

A dictionary of pronouns for Brazilian Portuguese

Flavio Carvalho

CEFET/RJ

Rio de Janeiro, Brazil
flavio.carvalho@eic.cefet-rj.br

Rafael G. Rodrigues

CEFET/RJ

Rio de Janeiro, Brazil
rafael.rodrigues@cefet-rj.br

Lilian Ferrari

UFRJ

Rio de Janeiro, Brazil
lilianferrari@uol.com.br

Gustavo P. Guedes

CEFET/RJ

Rio de Janeiro, Brazil
gustavo.guedes@cefet-rj.br

ABSTRACT

Function words like articles, prepositions, conjunctions, and especially pronouns are important to Text Analysis as they reveal much more about someone's psychological state than content words do. In this study, we present the development of a pronoun enriched dictionary for text analysis in Brazilian Portuguese (BP), named PronounBP, which is based on the Brazilian Portuguese language version of the LIWC dictionary. Pronouns have great importance in Text Analysis, so we included in PronounBP as many Portuguese pronouns as possible. In this work, we carry out two experiments. The first one presents a statistical comparison of percentage values in categories obtained for bilingual texts, using the English LIWC dictionary, PronounBP and the already available BP LIWC dictionary. It provides satisfactory results since higher correlation results were obtained using PronounBP. We also produced an empirical evaluation in a computational linguistic analysis task, and PronounBP has also achieved better results than the BP LIWC dictionary.

Author Keywords

Text Mining, Dictionary, Brazilian Portuguese, LIWC, Pronouns.

ACM Classification Keywords

I.2.7 [Natural Language Processing]: Text analysis

INTRODUCTION

People with access to social networks and messaging applications have been registering many feelings and emotions in text format [1]. Drawing on this great amount of text available, some studies have been carried out using text mining techniques to verify aspects of society, culture, governmental issues and individuals' everyday life. Text analysis can also be applied to analyze political homophily in social networks [4], to obtain information about social and economic status [20], to predict age and gender from social network users [16], among many other interesting tasks.

A widely used software for text analysis is Linguistic Inquiry and Word Count (LIWC) [18]. LIWC can be configured with language dictionaries to perform text mining tasks [8]. The available dictionary for Brazilian Portuguese (BP) language, named LIWC2007pt from here on, presented good results in detecting positive valence in texts in this language, considering the first published evaluations [2]. On the other hand, the authors state that the

performance observed in negative valence detection was not good.

LIWC2007pt has problems caused by the inclusion of many words with errors in spelling (e.g., 'ninguén') and inconsistencies because of words with the same sequence of adjacent characters before a wildcard character '*' (asterisk) [5]. Another problem is related to the improper association of a large number of words to categories. In the pronouns category (pronoun), it was possible to find 44 words, from a total of 128, that were improperly classified. Considering all the pronouns subcategories, the number of improperly included words is 113. Since LIWC2007pt has 337 words associated with these categories, this indicates that more than 33% of words in pronouns categories are not pronouns. Other words in the LIWC2007pt dictionary were not in the appropriate categories, for example, some words that are considered pronouns in language were not found in pronouns category.

The importance of pronouns as markers of attitude is well documented in the literature [6]. For example, studies such as Rude et al. [26] suggest that depressed students use more first-person singular pronouns than never-depressed students do. Other studies such [31] indicate that there is a strong correlation between pronoun use and political orientation. It is also important to note that people use fewer first person singular pronouns and greater first person plural pronouns with age (increase) [19].

Considering the importance of pronouns to Text Analysis [15,17,6], this work aims to produce a new pronoun dictionary, named PronounBP. To create this dictionary, we compared pronouns from LIWC2007pt with words present in a list of Portuguese pronouns we created and, after that, we (i) removed some pronouns from inadequate categories; (ii) included words in the pronouns categories.

In this work, we set up two experiments to validate the effectiveness of PronounBP. The first one presents a statistical comparison of percentage values in categories obtained for bilingual texts, using LIWC 2015 English version (LIWC2015), PronounBP and LIWC2007pt. It provides satisfactory results since most of the PronounBP categories showed higher correlation results to LIWC2015 than LIWC2007pt. We also produced a classification experiment to predict age from texts. The results indicate a better accuracy in predicting the author's age using PronounBP than using LIWC2007pt.

This work consists of five more sections: Section 2 describes LIWC; Section 3 presents works related to the translation of LIWC into other languages; Section 4 details the materials and procedures. Section 5 describes the data sets constructed in this work; Section 6 presents the results; finally, Section 7 concludes.

LIWC

LIWC is a computer-aided dictionary-based text analysis program, which was developed with the objective to analyze emotional, cognitive and structural components from texts [18]. It can be divided into two main parts: one is the main program, and the other is a dictionary, which contains words categorized into one or more categories that reflect linguistic, psychological, and social processes.

The main categories are divided into subcategories. For example, the category of pronouns (pronoun) is divided into two subcategories: personal pronouns (ppron) and impersonal pronouns (ipron). The subcategory ppron is divided into five more subcategories: first person singular (i), first person plural (we), second person (you), third person singular (shehe) and third person plural (they). LIWC2007pt has the following number of words for each pronoun category: pronoun (128), ppron (54), i (7), we (8), you (25), shehe (16), they (11) and ipron (88).

It is possible to obtain different types of information from social networks users with LIWC, for example, political tendencies [4], social and economic status [20], among others. LIWC is also used in a study that finds significant differences for the Alzheimer's disease group in the usage of several LIWC categories, including impersonal pronouns, suggesting that the method could be used for dementia screening [28].

RELATED WORK

Lexicon-based approaches can be used beyond the possibility of providing an efficient means to analyze open-ended texts and aid in identifying symptoms of depression [10] and suicidal tendencies [30]. Studies using lexicons to subjectivity classification can also target social networks [14] since social media plays a significant role in political discourse [3].

The literature contains many translations of the LIWC dictionary for European languages, (e.g., Dutch [29], Spanish [23], and French [21]). In addition to European languages that use the Latin alphabet, there are versions that use other writing systems, like Chinese [10] and Japanese [28]. In these studies, it is observed, as pointed out by van Wissen [29], that the task of translating a LIWC dictionary is not direct (i.e., as the translation of a list of words), going beyond merely associating equivalent English words.

However, the development of methods for text analysis in languages other than English is still needed [24]. An elaborated systematic review study describes seven

affective lexicons in BP [7]. The review shows that, along with LIWC2007pt, these lexicons were created between 2007 and 2016.

LIWC2007pt, the BP translation, is based on the 2007 version of the LIWC dictionary. It was a collaborative effort of three teams, using Portuguese-English bilingual Dictionaries to include 127,149 words in 64 categories. The authors stated that no revision of the translation's manual work was made and that it can be improved¹.

As for studies on the use of pronouns in texts, Pennebaker and Stone [19] examined the simple linear relationships between language use and age. They determined the degree to which age was generally related to each of the LIWC dimensions. The authors found that with increasing age, individuals are better ready to disconnect themselves from their written topics, as observed by a consistent drop in the use of the first-person singular. Schler and others [27] also found significant differences in writing style among authors of different ages.

The leading developer and researcher responsible for LIWC has already indicated the importance of this category of words in the textual analysis [17]. However, none of the related work stated directly the intention to include as many pronouns as possible. We set out to present a different development in which even deviant forms like 'vc' and 'vcs' were considered for inclusion since they were already present in LIWC2007pt (although not assigned to any pronoun-related category). In addition, we included even some archaisms, regionalisms and dialect forms (e.g., 'ancê', 'ocê' and 'vacê').

BP PRONOUNS DICTIONARY

The first task in developing PronounBP dictionary was creating a list of pronouns (LOP1), with the aim of including as many Portuguese pronouns as possible. We used the Brazilian Academy of Letters website², Online Dictionary Caldas Aulete³, Michaelis⁴ and a grammar reference book [12]. Three researchers checked and provided validation of LOP1: a psychologist and two linguists.

When comparing words in LIWC2007pt with words in LOP1, we identified several words incorrectly assigned (in LIWC2007pt) to pronouns category (pronoun) and the ones hierarchically below (e.g., ppron, you, ipron). Then, we removed 44 words incorrectly assigned to pronoun category. Table 1 exhibits the words removed from this category, as well as the other eight words from Personal

¹ <http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc>, as accessed in 2018-10-05

² <http://www.academia.org.br/nossa-lingua/busca-no-vocabulario>

³ <http://www.aulete.com.br/>

⁴ <http://michaelis.uol.com.br/moderno-portugues/>

pronouns (ppron) and 49 from Impersonal pronouns (ipron) sub-categories. As described in Section 2, the assignment of words can result in the same word appearing in more than one category. For example, ‘algos’ can be seen in pronoun and ipron categories.

Category Name (abbrev): words	Total
Pronouns (pronoun): acontecimento*, algos, algures, aproximadamente, assim, assunto, assuntos, bens, bobagem, bobagens, bugiganga*, coisa, coisas, diferente, diferentes, fato, fatos, fêmea, idéia, idéias, issos, materiais, material, matéria, matérias, menina, moça, mulher, negócio, negócios, noção, noções, objeto, objetos, pertences, sozinha, sozinho, substância, substâncias, tolice*, traste*, troço, troços, tão;	44
Personal pronouns (ppron): bora, fêmea, menina, moça, mulher, sozinha, sozinho, vamos;	8
Impersonal pronouns (ipron): acontecimento, acontecimentos, algos, algures, alternada, alternadas, alternado, alternados, aproximadamente, assim, assunto, assuntos, bens, bobagem, bobagens, bugiganga*, coisa, coisas, diferente, diferentes, ela, ele, fato, fatos, idéia, idéias, lhe, materiais, material, matéria, matérias, negócio, negócios, ninguém, noção, noções, objeto, objetos, pertences, se, substância, substâncias, tolice, tolices, traste, trastes, troço, troços, tão.	49

Table 1. List of words improperly included in the Pronouns category in LIWC2007pt ("*" specifies word stems).

In a further step, we added some words in the pronoun category and its subcategories. Table 2 shows a selection of words in LOPI that exists in LIWC2007pt but that are not classified in the correct pronoun-related category. For the purpose of presenting relevant modifications in a concise form, words from LOPI⁵ that do not exist in LIWC2007pt and were included in PronounBP are not shown in this table.

Category Name (abbrev): words	Total
Pronouns (pronoun): algum, alguma, algumas, alguns, bem, cada, certa, certas, certo, certos, consigo, dada, dado, dele, desse, desses, deste, destes, determinada, determinado, disso, mais, mesmas, mesmos, muita, muitas, muito, muitos, nada, nenhum, nenhuma, nenhuma, nenhuns, onde, pouca*, poucas, pouco, poucos, quanto, semelhante, si, tanto, toda, todo, um, uma, umas, uns, várias, vários, vc, vcs;	52

⁵LOPI is currently available at <https://github.com/LaCAfe/LOPI>

Category Name (abbrev): words	Total
Personal pronouns (ppron): consigo, ele, si, vc, vcs;	5
Impersonal pronouns (ipron): algum, alguma, algumas, alguns, bem, cada, certa, certas, certo, certos, dada, dado, desse, desses, deste, destes, determinada, determinado, disso, mais, mesmas, mesmos, muitas, muito, muitos, nada, nenhum, nenhuma, nenhuma, nenhuns, ninguém, o, onde, pouca*, poucas, pouco, poucos, quanto, semelhante, tanto, toda, todo, um, uma, umas, uns, várias, vários.	48

Table 2. List of words in LIWC2007pt dictionary that were reassigned into categories related to pronouns in PronounBP ("*" specifies word stems).

Assigning the words into the categories according to not only their linguistic but also to their psychological and social processes is a very challenging task. Thus, we also decided to take advantage of all the years of exploratory study of emotional, cognitive and structural components of speech samples already made [18]. Therefore, we used both the previous LIWC English dictionary and LIWC2015 for comparison in this effort.

In order to initialize PronounBP dictionary, we removed all the incorrectly assigned words from the categories in which they were improperly included. Then, we added the remaining pronouns (i.e., the correct ones) to PronounBP dictionary. We can observe that, after this step, PronounBP contains 84 words in Pronouns (pronoun) category, since 44 words were removed from it. Moreover, PronounBP has the following number of words for the modified categories: 46 for Personal Pronouns (ppron), 6 for First Person Plural (we), 21 for Second Person (you), 10 for Third Person Singular (she/he) and 39 for Impersonal pronouns (ipron). It is important to mention that the First Person Singular (i) and Third Person Plural (they) categories have not been modified.

After removing the words improperly included in the pronouns categories, we proceeded to check words in LIWC2007pt that should be added to these categories. To achieve this goal, we developed a list with all words that should be included in pronouns categories, considering if they could be found in LIWC2007pt. For each word in this list, we checked if it could be founded in LIWC2007pt dictionary. If so, we included it in PronounBP dictionary, associating it with the right category. If not, we also included it in PronounBP dictionary, adding the association with the respective pronouns category.

As a result, PronounBP presents more words in the pronoun category than LIWC2007pt dictionary. While LIWC2007pt has 128 words, PronounBP has 136, even though 44 words were removed from pronoun category. Also, it is worth noting that, in PronounBP, even after the removal of many

words from improperly assigned categories, the number of words in other pronouns category remained almost the same in most cases, since other words from LIWC2007pt were also chosen for reassignment. In Table 3 we compare the number of words in pronouns categories for LIWC2007pt and PronounBP. After the creation of the PronounBP, we conducted an evaluation procedure similar to that performed for the Dutch translation of LIWC dictionary [29].

Category Name (abbrev):	LIWC2007pt	PronounBP
Pronouns (pronoun)	128	234
Personal pronouns (ppron)	54	79
First Person Singular (i)	7	9
First Person Plural (we)	8	11
Second Person (you)	25	36
Third Person Singular (shehe)	16	17
Third Person Plural (they)	11	15
Impersonal pronouns (ipron)	88	157

Table 3. Number of words in pronouns categories in LIWC2007pt and PronounBP.

DATASETS

In this work, we use two data sets to evaluate PronounBP. The first one is MQD190k and the second is CorpusV2. The following subsections describe both data sets.

MQD190k

We collected data from a Brazilian social network named Meu Querido Diário (MQD)⁶. We choose MQD data because users write predominantly in Brazilian Portuguese. The data set, named MQD190k, consists of 190,000 entries, divided into three classes as done in [27]: 10s, 20s and 30s, containing users of both genders with ages from 13 to 17, 23 to 27 and 33 to 42, respectively. The 10-year interval between the start of classes is intended for a clearer differentiation. It is important to note that this data set is available at <https://github.com/LaCAfe/MQD190k>.

CORPUSV2

We also collected a set of texts in Brazilian Portuguese and its English equivalents to submit to PronounBP and LIWC2007pt. We organized a bilingual (English and Brazilian Portuguese) set of texts, which we named CorpusV2, with 36 files consisting of text samples from the Bible, not only because they are easily available, but also because it is a sample of rigorously translated texts with some variety of styles [25]. CorpusV2 is available at <https://github.com/LaCAfe/CorpusV2>.

To obtain the 36 text samples, we accessed a page that allows reading different versions and translations of the Bible⁷. We selected 18 excerpts from the ‘The New International Version’ (NIV) in English and the equivalent 18 excerpts from the ‘Nova Versão Internacional’ (NVI) in Portuguese translation of the Bible for inclusion in CorpusV2.

EXPERIMENTS

To evaluate PronounBP, we compared PronounBP with LIWC2007pt in an experiment to measure their correlation with LIWC2015 pronouns categories. In this experiment, we processed CorpusV2 to create tables with percentage values of the relative frequencies of words for PronounBP, LIWC2007pt and LIWC2015.

After that, we calculated the correlation score for the corresponding columns, e.g., ipron percentage values in the Portuguese texts with ipron percentage values in the English texts. We used the Kendall correlation coefficient [11], considering that not all data are normally distributed, as we can confirm with the results of the Shapiro-Wilk Test for Goodness of Fit [13] applied to data from LIWC2015, PronounBP, and LIWC2007pt.

Next, we compared the correlation coefficients from LIWC2015 with PronounBP, and from LIWC2015 with LIWC2007pt. The comparison of Kendall’s τ correlation coefficients helps to evaluate the development of this work, as shown in Table 4. In the evaluation of the LIWC2007pt using CorpusV2, the correlation values range from 0.190 to 0.895. PronounBP values range from 0.438 to 0.895 and present higher values in 5 out of the 8 categories that were modified. Taking out the tied value 0.895, the highest value for PronounBP is 0.739, obtained in pronoun category.

Category Name (abbrev):	$\tau 1$	$\tau 2$
Pronouns (pronoun)	0.712	0.739
Personal pronouns (ppron)	0.608	0.739
First Person Singular (i)	0.895	0.895
First Person Plural (we)	0.568	0.739
Second Person (you)	0.190	0.438
Third Person Singular (shehe)	0.595	0.516
Third Person Plural (they)	0.515	0.451
Impersonal pronouns (ipron)	0.477	0.556

Table 4. Comparison between LIWC x LIWC2007pt ($\tau 1$) and LIWC2015 x PronounBP ($\tau 2$), by looking at Kendall’s τ correlation coefficient.

Finally, we produced a classification experiment to predict the age of users in MQD190k data set. We conducted the experiments with both PronounBP and LIWC2007pt. In

⁶<http://www.meuqueridodiario.com.br>

⁷<https://www.bible.com/>

order to generate the results, we tested five classification algorithms: Naive Bayes (NB), NB Multinomial (NBM), Decision Tree, LMT, and J48. We adopted Weka [9] to produce the experiments using each of the algorithms with its default configuration settings.

We performed experiments using the cross-validation technique named k-fold validation with ten partitions. We adopted the F1-score measure, a weighted harmonic mean of recall and precision [22], to evaluate the results. Recall and precision represent the probabilities that a randomly chosen relevant instance or, respectively, a randomly chosen predicted instance would be relevant [22]. Table 5 shows the results for F1-score. In PronounBP dictionary, one can note that all algorithms perform better than in LIWC2007pt.

Classification algorithm	LIWC2007pt	PronounBP
NB	0,496	0,508
MNB	0,506	0,512
DT	0,491	0,498
LMT	0,500	0,503
J48	0,503	0,507

Table 5. Classification algorithms' F1 Score from inference of the age group of users of MQD, using exclusively the pronouns category and its subcategories.

DISCUSSION

The research we present in this article brings as the main contribution the development of a dictionary of pronouns for Brazilian Portuguese to be used in LIWC, named PronounBP⁸. It also brings an empirical evaluation of PronounBP by performing a statistical comparison of percentage values in categories obtained for bilingual texts produced by different versions of LIWC dictionaries. Lastly, we evaluated PronounBP with a classification task using Text Mining techniques, comparing PronounBP with the available BP LIWC dictionary (LIWC2007pt).

The dictionary of pronouns for Brazilian Portuguese (PronounBP) has a total of 8 categories and can be used with the 2015 version of the LIWC program. As a result of including as many Portuguese pronouns as possible, PronounBP presented better results for age classification than LIWC2007pt. It is a factor that stimulates this study to adjust other words to categories that appropriately match linguistic and psychological characteristics.

We are aware that the results of other studies with the LIWC2007pt show that the performance in negative valence detection using affective categories is not so good [2]. Therefore, in future developments, the goal is to

prioritize efforts to improve the affective categories. We will also thoroughly review each of the function word categories, as well as all of the categories below it. It is also important to note that this article also contributes to the evaluation of pronoun categories in LIWC2007pt since other studies only evaluate the valence using affective categories.

REFERENCES

1. Charu C Aggarwal. 2011. An introduction to social network data analytics. *Social network data analytics*, 1–15.
2. Pedro P Balage Filho, Thiago AS Pardo, and Sandra M Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*. 215–219.
3. Jonathan Bright, Scott A Hale, Bharath Ganesh, Andrew Bulovsky, Helen Margetts, and Phil Howard. 2017. Does Campaigning on Social Media Make a Difference? Evidence from candidate use of Twitter during the 2015 and 2017 UK Elections. arXiv:1710.07087.
4. Josemar Alves Caetano, Hélder Seixas Lima, Mateus Freira dos Santos, and Humberto Torres Marques-Neto. 2017. Utilizando Análise de Sentimentos para Definição da Homofilia Política dos Usuários do Twitter durante a Eleição Presidencial Americana de 2016. In *VI Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*.
5. Flavio Carvalho, Gabriel Santos and Gustavo Paiva Guedes. 2018. AffectPT-br: an Affective Lexicon based on LIWC 2015. In *37th International Conference of the Chilean Computer Science Society (SCCC 2018)*, Santiago, Chile.
6. Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* 1, 343–359.
7. Pedro Parreira Cruz, Rafael Guimarães Rodrigues, Kele Teixeira Belloze, and Gustavo Paiva Guedes. 2017. Uma Revisão Sistemática sobre Léxicos Afetivos para o Português do Brasil. In *Congresso Internacional de Informática Educativa (TISE 2017)*, Fortaleza.
8. Sergio Davalos, Altaf Merchant, and Greg Rose. 2015. Using big data to study psychological constructs: Nostalgia on facebook. *Journal of Psychology & Psychotherapy* 5, 6, 1.
9. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1, 10–18.

⁸ PronounBP is currently available for download and use at <https://github.com/LaCAfe/PronounBP>

10. Chin-Lan Huang, Cindy K Chung, Natalie Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben CP Lam, Wei-Chuan Chen, Michael H Bond, and James W Pennebaker. 2012. The development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology* 54, 2, 185-201.
11. Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2, 81-93.
12. Celso Pedro Luft, LM Averbuck, JA de Menezes, AN Ew, and AMR Filipouski. 1997. *Novo manual de português, gramática, ortografia oficial, literatura, redação, textos e testes*. Globo.
13. Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253, 68-78.
14. Sílvia MW Moraes, André LL Santos, Matheus Redecker, Rackel M Machado, and Felipe R Meneguzzi. 2016. Comparing approaches to subjectivity classification: A study on portuguese tweets. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 86-94.
15. Nir Ofek, Lior Rokach, Cornelia Caragea, and John Yen. 2015. The Importance of Pronouns to Sentiment Analysis: Online Cancer Survivor Network Case Study. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 83-84.
16. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 37-44.
17. James W Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211, 2828, 42-45.
18. James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. *Technical Report*.
19. James W Pennebaker and Lori D Stone. 2003. Words of wisdom: Language use over the life span. *Journal of personality and social psychology* 85, 2, 291.
20. Terry F Pettijohn and Donald F Sacco Jr. 2009. The language of lyrics: An analysis of popular Billboard songs across conditions of social and economic threat. *Journal of Language and Social Psychology* 28, 3, 297-311.
21. Annie Piolat, Roger Booth, CK Chung, M Davids, and JW Pennebaker. 2011. The French dictionary for LIWC: Modalities of construction and examples of use| La version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. *Psychologie française* 56, 3, 145-159.
22. David MW Powers. 2015. What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. arXiv preprint arXiv:1503.06410.
23. Nairán Ramírez-Esparza, James W Pennebaker, Florencia Andrea García, Raquel Suriá Martínez, et al. 2007. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología* 24, 1, 85-99.
24. Julio CS Reis, Pollyanna Gonçalves, Matheus Araújo, Adriano CM Pereira, and Fabricio Benevenuto. 2015. Uma abordagem multilingue para análise de sentimentos. In *IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015)*.
25. Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities* 33, 1-2, 129-153.
26. Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18, 8, 1121-1133.
27. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6. 199-205.
28. Daisaku Shibata, Shoko Wakamiya, Ayae Kinoshita, and Eiji Aramaki. 2016. Detecting Japanese patients with Alzheimer's disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop*. 78-85.
29. Leon van Wissen and Peter Boot. 2017. An Electronic Translation of the LIWC Dictionary into Dutch. In: *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. *Lexical Computing*, 703-715.
30. Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao Zhu. 2014. Using linguistic features to estimate suicide probability of Chinese microblog users. In *International Conference on Human Centered Computing*. Springer, 549-559.
31. Maayan Zhitomirsky-Geffet, Esther David, Moshe Koppel, and Hodaya Uzan. 2016. Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites. *Online Information Review* 40, 3, 362- 379.