RiCL Research in Corpus Linguistics

Building the *Great Recession News Corpus* (GRNC): A contemporary diachronic corpus of economy news in English

Javier Fernández-Cruz[ab] – Antonio Moreno-Ortiz[a]
Universidad de Málaga[a] / Spain
Pontificia Universidad Católica del Ecuador Sede Esmeraldas[b] / Ecuador

Abstract – The paper describes the process involved in developing the *Great Recession News Corpus* (GRNC), a specialized web corpus which contains a wide range of written texts obtained from the business section of *The Guardian* and *The New York Times* between 2007 and 2015. The corpus was compiled as the main resource in a sentiment analysis project on the economic/financial domain. A justification of the corpus design is provided, along with the methodology followed for the compilation process. To evaluate its usefulness, we include a sentiment analysis study on the evolution of the sentiment conveyed by the word *credit* during the years of the Great Recession.

Keywords – corpus linguistics; financial discourse; crisis studies; information retrieval; sentiment analysis

1. INTRODUCTION[1]

This paper describes and justifies the design and implementation of the *Great Recession News Corpus* (GRNC), a 21-million token compilation from 42,193 online business news articles of *The Guardian* and *The New York Times* published between 2007 and 2015. Our corpus serves as a useful linguistic resource for the scholarly research in multiple fields, such as English for Specific Purposes (ESP), comparative journalism or crisis studies, as it attempts to capture the impact that the Great Recession had on language.

The starting point of the GRNC is January 2007, coinciding with the emergence of the subprime crisis and the collapse of Lehman Brothers (2007–2008), which triggered a domino effect that was the foundation stone of the so-called 'credit crunch'. In addition, the corpus covers the Europe-centered aftershock, the 'sovereign debt crisis' in the

[1] This research was funded by the Pontifical Catholic University of Ecuador's Research Fund 2017: project LexiEcon. We would also like to thank the anonymous reviewers whose comments have greatly improved this manuscript.

European Union, which intensified from 2010 on and the impact of the "Whatever it takes" speech by Mario Draghi (European Central Bank 2012), and its final coverage coincides with the announcement of the European Central Bank's *Quantitative Easing Program* (European Central Bank 2015). Apart from including the major economic events, the text compilation offers a full coverage of the socio-political crisis provoked by this instability, and the social response to a massive decline of the living standards in the daily lives of common people around the world.

Economic news coverage exposes the causes and reactions generated by this crisis. According to Lischinsky (2011: 154), different discursive frames "can have major effects in public understanding and policy decisions." As a consequence, the GRNC allows to observe the discursive underpinnings of possible solutions from an ideological perspective at multiple linguistic levels (lexical, semantic, textual, etc.).

Another interesting factor here is the co-occurrence of the Great Recession with the decline of the so-called 'old media' or 'legacy media' (i.e., centralized printed newspapers or one-way broadcast technologies) and the rise of the age of digital media, which has been widely covered by the literature (Newman 2009; Huxford 2012; Franklin 2014). As the GRNC is composed entirely of web news, it may serve as witness to the innovation and radical changes across all aspects of a journalism already in search for alternative business models to start a sustainable journalism model for the future.

The origins of the GRNC go back to the research design of our main project: the development of a lexicon-based sentiment analysis (SA) system of financial texts with an appropriate treatment of the terminology in use during the Great Recession. In order to analyze these terms, a sentiment lexicon in the financial/economic domain, *SentiEcon*, was compiled from the corpus as a plugin lexicon for the *Lingmotif* sentiment analysis tool (Moreno-Ortiz 2017a, 2017b). Thus, corpus tools and techniques were used to (a) create a lexicon of sentiment words in the economic domain of the English language, and (b) to serve as a solid textual platform for observing the short-term diachronic — 'brachychronic' if we follow Renouf's (2002: 30) definition— evolution of sentiment in different lexical units within that domain.

In this paper we define the criteria used for selecting texts and we also explain the techniques we employed to process, organize, clean, and annotate the texts. For illustrative purposes, we also present a brief example of the research possibilities that this

resource offers in sentiment analysis. Finally, we discuss its limitations and future perspectives.

2. JUSTIFICATION

The GRNC consists entirely of journalistic articles from the business section of the newspapers *The Guardian* and *The New York Times*; thus, the textual typology corresponds to a specialized corpus. Following the classification of publicly available corpora used by McEnery *et al*. (2006: 59), our corpus is a diachronic written monolingual corpus of business news. Since our aim was to study the evolution of sentiment conveyed by financial-economic terms as a result of the economic crisis, the GRNC is annotated by time of publication, covering a nine-year span (2007–2015), and organized monthly, resulting in a time series of 9x12 data points.

The analysis of the GRNC may provide an authentic overview of how news texts contribute to the linguistic construction of social reality. Both dailies publish with a view to influencing not only their readers, but also the discourse of the international press. In accordance with Bednarek and Caple (2012: 20–25), this motivation is justified, on the one hand, by the abundance of texts and, on the other, by the great exposure that the public has to news. Social reality is linguistically constructed, and such a construction is largely shaped by the view of journalists (Schudson 1989). When examining the use of crisis-related terms in leading media such as *The Guardian* or *The New York Times*, we are confronted with a type of language which narrates events to the public using carefully selected terms that depart from the specialized domain of economics. Both publications are highly authoritative internationally, and recognized for their stylistic influence and their ability to set the agenda (Van Belle 2003; Golan 2006).

The link between print media and the language of ordinary people is as old as the print itself, and reviews the central role of the popular press as a social educator (Conboy 2006: 9). The emergence of social media and the spread of viral news (Al-Rawi 2019) has served as an accelerator for the dissemination of new uses and meanings of words and terms through the articles that opinion makers have generated since it became widely available.

The novelty of this medium is bidirectionality. The online versions of traditional newspapers have progressively adapted to the needs of their online readers, who have

become highly influential and, to a large extent, their discourse is modulated directly or indirectly by the mediation of its users through social media impact metrics or comments on social networks (Chung 2018). Nafría (2017: 236) argues that the challenge of adapting the headlines of *The New York Times* to Web 2.0 journalism implied a new discursive consolidation which required a combination of analytical journalism and simple language. As a result, newsrooms have designed social media policies to "guide newsworkers through the difficult intersection of traditional journalism and social media" (Duffy and Knight 2019: 932). Other relevant changes in online news exceed the textual level by incorporating multimodal items (e.g., animations or videos) or the appearance of new textual formats in their sections (i.e., blogs or microblogs). Paradoxically, the vast diversity of opinions that can be read on social networks (e.g., *Twitter*) has not diminished the influence of large media emporiums, but is thought to have increased the influence of traditional media on both the public and the stakeholders' opinion (for a thorough discussion, see Etter *et al*. 2017 and Blevins and Ragozzino 2019).

Due to the nature of our project, focused on sentiment analysis, news items are ideal for this task, as they are rich in evaluative language. Opinion is a key factor in the business sections of generalist media, since newswriters need to interpret macroeconomic figures and institutional statements in order to communicate this information to the public. Socioeconomic changes, as reflected in the texts, contribute to the construction of a value system, as understood by Thompson and Hunston (2000), which is built by the speaking community through evaluations. This system transcends as a component of ideology that permeates though the linguistic combinations and constructions of each of the texts. There is a vast array of definitions and discussions of the term 'evaluation' in linguistics and, in our view, Alba-Juez and Thompson's (2014: 13) is the most comprehensive one:

> a dynamical subsystem of language, permeating all linguistic levels and involving the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that s/he is talking about, which entails relational work including the (possible and prototypically expected and subsequent) response of the hearer or (potential) audience. This relational work is generally related to the speaker's and/or the hearer's personal, group or cultural set of values.

The next step is to describe the features of the GRNC in order to implement a solid research framework. Corpora must be defined in terms of size, representativeness and balance (Xiao 2010: 148–153). As for size, Bowker and Pearson (2002: 49) consider that

there is not a pre-established ideal number of words, since this depends mainly on the purpose of the study. While Sinclair's (1991: 18) maxim "a corpus should be as large as possible and should keep on growing" is still valid, even a small corpus can be a very useful resource if it is well designed. In particular, it is generally accepted that the size of a specialized corpus is generally smaller than that of a general corpus. Still, the GRNC is, however, significantly larger than other related corpora (see Section 3).

Representativeness is defined by Biber (1993: 243) as "the extent to which a sample includes the full range of variability in a population." Huan (2018: 57), however, questions this simplicity of operationalization, as different meanings of representativeness may emerge because, in contrast with general corpora, "most specialized corpora have already focused on special domain, time, and medium of the data." In our case, the main focus of the GRNC is hard news (domain) published online by two major British and American daily newspapers (medium) over the period between 2007 and 2015 (time).

The business section was selected because both newspapers fulfilled the following quality criteria: (1) the homogeneity of their language; (2) the editorial committees and the authors are representative experts of the domain; (3) an informative/didactic use of specialized language is made, so that it serves as a link between the specialist's discourse and the public; (4) the wide availability of the texts on the Internet; (5) the coverage of the main varieties of the English language; and (6) their online versions had free open access at the time of the compilation.

As for domain and medium representativeness, according to *ComScore* (2012), 644 million people worldwide accessed online newspaper sites in October 2012, representing 42.6% of the total Internet user base. Among reader popularity worldwide, *The New York Times* and *The Guardian* ranked second (48.7 million) and third (38.9 million), respectively. The business section of both dailies includes in-depth US/UK and international market news coverage, as well as company research tools. This section also includes international news involving political relations, finance and economy-related social issues. Texts include not only summaries of press conferences and economic reports, but also their interpretations, in the form of opinion columns, interviews, as well as live coverage of different events of interest and journalistic commentaries on the reactions of the public on social media. During the most turbulent events, both online sections included live coverage of major international events, such as meetings of the

Eurogroup. In addition to institutional coverage, possibly as a counterbalance, both media published crisis-related news related to the impact of the crisis on the common people, depicting social repercussions, such as the effects of mass unemployment, evictions, etc.

The aforementioned features do not qualify our corpus as representative, however. McEnery *et al*. (2006: 16) consider that specialized corpora are representative when the linguistic features at issue in the design are "subject to very limited variation beyond a certain point." In relation to this, Huan (2018: 57) argues that the previous consideration of representativeness in specialized texts does not occur without criticism, since it involves examining the "linguistic variability" (lexical, syntactic, etc.) of a corpus at the expense of "situational variability" (i.e., the range of genres and registers in the target population). In any case, the linguistic criterion can serve to test the skewness of a corpus collected in line with the situational criterion. Finally, Tognini-Bonelli (2001: 57–59) considers that the representativeness of a corpus can hardly be evaluated in objective terms, and ultimately relates to the question of balance.

For Sinclair (2005: Section 5), balance implies that "the proportions of different kinds of text it [a corpus] contains should correspond with informed and intuitive judgements." The balance of a corpus must be determined by the nature of the corpus and its intended research application (Xiao 2010: 149). McEnery *et al*. (2006: 16) debate the methodological problems behind Sinclair's definition and contend that a reliable scientific measure of corpus balance has not been set. They also consider that any statement of corpus balance in the literature is very much an act of faith rather than a factual statement. In addition, Douglas (2003: 34) considers the balance of a corpus to be secondary to good research practice and, consequently, the resulting compilation must address research questions adequately and offer transparency in the documentation.

The GRNC also attempts to cover the two main varieties of English equally. Thus, the texts in *The Guardian* (British English) account for 47% of the corpus, while the remaining 53% was extracted from *The New York Times* (American English).

3. Related work

Corpora from the domains of economy, business and finance are compiled for diverse purposes (e.g., language for specific purposes, terminology or natural language processing). The growing awareness of Great Recession-related corpus research has led

to different text compilations with a high disparity of sizes and purposes (i.e., discourse analysis, metaphor analysis, social network analysis, etc.). Nevertheless, to our knowledge, the GRNC fills an important gap, since no other English language corpus covers the main topics and features required for Great Recession journalistic or linguistic research. An array of examples of economics, business press and economic crisis-related corpora are reviewed synthetically in this section. In general terms, all prior work reviewed here can be included in one of these two genres: business communication or business news. The main business communication-related corpora are the following:

- The *Cambridge and Nottingham Spoken Business English Corpus* (CANBEC) (Handford 2010), one of the most widely distributed ESP corpora. It is an oral corpus that includes 912,734 words from 64 business meetings in 26 companies. It transcribes formal and informal meetings, presentations, phone conversations, etc.
- The *Hong Kong Financial Services Corpus* (HKFSC) (Li and Qian 2010) categorizes a total of 25 text types (among others, annual reports, fund description and speeches) in order to present a comprehensive picture of the written discourse in the financial services industry in Hong Kong. As of 2020, it is readily available online and includes more than 7 million words. It has been developed by the *Research Centre for Professional Communication in English* at the Department of English of the Hong Kong Polytechnic University.
- The *Malaysian Corpus of Financial English* (MaCFE) (Sadjirin *et al*. 2018) is a specialized online corpus which contains 4.3 million words from 1,472 electronic documents retrieved from banks and financial institutions' official websites.
- Diesner *et al*. (2005) created a complex network corpus containing 252,000 corporate emails in order to observe the characteristics and patterns of communicative behavior of Enron employees during the different stages of its collapse.
- Lischinsky (2011) built a corpus of 50 financial and corporate social responsibility reports of Swedish companies in 2008 totaling 1.5 million tokens.

As for business news corpora, the following are noteworthy:

- The *Reuters Corpus Volume 1* (Rose *et al*. 2002) is a freely available archive of 806,791 English language *Reuters* news between 1996 and 1997. It covers news from different economic subdomains: corporate/industrial, economics, government/social and markets.

- Schröter and Storjohann (2015: 50) built a 4-million token "thematically homogenous 'purpose-built' corpus" which includes the keyword *financial crisis* in British newspaper articles from 2009.
- Rojo and Orts Llopis' (2010) *English-Spanish Parallel Corpus* covers both *The Economist* and *El Economista* between two periods: the first one concerning the subprime crisis (June to November 2007), and the second one the era of the collapse of Lehman Brothers (September to December 2008).
- *Corpus de la Crisis Financiera* (CCF) (Botella *et al*. 2015) provide a snapshot of the opinion columns in Spanish daily papers *El País* and *El Mundo* throughout 2012.

4. METHOD FOR CORPUS COMPILATION

In order to extract the texts, we decided to employ a custom semi-automatic procedure, since, despite the existence of many scrapers and other information extractors, no tools were found to fully satisfy our needs. We also intend to provide a concise and clear description of this pipeline in order to offer a simple, step-by-step guide for all levels of expertise. Our procedure may be summarized as follows:

1. Extraction of the URLs of each news item using mixed techniques.
2. Scraping of HTML files.
3. Extraction and cleaning of texts.
4. Classification, labelling and post-processing of corpus files.

In the first step, all the public URLs of the business section of both digital editions were extracted. To obtain good results, we used a monthly *Google Advanced Search*[2] to find all URLs containing the /business/ pattern from the domain of each newspaper.[3] All URLs were extracted with *Link Klipper* (2017), a simple yet very powerful browser extension, and then exported to a text file.

In order to scrape HTML files, we used the *Linux wget* tool, a simple command line utility for downloading files from the Internet. As input, we used text files containing the extracted hyperlinks, which allowed us to download all HTML files containing the news

[2] We are aware that *Google's* personal search history can rearrange the order of matches. However, the influence of the said order is negligible since all the links were extracted.
[3] http://www.nytimes.com and http://www.guardian.co.uk.

items. Next, we used a custom shell script, available under demand, which classified the downloaded HTML files, both chronologically and by publisher, and discarded irrelevant (e.g., files containing no text) and repeated files. Finally, the files were cleaned automatically using the *BootCaT* utility (Baroni and Bernardini 2004) in order to keep labels such as <h1> or <p>, and discard all other irrelevant interface formatting elements. As a result, a typical corpus document contained headlines, sub-headlines and body text.

For an efficient search that allows us to observe the context of key terms chronologically, it was necessary to carry out a simple cataloguing of the texts. To this end, each of the files of the GRNC was named in a standard way to include the following coded metadata:

- The date of publication of the texts: encoded as four digits (YYMM, year-month). Thus, the date of a file published in August 2013 would be coded as "1308."
- The name of the newspaper. In this case there are two *The Guardian* (GU) and *The New York Times* (NYT).
- A numeric ID code to identify each article, so that each of the text files received a unique identification code.

An example of a filename would be 1303NYT103.txt. This file would correspond to an article published in March 2013 on *The New York Times* with 103 as an identification number.

All text files were uploaded to *The Sketch Engine* (Kilgarriff *et al*. 2014) and subsequently compiled. As a result, all texts were tokenized and parsed automatically with *Penn Treebank POS-tagging Sketch Grammar for English TreeTagger version 3.1* (Marcus *et al*.1993).

5. CORPUS DESCRIPTION AND PRESENTATION

Table 1 summarizes the final composition of the GRNC: 42,193 texts containing 21.27 million words and 24.87 million tokens (i.e. words, punctuation, digit, abbreviations, product names and clitics). As for the lexicon, the corpus includes 242,000 different words grouped into 942,000 sentences.

Source	Tokens	Words	Texts	Sentences	%
The Guardian	13,197,301	11,285,112	21,312	477,165	53.06
The New York Times	11,673,704	9,982,273	20,881	465,753	46.93
TOTAL	24,871,005	21,267,385	42,193	942,918	100

Table 1: Description of the GRNC corpus by source

Figure 1 provides a breakdown of the number of tokens by year and publisher.

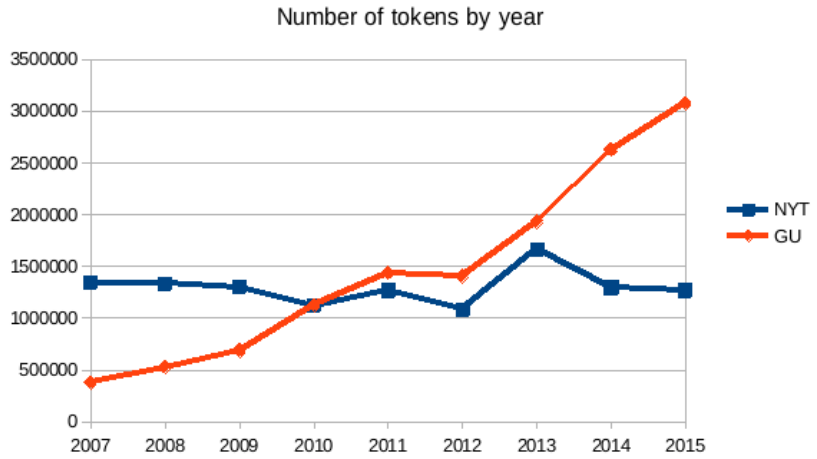

Figure 1: GRNC data number of tokens collected by year and publisher

The GRNC is available at *The Sketch Engine* by request for non-for-profit researchers. This platform was selected for its management, processing and dissemination possibilities, as well as the fact that our corpus can be used in combination with other 500 corpora in more than 90 languages that cover multiple language varieties.[4]

The GRNC can also be accessed as multiple subcorpora and, as a result, allow complex searches by year and publisher. For illustration purposes, a word frequency list from the corpus containing its most significant lexical items can be observed in Table 2.

[4] Visit http://tecnolengua.uma.es/grnc for more details. The corpus does not provide URLs as this data tend to change over time due to website rearrangements. For instance, *The New York Times* is currently behind a paywall.

Nouns		Adjectives		Verbs		Adverbs	
Lemma	**Freq.**	**Lemma**	**Freq.**	**Lemma**	**Freq.**	**Lemma**	**Freq.**
year	88,588	*more*	45,369	*be*	725,459	*not*	113,937
company	86,955	*new*	41,949	*have*	282,022	*also*	43,793
Mr.	57,976	*last*	39,677	*say*	182,133	*now*	27,440
business	55,435	*other*	33,712	*do*	71,287	*more*	27,437
market	38,291	*good*	23,764	*make*	52,053	*so*	24,910
people	35,747	*many*	23,179	*take*	35,327	*as*	22,641
bank	32,464	*big*	22,608	*go*	29,434	*just*	20,774
time	31,800	*chief*	21,624	*get*	27,609	*well*	19,668
price	29,283	*first*	20,579	*include*	27,217	*about*	19,171
sale	26,594	*high*	20,000	*use*	26,304	*even*	18,811
government	26,369	*financial*	19,682	*come*	24,870	*only*	17,179
percent	24,968	*large*	17,257	*work*	23,216	*still*	16,235
UK	24,502	*such*	16,084	*see*	21,385	*most*	13,463
month	24,187	*small*	13,654	*pay*	21,134	*very*	13,156
executive	23,704	*next*	13,219	*sell*	20,056	*then*	13,031
country	20,895	*economic*	12,472	*give*	19,138	*back*	11,939
share	20,586	*global*	12,139	*help*	18,766	*much*	11,294
industry	20,088	*low*	11,613	*need*	18,075	*too*	10,277
group	19,218	*own*	10,661	*want*	17,524	*already*	10,165

Table 2: Word frequency list in the GRNC

6. Sentiment analysis applications

We believe that this corpus offers a wide range of possibilities, from the observation of terms in use, or the analysis of new words or expressions in linguistics, to various applications in the digital humanities, such as contemporary historiography, and studies in behavioral economics, discourse analysis or compared media studies.

For illustration purposes, we briefly present here a study of the evolution of the sentiment conveyed by the term *credit*, an 'event word' (*mot événement*) during the Great Recession. According to Moirand (2007: 4), certain lexical units belong to a specific domain without connotations. After an event of a certain magnitude (e.g., the ongoing COVID19 crisis) that receives widespread media attention, these lexical units acquire connotative meanings related to this situation in particular.[5] As a consequence, these terms tend to appear in new contexts with new collocates that frequently may carry negative (or positive) sentiment and end up acquiring the sentiment of its collocates.

We extracted a data set of all the sentences from the corpus containing the keyword *credit* (*n*=6,764), and then proceeded to analyze it with the *Lingmotif* sentiment analysis software (Moreno-Ortiz 2017a, 2017b) in conjunction with the *SentiEcon* plugin lexicon

[5] Note the recent release of a brand new *Coronavirus Corpus*: https://www.english-corpora.org/corona/ (27 May, 2020.)

(Moreno-Ortiz *et al*. 2020). *SentiEcon* is a specialized sentiment lexicon on the financial domain. It contains 6,470 entries, both single and multi-word expressions, each with tags denoting their semantic orientation and intensity. It was extracted in its entirety from the GNRC. The main objective is to conduct a longitudinal study on the semantics of the word *credit*, from the sentiment perspective, by correlating the semantic orientation of the contexts in which this key term appears through the years that the corpus covers with the historical events that took place during that time.

In Figure 2 below, the resulting sentiment scores (first plot) and frequency trends (second plot) are presented in a time series. In order to smooth out random noise and seasonality in our plots, we calculated the yearly, rather than monthly, average of sentiment scores. We then used *The Sketch Engine* to extract the most frequent collocations of the term yearly. *LogDice* was selected as a statistic measure because it subsumes frequency and exclusivity of collocation. In addition, it is a standardized measure (range of 0–14) that avoids the bias produced by the different size of annual subcorpora (Gablasova *et al*. 2017: 164–166).

Figure 2: Sentiment and relative frequency plots for the term *credit*

In the sentiment plot, three clearly different trends can be observed:

a. The first stage corresponds to the dawn of the credit crisis in 2007, when *credit* was used in contexts with a positive semantic orientation. The analyzed sentences prioritized the use of specialized lexical items with neutral sentiment, including stable clusters (e.g., *credit card*, *credit line*) or other collocations with words such as *carbon*, *market* or *tax*, as illustrated in examples (1a) to (1d).

(1a) Gazprom, using Kyoto guidelines, plans to sell carbon CREDITS to Europe.
(1b) His business earns a tax CREDIT for hiring former prisoners.
(1c) Moreover, modern consumers love their CREDIT cards
(1d) Also, banks have traditionally had a monopoly on CREDIT and savings.

b. A second three-year phase (2008–2011) characterized by the sudden drop to a negative sentiment threshold. In addition, it can be observed that the relative frequency doubles in 2008 (0.57 per 1,000 words) compared to the previous year (0.28 per 1,000 words). These data correspond to the bursting of the housing bubble and the events that caused the credit system to freeze. The context of *credit* is characterized by more specific domain collocates that generally carry negative sentiment i.e., nouns: *default*, *squeeze* or *loss*, and adjectives such as *tight*, as illustrated in (2a) to (2d).

(2a) For a student, a default can destroy a CREDIT record, making it hard even to rent an apartment, let alone buy a home.
(2b) That simply doesn't compare to the 150% bubbles we saw in some of the countries that were hit by the CREDIT crunch.
(2c) As the fund was being wound down, UBS said about 70 percent of its losses came from exposure to CREDIT default swaps.
(2d) Just as in the mortgage markets, a sterling CREDIT rating –the bond insurer's seal of approval– is no longer trusted.

c. Semantic orientation is again reversed in 2012 and remains stable until 2015, while the relative frequency followed a slightly descending trend until the end of the series. It is pertinent to recall that by this time the journalistic machinery had set in motion a discourse in favor of reactivating credit from the central banks to the private banks. By then, the US Federal Reserve was consolidating its program of quantitative easing through the purchase of bank assets, and the new discursive paradigm following Mario Draghi's famous speech in 2012 (European Central Bank 2012) was underway. Here, among the collocates of *credit*, we can find

specific domain units and some previously absent positive items, e.g., *help, expand* or *cheap,* as illustrated in (3a) to (3d).

(3a) The SEC has disputed accusations that it has not done enough to tackle the individuals and companies that helped cause the credit crunch.

(3b) Legal to Censor, but Unwise Gabe Rottman, American Civil Liberties Union Pulling credit card services would help the haters and hurt free expression.

(3c) Cheap credit is essential when households and businesses are close to going bust.

(3d) The ECB has already taken steps to expand the supply of credit in an effort to drive down borrowing costs and ease pressure on household budgets.

It is then apparent that some level of correlation exists between the sentiment conveyed by the term *credit* and certain events that somehow determined its connotations. Of course, this simple study does not validate the corpus, but it certainly points to its usefulness as a research resource.

7. Conclusions, limitations and future perspectives

We have presented an ongoing project to design and build a diachronic, balanced, representative, and free-to-use corpus of economic-financial news from daily journals. Apart from our initial sentiment analysis application, the GRNC may be useful as a multipurpose resource, such as ESP, socio-economic studies, and diachronic linguistics.

Our corpus is still under development. Further research will shape the future of the GRNC, as our work is focused on the development of finely grained specific-domain sentiment analysis tools. One of the future goals is to expand its coverage to include (a) field-related texts from different journalistic sources and (b) non-journalistic sources, mainly social media and corporate reports.

The reason for expanding this corpus in these specific ways lies in the fact that the two newspapers which were used as sources share a similar liberal political angle. Future efforts will involve compiling other specialized publications of different ideologies, so that comparative language use can be performed. Another key factor in our future development is the question of the study of the expression of economic language from different levels of specialization.

On the other hand, integrating social media sources would allow us to compare the use that the public makes of economic language. Sources such as blogs, online comments in newspapers and social media would undoubtedly enhance the possibilities of the

current corpus. Other potential research possibilities would involve comparative studies on terminological trends in order to determine the level of influence of institutions and mainstream media into the general public.

Finally, the observation of highly specialized language from documents issued for specialists is a field of special interest, as is the case of internal corporate disclosures. In this way, the lexical, cognitive and affective divergences between different levels of specialization could be observed: specialist discourse, journalistic/informative language and public use of specialized terms.

REFERENCES

Al-Rawi, Ahmed. 2019. Viral news on social media. *Digital Journalism* 7/1: 63–79.

Alba-Juez, Laura and Geoff Thompson. 2014. The many faces and phases of evaluation. In Laura Alba-Juez and Geoff Thompson eds. *Evaluation in Context.* Amsterdam: John Benjamins, 3–24.

Baroni, Marco and Bernardini, Silvia. 2004. *BootCaT*: Bootstrapping corpora and terms from the web. In María Tersa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva eds. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Paris: European Language Resources Association, 1313–1316.

Bednarek, Monika and Helen Caple. 2012. *News Discourse*. London: A&C Black.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Blevins, Dane P. and Roberto Ragozzino. 2019. On social media and the formation of organizational reputation: How social media are increasing cohesion between organizational reputation and traditional media for stakeholders. *Academy of Management Review* 44/1: 219–222.

Botella, Ana, Keith Stuart and Lucía Gadea. 2015. A journalistic corpus: A methodology for the analysis of the financial crisis in Spain. *Procedia – Social and Behavioral Sciences* 198: 42–51.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Chung, Jae Eun. 2018. Peer influence of online comments in newspapers: Applying social norms and the social identification model of deindividuation effects (SIDE): *Social Science Computer Review* 36/5: 551–567.

ComScore. 2012. *Most Read Online Newspapers in the World: Mail Online, New York Times and The Guardian*. https://www.comscore.com/Insights/Infographics/Most-Read-Online-Newspapers-in-the-World-Mail-Online-New-York-Times-and-The-Guardian (4 May, 2020.)

Conboy, Martin. 2006. *Tabloid Britain: Constructing a Community through Language*. London: Routledge.

Diesner, Jana, Terril L. Frantz and Kathleen M. Carley. 2005. Communication networks from the *Enron Email Corpus* "It's always about the people. Enron is no Different." *Computational and Mathematical Organization Theory* 11/3: 201–228.

Douglas, Fiona M. 2003. *The Scottish Corpus of Texts and Speech*: Problems of corpus design. *Literary and Linguistic Computing* 18/1: 23–37.

Duffy, Andrew and Megan Knight. 2019. Don't be stupid. *Journalism Studies* 20/7: 932–951.

Etter, Michael, Davide Ravasi and Elanor Colleoni. 2017. Social media and the formation of organizational reputation. *Academy of Management Review* 44/1: 28–52.

European Central Bank. 2012. Verbatim of the remarks made by Mario Draghi. Speech given at the Global Investment Conference. London, 26 July 2012. https://www.ecb.europa.eu/press/key/date/2012/html/sp120726.en.html (25 May, 2020.)

European Central Bank. 2015. Introductory statement to the press conference (with Q&A) by Mario Draghi. Frankfurt am Main, 22 January 2015. https://www.ecb.europa.eu/press/pressconf/2015/html/is150122.en.html (25 May, 2020.)

Franklin, Bob. 2014. The future of journalism. *Journalism Studies* 15/5: 481–499.

Gablasova, Dana, Vaclav Brezina and Tony McEnery. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67/1: 155–179.

Golan, Guy. 2006. Inter-media agenda setting and global news coverage. *Journalism Studies* 7/2: 323–333.

Handford, Michael. 2010. *The Language of Business Meetings*. Cambridge: Cambridge University Press.

Huan, Changpeng. 2018. *Journalistic Stance in Chinese and Australian Hard News*. Shanghai: Springer.

Huxford, John. 2012. Reporting on recession: Journalism, prediction, and the economy. *International Business & Economics Research Journal (IBER)* 11/3: 343–356.

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. *The Sketch Engine*: Ten years on. *Lexicography* 1/1: 7–36.

Li, Yongyan and David D. Qian. 2010. Profiling the academic word list (AWL) in a financial corpus. *System* 38/3: 402–411.

Link Klipper 1.0.0. 2017. http://www.codebox.in/products/linkklipper/ (7 May, 2020.)

Lischinsky, Alon. 2011. In times of crisis: A corpus approach to the construction of the global financial crisis in annual reports. *Critical Discourse Studies* 8/3: 153–168.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19/2: 313–330.

McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Moirand, Sophie. 2007. *Les Discours de la Presse Quotidienne. Observer, Analyser, Comprendre*. Paris: Presses Universitaires de France.

Moreno-Ortiz, Antonio. 2017a. Lingmotif: A user-focused sentiment analysis tool. *Procesamiento del Lenguaje Natural* 58: 133–140.

Moreno-Ortiz, Antonio. 2017b. Lingmotif: Sentiment analysis for the digital humanities. In Mirella Lapata, Phil Blunsom and Alexander Koller eds. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia: Association for Computational Linguistics, 73–76.

Moreno-Ortiz, Antonio, Javier Fernández-Cruz and Chantal Pérez-Hernández. 2020. Design and evaluation of SentiEcon: A fine-grained

economic/financial sentiment lexicon from a corpus of business news. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asución Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille: European Language Resources Association, 5067–5074.

Nafría, Ismael. 2017. *La Reinvención del New York Times: Cómo la Dama Gris del Periodismo se está Adaptando.* Austin: Knight Center.

Newman, Nic. 2009. *The Rise of Social Media and its Impact on Mainstream Journalism*. Oxford: Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford.

Renouf, Antoinette. 2002. The time dimension in modern English corpus linguistics. In Bernhard Kettemann and Georg Marko eds. *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000*. Amsterdam: Brill/Rodopi, 27–41.

Rojo López, Ana María and María Ángeles Orts Llopis. 2010. Metaphorical pattern analysis in financial texts: Framing the crisis in positive or negative metaphorical terms. *Journal of Pragmatics* 42/12: 3300–3313.

Rose, Tony, Mark Stevenson and Miles Whitehead. 2002. *The Reuters Corpus Volume 1*– From yesterday's news to tomorrow's language resources. In Manuel González Rodríguez and Carmen Paz Suárez Araujo eds. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas de Gran Canaria: European Language Resources Association, 827–833.

Sadjirin, Roslan, Roslina Aziz, Nordin Abdul, Ismail Mohd Rozaidi and Norzie Diana Baharum. 2018. The development of *Malaysian Corpus of Financial English* (MaCFE). *Journal of Language Studies* 18/3: 73–100.

Schröter, Melani and Petra Storjohann. 2015. Patterns of discourse semantics: A corpus-assisted study of financial crisis in British newspaper discourse in 2009. *Pragmatics and Society* 6/1: 43–66.

Schudson, Michael. 1989. The sociology of news production. *Media, Culture & Society* 11/3: 263–282.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 2005. Corpus and text – Basic principles. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. http://users.ox.ac.uk/~martinw/dlc/index.htm (7 May, 2020.)

Thompson, Geoff and Susan Hunston. 2000. Evaluation: An introduction. In Susan Hunston and Geoff Thompson eds. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 1–26.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Van Belle, Douglas A. 2003. Bureaucratic responsiveness to the news media: Comparing the influence of *The New York Times* and network television news coverage on US foreign aid allocations. *Political Communication* 20/3: 263–285.

Xiao, Richard. 2010. Corpus creation. In Nitin Indurkhya and Frederick J. Damerau eds. *Handbook of Natural Language Processing.* Boca Raton: Chapman & Hall/CRC, 147–165.

Corresponding author:
Javier Fernández-Cruz
Departamento de Filología Inglesa, Francesa y Alemana
Escuela de Ingenierías Industriales
C/ Doctor Ortiz Ramos, s/n
29071 Málaga, Spain
fernandezcruz@uma.es

received: February 2020
accepted: June 2020