3. Methodology

The following section presents an overview of the research steps which were taken in order to design, propose and test the new approach to CDA research which is discussed here. Generally speaking, this approach was made up of four hierarchically arranged, interconnecting parts. The steps which were undertaken were the product of a series of brief pilot studies (see Section 3.1.1, p. 58) and were designed to offer the researcher authentic data while simultaneously limiting the role that they would have in attaining and analyzing it.

First, two corpora were constructed. These corpora were each made up of texts selected to form representative samples of particular language communities' (American English and Mexican Spanish) print media discourse during a given month. In the present investigation, newspaper reporting was the only genre of print media examined. Second, a statistical analysis of lexical frequency was undertaken and a frequency limit was established. This was then followed by an examination of the collocates of frequently occurring LIs (node words); these were selected using the frequency 'ceiling'; and finally, those collocations which were found to be statistically salient were analyzed alone, as part of lexical groupings as well as both inter- and intra-linguistically. This was done through the use of a variety of means including frequency analysis and semantic prosody analysis within and across both corpora.

This multi-faceted approach was designed to analyze a stratified random sample of language use from a specific language domain. The set of steps used in the current study allowed for a methodology in which critical language was presented *to* the researcher

through its mere presence in the corpus data instead of one in which the researcher deliberately searched *for* language which was deemed (by the researcher) to be 'critical.' This characteristic of the present investigation is not only a hallmark of positivist research, but marks the most obvious deviation from the typical procedures utilized in many previous CDA-based studies (see, e.g., Orpin, 2005; Fairclough, 1995).

3.1. Overview of the Methodology Used

The current project utilized two principle methodological approaches (CADS and SP analysis) in order to explore the potential of an approach to CDA in which textual DA could be carried out as objectively as possible. Together, these approaches allowed for testing and carrying out a new approach to CDA research. Due to the methodological weaknesses which are laid out in the above sections with regard to both CDA-based studies and—to a lesser extent—SP studies, the current research project combined the use of CADS (Freake et al., 2011) viewed from a critical standpoint (e.g. Baker et al., 2008; Orpin, 2005) with semantic prosody analysis (Oster, 2010; Louw, 2008) in order to address the themes usually examined by CDA researchers from a new and more objective perspective. This combination of approaches was seen as complimentary for a number of reasons (discussed in detail in Chapter 2, p. 29) and was chosen in order to exploit the strengths of both approaches rather than focusing on their weaknesses.

Semantic prosody, for one, has very seldom been used for anything except the analysis of single lexical manifestations across mega corpora. As such, its use on a smaller, more practical level was seen as an important tool for effectively analyzing the findings gleaned from corpus analyses—something missing or under-represented in previous explorations of

CDA-corpus linguistics combinations (Salama, 2011; Baker et al., 2008). For example, while researchers like Oster (2005) have looked at varied occurrences of LIs related to *fear* or the concept of *corruption* (Orpin, 2005) using SP analysis, few, if any, have used SP as a tool to aid in the practical analysis of language and discourse.

At the same time, the use of corpus assisted CDA (and SP) on a more focused scale allows for a more complete analysis. This emphasis is something missing in many works, which tend to select a specific lexical grouping in a mega corpus (oftentimes a lexical item or synonyms for expressing a concept—similar to SP analyses) and analyze it in as many contexts as possible with the end goal of laying out the underlying discourses surrounding the unit analyzed (see, e.g., Salama, 2011; Baker et al., 2008). Although the analysis of a mega corpus can provide interesting findings, the data obtained from the corpus still has to be subjected to the researcher's subjective interpretation and analysis. Additionally, these sorts of studies lack the ability to effectively analyze particular genres and environments on a macro scale. This is partly due to features such as 'seasonal collocates' (Baker et al., 2008) in which certain language uses, "…are very frequent in a small number of years," (p. 286). With these methodological weaknesses in mind, the use of corpora based on stratified random samples from a specific area was complemented by the use of SP in that both methodological tools were made more practical and the resulting study was provided with a more complete theoretical foundation. Together, the two methods serve as a potential solution to many of the problems found in traditional approaches to CDA research.

Through the fusion of these two approaches, the present investigation looked to address two oft-cited weak points in CDA research: text selection and analysis criteria. As an

answer to the arbitrary selection of texts pointed out by Prentice (2010), common in many traditional CDA studies, texts were selected based on minimal topic-related lexical characteristics and were included in the final corpora only if they complied with certain criteria set forth at the beginning of the research process. Furthermore, in order to facilitate the analysis of large quantities of text, AntConc 3.2.4 (Anthony, 2011), a corpus analysis program was utilized in the examination of the final corpora of texts. As a response to the types of textual analyses often employed in CDA research, an approach based on SP analysis (particularly addressing the presence of collocations) was chosen in order to analyze the overall discourse present in the corpora. This not only grounds the corpus findings in theory but also highlights the general connotations of each corpus and, as a result, the media discourse in each country regarding the topic shared by both corpora. Moreover, steps were taken in order to make the approach to traditional CDA which is presented here as balanced and objective as possible.

To facilitate the methodological fusion used in the present study, various intermediary steps were taken in piloting, designing, assembling, collecting data from and analyzing the corpora. These steps, as they relate to different phases of the research process, are laid out briefly here and are discussed individually in more detail below. This study relied on the use and analysis of two corpora. These two corpora were assembled using topically parallel newspaper articles (focused on 'drug-related violence' in Mexico and the United States). All articles had been published during a single, randomly selected month during 2011 in newspapers from the United States (published in English) and Mexico (published in Spanish).

After having been constructed, the corpora were analyzed for lexical frequency, collocations and the Mutual Information (MI) scores of the collocates of salient node words. The use of MI scores in corpus linguistics gives the researcher a numerical representation of the 'strength' of a collocation and has been used previously in studies combining CDA and corpus linguistics as well as in semantic prosody studies (see Salama, 2011; Oster, 2010 for discussion). Mutual Information (MI) scores determine, "…whether there is a higher-than-random probability of the two items [the node being examined and a given collocate] occurring together," (Mautner, 2009, p. 125). This was done for all statistically salient LIs in both corpora.

After having created the final corpora, a raw frequency list was created for each corpus and was normalized to account for the different sizes of the two corpora. Upon establishing raw counts of lexical frequency for each corpus, a frequency 'ceiling' was established. Due to the fact that the corpus was only representative of a very specific area of language use and that there were a large number of LIs which only appeared once, the data was not evenly distributed. Thus, the frequency 'ceiling' was defined as a frequency of occurrence greater than the mean frequency of lexical occurrence. This allowed for analysis to be focused on only those LIs with a statistically salient presence in the data. After having established the frequency 'ceiling' for each corpus, a concordance analysis was carried out using LIs which were identified as 'frequently occurring' (this set of LIs excluded function words).

Having defined what constituted a frequent occurrence in the corpora, as well as which LIs had particularly 'strong' collocates, salient LIs were analyzed for their semantic prosody by observing their presence in the corpora based on their appearance as parts of

collocations and through the use of the concordance tool in the AntConc corpus analysis program (Anthony, 2011). This was done using the selected collocates' MI scores; leading the research to a point at which a general analysis of the semantic prosody of salient collocations could be undertaken. This analysis set the groundwork for the final goal of the study: to examine both corpora inter- and intra-linguistically based on lexical collocations and semantic prosody.

What follows is a detailed account of the methodology used in the present study. Because the current project is intended to be used as a starting point for future research, the methodological description is not laid out in a 'traditional' format. That is, the steps which comprise the method which was used are laid out chronologically as procedural steps. This was done for two principal reasons. First, the methodology which was used is a logical set of steps and to present it in a 'standard' format would prove quite confusing; when presented as logical steps, however, the method as a whole (as well as the rationale which informed its creation) can be more easily understood by the reader. Second, since the methodology which was employed here is intended to lay the groundwork for similar studies in the future, a step-by-step description of the methods employed not only allows for simple, straightforward replication as well as further testing, but for a transparent means through which to see any flaws within the method proposed here; thus allowing for the effective implementation of productive changes.

The experimental design which informed and was used in the present study is presented here, chronologically, in its complete form. The initial piloting process is discussed as well as its connection to the final research design. Following this, requirements for text inclusion in the corpora are detailed and their presence in and importance to the current

study is made clear. Additionally, the actual data collection and corpus construction is discussed in detail, particular attention is paid to the collection of texts, the corpora construction and the general characteristics of the corpora analyzed. Finally, a discussion of the statistical measures employed in analyzing the data is presented, as well as a description of the text analysis process.

3.1.1. Pilot Study

In order to be able to find and address the necessities for and methodological shortcomings of the current project, the first step which was taken upon having proposed research was to conduct a pilot study. Because searching for and selecting corpus materials was a vital step in establishing as much objectivity as possible, piloting was considered important as it provided an opportunity to experiment with different search methods while simultaneously becoming acquainted with the construction and analysis of text corpora. Being as the current project placed great importance on a degree of objectivity within data collection and analysis, it was of the utmost importance that the search terms (and not the *searcher*) returned representative and authentic data. This in turn allowed for the construction of representative corpora with little or no interference on the part of the researcher. The piloting period not only helped to gain experience at carrying out data collection, but also exposed many weaknesses in search methods which were later addressed when constructing and carrying out the final research.

The piloting period took place from October 28th through November 18th of 2011. During this period, one data collection was carried out per week. Pilot corpora were searched, constructed and analyzed four times during this period (Friday, October 28th;

Friday, November 4th; Friday, November 11th; and Friday, November 18th). Corpora on these days were constructed from texts published on the day in question, and were selected using *Google News* (discussed below) between 9 and 10 o'clock p.m. Texts included were found using set search terms (which evolved during the course of piloting—see Table 1) and had been published on the day of the search; they additionally were all written by credited authors. Once a search was carried out, all articles found were included in the pilot corpora. The corpora were then analyzed for raw frequency using the TextStat[5] corpus analysis program (Hüning, 2012) and the overall content of the corpora was examined. As can be seen in Tables 1 and 2, the use of search terms changed for both corpora during the piloting period. Initially, so-called 'wild cards' were used in order to have a broader set of results; however, this was eliminated from the final methodology; additionally, the final set of search terms was selected after experimenting with different manners of conducting searches during the first and second pilot runs. Because of this, the search terms were changed to those which were used for the remainder of the pilot study as well as in the final data gathering.

Table 1. Search terms used for English corpus piloting

Pilot Date	Search Terms Used	Corpus Size
Oct. 28	drug*/viol* (Advanced search)	2,646
Nov. 4	drug OR drugs AND violent OR violence	2,467
Nov. 11	drug OR drugs AND violent OR violence	4,228
Nov. 18	drug OR drugs AND violent OR violence	6,349

[5]The TextSTAT program was only used during the piloting process as the most important area of piloting was examining raw frequencies. Upon initiating the study itself, corpus analyses were carried out using AntConc 3.2.4 (Anthony, 2011). This program was selected because it offered a wider range of features—most notably, tools for statistical analysis such as MI scores; something not possible in TextSTAT.

Table 1 shows the search terms utilized in carrying out the pilot data collections for the English corpus. As can be seen in the table, the search terms evolved during the course of piloting and as they did so, the data analyzed was pulled from increasingly large corpora. Table 2 (below) presents the same information for the Spanish corpus piloting and demonstrates similar patterns of change in the search terms used and in corpora size.

Table 2. Search terms used for Spanish corpus piloting

Pilot Date	Search Terms Used	Corpus Size
Oct. 28	droga*/violen* (Advanced search)	3,303
Nov. 4	droga OR drogas AND violente OR violencia	2,234
Nov. 11	droga OR drogas AND violente OR violencia	2,041
Nov. 18	droga OR drogas AND violente OR violencia	8,133

In addition, the use of the Advanced Search setting in *Google News* was abandoned beginning with the third pilot run. This was deemed necessary because it was found that the use of a date range returned less reliable results (fewer results in general and many of them not from the date range being searched). Interestingly, this was only found to be the case in utilizing the Mexican version of *Google News*; there was no difference in the search terms when searching in English and so the Spanish terms were changed. The combined size of the pilot corpora used in the pilot study (texts from all data collection dates) was comparable to that of the final corpora which were used. The English pilot corpus was made up of 15,590 words (an average of 3,897.5 words per data collection[6]) and the Spanish corpus consisted of 15,711 words (an average of 3,927.75 words per data collection).

[6] *Data collection* refers to the set of data obtained from one search. That is, all the data compiled during each piloting date.

As is pointed out above, the main purpose of carrying out a pilot study for the current project was to fine-tune the search method used so as to distance the researcher from the data collection process as much as possible and to minimize direct involvement in changing, selecting, and assigning the data included in the final corpora. Nonetheless, it is impossible to ever completely remove a researcher from the research process. That is to say that a researcher still sets search terms, selects texts, decides on a topic for study and eliminates texts which are not representative of the discourse arena being examined.

With this in mind, the pilot portion of the study additionally served to highlight a few of the types of texts which were consistently found with the broad search terms used in the current study. So, although piloting helped to make the present study's final methodology more objective, it additionally highlighted certain points in data collection which would need to be addressed manually after data had been collected and corpora had been built. One of the most critical points highlighted through piloting was the need for individual revision of the texts selected using the initial search. Because one of the strong points of the present research is that, using the search terms established, very little had to be done in order to select texts, once the search terms were entered, texts related to the topic were found. However, through piloting it became apparent that, due to the minimal search terms used—the very feature which allowed for a degree of objectivity in text selection—it was necessary to manually double check to assure that the texts selected based on the search were representative of the corpora being analyzed. This was done by establishing a set of minimal characteristics for text inclusion during the pilot period.

In this sense, piloting was one of the most crucial parts of the entire research process. While certain characteristics of the present project are reminiscent of similar studies (see,

for example, Salama, 2011; Baker et al., 2008; Schrøder, 2007), both the care taken in order to assure as much objectivity as possible in collecting and assembling texts and the intention of being able to identify a discourse pattern in texts with no expectation of underlying ideologies would appear to be fairly uncommon in many CDA-based studies. As such, the pilot portion of data collection aided the present research by solidifying the manner of searching for, criteria for inclusion of, and criteria for exclusion of the texts which formed the final corpora used in the research.

3.1.1.1. Texts

Because the present research was based entirely on the simultaneous analysis of two separate, topically parallel corpora, the corpora were built using two sets of print media texts (in this case, newspaper articles) which were topically similar. The corpora which were analyzed were made up entirely of articles from major Mexican and American newspapers. The newspaper articles included in the final corpora met a specific set of characteristics which were representative of a particular topic: 'drug-related violence' in the United States and Mexico. The characteristics which informed the selection of texts as well as the lexical search terms (see Tables 1 and 2, pp. 59 & 60) served to eliminate any bias that would result if the corpora were arbitrarily constructed based on the topic being analyzed. The use of two linguistically distinct corpora covering parallel issues allowed for comparisons between the discourses used in each. This then permitted a comparison of each country's print media discourse (in the vein of Freake et al. (2011)); as opposed to a comparison between a specific discourse and language use in a mega corpus (i.e., Salama, 2011).

The possibility of choosing texts based solely on topic was acknowledged from the beginning of research as a potential strategy for corpus construction but was avoided as much as possible. This was done due to the influence which researchers have been shown to have on the outcome of a study of this sort if they exercise unilateral control over the texts chosen for analysis—especially when chosen based on topic (as shown in Poole, 2010). With this potential short-coming in mind—what Stubbs (1997) refers to as 'circularity'—it was deemed important to design the current study in such a way as to have the search terms used in data collection be 'responsible' for the brunt of data collection. Thusly, search terms were selected which reflected the underlying themes which informed the greater media discourse being studied here (e.g. Orpin, 2005). In this case, variations on drug- and violence-related lexical units were used for searching for and collecting data. Although these search terms evolved during the course of the pilot study (see Tables 1 and 2, pp. 59 & 60), they were consistently centered on the presence of LIs related to drugs and violence within print news articles, thus serving to almost completely restrict the texts which were collected to stories centered on the current 'drug-related violence' in Mexico and the United States. While the steps which allowed for this text selection were, are and will continue to be imperfect, the initial piloting process was immensely important because it allowed for the design of the data collection process which was used in the final research.

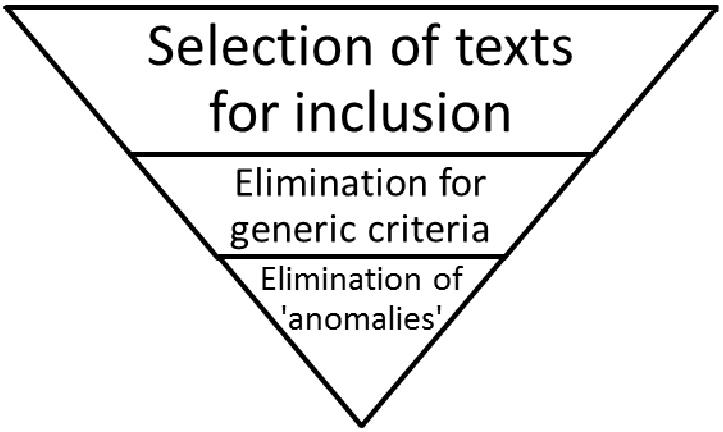
During the course of piloting it became apparent that there would need to be a further step employed in order to effectively limit the automatically collected data used in the final corpora. This was done only in the final data collection and was done in order to assure (based on findings from the pilot study) that the corpora were as representative of the language being studied as possible. Although some of the criteria for text selection were

established beforehand (the use of only ‘hard news’ stories, for instance) others were established based on methodological issues which were found during the course of the pilot study. The criteria used in compiling and constructing the final corpora are laid out in detail below.

3.1.1.2. Text Selection Criteria

In order to analyze the most representative corpora possible for the discourse use being studied, it was first important to establish the criteria which would be used to determine whether or not individual texts would or would not be included in the final corpora. This was done using categories which were established during the pilot study and which were applied after having first collected the data for the final corpora. That is, the search terms (established based on experimentation during the pilot study) were used to compile text corpora; following this, the criteria for text inclusion were applied and texts which fell outside of these criteria were excluded from the final corpora. In the interest of being redundant in selecting texts for the final analysis, the inclusion-exclusion process was carried out in three parts. Each of these parts was progressively less invasive and focused on more subjective criteria.

Figure 1. Process for determining inclusion in or exclusion from final corpora



The first part of the text selection process was the text search. This section of the process was the broadest and also the most objective part of the corpora building process in that it was dependent only on established search terms entered into *Google News*; further, all results which were brought up by this search were examined. Following the search portion of the research, texts were eliminated based on 'macro' criteria. The criteria which were employed in this step of the research process were intended to control for a variety of genre- and sample-based aspects of the final corpora. This was done in order to assure that the two corpora were as similar—and thus, comparable—as possible. While this section was more subjective than the previous step, since the criteria were based on the genre being examined (English- and Spanish-language hard news articles) and easily verifiable characteristics of the texts found in the first step, the researcher played a minimal role in actually 'deciding' on any inclusions or exclusions. In this sense, this step served as a sort of check list, articles were looked at individually and—based on the list of characteristics deemed necessary for corpus inclusion—were included or discarded based on whether or not they possessed the genre characteristics necessary for corpus inclusion (see Table 3, p. 67). Finally, after having constructed corpora which were relatively representative of the

generic and topical features being examined, a second round of eliminations was carried out before arriving at the final corpora (which were used for analysis). This round of eliminations (discussed below) was the most subjective step in the text selection process, but—once again—principally relied on the characteristics of the individual texts being looked at and not on the researcher's own opinions (at least to the extent to which such a thing is possible). Additionally, though more intrusive and subjective than previous steps, this step had the smallest impact on the overall makeup of the final corpora primarily because the majority of articles which could be eliminated already had been.

3.1.2. Sample Selection and Corpus Construction

In order to keep the corpora used in the current study as unbiased as possible, the most minimal criteria possible were utilized in establishing and building both corpora. This was done so as to avoid the very problems—characteristic of similar studies—which this project intended to confront (see Poole, 2010; Prentice, 2010 for discussion). However, these criteria were also established with care in order to ensure that the corpora remained representative of the topic being examined.

The articles which were examined were selected using *Google News* (see Section 3.1.1, p. 58). Through the use of a table of random numbers, a month in 2011 was chosen (October) and all articles included in both corpora had been published during that month. This was done because 2011 was the most recent full year. That is, all stories were as current as possible while still allowing for complete random selection in that articles could have been chosen from any month during 2011.

Upon having randomly selected October of 2011 as the sample month, articles were searched for and compiled using *Google News* by employing a set of parallel criteria for each corpus. Based on findings from the pilot stage, articles were searched day by day (that is, articles were not searched using a date range). This was done because during the pilot period a day by day search was found to yield more accurate and complete results than a search using a date range (i.e. an article search for publications between October 1st and October 31st). First, all articles which appeared on the *Google News* site for a given day were included in the initial sample. Following this, the list (all articles published in the month) was pared down using a set of generic criteria established for controlling the sample size and representativeness of the corpora.

Table 3. Criteria for inclusion in or exclusion from English and Spanish corpora

For Inclusion	For Exclusion
Published in Mexico or US	Published outside of US and Mexico
Published in October 2011	Non-newspaper (magazine, blog post, etc.)
Contain violen* and/or drug* (English)	Non-hard news (editorial, sports, human interest)
Contain violen* and/or droga* (Spanish)	Translations of stories from another country
Have a credited author	Published by wire service (AP, Reuters, 'redacción')

All stories which were included in the final corpora had to have been published online between October 1st and 31st of 2011. All stories included in the final corpora additionally had to have been written by credited writers. That is, no wire service stories, stories reprinted from wire services, or *redacciones* were included. The only exceptions to this stipulation were stories written by credited authors which were submitted to wire services and then distributed (i.e. a wire story with an author's name on the *by-line*). This was done in order to keep the corpora as representative of the discourse being studied as was

possible. Although wire stories obviously have writers, they go through a different editorial process and are written for a more general audience than are articles with credited reporters. Another exception was made for articles written by 'Staff' as this has begun to be a common practice in northern Mexico to preserve reporter anonymity due to violence perpetrated against journalists.

The only articles which were included in the construction of the final corpora were 'hard news' stories. These are stories which are commonly referred to as *headline* or *front page* articles. That is, no opinion, editorial, or human interest stories were included in the construction of the final corpora. Additionally, magazine articles, press releases, articles published by think-tanks, obituaries and stories published in countries outside of the United States and Mexico were excluded from the final corpora. This was done for reasons similar to those which informed the exclusion of wire stories from the current study.

Because the present research is being presented as a response to CDA research, it is important to focus on media discourse and the underlying social and political discourses which inform its development in the public eye. While opinion and human interest writing presents an accurate portrayal of current, popular discussions, they are too representative of 'colloquial' discussion to be included in the current research project. On the other hand, 'hard news' is—in theory, at least—a non-biased recounting of events. As such, these stories are more representative of the sort of discourse which CDA usually examines and, in isolation, represent a genre of print media in which (ideally) no underlying ideological discourse should exist. This is the case because of this 'representativeness,' as well as Carvalho's (2008) assertion that journalistic discourse intersects with every aspect of life

and that underlying discourse found in 'hard news' stories is important in that it will be taken at face value to be 'objective truth' by the audience being exposed to it.

In addition to these generic considerations, all articles used necessarily complied with the following micro-level lexical limitations which, in addition to allowing for an accurate and parallel method for searching and selecting articles, ensured that each corpus was comprised of articles reporting highly similar topics. This was done in order to control for the types of language and discourse topics being examined as much as possible. The first step in applying criteria to article selection was establishing search terms. Following a month-long period of weekly piloting (see Section 3.1.1, p. 58), during which time many approaches were taken to utilizing search terms within *Google News*, the following criteria were established for collecting Mexican and American newspaper articles.

- Mexican articles were searched using: *drogas OR droga AND violente OR violencia*
- American articles were searched using: *drugs OR drug AND violent OR violence*

The use of these particular search phrases was made necessary by the fact that Google does not allow one to search through the use of 'stemming.' That is, a search for *droga* will search every morphological variant of the lexical item (e.g. *droga*, *drogas*, and *drogada*). However, there is no way to search for variants of *droga* when it occurs as the stem of a given word. A search for *droga* will not yield results for a lexical item like *endrogada*, for example. Additionally, during the course of piloting the search terms to be used, it was found that searching *droga*—despite claims made by Google (2011)—limited results far more than if both *droga* and *drogas* were used. This was found to be the case when searching for *violente* and *violencia*, as well; this was the case when searching for

newspaper stories in English as well. The search for published articles containing these LIs formed the first part of the corpus construction process. The initial corpus size (all articles and headlines which appeared in a search—before delimiting them generically—see Table 3, p. 67) was 160 articles published in English and 160 articles published in Spanish.

Upon having selected articles from these initial corpora (based on the corpora selection steps laid out above), two separate corpora were constructed. The English corpus was made up of 33 articles with a total of 24,351 LIs. The Spanish corpus was made up of 24 articles with a total of 12,181 LIs.

Figure 2. Number of articles per day included in English and Spanish corpora

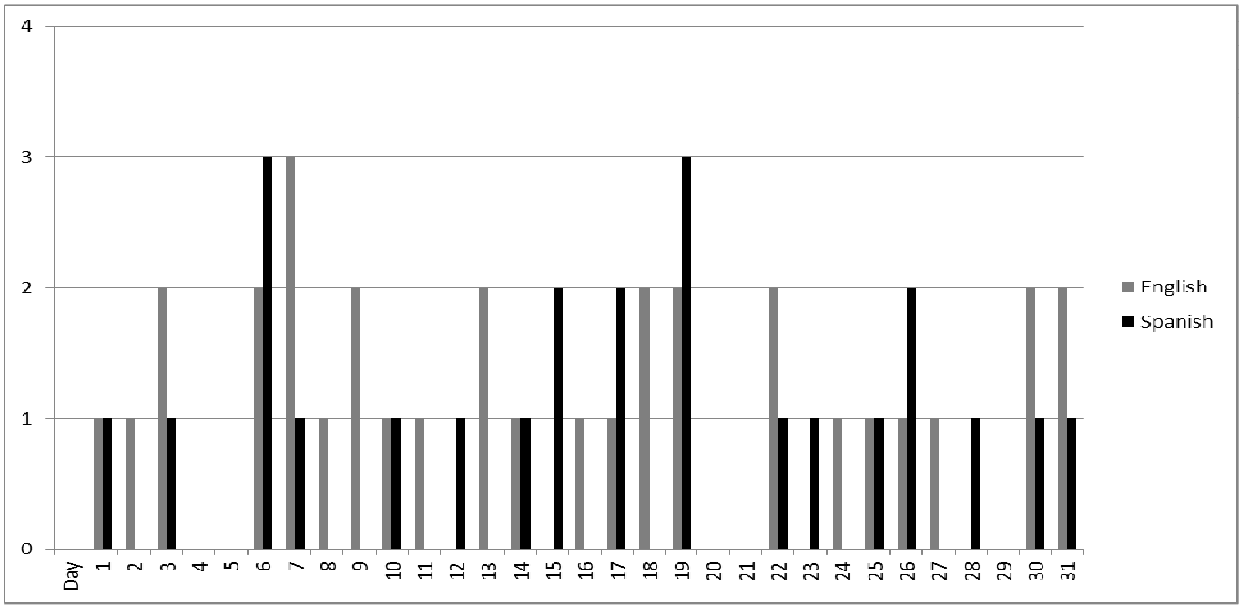


Figure 2 shows the number of articles used in the final corpora—according to the date of publication—before the elimination of 'anomalies' (discussed below). The articles featured in this phase of the corpus construction were those which had not been eliminated for generic features. In total, approximately 80.4% of English corpus articles and 85% of

Spanish articles were eliminated due to country of publication, genre, lack of a credited author, etc. (see Table 3, p. 67).

Based on the newspapers being searched as well as the search terms used, almost all of the texts encountered were related to current 'drug-related violence' in Mexico and the United States. However, there were some 'anomalies' within the texts collected which were eliminated prior to the construction of the final corpora. These 'anomalies' fell into two main categories which for the purposes of the current study were called *lexical* and *topical* anomalies. Although these would seem to be two separate categories of anomalies (and in some cases they were), most of the articles which were eliminated from the final corpora fell into both of these broad categories; lexical anomalies were articles which came up in the *Google News* search due to the presence of either *violen** or *drug** within the text, but not both. Though there were instances of lexical anomalies in which the article was still included in the final corpora (due to being on-topic), all of the texts which were eliminated for this reason were also topically anomalous (e.g. stories about prescription drugs, marijuana legalization or domestic violence). On the other hand, the texts which were eliminated for being topically anomalous were often only eliminated for topic. For example, two articles were eliminated from the final English corpus for being about a police operation to arrest members of a motorcycle gang. In this example, although the articles complied with the search terms established for inclusion in the corpora (both articles contained uses of *violen** and *drug**), topically the texts were not related to the media discourse being studied. As such, they were eliminated from the final corpora. All told, 11 articles (a total of 9,960 LIs) were eliminated from the final English corpus. No

articles were eliminated from the Spanish corpus, indicating that the news reported which featured one or both LIs had to do with the topic being studied.

Table 4. Lexical and topical elimination criteria for English and Spanish corpora

Lexical	Topical
drug*/droga* in non-illicit context	Occupy Wall Street "Bath Salts" Non-Mexican/American (same topic)
violen*/violen* in domestic abuse context	Afghanistan/Iraq Wars US-specific drug/crime stories

Table 4 shows the lexical and topical criteria for elimination from the final corpus. These were established during the pilot period based on the consistent appearance of certain topics. Some of the topical 'anomalies' were very closely related to the topic (e.g. US-specific stories), while others were found due to the lexical characteristics of the article(s) (e.g. those reporting on drug issues in other places). In the case of the latter three topical anomalies the topic being reported on was extremely similar to that being studied, but was eliminated (e.g. drug violence in Honduras). The first two topical anomalies, on the other hand, represented a form of what Baker et al. (2008) called 'seasonal collocates.' That is, topics which were briefly popular in the news media. In this case, both topical anomalies were found in the US corpus and were topics which were heavily reported during the fall months of 2011. Table 5 gives a brief overview of the actual articles eliminated from the final corpus due to being lexically or topically anomalous.

Table 5. Articles eliminated from final corpora

Number of Texts	Language	Topic
1	English	Soldier shooting
2	English	Arkansas drug arrests
2	English	Honduras crimes
1	English	Drug trial
2	English	Motorcycle gangs
1	English	Chinese boat traffic
1	English	Marijuana legalization
1	English	Occupy Wall Street
2	English	"Bath Salts"

Table 5 shows the articles eliminated from the final corpora according to language of publication and topic. These articles were eliminated due to both lexical and topical anomalies and included articles about the Occupy Wall Street movement, the use of 'Bath Salts' in the Northeastern United States, marijuana legalization and a Chinese military operation to stop the use of boats in the methamphetamine trade, among others. Interestingly, the American corpus was the only one which required articles be eliminated for these reasons. Although there were both topically and lexically anomalous texts in the Spanish corpus, they were all eliminated by the time this step was taken due to general criteria (see Table 3, p. 67) such as having been published outside of Mexico, or not having been written by a credited author.

3.1.3. Statistical Analysis of Corpora

Due to the nature of the present study, one of the most vital steps in carrying out a semantic prosody analysis of the text corpora was to first realize a statistical analysis of both corpora. This was deemed important to the current study in that a statistically based analysis serves as a response to many of the weaknesses present in past approaches to CDA

which have already been laid out and discussed. Beginning with the point of view put forth by Carvalho (2008), that an open reading (with limited pre-conceived notions of potential findings or goals of analysis) of texts "…allows for the identification of the most significant characteristics of the data…" (p. 166), it follows that a corpus analysis provides the researcher with a way to carry out this type of 'open' approach to analysis on a large scale. Further, the use of statistical analysis of corpora can permit the researcher to easily focus on LIs which occur with a statistically high frequency, thus allowing the research to more effectively examine the discourse present across texts by establishing 'pre-selection' (Oster, 2010). In this way, a consistent manner of text analysis could be established which was wholly dependent on lexical frequency and—by virtue—textual characteristics. Thus the researcher is able to study 'normal' occurrences within the text and not what they may have noticed or found interesting. This is an important step because it avoids the most common pitfall present in most CDA and semantic prosody studies in which the researcher sets out to analyze a particular word from the beginning of the investigation (thus, in one way or another, shaping the analysis itself) (see, e.g. Oster, 2010; Carvalho, 2008; Louw, 2008).

With these points in mind, the final corpora constructed for the present project were statistically analyzed for frequency and statistically salient LIs were then analyzed using the collocation search tool in the AntConc corpus analysis program (Anthony, 2011). This was done in order to focus on only those lexical features which were prominently featured in each corpus, thus avoiding the common CDA pitfall of "cherry-picking" which features to analyze (Mautner, 2009). Each corpus was subjected to two distinct statistical operations whose purpose was to bring lexical characteristics of each corpus to the surface. This was

done in order to permit pertinent data to 'appear' by virtue of its presence within the corpora. Statistical 'ceilings' were used for the initial frequency analysis and then for the collocation analysis (see section 3.1.4). Together, these 'ceilings' meant that the data examined could verifiably be seen as salient within the corpora.

3.1.4. Node and Collocation Selection

After having constructed both final text corpora, they were first analyzed for the raw frequency of each corpora's individual LIs. This was done using the AntConc corpus analysis program's 'Word List' feature (Anthony, 2011). Using AntConc, lexical frequency lists were obtained for each corpus using the 'treat all data as lowercase' setting in order to be as inclusive as possible in collecting data. In analyzing the corpora (particularly in regard to the frequency lists), the data was looked at in terms of 'types' and 'tokens.' This distinction, according to Kennedy (1998), is one primarily based on a word's underlying function within a corpus; 'types' being individual items and 'tokens' being occurrences of individual types. For example, two distinct morphological realizations of a single LI (e.g. *drug* and *drugs*) are seen to be two tokens of a single type. Nonetheless, for the purposes of the current study tokens were almost exclusively used in analysis. This was done because the broad presence of LIs was what was deemed most important and not merely whether they were present or not. For the analysis of American print media, the corpus used was made up of 14,391 tokens. The Spanish language corpus was comprised of 11,982 tokens.

A frequency list was made first in that it allowed for the remainder of the research steps to be taken effectively. After the initial frequency list was made, all single occurrences were eliminated from it since a single occurrence could not really offer

anything to a study of frequent occurrences. That is, while a LI which appeared once might serve an important purpose within a discourse, without multiple occurrences there was no way to extrapolate conclusions from its presence (for the purposes of the present methodology). Specifically, an item with a frequency of one could be seen as a random occurrence. LIs with a frequency of one were by far the most common features within both corpora. In the entire English corpus, 55.5% of LIs only occurred once; while in the entire Spanish corpus single occurrences accounted for 59.8% of the total number of frequent occurrences.

Table 6. Distribution of lexical items according to frequency in the English corpus

Freq. of Occurrence	Freq. per 1,000 words	% of Total
1 through 3	.0695-0.2085	79.87%
4 through 6	0.2770-0.4169	9.70%
7 through 9	0.4864-0.6524	3.40%
10 through 14	0.6949-0.9728	2.69%
15 through 20	1.0423-1.3898	1.42%
21 through 50	1.4593-3.4744	1.75%
51 through 1,000	3.6134-63.9288	1.13%

Table 6 shows the distribution of LI frequencies in the English corpus. As can be seen, the vast majority of LIs occurred between one and three times in the entire corpus. Table 7 (below) shows the same information for the Spanish corpus.

Table 7. Distribution of lexical items according to frequency in the Spanish corpus

Freq. of Occurrence	Freq. per 1,000 words	% of Total
1 through 3	0.0835-0.2504	85.50%
4 through 6	0.3338-0.5008	7.88%
7 through 9	0.5842-0.7511	2.61%
10 through 14	0.8346-1.1684	1.64%
15 through 20	1.2519-1.6692	0.57%
21 through 50	1.7526-3.6722	0.78%
51 through 1,000	4.7571-78.701	1.77%

After having eliminated all single occurrences from the corpora, the general lexical frequency of each corpus was determined in order to establish a frequency 'ceiling.' This was deemed to be the average frequency of lexical occurrence of the remaining tokens for each corpus (9.25 for the English corpus and 8.53 for the Spanish corpus). As such, the LIs included in the present study occurred with a frequency of ten or higher in the English corpus and nine or higher in the Spanish corpus (with a minimum of 0.6524 occurrences per 1,000 words in English, and 0.7511 in Spanish); having done this, the corpora used for final analysis were considered to be representative of the most statistically salient LIs in each language.

After finding the most statistically salient LIs for each corpus (115 LIs in English and 71 in Spanish), all function words and single letter tokens were eliminated from both corpora (Salama, 2011; Orpin, 2005), thus allowing for more focused analysis in that these LIs were, by and large, the most frequently occurring in each corpus. Both of these items were eliminated for similar reasons, namely that they served a syntactic purpose in the corpora but not a semantic one. The elimination of function words is standard practice in CDA research because they are devoid of meaning and thus serve little purpose in meaning-

based analyses (Mautner, 2009). Similarly, the majority of single letter tokens were grammatical (i.e. plural and possessive markers or parts of contractions).

Additionally, the elimination of these features helped to make both corpora vastly more manageable for analysis. For instance, the article *the*, in the frequency list for the English corpus, occurred nearly nine times as frequently (920 appearances) as did the most frequently occurring content word, *drug* (108 times). Similar characteristics were found in the Spanish corpus where the preposition *de* occurred more than ten times as frequently (943 times) as did the most common non-function word *México* (86 times). Due to the nature of the study being discussed here, both of these steps were included in order to make the total amount of data to be analyzed more manageable and also more lexically transparent. That is, the data which was analyzed could be said to (a) be representative of the most statistically common lexical uses in each corpus (through the use of a frequency 'ceiling') and (b) be representative of meaningful discourse (through the elimination of function words).

The initial frequency lists for the English and Spanish corpora contained 3,081 and 2,981 LIs, respectively. After having eliminated both single token occurrences and function words, the English frequency list contained 115 salient LIs and the Spanish list contained 71. The lists which were obtained at this point served as the corpora from which to draw salient nodes (see below) for the collocation and SP portion of the analysis. A list of the node words used can be found in Appendix A (see p. 135).

3.1.4.1. Sorting of Nodes and Collocations

Once the salient node words were selected from each corpus, the final phases of analysis could begin. This process involved the use of the AntConc corpus analysis program (Anthony, 2011) to search concordances for all node words, the establishment of a 'ceiling' for determining significance of collocate strength and the actual comparison and analysis of the language used in both corpora. The following sections present the methodological tools used as well as a discussion of their importance to and use in the present study.

3.1.4.1.1. Node Words

In all studies which examine the use of certain LIs as part of a KWIC search (such as the present study) the basest unit used in the analysis is the 'node word' (Baker et al., 2008; Xiao & McEnery, 2006; Sinclair, 1991), sometimes also called the 'key word' (see Orpin, 2005). The node word is a given LI which is studied alongside its collocates (Sinclair, 1991).

Although there are multiple ways to refer to a LI within a corpus as well as its classification within said corpus, the most widely used distinction is that which separates 'types' from 'tokens' (Kennedy, 1998). In the present study, the AntConc corpus program (Anthony, 2011) was used to analyze the presence of both types (all lexical features in the corpora) and tokens (those present in the frequency list). While this distinction is an important one in laying out the groundwork for a lexically based corpus analysis (such as in Kennedy's case), for the purposes of the current project, the brunt of the methodological focus was placed on what Kennedy (p. 251) calls the, "…target item, node word or search

item." Although Kennedy and many other authors have alternately used the term keyword as a synonym (Prentice, 2010), the terms 'node' and 'node word' will be employed here as they are most commonly used in studies which principally focus on concordance analysis (see Salama, 2011; Oster, 2010; Louw, 2008).

The node words which were found through the previous methodological steps (Section 3.1.4, p. 75) were used as the basis for collocation analysis. That is, they served as the nodes for the KWIC portion of the present study (see Tables 8 and 9). Tokens which are being examined in their capacity as node words are included in all capital letters.

Table 8. Top ten most frequently occurring nodes in the English corpus

Node Word	Raw Freq.	Freq. per 1,000 words
DRUG	108	7.50
MEXICO	99	6.88
BORDER	66	4.59
MEXICAN	61	4.24
CARTEL	60	4.17
STATE	52	3.61
OFFICIALS	45	3.13
POLICE	44	3.06
ONE	40	2.78
VERACRUZ	39	2.71

Table 9. Top ten most frequently occurring nodes in the Spanish corpus

Node Word	Raw Freq.	Freq. per 1,000 words
MÉXICO	86	7.18
VIOLENCIA	59	4.92
DROGAS	57	4.76
ESTADOS	44	3.67
PAÍS	39	3.25
UNIDOS	33	2.75
SECUESTRO	32	2.67
GRUPOS	28	2.34
AÑOS	27	2.25
ESTADO	27	2.25

Tables 8 and 9 show the ten most frequently occurring nodes in each corpus along with the raw frequency and a normalized frequency of occurrence for each of them; all nodes included here (as discussed above) were content words which occurred at an above average rate of frequency in each corpus (see Section 3.1.4, p. 75). Upon having drawn up a list of all the nodes to be used, the next step in the research process was a concordance analysis of these.

3.1.4.1.2. Concordance Analysis

Having found pertinent node words for each corpus, these were then analyzed for concordance features using the AntConc corpus analysis program (Anthony, 2011). This was a necessary step in carrying out a SP analysis of the data contained in both corpora (see Louw, 2008; Partington, 2004). As Partington points out, the cornerstone of SP theory hinges on the idea that most words have inherent evaluative meanings (i.e. negative, positive, or neutral connotations) and that these can and are used with specific ends depending on the author or speaker's intentions. Because of this, it is possible to examine the consistent in-text behavior of certain LIs using corpus analysis tools in order to observe the prosodic characteristics of a given word. One of the most frequently used ways of going about this is to examine the collocates of specific LIs to analyze patterns in their use. This has been done frequently by many authors (a thorough discussion of this can be found in Partington (2004)). The present study employed a concordance analysis using salient node words—the selection of which is described above.

Because semantic prosodies are formed based on the co-occurrence of certain LIs (Louw, 2008), it was first important to determine what some of the salient co-occurrence patterns were in the corpora used here. With this in mind, every node word in each corpus (135 English LIs and 65 Spanish LIs) was searched using the *Collocates* tool featured in the AntConc corpus analysis program (Anthony, 2011). This feature allowed for nodes to be searched within a given corpus and presented the collocates of each. The tool also can be adjusted to accommodate different spans of text around a node, and can show how frequently the node and collocate occurred together and how often a given collocate occurs before or after the node; the tool also features a statistical tool which can be used to analyze different facets of a given search term's co-occurrence patterns.

For the current research, the concordance analysis was comprised of two main methodological processes. As with previous research steps, these two points focused on separating the most salient data possible while simultaneously relying on as little subjective input from the researcher as possible. The node words were searched according to some of the stipulations set forth by Salama (2011, citing Hunston, 2002). First of all, collocates were searched for within a five-word window on either side of the node words. For example, in a search of the node MEXICO in the English corpus, the tokens which appear on either side of the node in the sample KWIC analysis presented below (Figure 3) represent the environment within which collocations were searched by the program.

Figure 3. Sample KWIC presentation from English corpus (MEXICO)

Beltran Leyva cartel, according to	**MEXICO**	City s Attorney General Miguel Ángel
dividing the city from neighboring	**MEXICO**	state along a busy road by the Defense
yet to find the bodies. While	**MEXICO**	's on-going drug war has made its way

As can be seen in Figure 3, the corpus analysis program shows the node word in context. Having done this it is then possible to examine the collocates which frequently co-occur with the node. In the case of Figure 3, one can easily see that the collocate *city*, for example, co-occurs twice with the node in three contexts, suggesting a strong collocation. Based on this type of analysis, once a search is entered, the AntConc program (Anthony, 2011) produces a list of frequent collocates along with all pertinent information about them, their presence in the corpus, and how they co-occur with the node word being searched.

The present study was carried out using a list of statistically determined node words from each corpus. These words were searched for one at a time using the AntConc corpus analysis program (Anthony, 2011). In order to search the node words effectively and in a parallel manner within each corpus, the following steps were taken and the settings employed are described. In the interest of including all possible tokens of a given type, LIs were searched using the 'treat all data as lowercase' setting in AntConc (Anthony, 2011). This means that a search for the node CITY would generate results for the LI's use as part of a name (as in, *Mexico City*) as well as a noun used to describe a populous grouping of inhabitants bigger than a 'town' (both instances of this node's use are present in the above KWIC—see Figure 3, p. 82). This insured that no use of a given LI would go unnoticed when working with the corpora. Additionally, the data shown and consequently used in the concordance analysis had to have been present two times or more within the corpora. This decision was very similar to the choice to eliminate single frequencies from the frequency list and was used in the data collection portion of the research for the same reasons. Namely that no matter how strong a given collocation is found to be, it becomes difficult to infer anything about its use should it only occur once in a corpus.

In order to analyze the strength of the individual collocations found through these steps, MI scores were used. MI scores are a form of 'association measures' (Bouma, 2009). The use of MI scores allows the researcher, "…to rank candidates extracted from a corpus…" (p. 58). Upon having selected the top ranking 'candidates,' these can be seen as co-occurring significantly and can then be analyzed. The use of MI scores has been used in many similar corpus analysis projects (see Salama, 2011; Oster, 2010; Xiao & McEnery, 2006) and has been found to be quite effective in highlighting patterns of co-occurrence in many differently sized corpora (Xiao & McEnery, 2006). The use of MI scores provides the researcher with a number corresponding to the 'strength' of a given collocation. In cases of MI's use in similar studies (see Salama, 2011; Xiao & McEnery, 2006) the number considered representative of significant collocation has been three or higher (indicating a co-occurrence which can be said to have not occurred by chance). In the interest of limiting the size of data output and being as methodologically rigorous as possible, the current study examined those collocates found to have an MI score of four or higher. Thus, all collocates analyzed here had both a MI score of four or higher and co-occurred with their corresponding node two or more times within the corpus. Upon having extracted only these collocates from the full list of collocates for each node, the actual analysis of each corpus' language use could begin.

3.1.5. Final Analysis of Corpora

After having constructed and statistically analyzed both corpora and having searched and sorted the collocates of all statistically salient node words found therein, the final step in conducting the analysis was to look at the data obtained from a SP perspective. This was done in two phases. First, data was looked at according to the raw frequency of certain LIs

(analyzed as nodes); and second, any node words which occurred in both corpora were examined side by side. Although when looked at as part of the entire analysis process (along with corpus construction, and both frequency and collocation analyses) the SP-based portion of the present study may seem to make up a small part of the methodology, it was the culmination of the preceding methodological steps.

In the present project, the goal was to locate instances of SP in the corpora themselves. This was accomplished by examining the data collected in both corpora using the AntConc corpus analysis program (Anthony, 2011). The first part of the SP analysis involved the examination of the most frequently occurring LIs and their collocates in both corpora. Following this, the same analysis was carried out using any LIs which had an equivalent in the other corpus (e.g. DRUG and DROGA or MEXICO and MÉXICO). SP characteristics were determined for each item based on linguistic competence. That is, since prosodic characteristics are natural in all words, competent speakers are able to determine these features based on experience. While this has been debated in SP circles (see Hunston, 2007), there is a wealth of information to substantiate the existence and consistency of SP's presence in language. SP can be seen in many simple examples of language use.

Obviously, a competent speaker of English will never say that a word like *happiness* has negative connotations and, likewise, a competent Spanish speaker will not say that a word like *muerte* has positive connotations unless speaking out of context or employing humor. However, this can be seen even clearer in looking at the use of linguistic irony. SP is particularly obvious in irony precisely because many uses of linguistic irony are accomplished by invoking the opposite prosody in a word (e.g. sarcasm) (see Partington (2007) for complete discussion).

The analysis of the SP characteristics of node words and their collocates was by far the most novel part of the present study in that not only has SP not been extensively studied, but it has rarely—if ever—been applied to CDA and corpora in the way that it was here. Essentially, once a set of collocates was extrapolated for a given node word, they were examined for their prosodic features, both on their own and in relation to the linguistic environment in which they were found to be present in the corpora. As in other studies utilizing SP as a methodological tool, prosodic characteristics were seen as part of a given word's "DNA" (Morley & Partington, 2009) that is, there is no discussion as to whether a word is positive or negative; it simply is. Being as there is no way—at least presently—for a researcher to determine a LI's prosodic characteristics without involving themselves in making a determination, that is precisely what was done here. As a competent speaker of both Spanish and English, the prosodic characteristics of individual LIs were determined by the researcher. Words known to have a positive prosody were deemed as such and words deemed to be 'negative' were treated likewise. Any words of which there was doubt were treated as neutral (these were principally titles and geographical terms). This methodological process, as well as the results which were obtained using it, is discussed in more detail in the following chapter.