Identification of Patterns and Document Ranking of Internet Texts: A Frequency-based Approach

Marcos Zampieri[1], Jürgen Hermes[2], Stephan Schwiebert[2]
Romance Philology Department[1]
Institute for Linguistics[2]
University of Cologne

Abstract

This paper presents an automatic method to process Portuguese internet data using the Text Engineering Software Laboratory, Tesla. The experiments presented here were carried out in two stages. On the first step, frequency information was used to identify salient lexical and orthographical features in internet language in comparison to a standard journalistic corpus. The results obtained in the first step were later used to rank texts according to the their features from standard to non-standard language. Results are presented and discussed.

1 Introduction

The Internet is seen as a huge repository of textual material used by (computational) linguists for corpus compilation and linguistic analysis. Researchers use internet data to compile large-sized corpora and then they investigate these corpora considering different aspects of human language such as syntax and morphology.

The research question, depends on the linguistic data. The Internet contains, on one hand, linguistic data published by on-line magazines, newspapers and news agencies aiming at general readership. These texts pass through an editorial process to ensure that the material contains standard language, similar to books and other printed media. On the other hand, the Internet also contains a vast amount of user-generated content published in *blogs*, *forums* and other computer-mediated contexts. The language used in computer-mediated communication possess unique linguistic features that will be explored in this paper.

Collecting and processing internet data for corpus compilation is not a trivial task. As will be discussed in section 1.1, user-generated content is non-standard in terms of lexicon, syntax and above all orthography. NLP tools such as tokenizers, part-of-speech taggers [1] and syntactic parsers [2] are designed to process standard contemporary language. When these resources are used on non-standard data, their performance is substantially worse.

Even though many challenges arise when processing non-standard data, it is, to a certain extent, understood that the Internet can be used to compile material for corpora. The same is not true, however, for the use of the Internet itself as a corpus. In this case, several methodological issues arise and the question of size, balance, sampling and representativeness of linguistic data [3] should be taken into account. Examples of tools to work with the web as a corpus include WebCorp [4].

As previously mentioned, the second aspect of using Internet data is the linguistic description of computer-mediated communication itself [5], [6]. This comprises the analysis of its most important linguistic features such as spelling and grammar, but also the study of human interaction. For this reason, the description of computer-mediated communication is interest not only to linguists but also to psychologists and cognitive scientists.

In this work, we aim to contribute to both of these aspects, corpus compilation and processing, and the linguistic analysis of user-generated content. Firstly, we discuss methodological issues of processing internet texts using the Text Engineering Software Laboratory, Tesla [7] and propose a new corpus-driven method for this task. Secondly, by applying these methods, we provide empirical evidence for the description of different aspects of Brazilian Portuguese internet language (e.g. lexicon, orthography). To our best knowledge, this work is the first attempt at applying a fully automatic frequency-based method to identify patterns in Brazilian Portuguese Internet data.

1.1 Internet Language: A Substandard?

Koch and Österreicher's model [8] from written and spoken language can also be used to explain phenomena related to internet language. Their model see communication as a continuum ranging from near *(Nähe)* to far *(Distanz)*. In *Nähe*, communication is done by face-to-face interactions and is driven by spontaneity and involvement, whereas *Distanz* communication acts are usually monologues and far less spontaneous.

Based on this model, it is possible to say that the discourse features of face-to-face speech diverge largely from those associated with traditional writing. Face-to-face speech is commonly dialogic, although in practice there may be multiple participants. Either way, the speaker knows who the participants are, and this kind of communication allows for immediate feedback. The properties of face-to-face spoken discourse versus traditional writing are important for understanding the kind of natural language used on the Internet, because it lies somewhere in the continuum between *Nähe* and *Distanz*.

There were a number of systematic attempts to describe the language used on the Internet such as Baron [9] and Crystal [10]. Crystal aims to investigate the main properties of internet language and coined the term *Netspeak*. The term itself is no longer widely used in the research community, but Crystal's contribution to the study of these phenomena still remains substantial.

It should be noted that there are several means and types of communication on the Internet and each of them presents different properties. Some examples of current internet communication include: *forums*, *e-mail*, *instant messages* and *blogs*. These media of communication might share common linguistic features that allow one to group them as representing a substandard language in respect to the standard language system, but in our opinion, each of them should be studied individually.

In this work we focus on one kind of interaction. The data processed here was retrieved from a question and answer forum from the Brazilian *Yahoo!* called *Yahoo! Perguntas*. The methods used can be applied to other types of computer-mediated communication, but the findings should be restricted to this particular medium.

1.2 Related Work and Motivation

This study was designed to fill a gap created by the lack of corpus-driven approaches to the study of internet language. There were a number of attempts to study internet language phenomena such as the aforementioned [9] and [10] but these studies are rather descriptive and not corpus-driven.

We believe that the lack of corpus-driven studies is due to the shortage of linguistic resources available. The compilation of internet corpora is a relatively recent phenomenon and poses many more difficulties to researchers by comparison to standard contemporary language. Some resources available include multilingual internet corpora compiled by Sharoff [11] and a couple of corpora available available on Sketch Engine [12], including a recently compiled Portuguese corpus. For the compilation of Internet corpora, computational tools such as the WebBootCat [13] help researchers to retrieve and prepare data (e.g. removing HTML tags and metadata).

Our study has two main goals. The first goal is to use probability estimation to calculate salient lexical features of the internet corpus. These salient features are calculated using the log likelihood estimation as described by Dunning [14]. The product of this calculation is represented by a keyword list, widely used in corpus linguistics [15]. The keyword list is calculated between a study and a reference corpus. In our experiments, we used the internet corpus as the study corpus and a standard contemporary journalistic corpus as the reference corpus. More on that in section 2.2.1.

Keywords were applied to several corpus-based and driven research topics ranging from terminology to discourse analysis [16], [17] and [18]. In this study, the salient features should be words that have different orthography if compared to the standard orthography, as well as lexical choices that are related to the medium (internet forums) or the kind of discourse.

Our second objective is to automatically rank texts based on their degree of *internetness*. This degree of *internetness* orders texts from standard (in our experimental setting represented by journalistic) to non-standard language as described in section 2.2.2. This document ranking may be useful to identify users that show preference for non-standard spelling or lexicon, or it may be used as a first step of a more sophisticated text classification task.

A known shortcoming of this approach is that there is no gold standard available for evaluation. Unlike other NLP tasks such as POS Tagging, Parsing or Text Classification, it is not possible to evaluate how accurate the method is on a purely quantitative basis. The evaluation should ideally take into account a careful linguistic analysis of the ranked tasks and the keyword list.

2 Methods

We compiled two Brazilian Portuguese corpora for our experiments. The reference corpus was compiled using material from *Folha de São Paulo* published in 2004. An internet corpus was collected in March 2012 by retrieving six threads of the Brazilian *Yahoo Perguntas*. An overview of these corpora is presented in section 2.1.

With these corpora, we carried out the main computational processing using Tesla components. Before using Tesla components a short preprocessing stage was carried out using Python scripts to remove URLs, user names, tags and other meta information. Although copyright restrictions do not apply to most public internet data, privacy is an important issue and we opted to anonymize all user names.

The pre-processing step prepared texts to be used in Tesla. Tesla components were then used to tokenize texts and identify its sentences and paragraphs. Subsequently, we used Tesla to calculate two frequency lists, one for each corpus. After these steps, two new components were developed in Tesla's infrastructure: the *Corpus Log Likelihood* and the *Document Rank* component. The first calculates the most important keywords in the study corpus in comparison to the standard corpus and the latter ranks texts on the study corpus from standard to non-standard.

2.1 Corpora

An overview of the two corpora used in the experiments is presented next:

Type	Source	Tokens	Texts
Journalistic	Folha de São Paulo	10,542,594	128,864
Internet	Yahoo Perguntas (Brazil)	969,805	15,605

Table 1: Corpora

According to what was described by Berber Sardinha[16] for keyword calculation, the reference corpus should be at least 5 times bigger than the study corpus. In our example, the reference corpus is more than ten times bigger than the study corpus.

2.2 Text Engineering Software Laboratory

The *Text Engineering Software Laboratory* (Tesla[1]) is a component framework designed for experimental computational linguistics. It shares several features with other frameworks like GATE[19] and UIMA[20], and it is used primarily in science, research, and prototyping.

Tesla uses an object-oriented annotation concept, the *Tesla Role System* [7] which allows developers to define linguistic roles by extending two Java-interfaces which determine the methods an annotation offers, and the way such annotations can be retrieved. A Tesla component consumes and produces one or more roles, allowing users to define workflows with compatible Java-interfaces only. The major benefit of the TRS is that developers can use the full coverage of the Java programming language to specify data structures (including parametrized methods, inheritance, etc), such that they can focus on the development of problem-specific algorithms, without the need to consider framework-related restrictions.

[1] http://tesla.spinfo.uni-koeln.de

Workflows generated with Tesla are virtual experiments, which can easily be published, adopted or modified to tackle related tasks. This simplifies the scientific exchange not only within work groups, but also through specialized online platforms, such as MyExperiment[2].

Tesla makes use of a client-server architecture: Language resources (such as corpora or dictionaries) are stored on server-side, where experiments are executed as well. The client consists of several plug-ins, which enhance the Eclipse[3] framework with several Tesla-specific views and functions. These include a graphical experiment editor, views to manage corpora, roles, components and experiments, and evaluation views, which provide different visualization mechanisms to analyze the results of an experiment (highlighted or bracketed text, tables, or a word cloud), or to export the results.

2.2.1 Corpus Log Likelihood Component

The *Corpus Log Likelihood* component takes the token frequencies of two corpora (one of them is assumed to be the reference/journalistic corpus, the other to be the study/internet corpus) as starting point and applies the *log likelihood ratio* (*LLR*)[4] [14] to compute the keywordness-score of all tokens (alternative: token n-grams) belonging to the study corpus.

The calculation of the scores is based on row and column sums of the following table:

	Tokens J	Tokens $\neg J$	
Tokens I	$I \wedge J$	$I \wedge \neg J$	Row 1 sum
Tokens $\neg I$	$\neg I \wedge J$	$\neg I \wedge \neg J$	Row 2 sum
	Column 1 sum	Column 2 sum	

Table 2: Table of token occurences

Tokens can occur in the journalistic corpus (J), in the internet corpus (I)

[2] http://www.myexperiment.org/workflows?query=tesla

[3] http://www.eclipse.org

[4] The Tesla component uses the implementation of the Apache Mahout Project, http://mahout.apache.org/

or in both. To compute the *squaredLLR* (also known as G), Dunning uses the Shannon Entropy (H), defined as $H = \sum_{i=1}^{n} p(x_i) * log(p(x_i))$; *TCount* is the absolute count of tokens in both corpora:

$$G = \sqrt{2 * TCount * (H(matrix) - H(rows.sums) - H(col.sums))} \quad (1)$$

The types were ranked according to their G scores. After running the experiment, the rankings can be studied in an export result table (each in csv and a LaTeX format). The G scores were also part of the input to the component described next.

2.2.2 Document Rank Component

The *Document Rank* component calculates the degree of a document's belonging to a special corpus. (*Dob*) is computed in a very simple manner by summing up the G scores (computed by the component described above) of each token contained in the document and dividing this sum by the length of the document:

$$Dob = \frac{\sum_{i=1}^{n} G(x_i)}{n} \quad (2)$$

The documents were ranked according to their *Dob* values and listed in a result table, that can be studied by the user or be taken as input for other components. If the study corpus consists of internet texts and the reference corpus consists of journalistic texts, high *Dob* values indicates documents with strong relationship to a feature we call *internetness*.

2.3 Experiment Setting

Tesla has a user-friendly interface that allows users to integrate and configure resources and components very quickly. The experimental setting explained briefly in section 2 can be best understood by looking at image number 1.

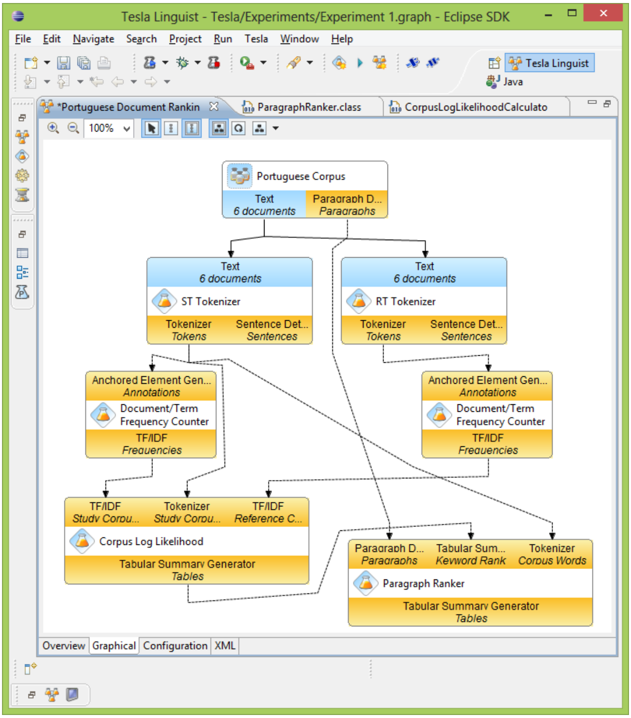

Figure 1: Tesla Experiment

In this figure, the two corpora are integrated to 6 different component boxes: 2 tokenizers (one for each corpus), 2 TF/IDF components and the new *Corpus Log Likelihood* and *Document Ranker*.

3 Results

The results of the experiments are presented in two tables. Table 3 presents the results of the log likelihood calculation for keywords and table 4 the first 15 ranked documents.

In table 3, the first ten keywords are listed as well as words that were ranked between 101st and 110th position as an example. The first column contains the position of the given type in the list, second column contains the type itself, and the third column presents the LL value for each token.

Rank	Type	LL Value	Rank	Type	LL Value
1	eu	163.20	101	oq	35.14
2	pão	132.95	102	esta	34.95
3	vc	123.36	103	youtube	34.62
4	pra	115.86	104	assistir	34.20
5	nao	103.73	105	Bleach	34.10
6	musica	96.20	106	tu	34.08
7	me	79.19	107	olha	33.99
8	você	78.09	108	tava	33.85
9	anime	75.85	109	note	33.79
10	sei	75.36	110	ver	33.70

Table 3: Wordlist: Log Likelihood

In table 4, we present the documents ranked from the 1st to the 15th position and the table has the same structure as the previous 3 table.

R	LL Value	Text
1	72.51	Eu tenho pq eu sou menina kkkkkkkkkkkkkkkâĂę Mas eu adoro rock :)
2	67.64	pão eu não como prefiro cerveja cerveja cerveja cerveja cerveja cerveja cerveja
3	61.32	OTAKUS !! QUAL O MELHOR ANIME SO PODE 1? qual o melhor anime pra vc?
4	59.85	eu gosto muito qual musica você mais gosta deles a minha é essa
5	58.72	Sei lá, eu poderia ser o Gaara. Mas mesmo assim, se eu pudesse ser eu, eu seria eu!
6	58.58	Cara eu só sei do EVA(Evangelion..) Eu ja vi o anime e adorei.
7	58.55	Qual a melhor musica de rebelde *-*? Eu gosto mais de "vc é o melhor pra mim"
8	58.07	Guns n Roses Guns n Roses Guns n Roses Guns n Roses Guns n Roses
9	57.71	Pão é muito bão... kkkkkkk, eu gosto muito de pão de queijo!!!
10	57.59	eu nao gosto da banda, eu gosto da musica, se ela for musica de merda eu paro
11	57.54	Eu respeito mto essa banda,mais nao gosto mto... Nao sei te explicar pq...
12	57.35	ta ai o melhor site pra baixar series q eu ja vi aproveite xD
13	57.21	eu também acho...mas mesmo assim ela é linda e eu adoro ela
14	57.20	Pra falar a verdade eu não lembro. Mas o filme é muito bom.
15	56.95	Sei 2 mas prefiro nao citar nomes pra nao arrumar pra mim...

Table 4: Top 15 Documents: Document ranker

After a few attempts, we arbitrarily set the minimum length of texts to 10 tokens. In our experiments, the first positions of the list using shorter texts (e.g. maximum 5 tokens) often presented repetitions of high ranked words. These texts are not interesting for linguistic analysis, as they reflect users' personal idiosyncrasies, rather than properties of the studied text type. Using larger texts tends to diminish this bias, although repetitions may appear as can be seen in texts number 2 and 8.

3.1 Discussion

The tables presented in the last section provided the reader a snapshot of the obtained results. For a more exhaustive analysis, Tesla allows tables to be exported with a larger number of rows, providing more data for linguistic analysis.

The 20 words in table 3 reflect different phenomena related to computer mediated communication. Spelling variation occurs in Portuguese with the absence of graphical signs such as in *não* - *nao* or in *música* - *musica*. Portuguese internet forms are often formed by the suppression of vowels to make typing faster and easier, these cases include *você* - *vc* or *o que* - *oq*.

Some lexical and lexico-syntactic choices related to the medium and text type may also be observed such as the case of the first person singular pronouns *eu* - *me*, which do not appear often in written journalistic texts [8]. Another aspect related to the medium are foreign words that might appear on the Internet such as *youtube* or *Bleach* but not often in journalistic prose.

As for the ranked documents, we may observe the aforementioned predominance of the first person singular discourse, most texts in this list were written in this form. Another feature that can be observed is the use of emoticons and concatenations of the letter *K*, which is the Portuguese equivalent to the English *LOL*, meaning laughing.

The results obtained in both tables call for a careful linguistic analysis. We aim to use the information obtained in these experiments to provide an accurate description of this particular Internet text type based on frequency information.

4 Conclusion

This paper dealt with the identification of patterns in Internet language applied to Portuguese data. It constitutes to our knowledge the first attempt at proposing a fully automatic corpus-driven method for this task.

As a concrete output of this research, two new components were integrated into Tesla: the *corpus log likelihood* component and *document rank component*. These components can be used to replicate our experiments for Portuguese or any other language, as well as to design new frequency-based experiments.

From a theoretical point of view, experiments show that it is possible to obtain keyword lists by comparing internet corpora to standard corpora. These keyword lists reflect different aspects of computer-mediated communication such as spelling variation, lexicon and syntax. Moreover, experiments also indicate that document ranking using frequency information may help to describe the extent to which internet data is closer to spoken and/or standard written language.

5 Future Perspectives

We plan to continue the experiments presented here using Tesla. This can be done by improving the quality and functionalities of the current components as well as by developing new components to perform other text processing tasks.

There are a couple of directions which can taken to expand this work such as the using other associative metrics on this dataset. Results can be compared to those obtained using log-likelihood to determine which metric performs best for this particular task. Another implementation might be a semi-automatic method for the identification and tokenization of special characters (e.g. emoticons).

We also aim to apply the same methods we applied to Brazilian Portuguese to European Portuguese internet data. Recent experiments [21] show that it is possible to distinguish Brazilian and European Portuguese standard texts with more than 95% accuracy at both lexical and character level. It would be interesting to see if this distinction is also possible when dealing with user-generated content.

Finally, the experiment configuration and a small snapshot of the dataset used here will be available in Tesla in the coming months.

Acknowledgements

The authors would like to thank Sascha Diwersy for the help with retrieving data from *Yahoo Perguntas*, and Jürgen Rolshoven and Martin Becker for insightful comments throughout the development of this work. We also thank *Folha de São Paulo* for the reference corpus.

References

[1] Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

[2] Bick, E. (1999) *The parsing system Palavras*. Aarhus University Press, Denmark.

[3] McEnery, T.; Hardie, A. (2011) *Corpus Linguistics*. Cambridge University Press.

[4] Renouf, A. (2009) Corpus Linguistics beyond Google: the WebCorp Linguist's Search Engine. *Digital Studies - Le Champ Numerique*. Vol 1.

[5] Beißwenger, M. (2009) Multimodale Analyse von Chat-Kommunikation. In Birkner, D.; Stukenbrock, A (eds.): *Die Arbeit mit Transkripten in Fortbildung, Lehre und Forschung*. Mannheim. p. 117-143.

[6] Herring, S. (2001) Computer-mediated discourse. In Schiffrin, D.; Tannen, D.; Hamilton, H. (eds.) *The Handbook of Discourse Analysis*. Oxford Blackwell Publishers. p. 612-634

[7] Hermes, J.; Schwiebert, S. (2009) Classification of text processing components: The Tesla Role System. In: Fink, A.; Lausen, B.; Seidel, W. and Ultsch, A. (Eds.). *Advances in Data Analysis, Data Handling and Business Intelligence. Studies in Classification, Data Analysis, and Knowledge Organization.* Volume 4. Springer. p. 285-294.

[8] Koch, P. Österreicher, W. (1985) Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch 36*. Berlin / New York. de Gruyter. p. 15-43.

[9] Baron, N. (2008) *Always On: Language in an Online and Mobile World*. Oxford University Press. New York.

[10] Crystal, D. (2001) *Language and the Internet*. Cambridge University Press.

[11] Sharoff, S. (2006) Creating general-purpose corpora using automated search engine queries. In Baroni, M.; Bernardini, S. (eds): *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

[12] Kilgarriff, A.; Rychly, P.; Smrz, P.; Tugwell, D. (2004) The Sketch Engine. *Proc EURALEX 2004*. Lorient, France. p 105-116.

[13] Baroni, M.; Bernardini, S. (2004) BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC*.

[14] Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics - Special Issue on Using Large Corpora.* Volume 19, Issue 1. MIT Press.

[15] Scott, M. (1997) PC Analysis of key words and key key words. *System*. n. 25. Elsevier. p. 233-245.

[16] Berber-Sardinha, T. (2000) Comparing corpora with WordSmith Tools: How large must the reference corpus be? *Proceedings of the Workshop on Comparing Corpora 9*. p. 7-13.

[17] Kemppanen, H. (2004) Keywords and ideology in translated history texts: A corpus-based analysis. *Across Languages and Cultures 5.1*. p. 89-106.

[18] McEnery, T. (2009) Keywords and moral panics: Mary Whitehouse and media censorship. in D. Archer (ed.) *What's in Word-list? Investigating Word Frequency and Keyword Extraction*. Oxford: Ashgate.

[19] Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V.; Aswani, N.; Roberts, I.; Gorrell, G.; Funk, A.; Roberts, A.; Damljanovic, D.; Heitz, T.; Greenwood, M.; Saggion, H.; Petrak, J.; Li, Y.; Peters, W. (2011) *Text Processing with GATE (Version 6)*. University of Sheffield. Department of Computer Science.

[20] Ferrucci, D.; Lally, A. (2004) UIMA: an architectural approach to unstructured information processing in the corporate research environmen., *Natural Language Engineering*. V.10, N.3-4. p.327-348.

[21] Zampieri, M.; Gebre, B. G. (2012) Automatic Identification of Language Varieties: The Case of Portuguese. *Proceedings of KONVENS 2012*. Vienna, Austria. p. 233-237