A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words

Nathan S. Hartmann, Lucas V. Avanço, Pedro P. Balage, Magali S. Duran, Maria G. V. Nunes, Thiago A. S. Pardo, Sandra M. Aluísio

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
nathanshartmann@gmail.com, avanco89@gmail.com, pedrobalage@gmail.com, magali.duran@uol.com.br,
gracan@icmc.usp.br, taspardo@icmc.usp.br, sandra@icmc.usp.br

Abstract

Web 2.0 has allowed a never imagined communication boom. With the widespread use of computational and mobile devices, anyone, in practically any language, may post comments in the web. As such, formal language is not necessarily used. In fact, in these communicative situations, language is marked by the absence of more complex syntactic structures and the presence of internet slang, with missing diacritics, repetitions of vowels, and the use of chat-speak style abbreviations, emoticons and colloquial expressions. Such language use poses severe new challenges for Natural Language Processing (NLP) tools and applications, which, so far, have focused on well-written texts. In this work, we report the construction of a large web corpus of product reviews in Brazilian Portuguese and the analysis of its lexical phenomena, which support the development of a lexical normalization tool for, in future work, subsidizing the use of standard NLP products for web opinion mining and summarization purposes.

Keywords: language in computer mediated communication, lexical normalization, lexical resources for Portuguese

1. Introduction

As it is widely known, corpus compilation and annotation is a very important task for research in practically any field of Natural Language Processing (NLP) and Computational Linguistics (CL), for both linguistic analysis and machine learning purposes. From the first individual initiatives, going through the use of web as corpus, to the current trends of crowdsourcing (largely made possible by Amazon Mechanical Turk), several challenges remain, as how to guarantee representativeness for the envisioned linguistic phenomena and how to select data and to train humans for reliably annotating it (see, e.g., Hovy et al., 2006; Hovy and Lavid, 2010). In addition, web has brought several new challenges regarding the quality of the collected material and language style in some specific communicative situations mediated by computers, which encompass talking at chats, social networks and other related Web 2.0 environments, in which tweets and SMS (Short Message Service) are pervasive.

Squires (2010) points out that, in these situations, language is marked by the absence of more complex syntactic structures (as subordinate constructions) and the presence of internet slang, with missing diacritics, repetitions of vowels, and the use of chat-speak style abbreviations, emoticons and colloquial expressions. Xue et al. (2011) reinforce that especially young users use this chat-speak style, resulting in a new form of writings that are very different from well-written texts.

Notwithstanding the success of such communication, it represents a severe challenge for most of the NLP and CL tools, which are mostly designed for dealing with well written texts, as early noticed by Ringlstetter et al. (2006). Several studies have tried to tackle the challenge of processing the language used in the web, mainly for dealing with out-of-vocabulary words in English texts, such as slang terms, acronyms and abbreviations used in websites, on Twitter, and Internet forums. For instance, several authors attempted to preprocess and normalize short text messages (or microtexts) (e.g., Han and Baldwin, 2011; Liu et al., 2011; Zhu et al., 2013), while others tried to automatically characterize the formality level of texts (see, e.g., Heylighen and Dewaele, 1999; Lahiri et al., 2011; Mosquera and Morada, 2011; Li et al., 2013). They usually apply simple rules, spelling checkers, and formal and informal (slang) language dependent dictionaries (e.g., NetLingo[1], NoSlang[2], and Internet Slang Words and Computer Slang[3]) to detect and correct internet slang and wrong words. On the other hand, Xue et al. (2011) propose a statistical approach to normalization based on the source channel model, in which four error models try to capture the way in which lexical variants are formed, namely, an orthographic factor, a phonetic factor, a contextual factor and acronym expansion.

Besides such problems, Portuguese product review texts also present recurrent use of technical jargon, mainly in English, related to electronics and computing, such as “cooler”, “wireless”, “home theater”, “webcam”, and “desktop” (see, e.g., a computer jargon glossary at www.jonstorm.com/glossary/).

To the best of our knowledge, there are no available tools and robust lexical resources for pre-processing all the phenomena of the language used in the web, more

[1] http://www.netlingo.com/dictionary/all.php
[2] http://www.noslang.com/dictionary/
[3] http://www.internetslang.com/

specifically, those posted in product review databases in Portuguese, although some theoretical discussions do exist (see, e.g., Bisognin (2008); Carvalho et al. (2009); Komesu and Tenani (2009)) as well as some restricted corpus linguistics initiatives (see, e.g., Gonzales, 2007).

The motivation of this work arose from the need to preprocess a Brazilian Portuguese web corpus constituted of product reviews, which will be used to train an opinion mining classifier and summarizer as end application. As our project includes the task of adding a layer of semantic role labeling (SRL) to the corpus and such annotation will be made over syntactic trees, we have to ensure that the layers of morphosyntactic and syntactic annotations be as similar as possible to those produced by taggers and parsers on well written texts.

In this paper we describe our methodological steps towards (i) the gathering of a web corpus of product reviews in Brazilian Portuguese, composed of relatively short texts carrying phenomena that occur in social media conversations, with errors and not following formal language structuring rules (Section 2); (ii) the corpus analysis and annotation process of the kinds of phenomena that give rise to the out-of-vocabulary words in such genre of texts in Portuguese (Section 3) and (iii) the compilation of six lexical resources made freely available at the website *Lexical Normalization of Product Reviews from the Web* [4], used to perform lexical normalization in order to allow further automatic processing by tools that require well written texts, e.g., taggers and parsers (Section 4). Finally, we present directions for future work in Section 5.

2. Corpus building and pre-processing

To build the corpus of product reviews, we have crawled databases from some of the most traditional online services in Portuguese, namely, Buscapé[5], Reclame Aqui[6] and Mercado Livre[7], where users post their pros and cons comments about several products, services and companies, besides reporting complaints about a company after a customer´s narrative about an incident. For this paper, we only report the results for Buscapé, a corpus large enough to present all the categories of problems with which we will have to deal.

We gathered 85,910 reviews from Buscapé database from a crawling in September 2013. They account for more than 681Mb, showing 4,097,905 tokens and 68,633 types. In general, the most frequent product categories in the reviews are (in descending order of frequency): TV, cell phones and smartphones, digital cameras, perfume, games, air conditioners, notebooks, and tablets. After removing stopwords, numbers and punctuation, the frequency list summed up 63,917 types.

As expected, the most frequent words are the inflections of the verb "ser" (to be), the nouns "produto" (product), "qualidade" (quality), and "preço" (price), and the adjective "bom" (good). Curiously, the least frequent ones include several ill-formed words, which are very common in such domain.

3. Corpus Analysis and Annotation

We carefully analyzed a large sample of the data in order to identify which kind of pre-processing tasks were necessary to be applied. Such analysis also provided input material for the lexical resources built to be used with the corpus. Therefore, to fulfill the requirements of both purposes, the sample had to consist of ill-formed words.

To select the sample to analyze, we first filtered out the well-formed words by matching them with a large and freely available lexicon for Portuguese, the Unitex-PB dictionary[8] (Muniz et al., 2005), composed approximately of 67,500 canonical forms, 880,000 inflected forms and 4,000 multiwords. Although the GNU Aspell Brazilian Portuguese dictionary [9] (307,726 canonical forms and 10,440,299 inflected forms) is bigger than Unitex-PB, we have decided to use an in-house dictionary in order to evaluate its weaknesses and be able to improve it during our ongoing project.

We obtained a list of unknown words containing 34,775 types, which was then processed with GNU Aspell running with the Brazilian Portuguese dictionary, in order to evaluate its benefits before applying the Palavras parser (Bick, 2000). To assess whether Aspell output words were correct or not, we matched them to the lexicon of Unitex-PB again. Aspell corrected 14,789 tokens (42,55%) of the 34,775 unknown words. Disregarding the correction of missing diacritic signals, which is the simplest case, Aspell corrected 12,004 (38,43%) of the tokens. The fact that the Aspell output words have not been attested in Unitex-Br does not mean they are any less correct. In fact, we have found a large number of diminutive words and computer jargon that is missing in Unitex-PB and set up a new project to enlarge this in-house dictionary.

Aiming to compare the results obtained by Aspell, we tested REGRA, the MS-Office spelling checker [10]. REGRA corrected 10.794 *tokens* (31,04%), thus 3,995 (11,51%) less than Aspell.

Next we measured the precision of the morphosyntactic and syntactic tags assigned by parser Palavras in nine texts from Buscapé database, comprising 369 tags. Figure 1 shows three of them (in Portuguese) to illustrate some problems faced by a parser based on well-written texts.

The parser precision was 83,73% in the original texts and raised to 84,28% after pre-processing the texts with Aspell. We noticed that the inconsistent capitalization was responsible for generating multiword proper names, since Palavras groups contiguous capitalized words into a single word. Palavras also depends on capitalization of

[4] http://nilc.icmc.usp.br/semanticnlp/LexicalNormalization
[5] http://www.buscape.com.br/
[6] http://www.reclameaqui.com.br/
[7] http://www.mercadolivre.com.br/
[8] www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html
[9] ftp://ftp.gnu.org/gnu/aspell/dict/0index.html#0.60
[10] released to NILC only for academic research, as NILC has been the first of its developers

acronyms in order to properly tag them.

Since the gain in performance was very low after using Aspell, we decided to thoroughly investigate the categories of the unknown words, beginning from the most frequent until the ones with 3 occurrences. We selected this range because the analysis of a sample of tokens with low frequency (2 occurrences) showed us how subjective the annotation task may become, as many ill-formed words are so spurious that it is hard to decide which tag to assign to them. The analysis consisted of a double blind annotation task that categorized 5,775 different tokens.

Found errors and their identification in the 3 texts below: abbreviation (AB), internet slang (IN), misspelling errors (X), acronym (SI), foreign word (ES)

ela e **[X: é]** muito escura quando vc **[AB: você]** esta [X: está] deitado se vc **[AB: você]** tiver **[IN: estiver]** sentado ela e **[X: é]** boa mas deitada nao **[X: não]** e **[X: é]** muito nao **[X: não]**. A Samsung inova o mercado de tv's **[AB: televisões]** com uma grande obra de arte que se adequa **[X: adéqua]** à qualquer ambiente. Esta tv **[AB: televisão]** possui excelente imagem quando ligada a uma fonte de dvd **[SI: DVD]** com hdmi **[SI: HDMI]** e na tv **[AB: televisão]** a cabo (Digital). O som é perfeito quando é personalizado pelo usuário. Ou seja, MENU, SOUD **[ES: SOUND]**, EQUALIZAR E ENTER **[ES]**.

Gostei dimais **[X: demais]** dessa câmera, comprei outra!Além de uma excelente e reconhecida marca, essa câmera tem um design **[ES]** super inovador e mtu **[IN: muito]** atraente...uma resolução mtu **[IN: muito]** boa e **[X: é]** td **[AB: tudo]** o que uma boa câmera SONY tem que ter!! Até hj **[AB: hoje]** nunca me deixou na mão... recomendo!!!

muto **[X: muito]** bom para manuziar **[X:manusear]** quando vc **[AB: você]** ta **[AB: está]** trabalhano **[X: trabalhando]** com este produto ,nao **[X: não]** tenho que recramar **[X: reclamar]** gostei mesmo parabéns. RECOMENDO O PRODUTO, FÁCIL DE USAR, ADOREI !

Figure 1: Examples of sentences with parsing problems.

For the annotation task, four pairs of annotators were selected. There were two training phases: the first with 100 tokens and the second with 75, which were individually annotated and then corrected with the help of the four pairs. After calibrating the annotation, each team received a set of 1,400 of the remaining 5,600 tokens. At all stages, the set of tokens assigned for annotation contained a proportional amount of the of the frequency distribution in the corpus.

An annotation manual (which was created to guide the annotation) describes and exemplifies (i) 8 categories (and corresponding tags) for the unknown words, (ii) the order to follow in their annotation, and (iii) several actions depending on the category. The categories are listed in what follows:

(1) Misspellings (X) include:

(1a) words with **missing diacritic** (e.g., the word “nao”, in Portuguese, which should be written as “não” - “not”, in English).

(1b) **spelling errors**, with missing or changed letters (e.g., “exelente” instead of “excelente” - “excellent”, in English; and “imagems”, instead of “imagens” - “images”, in English).

(1c) **correct words** not included in the Unitex-PB dictionary.

(2) Acronyms (SI) include:

acronyms, which could not be filtered out by the Unitex-PB dictionary, such as HDTVi and Wi-Fi, appearing with and without uppercase letters.

(3) Proper Names (NP) include:

use of **proper nouns**, mainly for companies and products that could not be filtered out by the Unitex-PB dictionary, appearing with and without initial uppercase letter, for example, “android” and “windows”.

(4) Abbreviations (AB) include:

abbreviations introduced by chat-speak style and measures of already established abbreviations of common words. An arbitrary rule should be followed: abbreviations must have the same first letter of the abbreviated word. The abbreviation is always a single word, while an acronym involves multiple words. For example, in Portuguese, it is very frequent to use “vc” instead of “você” (“you”, in English), “tbm” instead of “também” (“in addition/too”, in English).

(5) Internet Slang terms (IN) include:

language used in the web, i.e., slang terms, (e.g., in Portuguese, it is very frequent to use “naum” instead of “não” (“not”, in English), “produtu” instead of “produto” (“product”, in English), adoreiiiiiiiii instead of “adorei” (“I loved it”, in English). One of its features is to try to mimic the spoken language. It is usual to find letters that originally did not exist in the word they represent.

(6) Foreign Words (ES) include:

(6a) use of **foreign words from English** (e.g., “touch”).

(6b) use of **foreign words from languages other than English** (French and Spanish).

(6c) tokens that could be **truncated words or part of multiword expressions,** such as “blu”, which could be part of “blue-ray”.

They are common words in foreign languages that were borrowed for use in Portuguese.

(7) Units of Measure (UM) include:

a name used to express a **quantity**, such as “gigabytes”. They not include abbreviations of names of units of measure.

(8) Unrated (SC) or **doubtful tokens,** meaning that:

(8a) we were not able to assign any of the seven categories above.

(8b) the word has a different category in each presented context (ambiguity). For example, the word “oops” can be tagged with IN or NP (“Black Ops II”).

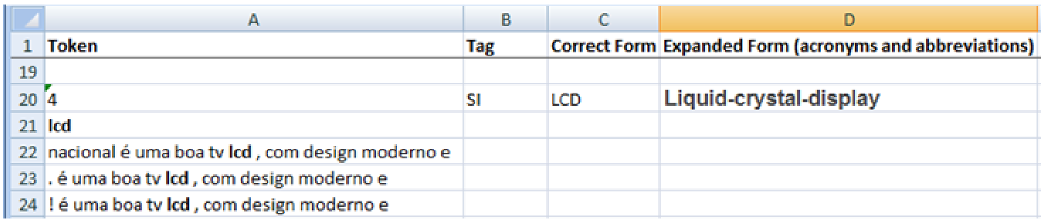
(8c) the word is part of a different multiword in each presented context. For example, “high” may appear in “high-quality”, “high-end” and “high-tech”.

To provide a context for meaningful annotation, 3 token occurrences (concordances) taken from the corpus texts were brought below the token to be tagged. Each token was analyzed with the support of Wikipedia, a web search engine and the output of Aspell, in order to understand the kind of occurrences and errors were made. The category Unrated (SC) requires further investigation, supported by more examples of corpus occurrences to solve their ambiguities.

Since many words may belong to two or more categories at the same time, we decided to establish an order of precedence among the categories. The order for the 8 tags was the following: (1) Misspellings (X), (2) Acronym (SI), (3) Proper Name (NP), (4) Abbreviation (AB), (5) Internet Slang (IN), (6) Foreign Word (ES), (7) Unit of Measure (UM) , and (8) Unrated (SC) for doubtful tokens. For example, the word “HTC” (High Tech Computer Corporation) is an acronym, a proper name, and is composed of foreign words. Since Acronyms (SI) have precedence over Proper Names (NP) and Foreign Words (ES), the word should be classified as an acronym. On the other hand, if it appears as “High Tech Computer Corporation”, this would be a Proper Name (NP).

With regards to actions, Acronyms had their case corrected and Abbreviations were expanded and had their case corrected; Proper Names had their case corrected; Internet Slang terms were translated into their corrected form; Foreign Words and Units of Measure had their form corrected, if necessary, and were capitalized. The last two actions were needed to improve the precision of the Palavras parser, since by default, the parser tags as verb any word not found in its lexicon. Figure 2 shows an entry of the list of tokens to be annotated in which the token “lcd” was tagged as Acronym (SI), corrected to “LCD", and expanded to “Liquid-crystal-display”. The way the words were annotated aimed to facilitate their future use in the tools designed to provide lexical normalization.

The Kappa agreement measure was computed for each pair of annotators, as Table 1 shows. The proportion of the 8 categories appearing in the set of annotation is also shown.



	A	B	C	D
1	Token	Tag	Correct Form	Expanded Form (acronyms and abbreviations)
19				
20	4	SI	LCD	Liquid-crystal-display
21	lcd			
22	nacional é uma boa tv lcd , com design moderno e			
23	. é uma boa tv lcd , com design moderno e			
24	! é uma boa tv lcd , com design moderno e			

Figure 2: Example of occurrences where a token to be annotated appears.

		Categories							
Pair	**Kappa**	**X**	**AB**	**ES**	**IN**	**NP**	**SC**	**SI**	**UM**
1	0,737	42,6%	2,5%	5,9%	3,3%	19,0%	21,8%	4,3%	0,7%
2	0,714	44,1%	2,4%	6,3%	3,2%	27,7%	11,6%	4,2%	0,5%
3	0,801	45,1%	1,9%	7,8%	5,0%	23,4%	11,6%	5,1%	0,2%
4	0,757	45,5%	1,9%	9,5%	4,8%	21,0%	13,0%	4,0%	0,4%
Average	0.752	44,33%	2,18%	7,38%	4,08%	22,78%	14,5%	4,4%	0,45%

Table 1: Kappa values and distribution of the 8 categories.

The average Kappa is 0.752, considered a substantial agreement (Carletta, 1996). To calculate the proportion of each category, we first calculated the average between the individual annotations in the category. In Table 1, one may notice that the highest concentration of errors is of type X.

Since the identification of spelling errors is more natural than the other categories to literate people, we removed annotation instances for which there was agreement in the annotation of errors of type X and computed again the Kappa. Table 2 shows the new values.

		Categories							
Pair	**Kappa**	**X**	**AB**	**ES**	**IN**	**NP**	**SC**	**SI**	**UM**
1	0,572	4,2%	4,2%	9,8%	5,5%	31,7%	36,3%	7,2%	1,1%
2	0,512	4,1%	4,1%	10,8%	5,5%	47,5%	19,9%	7,2%	0,9%
3	0,670	4,1%	3,3%	13,6%	8,7%	40,8%	20,2%	8,9%	0,4%
4	0,603	4,7%	3,3%	16,5%	8,4%	36,8%	22,7%	6,9%	0,7%
Average	0,589	4,3%	3,7%	12,7%	7,0%	39,2%	24,8%	7,6%	0,8%

Table 1: Kappa values and distribution of categories without considering when both annotators agreed on X.

In Table 2, we may see a small number of instances of the type X, on which there was disagreement in the annotation. In fact, the category X has a bias that tends to increase the value of the Kappa measure. After removing the instances where both annotators agreed on X, we have an average Kappa value of 0.589, which is considered moderate, but lower than the previous value in Table 1.

From the average values of Table 1, one may conclude that the main challenges for processing such kind of corpora include properly processing: (i) spelling errors, (ii) proper names of companies and products, (iii) foreign words (most from English), (iii) acronyms related to the domain of the products reviewed, and (v) the language used in the web.

4. The Lexical Normalization Website

Each of the five analyzed problems requires different solutions in order to increase the likelihood that the text files of the products reviews database be successfully parsed. As a first step towards lexical normalization of product reviews, we developed a workbench called *Lexical Normalization of Product Reviews from the Web*, bringing together six lexical resources (in UTF-8 encoding) compiled from the annotation reported in Section 3, which are freely downloadable. It contains: 172 Acronyms (SI); 1,023 Proper Names, mainly of companies and products; 77 Abbreviations (AB); 165 Internet Slang terms (IN); 265 Foreign Words (ES) and 12 Units of Measure (UM).

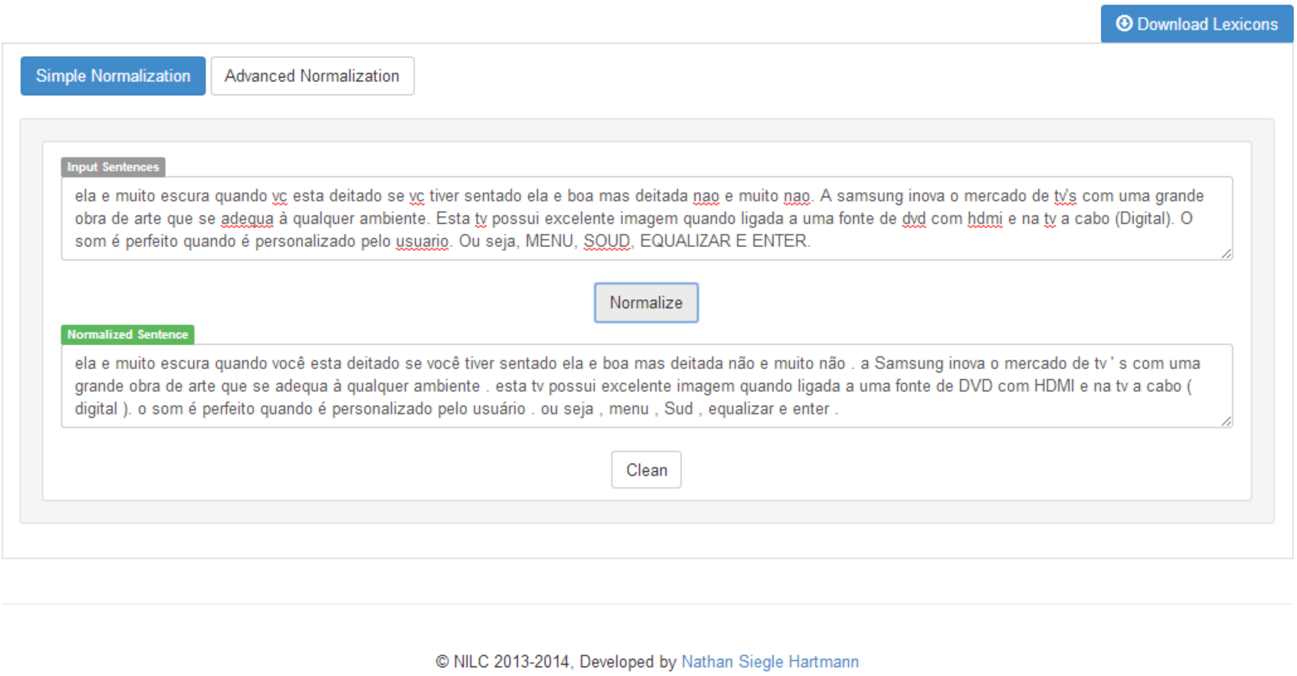
In this workbench, one may also evaluate the order of application of the dictionaries. We have used the following order to manually annotate tokens: (1) Misspellings (X), (2) Acronym (SI), (3) Proper Name (NP), (4) Abbreviation (AB), (5) Internet Slang (IN), (6) Foreign Word (ES), (7) Unit of Measure (UM) , and (8) Unrated (SC) for doubtful tokens. However, we believe that, in automatic application of a cascade process, the Aspell spelling checker should be the last resource applied, since it may improperly correct some words and "mask" the real problems, jeopardizing the action of the other lexical resources in the sequence In addition, in order to "protect" the words already recognized and corrected by our lexical resources, we plan to restrain the use of Aspell only for those words that did not suffer previous revisions.

The website makes available a toolkit for lexical normalization that works in two modes:

(i) the Simple Normalization mode applies a fixed number of lexical resources in a fixed order: Acronym, Proper Name, Abbreviation, Internet Slang, Foreign Word, Unit of Measure and finally the spelling checker Aspell.

(ii) the Advanced Normalization mode allows the user to choose the dictionaries and the order in which they shall be applied. This mode also includes the spelling checker Aspell and enables the visualization of the parsing performed by Palavras on the revised text.

Figure 3 shows a screen dump of the Simple Normalization mode, where a normalized product review may be seen.



ela e muito escura quando vc esta deitado se vc tiver sentado ela e boa mas deitada nao e muito nao. A samsung inova o mercado de tv's com uma grande obra de arte que se adequa à qualquer ambiente. Esta tv possui excelente imagem quando ligada a uma fonte de dvd com hdmi e na tv a cabo (Digital). O som é perfeito quando é personalizado pelo usuario. Ou seja, MENU, SOUD, EQUALIZAR E ENTER.

ela e muito escura quando você esta deitado se você tiver sentado ela e boa mas deitada não e muito não . a Samsung inova o mercado de tv ' s com uma grande obra de arte que se adequa à qualquer ambiente . esta tv possui excelente imagem quando ligada a uma fonte de DVD com HDMI e na tv a cabo (digital). o som é perfeito quando é personalizado pelo usuário . ou seja , menu , Sud , equalizar e enter .

© NILC 2013-2014, Developed by Nathan Siegle Hartmann

Figure 3: Example of a normalized product review using the mode "Simple Normalization".

5. Conclusion and Future Work

Handling out-of-vocabulary words, which, in this case, comes mainly from web writing, is only the first step towards proper language normalization. Several challenges remain.

In order to automatically correct some ill-formed proper names, a named entity recognition system need to be used. However, to match a correct proper name with its deviation is not an easy task. There are studies working with artificial generation of errors that may help to measure how similar they are in order to indicate some correction (see, e.g., Foster and Oistein (2009) and Dickinson (2010)). We believe it is an interesting path to pursue. We plan to start with the list of single proper names generated by our manual analysis. This list should be extended to consider multiword expressions. The toolkit mwetoolkit (Ramisch et al., 2010) may be used, in this case.

Regarding acronyms, the solution is similar to proper names, since we just have to turn them into uppercase. As for the language used in the web, by merging the results of our corpus analysis and the scarce resources we found for Portuguese, we may produce a larger dictionary of internet slang for Portuguese. Such dictionary is essential for properly pre-processing web language and should be enlarged to treat the dynamic nature of this form of expression that mix written and spoken register, besides creating a proper form of communication.

To identify English foreign words in our corpus related to the domain of the products reviewed and to correct spelling errors in English and Portuguese, we foresee the use of two spelling checkers: one for Portuguese and the other for English, since we are dealing with texts that bring words in both languages.

Finally, it is worth mentioning that a full normalization goes beyond lexical adaptations. In particular, it is apparent that, in several cases, syntactical transformations are necessary to produce well-formed sentences. However, in this paper, we have not dealt with this subject, which remains for future work.

6. Acknowledgements

This paper results from an academic agreement between University of São Paulo and Samsung Eletrônica da Amazônia Ltda, to whom the authors are grateful. The authors are also grateful to FAPESP and CNPq for supporting this work.

7. References

Bick, E. (2000). The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 505p. Ph.D. Thesis (Philosophy) – University of Aarhus, Denmark.

Bisognin, T. (2008) Do internetês ao léxico da escrita dos jovens no Orkut. Dissertação de Mestrado, 2008. Universidade Federal do Rio Grande do Sul. Instituto de Letras. Programa de Pós-Graduação em Letras. Available at http://hdl.handle.net/10183/14385. (In Portuguese).

Carletta, J. (1996) Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, vol. 22, n. 2, pp. 249--254.

Carvalho, D. B.; Nogueira, L. A. and Cabral, C. P. (2009) A LINGUAGEM ESCRITA DOS ADOLESCENTES FACE ÀS TECNOLOGIAS DA COMUNICAÇÃO. Available at: http://www.perspectivasonline.com.br/revista/2009vol3n12/volume3%2812%29artigo10.pdf (In Portuguese).

Dickinson, M. (2010) Generating Learner-Like Morphological Errors in Russian. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 259--267, Beijing, August, 2010.

Foster, J. and Oistein, A. E. (2009). GenERRate: Generic Errors for Use in Grammatical Error Detection. In Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications, pp. 82--90, Boulder, Colorado, June 2009.

Gonzalez, Z. M. G. (2007). Linguística de corpus na análise do internetês. Master Thesis in Linguistics, Catholic University of São Paulo, 123 p. (In Portuguese).

Ramisch, C.; Villavicencio, A. and Boitet, C. (2010) Multiword Expressions in the wild? The mwetoolkit comes in handy. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010) — Demonstrations, pp. 57--60, Beijing, China, August 2010.

Han, B. and Baldwin, T. (2011) Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 368--378, Portland, Oregon, June 19-24, 2011. Association for Computational Linguistics.

Heylighen, F. and Dewaele, J. (1999) Formality of Language: definition, measurement and behavioral determinants. Internal Report, Center "Leo Apostel", Free University of Brussels, 1999. Available at: http://pespmc1.vub.ac.be/papers/Formality.pdf.

Hovy, E. H. and Lavid, J. M. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation* 22: 1, Jan-Dec. 2010: 13--36.

Komesu, F. and Tenani, L. (2009). CONSIDERAÇÕES SOBRE O CONCEITO DE "INTERNETÊS" NOS ESTUDOS DA LINGUAGEM. In: Linguagem em (Dis)curso, Palhoça, SC, v. 9, n. 3, p. 621--643, set./dez. 2009. (In Portuguese). Available at: http://www.scielo.br/pdf/ld/v9n3/10.pdf.

Lahiri, S.; Mitra, P., and Lu, X. (2011). Informality judgment at sentence level and experiments with formality score. Computational Linguistics and Intelligent Text Processing *Lecture Notes in Computer Science* Volume 6609, 2011, pp. 446--457. Springer.

Li, H.; Cai, Z. and Graesser, A. C. (2013). Comparing Two Measures for Formality. In Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, pp. 220--225, 2013.

Liu, F.; Weng, F.; Wang, B. and Liu, Y. (2011). Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pp. 71--76, Portland, Oregon, June 19-24, 2011. Association for Computational Linguistics.

Mosquera, A. and Moreda, P. (2011). The Use of Metrics for Measuring Informality Levels in Web 2.0 Texts. In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pp. 184--188, Cuiabá, MT, Brazil, October 24-26, 2011. Sociedade Brasileira de Computação.

Muniz, M. C. M.; Nunes, M. G. V. and Laporte, E. (2005) UNITEX-PB, a set of flexible language resources for

Brazilian Portuguese. In Proceedings of the III Workshop on Technology of Information and Human Language (TIL), pp. 2059--2068, São Leopoldo, RS, Brazil, July 22-29, 2005. Available at http://www.nilc.icmc.usp.br/til/til2005/arq0102.pdf

Ringlstetter, C.; Schulz, K. U. and Mihov, S. (2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. *Computational Linguistics* Volume 32, Number 3, pp. 295--340. Association for Computational Linguistics.

Squires, L. (2010). Enregistering internet language. *Language in Society*, 39(04):457–492.

Xue, Z.; Yin, D.; Davison, B.D. and Davison, B.: Normalizing Microtext. In Proceedings of the 2011 AAAI, pp. 74--79. Association for the Advancement of Artificial Intelligence (2011).