Working Paper #1

The Brazilian Corpus on Urban Violence

Carmen Dayrell, CASS/Lancaster University

This working paper reports on the process of compiling the Brazilian Corpus on Urban Violence. It describes four relevant steps in the compilation of corpus: (i) selection of sources from which articles were collected and selection of individual texts, (ii) formatting of the retrieved texts so that they can be processed by corpus linguistics techniques, and (iii) clearing of undue noise, and (iv) standardisation of spelling. The report concludes with an overview of the corpus' content.

1. Selecting sources and individual texts

The newspaper articles included in the Brazilian Corpus on Urban violence were collected from *Factiva*, a news aggregator service that provides full-text access to newspapers, newswires, business journals, market research and analyst reports, and web sites from 118 countries. Here we focused on articles published between 01/Jan/2014 to 31/Dec/2014 by the following Brazilian newspapers: *Zero Hora*, *Pioneiro*, *Folha de São Paulo*, and *O Estado de São Paulo*. These are daily broadsheet papers with wide circulation in the states where they are based. The first two (*Zero Hora* and *Pioneiro*) are based in the Brazil's Southern state of Rio Grande do Sul, where the Brazilian researchers in this project are based and hence the focus of our study. The other two newspapers are published in São Paulo, the wealthiest and most populated state in Brazil. They were included in the corpus to allow comparison of the discourse around urban violence in different regions of Brazil.

To select individual texts, our initial approach was to apply Gabrielatos' (2007) method which is especially useful to determine query words or phrases which favour the retrieval of a wide range of relevant texts from a restricted-access database. Briefly, Gabrielatos (2007) suggests using a core query consisting of two or three words/phrases as a starting point to compile a pilot corpus. This pilot corpus is then used to identify additional relevant query words/phrases. These are words/phrases that tend to occur in texts where the core terms are also used, thus they are at least in principle closely associated with the core terms in a significant number of contexts. The ultimate purpose of applying Gabrielatos' (2007) method is to identify words/phrases that would return articles on the topic under investigation, even though core terms themselves are not used in them. At the same time, these additional terms should not create undue noise, that is, useful additional terms are those that retrieve a sufficient number of articles which do not contain the core terms but are still relevant.

Given the restricted time period examined in this study (2014 only), we opted for compiling an initial corpus using all articles published in the chosen four newspapers (*Folha de São Paulo*, *O Estado de São Paulo*, *Zero Hora* and *Pioneiro*) in the entire period (Jan-Dec/2014). This initial corpus would then be used to identify additional relevant query words/phrases as suggested by Gabrielatos (2007). Our first attempt was to use the Portuguese equivalent for *urban violence* (*violência urbana*) and *violence in cities/towns* (*violência na(s) cidade(s)*) as our core query terms. However, these two terms did not retrieve as many texts as one would expect in a country where urban violence is a major issue. Overall, *urban violence* (*violência urbana*) appeared in 66 articles and *violence in cities/towns* (*violência na(s) cidade(s)*) in 10 articles. Neither was *violence in the street(s)* (*violência na(s) rua(s)*) frequently used: 22

articles in total. In an attempt to identify search terms that would lead to a higher number of texts on urban violence, we then searched for *urban security* (*segurança urbana*) and *public security* (*segurança pública*). *Urban security* (*segurança urbana*) is not frequently used in Brazilian newspapers either: 50 articles in total. *Public security* (*segurança pública*) on the other hand is frequently mentioned: 1,809 articles in total.

Violência urbana (*urban violence*) and *segurança pública* (*public security*) were then used to compile a pilot corpus so that Gabrielatos' method could be applied to identify additional search terms. The method pointed to three additional terms: *criminalidade* (*criminality*), *homicídio* (*homicide*), and *roubo* (*robbery/theft*). While relevant, using *homicídio* (*homicide*), and *roubo* (*robbery/theft*) as query terms would result in a biased selection of texts that would inevitably favour texts about these two crimes specifically. This would not allow us to have a clear picture of what crimes are most frequently mentioned in Brazilian newspapers, the project's research question #1. Our decision was therefore to complement the list of query terms with crime names mentioned in official government statistics[1] as well as other crimes the researchers would intuitively deem important. Also, in an attempt to gather as many relevant texts as possible, we opted for expanding the collection of texts to all word forms related to the selected crimes names. Thus, for example, rather than using *roubo* (*robbery/theft*) as a query term, we used *roub** which retrieves texts containing *roubo* as well as *roubos* (*plural form*), *roubar* (to rob/steal), *roubou* (robbed/stole), *roubado* (robbed/stolen), etc. The full list of our query terms is presented in Figure 1, under "at least one of these words".

While useful to identify texts related to urban violence in Brazil, using crime related words as query terms has nevertheless introduced some undue noise. A number of texts in which these terms appeared referred to violence and crimes in other parts of the world, rather than in Brazil: *murders* in Iraq, *kidnapping* in Nigeria, *homicides* in war zones and so on. In addition, there were also a large number of texts referring to issues other than urban violence such as corruption, internet crimes and labour issues, in Brazil and somewhere else as well as articles related to cinema (especially thrillers) and crime fiction. To make matters more complicated, one cannot ignore the metaphorical nature of language. There was also a large number of texts in which our query terms were used metaphorically and not at all related to urban violence: *roubar a cena* (*steal the scene*), *roubar meu lugar* (*take over my place*), *furtar-se a fazer alguma coisa (avoid doing something)*, etc.

To minimize such noise, we have discarded a wide range of topics in the actual retrieval of texts from the *Factiva* news aggregator. The topics discarded are shown under the lave "subjects" in Figure 1. They were identified on the basis of a random analysis of the texts within such categories. We have also discarded texts containing one or more of the following words/phrases: *comissão da verdade* (*truth commission* – a committee established in 2012 to investigate violations of human rights by the Brazilian government between 18/Sep 1946 to 05/Oct 1988), *Bolsonaro* (a Brazilian congressman, infamous for his controversial comments on rape and human rights), *Petrobrás or Petrobras* (Brazilian oil company at the centre of a corruption scandal), *ditadura* (*dictatorship*), *ditador* (*dictator*), *Al-Quaeda*. These words are shown under "None of these words" in Figure 1. Also, within the *Factiva* search options, we have chosen to discard identical duplicates and also republished news, recurring pricing and market data, obituaries, sports, calendars.

[1] The following have been consulted: https://www.sinesp.gov.br/estatisticas-publicas (national statistics); http://www.ssp.rs.gov.br/?model=conteudo&menu=189 (Rio Grande do Sul); and http://www.ssp.sp.gov.br/novaestatistica/default.aspx (São Paulo).

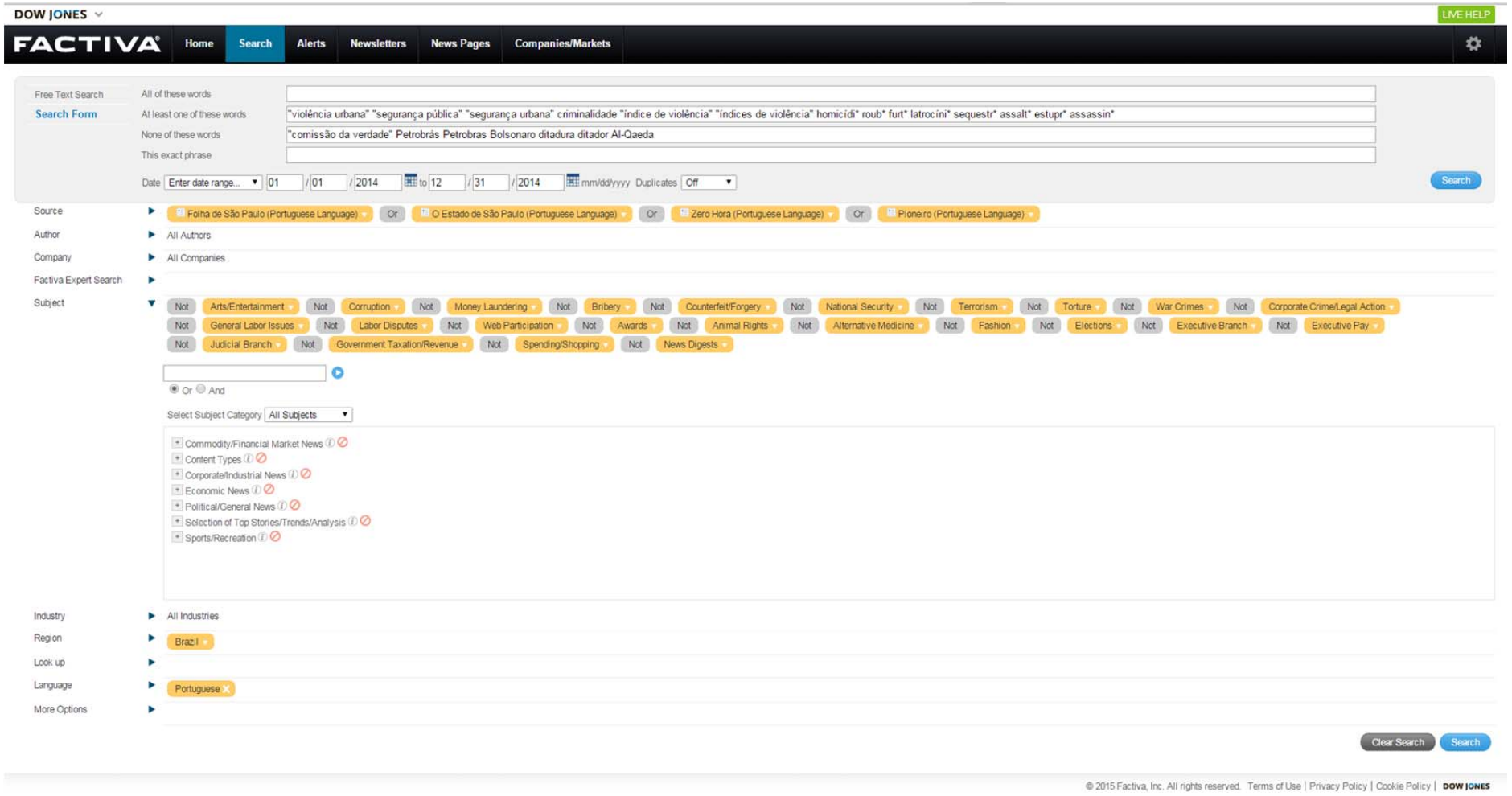


Figure 1: Criteria for selecting texts from Factiva and including in the Brazilian Corpus on Urban Violence

All texts meeting the criteria above were retrieved in full, including their headline(s). This means that there was not filtering according to the section of the newspaper in which the text was published. In other words, the corpus contains news reports as well as editorials, opinions, interviews, or any other text type. It is also important to stress that texts were selected irrespective of the number of query words/phrases it contained and their frequency within each text. This means that the texts included in the Brazilian Corpus on Urban violence vary in relation to the extent to which urban violence is discussed. Here, any reference to urban violence is considered relevant, even if urban violence is not the main topic discussed in the text. This enables us to look at both texts discussing urban violence issues in detail as well as those in which urban violence issues are mentioned in relation to another topic. Such approach broadens the scope of the analysis and enables us to examine situational contexts which are directly or indirectly associated with urban violence.

2. Formatting texts

Texts obtained from the *Factiva* database are in electronic format but require pre-processing in order to allow appropriate corpus analyses. This is mainly because the *Factiva* search tool allows the user to download 100 texts at once but they all get saved in a single file. In addition, each text includes not only the article itself and its headlines but also some extra-linguistic information about them such as an identification code, the newspaper where it was published, data of publication, and section of the newspaper. This extra-linguistic information is particularly relevant in the data analysis as it allows us for example to compare discourses across newspapers. It may also enable us to establish whether particular discourses are linked to specific authors or sections of a paper. However, in order to allow that, this information needs to be identified and annotated. This research follows the standard corpus linguistic procedure and inserts metadata within angle brackets. By doing so, corpus linguistics tools can access to it whenever necessary while at the same time overlooking it when processing the text.

To illustrate, below is a news report published by *Pioneiro* in October 2014. The numbers on the left of each line are not part of the actual text; they are used here as reference to illustrate what each line refers to. The headline appears in line (i) and the news report in (viii). The text also includes: (ii) its total number of words, (iii) date of publication, (iv) extended name of newspaper, (v) short name of newspaper, (vi) language, (vii) a copyright disclaimer, (ix) the corporate name of the media outlet, and (x) the document identification number. Although not included in the example below, texts may also include the name of the author(s), the section and/or page in the newspaper where the text was extracted from.

(i) *Homem que matou mulher é julgado*
(ii) *97 words*
(iii) *30 October 2014*
(iv) *Pioneiro*
(v) *PIONEI*
(vi) *Portuguese*
(vii) *© 2014 RBS Internet e Inovaçâo – Todos os direitos reservados.*
(viii) *Denunciado por homicídio qualificado em 2012 pelo Ministério Público, o metalúrgico Antônio Elton Correa de Oliveira, 53, será julgado hoje no Fórum. O réu responde pelo assassinato de Marcia Aparecida da Silva Reis, 38, com quem mantinha um relacionamento. Conforme inquérito policial, em março de 2012, Oliveira desferiu mais de 17 golpes de faca contra a mulher na moradia do casal, no bairro Arcobaleno. O crime, segundo o inquérito, foi cometido por motivo torpe, pois o réu estava insatisfeito com o relacionamento e atribuía à vítima o fracasso na vida pessoal.*

(ix) Grupo RBS
(x) Document PIONEI0020141030eaau0000o

Our first step was to identify the metadata included in each text. This process of annotating the metadata within each text was done automatically[2]. This is possible because such information tends to be presented in a given sequence. The problem is that this sequence may vary from one text to another. For example, in the case of the Brazilian Portuguese news reports, we may find the author of the text, a subheading, or the section of the newspaper in the second line of the text. In these cases, the number of words would be placed in the third line. To overcome this problem, the solution was to design the automatic annotation according to the structure of the vast majority of texts. For example, we considered any text before the number of words as the headline, even though it was not always necessarily the case. The downside of it is that this also means that the name(s) of the author(s) and section of the newspaper could not be automatically identified since they were shown in different positions within the text. If relevant, this information can be manually recovered during the course of the data analysis.

Table 1 lists the tags used to annotate the texts within the Brazilian Corpus on Urban violence and a short description of each.

Tag	Description
text id	Text identification
factiva_id value	Text identification in Factiva
copyrightStatement value	Copyright disclaimer
extent value	Number of words
publication value	Newspaper
pubDate value	Date of publication
edition value	Edition
pageDesc value	Page
fullTitle value	Full Headline
section value	Section within the newspaper
headline value	Headline
byline value	Author

Table 1: Tags used to annotate metadata in texts within the Brazilian Corpus on Urban violence

Here is how the news article above looks like after annotation. The notation </text> at the end was included to mark the end of each text.

```
<text id="PIONEI30 October 2014003797">
    <factiva_id value="PIONEI0020141030eaau0000o" />
    <copyrightStatement value="© 2014 RBS Internet e Inovação – Todos os direitos reservados." />
    <extent value="97 words" />
    <publication value="PIONEI" />
    <pubTitle value="Pioneiro" />
    <pubDate value="30 October 2014" />
    <edition value="NULL" />
    <pageDesc value="NULL" />
    <fullTitle value="Homem que matou mulher é julgado" />
    <section value="NULL" />
```

[2] We thank the help and assistance from our CASS colleague Andrew Hardie to automatically identify the meta-data within texts as well as to carry out the second stage of the pre-processing, to split files into individual texts.

<headline value="Homem que matou mulher é julgado" />
<byline value="NULL" />

Denunciado por homicídio qualificado em 2012 pelo Ministério Público, o metalúrgico Antônio Elton Correa de Oliveira, 53, será julgado hoje no Fórum. O réu responde pelo assassinato de Marcia Aparecida da Silva Reis, 38, com quem mantinha um relacionamento.
Conforme inquérito policial, em março de 2012, Oliveira desferiu mais de 17 golpes de faca contra a mulher na moradia do casal, no bairro Arcobaleno. O crime, segundo o inquérito, foi cometido por motivo torpe, pois o réu estava insatisfeito com o relacionamento e atribuía à vítima o fracasso na vida pessoal.

<Grupo RBS>

</text>

Once all texts within the 100-text Factiva file were annotated, our next step was to split each file according to the number of texts it contained so that we would have one newspaper text per file. This process was also carried out automatically using the notation </text> to identify the end of each text. To make it easier to identify the text from which a given instance comes from during the data analysis, each text was named as follows *xxx_aaaa_mm_n*, where:

- *xxx* – indicates the newspaper where it was published. The full list of the newspapers and their respective identification codes is presented below.

	Newspaper	**Identification Code**
1.	*Folha de São Paulo*	FLH
2.	*Estado de São Paulo*	EST
3.	*Zero Hora*	ZHR
4.	*Pioneiro*	PNR

- *aaaa_mm* – date of publication (year and month).
- *n* – texts were numbered so that those published within the same month could be distinguished one from another.

Thus, PNR_2014_06_0000003 indicates that text was published in *Pioneiro* in June 2014 and it is the text #3 for that particular month.

3. Further cleaning of undue noise

While the procedures described earlier were efficient to reduce undue noise in the corpus, irrelevant texts still got selected. This is an inevitable side-effect of any search that uses query words/phrases to automatically select texts to be retrieved. But, fortunately, one that is easily spotted in the initial stages of the data analysis and can be easily solved. In our case, the corpus still contained articles about murders, homicides, kidnapping and robbery in various places other than Brazil. It also included articles about a Brazilian banker (*Pizzolato*) accused of corruption and money-laundering and some remaining articles related to film releases.

Thus, to remove articles that were about violence and conflicts outside Brazil, we have discarded all articles from *O Estado de São Paulo* published in the section *International* and those from the *New York Times* published by *Folha de São Paulo*. *Zero Hora* and *Pioneiro* did not enable clear annotation of the journal section and hence they did not allow us to identify and remove texts using this criterion. Further articles about places other than Brazil were removed by using the following phrases: *Tay Thi* (Nigerian girl kidnapped by

extremists), *Hamas, Islâmico, Boko Haram, Islã, Muçulman*, judeus, Palestin*, Exército do Povo Paraguaio, Nicolás Maduro, Chávez, Bush, Obama,* and *Pizzolato.* The words *suspense* (*thriller*), *sinopse* (*synopsis*), *em cartaz* (*showing*) were used to removed texts related to film releases. Overall, 276 texts were extracted from the initial corpus selection.

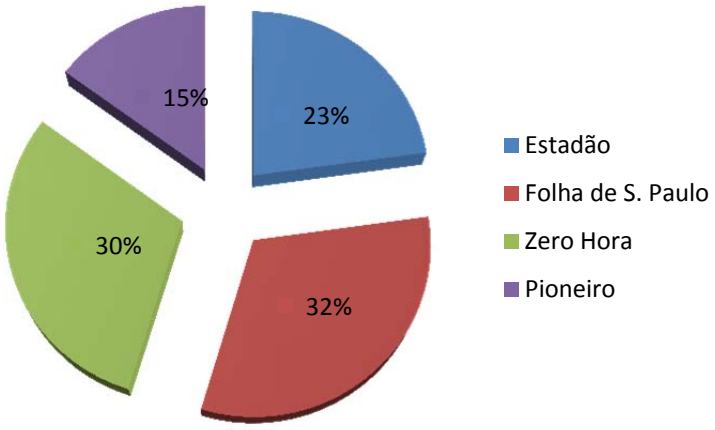
Additionally, most articles included a line at the very end with some information about the newspaper or the media conglomerate it is part of and, in some cases, copyright statements. These words/phrases were put between angle brackets using Sarant (Anthony 2014) so that they could be overlooked by AntConc (Anthony 2014) during the analysis. This was the case, for example, of <*Grupo RBS*> in the example above.

4. Standardising spelling

The last procedure in the pre-processing phase of the corpus compilation was the standardisation of spelling. This was necessary because a new spelling system for the Portuguese language was put in place in 2009 but the Brazilian government allowed a transition period between 1st January 2009 and 31st December 2012, which has been extended to 31st December 2015[3]. Texts published by newspapers in 2014 were written in either the old or the new spelling. An adequate corpus analysis requires all the data to follow one single system. All texts were converted into the new spelling system by means of the LINCE tool (Ferreira et al. 2012)[4], choosing the Brazilian variant of the language.

5. Composition of the corpus

This last section presents the composition of the Brazilian Corpus on Urban violence. The corpus contains 5,127 texts (1,778,282 words) published by four Brazilian broadsheet papers – *Zero Hora, Pioneiro, Folha de São Paulo,* and *O Estado de São Paulo* – between January and December 2014. Graph 1 shows the distribution of words in the Brazilian Corpus on Urban violence per newspaper. Raw figures are presented in Appendix I in terms of both number of articles and number of words.



Graph 1: Distribution of words in the Brazilian Corpus on Urban violence across newspapers.

[3] http://www2.planalto.gov.br/excluir-historico-nao-sera-migrado/acordo-ortografico-da-lingua-portuguesa-entrara-em-vigor-em-2016 (accessed on 21 Aug 2014).
[4] LINCE is accessible at http://www.portaldalinguaportuguesa.org/lince.html; accessed on 20 Oct 2015.

6. References

Ferreira, J. P.; Lourinho, A.; Correia, M. (2012). 'Lince, an End User Tool for the Implementation of the Spelling Reform of Portuguese'. In: Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (Eds.). *Computational Processing of the Portuguese Language.* Lecture Notes in Computer Science, 7243. Berlin-Heidelberg: Springer-Verlag, pp. 46-55

Gabrielatos C. (2007) 'Selecting query terms to build a specialised corpus from a restricted-access database'. *ICAME Journal* 31: 5–44.

Appendix I:

This Appendix presents the raw figures for the distribution of texts and words in the Brazilian Corpus on Urban Violence.

	Zero Hora	Pioneiro	Folha de São Paulo	Estado de São Paulo	Geral
Number of words	539,847	265,785	567,128	405,522	1,778,282
Number of articles	1,299	868	1,694	1,266	5,127