Lang Resources & Evaluation (2022) 56:857–879
https://doi.org/10.1007/s10579-021-09565-1

ORIGINAL PAPER

Broad coverage emotion annotation

Diana Santos[1] · **Alberto Simões**[2] · **Cristina Mota**[3]

Accepted: 30 September 2021 / Published online: 22 October 2021

Abstract In this paper we present the emotion annotation of 1.5 billion words Portuguese corpora, publicly available. We motivate the annotation process and detail the decisions made. The resource is evaluated, being applied to different areas: to study Lusophone literature, to obtain paraphrases, and to do genre comparison.

Keywords Emotions · Portuguese · Annotated corpora

1 Introduction

Although there has been a lot of interest in emotions and the detection of attitudes in text, giving the rise of the rich and exciting subarea of "sentiment analysis", there has been little work in describing emotions in different languages. In fact, usually the papers concerned with sentiment analysis in languages other than English start by explaining how they adopted or translated English emotion theories or lexicon. Some examples are (Grefenstette et al. 2006) for French or (Kajava et al. 2020) for Finnish and Italian.

The approach here described for the Portuguese language is radically different. It is inspired by the work of Wierzbicka (1999), who insists on looking at each

✉ Diana Santos
d.s.m.santos@ilos.uio.no

Alberto Simões
ambs@zbr.pt

[1] Linguateca and University of Oslo, Oslo, Norway

[2] Linguateca and 2Ai- School of Technology, IPCA, Barcelos, Portugal

[3] Linguateca and INESC-ID, Oslo, Norway

language by itself and whom claims that emotions and related words are language specific — see also the discussion of pain in English and its "equivalents" in other languages (Goddard and Wierzbicka 1994).

However, we did not start from scratch. Given the existence of a contrastive thesis, by Maia (1994), comparing Portuguese and English, which already had suggested a set of major emotion categories for Portuguese following Ortony et al.'s emotion framework (Ortony et al. 1988), we tried to only use Portuguese data (corpora, ontologies, dictionaries, and our native speaker intuitions) to create lexicons and emotion clusters that made sense for Portuguese, as a continuation of the work we have been doing on Portuguese corpora in the AC/DC project (Santos and Bick 2000) for colour, clothing, body, speech verbs, family and health. See Santos (2014a) for a little outdated history of this process, and Higuchi et al. (2019) for family and Santos (2019) for health annotation processes.

But we avow that emotion annotation poses harder problems, given that:

- There is no consensual set of emotion categories (neither for Portuguese nor for any other language), as has been argued by Santos and Maia (2018); Maia and Santos (2018).
- Since emotions are abstract, the verbs that describe them are often used in a concrete (and non emotional) sense (e.g. *reconhecer* (recognize) or *confiar* (trust), or *miss* in English).
- It is not easy to delimitate emotion from physical feelings (you can *sentir* (feel) both warmth and love, *dor* (pain) can be moral or physical) or from ethical valuations, the latter because very often (ethical) judgement causes the feeling: *vergonha* (shame) or *admirar* (admire) do presuppose that one judges the objects as respectively bad and good.
- Verbs may refer to emotions of several of their arguments, not only the subject: *enfurecer* (infuriate) or *humilhar* (humiliate) refer to the emotion of the object, while *troçar* (make fun of) can describe both disdain by the subject and humiliation of the object.
- Because emotions are personal and one has not access to other people's minds, there is a set of conventional actions, poses, and gestures that convey or demonstrate an emotion, but not always unambiguously. Examples are *chorar* (weep), *franzir o sobrolho* (frown) and *bater palmas* (clap).
- There are changes in "emotional colour" with time, and there are different strategies/interpretations in distinct varieties of Portuguese, something we have to be careful about in our analyses. For example, *amar* is much more widespread, while *gozar* is much more specific, in Brazilian Portuguese as compared to Portuguese from Portugal. Likewise, *bravo* meaning angry and *ilibado* meaning morally clean in Brazilian Portuguese (almost) do not exist in the language of Portugal, which means that a few annotation rules have to be variant-specific.

Although in the end the human interpretation is always necessary, no matter which semantic domain we annotate — something that is growingly recognized in corpus linguistics, see e.g. Lüdeling (2011) —, we believe that for emotions there is a much wider room for disagreement, as will be expounded in the next sections.

In addition, and as is the case for any annotation process, we cannot get it perfect by automatic means, and we need to create more specific rules for almost any emotion-denoting word in our emotion lexicon, following the model "human in the loop" advocated in Santos and Mota (2010). So, the annotation we present here is not fully revised, but is being continuously improved as time goes by. (Evaluation of the current state is described in Section 5)

It should also be mentioned that both the lexicon and the rules employed are publicly available, for transparency, and we strive for constantly improving documentation of the material.[1]

We present in the following sections more information about our goals and decisions for the annotation process.

2 Emotions in text

We start by insisting that annotation depends on the research goals and, although we offer this resource to the community in the hope that it is useful for better understanding of Portuguese and of emotion in general, we cannot claim or expect that all researchers share our opinions. We actually hope that, by clarifying the way we reasoned and proceeded, we allow other researchers with different opinions to make use of the material by filtering out or adding information.

2.1 Whose emotions?

We adhere to the tripartite distinction suggested by (Paltoglou et al. 2010, page 30), namely:

- Emotion content of the text
- Author's emotional state
- Intended reader's likely emotional state

All three kinds of annotations (and corresponding research) are surely relevant, but not necessarily similar, something which is especially noteworthy in literary texts, where one may contend that authors play with their readers and characters so that they induce different emotions than those they describe or feel.

The annotation we were interested in (possibly as a first step) is the emotions referred to in the text, that is, what we call the text's emotional content, and not those emotions elicited or felt by the writer, as is the rule in the field of sentiment analysis (Pang and Lee 2008), which in addition is often only concerned with detecting polarity (positive, negative or neutral).

Another relevant clarification is that we are concerned with reference to emotion, not the expression of it. So, putative expressions like *Oh my God!* or a four letter

[1] See https://www.linguateca.pt/Gramateca/Emocionario.html.

word, although clearly expressing a strong outburst of emotion, do not refer to it; while *Ai que medo!* (What a fear!), *desdenhou* (showed contempt) or *virou as costas* (turned his back) do.[2]

2.2 Which emotions?

As to which emotions to annotate, there are widely different emotion palettes (and corresponding emotion lexicons) that have been used.

There are large inventories (Grefenstette et al. 2006) as well as sets of a few basic emotions, such as Plutchik's eight (Plutchik 1991) or Ekman's six Ekman (1984) emotions (see e.g. Mohammad and Turney (2012); Strapparava and Mihalcea (2007)).

In our view, the only way to decide is to look to the language, as the empirical data that gives us grounds on which to choose (e.g., see Ramos and Freitas (2019)).

Emotions can, in addition, have a set of different attributes (such as valence, arousal and control, as prescribed in the affective norms suggested for English (Bradley and Lang 1999), and widely adapted into other languages, see e.g. for two varieties of Portuguese, both incidentally translated from English, see (Soares et al. 2012; Kristensen et al. 2011)). We can also claim the necessity of a different annotation tag for each emotion, making thus e.g. *zangado* (angry) something different from *irritado* (irritated), or *misery* different from a high amount of *sadness*. Our annotation clusters provide the first approach, but users can choose the second by simply querying emotions separately, or even joining them differently.

Furthermore, some researchers consider both simple emotions and more complex ones, which merge sets of simple ones, providing a hierarchical classification. *Respect* can thus be seen as a composite of *admiration* and *fear*, or an emotion in its own, even if one often feels either fear or admiration together with respect.

How can we deal with emotions that hard to distinguish (or clearly related), such as hope, faith and trust (*fé, esperança, confiança*), joy, contentedness and happiness (*alegria, contentamento, felicidade*), and jealousy and envy (*ciúmes, inveja*)? Should we attempt a single cluster for each, or separate them in two? This is ultimately a question of choice, although work is in progress to evaluate this kind of decisions.

Our clusters are presented in the next section, and there is a strong emphasis on documenting all choices taken. Basically, we chose to create only clusters which had large lexical diversity. So, in the three cases mentioned above we did join them in three clusters. But we note, again, that it is possible to distinguish them by listing each new group separately in the search interface.

It is also important to mention that there are emotions without clusters, technically in a group called OUTRA (other). This group contains all cases that, so far, have not been assigned to a specific cluster.

We also considered, just like for colour or clothes, that absence of emotion still concerns the semantic field of emotion (the cluster AUSÊNCIA (absence)), just like

[2] This last because we considered conventional bodily expressions as emotion descriptions. See Section 5.5 for more on this subject.

incolor (colourless) and *nu* (naked) belong to colour and clothing annotation, respectively.

2.3 What is an emotion?

Finally, what counts as an emotion? This is probably the most difficult question, and it is why we discuss it in the end.

The initial process was documented in Mota and Santos (2015), where we describe how we made use of synonyms, antonyms and hypernyms to gather a large set of emotion words, using two lexical ontologies for Portuguese (Maziero et al. 2008 and Gonçalo Oliveira et al. (2010)). This means we relied on lexicographical tradition and other researchers' intuitions about our language, which we then filtered out when in disagreement.

We used several empirical tests as well, especially for cases where there is theoretical discussion whether some concept is or not an emotion: the idea was to let the language decide. If the Portuguese language treated it as an emotion, we classified it as one, despite philosophical arguments to the contrary. Typical examples which may not be consensual are *coragem* (courage), *curiosidade* (curiosity), and *amizade* (friendship), which are fine in Portuguese in phrases like *senti uma grande...* (I felt large *X*), *um sentimento de...* (a feeling of *X*), *enchi-me de...* (I filled with *X*), *estou cheia de...* (I am full of *X*), etc.

These tests were often performed relying on our native speakers' competence; in other cases we used corpora, and interpreted the results.

One particular case we included in the emotion lexicon was what Strapparava and Mihalcea (2007) call "indirect affective words", like *inimigo* ('enemy'). This word denotes a person toward which one has a strong negative feeling, even if it can be used in a neutral context, such as *X and their enemies made peace*. But given that hate is a key ingredient of a word like *inimigo*, we have included these words as referring to emotions for now, also because they are often related to other words which refer more clearly to emotions, like *amante* (lover) and *amado* (loved one), derived from *amar* (to love).

On the other hand, and as already mentioned, we are not interested in "emotion-laden words" as defined by Pavlenko (2008). Even though they strongly inspire a negative emotion, we do not consider words like *cancer* or *murder* as emotional, as for example Subasic and Huettner (2001) did.

Finally, we believe that one of the most essential properties of natural language is vagueness, and therefore we expect words to be often vague between several related meanings. This is dealt with, in our annotation, by allowing a particular instance in context to receive more than one emotional interpretation, or also non-emotional interpretations. We hope that these cases will help to motivate either conflation or division of groups in a diachronic study, as well as to provide clues for further division or merging of the current clusters, something we are currently working on.

Table 1 Size of the emotion clusters, both in number of tokens, and in distinct lemmas, in general. If a case is vague between two emotions, it is counted twice

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		

3 The annotation scheme

The used annotation is simple and consistent with all other annotation projects at Linguateca:

- Any word considered to refer to an emotion is marked as `emo:GROUP`, being the group one of those in Table 1[3]
- If the word is considered to belong to several groups, it is marked with them all, separated by an underscore.
- When one emotion is expressed in several words we annotate just one: in general the most specific. So, expressions like *sentir saudades* (feel a longing), where one might mark both *sentir* and *saudades* as respectively `emo:gen` and `emo:saudade`, received only the annotation for the last word.
- For multi-word expressions where only one emotional word is present, we might change its emotion group, accordingly with the expression meaning. As an example, *sentir a falta* (literally, feel the lack, meaning “to miss”), where *sentir* is annotated as `emo:saudade`.
- Finally, in the cases where an expression refers to an emotion, but none of the words is originally an emotion (and would therefore not receive emotional annotation alone), we mark it on the first word of the expression, like in *franzir as sobrancelhas* (raise highbrows).

[3] For readability, we added diacritics to the group names, that are not present in the corpus. For instance, the corpus uses `odio` instead of *ódio*.

```
 LIST MEDO = "acobardar"
"acovardado"
"acovardamento"
"acovardar"
 (...)
ADD (%EMO:MEDO) TARGET MEDO;
ADD (%EMO:SURPRESA) TARGET ("choque") (-1 ("um")) (-2 (APANHAR));
ADD (%EMO:ESPERANCA) TARGET ("voto") (-1 ("fazer"));
ADD (%EMO:INFELIZ) TARGET ("pobre") (0 (@>N));
```

Fig. 1 Examples of VISL rules: lexical rules; conditional lexical rules *apanhar um choque* (get a shock) implies that *choque* is marked as surpresa (surprise), *fazer votos* adds esperança (hope) to *votos*, and *pobre* (poor) is considered infeliz (unhappiness) if it precedes the noun

We may, at a later stage, apply to emotions the (more complicated) scheme with the `mwe` label, which we use for body annotation (Freitas et al. 2015), which would assign possibly distinct emotion classifications to a multi-word expression and its constituents, but so far we did not deem it relevant.

As to the process we follow, we provide here some more detail, in order for the reader to appreciate the full workflow and understand how improvements were performed. The first step is the application of VISLCG3 (Bick and Didriksen 2015) rules that take care of the bulk of the annotation (see Fig. 1). They perform simple lexical annotations.

The second step, for a more linguist-friendly rule specification, is done with `corte-e-costura` (Santos and Mota 2010), which is basically a set of arbitrarily complex correction rules in a CQP-like format (Evert and Team 2021), in five phases, which can be tailored per sub-corpus. Some simple examples are shown in Fig. 2.[4]

Further examples of the use of both systems, together with detailed information about the initial (2013) emotion annotation are available from (Mota 2013). Moreover, all rules are public, even for the corpora which are only available for querying.

4 Quantitative description

In our emotion lexicon we have currently 26 clusters, corresponding to around 2800 different lemmas, and 8000 different word forms, a rough description of which is shown in Table 1, documenting how many cases have been detected of each emotion group in our material.

The names of the emotion groups can be thought as shortcuts to a much longer description. At first they can be confusing: for example, the group amor (lit. love) also includes the word *amizade* (friendship); and the group desespero (lit. despair) also includes the word *desconfiança* (mistrust). However, we still cling to these names given that a full list of terms would be impractical, and the use of numbers or

[4] For fairness' sake, we should avow that we devised `corte-e-costura` before having access to the VISLCG3 parser, otherwise we might have decided to do everything in VISLCG3. But when we started with emotion annotation, we had already a lot of `corte-e-costura` rules for other semantic domains, and had trained several people who were familiar with this setup.

```
# general negative rules
[lema="divisão"] a:[lema="acanhado"] >> a:[sema="0"]
# if the lemma "acanhado" (shy) follows the word "divisão" (room), it does not
# describe an emotion, but a size
# general positive rules
[lema="ter"] [word="muita|tanta|imensa"] a:[word="pena"] >> a:[sema="emo:pena"]
# when pena is preceded by the verb "ter" (have) followed by a quantifier, it
# means "sorrow"
# general specialization rules
[lema="admirar\+el[ea]s*"] >> [sema="emo:admirar"]
# when the clitic of the verb "admirar" (which can mean surprise or admire) is
# a 3rd person object pronoun, the emotion is ADMIRAR (admiration)
[lema="ser"] [pos="ADV.*"] a:[lema="admirar" & temcagr=".*PASSIVA.*"] >>
a:[sema="emo:admirar"]
# and the same happens when "admirar" is in the passive form with "ser".
```

Fig. 2 Examples of corte-e-costura rules: the comments were added for the readers' sake

letters (e.g. group 352, or group A) would be even less intuitive. But be warned: to interpret the results you have to look at the group elements.

The cluster GEN(érica) (generic) contains all words that name emotion without specifying which one, like *sentimento* (feeling) or *emoção* (emotion) itself.

It is important to stress that the vast majority of the cases have not been human revised, so we believe there is gross over-generation, especially for the generic group. For example both the word *sentir* and the word *raiva* could have been annotated in the multi-word expression *sentir raiva* (feel anger), instead of only *raiva* be marked.[5]

The one with the highest number of references is DESEJO (desire, wish) closely followed by AMOR (love), whereas AMOR (love) is the one with the highest lexical diversity, closely followed by FELIZ (happiness) and VERGONHA (shame).

This is not, however, a reliable quantitative estimate of the emotions in text, for several reasons.

4.1 Contextual issues

The first and most important one is that the words themselves do not always refer to an emotion, depending on its context. For example, *reconhecimento* often expresses the emotion of gratitude, but can also be used as a neutral term referring to remembering people, examining a piece of land, etc. Although we developed (and continue developing) rules that try to separate the different cases, it is unrealistic to expect, or promise, that we will have this perfect for billions of words.

To partially address this problem, we defined the **degree of lexical emotionality** of a word as the fraction of times it is used with an emotional meaning. Note that this measure does not represent how emotionally strong the word is when employed in its emotional meaning. A similar measure could be proposed for degree of lexical colourness, lexical bodiness, etc. Take the words *pena* and *fã*:

[5] Obviously, we devised rules to avoid this, as mentioned in the previous section, but we cannot yet solve complex cases like *sentiu, como de costume, uma raiva surda...* (s/he felt as usual a cold anger...).

- The first means sorrow, but also feather, pen (when people used feathers to write but also nowadays metaphorically), and punishment. So, its degree of lexical emotionality is quite low (current estimate in literary text: 13%);
- The second only means supporter or admirer, and so its emotionality degree is 100%.

Sometimes a distinction can be done by grammatical category alone: the word *pesar* as a noun means sorrow, while as a verb it means to weigh.[6]

Although these properties have to be estimated from humanly revised annotated corpora, and may change with time and genre, we believe they can be genuinely relevant indicators for estimating emotional content. They can be obtained directly from our resource, using the corpora where human revision has taken place, or estimated by users asking for random samples and classifying them.

In fact, we are also creating a lexicon of emotion words marked with their degree of emotionality, so that these numbers can be used, for example, to automatically correct purely lexically-based estimates of emotionality.[7] See Table 2 for some examples.

For the case of *dor*, where two numbers are presented, they correspond to cases unambiguously emotional and to vague cases, between emotional and physical pain.

4.2 Reference as opposed to statement

Secondly, note that the use of a word belonging to a particular emotion group does not necessarily describe that emotion in that particular context. After all, we are not doing sentiment analysis. So, for example, in negated contexts, the word may as well convey the opposite emotion group. In fact, some emotion groups are much more likely than others to be referred to negatively, as documented in Maia and Santos (2012), which is yet another characteristic worth looking into (and, incidentally, which can only be appreciated if we do annotate the initial, non-negated, reference).

To arrive at a better estimate of "real" presence of an emotion in text, one could select the negated contexts and associate them with the opposite emotion. But this is naïve in two ways:

- First, there is not always one opposite emotion. Some emotions, like SAUDADE (longing) or SURPRESA (surprise), do not have opposites, and others can have more than one: consider ORGULHO (pride), whose opposite can be HUMILDADE (humbleness) or VERGONHA (shame).
- Second, the negation of an emotion does not necessarily imply the opposite even when it obviously exists: for instance, *not particularly like* is not synonymous

[6] In fact, the verb *pesar* can also be used metaphorically, at least in the expression *pesar na consciência* (weigh in conscience), where it denotes remorse, but we considered this a case of a multi-word expression.

[7] This would be mainly for other corpora or texts not served by Linguateca, and therefore not annotated with our rules, but one can also use these estimates to check how far the annotation of a particular corpus has been improved.

Table 2 The degree for lexical emotionality (DLE) for some words, measured in random samples from a particular corpus genre. Some translations into English of the two kinds of meanings are presented

Word	DLE	Source	Emotional	Non-emotional
Pena	0.13	Literary	Pity	Feather, pen
Dor	0.51/0.60	Literary	Sorrow	Pain
fã	1.00	Newspaper	Fan	–
Raiva	0.73	Newspaper	Anger	Rabies
Reconhecimento	0.32	Newspaper	Gratitude	Recognition
Aterrar	0.03	Newspaper	Frighten	Land, ground

with *hate* or even *dislike*. In short, one can negate one emotion without stating another.

To sum up, one should be careful in the use and the conclusions one takes from (our, or any) emotion annotation. With our annotation, we were interested in assessing how often a particular emotion is referred to, independently of whether it is negated or not, and/or whether it is truly meant or sarcastically used. These are different, albeit relevant, concerns, which we believe belong to a next level of annotation and study.

5 Evaluation of the emotion annotation

It is not straightforward to evaluate a large body of annotation work albeit there are few techniques for this:[8]

- One can evaluate the degree of agreement among several annotators (or users) as to the correctness of the annotation result, but it is always a question of a finite, generally little, set of annotators, and for obvious reasons can only be done for few cases and contexts.
- One can evaluate the use of the annotation for specific tasks, called user-visible evaluation by Gaizauskas in Gaizauskas (1998), which is what we will be doing in this paper.

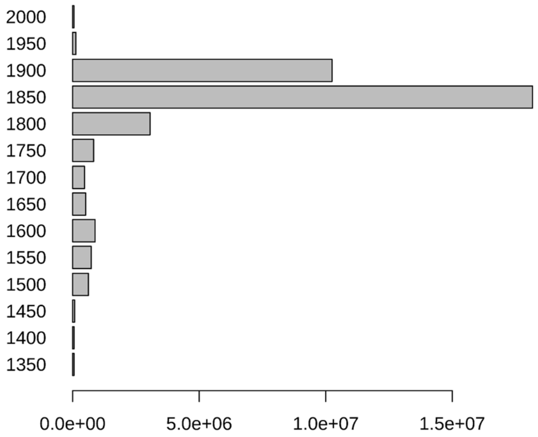
In what follows, we describe several tasks and/or research questions in which we made use of the emotional annotation. But first, we would like to comment on the absence of inter-annotation agreement measures.

[8] One could also compare our lexicon with other sentiment lexicons for Portuguese, but since their goal is quite different from ours, we decided not to present a comparison here.

Table 3 Literateca material by corpus origin

Corpus	Reference	Number of works	Number of million words
Vercial		340	15.8
OBras	Santos et al. (2018)	275	8.2
NOBRE		131	8.4
Tycho Brahe	Galves and Faria (2010)	39	1.8
COLONIA	Zampieri and Becker (2013)	36	1.6
PANTERA	Santos (2019b)	37	0.2

Fig. 3 Distribution of the material (number of words) by half-century



5.1 Inter-annotation agreement?

This is not a traditional human annotation project, where different human annotators painstakingly annotate different texts, and where it is therefore important to assess both whether the guidelines are clear enough and whether the human annotators have the same understanding in context.

Here we have machine annotation, following rules that have been agreed beforehand by a small team, who are very aware, as we hope the discussion above illustrates, that many of the decisions are controversial — or, at least, not consensual, and which are to be refined in the next iterations.

We have therefore chosen not only to make the annotation results but also the lexicon and the rules public, and we keep improving the annotation with more rules and contexts being explicitly observed and cleaned up.

We have done a pilot project of full human revision for the emotion group admirar in two corpora, described in Santos and Mota (2015). But what we are going to present in the rest of the paper are studies based on much larger corpora, mainly not human revised.

5.2 Studying literary language

We have collaboratively created the largest publicly available computational infrastructure for studying literature in Portuguese, Literateca (Santos 2019; Santos and Simúes 2019), which currently (version 7.1, 2th September, 2021) features 857 works and is being actively enhanced and extended.

At present it encompasses ca. 36 million words written by 245 authors — 26 million words from Portugal and 10 million words from Brazil — with linguistic annotation and metadata about each work and author (date, genre, literary school, author gender, and whether it is a translation — for the few cases where known authors have translated works into Portuguese). In addition to canonical works, we include also ca 100 texts in the public domain which have been forgotten by the modern public, see Schöch et al. (2021) for the rationale.

Table 3 shows the origin of the material present in Literateca.

As far as temporal distribution is concerned, see a rough categorization as a barplot in Fig. 3.

As far as type of text is concerned, one can divide the material in a first tripartition of poetry, prose and theater, as indicated in Table 4.[9]

Clearly, we do not aim for a balanced corpus in terms of time and authors, since there were much fewer writers in older times, and there are considerably fewer older texts than modern ones. We have concentrated on public domain works for obvious reasons.

5.3 Browsing literature using emotions

In Santos et al. (2020) we tried to identify whether we could predict literary school (and period) by using a set of semantic and syntactic features, including the number of words of each emotion group per work. Although emotions were just a subset of the features, some of them were found to be specifically discriminative for that task, which indirectly points to the usefulness of the classification.

Redoing the process for the (299) current novels which are classified by literary school, we get the results of Fig. 4, where the emotion groups HUMILDADE (humblenesss), GRATO (gratitude), ALÌVIO (relief) and AMOR (love) are highly discriminative.

5.4 Testing hypotheses about emotions in literature

There are some predictions about the density of reference to emotions in different literary genres, that we can straightforwardly test with our annotation:

- poetry refers to more emotions than prose;
- women refer more to emotion than men;
- romanticism is more prone to describe emotions than realism/naturalism.

[9] Verse included in prose is counted here as prose, while prefaces and author notes in poetry works have been counted as poetry. No distinction has been made between theater in verse or prose.

Table 4 Material by type of text

Type	Texts	Authors	Million tokens
Poetry	100	41	2.5
Theater	89	28	1.7
Prose	668	206	31.8
Total	862	246	36.0

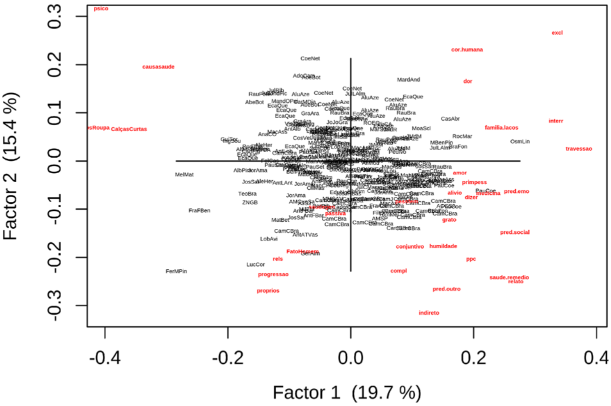


Fig. 4 Correspondence analysis of the novels

Although these are quite sweeping generalizations, and literary studies concern finer-grained questions, if our data can show these trends, the emotion annotation cannot be too much off the mark.

Fig. 5 confirms the two first expectations, namely that poetry and works authored by women show a higher density of emotions in text. However, Fig. 6 goes initially against our expectations, showing symbolism in the top for reference to emotions.

But looking closer at the data, this may be due to the fact that (so far?) we only have four works from that school. If we look at a specific emotion, AMOR (love), more associated with romanticism, we get what we expect, in Fig. 7.

5.5 The human body and the locus of emotions

Another recent work (Ramos et al. 2020) focused on the relationship between body — and expressions involving body — and emotions, attempts to take advantage of a previous annotation of the body semantic domain. The first thing we wanted to appreciate was to whether the emotion grid developed in our emotion annotation project would be appropriate to classify these expressions, or whether other clusters should be proposed.

The results were as follows: it was possible to get 70% of the body expressions in the emotion clusters, while 30% represented other more specific cases — which

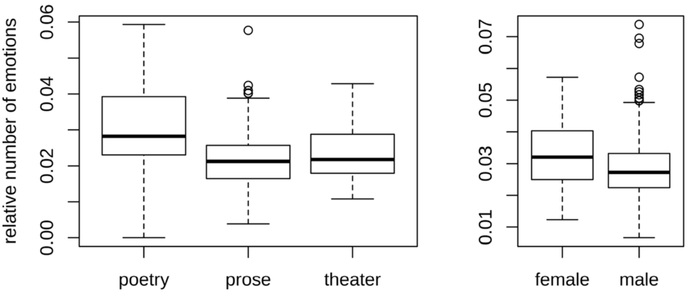


Fig. 5 Relative frequency of emotions per major literary genre, and per author's gender

Fig. 6 Relative frequency of emotions per literary school

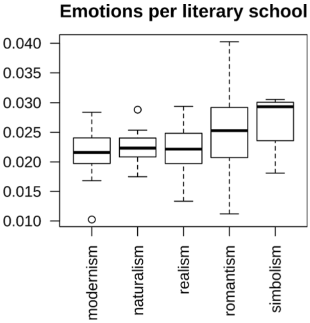
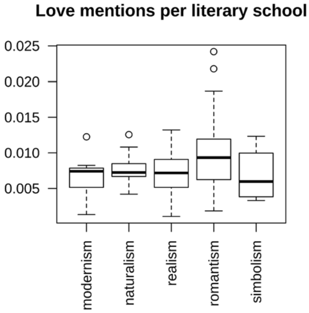


Fig. 7 Relative frequency of love per literary school



could apparently only be expressed with reference to body. As to body expressions related to judgement, only a fourth could be subsumed under the emotion clusters.

This shows the complex relationship between body and emotion and it provided us with more data for the emotion clusters, especially for the OUTRA (other) cluster. The investigation of body together with emotion also made us aware of several conventional ways to describe emotions that were initially neither in body nor emotion repertoires, such as *lágrimas* (tears) for sadness or *riso* (laughter) for happiness. By browsing the literary corpora with the search *lágrimas de* (tears of)

Table 5 Emotions in corpora of different genres (DHBB, Museu da Pessoa, Ciência Viva, CHAVE, Literateca): EmoDens stands for emotional density

Corpus	Size (millions)	EmoDens	Most common emotion(s)
Encyclopedic	14.5	3.8 e-3	fúria (anger), humildade (humbleness)
Interviews	1.5	10.0 e-3	desejo (desire), amor (love)
Science newspaper	0.7	5.4 e-3	desejo (desire), amor (love)
General newspaper	106.1	9.4 e-3	desejo (desire), amor (love)
Literary works	36.0	26.3 e-3	amor (love), desejo (desire)

we found 83 different emotions, both positive and negative[10]; and 34 emotional causes for *rir* (laugh) could be found in the same material.[11]

5.6 Emotions and different types of text

Having a large number of distinct text types in our corpora, we can also look at them in terms of emotional density and content. We have compared five different corpora with respect to: emotion density; and which emotions are more prevalent.

The results can be seen in Table 5. It is especially interesting to note that general newspaper, oral interviews and scientific newspapers share the two most common emotions, and differ widely from literary text (the text type with highest emotion density) and from political history encyclopedia (the text type with lowest emotion density).

5.7 Identifying emotion paraphrases

In Mota et al. (2021), the emotion annotation described here was applied to parallel text in Portuguese (parallel translations into varieties of Portuguese), and it allowed the authors to identify a significant number of paraphrases relative to emotion, as well as highlight different translation strategies and interpretations. This is another application of the annotation described here, although due to copyright restrictions, only the paraphrase emotions, and not the texts, are available.[12]

5.8 Comparing authors for their emotion use

In Santos et al. (2020), we looked at six authors with several works in our corpus, attempting to investigate whether they used the emotion palette differently. The results are graphically presented in Fig. 8.

The most conspicuous finding, in addition to the fact that they had comparable emotion density (as shown in Fig. 9), was that *Raul Brandão* had most mentions to

[10] See https://www.linguateca.pt/Gramateca/Lagrimas.html.

[11] See https://www.linguateca.pt/Gramateca/Riso.html.

[12] See https://www.linguateca.pt/Gramateca/ExemplosParafrasesEmocao.html.

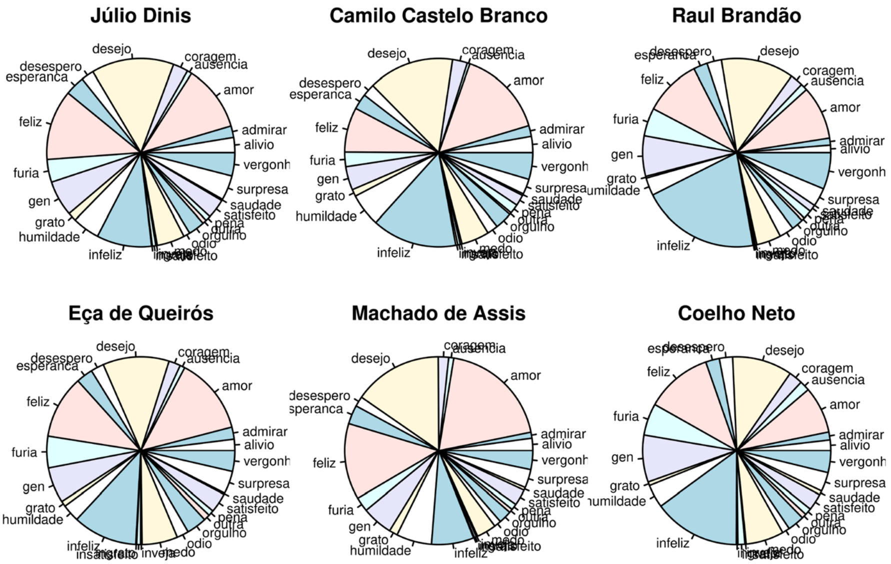


Fig. 8 How authors distribute their emotion reference

unhappiness, and *Machado de Assis* to love and happiness. But all six authors (from the nineteenth or early twentieth century) had a strikingly similar distribution.

5.9 Comparing emotions in translated and original text

Another pilot test we did using PANTERA, a Portuguese-Norwegian parallel corpus (Santos 2019b), was whether the emotions (in Portuguese) in both Portuguese and Norwegian originals had any interesting differences.

Of course, this depends on the particular works that are included in the corpus, so it could not be a definite verdict of the differences between the two cultures, but can be interesting to follow up some of the findings.

The preliminary findings, corresponding to excerpts from 118 works[13], give the following initial picture presented in Fig. 10.

What meets the eye is the relatively larger mention of happiness in Norwegian originals, and the much higher mention in original Portuguese of humildade (humbleness), as well as a slightly higher unhappiness in Portuguese.

As might be expected, generic mentions of emotion are more common in translated text than in original one. This happens for both the generic and the absence category.

In any case, we could not really find striking differences.

A closer scrutiny is, however, necessary, since the authors and texts included in PANTERA are very varied. For example, the fact that there is much more dramatic

[13] The interested readers can look at the detailed corpus contents in the PANTERA website, https://www.linguateca.pt/PANTERA/.

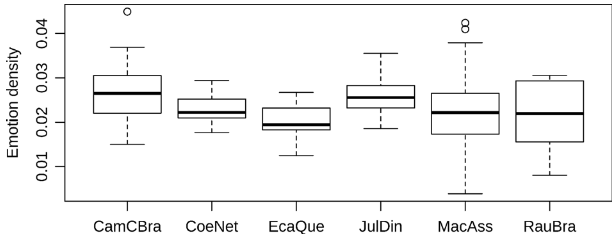


Fig. 9 Box-plots of reference to emotion in six authors: Camilo Castelo Branco, Coelho Neto, Eça de Queirós, Júlio Dinis, Machado de Assis and Raul Brandão

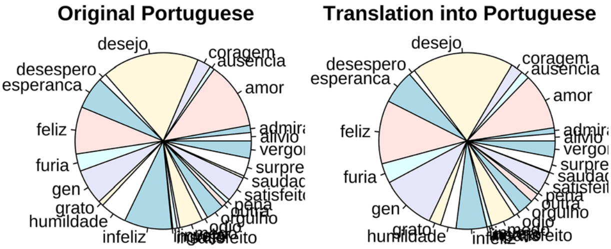


Fig. 10 Differences between Portuguese and Norwegian originals

text in Norwegian and many more short stories in Portuguese in PANTERA can also play a significant role. (In addition, and because of the alignment procedure employed in PANTERA, in translation the sentence can be repeated several times,[14] which produces some unbalance in the quantitative results.)

A quantitative summary of emotions in PANTERA is presented in Table 6.

We can see that there is higher diversity in the emotion lexicon in original Portuguese, and a slightly higher density of emotions. While this is expectable in terms of translation theory, which claims that translated texts tend to be less extreme – and can also be a by-product of repeated sentences in the translated part, or caused by the slightly less diversity of authors, none of this is very noteworthy.

6 Other approaches

Although we think our work is quite unique in scope and method, we provide here a short overview of similar or related work for other languages here. We will not review here the ways emotion as been described or categorized throughout history or in computer science, since this has recently been done in Maia and Santos (2018) and Santos and Maia (2018) respectively, but will deal solely with corpus annotation and its evaluation.

[14] In the case of two sentences in the original text being translated by one sentence, we chose to repeat the translated sentence in both alignments, instead of dividing it in two parts.

Table 6 Emotion numbers in the Portuguese side of PANTERA

	Original	Translated
Words	255,334	416,245
Texts	53	65
Authors	31	27
Emotions	5313	7146
Emotion density	0.0208	.01717
Emotion lemmas	749	692

Some of the pioneers of what one can call “emotion annotation” are Maia (1994) (for Portuguese and English) and Aman and Szpakowicz (2007) (for English), who defined clearly an “emotion annotation task” that is categorical.

The project which is closest in spirit and size to ours — although we only learned about it recently — is the one of Ptaszynski and colleagues (Ptaszynski et al. 2014), who annotated automatically a large scale corpus of Japanese blogs (5 billion words). As they claim, most previous work on emotion annotation targeted blog text, be it for Chinese, Bengali, English or Japanese (see references in their paper). They have a wide concept of emotion annotation, that borders with sentiment analysis, since they consider “emotion corpora” if they have emotion classes, but also if they simply (?) distinguish between emotive or non-emotive, or subjective and non-subjective sentences. Contrary to us, they do not worry about whose emotion it is.

Their corpus is annotated with two kinds of (emotional) information: emotive sentences (namely, sentences which indicate that the utterer is feeling some kind of emotion, by including “emotemes”), and emotive expressions, “words or phrases that directly describe emotional states”. It is only the second kind we are interested in.

They used the following 10 emotion classes: joy, dislike, fondness, fear, relief, excitement, surprise, gloom, anger, and shame (or better, these are the translations of Japanese concepts). Interestingly, we do not have a class for “excitement” in Portuguese. This may be related to the kind of user-generated content they are concerned with, but we should investigate whether at least the corresponding lexical items in Portuguese should be added (initially, to the other class).

They investigated whether lexical richness was correlated with annotation volume using Spearman’s rank correlation test, and found no correlation (ρ=0.38).

They propose an interesting evaluation form, namely the generation of what they called a “robust emotion object ontology” (Ptaszynski et al. 2014, page 52), where causes and consequences are extracted (*I’m sad because my girlfriend dumped me...*), as well as suggested a moral consequence retrieval agent.

Bostan and Klinger (2018) attempt a cross-classification of emotion corpora according to the following dimensions: annotation schemes (which emotion models); annotation procedures (expert annotation, crowd-sourcing, “distant supervision” — adding tags, etc.), data type and application (what they call “domains and topics”), and methods.

They claim that Twitter is the currently (2018) preferred object of research, but they also mention literature as another hot topic. (Although they cite research whose purpose is to identify whose emotion, Buechel and Hahn (2017), they do not distinguish it in the overview).

Trying to see our work from their perspective, we use as many genres as we have available, but little user-generated content (apart from blogs). The way we annotate may be classified as another kind of expert annotation, but rule-based, and incremental, in the sense that the more revision rules we write, the better the annotation becomes. As to the categorical model we use, it arose from the study of Portuguese and cannot be directly compared to English models, even if purported to be universal.

Seyeditabari et al. (2018), also doing a review, propose the label "emotion detection" to cover the identification of discrete emotions in text, suggesting that this task is a natural evolution of sentiment analysis. They are interested in the emotion conveyed by a sentence, paragraph or even document (just like in sentiment analysis), while we look only at reference to emotion. So in a way our work may be a prerequisite to their task. However, they do not distinguish at all about whose emotion one is concerned with, therefore conflating different problems.

Canales et al. (2020) suggest a computational paradigm, intensional learning, which is a bootstrapping approach from a set of seed emotional sentences, augmented with distributional semantic methods, on top of which a supervised classifier is then trained, to efficiently build up automatically annotated corpora. In a way, the first step is similar to what we did (for the lexicon), but instead of creating a machine learning classifier we developed rules that we subsequently apply. But it is important to stress that the authors, together with the previous ones we mentioned, are interested in overall emotion and not (linguistic) reference to emotion. Clearly, it is a rather different task.

If we look specifically for emotions in literary corpora, there are several works that deal with these, from different angles:

Alm et al. (2005), concerned with the "text-based emotion detection task", for text to speech applications, claim that "narrative text is often especially prone to having emotional contents" (page 579), and classify emotional text passages in children's stories according to eight emotion classes (one neutral). They explain that they target the feeler (writer/teller) emotion.

Mohammad (2012) does emotional tracking in emails and books, and compares novels and fairy tales according to six emotions (Eckman's), as well as visualizes plots showing how emotion-related words vary through the course of a book. Note that, differently from our work, it is not simply emotion words, but emotion-related words.

Acerbi et al. (2013) study the prevalence of categories of emotion in English using Google books, and show that they diminish with time.

Kim et al. (2017) investigate the relationship between five sub-genres of the novel (adventure, humor, mystery, romance, science fiction) in English and the emotional structure of the plot, and conclude that "emotions carry information that is highly relevant for distinguishing genres" — and that this approach performs as well as classifiers based on “rich, open-vocabulary lexical information”. This is an — albeit indirect — vindication of the usefulness of our kind of annotation.

For Spanish, Moreno-Jiménez and Torres-Moreno Moreno-Jimênez and Torres-Moreno (2020) report on a corpus of 500 literary sentences in Spanish annotated with five emotions, but its purpose is mainly the test of machine learning classifiers.

In conclusion, we still believe that what we purported to do for Portuguese was rather different, more linguistically focused and possibly less ambitious. We wanted to study the reference to emotion by linguistic means, not the emotions conveyed or felt. How the two tasks interact, correlate, or eventually can help each other, is something that remains to be ascertained.

7 Concluding remarks

We presented here a resource for the study of emotions in Portuguese text, that to our knowledge is unique, both in size and in scope, for any language. It is available for querying on the Web[15] and, for the corpora that we distribute as full text, also available as downloadable annotated corpora (amounting to more than 370 million tokens)[16].

In addition to describing how our resource was created and which assumptions underlie it, we also presented several use cases to demonstrate its usefulness, which should often be read as suggestions for further work.

We hope that this resource will contribute to the study of emotions in general and to the study of emotions in Portuguese, and that user feedback will help us improve the annotation significantly.

An important clarification is probably needed here. One of our main interests is the comparison of different languages, as motivated in Maia and Santos (2018); Santos and Maia (2018). But selecting arbitrarily some emotions, like for example Plaza del Arco et al. (2020) did, does not allow one to identify how different languages work. Rather, we believe we must have a thorough coverage of one language first so that we may then attempt such comparison. And it may be refreshing that for once it is not English which leads the way.

Acknowledgements We thank the whole Linguateca team with whom we have had discussions on these matters, especially Belinda Maia, Cláudia Freitas and Barbara Ramos. We are grateful to Eckhard Bick and Tino Didrichsen for help with vislcg3; to the Research Computing Services of Sigma, Norway, for cluster facilities, and to FCCN - Fundação para a Computação Científica Nacional for the allocation and maintenance of Linguateca’s servers. This project was partially funded by Portuguese national funds (PIDDAC), through the FCT - Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project UIDB/05549/2020 and the project NORTE-01-0145-FEDER-000045, supported by Northern Portugal Regional Operational Programme (Norte2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (FEDER).

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital).

[15] See https://www.linguateca.pt/ACDC/.

[16] See https://www.linguateca.pt/Gramateca/Emocionario.html.

References

Acerbi, A., Lampos, V., Garnett, P., & Bentley, R. A. (2013). The expression of emotions in 20th century books. *PLoS ONE, 8*(3), e90972. https://doi.org/10.1371/journal.pone.0059030

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 579–586). Association for Computational Linguistics.

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matousek & P. Mautner (Eds.), *Text, Speech and Dialogue (TSD)* (pp. 196–205). Berlin, Heidelberg: Springer.

Bick, E., & Didriksen, T. (2015). CG3 - beyond classical constraint grammar. In B. Megyesi (Ed.), *20th Nordic Conference of Computational Linguistics (NODALIDA)* (pp. 31–39). Linköping University Electronic Press.

Bostan, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *27th International Conference on Computational Linguistics* (pp. 2104–2119). Santa Fe, New Mexico, USA.

Bradley, M.M., Lang, P.J.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Tech. Rep. C-1, The Center for Research in Psychophysiology, University of Florida (1999). http://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf

Buechel, S., & Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In 15th *Conference of the European Chapter of the Association for Computational Linguistics* 2, (pp. 578–585). Valencia, Spain.

Canales, L., Strapparava, C., Boldrini, E., & Martínez-Barco, P. (2020). Intensional learning to efficiently build up automatically annotated emotion corpora. *IEEE Transactions on Affective Computing, 11*(2), 335–347. https://doi.org/10.1109/TAFFC.2017.2764470

Ekman, P. (1984). Expression and the Nature of Emotion. In K. Scherer, P. Ekman (Eds.), *Approaches to Emotion* (pp. 319–343). Lawrence Erlbaum.

Evert, S., Team, T.C.D.: CQP interface and query language manual - CWB version 3.5 (2021). https://cwb.sourceforge.io/files/CQP_Manual.pdf

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. *Revista de Estudos da Linguagem, 23*(3), 641–680.

Gaizauskas, R. (1998). Evaluation in language and speech technology. *Computer Speech and Language, 12*(4), 249–262.

Galves, C., Faria, P.: Tycho Brahe Parsed Corpus of Historical Portuguese (2010). http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html

Goddard, C., & Wierzbicka, A. (1994). Pain: is it a human universal? In C. Goddard & A. Wierzbicka (Eds.), *Semantic and Lexical Universals* (pp. 127–155). Amsterdam, The Netherlands: John Benjamin Publishing.

Gonçalo Oliveira, H., Santos, D., & Gomes, P. (2010). Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática, 2*(1), 77–94.

Grefenstette, G., Qu, Y., Evans, D. A., & Shanahan, J. G. (2006). Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In Y. Qu, J. Shanahan, & J. Wiebe (Eds.), *Exploring Attitude and Affect in Text: Theories and Applications (AAAI)* (pp. 93–107).

Higuchi, S., Santos, D., Freitas, C., & Rademaker, A. (2019). Distant reading Brazilian history. In *4th Conference of The Association Digital Humanities in the Nordic Countries* (pp. 190–200).

Kajava, K., Öhman, E., Hui, P., & Tiedemann, J. (2020). Emotion preservation in tranlation: Evaluating datasets from annotation projection. In S. Reinsone, I. Skadina, A. Baklane, & J. Daugavietis (Eds.), *Digital Humanities in the Nordic Countries, 5th Conference* (pp. 38–50).

Kim, E., Padó, S., & Klinger, R. (2017). Investigating the relationship between literary genres and emotional plot development In *Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 17–26). Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2203

Kristensen, C. H., de Azevedo Gomes, C. F., Justo, A. R., & Vieira, K. (2011). Normas brasileiras para o Affective Norms for English Words. *Trends in Psychiatry and Psychotherapy, 33*(3), 135–146.

Lüdeling, A. (2011). Corpora in Linguistics. In K. Grandin (Ed.), *Going Digital: Evolutionary and Revolutionary Aspects of Digitization* (vol. 147, pp. 220–243). Science History Publications.

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

Maia, B., Santos, D.: Who is afraid of ... what? - In English and in Portuguese. Aspects of corpus linguistics: compilation, annotation, analysis **12** (2012)

Maia, B., & Santos, D. (2018). Language, Emotion, and the emotions: multidiciplinary and linguistic background. *Language and Linguistics compass, 12*(5), 12280.

Maziero, E. G., Pardo, T., Felippo, A. D., & Dias-da Silva, B. C. (2008). A base de dados lexical e a interface web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *6th Workshop em Tecnologia da informação e da linguagem humana (TIL)* (pp. 390–392).

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.

Mohammad, S. M., & Turney, P. D. (2012). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.

Moreno-Jimênez, L.G., Torres-Moreno, J.M.: Lisss: A new multi-annotated multi-emotion corpus of literary spanish sentences. Computación y Sistemas **24**(3) (2020). https://doi.org/10.13053/cys-24-3-3474

Mota, C.: Anotação de emoções nos corpos do AC/DC. Tech. rep., Linguateca (2013). https://www.linguateca.pt/documentos/Mota2013.pdf

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

Mota, C., Santos, D., & Barreiro, A. (2021). Paraphrasing Emotions in Portuguese. In B. Bekavak, K. Kocijan, K. Sojat, & M. Silberztein (Eds.), *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities* (pp. 134–145). Berlin, Heidelberg: Springer.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge, UK: Cambridge University Press.

Paltoglou, G., Thelwall, M., & Buckley, K. (2010). Online textual communications annotated with grades of emotion strength. In *3th international workshop of Emotion: Corpora for research on emotion and affect* (pp. 25–31).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1–135.

Pavlenko, A. (2008). Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism: Language and Cognition, 11*(2), 147–164.

Plaza del Arco, F. M., Strapparava, C., López, L. A. U., & Valdivia, M. T. M. (2020). EmoEvent: A multilingual emotion corpus based on different events. In *12th Conference on Language Resources and Evaluation (LREC)* (pp 1492–1498).

Plutchik, R. (1991). *The Emotions, revised edition edn.* University Press of America.

Ptaszynski, M., Rzepka, R., Araki, K., & Momouchi, Y. (2014). Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. *Comput. Speech Lang., 28*(1), 38–55. https://doi.org/10.1016/j.csl.2013.04.010

Ramos, B., & Freitas, C. (2019). Sentimento de quê? uma lista de sentimentos para a Análise de Sentimentos. In *Symposium in Information and Human Language Technology (STIL)* (pp. 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.), *Digital Humanities and Natural Language*

Processing Workshop at the 14th International Conference on the Computational Processing of Portuguese Language.

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

Santos, D.: Doctors in lusophone literature (2019). https://dls.hypotheses.org/952. Blog post in Digital Literary Stylistics (SIG-DLS)

Santos, D. (2019). Literature studies in Literateca: between digital humanities and corpus linguistics. In M. Doerr, Øyvind Eide, O. Grønvik, B. Kjelsvik (Eds.), *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore* (pp. 89–109). Novus forlag

Santos, D. (2019). PANTERA: a parallel corpus to study translation between Portuguese and Norwegian. In J. Askeland, M. Gargiulo, S.O. Rosales (Eds.), *XX Scandinavian Romanist Conference Bells* Volume 10, no. 1.

Santos, D., & Bick, E. (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhauer (Eds.), *2nd International Conference on Language Resources and Evaluation (LREC)* (pp 205–210).

Santos, D., Freitas, C., Bick, E. (2018). OBras: a fully annotated and partially human-revised corpus of brazilian literary works in the public domain. In *OpenCor Workshop*. https://www.linguateca.pt/Diana/download/CorLex.pdf

Santos, D., & Maia, B. (2018). Language, emotion and emotions: a computational overview. *Linguistics and language compass, 12*(5), e12279.

Santos, D., & Mota, C. (2010). Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), *International Conference on Language Resources and Evaluation (LREC)* (pp. 1437–1444).

Santos, D., & Mota, C. (2015). A admiração à luz dos corpos. 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, 7(1), 57–77. OSLa.

Santos, D., Pires, E., Lopes, J. M., Fuão, R. S., & Freitas, C. (2020). Periodização automática: Estudos linguástico-estatásticos de literatura lusófona. *Linguamática, 12*(1), 80–95.

Santos, D., Simúes, A.: Towards a computational environment for studying literature in Portuguese (2019). https://www.linguateca.pt/Diana/download/PresentationBudapestSantosSimoes.pdf. Presentation at DH Budapest 2019, Digital Humanities Conference (Budapest, 25-27 September 2019)

Santos, D., Simúes, A., Mota, C.: Estudo de sentimentos: algumas direções. In: Workshop Empirical Research on Portuguese (2020). https://www.linguateca.pt/Diana/download/posterSantosetal2020.pdf

Schöch, C., Erjavec, T., Patras, R., & Santos, D. (2021). *Creating the European Literary Text Collection (ELTeC)*. Modern Languages Open. to appear.

Seyeditabari, A., Tabari, N., Zadrozny, W.: Emotion detection in text: a review (2018). https://arxiv.org/pdf/1806.00674.pdf

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods, 44*(1), 256–269.

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *4th International Workshop on Semantic Evaluations* (pp. 70–74). Association for Computational Linguistics.

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems, 9*(4), 483–496.

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge, UK: Cambridge University Press.

Zampieri, M., Becker, M. (2013). Colonia: Corpus of Historical Portuguese. *ZSM Studien* **5**. Special Volume on Non-Standard Data Sources in Corpus-Based Research.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.