

# Emotions in natural language: a broad-coverage perspective

**Cristina Mota**

Linguateca

d.s.m.santos@ilos.uio.no

**Diana Santos**

Linguateca & University of Oslo

cmota@ist.utl.pt

## Abstract

This paper presents five studies of emotions in Portuguese on a large corpus infrastructure that is offered as a public service: (i) negated emotions, (ii) attribution, (iii) assessment, (iv) objects, and (v) co-occurrences of emotions.

## 1 Presentation

Nowadays there is undeniably a strong interest in sentiment and emotion in natural language, mainly concerning attitude detection: positive vs. negative reviews or reputation, see e.g. Pang and Lee (2008) for an overview.<sup>1</sup>

In our work, initially sparked by a contrastive analysis between English and Portuguese, we study the different emotions and how they are expressed in Portuguese as a window to the basic categorization of the world in that language. Contrary to common belief, we believe that language is almost always subjective – see Ellis (1993) for linguistic and philosophical arguments for this position. Presentation of impartial facts almost always conceals shades of subjectivity and opinion that are relevant to a full understanding. (For example, appositions. They seem factual, but are a definite choice of the presenter to show the presentee in a given light. Plus, many examples of use of language seem to present facts at face value, but they employ irony, appeal to ridicule, use creativity, and say something quite different.)

<sup>1</sup>Pang and Lee (2008, page 6) explicitly stress that they “focus on information-access applications, as opposed to work of more purely linguistic interest”, and that the importance of the latter is absolutely not in question, but the fact is that computational works on emotion tend to cite Pang and Lee (2008) and not linguistic works as well.

Table 1: Names and abbreviations into English

abbrev	Portuguese	English
al	alívio	relief
ad	admirar	admiration
am	amor	love
co	coragem	courage
dj	desejo	desire
dp	desespero	despair
es	esperança	hope
fe	feliz	happy
fu	fúria	anger
gr	grato	grateful
inf	infeliz	unhappy
hu	humildade	humbleness
ins	insatisfeito	insatisfaction
me	medo	fear
or	orgulho	pride
sat	satisfeito	content
sau	saudade	missing, nostalgia
su	surpresa	surprise
ve	vergonha	shame

Our work has a strong empirical bias: To study the actual way people employ emotion and evaluative words in a particular language, broad-coverage annotation of emotions and feelings in different genres is necessary.

## 2 What is an emotion?

Already in (Maia, 1994)’s seminal work on the subject she discusses that “emotions” are far from consensual, and that the layman’s view (for example of *love* and *hate*) does not coincide with what psychologists or sociologists have defined as emotions. Her way out – as ours – is to use linguistic behaviour to single out emotion groups.

In fact, not only several distinctions are debatable – the distinction between intellectual/cognitive and purely emotional responses, and between dispositions or temporary states –, but words for emotions themselves are difficult to categorise clearly: e.g. does *disgusting* belong to DISLIKING or CONTEMPT? And *maravilhar* to ADMIRAR or to SURPRESA?

Staiano and Guerini (2014) have reported on the incomparability of different computational re-

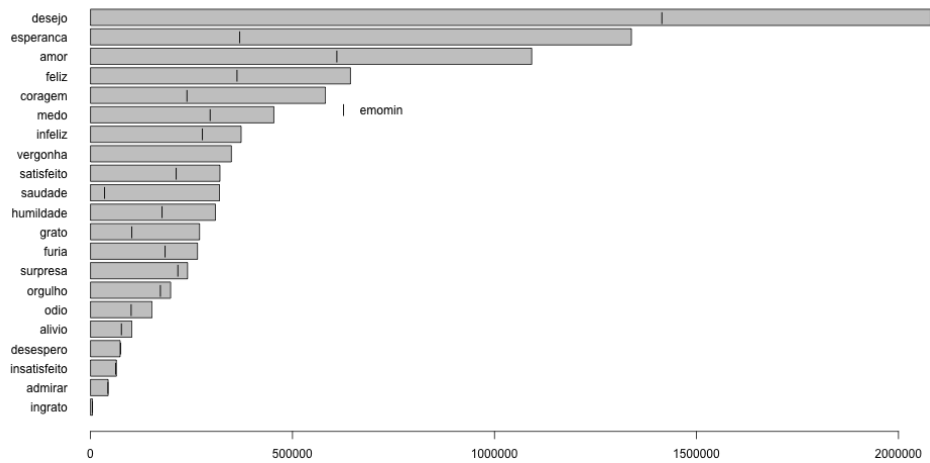


Figure 1: Emotion occurrences in the corpora, min and max

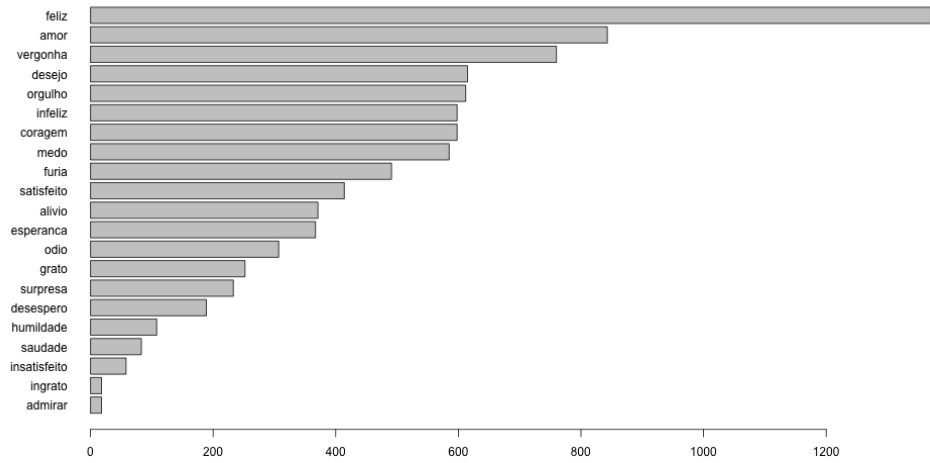


Figure 2: How many different words (lemmas) per emotion

sources. We had also problems to find boundaries or clear cut emotion (fields). What is different between the two works (after all, 20 years have gone by) is the size of the data involved: 11 fiction works in English and 11 in Portuguese in 1994, compared to 1,500 million words in many genres in Portuguese text now.

We were therefore not able to classify – as Maia did – every instance of emotion words, and had to rely on lexical clues (obtained automatically from a set of seeds from lexical resources, and then humanly revised). As a first way to identify the sheer volume of what we had to process later, we created two lists: the first, **emomin**, included those lexical items that indicated (almost) always emotion, the other more encompassing one, **emomax**, contained all those words that could be emotions but might also be something else. This allows a

more accurate idea of the true picture if we ever manage to have perfect annotation (in which the context disambiguates).

In Figure 1 we show the approximate number of emotions in our corpora, with the minimal number being named .min, while in Figure 2 we measure emotion diversity – the number of different lexical items belonging to each “emotion”.

The first thing to mention is that hardly any of these data can be taken at face value: For instance, the high proportion of GRATITUDE in the corpora (compared to INGRATITUDE) stems from the multipurpose and highly frequent verb (and noun) *reconhecer* (to recognize) and *reconhecimento* (recognition), which can be used to recognize favours, but the vast majority of the cases have nothing to do with gratitude. Or take the field of RELIEF, where in a previous version the most

frequent lexical item found, *facilitar* (to ease), may bring relief, but is hardly ever employed in an emotion context (so much so, that we simply took it out from the emotion lists). But we could not do the same in the field of SATISFACTION, where *realizar*, meaning in general “execute” or “implement” instead, is only related to satisfaction when reflexive. We have obviously tried to deal with this with emomin anotation, but want to illustrate the dangers of lexicon-driven annotation.

Another interesting observation was the high number of emotion words in legal documents, due to a large sharing of emotion words with legal actions, e.g. *confiar* (trust and legal deposit), *apreciar* (appreciate and legally appreciate). Genre (as well as subject/theme, since bank transactions or loans also use the same lexicon) may thus be used as a negative filter in future refinements.

Preliminary as these data are, the attempt to annotate all emotions showed: First, for some fields, there is no clear identification of positive or negative member, like in PRIDE/HUMBLENESS/HUMILIATION. Likewise, if we consider FEAR/COURAGE/DARING – the way you see the things that inspire the pride (or fear) decide whether the pride (or fear) is positive or negative (check e.g. *fear of God*, *fear the truth*). Secondly, other cases show no negative or positive poles, as SAUDADE (‘longing’), where it depends on the speech act, and on the actor focused upon (subject or object): To say that you miss something is positive about that “something”, but negative about your state of mind. To say that someone is missed is positive about her but also negative (the thought of her brings sadness). Or, *dar saudades a* is a conventional Portuguese greeting (roughly *send regards to X*) and is rather positive than negative, since it conveys that I like and think about X. Thirdly, where clear positive and negative poles can be identified, the frequency of each can be very different: Figure 5 presents the relative importance of the lexicon in the different polarized emotions, in all corpora. While (lexical) satisfaction is always more mentioned than insatisfaction, and love is more expressed than hate, the fields of fear/courage and of pride and its contraries are rather more complex, and apparently depend on genre as well.

This shows that, while the details of each emotion group require verification and improvement, to be able to see them at a glance and investigate

positive or negative predominance is interesting in itself. Anyway, it is important to stress that we are measuring here the lexical predominance of each pole, not the actual expression of positive or negative emotions, since negation reverses the polarity.

### 3 Negated emotions

It is common to use negation and double negation as hedges, and expressions like *I don’t like* are socially preferred to *I hate* or *I dislike*. Maia (1994, sec. 12.1) found out that, except for the LIKING and DISLIKING groups,

one of the most interesting points that emerged was that the negative very rarely applies. It would seem that we do not often feel the need to mention an emotion merely to negate it.

In addition, she makes the point that when negation is linguistic, as in

We can say *I don’t like Mary* and mean *I dislike Mary*, but one can also say it and then add *I love her*, or *I am simply indifferent to her*.

it is necessary to check which side of the emotion is meant, in case one were tempted to count negatives as belonging to the opposite pole, i.e. DISLIKING instead of LIKING in this case.

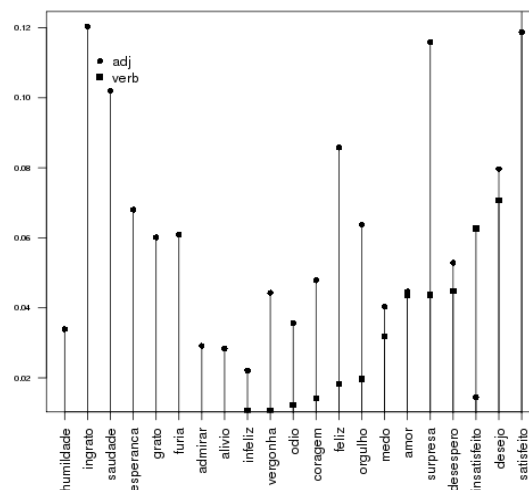


Figure 3: Which verbal and adjectival emotions (min) are mostly negated, ordered by the proportions of negated adjectives.

With our corpora we can measure which emotions are often associated with negation. First, we

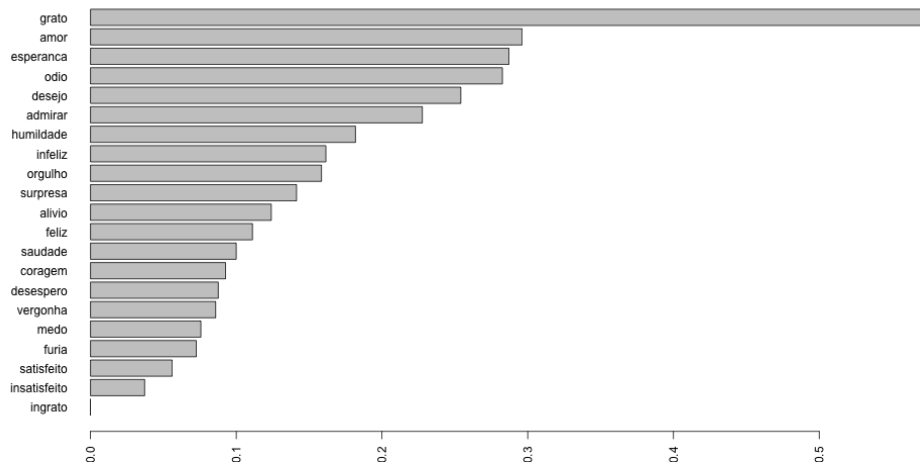


Figure 4: Percentage of emotion words in the first person

checked directly negated emotion verbs, and adjectives in copular sentences with *ser*, *estar* or *ficar* as copulas. Figure 3 displays the proportion of negated cases. Verbs denoting courage and dissatisfaction are the most negated, while satisfaction and wish adjectives were the most negated.

Second, we looked into negative verbs which are themselves negated, like *não desgostar*, (literally “not dislike”, which is a way of conveying a weak liking), as described by Santos (2008), who also discusses *não achar* and *não importar*, roughly corresponding to *do not think* and *do not mind*. Although the proportion of negated cases was higher than average: 20%, the numbers were too small (204 in 1019) to be conclusive results.

#### 4 Whose emotion?

Maia classifies all emotions in context as being senser or phenomenon oriented, and finds for different emotion (fields) different patterns of orientation. In her 11 different types, she separates emotions related to self from others. For us it was out of question to attempt such a detailed classification, but as a simple proxy to assess whether one is describing the emotions of others or one’s own, we used grammatical person: first person against non-first person, restricting the counts to finite verbs identifying emotion (verbs are marked for person in Portuguese). Figure 4 presents the results for all corpora.

Interestingly, the higher proportions of first person concern the expression of likes and dislikes and hopes<sup>2</sup> while happiness or unhappiness are not

<sup>2</sup>The huge percentage of GRATITUDE comes from thank-

so frequent in first person. This last result has to be taken with care, though, since some of these verbs are ergative-like, just like in English *it pleases me*, and therefore can be in the third person but relate to the first. Shame, pride, anger and despair are more used about others, as one would expect. Fear, courage and surprise are relatively high in the first person, and this may be due to their use as hedges, as illustrated in English by *I am afraid that*, *I dare say that* or *I was surprised to see...* to tone down the actual message. (Why SAUDADE and HUMILDADE have no 1st person is simply due to the fact that no verbs have been assigned that emotion.<sup>3</sup>)

An obvious follow-up on this would be to study differences between plural and singular (are there emotions that we feel or express together and others individually?), and tense of emotions.

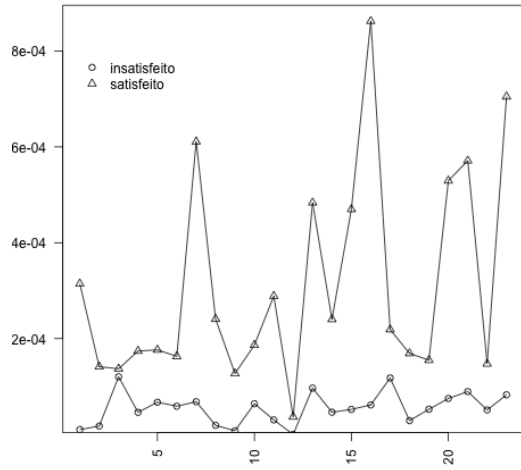
#### 5 Opinion, evaluation and emotions

In a sentence like *Ele estava cheio de medo*. (He was full of fear) it is hard to disentangle whether we face an assessment, a moral judgement, or a simple description of an emotion, or even all of them at the same time. But there is one case where we can be almost sure:

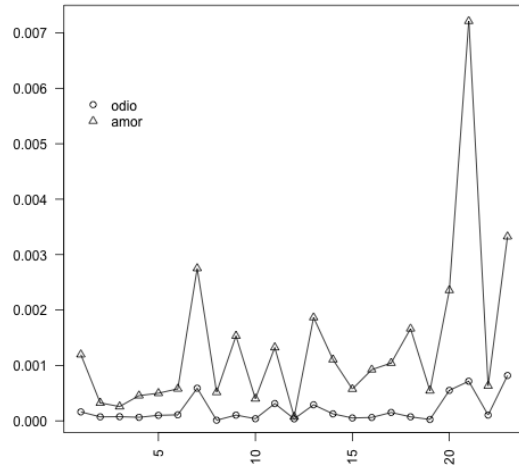
A particular kind of sentential adverbs in Portuguese always describe our (the speaker’s) judgement towards what is described, and not the feelings or attitudes involved in the action. The

ing, and one can discuss whether it expresses an emotion or just a conventional act.

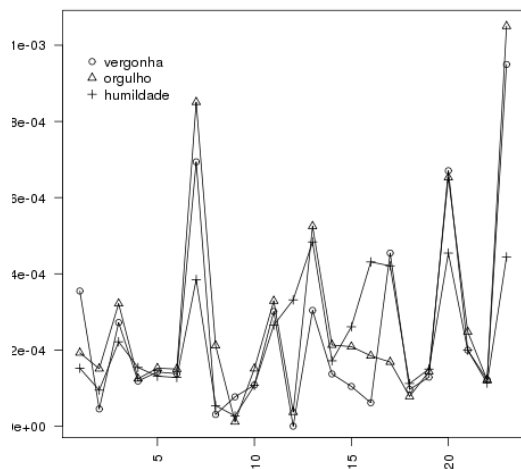
<sup>3</sup>This is an interesting difference compared to English or other Germanic languages, where the first emotion is mainly expressed by verbal means, as in *to miss*.



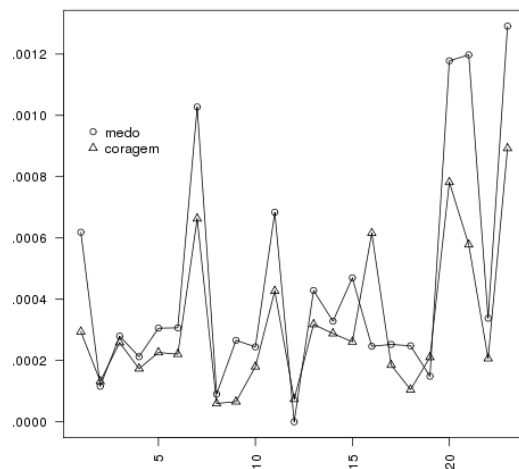
(a) (In)satisfaction



(b) Love/hate



(c) Shame/pride/humility



(d) Fear/courage

Figure 5: Positive or negative (in the emomin case) predominance? It depends on the emotion (field)

Table 2: Most frequent objects (lemmas) of emotion words

MEDO (83365)	que, ser, perder, fazer, poder, deixar, entrar, matar, represália, ter	(that, be, lose, do/make, can, let/leave, enter, kill, repr. have)
ESPERANÇA (12906)	que, ser, ter, poder, estar, haver, fazer, ir, ver, dever	(that, be, have, can, be, there be, do/make, go, see, should
AMOR (237619)	que, ser, fazer, ter, ficar, ver, falar, manter, trabalho	(work, appearance, that, culture, knowledge, experience)
envergonhar (1462)	país, nação, povo, brasileiro, país, governo, sociedade, família	(country, nation, Brazil, Brazilians, people, parents, government, society, family)
envergonhar-se com/de (1134)	mundo, pai, homem, dizer, país, filho	(world, parent(s), man, say, country, children)
zangar/enfurecer-se com (836)	particular, mundo, jogador, amigo, partido, pessoa, causa	(particular, world, player, friend, party, person, cause)
admirar/espantar-se com (2576)	fato/facto, quantidade, número, resultado, capacidade, preço, falta	(fact, amount, number, result, ability, price, lack)

most frequent are: *felizmente* (5234 occurrences), *graças a Deus* (2923) for positive attitude and *infelizmente* (17872), *lamentavelmente* (3668) for negative. Our corpora thus contain three times more expressions of negative judgement than positive (in this form).

## 6 Objects of emotion

What kind of (semantic) objects are the target of emotions, or their source? Inspired by the work on FEAR by Maia and Santos (2012), we selected the direct objects of some emotion verbs, presenting the most frequent in table 2.

A cursory look suggests some interesting remarks: speakers of Portuguese connect SHAME mostly with their country – note that we separated two forms of shame: causing shame on others (transitive *envergonhar*) and being ashamed of/on behalf of (reflexive *envergonhar-se*). And they are angry (or voice ANGER) at the world, players, (political) parties, and FEAR mostly situations or reactions, while HOPING for states, situations, actions and results.

This is, of course, highly genre dependent. When the same questions are posed by genre, a different picture emerges, although genre in itself is not easy to pin down. Figure 6 shows an initial distribution of emotions by genre.

## 7 Co-occurrence of emotions

Another interesting investigation is: which emotions tend to co-occur in language, and how often? Co-occurrence in language can be measured at the sentence level, at the paragraph level or even at the text level. We have started with co-occurrence at the sentence level, producing Table 5.

Note that we count also co-occurrence of two or more elements of the same emotions, so that a co-occurrence matrix has as diagonal the cases of more than one word of such emotion in the same sentence. The fact that the diagonal numbers are so high can be seen as a confirmation of the work of Justeson and Katz (1992), who made the case that semantic relations like antonymy were also

defined by their co-occurrence.

To ease interpretation of the co-occurrence table, the absolute numbers concerning each emotion are provided in Table 3.

Table 3: Size of emotions: when one word is vague between two emotions, it counts 1/2

admirar	39571.0	alivio	92215.0	amor	764303.0
coragem	525736.0	desejo	1424029.0	desespero	63529.0
esperanca	1158530.0	feliz	400689.0	furia	168687.0
grato	258320.0	humildade	175839.0	infeliz	303708.0
ingrato	4406.0	insatisfeito	58395.0	medo	403085.0
odio	136484.0	orgulho	181880.0	satisfeito	286328.0
saudade	294739.0	surpresa	212738.0	vergonha	302408.0

In fact, it is possible that percentages of co-occurrence, as presented in Table 4, might be more easily understandable, and no longer symmetric.

Table 4: Percentage of the occurrences of the emotion on the left that co-occur with the emotion on the top

	admirar	amor	odio	vergonha
admirar	2.6	<b>8.9</b>	1.0	1.1
amor	.46	<b>9.5</b>	1.4	1.3
odio	.82	<b>8.0</b>	5.0	2.4
vergonha	.37	.14	1.8	<b>3.8</b>

These numbers measure the proneness of co-occurrence, and are therefore not symmetric, since they are weighted by the absolute frequency of the emotion. It is thus clear that it is much more common for (mentions of) ODIO (hate) to be accompanied by (mentions of) AMOR (love) than vice-versa: 8.0 vs. 1.4.

## 8 Concluding remarks

This<sup>4</sup> is an interesting infrastructure for studying emotions in Portuguese. We believe an emotion map is language specific, and can thus be different from other languages; see Lewandowska-Tomaszczyk and Wilson (2001) or Lewis (2000). While we are still performing initial explorations, the data is available for everyone else to investigate as well.

Given that agreement on emotion is problematic (Volkova et al., 2010), we intend to cater also for user-generated emotion groups, so that more

<sup>4</sup>Details in the final paper.

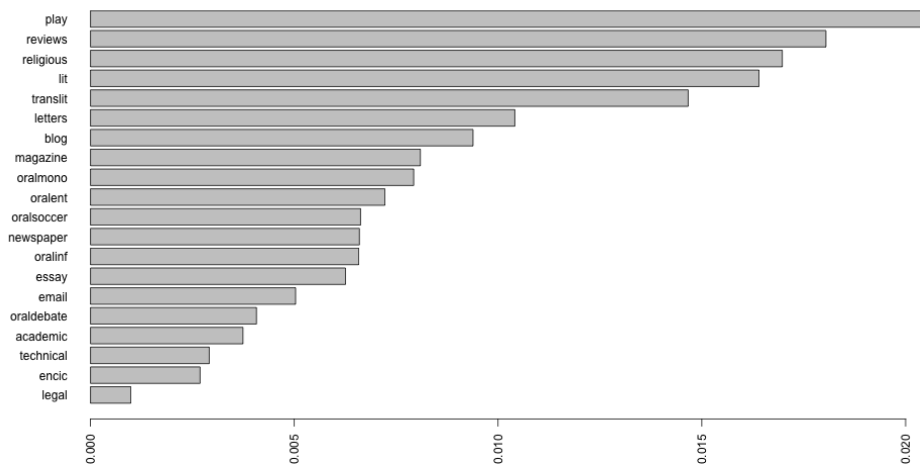


Figure 6: Distribution of emotions per genre

fine-grained – or alternative – studies can be conducted in the material, replicating studies as varied as those in Volkova et al. (2012; Mihalcea and Liu (2006; Hatzivassiloglou and McKeown (1997; Turney and Littman (2003), to cite just a few.

Users can use the annotated corpora to improve and evaluate a variety of lexical resources for Portuguese, such as WordnetAffectBR (Pasqualotti and Vieira, 2008) and Sentilex (Silva et al., 2012), or to ameliorate annotation schemes such as the one in ReLi (Freitas et al., 2014).

Also, literary-minded scholars might be happy with emotional signatures from different authors, using the subset of written fiction from our corpora, and those could be used for e.g. studying literary influence and/or characterizing style.

It could also provide ample material to develop an Hourglass of Emotions model (Cambria et al., 2012) for Portuguese.

Finally, we envisage a number of studies of specific emotions that can go deeper in emotional research and be of help for discourse analysis as well.

## References

- Erik Cambria, Andrew Livingstone, and Amir Husain. 2012. The hourglass of emotions. In *Proceedings of the 2011 International Conference on Cognitive Behavioural Systems, COST'11*, pages 144–157, Berlin, Heidelberg. Springer-Verlag.
- John M. Ellis. 1993. *Language, Thought and Logic*. Northwestern University Press, Evanston IL.
- Cláudia Freitas, Eduardo Motta, Ruy Luiz Milidiú, and Juliana César. 2014. Sparkling Vampire... lol! Annotating Opinions in a Book Review Corpus. In Sandra Aluísio and Stella E. O. Tagnin, editors, *New Language Technologies and Linguistic Research: A Two-Way Road*, pages 128–146.
- Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL (Madrid, Spain, July 1997)*, pages 174–181.
- John S. Justeson and Slava M. Katz. 1992. Redefining Antonymy: The Textual Structure of a Semantic Relation. *Literary and Linguistic Computing*, 7(3):176–184.
- Barbara Lewandowska-Tomaszczyk and Paul Wilson. 2001. Culture-based Conceptions of Emotion in Polish and English. In Peter Lang, editor, *Explorations across Languages and Corpora (PALC 2009)*, pages 229–239.
- Richard D. Lewis. 2000. *When Cultures Collide: Managing successfully across cultures*. Nicholas Brearely Publs.
- Belinda Maia and Diana Santos. 2012. Who is afraid of ... what? - In English and in Portuguese. In Signe Oksefjell Ebeling, Jarle Ebeling, and Hilde Hasselgård, editors, *Aspects of corpus linguistics*:

Table 5: Co-occurrence of emotions in the corpora

al	2301	150	3264	1650	5947	938	1825	419	2313	1120	728	1593	1234	818	1136
ad	150	1027	3512	1002	1739	184	478	557	761	414	699	869	460	891	439
am	3264	3512	72753	19976	291450	3260	8296	11139	13207	10873	13339	18358	9469	8085	9913
an	1650	1002	19976	29201	116374	1425	13343	6376	5734	3531	5111	5250	5479	3877	5846
ap	5947	2806	291450	32201	116374	4790	499922	14508	20722	9567	17384	22485	15619	11685	17307
aq	938	184	3260	1425	4790	1520	4609	423	3100	1311	846	1393	1592	999	1289
ar	1825	184	3260	1425	4790	1520	4609	423	3100	1311	846	1393	1592	999	1289
as	4515	1739	143528	21418	499922	4191	63338	32114	16879	5310	8508	15688	11633	9312	10522
at	3341	1765	34006	18227	45370	2965	5580	6581	17049	4105	9022	189787	6113	7486	6653
au	1825	478	8296	13343	11971	4609	8802	924	5119	42111	2725	2747	3012	2614	4060
av	679	557	11139	6376	14508	423	11068	3077	1811	690	2737	3739	2341	1847	1872
aw	519	905	9697	4163	11787	386	1262	3077	2180	920	5157	2664	1932	2500	2560
ax	2885	695	14787	5734	20722	4491	17049	4817	17040	3985	3377	8802	8141	3448	6861
ay	46	21	391	94	374	52	168	221	213	128	123	107	77	49	178
az	419	124	2474	891	4392	410	878	371	342	585	390	3205	1286	717	828
ba	2313	761	13207	8892	23015	3100	7584	1494	9416	6209	3381	4347	5405	9873	7100
bb	1120	414	10873	3531	9567	1311	42111	690	3985	6819	2156	2118	1364	1860	3281
bc	728	699	13339	5111	17384	846	22485	2747	3377	2156	4246	4246	2129	2010	5471
bd	1593	869	18358	5250	22485	1393	189787	3739	8802	2118	4246	14728	3471	4055	3205
be	1234	460	9469	5479	15619	1592	6113	3012	8141	1364	2129	3471	16097	2493	4712
bf	818	891	8085	3877	11685	999	7486	2614	3448	1860	2010	4055	2493	6759	2162
bg	1136	439	9913	5846	17307	1289	4060	1872	6861	3281	5471	3205	4712	2162	11436

compilation, annotation, analysis, number 12 in Studies in Variation, Contact and Change in English, Dezembro.

Belinda Maia. 1994. *A Contribution to the Study of the language of Emotion in English and Portuguese*. Ph.D. thesis, FLUP.

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *AAAI Spring Symposium on Computational Approaches to Weblogs*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1-135.

Paulo Pasqualotti and Renata Vieira. 2008. WordnetAffectBR: uma base lexical de emoções para a língua portuguesa. *RENTE. Revista Novas Tecnologias na Educação*, 6:1-10.

Diana Santos. 2008. *Corpos linguísticos da Linguatca: apresentação*, 3rd July. TaLC at TaLC: Teaching and Linguatca's (Portuguese language) Corpora.

Mário J. Silva, Paula Carvalho, and Luís Sarmento. 2012. Building a sentiment lexicon for social judgement mining. In *Proceedings of the 10th international conference on Computational Processing of the Portuguese Language, PROPOR'12*, pages 218-228, Berlin, Heidelberg. Springer-Verlag.

Jacopo Staiano and Marco Guerini. 2014. DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. In *LREC 2014*.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315-346.

Ekaterina P. Volkova, Betty J. Mohler, Detmar Meurers, Dale Gerdemann, and Heinrich H. Bulthoff. 2010. Emotional Perception of Fairy Tales: Achieving Agreement in Emotion Annotation of Text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10)*, pages 98-106. Association for Computational Linguistics.

Svitlana Volkova, William B. Dolan, and Theresa Wilson. 2012. Clex: a lexicon for exploring color, concept and emotion associations in language. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12*, pages 306-314, Stroudsburg, PA, USA. Association for Computational Linguistics.