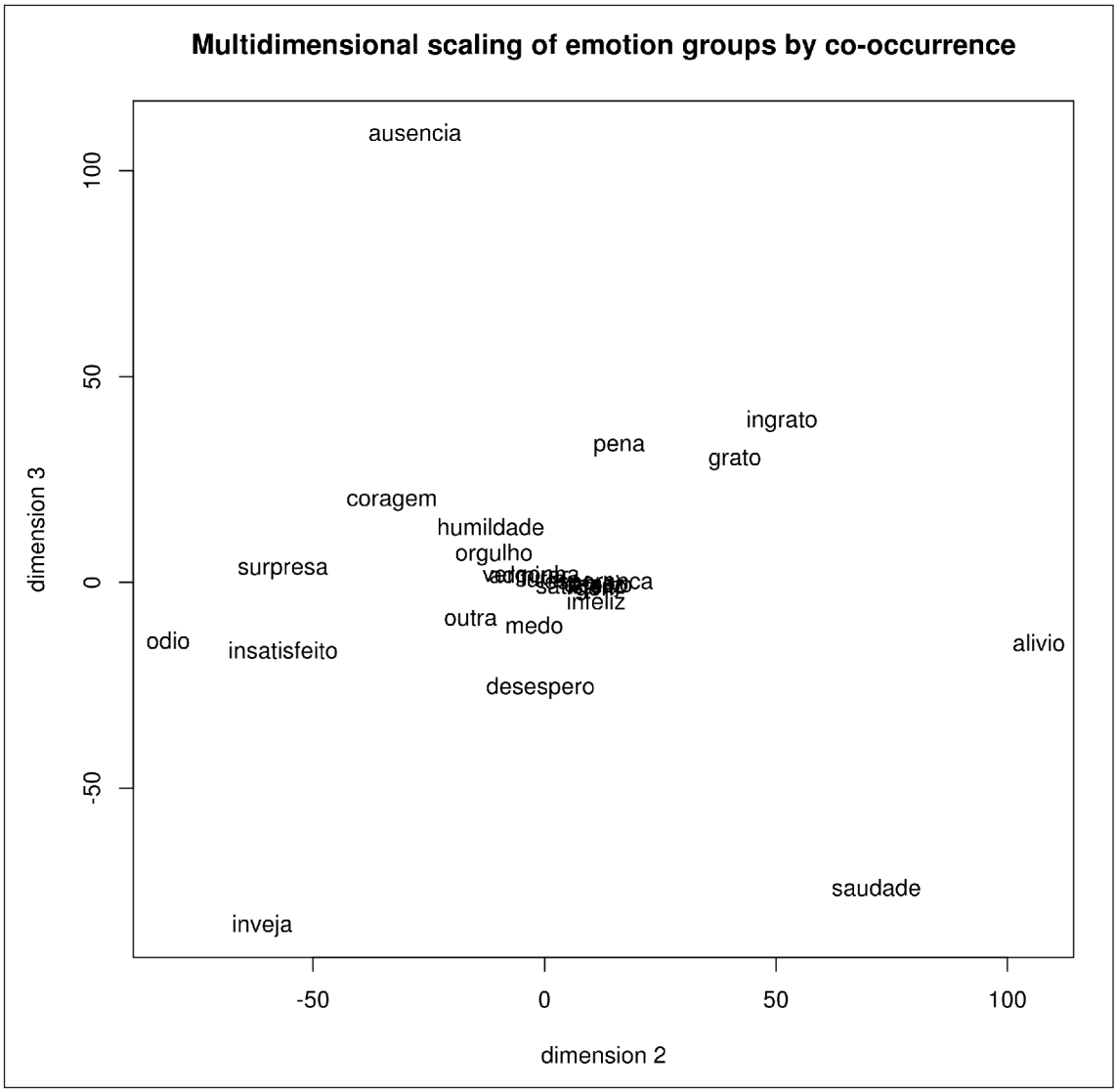

Figure 10: Multidimensional scaling of the data obtained by co-occurrence of the emotion groups in *Todos*, with 3 dimensions.

Trying hierarchical clustering (with R’s hclust command), we get a similar result, see **Figure 11**.

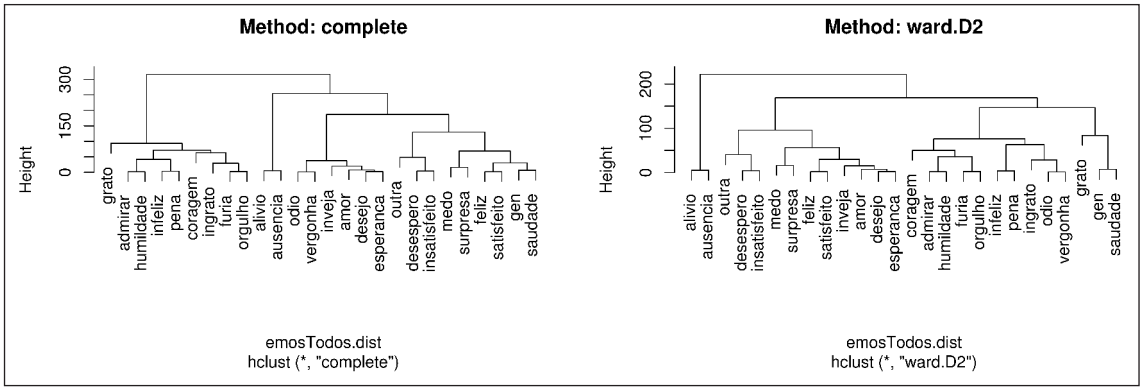

Figure 11: Hierarchical clustering of the data obtained by co-occurrence of the emotion groups in *Todos*, by two different methods.

Table 3a shows the most frequent emotion group co-occurrences.

group pair		**co-occurrence**
desejo	*esperanca*	517437
amor	*desejo*	289875
satisfeito	*feliz*	216041
orgulho	*furia*	202921
odio	*vergonha*	159740
amor	*esperanca*	154837
vergonha	*furia*	151276
odio	*furia*	146411
amor	*amor*	68827
admirar	*humildade*	66455

Table 3a: The ten most common co-occurrences between emotion groups in *Todos*.

Most of these numbers can be interpreted straightforwardly: one joins quasi-synonyms (*FELIZ* ('happy') and *SATISFEITO* ('satisfied')), another shows the cohesiveness of the same group (*AMOR* ('love')), while others join feelings that often come together, like *ODIO* ('hate') and *VERGONHA* ('shame'), *VERGONHA* ('shame') and *FURIA* ('anger'), or *DESEJO* ('desire') and *ESPERANCA*

('hope'), *AMOR* ('love') and *FELIZ* ('happiness'), and *ADMIRAR* ('admiration') and *HUMILDADE* ('humility'). Finally, one may also interpret *ORGULHO* ('pride') as a cause for *FURIA* ('anger'), although obviously not always. It is nevertheless interesting that no antonyms come to the fore: all pairs are either both positive or both negative.

One should also recall that there is a significant number of word occurrences that are marked as belonging to *AMOR* ('love') and *DESEJO* ('desire') or *AMOR* ('love') and *ESPERANÇA* ('hope'), and these would inflate (artificially, in fact) the number of the co-occurrences of the two categories. Namely, all words marked as belonging to a double or triple category count as co-occurrences among these categories. This is something that we have to deal with, and an alternative **Table 3b** was therefore created without those cases.

groups		**co-occurrence**
amor	*amor*	59064
amor	*desejo*	43803
desejo	*esperanca*	28780
amor	*esperanca*	27075
gen	*amor*	26917
infeliz	*amor*	25121
amor	*feliz*	24682
desejo	*desejo*	22582
medo	*medo*	20056
gen	*desejo*	19976

Table 3b: The ten most common co-occurrences between emotion groups in *Todos,* disregarding words with more than one category.

We see that the quantities are considerably smaller than those of **Table 3a**, showing that many of these co-occurrences involved (or were a product of) vague categories. In this new table, there are three emotion groups that co-occur with themselves: *AMOR*, *DESEJO* and *MEDO*. However, the categories that are included in vague classifications continue to be frequently co-occurring, which in a way vindicates the existence of words that convey both.

4. Investigating word embeddings

For almost a decade now, the technique of using large amounts of data to produce (static) word embeddings has been actively used in many different NLP tasks in order to provide a better representation of a word's meaning, and has also been applied in other linguistic and literary contexts (Antoniak & Mimno, 2018). Although there are fortunately several word embeddings for Portuguese (see Batista, 2019, for an overview of Hartmann et al., 2018; Rodrigues & Branco, 2018, Grave et al., 2018 and Yamada et al., 2016; and Santos, 2021, for a recent comparison among them), we decided to create our own embeddings based on precisely the data we wanted to analyse, also because, as explained in Section 2 above, we tried four kinds of word embeddings.

However, one thing that stood out was that there is a scarcity of research that uses clustering over word embeddings. Tang et al. (2014) claimed that poor results of clustering over word embeddings are due to the fact that traditional word embeddings are based on substitutability, not similarity, and so "they cannot distinguish words with similar context but opposite sentiment polarity (e.g., *good* and *bad*)" (Tang et al., 2014, p. 1563). This means that, in a word embedding representation, antonyms are closer than unrelated words, since they are often substitutable. Another property of antonymy has been pointed out by Justeson and Katz (1992), who suggested that corpus co-occurrence is a textual marker for the antonymy lexical-semantic relation. In other words, antonyms tend to co-occur in text.

Before clustering, we tried to exploit the information gathered by the word embeddings in several ways, as we describe in what follows.

4.1 Emotions near emotions?

We first set out to investigate whether words annotated as emotions also have emotions as their nearest neighbours (most similar words) in word embeddings. In order to do this, we computed the most similar words for the 3 embedding models where emotions were explicitly marked (using Gensim's (Rehurek & Sojka, 2010) method similar_by_word from its KeyedVectors module), and extracted the following statistics:

1. how many emotion words were included in the first 50 closest words
2. what was the position in the top 50 closest words of the first emotion (-1 if none was an emotion)
3. the sum of the inverse ranks of the 50 closest words which were considered an emotion.

Let us illustrate the third statistic with the help of **Figure12**, which lists the 50 nearest neighbours of *emo:amor*. The rankings of emotion words (in bold) are then 2, 3, 4, 5, 12, 13, 14, 16, 18, 19, 20, 21, 25, 29, 30, 31, 33, 34, 36, 41, 43, 45, and 46, and the statistic amounts to 1/2 + 1/3 +

1/4 + 1/5 + 1/12 + 1/13 + 1/14 + 1/16 + 1/18 + 1/19 + 1/20 + 1/21 + 1/25 + 1/29 + 1/30 + 1/31 + 1/33 + 1/34 + 1/36 + 1/41 + 1/43 + 1/45 + 1/46, equalling 2.10.

Generically, $\sum_{w} \frac{1}{\text{pos}(w)}$ where *w* is each one of the similar (emotion) words, and *pos(w)* is the rank of that word.

[('amor', 0.7659767270088196), ('**emo:afeto**', 0.7045239806175232), ('**emo:ódio**', 0.6937524080276489), ('**emo:prazer**', 0.6579146385192871), ('**emo:ciúme**', 0.6536670923233032), ('erotismo', 0.6392281651496887), ('encantamento', 0.6367102861404419), ('homem', 0.633893609046936), ('encanto', 0.631511390209198), ('egoísmo', 0.6305296421051025), ('sentimento', 0.6200258135795593), ('**emo:horror**', 0.6196451783180237), ('**emo:ternura**', 0.6182450652122498), ('**emo:compaixão**', 0.6177799701690674), ('esposo', 0.6098265051841736), ('**emo:desamor**', 0.608447253704071), ('deus', 0.6077516674995422), ('**emo:amante**', 0.6072775721549988), ('**emo:carinho**', 0.6070752143859863), ('**emo:heroísmo**', 0.6055108904838562), ('**emo:amado**', 0.603879988193512), ('instinto', 0.603283703327179), ('remorso', 0.6010612845420837), ('martírio', 0.5984621047973633), ('**emo:paixão**', 0.598348081111908), ('arrebatamento', 0.5970234274864197), ('sublime', 0.5958089232444763), ('pai', 0.5955429673194885), ('**emo:gozo**', 0.5941588282585144), ('**emo:desejo**', 0.5933477878570557), ('**emo:sentimento**', 0.5914780497550964), ('casamento', 0.5867379307746887), ('**emo:amar**', 0.5865211486816406), ('**emo:arrependimento**', 0.5804072618484497), ('marido', 0.5775448679924011), ('**emo:fascínio**', 0.576038658618927), ('lirismo', 0.5758034586906433), ('tédio', 0.5755293965339661), ('mistério', 0.5754537582397461), ('coração', 0.5726546049118042), ('affecto', 0.5726374387741089), ('**emo:desprezo**', 0.5715218782424927), ('infortúnio', 0.5671573877334595), ('**emo:sonho**', 0.5664375424385071), ('idealismo', 0.5649288296699524), ('**emo:sofrimento**', 0.5637087821960449), ('**emo:afeição**', 0.5631664991378784), ('demônio', 0.5607395768165588), ('noivo', 0.5600885152816772), ('sacrifício', 0.5595508217811584)]

Figure 12: The closest neighbours of *emo:amor* in the second kind of word embeddings, where we highlighted in bold those which were also marked as emotion by the prefix *emo:*. The corpora include spellings from all Portuguese varieties, and old orthographies.

This example at once shows several interesting features: the closest word to *emo:amor* is... *amor* itself! But this second *amor* was not considered an emotion in the context it appeared in – we can guess it probably was included in a proper noun like a movie or book title, and proper nouns were not annotated with emotions. Another observation is that *encantamento* ('enchantment'), *encanto* ('enchantment'), *remorso* ('remorse') and even *arrebatamento* ('rapture') do seem to us quite good emotion candidates, although they weren't considered as such. Therefore, we can give this kind of feedback to enhance the lexicon and/or the rules, so that the corpus annotators can include these terms in the next round. Finally, the word *coração* ('heart') is also quite interesting because of the well-known relationship between body organs and emotions (see e.g., Enfield & Wierzbicka, 2002). Any Portuguese native speaker is used to the metaphor that love is located in the heart (a metaphor which is also incidentally shared by English and many other languages). Should one have also tagged (some) body parts as emotions? We have tagged body parts as possibly meaning emotions in another project (Freitas et al., 2015), but making use of a different

syntax, and we have not so far converted that information into emotion annotation which could have been used in our experiments here.

These observations show that the results we obtain may not be final, and that several things could have be enhanced or done differently.

In order to make sense of all emotion words and not single out just one instance, we computed the three statistics given above for all the words of our 3 word embedding models. In **Figure 13**, we present a histogram of the results for the first statistic (how many words in the top 50 are also emotions), based on the second model, where words denoting emotion in context were prefixed with *emo:*

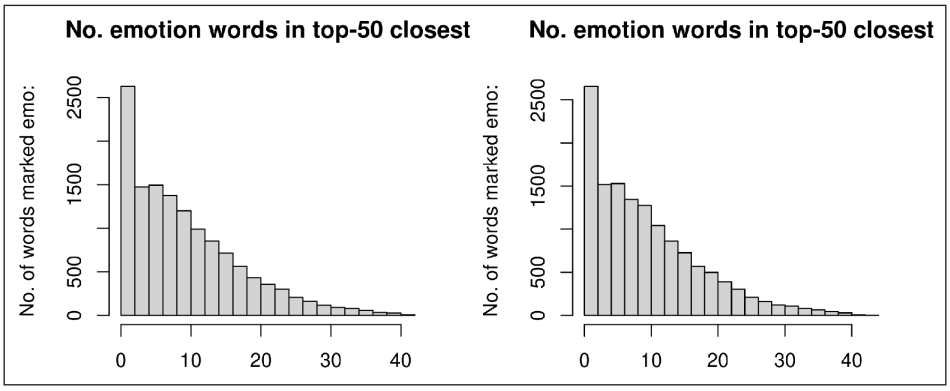

Figure 13: Distribution of the first statistic in the second (left) and third (right) kind of word embeddings, where words marked as emotion had the prefix *emo:*.

In **Figure 13**, we can see that there are wide differences between words. There are many emotion words (1,120 for the second word embeddings) that do not have an emotion among the closest neighbours, while others (216) have more than 40 out of 50. We assessed whether word frequency correlated with number of emotions as closest numbers, but the value of 0.11 for the Pearson's product moment correlation coefficient means that there is no correlation at all.

We have also created a corresponding histogram for the third kind of word embeddings, because we have slightly different *words* (remember that the *words* here have a prefix added, see Section 2 above). The lexicon gets larger (the same word can be in many different groups) and embeddings may differ. In fact, the random component underlying Word2Vec's algorithm has often been challenged with lack of stability (see Mihalcea, 2021), which means that different runs of the algorithm with the same parameters can actually produce wildly different results.

In **Figure 14**, we use the third kind of word embeddings (where we keep both word and group) and restricting our attention to those emotion words with at least one emotion in their closest 50 neighbours, we check whether they belong to the same emotion group. We create a fourth statistic that measures how many of the closest emotions belong to the same group.

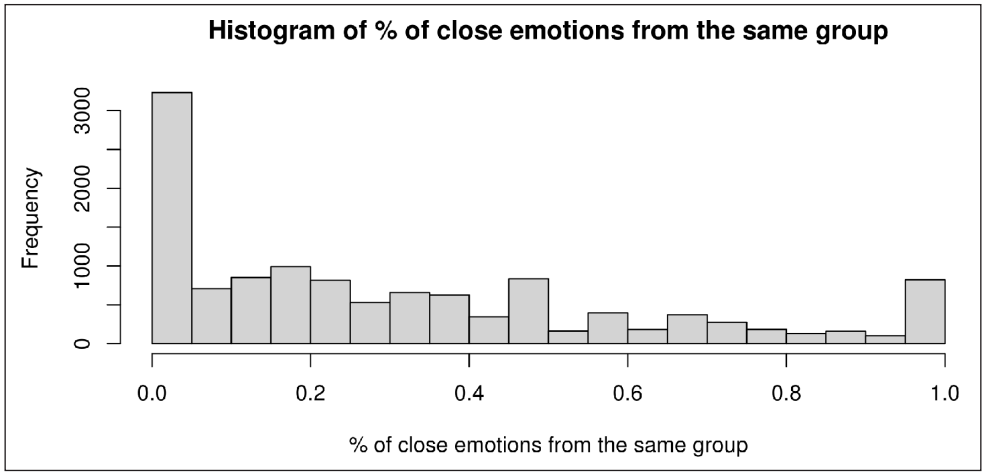

Figure 14: Distribution of the fourth statistic for the third kind of word embeddings, where words marked as emotion were prefixed by their emotion group. Only words with at least one emotion neighbour are taken into consideration.

The results leave room for improvement. We believe that a decisive factor could be the property of antonyms that has been already discussed: words tend to co-occur with their antonyms, which are therefore closer to them than a word taken at random. So as a next step for future work, one should find a way to mark antonym emotion groups and check whether the percentages in **Figure 14** would significantly increase.

4.2 Which emotion groups are the closest?

Using the fourth word embeddings, where we replaced all words of a particular group by their group name (keeping vague group names, like *AMOR_DESEJO* ('love_desire'), as separate groups), we can use this representation to obtain similarities between groups.

The group *DESEJO* ('desire') has as the closest and only neighbours the groups *AMOR_DESEJO_ESPERANCA* ('love_desire_hope'), *AMOR* ('love') and *DESEJO_ESPERANCA* ('desire_hope'), while *AMOR* ('LOVE') has many neighbours (16), in the following order: *FELIZ_SATISFEITO*, *FELIZ*, *INVEJA*, *DESEJO*, *ADMIRAR*, *ODIO*, *SATISFEITO*, GEN, *DESESPERO*, *ESPERANCA*, *SURPRESA*, *DESEJO_ESPERANCA*, *MEDO*, *ORGULHO*, *HUMILDADE_ADMIRAR* and *PENA*. The closest neighbours (21) of *DESESPERO* ('despair') are as follows: *MEDO*, *INFELIZ*, *SURPRESA*, *MEDO_SURPRESA*, *FURIA*, *INSATISFEITO_OUTRA*, *ALIVIO*, *ODIO*, *INVEJA*, *INFELIZ_DESESPERO*, GEN, *INFELIZ_INSATISFEITO*, *FELIZ*, *AUSENCIA*, *FURIA_ODIO*, *ESPERANÇA*, *VERGONHA*, *FELIZ_SATISFEITO*, *OUTRA*, *SAUDADE* and *INFELIZ_VERGONHA*.

These data are more difficult to interpret than those of the co-occurrence experiments, but we may hypothesize that the *DESESPERO* ('despair') group is the one with the most neighbours precisely because it has fewer occurrences and therefore it is more difficult to become

autonomous. Conversely, the fact that *DESEJO* ('desire') has only three group neighbours may show that either it is quite different/far away from the rest of emotions, and/or that it belongs to another (possible) cluster, namely that of volition.

One could use these similarities both to rank these groups in terms of their actual emotionality, and to measure their proximity with other emotions.

In **Table 4** we show the raw results (for these groups).

emotion group	**close groups**	**Gensim similarity**
emo:desejo	*emo:amor_desejo_esperanca*	0,74783319
	emo:amor	0,60807204
	emo:desejo_esperanca	0,59662843
emo:amor	*emo:feliz_satisfeito*	0,65850675
	emo:feliz	0,62952852
	emo:inveja	0,62798548
	emo:desejo	0,60807198
	emo:admirar	0,60569620
	emo:odio	0,59930396
	emo:satisfeito	0,55990100
	emo:gen	0,54665804
	emo:desespero	0,53727543
	emo:esperanca	0,51752120
	emo:surpresa	0,50839704
	emo:desejo_esperanca|	0,48341280
	emo:medo	0.47919095
	emo:orgulho	0,47487900
	emo:humildade_admirar	0,47156897
	emo:pena	0,46538004

(Contd.)

emotion group	close groups	Gensim similarity
emo:desespero	*emo:medo*	0,79399341
	emo:infeliz	0,69591373
	emo:surpresa	0,68163306
	emo:medo_surpresa	0,67511058
	emo:furia	0,66599011
	emo:insatisfeito_outra	0,66186285
	emo:alivio	0,64180046
	emo:odio	0,64122468
	emo:inveja	0,64011359
	emo:infeliz_desespero	0,63227987
	emo:gen	0,63055670
	emo:infeliz_insatisfeito	0,62733608
	emo:feliz	0,62732589
	emo:ausencia	0,61817032
	emo:furia_odio	0,61506921
	emo:esperanca	0,61258727
	emo:vergonha	0,60081607
	emo:feliz_satisfeito	0,57563734
	emo:outra	0,57479715
	emo:saudade	0,55613655
	emo:infeliz_vergonha	0,54775155

Table 4: The closeness among some groups using the fourth kind of word embeddings: after the group comes the Gensim similarity.

In order to provide a better way to understand these data, we have also tried to create partial pictures for each emotion, in **Figure 15**.

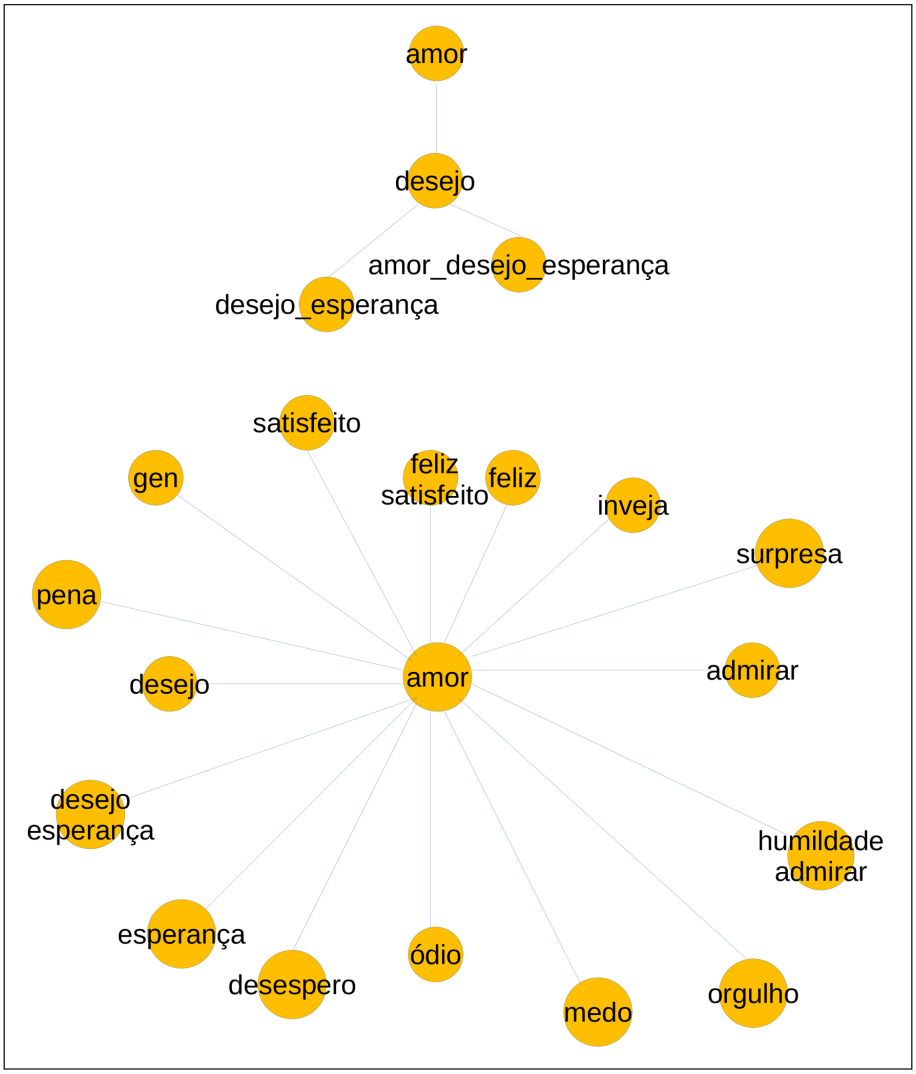

Figure 15: Representing the data of Table 4 in graph form, by (manually) drawing the other emotions with a distance inversely proportional to their similarity.

4.3 What about the pure word embeddings?

So far, we have only used our three word embedding models that make use of the annotation. But could anything be learned, or discovered, from the traditional word embeddings that require no corpus annotation? We could even compare the results with other word embeddings for Portuguese.

So, we decided to try them out by choosing the words *amor* and *desespero* (the ones that, after all, gave the names to the groups we have been looking at) and manually identify their close neighbours that are emotions, in **Figure 16**.

[('*ódio*', 0.6967464089393616), ('ciúme', 0.6639262437820435), ('afeto', 0.6538246273994446), ('encanto', 0.6484927535057068), ('amante', 0.6305267214775085), ('**martírio**', 0.6269147396087646), ('*ternura*', 0.6203771233558655), ('prazer', 0.6160613894462585), ('horror', 0.6149895191192627), ('**homem**', 0.6148678660392761), ('**esposo**', 0.6134668588638306), ('*desamor*', 0.6080464720726013), ('**deus**', 0.6049481630325317), ('encantamento', 0.6027072072029114), ('**coração**', 0.6024234890937805), ('**erotismo**', 0.5970407128334045), ('**pai**', 0.596354603767395), ('egoísmo', 0.5961049199104309), ('**anjo**', 0.5942661762237549), ('**sublime**', 0.592392683029174 8), ('*carinho*', 0.5912818908691406), ('remorso', 0.5890697240829468), ('**instinto**', 0.5875118374824524), ('compaixão', 0.5865410566329956), ('*amar*', 0.5843858122825623), ('**marido**', 0.5837520956993103), ('gozo', 0.5832496285438538), ('arrependimento', 0.5825256109237671), ('paixão', 0.5823038220405579), ('**pecado**', 0.5819383263587952), ('**casamento**', 0.5781994462013245), ('**sonho**', 0.5765835642814636), ('beijo', 0.5765777230262756), ('**demônio**', 0.5759337544441223), ('tédio', 0.5746782422065735), ('**poeta**', 0.5734331011772156), ('desprezo', 0.5716856718063354), ('**génio**', 0.5710428357124329), ('**mistério**', 0.5702476501464844), ('afecto', 0.5701011419296265), ('**noivo**', 0.5684205293655396), ('infortúnio', 0.568371057510376), ('desejo', 0.5682575106620789), ('**riso**', 0.5666162967681885), ('sentimento', 0.5647931694984436), ('êxtase', 0.5643293261528015), ('**lirismo**', 0.5605846643447876), ('drama', 0.5598238706588745), ('rancor', 0.5595110058784485), ('affecto', 0.5583791732788086)]

[('desalento', 0.7793346643447876), ('desânimo', 0.749446451663971), ('pavor', 0.7067726254463196), ('tédio', 0.7026311755180359), ('**delírio**', 0.6774165630340576), ('sofrimento', 0.6745300889015198), ('horror', 0.6716588735580444), ('ódio', 0.6674385666847229), ('aborrecimento', 0.6592799425125122), ('desgosto', 0.6571463942527771), ('pânico', 0.6562241911888123), ('infortúnio', 0.6560782790184021), ('desencanto', 0.6483257412910461), ('aflição', 0.6463565826416016), ('espanto', 0.6443071961402893), ('terror', 0.6340198516845703), ('desamparo', 0.6320211887359619), ('ressentimento', 0.6297689080238342), ('assombro', 0.6283164620399475), ('"panto", 0.6274684071540833), ('**cansaço**', 0.6268096566200256), ('remorso', 0.6251256465911865), ('tristeza', 0.624626874923706), ('rancor', 0.6213040947914124), ('medo', 0.617042064666748), ('angústia', 0.605400025844574), ('**desvario**', 0.5995854735374451), ('êxtase', 0.5976307988166809), ('**pesadelo**', 0.5912253856658936), ('desapontamento', 0.5910376310348511), ('**martírio**', 0.590295672416687), ('susto', 0.5895748138427734), ('ciúme', 0.5877094268798828), ('arrependimento', 0.583760678768158), ('nervosismo', 0.582927405834198), ('temor', 0.5791856646537781), ('arrebatamento', 0.5784072875976562), ('desassossego', 0.5769583582878113), ('tormento', 0.5766277313232422), ('desconsolo', 0.5752002000808716), ('júbilo', 0.5749205350875854), ('**riso**', 0.5746715664863586), ('**silêncio**', 0.5734441876411438), ('**desconcerto**', 0.5717650651931763), ('egoísmo', 0.5689138770103455), ('desprezo', 0.5676848888397217), ('furor', 0.567651093006134), ('**fanatismo**', 0.5669714212417603), ('**descrédito**', 0.5666866898536682), ('**pessimismo**', 0.5662282109260559)]

Figure 16: The 50 closest words with the first word embeddings for the words *amor* and *desespero*, where the non-emotions are in bold, and words from the same group or antonyms are in italic.

We see that 21 out of 50 are not emotional words for *amor*, while only 11 out of 50 do not describe emotion for *desespero*. This means that regular word embeddings fare very well compared with those created with explicit annotation. However, these test cases may not have been the best ones, since they correspond to cases where the word (*amor*, *desespero*) is always an emotion, as opposed to *pena* ('sorrow', 'feather', 'punishment', 'pen', etc.) or *reconhecer (*'be grateful', 'recognize', etc.*)*.

4.4 Clustering based on embeddings

Finally, we also attempted the direct use of the k-means algorithm, implemented in the scikit-learn Python package (Pedregosa et al., 2011), to group the embeddings. The k-means algorithm groups vectors on k clusters, where k is a predefined value. It works by classifying each vector in the cluster with the nearest mean. The main problem of this method is the requirement of predefining the number of desired clusters, see Lloyd (1982). This was performed directly, importing each word embedding vector for each one of the groups. With the idea of finding groups that could be merged, or that are semantically closer, we asked for 20 clusters, shown in **Table 5**.

1	*inveja, amor, ódio*
2	*amor_desejo_esperança, desejo, desejo_esperança*
3	*amor_desejo_orgulho, amor_orgulho, grato*
4	*infeliz, medo_surpresa, desespero, gen, furia_odio, medo, alivio*
5	*furia, vergonha, infeliz_pena, orgulho*
6	*infeliz_insastifeito_outra, outra, infeliz_insastifeito*
7	*humildade, odio_vergonha, pena, insatisfeito*
8	*amo_desejo*
9	*infeliz_desespero, feliz_orgulho, medo_infeliz, feliz_amor, infeliz_vergonha, coragem, orgulho_vergonha, coragem_furia, ingrato, orgulho_admirar, desespero_furia*
10	*furia_orgulho*
11	*coragem_ausencia, furia_outra, ausencia, gen_ausencia, alivio_ausencia*
12	*humildade_vergonha, pena_vergonha*
13	*humildade_admirar, admirar_humildade*
14	*furia_odio_vergonha*
15	*satisfeito, feliz_satisfeito, feliz*
16	*saudade_surpresa*
17	*infeliz_furia_insatisfeito, surpresa, insatisfeito_outra*
18	*desejo_inveja, amor_admirar, admirar*
19	*esperanca*
20	*saudade*

Table 5: Clusters created from the fourth word-embeddings using k-means with $k = 20$.

Although some clusters are difficult to interpret, some interesting details emerge. *Amor* ('love') and *Odio* ('hate') are in the same cluster (1), although why *Inveja* ('envy') is with them is less clear. The same happens for *Orgulho* ('pride') and *Vergonha* ('shame') in cluster 5, also accompanied by others. Cluster 11 correctly joins all classes which have absence of emotion, although it also comprises the group *Furia_Outra* ('anger_other'). One would expect that with more required clusters *Furia_Outra* ('anger_other') would move to another cluster. Cluster 15 links *Satisfeito* ('satisfaction') and *Feliz* ('happiness') (recall that this was one of the cases where we wondered whether merging made sense), and cluster 2 is about *Desejo* ('desire'). Cluster 6 showed a spelling error in *Insastifeito*, which is not a group, it occurs most probably as the annotation of one word only, which was joined with *Outra* ('other'). Finally, cluster 13 uncovers yet another problem of the original annotation, namely the fact that there are cases marked *Humildade_Admirar* and cases marked *Admirar_Humildade* in the material, while there should be only one way of encoding this group. This cluster also shows that clustering was useful.

After this clustering attempt, we tried using fuzzy clustering, following Atakishiyev and Reformat (2020), who have also applied it to word embeddings. Fuzzy clustering is an extension of clustering in which each data point can belong to more than one cluster. Membership is then a likelihood. This was motivated by the following property of the corpus annotation: we allowed more than one emotion group to be assigned to a particular word in context. These cases seem to require a different kind of clustering, where membership is not crisp, but partial. However, we can only report negative results in that respect. Our preliminary attempts did not show any improvement over k-means.

5. Related work

We report briefly on three different alleys in related work: looking at emotions in text and not only polarity; clustering emotions; and interpreting results with word embeddings.

Parallel to sentiment analysis, there is a (much smaller) research alley that pursues what has been called "emotion annotation" or "emotion detection", represented by Maia (1994), Aman and Szpakowicz (2007), and Ptaszynski et al. (2014). For a review, see Seyeditabari et al. (2018). There has also been some work on precisely the literary domain to identify emotions, examples of which are Mohammad (2012) and Kim et al. (2017). Our work is definitely included in this tradition. Compared to the above works, we are using considerably larger amounts of text.

As to clustering emotions, that is, using empirical methods to identify how emotion is structured in a language, we are only aware of a few works: Feng et al. (2011) use a sentiment lexicon in order to cluster blogs, and extract the most common words in each of their (eight) clusters; Hu et al. (2009) classify Chinese lyrics by their emotions, using fuzzy clustering. Again, our work addresses far more textual data and more text genres.

The work closest to ours, Tang et al. (2014), creates what they call sentiment-specific word embeddings (SSWE). But, as the name indicates, they use "sentiment" (positive or negative) and not emotion. They also use the top-100 closest words in their word embeddings to evaluate the polarity consistency of different sentiment lexicons. Other researchers trying to adapt word embeddings to emotion or sentiment use techniques like retro-fitting or counter-fitting emotion lexicons onto the "ordinary" word embeddings, producing other kinds of models, see e.g. Speer and Chin (2019). Although these works are valuable to suggest new alleys for using word embeddings in the study of emotions, we believe our different attempts are also worth pursuing, and enrich this relatively young area.

6. Concluding remarks

These exploratory experiments only scratch the surface of what can be done having such a rich resource at our fingertips, the purpose of which is to provide a wide picture of the reference to emotions in Portuguese. We tried to explore this resource with several big data techniques, but we hope that this is just the beginning of a new research area in the years to come, especially because we have made the resources and the methods we employed in the explorations described here publicly available,[4] so we hope to see others follow suit.

There are a number of experiments we still wish to perform, from merely testing different word embeddings approaches like FastText or GloVe, experimenting with different preprocessing strategies, adding other annotation information (part-of-speech, functional dependency, morphological or other semantic properties), investigating text genre, time period, and so on.

Also, there are other approaches in the word embeddings world that seem worth trying, from box embeddings (Abboud et al., 2020) to comparison with other public word embeddings for Portuguese, as well as the creation of word embeddings per genre, which has often been proposed in the literature (Tshitoyan et al., 2019).

Likewise, we only looked at the emotion groups *Amor* ('love') and *Desespero* ('despair'). More than twenty further groups, as well as their possible mergings (for example, by joining together antonyms), could and should be investigated in order to learn more about emotion in Portuguese, and to evaluate thoroughly the particular annotation available.

Concerning the three initial motivating questions, our preliminary conclusions based on the available material are as follows:

- There is evidence to merge *Feliz* ('happiness') and *Satisfeito* ('satisfaction'), from clustering word embeddings.

[4] From https://www.linguateca.pt/documentacao/artigoClusteringEmotions.html.

- The *DESESPERO* (‘despair’) group is apparently extremely wide, since it displays connections with many different emotions, which may mean that it joins disparate things and should be divided.
- There seems to be no reason to separate friendship and love. They seem to belong to the same group, called *AMOR* (‘love’).

As one reviewer noted, there are no predefined predictions from previous studies of what emotions groups might look like in Portuguese, so we cannot contrast our groups with others in the literature.

Finally, and concerning the limitations of our approach to comparative work, clearly we are only discussing Portuguese here. Similar studies have to be done for other languages before one can contrast Portuguese with them. It is noteworthy to emphasize that we, with Wierzbicka (1999) and many others (Jackson et al., 2019), do not believe emotion concepts work similarly in different languages.[5] Otherwise, there would be no point in doing emotion studies in Portuguese, given that there are more resources in general for English linguistics.

[5] Although Wierzbicka assumes a universal framework of semantic primitives, she takes special care to explain that (most) emotions are culture-specific, albeit built from common primitives, see also her work on pain (Goddard & Wierzbicka, 1994).

Acknowledgements

We are grateful to *Fundação Científica para a Computação Nacional*, FCCN, for maintaining Linguateca's servers, and to NRIS – Norwegian research infrastructure services for access to the saga cluster in Norway. We thank the anonymous reviewers of our paper for excellent feedback and editorial help, and we thank every researcher at *Linguateca* which contributed to the resources used here. A special thanks goes to Cristina Mota, with whom we discussed almost every line of this paper, and who contributed therefore significantly to its final form.

Competing Interests

The authors have no competing interests to declare.

References

Abboud, R., Ceylan, I. I., Lukasiewicz, T., & **Salvator, T.** (2020). BoxE: A Box Embedding Model for Knowledge Base Completion. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (pp. 9649–9661). Vancouver, Canada: Curran Associates Inc. DOI: https://doi.org/10.48550/arXiv.2007.06267

Aman, S., & **Szpakowicz, S.** (2007). Identifying Expressions of Emotion in Text. In V. Matousek & P. Mautner (Eds.), *TSD 2007: Text, Speech and Dialogue* (pp. 196–205), Springer. DOI: https://doi.org/10.1007/978-3-540-74628-7_27

Antoniak, M., & **Mimno, D.** (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119. DOI: https://doi.org/10.1162/tacl_a_00008

Atakishiyev, S., & **Reformat, M. Z.** (2020). Analysis of Word Embeddings Using Fuzzy Clustering. In S. Shahbazova, J. Kacprzyk, V. Balas, & V. Kreinovich (Eds.), *Recent Developments and the New Direction in Soft-Computing Foundations and Applications. Studies in Fuzziness and Soft Computing* (pp. 539–551). Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-47124-8_44

Barrett, L. F., Lewis, M., & **Haviland-Jones, J. M.** (Eds.) (2018). *Handbook of Emotions: Fourth Edition*. Guilford Press.

Batista, D. S. (2019). Portuguese Word Embeddings. http://www.davidsbatista.net/blog/2019/11/03/Portuguese-Embeddings/

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Bick, E. (2007). Automatic Semantic Role Annotation for Portuguese. *TIL, V Workshop em Tecnologia da Informação e da Linguagem Humana,* 1715–1719.

Bick, E. (2014). PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. In T. B. Sardinha, & T. L. S. B. Ferreira (Eds.), *Working with Portuguese Corpora* (pp. 279–302). London/New York: Bloomsbury Academic.

Birjali, M., Kasri, M., & **Beni-Hssane, A.** (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *26*(107134). DOI: https://doi.org/10.1016/j.knosys.2021.107134

Boddice, R. (2018). *The history of emotions*. Manchester: Manchester University Press.

Bojanowski, P., Grave, E., Joulin, A., & **Mikolov, T.** (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5,* 135–146. DOI: https://doi.org/10.1162/tacl_a_00051

Csardi, G., & **Nepusz, T.** (2006). The igraph software package for complex network research. *InterJournal, Complex Systems, 1695*(5), 1–9. http://igraph.org

Enfield, N. J., & **Wierzbicka, A.** (2002). Introduction: The body in description of emotion. *Pragmatics & Cognition*, *10*(1/2), 1–25. DOI: https://doi.org/10.1075/pc.10.1-2.02enf

Feng, S., Wang, D., Yu, G., Gao, W., & **Wong, K.-F.** (2011). Extracting common emotions from blogs based on fine-grained sentiment clustering. *Knowledge Information Systems, 27*, 281–302. DOI: https://doi.org/10.1007/s10115-010-0325-9

Freitas, C., Santos, D., Mota, C., Carriço, B., & **Jansen, H.** (2015). O léxico do corpo e anotação de sentidos em grandes corpora: o projeto Esqueleto [The lexicon of the body and sense annotation on large corpora]. *Revista de Estudos da Linguagem, 23*(3), 641–680. DOI: https://doi.org/10.17851/2237-2083.23.3.641-680

Gensim: Topic modelling for humans. https://radimrehurek.com/gensim/ [last visit: 5 October 20

Goddard, C., & **Wierzbicka, A.** (1994). Pain: is it a human universal? In C. Goddard & A. Wierzbicka (Eds.), *Semantic and Lexical Universals* (pp. 127–155). John Benjamin Publishing. DOI: https://doi.org/10.1075/slcs.25

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & **Mikolov, T.** (2018). Learning Word Vectors for 157 Languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 3483–3487. https://aclanthology.org/L18-1550

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., & **Aluísio, S. M.** (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. *Proceedings of 11th Brazilian Symposium in Information and Human Language Technology*, 122–131. https://aclanthology.org/W17-6615/

Hu, Y., Chen, X., & **Yang, D.** (2009). Lyric-based song emotion detection with afective lexicon and fuzzy ckustering method. *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 123–128.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M, Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D. & **Lindqvist, K. A.** (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366* (6472), 1517–1522. DOI: https://doi.org/10.1126/science.aaw8160

Justeson, J. S., & **Katz, S. M.** (1992). Redefining Antonymy: The Textual Structure of a Semantic Relation. *Literary and Linguistic Computing, 7*(3), 176–184. DOI: https://doi.org/10.1093/llc/7.3.176

Kim, E., Padó, S., & **Klinger, R.** (2017). Investigating the Relationship between Literary Genres and Emotional Plot Development. *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature,* 17–26. https://aclanthology.org/W17-2203/. DOI: https://doi.org/10.18653/v1/W17-2203

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. DOI: https://doi.org/10.1109/TIT.1982.1056489

Maia, B. (1994). A Contribution to the Study of the language of Emotion in English and Portuguese. Porto: Faculdade de Letras da Universidade do Porto. Revised version: 1996 http://web.letras.up.pt/bhsmaia/belinda/pubs/thesis.htm

Maia, B., & **Santos, D.** (2018). Language, emotion, and the emotions: The multidisciplinary and linguistic background. *Language and Linguistics compass, 12*(5). DOI: https://doi.org/10.1111/lnc3.12280

Mihalcea, R. (2021, 19 May). *The Ups and Downs of Word Embeddings.* 2021. https://www.youtube.com/watch?v=33XtLnPDOC0

Mikolov, T., Chen, K., Corrado, G., & **Dean, J.** (2013). Efficient Estimation of Word Representations in Vector Space. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Currant Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, *53*(4), 730–741. DOI: https://doi.org/10.1016/j.dss.2012.05.030

Mota, C., & **Santos, D.** (2015). *Emotions in natural language: a broad-coverage perspective*. http://www.linguateca.pt/acesso/EmotionsBC.pdf

Pang, B., & **Lee, L.** (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. DOI: https://doi.org/10.1561/1500000011

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, M. B., Perrot, M., & **Duchesnay, E.** (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Pennington, J., Socher, R., & **Manning, C.** (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014,* 1532–1543. DOI: https://doi.org/10.3115/v1/D14-1162

Ptaszynski, M., Rzepka, R., Araki, K., & **Momouchi, Y.** (2014). Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. *Computer Speech and Language*, *28*, 38–55. DOI: https://doi.org/10.1016/j.csl.2013.04.010

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org/

Ramos, B. C. (2021). Descrição de uma metodologia desenvolvida para revisão de um léxico de palavras de emoção [Description of a methodology developed to revise an emotion lexicon]. *Jornadas de Descrição do Português, STIL 2021* (pp. 389–397). DOI: https://doi.org/10.5753/stil.2021.17819

Ramos, B. C., & **Freitas, C.** (2019). "Sentimento de quê?" uma lista de sentimentos para a Análise de Sentimentos [Feeling of what? A list of feelings for emotion analysis]. *STIL – Symposium in Information and Human Language Technology,* Salvador, BA, 38–47.

Ramos, B., Santos, D., & **Freitas, C.** (2020). Looking at body expressions to enrich emotion clusters. In M. J. B. Finatto, S. Luz, S. Pollak, & R. Vieira (Eds.), *Proceedings of the Digital Humanities and Natural Language Processing Workshop at the 14th International Conference on the Computational Processing of Portuguese Language* (pp. 57–62). http://hdl.handle.net/10400.26/35280

Rehurek, R., & **Sojka, P.** (2010, May). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45–50. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf

Rodrigues, J., & **Branco, A.** (2018). Finely Tuned, 2 Billion Token Based Word Embeddings for Portuguese. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2403–2409. https://aclanthology.org/L18-1382.pdf

Romanov, V., & **Khusainova, A.** (2021). Evaluation of Morphological Embeddings for English and Russian Languages. *NLPIR 2019: Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval,* 144–148. DOI: https://doi.org/10.1145/3342827.3342846

Santos, D. (2014). Corpora at Linguateca: Vision and roads taken. In T. B. Sardinha, & T. L. S. B. Ferreira (Eds.), *Working with Portuguese Corpora* (pp. 219–236). Bloomsbury.

Santos, D. (2016). Comparando corpos orais (transcritos) e escritos na Gramateca [Comparing (transcribed) oral corpora with written corpora in Gramateca]. In C. Bardel & A. De Meo, *Parler les langues romanes/Parlare le lingue romanze/Hablar las lenguas romances/Falando línguas românicas. Atti del Convegno Internazionale GSCP 2014* (pp. 127–142). Napoli: Università di Napoli L'Orientale, Il Torcoliere.

Santos, D. (2021). Natural and artificial intelligence; natural and artificial language. In R. Queirós, M. Pinto, A. Simões, F. Portela, & M. J. Pereira (Eds.), *10th Symposium on Languages, Applications and Technologies (SLATE 2021)* (pp. 1:1–1:11), OASIcs – OpenAccess Series in Informatics (vol. 94). DOI: https://doi.org/10.4230/OASIcs.SLATE.2021.1

Santos, D., & **Maia, B.** (2018). Language, emotion, and the emotions: A computational introduction. *Language and Linguistics compass, 12*(6). DOI: https://doi.org/10.1111/lnc3.12279

Santos, D., & **Mota, C.** (2010). Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010),* 1437–1444. http://www.lrec-conf.org/proceedings/lrec2010/pdf/457_Paper.pdf

Santos, D., & **Mota, C.** (2015). A admiração à luz dos corpos [Admiration illuminated by corpora]. In A. Simões, A. Barreiro, D. Santos, R. Sousa-Silva, & S. E.O. Tagnin (Eds.), *Linguística, Informática e Tradução: Mundos que se Cruzam. Homenagem a Belinda Maia, Oslo Studies in Language,*

7(1), 57–77. https://journals.uio.no/public/journals/1/images/osla-7-1.pdf. DOI: https://doi.org/10.5617/osla.1466

Santos, D., Simões, A., & **Mota, C.** (2022). Broad coverage emotion annotation, *Language Resources and Evaluation, 56,* 857–879. DOI: https://doi.org/10.1007/s10579-021-09565-1

Seyeditabari, A., Tabari, N., & **Zadrozny, W.** (2018). *Emotion Detection in Text: A Review*. https://arxiv.org/pdf/1806.00674.pdf

Speer, R., & **Chin, J.** (2019). *An Ensemble Method to Produce High-Quality Word Embeddings*. https://arxiv.org/pdf/1604.01692.pdf

Stack Exchange. (2018). *K-means clustering of word embedding gives strange results*. https://bit.ly/[illegible]mWEM

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & **Qin, B.** (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1555–1565. https://aclanthology.org/P14-1146.pdf. DOI: https://doi.org/10.3115/v1/P14-1146

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & **Jain, A.** (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*, 95–98. DOI: https://doi.org/10.1038/s41586-019-1335-8

Wierzbicka, A. (1999). *Emotions across Languages and Cultures: Diversity and Universals*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511521256

Yamada, I., Shindo, H., Takeda, H., & **Takefuji, Y.** (2016). Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, 250–259. DOI: https://doi.org/10.18653/v1/K16-1025

Typesetting queries

1. The following items have been included within the reference list, but are not cited within the text. For each un-cited reference, please advise where it should be cited in the text, or confirm that it can be removed from the reference list.
    a. Ref. " Gensim: Topic modelling for humans. "
    b. Ref. " Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., & Aluísio, S. M. (2017). "
    c. Ref. " Santos, D., Simões, A., & Mota, C. (2022). "
    d. Ref. " Stack Exchange. (2018). "