UNIVERSITÀ DEGLI STUDI DI NAPOLI

FEDERICO II

DIPARTIMENTO DI SCIENZE POLITICHE

SCUOLA DI DOTTORATO IN SCIENZE PSICOLOGICHE, PEDAGOGICHE E LINGUISTICHE

DOTTORATO DI RICERCA IN

LINGUA INGLESE PER SCOPI SPECIALI

XXVIII CICLO

TESI DI DOTTORATO

Crawling in the deep.
A corpus-based genre analysis of news tickers

CANDIDATO

dott. Antonio Fruttaldo

Relatore

Prof. Marco Venuti

Coordinatore del Dottorato

Prof.ssa Gabriella Di Martino

NAPOLI 2016

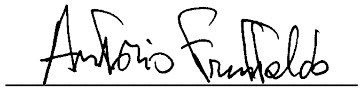
Declaration of Authorship

The present thesis was written by myself and the work contained therein is my own work, unless explicitly stated otherwise in the text. Wherever contributions of others are involved, every effort has been made in order to indicate this clearly, with due reference to the literature.

However, I would like to underline that some of the observations, definitions, and examples offered in Section 3.1 of this dissertation are drawn on my personal notes and materials taken and distributed during the English Corpus Linguistics 2014 Summer School held at the University College of London, the UCREL Summer School in Corpus Linguistics 2014 held at Lancaster University, and the online MOOC in *Corpus linguistics: method, analysis, interpretation* offered by Lancaster University via the FutureLearn online e-learning platform. Thus, even though they are not directly acknowledged, I will draw on them for the brief introduction to Corpus Linguistics introduced in this contribution.

I am very grateful to the news organizations around the world that lent support for this investigation by giving me the permission to reproduce news stories and images in this dissertation. I am particularly thankful to the BBC World News which waived copyright fees in the collection of the corpora under investigation. Copyright for all news stories remains with the organization. All sources are fully acknowledged throughout this work by clearly identifying where each example was first published/broadcasted (including outlet and date of publication).

Naples, March 31, 2016



A note to the reader

Before introducing my work, I would like to address the reader of this dissertation by underlining that, throughout the thesis, I have decided to refer to my own person through the use of the pronoun ‘we’. This may be in contrast with the Declaration of Authorship previously introduced, but the choice of this pronoun is linked to both my personal belief that in ‘research’ there is no ‘I’ and, more importantly, that during my work, different ‘voices’ have implemented and perfected this research thanks to their constant support and feedback. In particular, I am referring to my supervisor, Marco Venuti, the members of the PhD board, my fellow colleagues, and all those people that during conferences, where I have presented the preliminary results of my investigation, have offered their feedback and pointers to a better understanding of the genre under investigation. Thus, in line with the observation made by Isaac Newton in his famous statement “If I have seen further it is by standing on the shoulders of Giants”, the use of the pronoun ‘we’ is nothing but my own personal and modest tribute to all those people that, in some way, have shaped and contributed to this investigation.

Additionally, I would also like to point out that, in this dissertation, I have decided to adopt the use of the singular ‘they’ when referring to antecedents that are grammatically singular, in order to neither reinforce nor perpetuate any form of gender binarism. This is in line with my personal belief that, in Academia, we should all work towards a more gender-inclusive environment, and since binary systems tend to exclude some of the ‘voices’ that may enrich with their extraordinary lives our professions, this is my own personal tribute to all those people that, every day, in the academic world, struggle in order to be heard and taken into consideration for their knowledge and their commitment to research, and not solely judged on the basis of the way they identify themselves. I do apologise to the reader for this personal (and political) stance towards these issues, and I confide in their understanding and open-mindedness in reading this contribution.

Abstract

This thesis investigates a relatively new textual genre found in TV news broadcasts, which made its first appearance on the morning of 9/11. The genre under investigation is that of news tickers (or crawlers), the string of texts that scroll in the lower third space of a TV screen on certain TV news channels and programmes, displaying a summary of the major news stories or announcing a given breaking news story. Thus, through a corpus-based analysis, we will try to better define the genre under investigation. In particular, our corpus-based analysis has allowed us to underline specific phenomena of hybridisation in the genre of news tickers as displayed on the BBC World News. These phenomena can be ascribed to what Meech (1999) defines as brandcasting, which refers to the vast array of corporate branding techniques that broadcasters use in order to project their brand identity. These branding techniques are highly frequent in the corpus of news tickers taken from the BBC World News and, while some of them may be classified as overt promotional strategies (Clearly and Coffey 2008, 2011), others may be seen as subtly achieving the same purpose. Additionally, thanks to the use of corpus linguistic methodologies, our investigation has also highlighted the mixed nature of the genre of news tickers, proven by the merging of two functions traditionally belonging to the journalistic genres of headlines and lead paragraphs. Finally, by focusing on how specific news values are enhanced in the genre under investigation, we have further confirmed our hypotheses on the genre of news tickers.

In Chapter 1, we will introduce the topic of our research project and we will further set out the aims of our investigation.

In Chapter 2, we will briefly outline an overview of the social and professional context where news tickers have slowly shaped themselves in the format that nowadays we see on major TV news broadcasts. In particular, we will offer an overview of the various works that have investigated this genre, underlining the absence of a specific literature in discourse studies on graphic elements displayed on TV news channels and TV news programmes.

In Chapter 3, we will offer an introduction to the methodological framework used for our analysis, which adopts corpus linguistic methodologies as its main

analytical tool in order to highlight given linguistic patterns in the genre of news tickers. Additionally, this chapter also offers an overview of genre analysis (Biber 1988; Tribble 1999; Swales 1990; Bhatia 1993, 2004; Xiao and McEnery 2005), lexical priming (Hoey 2005; O'Donnell *et al.* 2012), and news values (Bell 1991; Bednarek 2006; Bednarek and Caple 2012a, 2012b, 2014; Caple and Bednarek 2015; Potts, Bednarek and Caple 2015).

Chapter 4, on the other hand, offers a step-by-step description of the way the corpora that we have used in order to carry out our investigation have been collected and prepared for our analysis. Additionally, the chapter also describes the software that we have used in order to carry out our investigation and, by doing so, a preliminary analysis of the genre will also be introduced in the form of an initial investigation into the lexical richness of the genre of news tickers when compared to other genres found in the same professional environment.

In Chapter 5, we will introduce the analysis carried out thanks to corpus linguistic methodologies of the genre of news tickers. We will start with the examination of the textual realisation of this genre and, then, thanks to a keyword analysis of the corpus of news tickers, we will both underline the hybrid and mixed nature of this genre, demonstrated by specific linguistic patternings. Additionally, we will further test our claims by analysing the way specific news values are enhanced in news tickers, thus, highlighting how this type of analysis, which combines both a quantitative and a qualitative approach to the data, can reveal which values are particularly enhanced by editors in presenting news stories in this format.

de La Hyre, L. 1650. *Allégorie de la Grammaire*. London: National Gallery.

Dadme albricias, buenos señores, de que ya yo no soy don Quijote de la Mancha, sino Alonso Quijano, a quien mis costumbres me dieron renombre de «bueno». Ya soy enemigo de Amadís de Gaula y de toda la infinita caterva de su linaje; ya me son odiosas todas las historias profanas de la andante caballería; ya conozco mi necedad y el peligro en que me pusieron haberlas leído; ya, por misericordia de Dios escarmentando en cabeza propia, las abomino.

de Cervantes, M.

Don Quijote de la Mancha (II, 74)

Contents

Acknowledgements

According to some of the major style manuals, the acknowledgement section in a contribution should clearly thank or mention those people or institutions that have contributed in some way to a specific manuscript, underlining that, while being a personal acknowledgement of the way people have influenced or helped with your work, its register should not be informal, because one day you may regret some of the things you have written. So, to my future self: I am so very disappointed in you, if you have lost the ability to look at the world the way you always do. As you may have forgotten, I am truly bad at keeping my feelings aside and I always display a tendency towards personalisation. Thus, I am truly sorry if you think that these acknowledgments should have been written in a different way, by using a certain degree of formality. And I do pity you if you have forgotten where all this is coming from.

So, without further ado, I would like to start by saying that I have never believed in an afterlife or in any kind of supreme force looking down at me from above and forcing me to judge others in a given way or in another. But I do believe that we are the sum of all the things we have done in our life. And in sum, I do cherish above all the memories that I have of the people that inhabit or have inhabited my existence because, sometimes, the smallest things, the things caught in a specific moment, are worth the sum of all things. Paraphrasing a monologue taken from a not-so-famous TV series (*Being Human*, BBC 3, 2008–2013), in our life, we meet people and, when we part, they leave marks for us to remember them by. They sculpt us. They define us, for better or worse. They linger inside us, invisible, but always there.

And, amongst the various people that during these three years of my PhD have in some way shaped my very existence, I would like to express my deepest and sincerest gratitude to my supervisor, Marco Venuti. He has constantly encouraged me, guided me through the dark allies of these years, comforted and motivated me when things seemed unbearable. Looking at him working his ‘magic’ with corpus linguistic methodologies has been an invaluable experience, and I shall always cherish all the moments spent together in his office trying to figure out some way to solve given problems in the data collection and analysis or, more simply, discussing the last

episode of *Game of Thrones*. I thank him so much for his incredible patience and for these wonderful moments, and if this 'dragonian' dissertation, hatched from the fire of his inspiration, may show any weaknesses and, therefore, should 'bite my hand off', the sole responsibility is on me.

I would also like to express my very great appreciation to Professor Giuditta Caliendo. During these years, the echo of our conversations, her enthusiastic encouragements and useful critiques, her valuable and constructive suggestions, and her willingness to give her time so generously during the initial planning and development of this research work have been very much appreciated. I always keep in my pocket during conferences a token of her appreciation, a good-luck charm that she gave me when I presented my MA research. It did not fail me so far during these years as a PhD student and, every time I take it from my pocket and carefully put it back in my wallet, I am always reminded of this incredible person that has in so many ways changed my life.

My grateful thanks are also extended to Professor Gabriella Di Martino for her constant encouragements and, more importantly, for the kindness and generosity she has shown during the years towards 'her' PhD students. Every time I smile, I am always reminded of the nickname she has given me (her Cheshire cat), and I think there is nothing more rewarding than seeing your PhD coordinator treating you with the affection and love she truly feels for all her students. She has taught me so many things during these years and I will forever be indebted to her for the valuable advices and suggestions linked to my PhD experience.

I would like to offer my special thanks also to Professor Cristina Pennarola and Professor Vanda Polese. They have given me the opportunity to share some of my knowledge with their students, and these invaluable experiences have helped me work on my confidence. Their feedback during official PhD meetings and informal discussions, and the way they have 'cuddled' and spoiled me during these years are something that I shall never be able to pay back. In particular, I would like to express my deepest gratitude to Professor Vanda Polese, who was initially appointed as my supervisor. I thank her because she helped me increasingly 'fall in love' with corpus linguistic methodologies. But, more importantly, I thank her for our animated discussions and arguments, after which we are always able to work out a common

strategy. I thank her for all she has done during these years, and I hope we will never stop ‘bickering’ about things, because it was thanks to these precious moments that I have truly learned something from her incredible experience.

I would also like to express my greatest gratitude to Paolo Donadio, who is always trying to run me over with his Vespa when crossing the street that takes me to the Department of Political Science. Leaving aside our funny ways of showing our appreciation for each other, I thank him for the incredible support shown during these years and for his knowledge of media discourse, which has allowed me to further explore given peculiarities in the genre I have investigated.

I am also grateful to the administrative staff of the Department of Political Science at the University of Naples Federico II and, in particular, to Mena Vilardi and Cinzia della Monica for their constant and untiring help and support. They have always been so very kind to me and I will always treasure the moments spent together in their office talking about our lives and discussing the last conferences and/or events in our PhD course.

I would like to express my very great appreciation to my fellow PhD colleagues. In particular, I would like to thank Antonio Compagnone whom I always look up to with the highest and most profound admiration, and who has constantly provided his help when I was a little bit ‘lost’ in the jungle of the academic world. I would like to also extend my gratitude to Eleonora Esposito, Cristina Aiezza, and Adriano Laudisio for making me feel at home, including me in their lives and allowing me to ‘eavesdrop’ on their researches. My special thanks are extended also to Chiara Nasti, Cristina Nisco, and Alba Sole Zollo, who have been so very kind to me. I was a stranger to them and, nonetheless, they have encouraged me, helped me, and supported me with their words and their constant presence. In particular, I would like to thank Cristina Nisco for her silent presence by my side when I was writing this dissertation. Her work has been my ‘blanket’ that has constantly comforted me when I felt hopeless, and I shall always thank her for this. My deepest gratitude goes also to Fabrizio Esposito, Annarita Magliacane, and Angela Zottola. They have been the most important presence in my life in these three years and I think there are no words to express how deeply and irrevocably I cherish each moment spent together. Last but not least, I would like to thank Francesca Raffi. Having her by my side during these years has

been a real privilege. The last time we saw each other was during *Languaging Diversity 2016* in Macerata and, when parting, I could not help but cry because it was the last conference that we both attended as PhD students. And if these years have been truly and incredibly happy is because of those moments we shared together talking on the phone or walking by the streets of Catania, laughing our heads off about our banters. I thank her for the happiness she has brought in the darkest moments of my PhD and I can only say to her: see you in prison, darling.

As a way of concluding this Acknowledgement section, I would like to dedicate this dissertation to my mother, my father, and my brother, for their support during these years and for their patience. And I would also like to dedicate this dissertation to Davide Bizjak. We met each other when we were in high school, and he is the kindest, most decent, unfailingly generous person I could have ever met in my life. His advices and the way he knows me sometimes more than myself makes him the most important presence in my life.

List of tables

List of figures

1. Introduction

Journalistic practices are undergoing, in the last few years, a radical change due to the increasing pressure of new digital media on the professional practice (Bivens 2014). The ever-growing development of new technologies challenges traditional genres found in this context. Indeed, as our lives and social institutions are constantly in flux, creating "a society in which the conditions under which its members act change faster than it takes the ways of acting to consolidate into habits and routines" (Bauman 2005: 1), the ceaseless fluctuation of social practices has inevitable consequences on the genres and discourses created by social institutions, since genres and discourses are socially and linguistically significant entities (Fairclough 2011).

Thus, as our society becomes characterised more and more by forms of "liquidity" (Bauman 2000, 2005), social practices are also changing in order to stay up-to-date within this constant state of flux. Given this picture, journalistic practices and genres "should be understood within the wider context of liquidity" (Bivens 2014: 77), as practices which try to incorporate in their routines and in their genres the liquidity of contemporary society. However, since liquid modernity is unrestrainable, journalistic practices try to convey this flow of ever-changing information by relying on their traditional boundaries and formats.

Indeed, "journalism still depends on its established mode of production, through which it largely (and unreflexively) reproduces the institutional contours of high (or "solid") modernity" (Deuze 2008: 856). Therefore, contemporary journalism is at the mercy of two opposite forces. The first one constraints journalism within its traditional norms of production and reproduction, while the second one leads it to new forms of fluid contents and the implementation of digital media.

Furthermore, the challenge represented by liquid modernity to social and, consequently, professional practices has inevitable consequences also on the traditional frameworks developed in order to analyse genres in the professional environment. Indeed, a dynamic environment such as that of contemporary journalism calls into question the very nature of genre analysis.

Genres have been traditionally analysed on the basis of "the use of language in conventionalized communicative settings, which give rise to specific set of communicative goals to specialized disciplinary and social groups, which in turn establish relatively stable structural forms" (Bhatia 1996: 47). On the contrary, in a fluid social context (Deuze 2008), genres are increasingly becoming dynamic rhetorical configurations, whose conventions can be exploited to achieve new goals. In the words of Berkenkotter and Huckin (1995: 6):

> Genres […] are always sites of contention between stability and change. They are inherently dynamic, constantly (if gradually) changing over time in response to sociocognitive needs of individual users.

Thus, the ultimate aim of genre analysis is becoming that of dynamically explaining the way language users manipulate generic conventions to achieve a variety of complex goals (Bhatia 2004). Mixed or hybrid forms are most frequently the results of these manipulations, particularly due to the competitive professional environment, where users exploit genre-mixing or hybrid genres "to achieve private intentions within the context of socially recognized communicative purposes" (Bhatia 1996: 51).

These private intentions, however, are not detectable at first hand, since they are blended in the social context where the mixed or hybrid form was created. Kress (1987) explains this by referring to the so-called appropriate authority to innovate, which depends on the likelihood of developing new generic forms on the basis of social change. In other words, "unless there is change in the social structures – and in the kinds of social occasions in which texts are produced – the new generic forms are unlikely to succeed" (Kress 1987: 41-42). Thus, if genre-mixing, defined as the "mixture of two or more communicative purposes through the same generic form" (Bhatia 2002: 11), does not meet the appropriate social environment, such forms are less likely to flourish and they will soon perish.

Given the ever-changing social context where journalistic practices operate, they are constantly exploiting new forms of hybridity and genre-mixing in order to compete with new ways of delivering the news. For instance, as new media technologies are introduced, we can notice that the "boundaries between news,

entertainment, public relations and advertising, always fluid historically, are now becoming almost invisible" (Schiller 1986: 21). This intensifying pressure on traditional media has given rise to a variety of hybrid and mixed-generic forms, among which we are going to focus on a relatively new genre of TV news broadcast, generally referred to as news tickers (or crawlers).

1.1 Aims of the research

The genre of news tickers, which made its first appearance on 9/11 in order to deal with the enormous amount of information coming from the American news agencies, has been adopted by various TV news channels and programmes in order to constantly deliver to viewers a summary of the major news stories of the day or to alert viewers of particular breaking news stories. However, during the years and given the increasing pressure on TV journalism to allure viewers, the genre of news tickers has been slowly appropriating certain generic conventions from other genres to serve this purpose. Indeed, given "the growing ability of viewers to avoid or ignore traditional commercials" (Elliott 2009), TV news networks have found in news tickers a subtle way to market their products, "due to the ticker's location at the bottom of the screen, and its format, which does not interrupt programming" (Coffey and Clearly 2008: 896). Genre analysis (Swales 1990; Bhatia 1993, 2004) can, thus, highlight these textual cues that can reveal how news tickers are purposefully being exploited in order to brand the TV news network or programme. However, since genre analysis is increasingly changing in order to stay up-to-date with the dynamically changing context of contemporary society, this social context has demanded a reshaping of its conventional approach to textual analysis, since genres are progressively becoming fluid entities, open to unexpected innovations by borrowing structural conventions and rhetorical configurations from other generic forms. These challenges to genre analysis, however, can be easily overcome by the increasing availability of corpora to researchers. Thus, changes in professional practices can be successfully highlighted by the use of corpus linguistic methodologies.

As we will see, the availability of ready-made corpora may, nonetheless, cause some disadvantages on the behalf of researchers interested in particular areas of

human communications, since "a corpus is always designed for a particular purpose" and the majority of them "are created for specific research projects" (Xiao 2008: 383), thus, focusing only on specific genres, while others remain unexplored.

In order to study very specific instances of language in use of a particular discourse community, most of the time, researchers have to create their own specialised corpora, and this is particularly the case of news tickers, given the unavailability of an already-built corpus but, more importantly, no database with instances of this genre.

Thus, in the following paragraphs, thanks to a corpus-based linguistic analysis, we are going to focus on if and how the BBC World News uses its news tickers in order to promote itself and its products. In this, corpus-based methodologies have been of great help, since "[t]he computational analysis of language is often able to reveal patterns of form and use in particular genres […] that are unsuspected by the researcher and difficult to perceive in other ways" (Bhatia 2002: 13). This is the reason why a bottom-up approach to the analysis of these strategies has been adopted, since "one cannot detect these functions without first noticing a pattern of forms" (Berkenkotter and Huckin 1995: 43), which Corpus Linguistics allows us to do.

Indeed, through a corpus-based genre analysis, we will try to better define the generic status of the news tickers displayed on the BBC World News by, firstly, introducing a keyword analysis of the genre under investigation, which has allowed us to highlight the presence of given strategies of marketisation (Fairclough 1989, 1992) in the comparison with a reference corpus of headlines and lead paragraphs taken from the BBC website. In this, a textual colligation analysis (Hoey 2005; O'Donnell *et al.* 2012) has further helped us highlight the tendencies for these strategies to occupy non-initial sentence positions, therefore, underlining the peculiar textual realisation of the phenomenon highlighted in the corpus under investigation when compared to other genres found in the same professional environment. This has demonstrated the fact that, in the migration of contents from one media platform to the other, the BBC uses given self-promotion marketing strategies that are textually specific to the genre they are realised in. Thus, as Coffey and Clearly (2008, 2011)

maintain, also in the context of British news broadcasts, news tickers are particularly used in order to promote the news network and its contents.

Additionally, as we will see, the analysis of the genre under investigation has also proven that, in the comparison with a reference corpus of headlines and lead paragraphs, news tickers can be seen as a mixed (sub-)genre in the context of TV news broadcast. Indeed, the combination of linguistic elements of headlines and lead paragraphs, from a (Critical) Genre Analysis point of view (Bhatia 2004, 2007, 2008, 2012), underlines a specific private intention in the context of the BBC, that is to say, using the 'offline' contents in order to lead viewers to their online platform, thus, increasing their visibility on the Internet as a leading news company.

Finally, the analysis of the news values (Bell 1991; Bednarek and Caple 2012a, 2012b, 2014; Potts, Bednarek and Caple 2015) discursively realised in news tickers has additionally confirmed the previous hypotheses on the hybrid and mixed nature of news tickers but, more importantly, on their subtle purpose to lead viewers towards the BBC online platform.

1.2 Outline of the research

As we have previously underlined, our work mainly focuses on a corpus-based genre analysis of news tickers. Thus, in order to better define the generic status of the genre under investigation, in Chapter 2 we will briefly outline an overview of the social and professional context where this textual realisation has slowly shaped itself in the format that nowadays we see on major TV news broadcasts. In particular, we will offer an overview of the various works that have investigated this genre, underlining the absence of a specific literature in discourse studies on graphic elements displayed on TV news channels and TV news programmes. Indeed, as we will see, while the increased use of TV news graphics has attracted the attention of some scholars belonging to the field of reception studies, very few attempts, from a discourse analysis point of view, have been made in trying to investigate these genres found in media discourse. Thus, we will firstly introduce a summary of the major works in reception studies on TV news graphics, which will underline how, from a visual behaviour point of view, news tickers play a major role in catching viewers' attention during TV news broadcasts. We will then proceed to the illustration of a brief history

of news tickers, in order to better understand how this genre, during the years, has slowly changed the purposes that it served and, thus, how it slowly has been appropriated by TV news broadcasts.

In Chapter 3, we will offer an introduction to the methodological framework used for our analysis, which adopts corpus linguistic methodologies as its main analytical tool in order to highlight given linguistic patterns in the genre of news tickers. Additionally, this chapter also offers an overview of genre analysis (Biber 1988; Tribble 1999; Swales 1990; Bhatia 1993, 2004; Xiao and McEnery 2005), lexical priming (more specifically, as we will see, we will focus on the notion of textual colligation as introduced in the context of Hoey's (2005) approach to textual analysis), and news values (Bell 1991; Bednarek 2006; Bednarek and Caple 2012a, 2012b, 2014; Caple and Bednarek 2015; Potts, Bednarek and Caple 2015).

Chapter 4, on the other hand, offers a step-by-step description of the way the corpora that we have used in order to carry out our investigation (i.e., the NTC and the bw_14) have been collected and prepared for our analysis. In particular, we will introduce the annotation scheme used in order to tag our corpora and, by doing so, we will offer a preliminary introduction to the genre of news tickers, since the codification of structural elements in the corpus of news tickers provides an initial overview of the genre. The chapter also describes the software that we have used in order to carry out our investigation and, by doing so, a preliminary analysis of the genre will also be introduced in the form of an initial investigation into the lexical richness of the genre of news tickers when compared to other genres found in the same professional environment.

In Chapter 5, we will introduce the analysis carried out thanks to corpus linguistic methodologies of the genre of news tickers. We will start with the examination of the textual realisation of this genre, since the analysis of the formal structure of the crawlers collected in our corpus will provide some interesting insights that will be further investigated in their linguistic realisation. Indeed, thanks to a keyword analysis of the corpus of news tickers, we will both underline the hybrid and mixed nature of this genre, demonstrated by specific linguistic patternings. Additionally, we will further test our claims by analysing the way specific news values are enhanced in news tickers, thus, highlighting how this type of

analysis, which combines both a quantitative and a qualitative approach to the data, can reveal which values are particularly enhanced by editors in presenting news stories in this format.

This contribution concludes with an Appendix section. In Appendix 1, the data computed thanks to a keyword extraction from the comparison between the corpus of news tickers collected from the BBC World News and the corpus of headlines and lead paragraphs taken from the BBC website will be introduced. In Appendix 2, on the other hand, a tag keyword extraction in the comparison between the corpus of news tickers and the corpus of headlines and lead paragraphs is offered. Appendix 3 lists the tag keyword extraction in the comparison between the corpus of news tickers and the headline component of the corpus of headlines and lead paragraphs taken from the BBC website. Conversely, Appendix 4 shows the tag keyword extraction in the comparison between the corpus of news tickers and the lead paragraph component of the corpus of headlines and lead paragraphs taken from the BBC website. In line with one of the hypothesis tested in Chapter 5, Appendix 5 provides empirical evidence to this claim by listing all the adverbs occurring in the corpus of news tickers displayed according to their raw frequency. Finally, Appendix 6 lists all the pointers to newsworthiness in the corpus of news tickers.

2. Graphic elements of TV news broadcasts

As previously underlined, contemporary journalism is at the mercy of two divergent forces. The first one constraints journalism within its traditional norms of production, while the second one leads it to new forms of fluid contents and the implementation of digital media.

In TV news journalism, a compromise between these two forces has been found in the increasing implementation of the graphic layout of TV newscasts. This information reception context has increased in its complexity in the last 30 years in TV news programmes, "such that visual stimuli are presented diversely in the visual space" in order to "provide extra information or additional messages to complement the anchor and news video" (Rodrigues, Veloso and Mealha 2012: 357). Indeed, just as an online webpage, TV news broadcasts tend to assign to designated areas of the TV screen given functions, since viewers need to recognise immediately what they are looking at from the place where the textual element is displayed and/or they can easily find certain information they are looking for thanks to a routinely established placement of this information in certain 'places' of the TV screen (see Figure 1).

Figure 1 A screenshot taken from the programme *GMT* aired on February 3, 2014, showing the different graphic elements typical of the BBC World News channel.

Given the strict placement of textual elements in the TV news graphic layout for the reasons previously highlighted, in the following pages, we are going to focus on the most frequently displayed graphic elements, in order to place them in a specific area of the TV screen but, more importantly, in order to define their function(s).

Additionally, in this Chapter, we will also offer an overview of the various works that have investigated TV news graphics in the context of reception studies, underlining the absence of a specific literature on the topic in discourse studies. Our review of the major works on TV news graphics will also underline the major role that news tickers play from a visual behaviour point of view in catching viewers' attention. Finally, in order to better define them, this Chapter also offers a brief history of news tickers.

2.1 Media discourse and digital currents: reshaping media conventions

As we have previously highlighted, TV news journalism has seen in the last 30 years an increasing implementation of the graphic layout of TV newscasts. And as a way of justifying this statement, in this Section, we would like to briefly introduce the various graphic elements that are usually and routinely implemented during TV news broadcasts, in order to place them in a specific area of the TV screen but, more importantly, in order to define their function(s).

Firstly, the most common and known graphic element of TV news broadcasts is represented by lower thirds, also referred in the US as 'superbars' (or simply 'supers') or 'chyrons', "due to the popularity of Chyron Corporation's *Chiron I* character generator, an early digital solution developed in the 1970s for rendering lower thirds" (Lower Third 2005). Even though the name of this graphic element refers to the space at the bottom of the TV screen, lower thirds rarely occupy this entire space. Typically, we can distinguish between three types of lower thirds (Rodrigues, Veloso and Mealha 2012):

a. one-tier lower thirds: they usually present a headline of the news story currently being presented by the anchor. One-tier lower thirds are also used in order to display the name of the anchor or correspondent;
b. two-tier lower thirds: they are comprised of two lines. In the first one, the headline of the news story being presented by the anchor is displayed, while

in the second one a subheadline is placed in order to add additional information about the news story. Two-tier lower thirds may also be used in order to display the name of the anchor or correspondent (in the first line) and their affiliation (in the second line);

c. three-tier lower thirds: even though nowadays they are rarely used during TV news broadcasts, since they seem to clutter the screen with too much information, three-tier lower thirds typically show, in the first line, the headline of the news story being presented by the anchor; in the second line, a subhead is added to elaborate on the main headline; and, finally, in the third line, the place and/or time the news report was originally broadcasted is displayed.

One-tier, two-tier, and three-tier lower thirds are also known as local ticker texts (Jindal, Tiwari and Ghosh 2011). This name does not refer to the relevance of the news stories from a geopolitical point of view. Indeed, the adjective 'local' is used in order to underline that the graphic content being displayed is in sync with the news story being presented by the anchor. In other words, in the case of local ticker texts, the aural and visual channels work together in delivering the news.

Global ticker texts, on the other hand, are defined by Jindal, Tiwari and Ghosh (2011: 460) as displaying "the highlights of all important stories in the news program", while scrolling texts provide "the gist of relatively unimportant news". However, the definitions provided for global ticker texts and scrolling texts are too specific to the TV news channel (i.e., the Indian TV channel *Times Now*) the authors have chosen to analyse. More importantly, in the case of scrolling texts, we can also notice a bias towards this graphic element, since the newsworthiness of given news stories should not be used as a parameter in defining a graphic element. Indeed, the degree of newsworthiness of the news stories conveyed by scrolling ticker texts is bound to the journalistic practices specific to each network station. Thus, from the observation of a corpus of newscasts collected from various TV news channels and programmes (i.e., BBC World News, Fox News, CNN, ABC's programme *Good*

Morning America and CBS' programme *This Morning*)[1], we suggest that a more general definition of global ticker texts is needed.

In this respect, they can be defined as all the graphic elements that display news stories which are not directly related to and/or in sync with the news story being presented by the anchor. Scrolling ticker texts (also known as news tickers, crawlers or ticker tape) are a particular form of global ticker texts, which can be identified by the way they are displayed on the screen, that is, as graphic elements that scroll from the right to the left bottom of the screen (in contrast with other global ticker texts such as flippers, which present one news item after the other by 'flipping' at the bottom of the TV screen).

Other typical graphic elements of TV news layout are the so-called DOGs and the locator. The DOG, which stands for Digital On-screen Graphic (Porter 2007), is generally used in order to display the TV channel's logo and/or the name of a particular TV news programme (Meech 1999). Since they are static graphic elements, DOGs are also referred to as 'bugs', a term which metaphorically refers to their overlaying a given screen-area. DOGs are typically found at the bottom of the screen, near or as part of the lower third area. The locator, on the other hand, is a graphic element displaying the location and/or the local time of a correspondent being interviewed by the anchor (typically placed up to the left or right screen corner).

Given this picture on the graphic elements that are typically used during TV newscasts, our previous comment on how TV news layout has become more complex in its implementation of the information conveyed by the anchor and the news video is quite confirmed. However, we must underline that, while TV news channels and TV news programmes have increasingly been using graphic elements in the past few

[1] The corpus collected was part of a preliminary pilot study on the genre of news tickers (see Section 4.1 for further information). As we will see in the paragraph on the methodology used in order to collect the data, a week-long observation of the way these TV news channels and programmes used this genre has allowed us to focus our attention on the BBC World News channel, which in the comparison with the others consistently used news tickers in its daily journalistic routine, while the other channels or programmes did not seem to use them as consistently and routinely (e.g., the normal flow of information conveyed in the news tickers was abruptly interrupted by commercials and did not pick up from where it stopped; during weather emergencies, news tickers were absent in order to leave the graphic 'stage' to messages linked to school closing; etc.).

years, from a discourse analysis point of view, very few studies have been conducted on the topic.

An exception is represented by the work of Montgomery (2007), who focuses on the discourse of two sets of headlines in TV newscasts: the first one represented by "the opening of a standard evening bulletin programme, ITN's *News at Ten*", and the second one by "a lunchtime BBC news programme" (Montgomery 2007: 78). In analysing these two sets of TV headlines, Montgomery (2007) offers some interesting generalisations on the semantic and lexicogrammatical status of the genre as displayed on TV news broadcasts. Another important analysis of TV news headlines is represented by the work of Bednarek and Caple (2012a), who offer some examples of linguistic structures typical of TV headlines compared to the ones found in print newspapers. In their words (Bednarek and Caple 2012a: 105):

> While they [TV headlines] feature some of the characteristics of print news headlines (such as deletion of verbs or articles, use of nominalization), they can also be full sentences with only the verbal group omitted or reduced to a non-finite form or even consist of several sentences with full verbal groups.

Even though these analyses do offer some important insights on the nature of TV news headlines, they do not regard other graphic elements, such as subheadlines, news tickers, breaking news headlines, etc. This may be due to two reasons. Firstly, since there are no available OCR technologies (at the time of writing) that allow researchers to automatically collect corpora of TV news graphics, their analyses tend to focus only on particular case studies, which cannot offer significant generalisations on TV news journalistic practices in the context of TV news graphic elements. Secondly, because some graphic elements (e.g., headlines and subheadlines) are also found in other genres (i.e., print and online newspapers), researchers tend not to regard them as elements of analysis. This can be ascribed to the fact that they misleadingly hypothesise that the observations made for print newspapers' headlines, for instance, will also be valid for TV headlines. This is particularly clear in what Cotter (2010: 16) defines as "modality bleeds", which cause "seepages of one media form to another, […] that come about through changes in media technology". The concept of modality bleeds, however, disregards the fact

that different media will require different genres, since their communicative purposes may vary. Thus, while the concept of modality bleeds generally tends to be used as a way to highlight the similarities found in the different media, little attention has been paid to the sometimes imperceptible variations due to the different environments where the leaked fluids form a new pool.

And while, in their discourse-based account of news discourse, Bednarek and Caple (2012a: 107) relegate to future research how variation across platforms can be analysed in new media, the increased use of TV news graphics has attracted the attention of some scholars belonging to the field of reception studies, where the analysis of TV news graphic layout has been most flourishing in the last few years.

2.2 When is enough… enough? Reception studies in the context of TV news graphic layout

As previously said, while discourse analysts have quite neglected the analysis of TV news graphic layout or focused their attention only on specific instances, in the field of reception studies these elements of broadcast journalism have been at the centre of numerous research studies.

Josephson and Holmes (2006), for instance, have investigated whether the attention spent by participants in different areas of the TV screen (crawler, headline, title, globe, and main area) varied in three versions of the same news story. In the first version, the news story was presented without any textual contents. In the second one, the video and audio of the news story was accompanied by unrelated textual contents (i.e., crawlers). Finally, in the third version of the news story, the video and audio was presented with both related and unrelated textual contents (i.e., headlines and crawlers). From the analysis of the data collected from an eye tracker, the results of this study suggested that the presence of unrelated textual contents (i.e., crawlers) "produced more fixation time at bottom of the screen", while the presence of related textual contents (i.e., headlines) "drew more visual attention to that area of the screen" (Josephson and Holmes 2006: 161). However, in both cases, the presence of related or unrelated items draw away the attention of the viewers from the main screen area, that is, from the anchor and, consequently, the audio content.

Given these results, Josephson and Holmes (2006) also tested whether the information recall for audio contents of the TV news story was influenced by the presence of on-screen visual enhancements. The results of this part of their research suggested that the presence of related textual elements (i.e., headlines) enhanced the recall of key information in the news story, while unrelated textual elements (i.e., crawlers) did not diminish the recall of key information in the news story. However, as Josephson and Holmes (2006) argue, compared to the enhancement effect of related textual elements, "diminished recall of non-headline content suggests an interference effect as well" (Josephson and Holmes 2006: 161). In other words, the more the screen is cluttered, the more difficult it is for viewers to recall other story points, exhibiting an information interference effect.

Josephson and Holmes' (2006) results are confirmed by Matsukawa, Miyata and Ueda's (2009) study, whose purpose was to better understand the information redundancy effect of graphic elements in TV news programmes on viewers by using eye-tracking technology. The study was comprised of two moments: in the first one, viewers were asked to watch a TV news broadcast with eye tracking technology; in the second one, the degree to which viewers understood the contents presented in the TV news broadcast was tested. The participants to this study were divided into two groups: a first group watched the TV news broadcast with telops[2], while the second group watched the TV news broadcast without any telops. The analysis of the understanding of the news stories showed that the correct answer rates of the telop group were higher than those of the no-telop group. However, Matsukawa, Miyata and Ueda (2009) noticed that telops induced some misunderstandings of some questions when they tended to clutter the TV screen (for instance, when telops were accompanied by other textual elements, such as supers and/or flippers), thus, confirming Josephson and Holmes' (2006) findings.

In the studies reviewed so far, we have concentrated our attention on whether TV news layout affects viewers' understanding of the news stories being presented

[2] In Japan, the term is generally used to indicate texts superimposed on a screen: "[...] derived from the then widely employed American *Television Opaque Projector* equipment – a device to transmit separately prepared text or graphics directly on the TV screen without the use of camera [...], the term *telop* remains in common use to refer to any text separately added to the TV screen image" (O'Hagan 2010: 73, original emphasis).

by the anchor. However, amongst the various TV news graphics at the centre of these studies, no distinction was made in terms of which ones are viewed and, thus, used the most by viewers during TV newscasts. This is the reason why we would like to briefly review the research study conducted by Rodrigues, Veloso and Mealha (2012), who investigated which graphics were viewed the most during TV news broadcasts in terms of number of fixations as well as fixation time by analysing the data gathered from an eye tracker. The result of this study was that, in terms of visual behaviour, viewers spend more visual attention on the graphic elements that move, that is to say, the anchor (fixation points: 30.1%; fixation time: 41%) and the news ticker (fixation points: 28.6%; fixation time: 15.3%). Thus, the analysis offered by Rodrigues, Veloso and Mealha (2012) seems to point out that, from a visual behaviour point of view, one of the most viewed and used graphic during TV news broadcasts is represented by news tickers, the graphic element at the centre of our investigation. This important insight justifies our interest in them. Indeed, since viewers' attention is partly caught by crawlers, these elements of TV newscasts seem to work together with the aural channel in order to convey the news stories presented by TV news channels and programmes. Thus, an analysis of this genre is needed so as to better understand what peculiarities are identifiable in the genre under investigation.

However, since a specific genre can be seen as "a socially ratified way of using language in connection with a particular type of social activity" (Fairclough 1995a: 14), used by "communities of practice" with specific "functions" (Swales 1990), their concrete realisation, distinguished by Fairclough (1995a) as text type, is "situationally and historically quite particular" (Fairclough 1995a: 15). Thus, in order to better understand some of the characteristics of the genre of news tickers and its textual realization, a brief history of crawlers' development will be drawn. This little excursus in their evolution and development during the years will help us uncover in the next sections some of the most interesting aspects of this genre.

2.3 Send in the tickers: The day crawlers made their first appearance

On the morning of 9/11, there was more news than could fit on a TV screen (Poniewozik 2010). Indeed, news reports and rumours abounded of new attacks,

blood donations wanted for the Red Cross, buildings being evacuated, and so on. In other words, on TV newscasts there was too much information to send through the aural channel alone and the typical graphic layout of TV news screens was too small to show all this information at once.

Thus, in order to deal with the problem of information overload, each of the 24-hour American news channels implemented news tickers at the bottom of their screens. CNN was working on creating this format since the very beginning of 2001 (Keefe-Feldman 2007: 14), and a first test-run was displayed in the late summer of 2001. However, when 9/11 came about, CNN was preceded by Fox News in implementing news tickers on their screen, starting a running ticker at 10:49 a.m. (see Figure 2). CNN followed approximately at 11:11 a.m. (see Figure 3) and MSNBC at 2:00 p.m. (Moore 2001).

Figure 2 A screenshot taken from the Fox News live coverage of 9/11.

Figure 3 A screenshot taken from the CNN live coverage of 9/11.

After 9/11, news tickers remained at the bottom of certain TV newscasts, becoming a way to convey a sense of never-ending emergency to the audience, which keeps viewers in front of the TV screen, waiting for more, waiting for a breaking news story, which may or may not appear. In this respect, news tickers seem to compete perfectly with digital media, such as social networking systems. In particular, news tickers can be seen as the predecessors of Twitter (The truth about news tickers 2011), since they both share the basic purpose to give viewers and followers the most information in the least amount of time and space. And in a social context where digital media are increasingly becoming new ways of delivering the news to viewers, news tickers seem to emulate them on a TV screen, providing a "defence against television networks losing relevance with news on the Internet" (The truth about news tickers 2011) and creating an illusion of real-time information typical of news digital media.

Finally, news tickers seem to imply tacitly that there is simply too much information to send through one channel alone. In a society where attention span is at stake since we are more and more accustomed to jump from one subject to another, news tickers can be compared to multiple windows opened on a computer screen

(Poniewozik 2010). This allows viewers to listen to the anchor, read the headlines of each news report and, finally, read the news tickers that are crawling at the bottom of the screen at the same time. This process seems to keep viewers in a constant state of frenzy and hunger for more information that must be satisfied in any which way. Additionally, as "news consumers are confronted with television news in more locations, such as airports or busy intersections, on trains and planes" (Bivens 2014: 4), textual elements and, in particular, news tickers are increasingly being used in order to claim these public spaces and deliver the news to consumers.

As previously said, news tickers' modern layout was first displayed on TV news broadcasts on the morning of 9/11. However, their development has been slow and has taken half a century to become what we now see during TV newscasts. And even though Sella (2001) argues that "the Crawl's origins are a minor mystery in the news world" (Sella 2001), in the following pages, we will try to solve this mystery. And we can start by explaining why they are also generally referred to as 'news tickers'[3].

This term draws its origins from the ticker tape, the first mechanical instrument through which stock prices were conveyed over long distance telegraph wiring (Deese 2011). As Deese (2011: 75) explains, in 1870 Thomas Edison "used his knowledge of telegraphy to construct the Universal Stock Ticker", the first mechanical instrument through which stock prices were conveyed over long distance telegraph wiring. Once stock prices were telegraphed, the Universal Stock Ticker printed them at the speed of one Morse symbol per second on a rolling piece of paper called the ticker tape (Ticker Tape 2014).

Thus, the idea behind news tickers was to reproduce the tick-tacking of the Universal Stock Ticker in order to convey the most important news stories to viewers on a daily basis. The Motograph News Bulletin, better known as the 'zipper',

[3] In this contribution, the terms 'news tickers' and 'crawlers' will be used interchangeably. However, we believe that the latter is preferable since, in its meaning, it entails the way the textual elements are displayed on the TV screen. Indeed, just as in the case of flippers, the term 'crawlers' refers back to the way they scroll from the right to the left bottom corners of the screen. The term 'news tickers', on the other hand, seems to refer to the technologies used in the early 1980s in order to display static textual elements on the TV screen. However, during a private interview with one of the journalists working for the BBC, we have noticed that in the journalistic environment of the British broadcasting company the term 'ticker tape' is preferred.

represented one of the first examples of real-life ticker tape. Built in 1928 on the One Times Square skyscraper, it was used by the New York Times to display major news headlines (McKendry 2012). In this way, the genre of news tickers was slowly taking its form.

Nonetheless, only on January 14, 1952 news tickers were used for the first time on a TV news programme. Indeed, the NBC's *Today Show*, during its debut, used a news ticker bar in order to display a summary of the news stories presented by the anchor. This earliest form of crawlers consisted of typewritten headlines on a piece of semi-transparent paper superimposed on the bottom third of the TV screen during the live show (Deese 2011). Thus, this initial form of news tickers was in sync with the video content and, due to lack of technological development, key features of their modern layout (i.e., their scrolling from the right to the left bottom corner of the TV screen) were absent. However, since this initial form of news tickers lacked in popularity, it was soon dropped by the NBC.

We will have to wait until the early 1980s to see news tickers brought back on TV news channels and programmes, but they once again evolved and fulfilled a new function (The truth about news tickers 2011). Indeed, in northern parts of the US, many local television stations started using a news ticker during their newscasts to pass along information on, for instance, school closing due to severe weather conditions. The start of the ticker's cycle was accompanied by an audio signal, such as warning tones or a small jingle from the station's news theme. Two important changes in news tickers due to this new evolution of the genre must be underlined. Firstly, while news tickers were initially superimposed during TV newscasts, their migration in order to display emergency alerts brought a change in the space they were assigned to. Indeed, since emergency alerts were generally announced in the ways we have previously described (i.e., an audio signal followed by the emergency message), news tickers were forced to occupy their own space at the end of a newscast or they interrupted a newscast to convey a weather alert. Thus, they were presented as a sort of breaking news story that interrupted the scripted flow of news delivering. In this way, news tickers started to be associated with the idea of unrelated items to the news stories being presented by the anchor. A second change in their nature is linked to viewers' perception of news tickers. Indeed, since they

were presented as breaking news stories, viewers associated them with the idea of urgency and abruptness. Therefore, news tickers were viewed and perceived as linked to rapid and unforeseen changes in social routines.

This explains why their future developments are connected to rapid economic changes (i.e., stock market indexes), sports, and election results. Indeed, while ESPN in the mid-1980s brought back tickers in order to display sports results and news updates at the beginning or in the middle of each hour (the so-called ':28/:58 update'; Deese 2011), in the late 1980s, CNN's *Headline News* started using a continuous ticker, featuring stock prices during trading hours in order to compete with the Financial News Network. In this way, news tickers migrated again to their original environment, that is, they were once more used to convey stock prices over long distance. However, up until 1996, these stock tickers were manually transcribed, thus, showing recurring human mistakes. Only in 1996, the CNNfn network started using a fully automated stock ticker.

In December 1993, HLN Sports tested the first 24/7 ticker on the GCTV cable system. The most important innovation of the HLN Sports ticker was the fully automated combination of computer-generated graphics with a wire news service that worked as follows: a computer-based software recognized the conventions and labels in the wire service data and converted them into the words and symbols displayed along the screen. Following the same path taken by HLN Sports, in 1996, ESPN2 debuted a ticker, presented as the 'BottomLine', which featured non-stop sport results and news.

However, we must underline that news tickers' layout, up until 9/11, was more similar to news' headlines, popping up from the bottom of the screen with major news stories or sport scores. Only on 9/11 their typical layout was developed, that is to say, a string of texts 'crawling' at the bottom of the TV screen from the right to the left corner (reproducing the 'zipper' that crowns One Times Square in New York City).

In the UK, the development and appearance of news tickers on TV news networks and programmes was quite late compared to the US scenario. Indeed, only on November 15, 2004 the BBC World News implemented crawlers during news broadcasting due to a radical makeover started in 2003.

While the genre of news tickers was slowly making its appearance on various TV networks, the demonstration that its conventions were recognised by viewers and associated with specific functions of the daily routine of TV newscasts is demonstrated by some examples where the genre was parodically used on TV satire programmes.

When introducing the notion of intertextuality, Fairclough (1992: 103) argues:

> [T]he intertextuality of a text can be seen as incorporating the potentially complex relationships it has with the conventions [...] which are structured together to constitute an order of discourse. [...] [T]exts may not only draw upon such conventions in a relatively straightforward way, but may also 'reaccentuate' them by, for example, using them ironically, parodically, or irreverently.

Therefore, parodying a genre means recognising its conventions and use them in an exacerbated way in order to provoke a comic reaction. In the words of Gray (2006), "[t]o laugh at parody is to acknowledge comprehension of those conventions under attack, and hence is also an acknowledgement of a genre's artificiality" (Gray 2006: 47).

The Australian TV programme *CNNNN* represents one of the most famous examples of a parodic use of news tickers. The programme, aired from September 19, 2002 to October 23, 2003 on ABC Television, represented a parody of American news channels CNN and Fox News. Besides overtly mocking their journalistic practices, *CNNNN* also featured a 'newsbar', where unrelated news tickers displayed fake news stories (e.g., one of them announced to viewers that "PINOCCHIO STUCK IN ABUSIVE RELATIONSHIP WITH TERMITE"; another example can be seen in Figure 4).

Figure 4 A screenshot taken from the first episode of the TV programme *CNNNN* aired on September 19, 2002 on ABC Television.

One of the most important characteristics of news tickers which was parodied in this TV programme is represented by the exacerbation of their typical purpose of displaying ancillary news stories which, conversely, were perceived as unrelated items compared to what the news anchor was saying.

The example of the *CNNNN*'s parody of the genre of news tickers is, therefore, a demonstration that this genre, already in 2002, was recognised as part of the genres of TV newscasts, so blended in the news routine that in parodying one the other was subsequently parodied.

3. Corpus Linguistics and Genre Analysis

Introducing the methodological framework

When deciding how to analyse given instances of naturally occurring linguistic data, researchers are increasingly facing the hard decision of whether to use a qualitative or a quantitative approach to their analyses, or a mixture of both of them (usually referred to as mixed approaches[4]). Additionally, micro and macro level of investigation come into play, in particular, when deciding on whether to focus the analysis on structural units of a given language (morphemes, words, phrases, etc.) or on how speakers and writers strategically use their linguistic resources to achieve their goals (Biber *et al.* 1998).

In the next paragraphs, we will briefly introduce the methodological framework of our investigation, which adopts corpus linguistics methodologies as its main analytical tool in order to highlight linguistic patterns that can be seen as the fingerprints left behind of the particular strategies journalists want to achieve in the genre of news tickers[5]. More precisely, while presenting the methodologies used throughout the research, this chapter also offers an overview of corpus linguistics methodologies (McEnery and Wilson 1996; Biber *et al.* 1998; Hunston 2002; McEnery *et al.* 2006; Baker 2006, 2010; McEnery and Hardie 2012), genre analysis (Biber 1988; Tribble 1999; Swales 1990; Bhatia 1993, 2004; Xiao and McEnery 2005), lexical priming (more specifically, as we will see, we will focus on the notion of textual colligation as introduced in the context of Hoey's (2005) approach to textual analysis), and news values (Bell 1991; Bednarek 2006; Bednarek and Caple 2012a, 2012b, 2014; Caple and Bednarek 2015; Potts, Bednarek and Caple 2015).

---

[4] See Christensen and Johnson (2000: 31-56) for an overview of the characteristics of the methodologies highlighted above.

[5] Some of the observations, definitions and examples offered in the following paragraphs are part of the notes taken by the author of this dissertation during the English Corpus Linguistics 2014 Summer School held at the University College of London, the UCREL Summer School in Corpus Linguistics 2014 held at Lancaster University, and the online MOOC in *Corpus linguistics: method, analysis, interpretation* offered by Lancaster University via the FutureLearn online e-learning platform (FutureLearn can be reached at the following address: `https://www.futurelearn.com/`). Thus, even though they are not directly acknowledged, we will draw on them for this brief introduction to the methodology used for our linguistic investigation.

3.1 Introducing Corpus Linguistics

The term, which was firstly used in the title of the book edited by Aarts and Mejis (1984), refers to that linguistic approach that deals "with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions" (McEnery and Hardie 2012: 1). In this sense, corpus linguistics is not the study of a particular aspect of language nor a particular branch of linguistics (such as semantics, syntax or pragmatics) but, rather, it can be defined as an area of investigation that focuses on a particular set of methods to study examples of naturally occurring texts.

However, while it can be argued that corpus linguistics is not an area of linguistic enquiry in itself, "it does, at least, allow us to discriminate between methodological approaches taken to the same area of enquiry by different groups, individuals or studies" (McEnery and Wilson 1996: 2). In other words, while corpus linguistics does not openly delimit an area of linguistics in itself, it nevertheless does so as a methodological approach in a particular linguistic field (in this sense, corpus-based semantic enquiries can be opposed to non-corpus-based semantic enquiries, for instance).

Given the versatility of its applications in various areas of linguistic enquiry, corpus linguistics is not a monolithic methodology, offering a consensually agreed set of methods for the exploration of linguistic phenomena (McEnery and Hardie 2012). On the contrary, it can be seen as a heterogeneous field of enquiry, where differences in the methods and procedures give rise to separations and subcategories in the use of corpus data. The most significant difference is the one highlighted by

Tognini-Bonelli (2001) between corpus-based and corpus-driven approaches[6]. The former approach can be defined as follows (Tognini-Bonelli 2001: 65):

> [...] a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study.

Thus, corpus-based approaches generally use corpus data in order to test a given theory or hypothesis and, according to Tognini-Bonelli (2001), this creates a series of limitations in corpus linguistic investigations. Indeed, analysts using corpus-based approaches can be "accused of not being fully and strictly committed to corpus data as a whole as they have been said to discard inconvenient evidence […] by 'insulation', 'standardisation' and 'instantiation', typically by means of annotating a corpus" (McEnery *et al.* 2006: 8). Insulation (Tognini-Bonelli 2001: 68-71) refers to making the data fit the adopted theory by, for instance, annotating the corpus according to the theoretical framework chosen by the researcher. Standardisation (Tognini-Bonelli 2001: 71-74), on the other hand, reduces the data to "a set of orderly categories which are tractable within existing descriptive systems" (Tognini-Bonelli 2001: 68), and this can be again achieved by annotating the corpus. As McEnery and Gabrielatos (2006) argue, the criticism towards this second technique used in corpus-based approaches to cope with corpus data refers both to the fact that

[6] In the framework developed by Partington (2004, 2009, 2015) and Partington, Duguid and Taylor (2013), we can also distinguish a third approach to corpus data, referred to as corpus-assisted discourse studies (CADS). As Taylor (2009) explains, "[c]orpus-assisted discourse studies tends to reject the dichotomy of corpus-based or corpus-driven linguistics [...], in favour of combining them, in an approach which uses both theory and data as starting points for the research, and therefore shunts between qualitative/quantitative analyses and theoretical interpretative frameworks as the research progresses" (Taylor 2009: 214). Corpus-based approaches, however, have slowly if not steadily included this type of approach and its methodologies in their framework of analysis and, thus, the distinction, while still functional when referring to Partington's (2004, 2009, 2015) framework of analysis, has been slowly disappearing. This is the reason why, even though our approach is strictly related to Partinton's (Partington, Duguid and Taylor 2013) claim that in analysing language in context, one should never "treat the corpus as an isolated black box" and the examination of corpus-external data can be helpful "to try and interpret and explain our data and also as a means of identifying areas for analysis" (Partington, Duguid and Taylor 2013: 11), we prefer to define our approach as corpus-based, in order not to strictly link it to any particular methodological framework.

"the annotation scheme is based on a pre-conceived theory, and the manual annotation of the training corpus is influenced by both the theory and the annotator's intuitions" (McEnery and Gabrielatos 2006: 36). Finally, instantiation (Tognini-Bonelli 2001: 74-77) refers to "building the data into a system of abstract possibilities, a set of paradigmatic choices available at any point in the text" (Tognini-Bonelli 2001: 74). Associated with Halliday's (1991, 1992) view of grammar, this technique is criticized mainly because "its focus is predominantly paradigmatic rather than syntagmatic, that is, it is concerned with grammar rather than lexis" (McEnery and Gabrielatos 2006: 36), and since "grammatical patterns […] are not easily retrievable from a corpus unless it is annotated" (Tognini-Bonelli 2001: 77), also in this case the drawback of this technique is strictly related to the annotation scheme used in the corpus data.

In corpus-driven approaches, on the other hand, "the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence" (Tognini-Bonelli 2001: 84). This approach to corpus data sees the corpus as "the sole source of our hypotheses about language" (McEnery and Hardie 2012: 6). Thus, while the definition of corpus linguistics as a methodology is more appropriate in a corpus-based approach, "[c]orpus-driven linguistics rejects the characterisation of corpus linguistics as a method and claims instead that […] the corpus itself embodies its own theory of language" (McEnery and Hardie 2012: 6). However, corpus-driven approaches also display some limitations. Indeed, if the corpus is the sole source of our theory of language, and if natural languages are infinite, can the corpus data capture all the instances of a language? And if not, are corpus-driven studies involuntarily and inevitably sampling the linguistic data? Additionally, as Stubbs (1996: 47) argues:

> [T]he concept of data-driven linguistics must confront the classic problem that there is no such thing as pure induction […]. The linguist always approaches data with hypotheses and hunches, however vague.

Thus, corpus-driven approaches might be seen as an idealized extreme (McEnery and Gabrielatos 2006), since preconceptions are inherently part of the analyst or are subtly built in the corpus construction.

In conclusion, whether one chooses to adopt a corpus-based or a corpus-driven approach to the analysis of the linguistic data, one should always be aware of the limitations of both approaches, and claims regarding the superiority of one approach compared to the other should always be avoided. However, both approaches seem to highlight the importance linked to the way the linguistic data are collected. Thus, in the following section, we would like to provide a definition of the core element of corpus linguistics, that is to say, the corpus itself.

3.1.1 What is a corpus?

As we have seen in the previous section, corpus linguistic approaches the study of language in use through corpora[7]. A corpus can be roughly defined as a collection of naturally occurring examples of language stored on a computer. More precisely, a corpus can be defined as (Tognini-Bonelli 2001: 55, emphasis added):

> […] a *computerised* collection of *authentic* texts, amenable to automatic or semiautomatic processing or analysis. The texts are selected according to *explicit criteria* in order to capture the *regularities of a language, a language variety or a sub-language*.

However, it must be promptly underlined that these examples of language in use should usually be of a size that defies analysis by hand and eye alone, and this is the reason why they should be in a machine-readable format. Indeed, as McEnery and Hardie argue (2012: 2), "[u]nless we use a computer to read, search and manipulate the data, working with extremely large datasets is not feasible because of the time it

[7] We must underline that, as Gries (2011) argues, the term 'Corpus Linguistics' may refer to "(i) the study of the properties of corpora or (ii) the study of language on the basis of corpus data" (Gries 2011: 83). We can refer to these two different uses of corpus linguistics as corpora for corpora's sake and corpora for language's sake. According to the author, this distinction is necessary in order to distinguish between those approaches that focus on the frequency, for instance, of linguistic phenomena in a corpus and those approaches that, on the other hand, focus on methodological corpus issues, such as "how efficient different approaches to tagging a corpus are, or determining which kind of clustering algorithm best distinguishes different registers on the basis of *n*-gram frequencies" (Gries 2011: 83). Thus, while the focus in this contribution is on what corpus linguistics can do when applied to language studies, it must be acknowledged that corpus linguistics can also be used by researchers who are not necessarily interested in linguistic issues.

would take a human analyst". Additionally, manually working with large corpora increases the chances of human mistakes in the analysis of the data. Thus, in corpus linguistic approaches, "[c]orpora are invariably exploited using tools which allow users to search through them rapidly and reliably" (McEnery and Hardie 2012: 2). Computer software enable frequency-based statistical analyses (among others) to be carried out, as well as presenting the data so that patterns can be more easily observed. However, these analyses can be carried out insofar as the computer is able to search through the data (see Section 3.1.4).

The use of machine-readable linguistic data also refers back to the sometimes controversial debate between rationalists and empiricists. Indeed, by using naturally occurring examples of language in use, corpus linguistics is inevitably an empirical methodology, in contrast with intuition-based approaches, where researchers can instantly come up with examples to support or discard given hypotheses (see McEnery and Wilson 1996: 5-13 for a more detailed discussion[8]). However, intuition must be used with caution since, as McEnery *et al.* (2006) highlight, the data might be influenced by the speaker's dialect/sociolect; the invented examples might be regarded as unauthentic, as they have been monitored by the speaker in order to produce them; or, finally, data based on introspection are difficult to verify as introspection is not observable. Thus, in contrast with intuition-based approaches, corpus-based approaches draw upon authentic texts that can help researchers retrieve differences that intuition alone cannot perceive by providing examples of reliable quantitative data. Additionally, since the corpus is normally a collection of what speakers produce and, thus, their utterances can be regarded as grammatical and/or

[8] In addition to McEnery and Wilson's (1996) discussion on the distinction between corpus-based and intuition-based approaches to linguistic analysis, we must also acknowledge Partington's (2008) comment on the distinction between the "armchair linguist" (i.e., intuition-based approaches; Partington 2008: 95) and the "one-dimensional, lazy-minded corpus linguist" (i.e., corpus-based approaches; Partington 2008: 95), which was firstly introduced by Fillmore (1992). Partington comments that both approaches must be used with caution and, more importantly, that "[g]ood corpus linguists [...] exploit the interaction of intuition and data, giving balanced attention to analysis, description, interpretation, explanation" (Partington 2008: 96). In this sense, corpus linguists should not avoid the combination of theory and observational data and, thus, after observing the empirical data, "the corpus linguist may well then retire to his or her armchair to reflect upon them" (Partington 2008: 96).

acceptable, corpus-based approaches can also offer researchers instances of what speakers in general believe to be acceptable utterances in their language.

However, intuition is not completely rejected in corpus-based approaches. Indeed, "[t]he key to using corpus data is to find the balance between the use of corpus data and the use of one's intuition" (McEnery *et al.* 2006: 7). Thus, the two approaches are not mutually exclusive, but they are complementary. In this sense, corpus linguistics can be seen as a synthesis of both approaches, "relying on a mix of artificial and natural observation" (McEnery and Wilson 1996: 19).

As seen in Tognini-Bonelli's (2001) definition, the data collected in a corpus should be selected according to explicit criteria, that is to say, the corpus data that we have collected are strictly linked to the intended use of the corpus itself. Thus, the criteria in collecting the corpus are generally external to the texts and they are usually selected to explore a research question. Thus, the data collected must be well-matched for that research question. As McEnery and Hardie (2012) explain, "we cannot (or can only with some caution) make general claims about the nature of a given language based on a corpus containing only one type of text or a limited number of types of text" (McEnery and Hardie 2012: 2). Thus, the criteria in building a corpus should always be clear and the generalisations made on the basis of the linguistic data should always match the sample we have set out to explore.

Another crucial characteristic of a corpus is represented by the fact that it contains data that can be regarded as authentic, that is, as natural occurrences of language in use. However, does this mean that every kind of naturally occurring language can be considered as a corpus? Sinclair (2005: 15) excludes some categories as linguistic corpora, such as a single text[9], an archive, and the World Wide Web. In the specific case of the Web as corpus (Kilgarriff and Grefenstette 2003), Sinclair (2005) discards it as a corpus since "its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective" (Sinclair 2005: 15). Kilgarriff and Grefenstette (2003), on the other hand, by way of advocating the study of data collected from the World Wide Web,

[9] Tognini-Bonelli (2001: 3) defines a text as strictly linked to "a unique communicative context" and it can be seen "as a single, unified language event mediated between two (sets of) participants". For a detailed analysis of the differences between a text and a corpus, see Tognini-Bonelli (2001: 2-4).

define a corpus as "*a collection of texts when considered as an object of language or literary study*" (Kilgarriff and Grefenstette 2003: 334, emphasis in the original). Thus, authenticity in this last definition becomes something quite blurred, while the research question at the very heart of a given linguistic investigation becomes crucial in the analysis of the data collected. As Gries and Newman (2013) argue, "some collections of language can diverge from the prototypical property of being "naturally occurring language," and yet are still happily referred to as corpora by their creators" (Gries and Newman 2013: 258). In order to support this claim, the authors offer as an example the TIMIT Acoustic-Phonetic Continuous Speech Corpus[10], which contains broadband recordings of 630 speakers of eight major dialects of American English each reading ten phonetically rich sentences. While it represents "a uniquely valuable resource for the study of acoustic properties of American English" (Gries and Newman 2013: 258), the corpus cannot be considered as a collection of naturally occurring language. However, since the focus of this research was on capturing the phonetic variation of American dialects, this collection can rightfully be called a corpus.

Closely related to issues of authenticity is the "vexed question" (Tognini-Bonelli 2001: 57) of the representativeness of a corpus, that is, the way the data have been collected in order to be "representative of a given language, dialect, or other subset of a language" (Francis 1992: 17). While we will focus our attention on this controversial aspect of a corpus more extensively in Section 3.1.3, for the moment, in order to better understand why this concept is crucial in defining a corpus, we would like to offer the following quotation by Biber *et al.* (1998: 246), which perfectly 'bottles up' the essence of representativeness of a corpus:

> A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research

[10] The TIMIT Acoustic-Phonetic Continuous Speech Corpus is available for online purchase through the Linguistic Data Consortium at `https://catalog.ldc.upenn.edu/LDC93S1`

> questions that can be addressed and the generalizability of the results of the research.

Thus, representativeness constitutes a key element of what makes a corpus and, consequently, what makes a corpus is strictly linked to what it wants to capture. However, achieving representativeness is quite an elusive and controversial goal, as we will see. But, for the moment, we would like to stress the fact that since representativeness, as Biber *et al.* (1998) highlight, shapes the way the corpus has been collected and built and, additionally, since it is strictly related to the research questions linguists want to answer when they built a corpus, it can be seen as the very foundation of any corpus to be.

As we have seen in this section, a corpus is not simply a collection of words put together in a computer-readable format: a corpus should at least comply with some of the characteristics highlighted previously. We have also seen that there is not much consent on some of these characteristics. As Gries and Newman (2013: 258) argue, "[b]eyond being a body of naturally occurring language, then, it is difficult to agree on any more particular definition of what a corpus is or is not"[11]. In this contribution, however, we adopt Tognini-Bonelli's (2001: 55) definition of what a corpus is, since it has been used as our standard in building the corpus at the centre of our investigation.

3.1.2 Why use a corpus?

In the previous sections, we have advocated for a corpus-based approach without, however, offering any reasons why a corpus linguistic approach should be preferred to any other approach. Thus, in this section, we would like to underline some of the advantages (and disadvantages) that corpus-based approaches can provide researchers when investigating language in use.

Firstly, a corpus can be used in order to look at large amounts of data, which tell us, as we have previously argued, about tendencies and what is normal or typical

[11] However, we have also seen in Gries and Newman (2013) that the property of being a collection of naturally occurring instances of language in use comes into question when a corpus is shaped so as to, for instance, elicit given responses in participants in the case of phonetic variation investigations.

in real-life language use. Indeed, corpora show us things that we are doing routinely on a daily basis and that we find it very hard to imagine that we are doing. In other words, this brings us back to the dichotomy between intuition-based and corpus-based approaches. And in order to better understand this, we would like to offer an example of the investigation carried out by Subtirelu (2014) on the gendered nature of the adjective 'bossy'. Starting from the fact that language shapes the way we construct given identities, in particular, the way words are used to label young girls' behaviours and how the labels are often applied differently to young girls but not to young boys, Subtirelu (2014) used the case of the adjective 'bossy' in order to prove this. By using the Corpus of Contemporary American English (COCA)[12], Subtirelu (2014) investigated "to whom the word was applied most frequently, specifically to people of which gender" (Subtirelu 2014). The picture below summarises the findings of his investigation:

[12] As we will see in Section 3.1.5, the BYU Corpus of Contemporary American English (available at `http://corpus.byu.edu/coca/`), at the time of writing, contains 450 million words sampled from five sources (spoken, fiction, popular magazines, newspapers, and academic journals), with 20 million words for each of the 17 years during the period 1990-2012. The corpus is freely accessible to users and represents what is usually defined as a monitor corpus, that is, a corpus that dynamically "grows" in time, with new linguistic samples of the genres included in the corpus added each week, month or year.

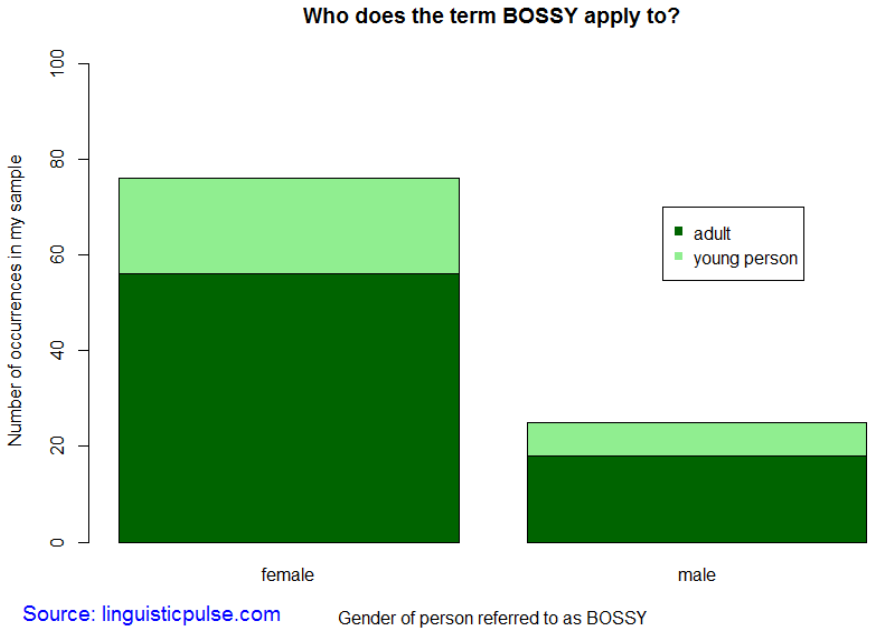


Figure 5 Subtirelu's (2014) investigation on the gendered nature of the adjective 'bossy'.

As we can see from Figure 5, the initial hypotheses were confirmed by looking at a corpus containing large amount of data. This has revealed the gendered nature of the word 'bossy', which seems to typically be used to refer to women, something that intuition alone would not be able to underline or prove.

The example provided here refers back to one of Baker's (2006) advantages listed by the author in combining corpus-based approaches to discourse analysis. More specifically, Subtirelu's (2014) example proves that corpus linguistics methodologies can be useful to discourse analysts in order to highlight the incremental effect of discourse, that is, they can be used in order to uncover "how language is employed, often in quite subtle ways, to reveal underlying discourses" (Baker 2006: 10). In other words, corpus-based approaches to discourse analysis can help us become "more aware of how language is drawn on to construct discourses or various ways of looking at the world" (Baker 2006: 10), and uncover how producers of texts can "manipulate us by suggesting to us what is 'common-sense' or 'accepted wisdom'" (Baker 2006: 10)[13].

[13] For a complete list of the advantages and limitations of a corpus-based approach to discourse analysis, see Baker (2006: 10–17).

Using corpora can, additionally, reveal instances of very rare or exceptional cases that we would not get from looking at single texts or by relying only on introspection. This last advantage in using corpus linguistic methodologies is particularly true in the case of researchers working in lexicography or second language acquisition (SLA). Indeed, from a lexical point of view, by highlighting rare cases, we can discover, for instance, new spelling variations in a given language, new words that have just entered the vocabulary, and so on. From an SLA perspective, on the other hand, researchers can highlight linguistic phenomena that can be regarded as non-native and, thus, ways to improve teaching methodologies in second language acquisition.

Another advantage in using corpora in order to carry out a linguistic analysis refers back to the question of accuracy and the reliance on computer technologies in order to analyse linguistic data. Indeed, it can be universally acknowledged that human researchers make mistakes and are slow. Computers, on the other hand, are much quicker and more accurate, insofar as the computer is able to search through the data. Thus, using a corpus-based approach to the analysis of language can improve the reliability of the results of a given investigation.

Also in this case, we can refer back to the advantages in using corpus-based approaches to do discourse analysis provided by Baker (2006) in order to add that using large amounts of data can also reduce researchers' bias. Indeed, as the author argues (Baker 2006: 12), "[b]y using a corpus, we at least are able to place a number of restrictions on our cognitive biases". And even though biases can still be present in the theoretical framework that we have chosen in order to analyse the corpus data, "[i]t becomes less easy to be selective about a single newspaper article when we are looking at hundreds of articles" (Baker 2006: 12).

Finally, among the various advantages highlighted here, we would also like to include one that is strictly related to our investigation. Indeed, as Bhatia (2002) argues, corpora can help us "reveal patterns of form and use in particular genres and areas of language that are unsuspected by the researcher and difficult to perceive in other ways" (Bhatia 2002: 13). While Swales (2002) has initially expressed some concerns regarding the use of corpora, since the bottom-up approach (in particular, the one linked to using concordances to study search items) seems to clash with the

typical top-down approach of his move structure analytic framework (Lee 2008), during the years, he has changed his view on corpus linguistic methodologies, proven by his work with Lee (Lee and Swales 2006). In this paper, they used specialised corpora of academic writing and speaking made available to doctoral students as tools for directed learning as well as self-learning of the rhetorical strategies used in their academic fields and sub-fields of research. Corpus linguistic methodologies were, thus, used in order to reveal how particular moves were linguistically realised and, more importantly, how a quantitative approach helped develop a sense of self-awareness in doctoral students of the way given linguistic features were peculiar of some genres when compared to others.

Another example of how corpus linguistic methodologies can be used in genre analysis is provided by Kanoksilapatham (2003). In her work, she uses a corpus of 60 research articles (320,000 tokens) taken from the top five journals in biochemistry in order to identify what rhetorical moves are commonly used in this specialised setting and how these moves are expressed. This analysis was carried out by combining Swales' (1990) qualitative move analysis (see Section 3.2.2) with Biber's (1988) quantitative multidimensional framework (see Section 3.2.1). The combination of these two approaches helped uncover a basic template of the structure of scientific research articles and identify sets of linguistic features commonly used to express scientific ideas.

As we have seen in this section, various are the advantages that can be highlighted in using a corpus-based approach and these represent the reason why a corpus should be taken into consideration when approaching the study of language. However, there are also many disadvantages in using a corpus-based approach. One of them is represented, for instance, by the fact that, since the corpus only focuses on language, all other semiotic realisations that also contribute to the understanding of messages (e.g., given pitches of the voice, using particular gestures, etc.) are excluded from the analysis. While multimodal corpora can be used to overcome this

limitation, as Knight (2011) argues, they are still in their infancy and, thus, corpus-based approaches are still restricted to verbal semiotics[14].

Another limitation of corpus-based approaches is represented by the fact that the social context where the linguistic data were retrieved are sometimes excluded from the corpus. This limitation can be overcome if these pieces of information are encoded by researchers through corpus annotation (see Section 3.1.4). However, since these procedures are sometimes time consuming, such information may be inevitably lost and we may, thus, be facing corpora containing "decontextualized examples of language" (Baker 2006: 18).

Some of the limitations highlighted here can be overcome by following given criteria in the collection of a corpus. This is the reason why we turn our attention on these criteria as they help researchers spell out the limitations of their approach and, thus, embrace them by making them clear to readers (something that qualitative approaches, for instance, tend to defocus on). Indeed, as Hunston (2002) argues, "[b]ecause data in corpora are de-contextualised, the researcher is encouraged to spell out the steps that lie between what is observed and the interpretation placed on those observations" (Hunston 2002: 123).

3.1.3 Criteria in choosing or building a corpus

While some of the criteria in choosing or building a corpus have already been introduced in Section 3.1.1, in the following paragraphs, we would like to introduce some other characteristics that researchers must take under consideration when shaping the corpus they are building or when selecting an already-built corpus to use for their investigation.

In order to underline the basic criteria that researchers should have in mind when deciding to collect or choose a corpus, we would like to offer the following quotation taken from Tognini-Bonelli (2001: 2, emphasis added) that she uses in order to define what a corpus is:

[14] An interesting and fascinating combination of corpus-based approaches and multimodal analysis in order to investigate how news values are constructed in newspapers articles is offered by Bednarek and Caple (2012a, 2012b, 2014) and Caple and Bednarek (2015).

> […] a collection of texts assumed to be *representative* of a given language put together so that it can be used for *linguistic analysis*. Usually the assumption is that the language stored in a corpus is *naturally-occurring*, that is, gathered according to *explicit design criteria*, with *a specific purpose in mind*, and with a claim to represent natural chunks of language selected according to *specific typology*.

Thus, when building a corpus (in the words of McEnery *et al.* 2006: 7, a DIY corpus, that is, a 'do-it-yourself' collection of data) or choosing the corpus to use, we should think above all of the purpose of the use of that corpus. The research questions and hypotheses crucially shape the corpus that we are collecting or choosing; they determine whether that corpus is going to be useful, if we are selecting it 'off the shelf'; or they influence the way the corpus should look like, if we are going to be building it. In the latter case, the criteria followed in order to collect these data should be explicitly acknowledged by the researchers in order to highlight how they are related to the hypotheses or research questions at the very heart of their linguistic investigation.

Another important criterion when building or selecting a corpus for a particular linguistic analysis is the already mentioned and quite controversial notion of the representativeness of a corpus. As we have already seen in Section 3.1.1, according to this criterion, the corpus must be representative of the language or genre of that language: in other words, the corpus must be representative of some type of language so that we can start to measure it up against the research questions we have posited. In this way, the corpus can act as some type of nominal standard reference about what is typical in a language, if the corpus is built to be broadly representative of that language, or about what is typical in a given genre. As we have repeatedly underlined, the notion of representativeness is quite controversial, since its definition varies on the basis of the research questions or hypotheses of given scholars, or on the basis of the researchers' approach (i.e., corpus-based vs. corpus-driven approaches to Corpus Linguistics; see Section 3.1). For instance, as Biber (1993b) argues, "[t]ypically researchers focus on sample size as the most important consideration in achieving representativeness: how many texts must be included in the corpus, and how many words per text sample" (Biber 1993b: 243). However,

while size is certainly important while building our corpus (as we have seen in Section 3.1.1), it is not indicative *per se* of the representativeness of the corpus: rather, it is indicative of the generalisations that we can make about the type of language stored in that corpus. Besides, putting it more simply, as in the case of the age-old question: 'how much wood would a woodchuck chuck if a woodchuck could chuck wood?', the same goes for corpus size: 'specifically, how many words and/or genres should the corpus include in order to become representative?'. Quite the conundrum. Additionally, representativeness should be interpreted in a loose sense (Hunston 2002), since "[t]he problem is that 'being representative' inevitably involves knowing what the character of the 'whole' is" (Hunston 2002: 28) and, given the fact that language is not finite, but infinite, knowing the 'whole' is impossible and, thus, representativeness can be regarded only as "the extent that findings based on its [the corpus'] contents can be generalized to a larger hypothetical corpus" (Leech 1991: 27). Thus, in defining a corpus as representative of a given language or genre, "a thorough definition of the target population and decisions concerning the method of sampling are prior considerations" (Biber 1993b: 243).

In this sense, representativeness is driven by text external criteria, that is, according to situational and not purely linguistic criteria. External criteria are defined as "situationally irrespective of the distribution of linguistic features" (McEnery *et al.* 2006: 14), while internal criteria are "defined linguistically, taking into account the distribution of such features" (McEnery *et al.* 2006: 14)[15]. In building a corpus, thus, one can choose to structure it by following text internal or text external criteria. However, the latter are generally the best suited, since they enrich the corpus with additional information that linguistic elements alone cannot provide (i.e., the so-called context of situation, that is, the "environment in which meanings are being exchanged", Halliday and Hasan 1989: 12). Text internal criteria, on the other hand, should be avoided since, as McEnery *et al.* (2006) argue, they create an inexorable circularity in the linguistic analysis. Indeed, if a corpus has been collected to study given linguistic distributions, designing it according to this distribution compromises

[15] On the basis of this distinction, Biber (1988) refers to situationally defined text categories as registers or genres, while linguistically defined text categories are referred to as text types. See Section 3.2.1 for an overview of Biber's multidimensional approach to genre analysis.

the analysis, since "[i]f the distribution of linguistic features is predetermined when the corpus is designed, there is no point in analysing such a corpus to discover naturally occurring linguistic feature distribution" (McEnery *et al.* 2006: 14). In other words, following internal criteria in designing a corpus clashes with one of its main characteristics, that it, the fact that it should contain naturally-occurring instances of language in use.

Finally, another important criterion that can be taken under consideration specifically when building a corpus is whether it should include metadata, annotations, and mark-ups. As we have previously seen, the choice between using or not a raw corpus, without any encoded information of the original context where the data were retrieved or linguistic information based on a given theory of language, is strictly linked to the difference between a corpus-based and a corpus-driven approach[16]. Thus, while we have included annotations to the list of criteria when choosing/building a corpus, it should be considered as a choice that lies upon researchers and the approach selected when analysing the data. However, since our investigation relies on a corpus-based approach to the analysis of the genre of news tickers, we agree with Leech (2005) when he states that "annotation is a means to make a corpus much more useful – an enrichment of the original raw corpus" (Leech 2005: 17). While further information on the types of metadata, annotations, and mark-ups that can be coded in a corpus will be offered in Section 3.1.4, for the moment, we would like to introduce some of the advantages and disadvantages in encoding such information in a corpus. Firstly, as Leech (2005) argues, "adding annotation to a corpus is giving 'added value', which can be used for research by the individual or team that carried out the annotation, but which can also be passed on to others who may find it useful for their own purposes" (Leech 2005: 17). Thus, annotating the corpus can be seen as a way of ensuring the replicability of the analyses carried out by researchers and, thus, making the corpus 'reusable'. However, while the corpus may be annotated for a given purpose, this does not entail

[16] As a way of reminding the stance of corpus-driven approaches towards corpus annotation, we would like to offer the following quotation taken from Sinclair (2004), who states that "[i]n corpus-driven linguistics you do not use pre-tagged text, but you process the raw text directly and then patterns of this uncontaminated text are able to be observed" (Sinclair 2004: 191). The focus of this quotation is on the adjective 'uncontaminated', meaning that introducing external information in the corpus is bound to corrupt its integrity.

that it cannot be used for another one. Indeed, the type of annotation chosen in a given corpus can be re-used in order to answer other research questions or prove/discard different hypotheses[17]. Additionally, corpus annotation records a linguistic analysis explicitly and, thus, it "stands as a clear and objective record of analysis that is open to scrutiny and criticism" (McEnery *et al.* 2006: 30) or so that researchers can replicate given studies. In this sense, annotations also provide a standard reference resource, "so that successive studies can be compared and contrasted on a common basis" (McEnery *et al.* 2006: 30). Thus, corpus annotation opens up a dialogue between researchers and, while it can be simplistically seen as adding value to the corpus, it rather adds value to future generations of researchers, creating a storyline in the analysis of given phenomena.

Corpus annotation, however, does have its limitations and disadvantages. Indeed, as Hunston (2002) argues, annotations may produce cluttered corpora, since "[h]owever much annotation is added to a text, it is important for the researcher to be able to see the plain text, uncluttered by annotational labels" (Hunston 2002: 94). Since most corpus exploration tools allow users to hide annotations, this criticism is directed more to those researchers who base their sole analysis, for instance, on the frequency of given tags encoded in the corpus, without browsing through the corpus in order to verify that the tags do reveal certain patterns. This should always be done since, in the specific case of automatic corpus annotation, for instance, "an automatic annotation program is unlikely to produce results that are 100% in accordance with what a human researcher would produce; in other words, there are likely to be errors" (Hunston 2002: 91). The same can be said of computer-assisted or manual corpus annotation (see Section 3.1.4), since human beings do generally make mistakes. Thus, reading the raw corpus in order to see if the annotation scheme has truly revealed given patterns is something that should not be avoided.

[17] The corpus annotation scheme encoded in the British National Corpus (BNC) has not precluded, indeed, the development of an enormous amount of linguistic investigations based on different linguistic theories. See the BYU website for some examples of researches carried out by using their web-interface of the BNC (available at `http://corpus.byu.edu/publicationSearch.asp?c=bnc`).

Leech (1993) also underlines the importance of being able to reverse back to the raw corpus in his so-called seven 'maxims' of corpus annotation[18], stating that "[i]t should always be easy to dispense with annotations, and revert to the raw corpus. The raw corpus should be recoverable" (Leech 1993: 275). This is the reason why the type of annotations used and their meanings should be stored separately by researchers, thus, allowing their deletion, if need be, for future researches based on that corpus (Leech 1993).

Going back to the criticisms towards corpus annotation, one of the major drawbacks in annotating a corpus refers back to the objection made by those who prefer a corpus-driven approach to the data collected, that is, the fact that corpus annotation imposes a given linguistic analysis on a corpus. More importantly, as Hunston (2002) argues, "the categories used to annotate a corpus are typically determined before any corpus analysis is carried out, which in turn tends to limit, not the kind of question that can be asked, but the kind of question that usually is asked" (Hunston 2002: 93). However, while users may ignore the annotation scheme and, thus, the type of linguistic approach subtly encoded in the annotation process, McEnery *et al.* (2006) additionally consider this drawback as an advantage rather than a weakness of a corpus, since annotation, as we have previously seen, "provides an objective record of an explicit analysis open for scrutiny" (McEnery *et al.* 2006: 31). In other words, corpus annotation makes the type of linguistic approach and methodology used in analysing the data clearly available to other users. Indeed, while a raw corpus can be easily re-used for other linguistic investigation, it can also make it impossible to recover what a given researcher has done in approaching the data, thus, challenging that 'communication channel' open to confutation that is generally referred to as research.

A final criticism that can be moved towards corpus annotation is that it can 'overvalue' (McEnery *et al.* 2006) the corpus: in other words, while making the corpus more useful and exploitable, corpus annotation conversely makes the corpus "less readily updated, expanded or discarded" (Hunston 2002: 93). The corpus is thus frozen in time and less accessible, and this limitation should be kept in mind in the

[18] For an exemplification of the seven maxims of corpus annotation introduced by Leech (1993), see McEnery and Wilson (1996: 33-34).

type of research questions that we want to answer when using or collecting a given corpus.

3.1.4 The mark-up of a text: Metadata, annotations and mark-ups

In the last paragraphs of Section 3.1.3, we have argued that one of the criteria that researchers should follow in building or using a corpus is represented by corpus annotation. In this section, we are going to expand on the notion of metadata, annotations, and mark-ups. And by way of introducing them, we would like to start by highlighting that usually, in the literature, no difference is made between these, since they are generally referred to as corpus annotation encoding, that is, the process whereby "textual/contextual information and interpretative linguistic analysis" (McEnery *et al.* 2006: 29) are encoded in the corpus. In this sense, annotating the corpus in a broad sense means adding information to a corpus file that are not really part of the collected data: they are a way of encoding the original context where the data were collected or adding linguistic information about the data collected. However, in this contribution, we follow McEnery and Hardie's (2012) distinction between three types of information that can be encoded in a corpus and that can be helpful in the investigation of the data: metadata, mark-ups, and annotations.

Metadata usually represent valuable information that tells us something about the text itself (McEnery and Hardie 2012). In a written text, for instance, metadata may include information about the author of that text, the date when the text was written, and other bibliographic information about the original text. Metadata can also provide interpretative categories (Burnard 2005), that is, the genre, domain, etc., of the original text. Mark-ups, on the other hand, are codes inserted into a corpus file to indicate structural features of the original text other than the actual words of the text. Again, in a written text, for example, mark-ups might include paragraph breaks, sentence boundaries, omitted pictures, and other aspects of the layout of the original text. Finally, annotations in a narrow sense refer to the encoding of linguistic analyses in the data, and the most common form of corpus annotation is represented

by part of speech tagging (POS-tagging)[19], that is, "including […] tags which label each word in a corpus as to its grammatical category (e.g. noun, adjective, adverb, etc.)" (Reppen 2010: 35).

Metadata, annotations, and mark-ups are used in order to communicate with our computers. Indeed, a computer does not have all the knowledge of language, of reading text, and of the conventions of setting up the text that we have. So, almost anything that we might want to say about that text must be made explicit to the computer. In order to encode these information in a way that a computer may be able to look through them and, more importantly, allow researchers to look in given sections of the corpus or retrieve given linguistic phenomena, the most used encoding system is represented by the *eXtensible Markup Language*, also known as XML, adopted by the Text Encoding Initiative (TEI) as a way to "facilitate data exchange by standardizing the mark-up or encoding of information stored in electronic form" (McEnery *et al.* 2006: 24). As we have previously seen, metadata and mark-ups allow us to tell the computer how the text looked like when we collected it and how the text was originally structured. So, to indicate, for instance, that a given clause is a newspaper headline and that the clause is a stand-alone sentence, we can use the following XML encoding (Hardie 2014: 95)[20]:

[19] POS-tagging is just one of the many ways a corpus can be annotated. The type of corpus annotation chosen depends on the type of research questions that researchers want to investigate through a certain corpus. For instance, if our research questions are strictly related to phonological issues in contemporary British English, thus, POS-tagging may not be useful in order to answer our research questions and phonetic or prosodic annotations may be preferred. If we are interested, on the other hand, in syntactic phenomena captured by our corpus, thus, parsing, treebanking or bracketing (McEnery and Wilson 1996: 53-61) might be a better way to annotate our corpus (see, for instance, the ICE-family of corpora: `http://ice-corpora.net/ice/index.htm`). Finally, if we are interested in a semantic analysis of our corpus, semantic annotation is thus preferred to POS-tagging (for an automatic semantic tagger, see the UCREL Semantic Analysis System (USAS): `http://ucrel.lancs.ac.uk/usas/`).

[20] Hardie's (2014) paper represents a lightweight edition of the TEI's guidelines. As he argues (Hardie 2014), the standards offered by the Text Encoding Initiative are "over-weighty in that the standards themselves are extremely complex documents" (Hardie 2014: 77). Thus, in his contribution, he simply covers those standards that offer a "*sufficient knowledge about XML for most corpus linguists' day-to-day needs*" (Hardie 2014: 73, emphasis in the original), such as use of tags, adding attribute-value pairs, nesting tags, encoding of special characters, XML well-formedness, etc.

(1) <head><s>MAN BITES DOG</s></head>

In this examples, the <head> tag is used in order to signal that the text that is found between the tag <head> and the tag </head> is a newspaper headline (with all its formal characteristics, e.g. bold character, etc.), while the <s> and </s> tags are used in order to indicate the start and end of a sentence. Additionally, we can also use an automatic POS-tagger in order to linguistically annotate example (1)[21]:

(2) <head><s><w id="2.1" pos="NN1">MAN</w> <w id="2.2" pos="VVZ">BITES</w> <w id="2.3" pos="NN1">DOG</w> </s></head>

As we can see, grammatical roles have been automatically encoded in example (2), where the <w id> tag indicates the position that a given word occupies in a sentence, while the <pos> tag indicates its role in the sentence (i.e., NN1 indicates that the word MAN and the word DOG are singular nouns; and VVZ indicates that BITES is the third person singular of the present simple of the lexical verb 'to bite').

The type of encoding shown in example (1) and (2) can help us to search the corpus, for instance, for a particular linguistic phenomenon occurring only in the headlines of a given newspaper or, more generally, corpus annotation can help us highlight given linguistic phenomena occurring only in headlines, thus, defining the lexicogrammatical characteristics of the genre.

Metadata, annotations, and mark-ups can be introduced into a corpus through purely manual, semi-automatic or purely automatic annotation. As we have previously underlined, the annotation process is, in general, not error free. Indeed, in the case of purely automatic annotation, Garside and Smith (1997) acknowledge an accuracy of 97% or more in the case of the automatic tagger CLAWS4. Semi-automatic or purely manual annotation, on the other hand, have to take into account

[21] For this example, the online automatic POS-tagger CLAWS (Garside *et al.* 1987) has been used, which is available at http://ucrel.lancs.ac.uk/claws/ We must underline that this automatic tagger was not used for our linguistic investigation since, as we will see, Sketch Engine (Kilgarriff *et al.* 2004; Kilgarriff *et al.* 2014), the online corpus analysis tool that we have used in order to analyse our data, provides its own automatic POS-tagger (see Section 4.2).

the human factor and, thus, the fact that the researcher tagging the corpus may make some mistakes. However, the type of annotation procedure chosen is strictly related to the type of corpus that we want to build. In the case, for instance, of a large corpus, semi-automatic or purely automatic annotation is preferred since they rapidly allow researchers to tag the corpus. However, when it comes to small and specialised corpora, manual annotation may be preferred, since the text can be easily processed and tags can be reliably checked. Additionally, some types of annotations (such as pragmatic or discourse annotations) are preferably encoded into a text manually, since the theoretical approach at the very heart of this type of annotations cannot dispense of the researchers' interpretation of the data.

To sum up, through metadata, annotations, and mark-ups we are able to undertake quite sophisticated linguistically motivated queries of the data collected, thus, improving the quality of our searches through the data. Annotations also help researchers preserve information, thus, encoding in the corpus their interpretation of the data and making them available to other researchers to confute and, sometimes, rebut. Finally, annotations are useful tools in order to re-create in the corpus features of the original context where the data were collected, thus, preserving the authenticity of their original format.

3.1.5 Types of corpora

According to their representativeness and their purpose, corpora may come in different shapes and sizes. As Bowker and Pearson (2002) ague, there are as many corpora out there as there are types of investigations. Additionally, since language is so diverse and dynamic, "it would be hard to imagine a single corpus that could be used as a representative sample of all language" (Bowker and Pearson 2002: 11).

In the following paragraphs, thus, we would like to offer a list of different types of available corpora. The purpose of this list is not that of being exhaustive in presenting the number of corpora available to researchers[22]. Rather, it is a way of classifying corpora on the basis of their potential use. In doing so, the characteristics

[22] For a useful list of well-known and influential corpora and their description, see Xiao (2008).

highlighted of the different types of corpora can help us better define the corpus collected for our investigation.

Generally speaking, the types of corpora available to researchers can be introduced in a contrastive way, that is, they can be presented as the opposite of another type of corpora. For instance, general corpora can be contrasted with specialised corpora; multilingual corpora with monolingual, and so on.

The types of corpora introduced as general corpora refer to those that can be seen as "representative of a given language as a whole and can therefore be used to make general observations about that particular language" (Bowker and Pearson 2002: 12). Thus, general corpora are usually very large corpora, including a series of sub-corpora focusing on particular genres in a given language in a specific time-span. An example of a general corpus is the British National Corpus (BNC), which includes approximately 100 million words of spoken (nearly 10 million words) and written (nearly 88 million words) British English. The written component of the BNC is composed of a large variety of written genres, and they are distributed in the following way[23]:

MODE	TEXT CATEGORY AND DESCRIPTION	NUMBER OF WORDS
Written (87.8 million words)	Informative writing: 1) world affairs; 2) leisure; 3) arts; 4) commerce and finance; 5) belief and thought; 6) social science; 7) applied science; 8) natural and pure science.	71.4 million
	Imaginative writing: 9) fiction.	16.4 million

Table 1 Overview of the written component of the BNC.

As for the spoken component of the BNC, it is broadly split in two, the so-called spoken demographic, a collection of informal conversations collected across the UK

[23] The information provided in the following tables can be found in the section of the online user reference guide of the BNC dealing with the corpus design (available at http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html).

and across a range of social classes, ages, and genders (male and female); and the spoken context governed, a collection of more task-centred speech recorded in specific locations (e.g., an interview with a bank manager, family talks, etc.):

MODE	TEXT CATEGORY AND DESCRIPTION	NUMBER OF WORDS
Spoken (10.3 million words)	Spoken demographic	4.2 million
	Spoken context governed	6.1 million

Table 2 Overview of the spoken component of the BNC.

General corpora such as the BNC can be well suited in order to carry out certain types of linguistic analyses, while other are best performed by using other types of corpora. Thus, in the specific case of general corpora, they can be used first and foremost in order to study general lexicography, becoming an important tool in the creation of dictionaries of a given language. Additionally, general corpora can be used in order to offer a grammatical description of the language captured by that corpus. Finally, general corpora can also help researchers highlight linguistic change in the time span covered by the corpus. However, general corpora are not well-suited, for instance, to carry out domain-specific vocabulary investigation, since the peculiarities of a given genre may be lost in the large amount of data collected in the corpus. Thus, general corpora are also ill-suited in order to carry out genre analysis, since the specificities of given practices may be lost or they may be mixed with those of other communities of practice. Moreover, as Flowerdew (2004: 14) highlights:

> [G]eneral-purpose corpora have been compiled for their representativeness of the language as a whole and carefully balanced among the different types of text for reception and production to reflect their importance in the culture, which means that there will be a limited representation of some genres.

Thus, as Aston (2001) argues, since general corpora have been collected in order to offer generalisations on a given language as a whole, if we want to investigate a specific genre, a specialised corpus (see below) is preferable. And even though

general corpora do offer the possibility to search for given phenomena in a particular genre, Flowerdew (2004) highlights that sometimes it is difficult to access these sections of the general corpus since "the search fields have not been set up with this purpose in mind" (Flowerdew 2004: 15). But, more importantly, general corpora sometimes "comprise text segments of 2,000 words rather than full texts, which has implications for analysis" (Flowerdew 2004: 15), especially in the case of genre-based analyses.

Finally, and strictly related to our previous observation, we would like to underline that general corpora should not be taken into consideration also when we want to study languages for specific purposes, since the linguistic practices and routines of given professional communities may be lost in the 'noise' generated by the large amount of data collected.

The limitations underlined in the use of general corpora can be overcome by the use of a specialised corpus, which can be defined as "one that focuses on a particular aspect of a language", and "[i]t could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group" (Bowker and Pearson 2002: 12). As Hunston (2002) underlines, a specialised corpus "might be restricted to a time frame, consisting of texts from a particular century, or to a social setting, such as conversations taking place in a bookshop, or to a given topic, such as newspaper articles dealing with the European Union" (Hunston 2002: 14). Therefore, specialised corpora must be:

- representative of a specific domain or genre (e.g., newspapers materials);
- cover a well-defined time-span (e.g., from 2010 to 2015);
- representative of a particular variety of language (e.g., British newspapers).

In other words, a specialised corpus must be representative of a particular text type in a given time and a given space. The most comprehensive overview of what can be defined as a specialised corpus is offered by Flowerdew (2004: 21), whose summary of the most important characteristics of a specialised corpus is offered in the following table (Table 3):

PARAMETERS	DETAILS / EXAMPLES
Specific purpose for compilation:	To investigate particular grammatical, lexical, lexicogrammatical, discoursal or rhetorical features
Contextualization:	- Setting (e.g. lecture hall) - Participants (role of speaker / listener; writer / reader) - Communicative purpose (e.g. promote, instruct)
Size: - whole corpus - subcorpus or small-scale corpus	 - 1-5 million words - 20,000-250,000 words
Genre:	Promotional (grant proposals, sales letters)
Type of text / discourse:	Biology textbooks, casual conversations
Subject matter / topic:	Economics, the weather
Variety of English:	Learner, non-standard (e.g. Indian, Singaporean)

Table 3 Table taken from Flowerdew (2004: 21), where she summarises and exemplifies the most important characteristics of a specialised corpus.

While specialised corpora are usually built *ad hoc* by researchers who wish to investigate specific areas of language, among the most well-known specialised corpora, we would like to mention the Michigan Corpus of Academic Spoken English (MICASE)[24], which contains 152 transcripts (approximately 1.8 million

[24] From 2007, the corpus can be accessed online at `http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple`

words) of nearly 200 hours of university speech given at the University of Michigan[25]. Table 4 below shows the structure of the corpus[26]:

CRITERION	DISTRIBUTION
Speaker gender	Male (46%); Female (54%)
Age groups	17-23; 24-30; 31-50; 51+
Academic role	Faculty (49%); Students (44%); Other (7%)
Language status	Native speakers (88%); Non-native speakers (12%)
Academic division	Humanities and Arts (26%); Social Sciences and Education (25%); Biological and Health Sciences (19%); Physical Sciences and Engineering (21%); Other (9%)
Primary discourse mode	Monologue (33%); Panel (8%); Interactive (42%); Mixed (17%)
Speech event type	Advising (3.5%); Colloquia (8.9%); Discussion sections (4.4%); Dissertation defences (3.4%); Interviews (0.8%); Labs (4.4%); Large lectures (15.2%); Small lectures (18.9%); Meetings (4.1%); Office hours (7.1%); Seminars (8.9%); Study groups (7.7%); Student presentations (8.5%); Service encounters (1.5%); Tours (1.3%); Tutorials (1.6%).

Table 4 Composition of the MICASE corpus taken from Xiao (2008: 417).

[25] In the context of academic English, we must also acknowledge the existence of two British counterparts to the MICASE corpus, that is, the British Academic Spoken English (BASE) corpus and the British Academic Written English (BAWE) corpus. The BASE corpus comprises 160 lectures and 40 seminars (1,644,942 tokens in total) recorded in a variety of departments at the University of Warwick and at the University of Reading (in the case of the data collected at the University of Warwick, lectures and seminars were also video-recorded). In the case of the BAWE corpus, on the other hand, the corpus was created as part of a collaboration between the Universities of Warwick, Reading and Oxford Brookes. The BAWE corpus contains a selection of about 2,761 student assignments (both undergraduate and postgraduate) from four disciplines (Arts and Humanities, Life Sciences, Physical Sciences, and Social Sciences), each ranging in length from about 500 to about 5,000 words. Further information on the two corpora can be found at `http://www2.warwick.ac.uk/fac/soc/al/research/collections/`, while the two corpora can be accessed through the online platform for corpus analysis Sketch Engine (see Section 4.3).

[26] Further information on the MICASE corpus can be found at `http://micase.elicorpora.info/micase-manual-pdf`

As for general corpora, specialised corpora are well-suited for certain types of investigations, while they should be used with caution or should not be taken under consideration for others. As we have seen, since they capture the linguistic practices of specific communities, working with a specialised corpus may help researchers investigate the domain-specific vocabulary of those communities. In this case, general corpora can be used in this type of investigation in order to highlight the specificities in the vocabulary of the specialised corpus under scrutiny, since statistical measures, such as keyword analysis (see Section 5.2), can highlight quantitatively what lexical phenomena are peculiar in the specialised corpus.

This type of corpora is also well-suited when analysing a particular genre, since it records the linguistic practices that routinely take place in a well-defined community of practice. In this sense, specialised corpora can be also perfectly used in LSP investigations.

Additionally, researchers working on discourse and pragmatic analyses tend to investigate specialised corpora since the contextualization information encoded in the corpus can help better characterise pragmatically or discursively a particular context of practice. It must be also underlined that, while specialised corpora are not always small corpora (Flowerdew 2004), the amount of data is certainly less abundant than that of a general corpus and, thus, linguistic phenomena captured by a specialised corpus can be semi-automatically or manually tagged with pragmatic or discursive information.

Using a specialised corpus, however, does have its limitations and drawbacks, first and foremost, in the type of generalisations that can be made about the language represented in the corpus. Indeed, since a specialised corpus is simply a snapshot of a very specific context of practice, we cannot make general assumptions on the linguistic behaviour of speakers of that language as a whole: we can only make generalisations on the prototypicality of given phenomena in the particular context represented in that corpus. As Sinclair (2004: 189) underlines, "[s]mall is not beautiful; it is simply a limitation", meaning that the phenomena observed will be extremely limited by the very nature of the corpus and, thus, also the range of features that can be observed. Consequently, specialised corpora are not well-suited for the study of general lexicography. Indeed, while a specialised corpus can indicate

the lexical specificities of a given community of practice, it may not help us describe the general vocabulary of the language represented by that corpus. Thus, specialised corpora are also ill-suited if we want to offer a grammatical description of a given language or if we want to analyse language change. In these cases, the use of general corpora is preferable.

Going back to our description of the types of corpora distinguishable on the basis of their characteristics, another important distinction can be made between monolingual and multilingual corpora. The formers can be straightforwardly defined as those that contain data in a single language, while multilingual corpora are collections of texts in two or more languages.

In the specific case of multilingual corpora, however, we must also distinguish between comparable and parallel corpora. Comparable corpora are those that contain different languages or different varieties of a language (Hunston 2002) collected according to the same sampling techniques (McEnery *et al.* 2006), allowing us to look at the differences and/or equivalences in two (or more) languages or varieties of a language. The two or more languages may be contained in the very same corpus but encoded as two different subcorpora or they may be collected in two (or more) independent corpora that are later contrasted. However, we must underline that the languages contained in the subsections of the corpus or in the two (or more) corpora must be taken from the same genres. In this sense, the contrast between the two or more languages must be productive: we cannot collect a corpus of, say, English that is representative of a particular genre and, then, compare it with a corpus of Spanish that is representative of another particular genre. Thus, comparable corpora must "contain the same proportions of newspaper texts, novels, casual conversation, and so on" (Hunston 2002: 15).

Parallel corpora, on the other hand, contain original texts in language L1 and their translations into a set of languages L2, L3, L_n, etc. Thus, they allow us to look at what, for instance, the process of translation does when we translate from a language L1 to a language L2, highlighting the translation strategies used by or in a given professional context.

An example of comparable corpora is the International Corpus of English (ICE), designed for the synchronic study of world Englishes[27], while amongst the various examples of parallel corpora, the initiative undertaken by the OPUS project (Tiedemann 2012) is worth mentioning. Indeed, the aim of this project is to convert, align and tag free online data, thus, providing publicly available parallel corpora to research given communities of practice. One of the most influential parallel corpora developed by this initiative is the OpenSubtitles group of corpora, a series of monitor corpora (see below) of subtitles aligned with their translations in various languages[28]. This group of corpora may allow us to study how languages change in translation but, more importantly, it also might allow us, for instance, to have a look, so to speak, in the funsubbing community of practice, thus, becoming an important source of investigation of an ever-growing linguistic community.

Another important group of corpora is represented by those that can be defined as learner corpora, that is, "a collection of the writing or speech of learners acquiring a second language (L2)" (McEnery *et al.* 2006: 65)[29]. These types of corpora allow us to "identify in what respect learners differ from each other and from the language of native speakers, for which a comparable corpus of native-speaker texts is required" (Hunston 2002: 15). An example of learner corpus is represented by the International Corpus of Learner English (ICLE), a project started by Granger (2003)

[27] At the time of writing, the following parts (and annotations scheme performed on them) of the ICE family of corpora are available: Canada (ICE-CAN – 1 million words, wordclass tags); Jamaica (ICE-JA – 1 million words, wordclass tags); Hong Kong (ICE-HK – 1 million words, wordclass tags); East Africa (ICE-EA – Kenya and Tanzania, 1 million words, wordclass tags); India (ICE-IND – 1 million words, wordclass tags); Singapore (ICE-SIN – 1 million words, wordclass tags); Philippines (ICE-PHI – 1 million words, wordclass tags); USA (ICE-USA, only the written component – c.400,000 words, wordclass tags); Ireland (ICE-IRL – 1 million words, wordclass tags); SPICE-Ireland (SPICE-IRL – c.600,000 words of the spoken component of the ICE-IRL with prosodic and pragmatic annotation); Great Britain (ICE-GB – 1 million words, POS-tagged and parsed, distributed with ICECUP 3.1 retrieval software); New Zealand (ICE-NZ – 1 million words, wordclass tags); Sri Lanka (ICE-SL – written component only; wordclass tags and POS-tagged with CLAWS C7 tagset); Nigeria (ICE-NG – 1 million words, wordclass tags). More information and updates on the ICE family of corpora can be found at `http://ice-corpora.net/ice/`

[28] The subtitles collected in the group of corpora were taken from the online website `http://www.opensubtitles.org/`

[29] As McEnery *et al.* (2006) underline, the term 'learner corpora' is used in opposition to 'developmental corpora', that is, corpora that comprise "data produced by children acquiring their first language (L1)" (McEnery *et al.* 2006: 65) and, thus, the term 'learner' is used in order to highlight that in this type of corpora we will find examples of L2 data.

at the University of Louvain in 1990. The corpus, in its latest release[30], contains 3.7 million words of argumentative essays produced by higher-intermediate to advanced EFL learners from 16 mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, and Turkish).

Corpora can also be distinguished on the basis of whether they offer "a snapshot of language use during a limited time frame" (Bowker and Pearson 2002: 12) or, rather, offer the possibility to investigate language over a long period of time, thus, using the corpus, for instance, in order to "trace the development of aspects of a language over time" (Hunston 2002: 16). In the first case, we talk about synchronic corpora, while in the latter we are dealing with diachronic corpora. The ICE-GB, that is, the Great Britain component of the ICE family of corpora, represents an example of a synchronic corpus, since it was collected in a specific time span (1990-1993) and, thus, portrays perfectly the type of language spoken at that time and in that specific place[31]. One of the best-known diachronic corpora, on the other hand, is the Helsinki Corpus of English Texts, which contains approximately 1.5 million words of English texts, dating from the 8th to the 18th centuries (dividing them in three periods: Old, Middle, and Early Modern English)[32].

Finally, corpora can also be distinguished on the basis of whether they do not add further data to their collection or, rather, they weekly, monthly or yearly add new data to their collection. Thus, in the two cases, we respectively talk about closed and monitor corpora (Bowker and Pearson 2002). Most early corpora can be defined as closed corpora, since they covered a relatively short period of time and, then, they were closed and no other data were added to the corpus. The BNC, thus, is of this type. An example of monitor corpus, on the other hand, is represented by the Corpus

[30] The International Corpus of Learner Language (ICLE) was firstly released in 2002 and it contained approximately 2.5 million words of EFL learners from 11 mother tongue backgrounds, namely, Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. A second version of the corpus was released in 2009 and it can be ordered at `https://www.uclouvain.be/en-277586.html`, where further information on the corpus can also be found.

[31] It must be underlined that synchronic corpora can become diachronic corpora when the data they represent no longer capture synchronic aspects of our contemporary society.

[32] For more information on the Helsinki Corpus of English Texts, see Kytö (1996).

of Contemporary American English (COCA), created by Davies (2010) of Brigham Young University. At the time of writing, the COCA contains more than 450 million words (20 million words added each year from 1990-2012), equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. As a monitor corpus, the COCA is also updated regularly (the most recent texts are from the summer of 2012)[33].

As we have seen from this brief overview of the types of corpora distinguishable by their key features, the purpose and the type of approach selected in the collection of data shapes the type of corpus that we want to investigate and/or collect. Thus, as we have previously underlined when discussing the representativeness of a corpus, the research question is the key ingredient in the way the corpus is going to be shaped or chosen: the research question that one has in mind when building or choosing a corpus determines what the corpus is going to look like and the characteristics that it is going to have. And this, consequently, provides a first step towards a certain degree of representativeness of the corpus.

3.2 Introducing a corpus-based approach to Genre Analysis

As we have argued in the Introduction to this dissertation, the nature of our investigation is mainly corpus-based and, thanks to the use of these methodologies, we will be able to highlight in Chapter 5 the main linguistic patterns found in the corpus under investigation. In this sense, corpus linguistic methodologies will help us better define the genre of news tickers.

However, before undertaking this investigation, and since our investigation makes use of corpus linguistic methodologies in order to perform a genre analysis of news tickers, we must briefly introduce our view on genre analysis. As we will see in the following Sections, our approach to the investigation of the genre of news tickers is strictly linked to Swales' (1990, 2002) approach to genre analysis and the subsequent application of this framework by Bhatia (1993, 1996, 2002, 2004) to professional practices. In the context of corpus-based approaches to genre analysis,

[33] The Corpus of Contemporary American English (COCA) can be freely accessed at `http://corpus.byu.edu/coca/`, where further information on the corpus structure can also be found.

however, we must also briefly acknowledge the remarkable contribution made by Biber (1988) to this field of enquiry. Thus, we will start by introducing his multidimensional approach to genre analysis, whose insights and limitations will be both highlighted. Additionally, Tribble's (1999) keyword approach to genre analysis will also be introduced. His methodology has been preferred to the one outlined by Biber (1988) for reasons that will be explained. We will, then, move on to Bhatia's (1993, 2004) approach to genre analysis, which he has defined as Critical (Bhatia 2007, 2008, 2012), in the sense of highlighting in genres given regularities that respond to given purposes in specific professional environments. After this, we will briefly introduce Hoey's (2005) and O'Donnell *et al.*'s (2012) notion of textual colligation. In line with Biber's (1988) view that genres are the result of given text types, whose linguistic patterns are clearly identifiable, Hoey's (2005) approach to textual analysis, further developed in the context of news discourse by O'Donnell *et al.* (2012), will be introduced since it was used in this investigation to highlight a particular textual phenomenon occurring in our corpus data. Finally, the notion of news values (Bell 1991; Cotter 2010; Bednarek and Caple 2012a, 2012b; Caple and Bednarek 2015) and a quantitative approach to the automatic extraction from a corpus of these values encoded in news discourse (Bednarek and Caple 2014; Potts, Bednarek and Caple 2015) will be introduced.

3.2.1 Corpus approaches to genre analysis

In the following paragraphs of this section, two of the main corpus-based approaches to genre analysis will be briefly introduced. The first one, outlined in the context of Biber's (1988) multidimensional approach (MDA), takes its moves from the distinction between genres and text types. The former are defined as "text categorizations made on the basis of external criteria" (Biber 1988: 68), while the latter refer to the "groupings of texts that are similar with respect to their linguistic form" (Biber 1988: 70), irrespective of genre categories, that is, these linguistic patterns may cut across various genres, and respond to given purposes in the context of that particular communicative event that the genre they are found in is representative of. From this first distinction, Biber (1988) proceeds to his analysis, which can be summarised in the following steps:

1) collection or selection of corpora, "providing a standardized data base and ready access to a wide range of variation in communicative situations and purposes" (Biber 1988: 63), thus, representative of various communicative events and well-defined genres, distinguished on the basis of external criteria (speaker, intended audience, purpose, activity type, etc.). In other words, "conventional, culturally recognised grouping of texts based on properties other than lexical or grammatical (co-)occurrence features" (Lee 2001: 38). The collection of various instances of genres is linked to Biber (1993b) and Biber *et al.*'s (1998) view on corpus representativeness (see Section 3.1.3 of this contribution for a more detailed discussion on this issue);
2) using computer-based software in order to look at the frequency and distribution of linguistic features across the genres under investigation;
3) using factor analysis, in order to "determine co-occurrence relations among the linguistic features" (Biber 1988: 63);
4) using microscopic analyses or, in the words of Biber (1988), "functional analyses" in order to "interpret the functional parameters underlying the quantitatively identified co-occurrence patterns" (Biber 1988: 63).

The table below (Table 5), taken from Biber (1988: 64), concisely summarises the methodological framework developed by the author in approaching the analysis of 21 genres of spoken and written British English on the basis of sixty-seven linguistic

features, identified in 481 texts from the Lancaster/Oslo-Bergen (LOB) corpus[34] and the London-Lund (LLC) corpus[35].

Preliminary analysis - review of previous research to identify potentially important linguistic features; - collection of texts and conversion to machine-readable form; review previous research to insure that all important situational distinctions are included; - count occurrence of features in the texts (through the use of computer programs).
Step 1: Factor analysis - clustering of linguistic features into groups of features that co-occur with a high frequency in texts; - interpretation of the factors as textual dimensions through assessment of the communicative function(s) most widely shared by the features constituting each factor.
Step 2: Factor scores as operational representatives of the textual dimensions - for each factor, compute a factor score for each text; - compute an average factor score for the texts within each genre; - comparison of the average factor scores for each genre; - further interpretation of the textual dimensions in light of the relations among genres with respect to the factor scores.

Table 5 A summary of Biber's (1988) multidimensional approach to the investigation of genres.

As we can see from this summary of Biber's (1988) multidimensional approach, after the selection of the linguistic data to be analysed, he proceeds to the identification of

[34] The Lancaster/Oslo-Bergen (LOB) corpus is "a British match for the Brown corpus" (McEnery *et al.* 2006: 61) and it was created in order to "represent written British English used in 1961" (McEnery *et al.* 2006: 61). The LOB corpus is the result of the cooperation between Lancaster University, the University of Oslo, and the Norwegian Computing Centre for the Humanities at Bergen, and like the Brown corpus, it comprises 500 written texts, for a total number of one million words in all. Further information on the LOB corpus can be found online from the ICAME association (the International Computer Archive of Modern and Medieval English, an international organisation of linguists and computational scientists working with English corpora) at `http://clu.uni.no/icame/manuals/LOB/INDEX.HTM`

[35] The London/Lund Corpus (LLC) is a corpus of spontaneous spoken British English recorded from 1953-1987. As part of two projects, that is, the Survey of English Usage at the University College of London (UCL) and the Survey of Spoken English at Lund University, the full LLC is comprised of 100 texts, for a total number of 1 million words. More information on the corpus can be found online from the ICAME association at `http://clu.uni.no/icame/manuals/LONDLUND/INDEX.HTM` Additionally, part of the LLC (a selection of 400,000 words) is also available in the Diachronic Corpus of Present-Day Spoken English (DCPSE), developed by the Survey of English Usage at the UCL and containing also the spoken component of the ICE-GB, thus, allowing the diachronic investigation over a period spanning 25-30 years of spoken British English. The DCPSE can be ordered online at `http://www.ucl.ac.uk/english-usage/projects/dcpse/`

given linguistic patterns across the various genres represented in the data. Factor analysis is then applied to these linguistic features in order to statistically identify which features tend to co-occur, on the basis of the assumption that "frequently co-occurring linguistic features have at least one shared communicative function" (Biber 1988: 63). Based on this observation, it is then possible to identify "a unified dimension underlying each set of co-occurring linguistic features" (Biber 1988: 64). The purpose of these various steps taken in order to identify co-occurring linguistic patterns in given genres is that of moving from a micro-level of observation, represented by the text-types presented in a genre, to a macro-level, that is, to dimensions of abstraction that identify a genre as a whole. Indeed, as Biber (1988: 64) argues:

> [...] I am using factor analysis as it is commonly used in other social and behavioral sciences: to summarize the interrelationships among a large group of variables in a concise fashion; to build underlying dimensions (or constructs) that are conceptually clearer than the many linguistic measures considered individually.

Thus, by following the methodological approach so far outlined, Biber (1988) identifies the following five dimensions[36] (Table 6, taken from Biber 1993a: 335), on the basis of the linguistic features that comprise each factor. These features are displayed in a complementary distributional pattern, that is, based on the positive or negative weight they display (i.e., the extent to which linguistic features vary together), "when a text has several occurrences of the features with negative weights it will likely have few of the features with positive weights, and vice versa" (Biber 1988: 101).

[36] Biber (1988) and Conrad and Biber (2001) also include other two dimensions, Dimension 6 (On-line Informational Elaboration Marking Stance) and Dimension 7 (Academic Hedging). The interpretation of these further dimensions, however, as Biber (1988: 114) highlights, is "extremely tentative", since they display few features with important loading (Conrad and Biber 2001), thus, "requiring more research into the use of the features" (Conrad and Biber 2001: 39).

Dimension 1 **Informational vs. Involved**	
Positive features nouns word length prepositional phrases type/token ratio attributive adjs.	Negative features private verbs *that* deletions contractions present tense verbs 2nd person pronouns *do* as pro-verb analytic negation demonstrative pronouns general emphatics first person pronouns pronoun *it* *be* as main verb causative subordination discourse particles indefinite pronouns general hedges amplifiers sentence relatives WH questions possibility modals non-phrasal coordination WH clauses final prepositions
Dimension 2 **Narrative vs. Non-Narrative**	
Positive features past tense verbs third person pronouns perfect aspect verbs public verbs synthetic negation present participial clauses	Negative features present tense verbs attributive adjs.
Dimension 3 **Elaborated vs. Situated Reference**	
Positive features WH relative clauses on object positions pied piping constructions WH relative clauses on subject position phrasal coordination nominalizations	Negative features time adverbials place adverbials other adverbs
Dimension 4 **Overt Expression of Persuasion**	
Positive features infinitives prediction modals suasive verbs conditional subordination necessity modals split auxiliaries possibility modals	Negative features [No complementary features]

Dimension 5 Abstract vs. Non-Abstract Style	
Positive features conjuncts agentless passives past participial clauses BY-passives past participial WHIZ deletions other adverbial subordinators	Negative features [No complementary features]

Table 6 A summary of the dimensions and their features identified by Biber (1988).

As we have previously highlighted, Biber's approach moves from a micro- to a macro-level of investigation in analysing the features that characterise specific genres. The result of this movement from a microscopic level to that of the genre as a whole is represented by the dimensions that a genre can be representative of. Dimensions, in this sense, should be seen as "continuums of variation rather than discrete poles" (Biber 1988: 9). In other words, the aim of a multi-dimensional analysis is that of describing a text as more or less informational, narrative, abstract, etc., and not as either formal or informal, for instance. A genre, in this view, is thus a degree of variation along the dimensions previously highlighted. And the study of how these dimensions may vary and become accentuated, reduced, or introduced in certain genres may reveal the variation that a genre may undergo during its evolution.

We must, however, acknowledge some of the limitations of this approach. Firstly, Biber's (1988) methodology starts with a preliminary review of previous researches to identify potentially important linguistic features, which are then tagged in the corpus and processed in the factor analysis. This preliminary investigation may, thus, exclude all those linguistic features which the literature on a specific genre has not highlighted[37]. Additionally, if a genre has not been taken under consideration by the linguistic community, it will be difficult to make a preliminary

[37] As Culpeper (in Hoover, Culpeper and O'Halloran 2014) highlights in the comparison between Biber's approach and Tribble's keyword approach to genre analysis (see the following paragraphs), "multidimensional analysis involves a priori decisions about what to count" (Hoover, Culpeper and O'Halloran 2014: 32), and thus, some linguistic phenomena may be overlooked due to this preliminary decision to focus only on a given set of linguistic patterns.

observation of the linguistic features characterising that genre. This is particularly the case of the genre under investigation. In order to solve this problem, we could apply Biber's (1988) highlighted factors to our investigation but, again, this could lead to an involuntary bending of the conventions of specific genres to another genre that may not share the same linguistic purposes. Thus, given traits may be insulated in the genre under investigation as an echo of other genres, becoming something that was not meant to become or highlighting characteristics and purposes that are simply not part of its galaxy: a memory of an imploding star from another galaxy, whose light should not be confused with the actual existence and presence of that star in another, different galaxy.

Another important limitation is highlighted by Stubbs (1996), who argues that while Biber's multidimensional approach "provides empirical comparisons of different genres, and can therefore provide a powerful interpretative background for the analysis of individual texts", since it can be used to "document the 'unconscious background' of textual variability against which individual texts are interpreted" (Stubbs 1996: 34), it is however carried out "across representative samples of genres and sub-genres, with no analysis of the discourse structure of individual instances of the genres" (Stubbs 1996: 34). Indeed, the author argues that "[t]he most powerful interpretation emerges if comparisons of texts across corpora are combined with the analysis of the organization of individual texts" (Stubbs 1996: 34). In other words, Biber's approach, in Stubbs' (1996) view, inevitably, if not inexorably, moves towards various degree of abstractions, thus, slowly doing without the original textual organisation of the genres under investigation. While a certain degree of abstraction is necessary in analysing genres, researchers should not completely forget where specific genres originally come from and Biber's approach may face this risk. His methodology, however, still is a powerful tool in analysing a representative collection of genres and reveal peculiarities in them.

A different method used in order to reveal not only the 'aboutness' (Phillips 1989; Scott and Tribble 2006; Scott 2015) of a given genre, but also its stylistic information, thus, becoming a powerful tool in identifying its generic status, is represented by Tribble's (1999) and Scott and Tribble's (2006) keyword approach to genre analysis. As McEnery *et al.* (2006) underline, Biber's multidimensional

approach "involves very sophisticated statistical analysis" and, thus, it may be seen as "extremely time-consuming" (McEnery *et al.* 2006: 308). Therefore, Tribble's keyword approach to genre analysis can achieve "an approximate effect of Biber's MF/MD approach" (McEnery *et al.* 2006: 308) in a more user-friendly way. Indeed, as Tribble's (1999) maintains, the methodology developed by Biber (1988) can be regarded as "lengthy and complex, and requires carefully marked up corpora" (Tribble 1999: 173). Since Tribble is particularly interested in the pedagogical issues linked to genre analysis, Biber's approach does seem too complicated to be applied, particularly, when using un-tagged texts. Thus, a function so easily found in every modern-day concordancers such as the keyword list can help researchers identify "the salient features which are functionally related to that genre" (McEnery *et al.* 2006: 308). Thus, Tribble (1999), in his dissertation, proceeds to the investigation of the genre of project proposals by analysing 14 texts belonging to this genre and using different reference corpora (in particular, reference word lists taken from the written component of the BNC and from a 95-million-word data set of *The Guardian*; Tribble 1999). The different reference corpora were used in order to test if, by choosing one reference corpus over the other, the keyword list would have shown some degree of variation, something that was proven wrong by the data, since "the top five words in each are identical, and are in the same frequency order" (Tribble 1999: 170).

The methodology developed by Tribble (1999) and further extended in Scott and Tribble (2006), led to the analytical framework summarised in Table 7 (taken from Tribble 2002: 133). As we can see, it slightly resembles Biber's (1988) distinction between two level of analysis of a given genre, that is, the genre as defined on the basis of its external criteria, and as groupings of texts that are similar on the basis of their linguistic form.

CONTEXTUAL Analysis	
1. name	What is the name of the genre of which this text is an exemplar?
2. social context	In what social setting is this kind of text typically produced? What constraints and obligations does this setting impose on writers and readers?
3. communicative purpose	What is the communicative purpose of this text?
4. roles	What roles may be required of writers and readers in this genre?
5. cultural values	What shared cultural values may be required of writers and readers in this genre?
6. text context	What knowledge of other texts may be required of writers and readers in this genre?
7. formal text features	What shared knowledge of formal text features (conventions) is required to write effectively in this genre?
LINGUISTIC analysis	
8. lexico-grammatical features	What lexico-grammatical features of the text are statistically prominent and stylistically salient?
9. text relations/textual patterning	Can textual patterns be identified in the text? What is the reason for such textual patterning?
10. text structure	How is the text organised as a series of units of meaning? What is the reason for this organisation?

Table 7 Tribble's (2002: 133) analytical framework in approaching genres.

Inscribed in the developing tradition of genre analysis (Swales 1990; Bhatia 1993, 2004), Tribble's (2002) analytical framework can be seen as a powerful tool of analysis. As the author maintains (Tribble 2002: 133):

> [...] such an approach provides a useful basis for contextual and linguistic awareness raising during an EAP course, and offers a coherent basis for the development of curricula for writing instruction and the evaluation of written production.

Additionally, as Xiao and McEnery (2005) demonstrate, from a methodological point of view, both Biber's MDA and Tribble's keyword framework of analysis seem to yield the same results when applied to genre analysis. Indeed, the only difference highlighted by Xiao and McEnery (2005) is that, while Biber's approach "requires

considerable expertise in data extraction and statistical analysis", Tribble's keyword analysis "provides a less demanding approach to genre analysis" (Xiao and McEnery 2005: 77). However, as the authors argue (Xiao and McEnery 2005), Tribble's keyword approach to genre analysis must not be seen as a substitute for Biber's MDA approach. Indeed, as Xiao and McEnery (2005) state, the keyword approach to genre analysis "provides a less comprehensive contrast of genres and may not work for more fine-grained types of genre analysis" (Xiao and McEnery 2005: 77). In other words, Tribble's approach can be useful in those cases where the contextual features of the genres are well-defined and can, thus, help researchers offer given explanations to the lexicogrammatical features identified by the corpus linguistic methodologies. However, when "the analyst is unfamiliar with the institutionalized practices of the genre, then it is a more difficult task to 'read off' and infer the discursive practice of the genre from the concordance data" (Flowerdew 2008: 125). Therefore, in these cases, Biber's approach is better suited, since it provides given baselines to approach different types of genres and, once decided the features that the researchers want to investigate in those genres, see where they are more statistically relevant and, thus, highlight given dimensions. Tribble's approach, in this sense, should not be taken under consideration since, as Culpeper (in Hoover, Culpeper and O'Halloran 2014) argues, keyword analyses may reveal the lexicogrammatical peculiarities of a given text, thus, raising "issues of comparability" among genres (Hoover, Culpeper and O'Halloran 2014: 32).

However, we must also acknowledge that, in the case of well-defined genres, whose contextual characteristics are quite clear to researchers, a keyword analysis is preferable to Biber's approach since, as Culpeper (in Hoover, Culpeper and O'Halloran 2014: 9–34) states, it does not involve decisions about what to take under consideration and what to discard. The author offers as an example of this limitation the use of interjections in discourse, which were not taken under consideration as linguistic items by Biber (1988) and, therefore, did not count as discourse markers in his multidimensional approach. On the other hand, Xiao and McEnery (2005), by using a keyword approach, discovered that interjections played an important role in "both conversation and monologic speech" (Culpeper in Hoover, Culpeper and O'Halloran 2014: 33), thus, proving that "while interjections are not included as a

relevant linguistic feature in MDA, they are an important feature in a keyword analysis" (Xiao and McEnery 2005: 74).

The brief overview of these two methodological approaches to corpus-based genre analysis offered in this Section has shown two different ways researchers can approach genre analysis by using empirical data provided by a corpus collection. We have also tried to highlight both the advantages and limitations of Biber's (1988) multidimensional approach and Tribble's (1999) keyword approach to genre analysis. Thus, in conclusion, by reviewing these two approaches and acknowledging their importance, we however have chosen to use Tribble's approach in the context of this dissertation, since it is more in line with the aim of our study and, more importantly, in line with the type of corpus collected. Indeed, being an extremely specialised corpus, collected in a well-defined context of practice, Tribble's (2002) analytical framework will be used in order to define the lexicogrammatical status of the genre of news tickers. Tribble's approach has also been preferred to Biber's since it is more consistent with the genre analysis tradition this contribution aligns with (Swales 1990; Bhatia 1993, 2004), which will be briefly outlined in the following Section.

3.2.2 Critical Genre Analysis (CGA)

As we have underlined at the end of the previous Section, the corpus-based genre analysis carried out in this contribution is imbued in the genre analysis tradition as developed by Swales (1990) and further extended by Bhatia (1993, 2004) to the genres found in professional practices.

This tradition takes its moves from Miller's (1994) outstanding contribution to genre theory, where she argues that a theoretically sound definition of genre must be centred not only "on the substance or the form of discourse" but also "on the action it is used to accomplish" (Miller 1994: 20). This implication leads to a new definition of genre, seen as "a rhetorical means for mediating private intentions and social exigence", motivated by "connecting the private with the public, the singular with the recurrent" (Miller 1994: 31). In this view, genres embody "aspect of cultural rationality" and can, thus, serve as "an index to cultural patterns and as tools for exploring the achievements of particular speakers and writers" (Miller 1994: 32).

But, more importantly, genres can serve as "keys to understanding how to participate in the actions of a community" (Miller 1994: 32). In other words, exploring genres can help researchers highlight the linguistic practices taking place in a given discourse community.

Swales (1990) further defines genres on the basis of the notion of discourse community, defined as "sociorhetorical networks that form in order to work towards sets of common goals" (Swales 1990: 9). In this view, genres are thus seen as (Swales 1990: 58):

> [...] a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constraints choice of content and style.

As we can see from the quotation above, a genre is firstly defined as a communicative event, that is, an event where "language plays both a significant and an indispensable role" (Swales 1990: 45). This event, however, comprises "not only the discourse itself and its participants, but also the role of that discourse and the environment of its production and reception, including its historical and cultural associations" (Swales 1990: 46). In this sense, a communicative event can be seen as a quite broad category and, thus, encompassing all situations involving language in use.

Thus, a key element in specifying a genre is represented by the communicative purpose that it serves, since "genres are communicative vehicles for the achievement of goals" (Swales 1990: 46). These aims may be overtly acknowledged by the discourse community, or they may be implicitly achieved. Nonetheless, they are embedded in the discourse thanks to given cues, referred in the quotation above as the rationale for the genre, which "establishes constraints on allowable contributions in terms of their content, positioning and form" (Swales 1990: 52). In other words, the recognition of given purposes in the discourse community provides the rationale which, in turn, "gives rise to constraining conventions" (Swales 1990: 53).

While Swales (1990) acknowledges that "conventions [...] are constantly evolving and indeed can be directly challenged", they nonetheless "continue to exert influence" (Swales 1990: 53). The textual 'places' where given purposes are realised and given conventions used in order to achieve these purposes are identified by Swales as "moves" (Swales 1990: 140–148). Thus, his framework, as we have seen in Section 3.1.2, can be defined as a top-down approach to genres since, by identifying given purposes in the genre under investigation, researchers then proceed to the investigation of their realisations in given textual conventions.

In this way, as Bhatia (1993) states, this framework of analysis of genres seems to "encourage prescription rather than creativity in application" (Bhatia 1993: 40). In other words, while acknowledging, on the one hand, Swales's outstanding contribution to genre analysis, on the other, his framework of analysis, according to Bhatia (1993), seems to pay too much attention on strict conventions realised in given genres or in given textual places of a genre. Thus, as Hart (1986) points out, this approach to genre analysis seems to suggest a methodology that is pattern imposing rather than pattern seeking.

On the basis of genres viewed as places of instability and conflict (Bhatia 2000), firstly highlighted by Berkenkotter and Huckin (1995), Bhatia (1993, 2004) further develops Swales' approach to genre analysis, by firstly defining a genre as (Bhatia 1993: 13):

> [...] a recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs. Most often it is highly structured and conventionalized with constraints on allowable contributions in terms of their intent, positioning, form and functional value. These constraints, however, are often exploited by the expert members of the discourse community to achieve private intentions within the framework of socially recognized purpose(s).

As we can read at the very beginning of this quotation, Bhatia's focus on the importance of the communicative purpose(s) in defining a genre echoes Swales' (1990) approach. Indeed, the communicative purpose of a genre again plays a key

role since, in the words of Bhatia (1993), the nature and construction of a genre is "primarily characterized by the communicative purpose(s) that it is intended to fulfil" (Bhatia 1993: 13). The set of communicative purpose(s) in a genre shapes it and gives it its internal structure. But, more importantly, and this goes back to the instability between standardisation and innovation in genres, Bhatia (1993) highlights that "[a]ny major change in the communicative purpose(s) is likely to give us a different genre", while "minor changes or modifications help us distinguish sub-genres" (Bhatia 1993: 13).

Therefore, in this view, while genres are identified on the basis of conventionalised features, they are not static, and "they continually develop and change" (Bhatia 2004: 29). Indeed, while referring to genres as "language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution", giving rise to "stable structural forms by imposing constrains on the use of lexico-grammatical as well as discoursal resources" (Bhatia 2004: 27), Bhatia (2004) also acknowledges the fact that "people, especially those who have the expertise and the power, do not hesitate to exploit some of these conventions to create new forms" (Bhatia 2004: 29). This, again, refers back to Berkenkotter and Huckin's (1995) view of genres, considered as (Berkenkotter and Huckin 1995: 3):

> [...] inherently dynamic rhetorical structures that can be manipulated according to the conditions of use, and [...] genre knowledge is therefore best conceptualised as a form of situated cognition embedded in disciplinary activities.

Hence, genres' integrity is always unstable, since members of a particular discourse community, in communicating with a larger set of people outside the community itself, may "appropriate rhetorical resources and other generic conventions across genres" (Bhatia 2002: 11), giving rise to hybrid, mixed, and embedded forms.

This propensity to manipulation of genre conventions is strictly linked to the fact that "the members of the professional or academic community have greater knowledge of the conventional purpose(s), construction and use of specific genres than those who are non-specialists". Thus, as Bhatia (2004) maintains, while the

general propensity is to see genres "in pure forms" (Bhatia 2004: 34), in the real world, genres "can be exploited to respond to novel rhetorical contexts, and thus have propensity for innovation and further development" (Bhatia 2004: 34). Novel rhetorical contexts can be the very result of new socially recognised communicative purposes and, thus, in this view, genres also change if changes in the social context where they are created occur.

In the previous paragraphs, we have briefly tried to highlight the major aspects of Bhatia's contribution to genre analysis. But the author himself (Bhatia 2004: 22, originally introduced in Bhatia 2002: 16) offers a quite interesting schematisation of the proposed theoretical applied genre analytical model (see Figure 6), which entails three perspectives on discourse: the textual perspective, the socio-cognitive perspective, and the socio-critical perspective.

Figure 6 The basis of Bhatia's theoretical applied genre analytical model (Bhatia 2004: 22, originally introduced in Bhatia 2002: 16).

Bhatia (2004) adopts a general definition of 'discourse', defining it as "any instance of the use of [...] language to communicate meaning in a particular context, irrespective of any particular framework for analysis" (Bhatia 2004: 22). Discourse,

however, may be analysed from a text, genre (or professional practice), and social practice point of view.

Thus, discourse as text refers to "the analysis of language use that is confined to the surface level properties of discourse, which include formal, as well as functional aspects of discourse" (Bhatia 2002: 17). In other words, discourse as text entails the analysis, for instance, of phonological, lexicogrammatical, semantic, organizational or information structures aspects of discourse. This level of investigation, as Bhatia (2004) underlines, does not necessarily take into consideration the text-external context where the discourse has been produced, thus, "merely taking into account what is known as co-text" (Bhatia 2004: 23).

Discourse as genre, on the other hand, further extends the analysis of discourse beyond the textual level and, thus, it incorporates the context of production, thus, taking into account "not only the way text is constructed, but also [...] the way it is often interpreted, used and exploited in specific institutional or more narrowly professional contexts to achieve specific disciplinary goals" (Bhatia 2004: 23). At this level of investigation, researchers are interested not only in linguistic, but also in socio-cognitive and ethnographic questions related to the discourse production. Additionally, a certain degree of awareness and understanding of the "shared practices of professional and discourse communities and their choices of genres in order to perform their everyday tasks" (Bhatia 2004: 23) is needed in order to understand the way discourse is shaped. Thus, Bhatia (2004) argues that, at this level, genres generally operate in what he refers to as tactical space, which allows "established members of discourse communities to exploit generic resources to respond to recurring or often novel situational contexts" (Bhatia 2004: 23).

Finally, discourse analysed as social practice takes a little further the analysis of contextual features of discourse, thus, taking into account, for instance, "the changing identities of the participants, the social structures or professional relationships the genres are likely to maintain or change, and the advantages or disadvantages such genres are likely to bring to a particular set of readers" (Bhatia 2004: 24).

Although they can be carried out separately, the levels of analysis thus underlined by Bhatia (2004) are developed by the author as "complementary to each

other" (Bhatia 2002: 18), thus, creating a framework of genre analysis that looks at discourse as "genre within a socio-cognitive space" (Bhatia 2004: 25), where the analysis of their lexicogrammatical features and, in particular, the textualisation of these features, on the one hand, and the analysis of features linked to professional and social practices, on the other, are combined together in order to explain the realisation of specific genres. Additionally, when approaching the analysis of genres, Bhatia (2007) argues that researchers can move along the two continuums: from text to context, and vice versa; and, consequently, from discursive to professional practice, and vice versa.

In particular, when approaching the analysis of professional practices, Bhatia (2007, 2008, 2012) underlines how this framework of analysis can help researchers highlight the conflict underlying genres found in professional contexts between genre integrity and genre bending, that is, "the appropriation of generic resources to achieve 'private intentions' within the context of 'socially accepted generic norms'" (Bhatia 2008: 175). The clash between these two divergent forces is mediated, according to the author, through forms of interdiscursivity, which play a major role in genre construction, appropriation, and interpretation.

Drawing on the work of Fairclough (1992, 1995a), Bhatia (2004, 2007, 2008) introduces the notion of interdiscursivity "in order to develop a comprehensive and evidence-based awareness of the motives and intentions of [...] disciplinary and professional practices" (Bhatia 2007: 393), and he defines it as "innovative attempts to create hybrid or relatively novel constructs by appropriating or exploiting established conventions or resources associated with other genres" (Bhatia 2008: 175). According to Bhatia (2012), processes of interdiscursivity are realised in genres as we can see in Figure 7 (Bhatia 2012: 25):

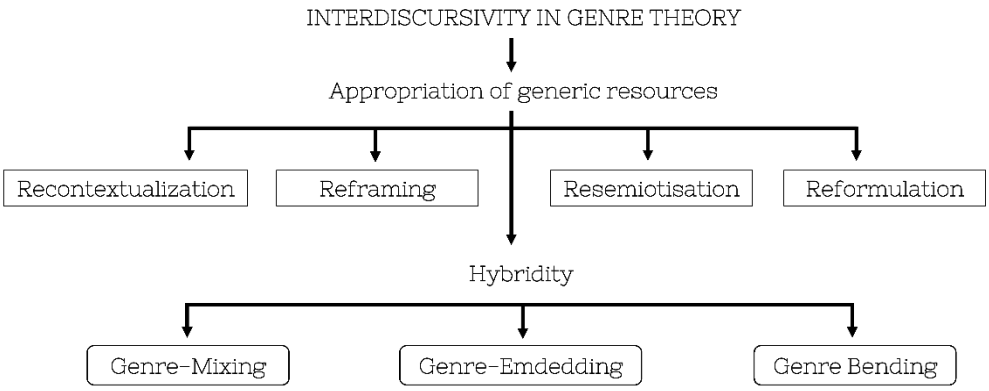


Figure 7 Interdiscursivity in Genre Theory (Bhatia 2012: 25).

Interdiscursivity, thus, can be traced in a variety of discursive processes, some of which, as we have previously highlighted, include "mixing, embedding and bending of generic norms in professional practice" (Bhatia 2008: 175). In Bhatia (2007), the author offers three examples focusing on the interdiscursive nature of genres found in professional contexts:

a) the colonization of accounting discourse within the corporate disclosure genre of annual reports, whose purpose is "to report accurately and factually on the basis of financial evidence of the past corporate performance" (Bhatia 2007: 395), by public relations discourse, whose purpose is to "promote a positive image of the company to its shareholders and other stakeholders in order to sustain their confidence in the future corporate performance" (Bhatia 2007: 395);
b) the colonization of arbitration by litigation practices, "threatening the integrity of arbitration practice to resolve disputes outside of the courts, and thus going contrary to the spirit of arbitration as a non-legal practice" (Bhatia 2007: 397);
c) the appropriation by philanthropic fundraising discourses of the corporate culture of marketing and advising, thus, challenging their integrity.

In all these cases, "the use of specific lexico-syntactic as well as socio-pragmatic resources, are cleverly exploited to 'bend' the norms and conventions" (Bhatia 2007: 395) of the genres under investigation. In this context, Bhatia (2007) calls for "a

critical study of discursive activities of professional cultures" (Bhatia 2007: 399), thus, introducing his Critical Genre Analysis (CGA) approach.

Critical Genre Analysis is the result of the combination of two fields of enquiry, that is, Genre Analysis and Critical Discourse Analysis (CDA). As Motta-Roth (cited in Bonini 2010) underlines, these two fields have slowly if not increasingly become closer in recent years (Motta-Roth cited in Bonini 2010: 487):

> In the case of Genre Analysis, Bakhtin and Fairclough, initially absent, appear in the more recent books of Swales (1990, 2004) and Bhatia (1993, 2004). [...] [And] if the sociological or sociohistorical thought represented by Fairclough has been settling in the discussions on genre, the use of the word "genre" by Fairclough has also become increasingly frequent. Used as a tool for theorizing and explanation, "genre" appears in an increasing number of pages in the table of contents of Fairclough's works over the years.

As Bhatia (2012) states, CGA is "an attempt to extend genre theory beyond the analyses of semiotic resources used in professional genres to understand and clarify professional practices or actions in typical academic and professional contexts" (Bhatia 2012: 22). As we have seen, this view was already present in his earlier works (Bhatia 1993, 1996, 2000, 2002, 2004), but in his later publications (Bhatia 2007, 2008, 2012) genre analysis further develops, setting out the following aim (Bhatia 2012: 23):

> [...] 'demystifying' professional practice [...] [by focusing] as much on generic artefacts, as on professional practices, as much on what is explicitly or implicitly said in genres, as on what is not said, as much on socially recognized communicative purposes, as on "private intentions" [...] that professional writers tend to express in order to understand professional practices or actions of the members of corporations, institutions and professional organizations.

The key element in this view is that professional practices are not assumed but, as it was initially clear in Bhatia (2000), they are a site of struggles and, thus, they are

always negotiated through and outside language. In this way, private intentions are incorporated "within the concepts of professionally shared values, genre conventions, and professional cultures" (Bhatia 2012: 23). Professional practices give shape to actions in specific professional contexts and, since language plays a key role in establishing specific conventions, "CGA makes a commitment [...] to explain, clarify, and "demystify" professional practice" (Bhatia 2012: 23–24), highlighting the lexicogrammatical cues present in genres, following Bhatia's theoretical framework previously introduced.

We must, however, underline that, while the purpose of CGA is to 'demystify' given professional practices by making apparent the purposes that lie behind given conventions, it differs from CDA with respect to the fact that it is not "an initiative to change professional practices of individual disciplinary, institutional, and corporate communities, but to understand how professional writers use language to achieve the objectives of their professions" (Bhatia 2012: 24), something that can be seen as quite the opposite to Fairclough's (1989, 1992, 1995a) approach to discourse analysis. Indeed, according to the author (Fairclough 1989, 1992), CDA's aim is to make apparent those ideologies that lie behind given discursive practices. More importantly, a critical approach to discourse analysis implies intervention (Fairclough 1992). Indeed, not only does CDA show "how discourse is shaped by relations of power and ideologies, and the constructive effects discourse has upon social identities, social relations and systems of knowledge and belief" (Fairclough 1992: 12), but also as a form of intervention in order to make such ideologies apparent and, thus, making people aware of them. CGA, on the other hand, loses this 'political' stance towards genre analysis, and it simply provides a fuller understanding of professional practices by uncovering the private intentions that lie behind given linguistic choices.

In conclusion, by adopting a CGA corpus-based approach to the study of the genre of news tickers, our aim is to explain, thanks to a bottom-up analysis, that is, by firstly retrieving given linguistic patterns in the genre under investigation, how these are indicative of specific private intentions in the context of the professional environment of the BBC. These private intentions are representative of a change in the professional practice of the news network company. Thus, the combination of

corpus linguistic methodologies and CGA approach to genre analysis have helped us 'demystify' these private intentions.

In the next section, we will briefly introduce the notion of textual colligation. As we have seen, Bhatia's approach to genre analysis investigates the socio-cognitive space of discourse on the basis of lexicogrammatical features and, in particular, by exploring how these features are 'textualise' in the realisation of given genres. In line with this approach, the concept of textual colligation can help us achieve this aim in the analysis of news tickers.

3.3 Lexical priming and textual colligation

The theory of Lexical Priming, as introduced by Hoey (2005), combines both corpus approaches to the analysis of language with the psycholinguistic notion of semantic priming, according to which "a 'priming' word may provoke a particular 'target' word" (Hoey 2005: 8). While the focus of Psycholinguistics is on the relationship between the prime and the target, Hoey (2005) argues that priming, in his theoretical framework, is particularly concerned with the "property of the word and what is primed to occur [...] seen as shedding light upon the priming item rather than the other way round" (Hoey 2005: 8). In other words, Hoey's concern is with the collocational use of given words, that is, how specific lexical items are mentally associated with other specific words. This is based on his belief that "[a]s a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context" (Hoey 2005: 8).

As we can see, the focus of Hoey's investigation is on lexical aspects of language, thus, prioritising them over grammatical features of discourse. In this view, priming is primarily realised in terms of phraseology, not grammar. This approach to language is linked to Hoey's (2005) argument that "[t]here is not [...] a single grammar to the language (indeed there is not a single language), but a multiplicity of overlapping grammars that are the product of the attempt to generalise out of primed collocations" (Hoey 2005: 47). Once a word has been acquired to be primed with another, this word sequence is in itself primed, giving rise to what Hoey (2005)

refers to as the process of nesting, whereby "the product of a priming becomes in itself primed in ways that do not apply to the individual words making up the combination" (Hoey 2005: 8). Such a view on language, thus, implies that each individual element reproduces the structure and the structure shapes the individual element.

The process of priming, however, can be seen as static and monolithic from the description so far offered. However, Hoey (2005) underlines that it is not a permanent feature of a word or word sequence. On the contrary, as the author maintains (Hoey 2005: 9):

> Every time we use a word, and every time we encounter it anew, the experience either reinforces the priming by confirming an existing association between the word and its co-texts and contexts, or it weakens the priming, if the encounter introduces the word in an unfamiliar context or co-text or if we have chosen in our own use of it to override its current priming.

Thus, lexical priming is always linked to confirmations of previously primed systems or to innovations to the mental representation of given words or word sequences, which may entail temporary or permanent language change. Thus, in this view, "[e]very word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word" (Hoey 2005: 13). From this observation, the author makes ten fundamental claims linked to the priming hypothesis (Hoey 2005: 13):

1) every word is primed to occur with particular other words; these are its collocates;
2) every word is primed to occur with particular semantic sets; these are its semantic associations;
3) every word is primed to occur in association with particular pragmatic functions; these are its pragmatic associations;
4) every word is primed to occur in (or avoid) certain grammatical positions, and to occur in (or avoid) certain grammatical functions; these are its colligations;

5) co-hyponyms and synonyms differ with respect to their collocations, semantic associations and colligations;
6) when a word is polysemous, the collocations, semantic associations and colligations of one sense of the word differ from those of its other senses;
7) every word is primed for use in one or more grammatical roles; these are its grammatical categories;
8) every word is primed to participate in, or avoid, particular types of cohesive relation in a discourse; these are its textual collocations;
9) every word is primed to occur in particular semantic relations in the discourse; these are its textual semantic associations;
10) every word is primed to occur in, or avoid, certain positions within the discourse; these are its textual colligations.

The author, then, proceeds to test these hypotheses by using a corpus of news stories taken from *The Guardian*. For instance, in the case of the first three claims of the priming hypothesis, Hoey (2005) investigates how the word 'consequence' is primed in the corpus under investigation in order to illustrate the collocates that are associated with the word and the semantic and pragmatic implications that they seem to imply. From this corpus-based investigation, Hoey (2005: 24-37) highlights four major types of semantic and pragmatic association regarding the word 'consequence' on the basis of the analysis of its collocates, that is, logical association (suggested by the collocates 'logical', 'ineluctable', 'direct'), negative evaluation (on the basis of the collocates 'awful', 'fire', 'appalling'), seriousness of the 'consequence' (linked to the collocates 'important', 'serious', 'significant'), and unexpectedness of the 'consequence' (realised by the collocates 'unforeseen', 'curious', 'surprising').

However, from the list of priming hypotheses previously highlighted, and further investigated and confirmed by Hoey (2005) during his discussion on priming, we would like to draw our attention on Claim 10, since this priming hypothesis will be part of our analysis of news tickers.

Introduced in order to explain the ambiguity emerging from the conflict between local priming and textual patterns (i.e., priming is realised not only at the sentence-internal level, but it must also take under consideration the textual dimensions where given items occur), this hypothesis is based on the assumption that

"[w]ords (or nested combinations) may be primed to occur (or to avoid occurring) at the beginning or end of independently recognised discourse units, e.g. the sentence, the paragraph, the speech turn" (Hoey 2005: 115), thus, introducing the notion of textual colligation. The author, then, offers some examples of textual colligations, starting with the 'consequence' example previously highlighted. Indeed, by looking at the corpus data under investigation, Hoey (2005) proves that the word 'consequence' is "primed to avoid paragraph-initial and text-initial position" (Hoey 2005: 130), while the plural 'consequences' is "positively primed to be paragraph-initial, though it shares the aversion of *consequence* for being text-initial" (Hoey 2005: 130, emphasis in the original).

The notion of textual colligation and the insightful observation it may unveil about the preference of given items to occur in particular 'places' of the text are further investigated in O'Donnell *et al.* (2012), where the methodology to identify how words are primed to occupy text-initial, paragraph-initial and sentence-initial positions is outlined and applied in the analysis of a corpus of newspaper articles (52 million words) taken from the 'Home News' section of *The Guardian* from 1998 to 2004. The methodology thus developed in O'Donnell *et al.* (2012) achieves the goal of "providing a systematic and comprehensive account of the associations that hold between items [...] and particular sentence positions within newspaper text" (O'Donnell *et al.* 2012: 81).

And if, as Hoey (2005) argues, priming is constrained by domain and/or genre, since it is linked to "the way language is acquired and used in specific situations [...] because we prime word or word sequences [...] in a range of social contexts and the priming [...] takes account of who is speaking or writing, what is spoken or written about and what genre is being participated in" (Hoey 2005: 13), thus, textual colligation can be considered as one of the many tools researchers may use in order to investigate the peculiarities of a genre. Indeed, since each genre may be seen as a correlation of the features found to characterise a given text (Hoey 2001), using textual colligation may help researchers highlight these correlations and, thus, identify how specific words or word sequences are primed in a particular genre. In line with the CGA approach used in this dissertation to investigate news tickers, textual colligation can thus be used to 'demystify' the private intentions behind given

textual patternings. Indeed, if in the context of news tickers, specific textual colligational patternings can be highlighted in the comparison with other genres found in the professional environment of the BBC, these can hide strategic purposes that the company wants to achieve when organising information in the way it does in the genre of news tickers. In this way, by encoding in the NTC and the bw_14 sentence-initial elements (see Section 4.2), we have been able to highlight a word sequence priming occurring in news tickers, which can be considered as a symptomatic element of a given policy hiding behind this choice of textual organisation.

3.4 News values and corpus linguistics

As previously highlighted when introducing Bhatia's approach to genre analysis, a key element in defining a genre is represented by the purpose(s) that it is intended to serve, and the way these purposes are achieved textually define the structure of a genre. Additionally, as we have also previously underlined, in the context of professional practices, these purposes take the form of private intentions that are incorporated "within the concepts of professionally shared values, genre conventions, and professional cultures" (Bhatia 2012: 23) that are at the very heart of the process of shaping the policies of a given professional practice. In this section, we would like to focus our attention on the very concept of "professionally *shared values*" (Bhatia 2012: 23, emphasis added) in the context of the news industry since, as Bell (1991) argues, in the news profession, there is a series of factors that editors take under consideration when deciding which events can be taken into the news and how they should be presented. These values, that govern in a certain way the professional practice of media production, are generally referred to as news values.

In the Journalism and Communication Studies literature, news values are usually defined as "properties of events or stories or as criteria/principles that are applied by news workers in order to select events or stories as news or to choose the structure and order of reporting" (Bednarek and Caple 2014: 2). Since they are generally referred to as 'values', van Dijk (1988a) underlines their location in social cognition, as "[t]hey are values about the newsworthiness of events or discourse,

shared by professionals [...], and indirectly by the public of the news media" (van Dijk 1988a: 119).

Firstly introduced by Galtung and Ruge (1965) in their foundational work on how events become news, the authors explain the concept of news values by using a metaphor, thus, comparing the world to "an enormous set of broadcasting stations, each one emitting its signal or its program at its proper wavelength" (Galtung and Ruge 1965: 65). In this metaphor, the set of world events can, thus, be seen as "the cacophony of sound one gets by scanning the dial of one's radio receiver" (Galtung and Ruge 1965: 65), and for this reason, the selection of events to be made into news can be quite challenging, since the media industry must choose among this continuous cacophonic sound of emitting waves. Therefore, in order to 'pick up' the right frequencies and maximise them in a given media outlet, the media industry follows certain values, presented by Galtung and Ruge (1965: 70) on the basis of this metaphor[38]. The list of values offered by the authors was, then, further developed by Bell (1991) and Bednarek and Caple (2012a). In particular, Bell (1991) classifies news values in three macro-categories:

1) values in the news text;
2) values in the news process;
3) values in news actors and events.

Values in the news text are defined by Bell (1991) as factors that contribute "in the quality or style of the news text" (Bell 1991: 160), and they include clarity (i.e., "an event with a clear interpretation, free from ambiguities in its meaning, is preferred to the highly ambiguous event from which many and inconsistent implications can and will be made" (Galtung and Ruge 1965: 66)), brevity (i.e., reducing word use and keeping sentences and paragraphs short, without compromising clarity), and colour (i.e., offering a particular stance/point of view on the news story). Since these values are more related to the "general characteristics demanded of a news story in order to

[38] The limitation entailed by this metaphor, however, must be highlighted. Indeed, this view does not take into account the fact that the process of writing is bidirectional, in the sense that, once a given frequency has been picked up by journalists, they will adjust the event in line with the expectations of their readership. Additionally, in a society where digital online products allow readers to increasingly participate in the newsroom activities (Bivens 2014) thanks to the comment section at the end of a given news story, journalists may change the way given stories have been initially presented thanks to this series of external stimuli.

be included" (Bednarek and Caple 2012a: 40), Bednarek and Caple (2012a) refer to them as "news writing objectives" (Bednarek and Caple 2012a: 40), since they can be considered as "operating factors behind the production of all stories" (Cotter 2010: 171).

Values in the news process, on the other hand, include the following ones:

- continuity: "once a story appears as news, it continues as news" (Bednarek and Caple 2012a: 40); or, in the words of Bell (1991: 159), "news breeds news";
- competition: "the competition among news institutions for scoops, the competition among stories for coverage" (Bednarek and Caple 2012a: 40); or, as Bell (1991) argues, the search for an exclusive angle of a news story;
- co-option: "associating one story with a more newsworthy one" (Bednarek and Caple 2012a: 40); or, as Bell (1991) exemplifies it, "a story which is only tangentially related can be interpreted and presented in terms of a high-profile continuing story" (Bell 1991: 159);
- composition: "the mix of different kinds of stories in the overall news bulletin or newspaper" (Bednarek and Caple 2012a: 40); or, in the words of Bell (1991), "in making up a newspaper or broadcast bulletin, editors want both a mixture of different kinds of news and some common threads" (Bell 1991: 159);
- predictability: "the scheduling of events, such as press conferences to fit the news cycle" (Bednarek and Caple 2012a: 40); this is particularly due to the fact that "[i]f an event can be prescheduled for journalists it is more likely to be covered than if it turns up unheralded" (Bell 1991: 159). However, we must underline the fact that the increasing availability of online media platforms has allowed journalists to challenge the concept of predictability in favour of the search of breaking news stories. Thus, unpredictability is increasingly becoming a news values in the news process that journalists search for in order to present their profession always on the move;
- prefabrication: "the existence of prefabricated input sources" (Bednarek and Caple 2012a: 40); or, in the words of Bell (1991), "[t]he existence of ready-made text which journalists can take over or process rapidly into a story

greatly enhances the likelihood of something appearing in the news" (Bell 1991: 159-160).

Bell's (1991) values in the news process are referred to by Bednarek and Caple (2012a) as "news cycle/market factors" (Bednarek and Caple 2012a: 41), since they represent the operating factors behind the selection of given news stories.

Finally, as for the values in news actors and events, Bell (1991), drawing again on the work of Galtung and Ruge (1965), identifies the following ones:

- negativity: "[n]ews stories very frequently concern bad happenings such as conflicts, accidents, damages, injuries, disasters or wars" (Bednarek and Caple 2012a: 42); in this sense, Bell (1991) defines negativity as "the basic news value" (Bell 1991: 156), since negative events can be seen as "the basic stuff of 'spot' news" (Bell 1991: 156). Positive news stories are nonetheless featured in the news, but they can be seen as fulfilling a stabilising function, always compared to an assumption of destabilisation. In other words, 'feelgood' stories seem to balance out the media outlet, which basically enhances negative news stories. Thus, positive news stories acquire newsworthiness because they contradict readers' or viewers' negative expectations. In this context, as Bell (1991) highlights, deviance also plays an important role. Indeed, as van Dijk (1988a) argues, since cognitive models are about everyday routine situations and actions, "[i]nformation about deviant and negative situations provides deviant models, which can be better retrieved and recalled due to their distinctiveness" (van Dijk 1988a: 123);
- recency (or timeliness): "[m]ore recent events are often more newsworthy" (Bednarek and Caple 2012a: 42). While Bell (1991) prefers to refer to this news values as "recency" (Bell 1991: 156–157), seeing it as the relevance of given events in terms of time, since "the best news is something which has only just happened" (Bell 1991: 156), Bednarek and Caple (2012a) refer to this value as "timeliness", since it entails how temporally relevant an event is to readers/viewers and, thus, as a value, it can also be constructed in discourse in order to project the temporal deictic centre in that of the viewers'/readers';

- proximity: “[w]hat is newsworthy usually concerns the country, region or city in which the news is published” (Bednarek and Caple 2012a: 42). In this sense, “geographical closeness can enhance news value” (Bell 1991: 157). However, as Bell (1991) points out, proximity is related to Galtung and Ruge’s (1965) news value of “meaningfulness” (Galtung and Ruge 1965: 67), which does not only entail the geographical distance between the event and the country, region or city in which the news story is published, but also “the cultural familiarity and similarity of one country with another” (Bell 1991: 157);
- prominence/eliteness: “[s]tories about ‘elite’ individuals or celebrities are more newsworthy than stories about ordinary people” (Bednarek and Caple 2012a: 43). Thus, “[r]eference to elite persons such as politicians or film stars can make news out of something which would be ignored about ordinary people” (Bell 1991: 158). Eliteness, however, plays also an important role in the selection of given sources over others. Thus, a news story coming from an established news media company will be more newsworthy if compared to a news story coming from a small and local media company. In the case of the prominence of sources, Bednarek and Caple (2012a) argue that it is sometimes given its own news value label, that is, “attribution” (Bednarek and Caple 2012a: 43; see also Bednarek 2015). Bell (1991) introduces it as a separate news value and argues that the eliteness of a news story’s source “can be crucial in its news chance” (Bell 1991: 158). Indeed, socially validated authorities have better chances to see their ‘voices’ being represented in the media, while “[t]he unaffiliated individual is not well regarded as a source” (Bell 1991: 158);
- consonance: “[t]he extent to which aspects of a story fit in with stereotypes that people may hold about the events and people portrayed in it” (Bednarek and Caple 2012a: 41). This news value seems to confirm what van Dijk (1988a) underlines about news values as located and responding to a social cognition, which builds on expectations and ways to perceive specific events. Thus, consonance of a news story refers to “its compatibility with preconceptions about the social group or nation from which the news actors

come" (Bell 1991: 157), therefore, meeting the stereotypes that editors have about given events and actors, and consequently reproducing these stereotypes in viewers/readers;

- impact: "[t]he effects or consequences of an event are aspects of a story that are newsworthy, especially, if they involve serious repercussions or have a more global impact" (Bednarek and Caple 2012a: 43). Bell (1991), in line with van Dijk (1988a), refers to this news value as "relevance", since it is linked to "the effect on the audience's own lives or closeness to their experience" (Bell 1991: 157). Van Dijk (1988a), however, highlights how "relevance must be defined in terms of large or powerful groups" (van Dijk 1988a: 122). Indeed, since there may be different groups of viewers/readers, with different interests and expectations, "[m]inority relevance is much less emphasized" (van Dijk 1988a: 122), in the sense that effects on minority groups in society may be overlooked by the media, while effects on relevant social classes are much more newsworthy. This is also linked to local and global impact of given events: the more global the relevance, the more newsworthy the event;
- novelty: "[n]ews stories are frequently about happenings that surprise us, that are unusual or rare" (Bednarek and Caple 2012a: 43). In other words, the more unexpected, the more newsworthy the event is. Bell (1991), however, prefers to distinguish between the news values of unexpectedness and novelty since, while the former refers more specifically to the fact that "the unpredictable or the rare is more newsworthy than the routine" (Bell 1991: 157), the latter enhances the 'breaking news' aspect of a given event. Thus, while not being completely 'unexpected' (for instance, in the case of pre-scheduled events), an event may be presented by enhancing its new aspects;
- superlativeness: "news stories usually focus on maximizing or intensifying particular aspects of an event" (Bednarek and Caple 2012a: 44). In this way, given aspects of an event are maximised or intensified in order to enhance their newsworthiness. Galtung and Ruge (1965) do not refer explicitly to this news value, but rather, by continuing their metaphor linked to radio

waves, they introduce the notion of threshold, according to which "[t]he stronger the signal, the greater the amplitude, the more probable that it will be recorded as worth listening to" (Galtung and Ruge 1965: 66);

- personalisation: "[n]ews stories that are personalized attract audiences more than the portrayal of generalized concepts or processes" (Bednarek and Caple 2012a: 44). Thus, the personal or human side of an event is more newsworthy than an abstract concept. This tendency in media discourse has been particularly studied by Fairclough (1992), who refers to it as the widespread tendency to "synthetic personalization" (Fairclough 1992: 98) in news discourse, that is, "the simulation of private, face-to-face, discourse in public mass-audience discourse [...] linked to a spread of conversational discourse from the private domain of the lifeworld into institutional domains" (Fairclough 1992: 98);
- facticity: "the degree to which a story contains the kinds of facts and figures on which hard news thrives: locations, names, sums of money, numbers of all kinds" (Bell 1991: 158). Thus, figures and numbers play an important role in enhancing the newsworthiness of a given event. Indeed, as Bednarek and Caple (2012a) maintain, while, on the one hand, these elements provide 'facts' to journalists, thus, making the news story seem more objective, on the other, they also construe, at the same time, Superlativeness (and, sometimes, Impact), thus, enhancing the newsworthiness of the event.

The list of news values that fall under Bell's macro-category of values in news actors and events seems to play a key role in enhancing the newsworthiness of a news story and, while Bell (1991) categorises them under this label, Bednarek and Caple (2012a) prefer to refer to them as core news values since it is the authors' belief that, as "the three categories that Bell distinguishes are very different in kind" (Bednarek and Caple 2012a: 41), they should not be labelled altogether under the umbrella term of 'news values', because this diminishes the purposes that each one serves in enhancing newsworthiness. Thus, as we can see in Table 8, Bell's (1991) macro-categories under which news values have been organised are re-labelled in the framework of analysis developed by Bednarek and Caple (2012a).

Bell's (1991) categories	Bednarek and Cable's (2012a) categories
Values in the news text	News writing objectives
Values in the news process	News cycle/market factors
Values in news actors and events	News values

Table 8 Bell's (1991) macro-categories of news values and Bednarek and Caple's (2012a) corresponding categories.

Additionally, we must also underline the fact that Bednarek and Caple's (2012a) view is particularly linked to their discursive approach to news values, which investigates "how newsworthiness is construed and established through discourse" (Bednarek and Caple 2012b: 104). Indeed, according to the authors (Bednarek and Caple 2012b, 2014; Caple and Bednarek 2015), news values can be conceptualised in terms of "how newsworthiness is construed or established through discourse (both language and image)" (Bednarek and Caple 2012b: 104). Thus, a discursive perspective sees news values as "quality of *texts*" (Caple and Bednarek 2015: 13, emphasis in the original), and their analysis can allow us to "systematically investigate how these values are constructed in the different types of textual material involved in the news process" (Bednarek and Caple 2012b: 104). This approach to news values allows the authors to highlight given textual traces, referred to as "pointers" to newsworthiness (Bednarek and Caple 2014: 11)[39], that can let us glimpse at how they are realised in news discourse, as we can see from Table 9 (Bednarek and Caple 2012a: 55-56, 2012b: 106).

[39] As Bednarek and Caple (2012a) explain, they use the term 'pointer' because they have not "examined each occurrence in its co-text" (Bednarek and Caple 2012a: 196), meaning that, by looking at corpus word lists, they have focused on those items potentially embodying a specific news value, without further investigating its collocates, for instance.

NEWS VALUES	EXAMPLES
NEGATIVITY	***terrible*** *news, a* ***tragedy***, ***distraught***, ***worried***, ***breaking our hearts***, ***killed***, ***deaths***, ***bodies***, etc.
TIMELINESS	***breaking news***, ***today***, ***yesterday***, [use of tenses that express that an event has only just happened, is still ongoing or will happen in the (near) future], etc.
PROXIMITY	[geographical names or cultural references]
PROMINENCE / ELITENESS	*pop* ***star***, ***celebrity*** *bad boy*, ***President***, ***MP***, etc.
CONSONANCE	***legendary***, ***notorious***, *a* ***flood*** *of immigrants*, ***yet another*** *personal scandal*, etc.
IMPACT	*a potentially* ***momentous*** *day, a terror that* ***took their breath away***, ***thousands*** *of people*, etc.
NOVELTY	*a very* ***different*** *sort of disaster, a* ***new*** *discovery*, ***unusual***, etc.
SUPERLATIVENESS	*they were* ***petrified***, *a* ***giant*** *storm, a tragedy of* ***epic*** *proportions*, etc.
PERSONALISATION	[an 'ordinary' person telling their story]

Table 9 Summary offered by Bednarek and Caple (2012a: 55-56, 2012b: 106) of the linguistic cues that can be used in order to construe news values.

The list of pointers to newsworthiness should not be restricted to the examples given in Table 9 and, furthermore, as Bednarek and Caple (2012b) maintain, certain textual devices can "simultaneously construe more than one news value and hence contribute significantly to rendering the story newsworthy" (Bednarek and Caple 2012b: 106). Additionally, Bednarek and Caple (2014) argue that "an analysis of how news values are discursively constructed in texts should be both 'manual' and 'multimodal'" (Bednarek and Caple 2014: 6). Indeed, since news discourse is increasingly relying on both textual and visual elements in delivering news stories to readers/viewers, images and other multimodal elements should also be analysed in order to see if the news values constructed through these media further enhance or contrast with the news values found in the textual elements of a news story (Caple and Bednarek 2015). Thus, such an analytical framework inevitably forces researchers to adopt a manual approach to the analysis of news values, since "only

through close analysis of texts can we find out what values are emphasised (foregrounded), rare or absent (backgrounded)" (Bednarek and Caple 2014: 6).

However, when approaching large amount of data, Bednarek and Caple (2014) and Potts, Bednarek and Caple (2015) argue that corpus linguistic techniques (such as the analysis of frequencies (word forms, lemmas, clusters), the analysis of keywords or grammatical/semantic tags, dispersion analysis, etc.) can help with news values analysis, especially in those cases where a non-topic-specific corpus is under investigation. Indeed, through the use of corpus linguistic methodologies, we can gain "first insights into a conventionalised repertoire of rhetoric of newsworthiness" (Bednarek and Caple 2014: 14) in the case of corpora representative of specific media genres. Thus, if "every journalist and every editor will have a different interpretation of what is newsworthy" (Rau 2010: 15), corpus linguistic techniques can help researchers identify "what kind of discursive devices are repeatedly used [...] to construct different news values" (Bednarek and Caple 2014: 16) and, consequently, they can take us to the backstage of the news production process. In this way, the combination of news values analysis and corpus linguistic methodologies can be used to better define a genre since, by underlining what is newsworthy for a specific news organisation, they can help researchers 'sneak a peek' into the professional practices at the very heart of the news production process. And, as we will see in Chapter 5, the analysis of the news values foregrounded in news tickers has helped us confirm one of the main characteristics of this genre, thus, demonstrating its usefulness in the analysis of the perpetuation of specific purposes in news tickers.

As we have seen in this Chapter, various are the methodologies used in order to analyse the genre under investigation. While the initial introduction to Corpus Linguistics has been outlined in order to better understand the type of corpus collected and its characteristics, in the Section dedicated to genre analysis we have briefly described the approach to text analysis that we have decided to follow in this investigation on the nature of news tickers, an approach that wants to focus on the linguistic cues that are representative of the professional practice behind the production of this genre. In this, the theory of lexical priming and, more specifically, the concept of textual colligation combined with the identification of specific news

values enhanced in news tickers have been used in order to demystify the practices that are textually embodied in news tickers. In the following Chapter, we will now move on to the description of the way the corpus under investigation was collected. Additionally, we will also introduce the tools used in order to carry out our analysis of news tickers.

4. Corpus collection and description

As previously underlined in the introduction to this dissertation, genre analysis is increasingly changing in order to stay up-to-date with the dynamically changing context of contemporary society. This social context has demanded a reshaping of its conventional approach to textual analysis, since genres are progressively becoming fluid entities open to unexpected innovations by borrowing structural conventions and rhetorical configurations from other generic forms. This challenge to genre analysis, however, can be easily overcome by the increasing availability of corpora to researchers. Thus, changes in professional practices can be successfully highlighted by the use of corpus linguistic methodologies.

However, the availability of ready-made corpora may cause some disadvantages on behalf of the researcher interested in particular areas of human communications, since "a corpus is always designed for a particular purpose", and the majority of them "are created for specific research projects" (Xiao 2008: 383), thus, focusing only on specific genres, while others remain unexplored. We have also seen that, in the case of general corpora, while they are increasingly becoming freely open to researchers, they can be considered as ill-suited for investigations focusing on genre analysis, since they have been collected in order to be representative of a language as a whole, and thus they may contain only fragments of given genres, while disregarding others.

Given these premises, in order to study very specific instances of language in use of a particular discourse community, most of the time, researchers have to create their own specialised corpora, and this is particularly true in the case of news tickers, given the unavailability of an already-built corpus but, more importantly, no database with instances of this genre[40]. Thus, the lack of any trace of this genre has forced us

[40] In a private interview with one of the journalists working at the BBC Broadcasting House in London, we have also discovered that the news network itself does not archive news tickers in its database, since they are automatically generated from news stories published on the BBC website. An editor, then, proofreads the news tickers thus created and sends them to the graphic editor, who will further revise them in order to implement them on the TV screen. The automatic extraction of news stories, however, does not leave any trace and, for this reason, news tickers are not archived.

to, first and foremost, collect the data by following the procedure that will be outlined in the following Sections.

4.1 Collecting the NTC and the bw_14

In order to observe the genre in its own 'environment', a week-long preliminary observation and recording of crawlers displayed on the BBC World News, Fox News, CNN, ABC's programme *Good Morning America*, and CBS' programme *This Morning*[41] was undertaken. The BBC World News was accessed from Italy thanks to its official online streaming website for viewers living outside of the UK, Livestation[42]. Initially, the BBC World News was recorded by using, for a limited period of time, the freeware Screencast-O-Matic[43], which allows to capture video and audio contents that are being played on a computer screen. However, due to some limitations of this software (for instance, the low quality of the video recorded, something that was fundamental, since our study focuses on a particular textual element of TV news broadcasting), Camtasia Studio 7.1.1 (TechSmith Corporation 2011) was then employed, which allows users to record the video and audio being played on a computer screen, but also offers the possibility to edit the recorded audio/video contents and save the file in an HD format, increasing the quality of the recordings.

---

[41] In the specific case of the ABC's programme *Good Morning America* and CBS' programme *This Morning*, at the time of recording, these were the only TV programmes during which news tickers were displayed on the screen of the ABC and CBS TV networks. This is the reason why we have focused our attention on them, while in the case of the BBC World News, Fox News and CNN, crawlers were constantly present on the TV screen.

[42] Livestation, which allows people to legally and freely watch various TV news channels online (such as Al Jazeera America, Al Jazeera English, BBC Arabic, BBC World News, CNBC EMEA, CNN International, Euronews English, France 24, Reuters, Sky News Arabia, VoA Persian, etc.), can be reached at `http://www.livestation.com` However, even though the majority of TV news channels can be streamed freely online on Livestation, at the time of recording, a few Premium Stations (such as BBC World News and CNN International) could be accessed at a cost of $1.99 per month, sharing this fee with the TV stations that provide this service.

[43] Screencast-O-Matic can be freely used at the following address: `http://screencast-o-matic.com/home` The freeware both features an online cloud platform, thus, providing the possibility not to install the software on the computer; otherwise, it also features an offline software that can be easily installed on a computer. In its premium version, further features of the software can be unlocked, such as the possibility to edit the recorded video.

As for the American TV news channels, another online streaming platform was used, that is, USTV Now[44], a streaming service developed specifically for American soldiers around the world, thus, allowing people to access from other countries the American cable TV at a monthly price.

This preliminary collection of data was based on an initial pilot study on the comparison between news tickers displayed on the UK and US TV news channels. However, the initial pilot study based on this comparison was soon abandoned since the context where the data were collected was due to influence the analysis and its conclusions in itself. Indeed, since the American component of the corpus did not include public TV channels, while at the time of the recording the only UK TV news channel using news tickers was public, this would have inevitably skewed the data. Indeed, the differences between the two corpora were already built in in the different professional contexts represented in the data. Additionally, this preliminary observation of the journalistic practices of the previously mentioned UK and US TV news channels also revealed something crucial to our decision to focus only on the BBC World News. Indeed, while on the American news channels crawlers seemed to play a secondary role, thus, they were suddenly cut off due to commercial breaks (and, more importantly, they did not continue from where they were interrupted); or they were not displayed in case of weather emergencies, during which a flipper with information regarding schools closed due to the adverse weather conditions substituted them (this was particularly true in the case of the ABC's programme *Good Morning America* and CBS' programme *This Morning*); in the case of the BBC World News, crawlers were constantly present on the TV screen and, thus, embedded in the journalistic practice of the TV news channel. Indeed, even in the case of breaking news stories, news tickers did not disappear from the TV screen, a difference we have noticed, for instance, in the comparison with the CNN and Fox News, which during breaking news stories 'silenced', so to speak, news tickers, in

[44] USTV Now can be reached at the following address: `http://www.ustvnow.com/` The website offers the unique possibility to DVR the TV programmes that users may be interested in for a monthly price (at the time of writing) of $29 (after 3 months, $39 per month). This has allowed us to easily collect the data from the American news channels previously mentioned. However, due to copyright issues, it was not possible to download the DVRed videos and, thus, in order to save them, the same procedure used in order to collect the video live-streamed online of the BBC World News was followed.

order to give more relevance to supers. In the case of the BBC World News, on the other hand, crawlers played a key role in displaying the development of the breaking news story, thus, confirming their importance in the journalistic routines of the British news channel. Additionally, we have decided to focus our attention on the British media panorama since, while in the scarce literature on news tickers most of the examples were taken from the American context of production, no example featured crawlers taken from the UK context. And since the BBC (together with the CNN) represents nowadays a standard reference in TV news production (Bivens 2014), investigating news tickers in this journalistic environment can reveal general tendencies in the genre as a whole.

Once the American component of our corpus was discarded for the reasons previously underlined, this gave us the possibility to focus on the British part of the corpus and, thus, on the BBC World News. However, a new problem arose when facing the BBC data collection. Since news tickers were constantly displayed on the BBC World News during the day[45], a decision was to be made on the part of the day during which we would record the genre under investigation, in order to decrease the redundancy of news tickers (i.e., news tickers displaying the same information and textual structure) and lower the likelihood of a singular event to dominate the scene. Thus, another week-long observation of the BBC World News was undertaken. During this time-frame, news tickers displayed on the BBC World News channel were recorded in three parts of the day (i.e., at 8:00 a.m., at 12:00 p.m. and at 8:00 p.m. GMT). From this additional observation of the genre, we have decided to focus our attention on the news tickers displayed at 12:00 p.m. during the BBC World

[45] During the period of our investigation, we have noticed that news tickers were not displayed on the TV screen of the BBC World News only during weather forecasts, commercial breaks about the BBC News programming and, more importantly, during documentaries, where their presence might have distracted viewers, thus, subtly acknowledging what Josephson and Holmes (2006) discovered in their study on the interference effect of on-screen visual enhancements on information recall of TV news stories.

News' programme *GMT*[46], since they contained both news stories reported in the morning bulletin and those news stories regarded as still newsworthy by the BBC's editorial board displayed in the 8:00 p.m. bulletin of the previous day. We have, then, daily recorded and transcribed in a .txt file thanks to the software Dragon NaturallySpeaking 12.5 (Nuance Communications 2013)[47] the news tickers displayed during this TV news programme from March 12, 2013 to April 11, 2014 (for a total of 365 days), thus, creating the News Tickers Corpus (NTC), which is comprised of 168,513 tokens (for a total number of 6,937 news tickers). The corpus was, then, annotated through XML encoding (Hardie 2014), which gives researchers enough freedom to develop their own coding given the specificities of the genres under investigation[48].

In order to highlight some of the peculiarities found in the NTC, a reference corpus was also collected of all the headlines and lead paragraphs found on the BBC

[46] While this has generally been the procedure followed in collecting our data, a degree of variation was also included, both for personal (e.g., impossibility to access the Internet) and technical reasons (e.g., the Livestation website was down for maintenance and, thus, could not be accessed at that particular time of the day). However, while in some cases we have found a way around these 'obstacles' (e.g., by recording, if available, the BBC World News' *GMT* programme broadcasted on BBC America and archived on the non-profit internet library *The Internet Archive*, available at `https://archive.org`), we have looked at this degree of variability as an opportunity rather than a limitation in our data. Indeed, the fact that the data were sometimes collected at different times of the day proves that the observations that we will make on crawlers can be applied in general to the genre of news tickers as displayed on the BBC World News as a whole rather than during a particular TV news programme.

[47] Dragon NaturallySpeaking (Nuance Communications 2013) is a speech recognition software that offers, amongst its various features, the possibility to dictate, edit, and format documents all by voice. This has allowed us, after a preliminary training period, to read aloud the tickers displayed on the BBC World News and, by doing so, to automatically transcribe them, thus, improving the accuracy of the transcriptions (thanks to Dragon's built-in English (UK) dictionary and its training system) and reducing the time spent transcribing the data by hand alone.

[48] For more information on the annotation scheme encoded in the NTC, see Section 4.2.

news website thanks to the online database LexisNexis[49] from June 1, 2014 to July 31, 2014, thus, creating the bw_14 corpus[50]. As Baker (2006: 30) explains, a reference corpus can be defined as a much bigger collection of data (when compared to the target corpus) representative of a given language variety. Additionally, a reference corpus may be seen as a "corpus which is not under scrutiny in a particular analysis" but is "used for inter-textual analysis" (Tabbert 2015: 59), that is, as a data collection used to compare and highlight peculiarities in a target corpus (McIntyre 2013), in contrast with intra-textual analyses that only focus on a target corpus without the use of a reference corpus (Adolphs 2006: 66-68).

As previously said, the reference corpus we have decided to create in order to highlight certain peculiarities in the NTC has been cleaned so as to include only the headlines and lead paragraphs (in other words, the nucleus; White 1997, 2000) of the news stories published on the BBC website in the aforementioned time-span. The reason behind this choice is linked to our conviction that, from a textual point of view, headlines and lead paragraphs are the most similar in length and function to news tickers, and thus they can help us highlight differences/similarities in the genre under investigation. Our assumption on the similarity between the textual genres represented in the NTC and the bw_14 is, indeed, supported by White (1997: 122), who argues:

> The opening headline/lead 'nucleus' casts the reader abruptly into the core subject matter of the report [...]. And perhaps most tellingly, the opening 'nucleus' goes directly to those aspects of the event or state-of-affairs which are assessed as constituting the peak or climax of social-order disruption. That is, it singles out those aspects of the event or issue at hand which pose

[49] As we can read on its official website (`http://www.lexisnexis.com`), LexisNexis offers the possibility to quickly access a wide range of full-text documents from over 17,000 sources from around the world and download them for a wide range of academic research projects. It allows users to download up to 500 documents at a time and save them in different formats (.doc, .txt, etc.). In the specific case of news research sources, it currently features more than 3,000 newspapers from around the globe and more than 2,000 magazines, journals, and newsletters. Additionally, it also offers the possibility to download broadcast transcripts, wires services updated, and blogs and video blogs from different news organisations.

[50] More information on the annotation scheme used in the bw_14 can be found in Section 4.2.

> the greatest threat to the material, power-relational or normative status-quo, extracting them from their original chronological or logical context and thus compelling the reader to engage immediately with some crisis point of social order disequilibrium.

The reference corpus thus created is comprised of 617,311 tokens (for a total number of 20,205 headlines and their accompanying lead paragraphs) and its selection as a reference corpus was based on the following hypothesis: given the same professional environment, what changes can be highlighted when contents migrate from one textual genre to the other and, more importantly, from one platform to the other? Thus, the focus here is not on the genres *per se*, but on how the different platforms influence those genres, what they can reveal about them and, in the case of news tickers, how traditional genres can help us highlight what peculiarities can be found in them.

Additionally, we must also underline that the time discrepancy in collecting the NTC and the bw_14 corpus was also driven by the need to lower the chances that structural similarities were due to identical news contents. Indeed, the same news content might have forced a similarity that was only due to its related stance towards a given news story. Thus, by chronologically separating the collection of the two corpora, we have tried to ensure that similarities or differences were only due to the different media environment taken into account, so as to help us better define the nature of news tickers.

Finally, we have decided to collect only the headlines and lead paragraphs of the news stories published during the time-span that goes from June 1, 2014 to July 31, 2014 in line with Berber-Sardinha's (2000) empirical suggestions on reference corpus size when it comes to robust statistical outcomes in a keyword analysis context (see Section 5.2). Reference corpus size seems to be a quite disputed issue. Indeed, while Xiao and McEnery (2005) illustrate that it is irrelevant in measuring the keyness of given items in a target corpus, since "the size of a reference corpus is

not very important in making a keyword list" (Xiao and McEnery 2005: 70)[51], Scott and Tribble (2006: 64) argue that "above a certain size, the procedure [keyword analysis] throws up a robust core of KWs whichever the reference corpus used". The authors claim this on the basis of Berber-Sardinha's (2000) investigation on the ideal size of a reference corpus in order to make robust statistical claims on a given target corpus. In his study (Berber-Sardinha 2000), five English corpora were collected and keywords were compiled in the comparison with reference corpora of various sizes (from two to one-hundred times larger than the target corpora). From this comparison, the author demonstrated that, while two, three and four times bigger reference corpora yielded keyword lists that were limited and tended to vary from each other, when using reference corpora that were five times larger than the target corpus, the keyword list offered a larger number of keywords and, more importantly, a certain degree of stability in the statistical outcome of the keyword analysis was achieved, since the reference corpus yielded similar results and amounts of keywords as the ones provided by using larger reference corpora. To sum up, Berber-Sardinha's (2000) enquiry in the ideal size of a reference corpus indicated that "a larger reference corpus is not always better than a smaller one, [...] while a reference corpus that is less than five times the size of the study corpus may not be reliable" (Berber-Sardinha 2000: 7). However, "[t]here seems to be no need for using extremely large reference corpora, given that the number of keywords yielded do not seem to change by using corpora larger than five times the size of the study corpus" (Berber-Sardinha 2000: 7). Hence, as Scott and Tribble (2006) and Scott (2009) argue, a more focused reference corpus selection can help researchers overcome difficulties linked to the disadvantages of using reference corpora whose context is not clear or can be time-consuming to retrieve. Thus, the reason why we have decided to focus our attention on the time span previously highlighted for the bw_14 reference corpus is strictly linked to its size. Indeed, even though we have also collected data that go from August 1, 2014 to August 31, 2014, we have decided to

[51] Even though McEnery *et al.* (2006) argue, in agreement with Tribble (1999), that "[t]here is little advantage in using relatively large reference corpus" (McEnery *et al.* 2006: 318), especially in the analysis of negative keywords (i.e., words that are relatively infrequent in the target corpus in the comparison with a reference corpus), since in order to understand the reason behind their statistical infrequency, we must refer back to the reference corpus and to its context.

exclude these data from the reference corpus, since the size of the corpus was already big enough for a robust statistical keyword analysis of the NTC corpus.

4.2 Annotating the NTC and the bw_14

Once the NTC corpus was completely transcribed, we have proceeded to its annotation since, as we have argued in Section 3.1.4, metadata, mark-ups, and annotations allow us to communicate with our computers by encoding information that were lost from the original context the data were retrieved and by adding given linguistic information to the data, in order to easily reveal patterns in the texts under investigation.

As for the metadata and mark-ups, in order to encode them in the NTC corpus, a semi-automatic annotation procedure was followed by using TextPad (Helios Software Solutions 2014), a Windows based software for text-editing that features the possibility to easily manipulate large amounts of data thanks to its multiple tools[52]. The textual regularities in the presentation of news stories in news tickers (in particular, the fact that each section was introduced by a given label[53]) has allowed us, in particular, to fruitfully exploit the regular expression (regex) tool offered in the "Replace" engine of TextPad. This has allowed us to, firstly, add metadata information about the date and time of each transcribed news ticker recording, which were encoded in XML language and, more specifically, coded as `<date value="2013-03-12" time="12:00 p.m.">` (this string was closed by using the code `</date>`). The texts enclosed in this XML code, thus, correspond to the round of news tickers played on the news network on the day and at the specific time encoded in the XML string. However, we must promptly underline that the time codes refer to the time span that goes from the beginning of the time code in question to the next (i.e., the time code `<time="12:00 p.m.">` refers to time span that goes from 12:00 p.m. to 12:59 p.m.): in other words, the news ticker enclosed in this XML string was collected in this time span. This is due to the fact that journalistic schedules may not be as precise as they might have been reported on their official TV

[52] A full-featured trial version of the software can be easily and freely downloaded at `https://www.textpad.com`

[53] See Chapter 5 for further information on the structural presentation of news stories in news tickers.

schedule, due to breaking news stories or any other type of delay that may influence the regular schedule of news programmes. For this reason, we have decided to underline this aspect of our data collection.

As we will see in Section 5.1, up until December 18, 2013, in the BBC World News' news tickers, a complete round of news tickers was comprised of two rounds of them. In the first round, five sections could be differentiated. In the section HEADLINES, a number of six major news stories was displayed. The HEADLINES section was followed by the BUSINESS section, where three major economic news stories were displayed. But the peculiarity of this first part of a complete round of tickers is represented by the sections MARKETS and CURRENCIES, where stock indexes and the value of foreign exchange rates were presented. Finally, this first part ended with the SPORT section, where two major sport updates were shown. The first round of tickers was closed by the following message:

(3) WEBSITE
MORE ON ALL THESE STORIES AT bbc.com/news
TWITTER
FOR LATEST FOLLOW US VIA @bbcworld AND @bbcbreaking

Once this message scrolled, announcing the end of the first round of tickers, a second round was introduced, showing again some of the sections displayed in the first one. The section HEADLINES of the second round usually introduced seven major news stories. The first two news stories were the same first two displayed in the first round of tickers, while the third news story of the second round was the last one presented in the section HEADLINES of the first round of tickers. These three news stories were followed by further four news updates that had not been previously introduced. The section BUSINESS in the second round of tickers displayed as its first news story the first one presented in the section BUSINESS of the first round of tickers, followed by two new economic updates that had not been previously displayed. Finally, the BUSINESS section was followed directly by the SPORT section, where two new sport updates were displayed. The second round of tickers was closed by the following message:

(4) CONTACT US
HAVE YOUR SAY AT facebook/bbcworldnews
- WEBSITE: bbc.com/haveyoursay
- EMAIL: haveyoursay@bbc.co.uk
- SEND YOUR VIDEOS TO: whysvideo@bbc.co.uk
- FOR TERMS ON SENDING PICTURES AND VIDEOS: bbc.com/terms

This quite complicated routine, which will be better exemplified and explained in Section 5.1, was dropped on December 18, 2013, when a single, simpler, and much shorter round of tickers was developed and displayed. Indeed, the MARKET, CURRENCIES, and SPORT sections were dropped and only the sections HEADLINES and BUSINESS survived in this new format of news tickers.

In order to encode these differences in the rounds of tickers previously highlighted, we have decided, once again, to use XML scripts so as to, additionally, lower the degree of repetition in the collection of our data. Indeed, instead of daily reporting the two messages at the end of each round of tickers that signalled, thus, their chronological sequence, all the tickers included in the code `<div1 id="1">` belonged to the first round of tickers (this string was closed by using the code `</div1>`). On the other hand, all the tickers appearing in the second round of tickers were enclosed in the code `<div1 id="2">` (this string was once again closed by using the code `</div1>`). When, from December 18, 2013 on, this complicated routine was abandoned, the new type of structural presentation of news tickers was encoded as `<div1 id="4">` (again, closed by the code `</div1>`). And, finally, the NTC corpus also contains some examples of breaking news crawlers, in order to see if they differ from regular news tickers. These crawlers were annotated as `<div1 id="3">` (again, this string was closed by using the code `</div1>`).

In order to better understand the annotation schemes presented so far (and introduce some of the mark-ups that will be soon presented), we would like to offer the following example randomly taken from the NTC corpus:

(5)
```
<date value="2013-06-03" time="12:00 p.m.">
<div1 id="1">
<div2 type="news">
<head type="ticker"> <s type="first">A FIRE HAS KILLED AT LEAST 119 PEOPLE AT A POULTRY PROCESSING PLANT IN NORTH-EAST CHINA'S JILIN PROVINCE, OFFICIALS SAY</s> </head>
<head type="ticker"> <s type="first">PM RECEP TAYYIP ERDOGAN AGAIN CONDEMNS THE ANTI-GOVERNMENT PROTESTS IN TURKEY, NOW IN THEIR FOURTH DAY, SAYING THEY DO NOT CONSTITUTE A TURKISH SPRING</s> </head>
```

<head type="ticker"> <s type="first">THOUSANDS OF PEOPLE FLEE THEIR HOMES ACROSS SOUTHERN GERMANY, THE CZECH REPUBLIC AND AUSTRIA AS DEADLY FLOOD WATERS CONTINUE TO RISE</s> </head>
<head type="ticker"> <s type="first">TAIWAN'S IMPRISONED FORMER PRESIDENT CHEN SHUI-BIAN TRIES TO TAKE HIS OWN LIFE, THE JUSTICE MINISTRY SAYS</s> </head>
<head type="ticker"> <s type="first">AT LEAST 10 CHILDREN ARE AMONG 13 KILLED AS A SUICIDE BOMBER ATTACKS A MILITARY PATROL IN EASTERN AFGHANISTAN, SAY POLICE AND NATO OFFICIALS</s> </head>
<head type="ticker"> <s type="first">US SOLDIER BRADLEY MANNING, ARRESTED THREE YEARS AGO ON SUSPICION OF LEAKING MILITARY SECRETS TO WIKILEAKS, IS DUE AT A MILITARY COURT FOR TRIAL</s> </head>
</div2>
<div2 type="business">
<head type="ticker"> <s type="first">THE PACE OF DECLINE IN THE EUROZONE'S MANUFACTURING SECTOR EASED LAST MONTH, ACCORDING TO A CLOSELY-WATCHED SURVEY</s> </head>
<head type="ticker"> <s type="first">TURKEY'S MAIN SHARE INDEX FALLS 6% FOLLOWING THE ESCALATION OF ANTI-GOVERNMENT PROTESTS OVER THE WEEKEND</s> </head>
<head type="ticker"> <s type="first">TECHNOLOGY GIANT APPLE IS TO BEGIN ITS DEFENCE AGAINST CHARGES BY THE US GOVERNMENT THAT IT COLLUDED TO SET THE PRICES OF E-BOOKS</s> </head>
</div2>
<div2 type="sport">
<head type="ticker"> <s type="first">JOSE MOURINHO TELLS SPANISH TELEVISION HE EXPECTS TO BE REAPPOINTED AS CHELSEA MANAGER BY THE END OF THE WEEK</s> </head>
<head type="ticker"> <s type="first">BRAZIL V ENGLAND REACTION, LATEST ON JOSE MORINHO, FOOTBALL TRANSFER UPDATES AND OTHER BREAKING NEWS</s> </head>
</div2>
</div1>
<div1 id="2">
<div2 type="news">
<head type="ticker"> <s type="first">A FIRE HAS KILLED AT LEAST 119 PEOPLE AT A POULTRY PROCESSING PLANT IN NORTH-EAST CHINA'S JILIN PROVINCE, OFFICIALS SAY</s> </head>
<head type="ticker"> <s type="first">PM RECEP TAYYIP ERDOGAN AGAIN CONDEMNS THE ANTI-GOVERNMENT PROTESTS IN TURKEY, NOW IN THEIR FOURTH DAY, SAYING THEY DO NOT CONSTITUTE A TURKISH SPRING</s> </head>
<head type="ticker"> <s type="first">US SOLDIER BRADLEY MANNING, ARRESTED THREE YEARS AGO ON SUSPICION OF LEAKING MILITARY SECRETS TO WIKILEAKS, IS DUE AT A MILITARY COURT FOR TRIAL</s> </head>
<head type="ticker"> <s type="first">ONE OF AUSTRALIA'S LEADING INDIGENOUS FIGURES, THE SINGER YUNUPINGU, WHOSE MUSIC HELPED BRIDGE THE DIVIDE BETWEEN WHITE AND BLACK AUSTRALIANS, HAS DIED AGED 56</s> </head>
<head type="ticker"> <s type="first">LEFT-WING FARC REBELS IN COLOMBIA DENY ISSUING DEATH THREATS AGAINST TRADE UNIONISTS AND BLAME "RIGHT-WING FORCES" FOR THE THREATS</s> </head>
<head type="ticker"> <s type="first">THE FAMILY OF SOUTH AFRICAN ATHLETE OSCAR PISTORIUS IS "SHAKEN" BY LEAKED PHOTOS OF THE BATHROOM WHERE HE SHOT DEAD HIS GIRLFRIEND IN FEBRUARY, A SPOKESMAN SAYS</s> </head>
<head type="ticker"> <s type="first">PALESTINIAN AUTHORITY PRESIDENT MAHMOUD ABBAS APPOINTS RAMI HAMDALLAH AS PRIME MINISTER FOLLOWING THE RESIGNATION OF SALAM FAYYAD</s> </head>
</div2>
<div2 type="business">
<head type="ticker"> <s type="first">THE PACE OF DECLINE IN THE EUROZONE'S MANUFACTURING SECTOR EASED LAST MONTH, ACCORDING TO A CLOSELY-WATCHED SURVEY</s> </head>
<head type="ticker"> <s type="first">JAPANESE SHARES CONTINUE THEIR RECENT DECLINE, HURT BY WEAK MANUFACTURING DATA FROM CHINA AND FEARS OVER THE US SCALING BACK KEY STIMULUS MEASURES</s> </head>
<head type="ticker"> <s type="first">MARIO DRAGHI, THE HEAD OF THE EUROPEAN CENTRAL BANK (ECB), DEFENDS ITS BOND-BUYING PROGRAMME AHEAD OF A COURT HEARING AGAINST IT</s> </head>
</div2>
<div2 type="sport">

<head type="ticker"> <s type="first">BRIAN O'DRISCOLL WILL CAPTAIN THE BRITISH AND IRISH LIONS AGAINST WESTERN FORCE ON WEDNESDAY, AND MANU TUILAGI ALSO STARTS</s> </head>
<head type="ticker"> <s type="first">BEACH VOLLEYBAL WILL NOT RETURN TO LONDON IN 2013 AFTER FUNDING SHORTFALL SEES AN EVENT CANCELLED</s> </head>
</div2>
</div1>
</date>

As we can notice, we have also decided, in order to reduce the redundancy effect that the section labels would have insulated in the corpus, to code them as follows:

- HEADLINES: <div2 type="news">;
- BUSINESS: <div2 type="business">;
- SPORT: <div2 type="sport">;
- BREAKING: <div2 type="news" specifics="breaking">.

In order to close these strings, the code </div2> was used. Additionally, as we can also see from example (5), in order to identify each news ticker in each section, the mark-up <head type="ticker"> was used (closed by the code </head>). This was done in order to reproduce in the corpus the separation symbol used in the original context in order to signal the end of a news ticker (originally encoded on the TV screen by using the symbol •). Finally, in order to signal sentence boundaries, the code <s type="first"> was used (closed by the code </s>)[54]. While usually, as we have seen in Section 3.1.4, sentence boundaries are signalled by using the XML codes <s> </s>, in the specific case of the NTC corpus we have used a different code because it refers back to Hoey' (2005) and O'Donnell *et al.*'s (2012) concept of textual colligation (see Section 3.3). Indeed, as we will see, encoding this type of information in the NTC corpus (and in the bw_14 corpus) has helped us unveil one of its major peculiarities from a lexicogrammatical point of view, a syndrome (Halliday 2005), so to speak, of the genre's textual characteristics and, more generally, of the journalistic practices of the BBC.

The NTC corpus thus annotated is comprised of different .txt files (Code set: UTF-8), each containing a month long collection of data. The files were then uploaded on the online corpus analysis tool Sketch Engine (Kilgarriff *et al.* 2004; Kilgarriff *et al.* 2014), which was used as our main analysis tool for this

[54] See the paragraphs below for a complete account of the encoding of such string in the NTC and bw_14.

investigation[55]. Sketch Engine also allowed us to automatically encode linguistic information in the NTC thanks to its Penn Treebank tagset (Marcus, Santorini and Marcinkiewicz 1993). Based on the Brown Corpus[56] tagset (181 (both simple and compound) POS tags and 6 tags for punctuation), the Penn Treebank Tagset (36 simple POS tags and 12 tags for punctuation and currency symbols) represents its simplification, since some of the tags used in the Brown Corpus could be reduced "by taking into account both lexical and syntactic information" (Marcus, Santorini and Marcinkiewicz 1993: 314). This can be seen as an advantage in the automatic annotation of a corpus, since the more the annotation tags introduced in a given scheme, the more the tagger is prone to mismatch the tags. Thus, the Penn Treebank tagset for automatic POS-tagging available on the Sketch Engine platform can be considered as an invaluable resource for corpus annotation. However, we must underline that, in the case of the NTC, in order to 'help' the tagger rightfully assign to given words their specific tags, all words have been encoded in small letter case[57]. Additionally, it must be underlined that the Sketch Engine compile tool, which

[55] See Section 4.3 for further information on the corpus analysis software used in this research project.

[56] The Brown University Standard Corpus of Present-Day American English (usually referred to as just the Brown Corpus) was originally compiled by Kučera and Francis (1967) as a general corpus of written American English comprised of 500 samples (roughly 2000 words each for a total number of 1 million words) taken from texts published in the United States in 1961 (or earlier). The corpus was built, as Xiao (2008: 395) underlines, "with comparative studies in mind", that is, as a standard to look up to in order to create corpora that could have been compared to it. And, indeed, its standards have been used widely in the construction of a series of corpora for synchronic or diachronic analyses, or language varieties comparisons.

[57] The NTC corpus was originally collected respecting the actual spelling of the words displayed in the news tickers of the BBC World News, that is, by transcribing all the words in the corpus in capital letters. However, when we first uploaded the NTC on the Sketch Engine platform and run the automatic POS-tagger, we noticed that, for instance, some verbs were annotated as nouns (NN or NP), while others in the simple past form were annotated as adjectives (JJ). We, thus, hypothesised that the tagger was attributing tags of a given genre to the words found in our corpus on the basis of the way they were spelled: since headlines in newspapers are generally written in capital letters and make use profusely of noun phrases and adjectival premodifiers (Bell 1991; Bednarek and Caple 2012a; Isani 2011), the tagger 'thought' it recognised these conventions, additionally in light of the capital spelling of the words. Thus, we have decided to 'help' the tagger by eluding the comparison to such genre, thus, transforming all capital letters in small ones. Our hypothesis was proven right, since the tagger, after this alteration to the corpus, started to assign tags more precisely to the words in the NTC.

automatically assigns POS-tags, also encodes structural elements to the corpus, more specifically, sentence boundaries and paragraph breaks. Since sentence boundaries were already encoded manually in the NTC at the beginning (and end) of each news tickers, we found ourselves with a corpus where at the beginning of a sentence we had the code <s><s> (and at the end of the sentence the code </s></s>), where the first <s> was automatically encoded by the Sketch Engine compile tool, while the second <s> was part of the original structural annotation scheme encoded semi-automatically in the corpus thanks to the use of regular expressions on TextPad, which allowed us to insert at the beginning of each line the code <s> and at the end of each line the code </s> (with each line representing a single news ticker). To better understand this complication in the corpus annotation, we would like to offer an example taken from the original NTC corpus (i.e., the NTC before it was uploaded on the Sketch Engine platform):

(6) <head type="ticker"> <s>JAPAN IS TO INVEST HUNDREDS OF MILLIONS OF DOLLARS INTO BUILDING A FROZEN WALL AROUND THE FUKUSHIMA NUCLEAR PLANT TO STOP LEAKS OF RADIOACTIVE WATER</s> </head>

However, the automatic codification of structural elements offered by Sketch Engine, as we have previously underlined, created the following 'anomalies' in the annotation of the corpus once it was uploaded and automatically annotated by the online platform (the 'anomalies' created in the automatic annotation process have been emphasised in bold type):

(7)
<head type="ticker">
<s><s>

japan	NN	japan-n
is	VBZ	be-v
to	TO	to-x
invest	VV	invest-v
hundreds	NNS	hundred-n
of	IN	of-i
millions	NNS	millions-n
of	IN	of-i
dollars	NNS	dollar-n
into	IN	into-i
building	VVG	build-v
a	DT	a-x
frozen	JJ	frozen-j
wall	NN	wall-n
around	IN	around-i
the	DT	the-x
fukushima	NN	fukushima-n
nuclear	JJ	nuclear-j

plant	NN	plant-n
to	TO	to-x
stop	VV	stop-v
leaks	NNS	leak-n
of	IN	of-i
radioactive	JJ	radioactive-j
water	NN	water-n
</s></s>		
</head>		

While the repeated codification of structural elements was initially seen as a drawback in originally encoding this information semi-automatically in the corpus before uploading it on Sketch Engine, it however became a useful resource in distinguishing text-initial from non-text-initial sentences (Hoey 2005; O'Donnell *et al.* 2012). Indeed, those sentences that began with a double <s><s>, since they were encoded as representing the very first sentence of a given paragraph, were substituted thanks to TextPad with the code <s type="first">, while all sentences starting with a single <s> were substituted with the code </s><s type="other">[58] (the double repetition of the closing element </s></s> was simply substituted by a single </s>):

[58] The closing element </s> in the string </s><s type="other"> was used in order to close the string <s type="first">. Thus, as we can see, by using regular expressions and 'anchorage points' in the codification of the NTC, we have been able to accurately encode given information in the corpus.

(8)

```
<head type="ticker">
<s type="first">
japan                NN          japan-n
is                   VBZ         be-v
to                   TO          to-x
invest               VV          invest-v
hundreds             NNS         hundred-n
of                   IN          of-i
millions             NNS         millions-n
of                   IN          of-i
dollars              NNS         dollar-n
into                 IN          into-i
building             VVG         build-v
a                    DT          a-x
frozen               JJ          frozen-j
wall                 NN          wall-n
around               IN          around-i
the                  DT          the-x
fukushima            NN          fukushima-n
nuclear              JJ          nuclear-j
plant                NN          plant-n
to                   TO          to-x
stop                 VV          stop-v
leaks                NNS         leak-n
of                   IN          of-i
radioactive          JJ          radioactive-j
water                NN          water-n
</s>
</head>
```

This procedure was particularly fruitful in the annotation of the reference corpus bw_14, since in the NTC no cases of `<s type="other">` were found. This was due to the fact that each ticker represents a single complete sentence and, thus, no non-initial elements were highlighted in the automatic annotation of the corpus. Things were different, however, for the reference corpus, whose annotation scheme is to be introduced in the following paragraphs.

As we have previously seen, the bw_14 corpus was collected by downloading all the news stories published on the BBC website from June 1, 2014 to July 31, 2014, thanks to the online database LexisNexis. The corpus thus collected was, firstly, saved in different .txt files (Code set: UTF-8), each containing a daily long collection of data. The corpus was then additionally cleaned in order to eliminate all the unnecessary information added by LexisNexis (i.e., the name of the source where the data were collected (BBC News); the number of news stories retrieved in a given time span and the identification number for each news story (e.g., 1 of 207 DOCUMENTS, 2 of 207 DOCUMENTS, etc.); the number of words in a given news story; etc.). But, more importantly, since the purpose of our research project was that of analysing what happens linguistically to news contents when they migrate from one media platform to the other, thus, allowing us to underline the peculiarities of the

genre under investigation, we have decided also to focus only on the headlines and the lead paragraphs of the news stories downloaded from LexisNexis. Indeed, as we will see in Chapter 5, since the nucleus (White 2000) and news tickers fulfil similar purposes (i.e., summarising and framing the event, for instance), we have thus decided to use as a reference corpus a collection of data that would have provided us with a clear picture of what linguistic characteristics the nucleus has in the online environment of the BBC, thus, allowing us to consequently see, in such comparison, what peculiarities could be highlighted in a genre that shares with it similar purposes.

Once the bw_14 corpus was cleaned of all the unnecessary 'noise' and, so, 'calibrated' in order to represent the purpose of our investigation, the corpus looked as we can see in the following example:

(9) `http://www.bbc.co.uk/news/uk-27653861`

Pension plan set for Queen's Speech

The government is expected to unveil plans to allow pension providers in the UK to pool investments into shared funds for the first time. Under the changes - set to be unveiled in this week's Queen's Speech - workers would be able to pay into Dutch-style "collective pensions".

We have, then, moved on to the corpus annotation, thus, encoding metadata about the date when the news story was published on the BBC website (i.e., `<date value="2014-06-01"> </date>`) and the 'location' of the news story on the BBC website, retrievable from the URL (Uniform Resource Locator) of the news story[59]. Additionally, in order to distinguish headlines from lead paragraphs, the former were annotated as `<head type="main"> </head>`, while the latter as `<p type="lead"> </p>`. Finally, sentence boundaries were also encoded in the corpus as `<s> </s>`.

[59] Looking, for instance, at example (9), if the URL of the news story is `http://www.bbc.co.uk/news/uk-27653861`, the protocol and the hostname of the BBC news website (`http://www.bbc.co.uk/`) was discarded as metadata, while the file path syntax, comprised of the information about the section were the news story was originally published (in the case of the URL in example (9), the directory `/news/`), the information about the geographical area the news story focused on (again, in the case of the URL in example (9), the directory `/uk`), and the identification number of a given news story (in the case of the URL in example (9), the identification code `27653861`), were encoded as follows: `<div1 id="27653861"> </div1> <div2 type="news" specifics="uk"> </div2>`

After this preliminary semi-automatic annotation carried out by using TextPad, the bw_14 looked as we can see in the following example:

(10)
```
<date value="2014-06-01">
<div1 id="27653861">
<div2 type="news" specifics="uk">
<head type="main"> <s>Pension plan set for Queen's Speech</s> </head>
<p type="lead"> <s>The government is expected to unveil plans to allow pension providers in the UK to pool investments into shared funds for the first time. Under the changes - set to be unveiled in this week's Queen's Speech - workers would be able to pay into Dutch-style "collective pensions".</s> </p>
</div2>
</div1>
</date>
```

The corpus thus annotated was then uploaded to the Sketch Engine platform. However, as we have previously underlined, the following 'anomalies' presented in the additional automatic codification of structural elements by Sketch Engine (the 'anomalies' created in the automatic annotation process have been emphasised in bold type):

(11)
```
<date value="2014-06-01">
<div1 id="27653861">
<div2 type="news" specifics="uk">
<head type="main">
<s><s>
Pension                 NN      pension-n
plan                    NN      plan-n
set                     VVN     set-v
for                     IN      for-i
Queen                   NP      Queen-n
<g/>
's                      POS     's-x
Speech                  NP      Speech-n
</s></s>
</head>
<p type="lead">
<s><s>
The                     DT      the-x
government              NN      government-n
is                      VBZ     be-v
expected                VVN     expect-v
to                      TO      to-x
unveil                  VV      unveil-v
plans                   NNS     plan-n
to                      TO      to-x
allow                   VV      allow-v
pension                 NN      pension-n
providers               NNS     provider-n
in                      IN      in-i
the                     DT      the-x
UK                      NP      UK-n
to                      TO      to-x
pool                    VV      pool-v
investments             NNS     investment-n
into                    IN      into-i
shared                  JJ      shared-j
funds                   NNS     fund-n
for                     IN      for-i
```

the	DT	the-x
first	JJ	first-j
time	NN	time-n
<g/>		
.	SENT	.-x
</s>		
<s>		
Under	IN	under-i
the	DT	the-x
changes	NNS	change-n
-	:	--x
set	VV	set-v
to	TO	to-x
be	VB	be-v
unveiled	VVN	unveil-v
in	IN	in-i
this	DT	this-x
week	NN	week-n
<g/>		
's	POS	's-x
Queen	NP	Queen-n
<g/>		
's	POS	's-x
Speech	NP	Speech-n
-	:	--x
workers	NNS	worker-n
would	MD	would-x
be	VB	be-v
able	JJ	able-j
to	TO	to-x
pay	VV	pay-v
into	IN	into-i
Dutch-style	NP	Dutch-style-n
"	''	"-x
<g/>		
collective	JJ	collective-j
pensions	NNS	pension-n
<g/>		
"	''	"-x
<g/>		
.	SENT	.-x
</s></s>		
</p>		
</div2>		
</div1>		
</date>		

As we have previously said, since sentences starting with <s><s> are to be considered as paragraph initial sentences, while those followed by the code <s> (automatically encoded by Sketch Engine in recognising the opening of a new sentence) represent non-initial text elements, after downloading once more the corpus thus annotated from the Sketch Engine platform, we have used TextPad in order to substitute all the <s><s> with the code <s type="first">, and all the <s> with the code </s><s type="other"> (while all the sentence XML closing tags </s></s> were substituted with a single </s>), as we can see in the following example:

(12)

```
<date value="2014-06-01">
<div1 id="27653861">
<div2 type="news" specifics="uk">
<head type="main">
<s type="first">
Pension                         NN      pension-n
plan                            NN      plan-n
set                             VVN     set-v
for                             IN      for-i
Queen                           NP      Queen-n
<g/>
's                              POS     's-x
Speech                          NP      Speech-n
</s>
</head>
<p type="lead">
<s type="first">
The                             DT      the-x
government                      NN      government-n
is                              VBZ     be-v
expected                        VVN     expect-v
to                              TO      to-x
unveil                          VV      unveil-v
plans                           NNS     plan-n
to                              TO      to-x
allow                           VV      allow-v
pension                         NN      pension-n
providers                       NNS     provider-n
in                              IN      in-i
the                             DT      the-x
UK                              NP      UK-n
to                              TO      to-x
pool                            VV      pool-v
investments                     NNS     investment-n
into                            IN      into-i
shared                          JJ      shared-j
funds                           NNS     fund-n
for                             IN      for-i
the                             DT      the-x
first                           JJ      first-j
time                            NN      time-n
<g/>
.                               SENT    .-x
</s>
<s type="other">
Under                           IN      under-i
the                             DT      the-x
changes                         NNS     change-n
-                               :       --x
set                             VV      set-v
to                              TO      to-x
be                              VB      be-v
unveiled                        VVN     unveil-v
in                              IN      in-i
this                            DT      this-x
week                            NN      week-n
<g/>
's                              POS     's-x
Queen                           NP      Queen-n
<g/>
's                              POS     's-x
Speech                          NP      Speech-n
-                               :       --x
workers                         NNS     worker-n
would                           MD      would-x
be                              VB      be-v
able                            JJ      able-j
to                              TO      to-x
pay                             VV      pay-v
into                            IN      into-i
Dutch-style                     NP      Dutch-style-n
"                               ''      "-x
```

```
<g/>
collective                  JJ      collective-j
pensions                    NNS     pension-n
<g/>
"                           ''      "-x
<g/>
.                           SENT    .-x
</s>
</p>
</div2>
</div1>
</date>
```

Before uploading once more the NTC and the bw_14 on Sketch Engine after this codification improvements, in order to verify the well-formedness of the XML structure of the two corpora, Oxygen XML Editor (Syncro Soft 2015) was used. The software, thus, highlighted those elements that were not well-formed in the codification of the XML structure of the two corpora, helping us validate the XML syntax rules encoded in the corpora under investigation.

4.3 Analysing the NTC and the bw_14 corpora: Introducing Sketch Engine and WordSmith Tools

As we have previously seen, one of the main tools used in annotating the corpus has been the Sketch Engine online platform for corpus analysis. The Sketch Engine, firstly launched in 2004, was initially developed by Kilgarriff and his team (Kilgarriff *et al.* 2004) in order to create word sketches, that is, "summaries of a word's grammatical and collocational behaviour" (Kilgarriff *et al.* 2004: 105). However, during the years, it has slowly become one of the main software tools used by corpus linguists in order to create, query, and manage their data or take advantage of the online pre-loaded and available corpora on the Sketch Engine web service. Its success is mainly linked to the fact that, while corpus tools such as WordSmith Tools (Scott 2014) and AntConc (Anthony 2014) can be used only on a personal computer, Sketch Engine, thanks to its online platform, can be used wherever and whenever it is necessary, as long as an internet connection is available. In this sense, corpora built by researchers can be accessed online, without the corpus file actually being on the computer used to carry out a given investigation.

Among the main functions that the online software offers, we can start with the already-mentioned word sketch (see Figure 8), "a one-page summary of a word's grammatical and collocational behaviour" (Kilgarriff *et al.* 2014: 9).

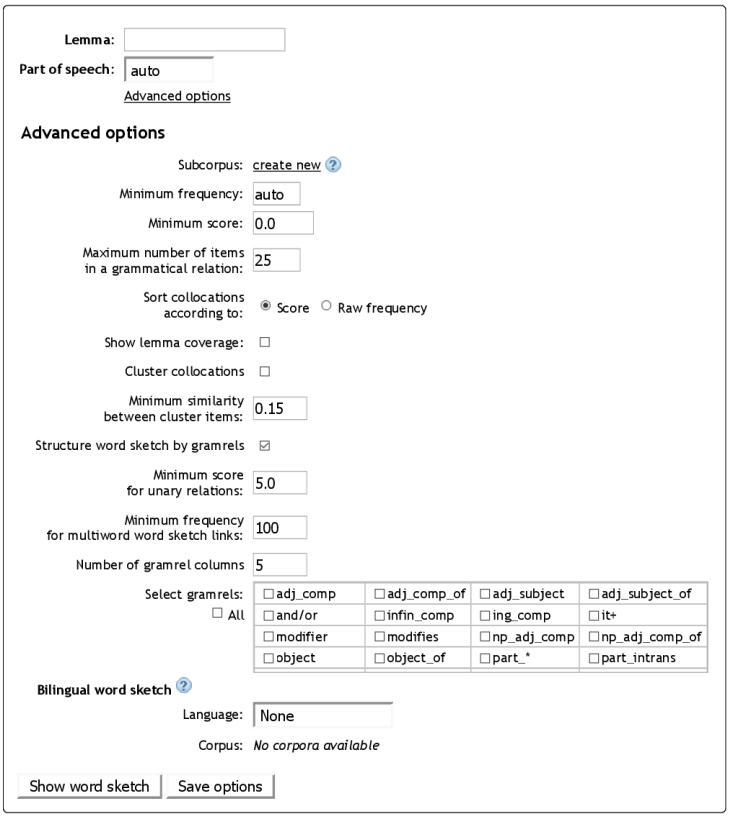


Figure 8 The word sketch tool on the Sketch Engine online platform.

This tool can be seen as a dictionary entry of a given word based on the linguistic context of the corpus that we are analysing. Indeed, as Kilgarriff *et al.* (2014) explain, the software "has worked its way through the corpus to find all the recurring patterns for the word and has organised them, ready for the lexicographer to edit, elucidate, and publish" (Kilgarriff *et al.* 2014: 10; see Figure 9 for an example created by using the BNC, one of the corpora made available on the Sketch Engine service).

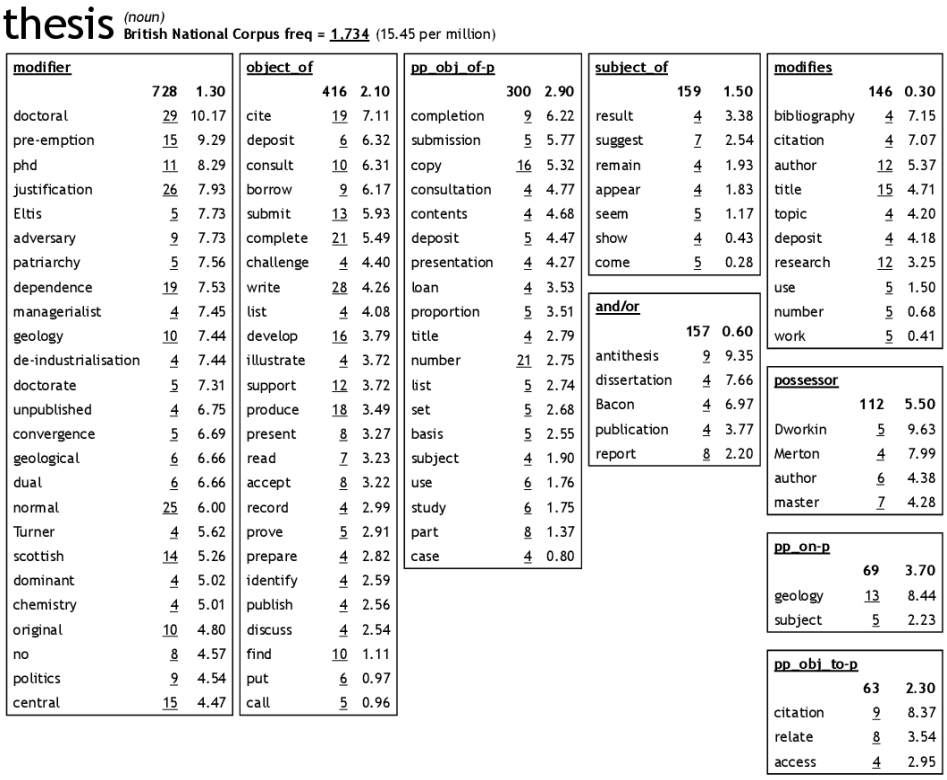


Figure 9 An example of word sketch generated on the Sketch Engine platform by using the BNC in order to see the grammatical and collocational 'surroundings' of the word 'thesis'.

In this sense, the word sketch is particularly useful to those researchers interested in lexicography. However, since this tool reveals the linguistic 'surroundings' of a given word, it can also be fruitfully used in discourse analysis approaches. Indeed, since, as Firth (1957) maintains, "[c]ollocations of a given word are statements of the habitual or customary places of that word" (Firth 1957: 181), word sketches can reveal given representations of the world in the meanings associated with given words. In the case of Critical Discourse Studies, this tool can help researchers highlight certain ideologies hidden in the collocational patterns associated with a specific lexical item, therefore, underlining the semantic prosody of a word, that is, "a form of meaning which is established through the proximity of a consistent series of collocates" (Louw 2000: 57).

Sketch Engine also features a concordance tool that can be either accessed in the word sketch and, thus, allowing to look at the concordances of the word at the heart of a word sketch, or it can be accessed by simply clicking on the corpus we

want to query for that particular word. The concordance tool offers two kinds of queries: a basic and an advanced search. The basic search (see Figure 10) is characterised by the following aspects:

- it is case-insensitive (thus, for instance, it searches for ‘cat’, ‘Cat’, ‘CAT’, and so on);
- it searches for either word form or lemma (so, for instance, it searches for ‘think’, ‘thinking’, ‘thinks’, ‘thought’);
- it can search for sequences of words (for instance, it can search for the sequence ‘think that’, if we are interested, for example, in seeing if it occurs more in written or spoken English, if the corpus queried for this sequence is representative of these registers).

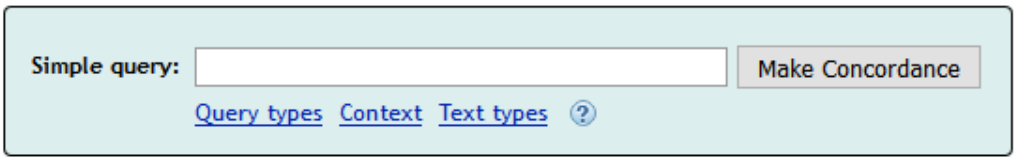


Figure 10 The simple query option on Sketch Engine.

The advanced method (see Figure 11) to show the concordances of a given node can be accessed by clicking on ‘Query types’, where we can choose to search for lemmas (also specifying the word class, e.g., verb, noun, adjective, etc., if the corpus has been previously POS-tagged, for instance, by clicking on the ‘Compile’ option when uploading a given corpus on Sketch Engine) or a phrase or word form. We can also search the corpus by using a CQL query[60].

[60] Developed and formalised at the University of Stuttgart in the early 1990s (Christ and Schulze 1994; Christ 1994), the Corpus Query Language (CQL) can be defined as “[a] general-purpose query language, which treats the whole corpus as a structured knowledge source and allows to express queries involving all knowledge sources declared for a specific corpus (no matter how the knowledge is accessed physically)” (Christ 1994). In other words, CQL allows queries linked to all the information encoded in the corpus, thus, avoiding over-specifications in search of given phenomena. For instance, instead of searching for every adjective in its lexical form in a pre-tagged corpus, researchers can simply insert in the CQL corpus query box on Sketch Engine the string `[tag="JJ"]` and all the adjectives in the corpus will be found. CQL can also allow the implementation of regex in the corpus query.

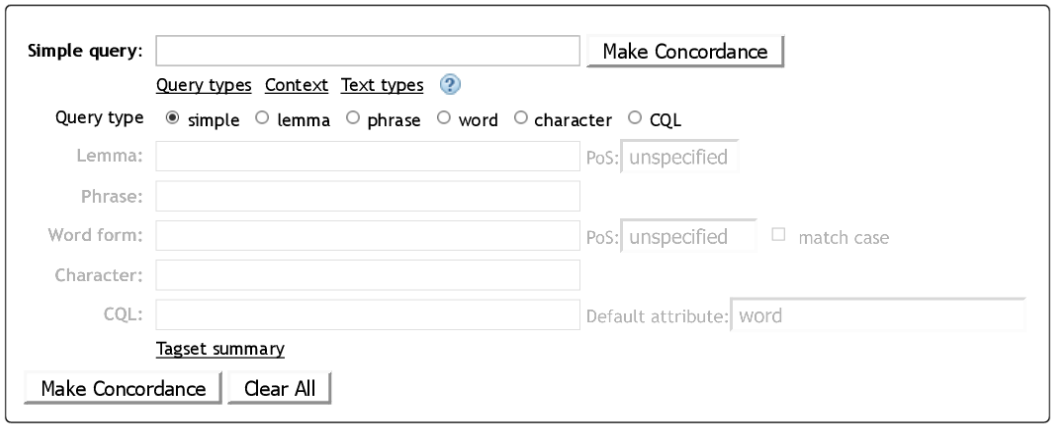


Figure 11 The Query type option in the concordance tool of the Sketch Engine platform, where searches can be carried out according to lemmas, phrases, word forms, characters, or by using CQL queries.

Additionally, both the basic and advanced concordance search tools on Sketch Engine allow researchers to investigate a particular pattern of use thanks to the Context options (see Figure 12). Thanks to these, in a concordance search, it is possible to specify the number of words to the left and to the right (or only to the left and/or only to the right) of the node to be displayed; search for a specific lemma co-occurring with the node word; and, finally, specify the particular grammatical role fulfilled by the words surrounding the node thanks to the POS filter.

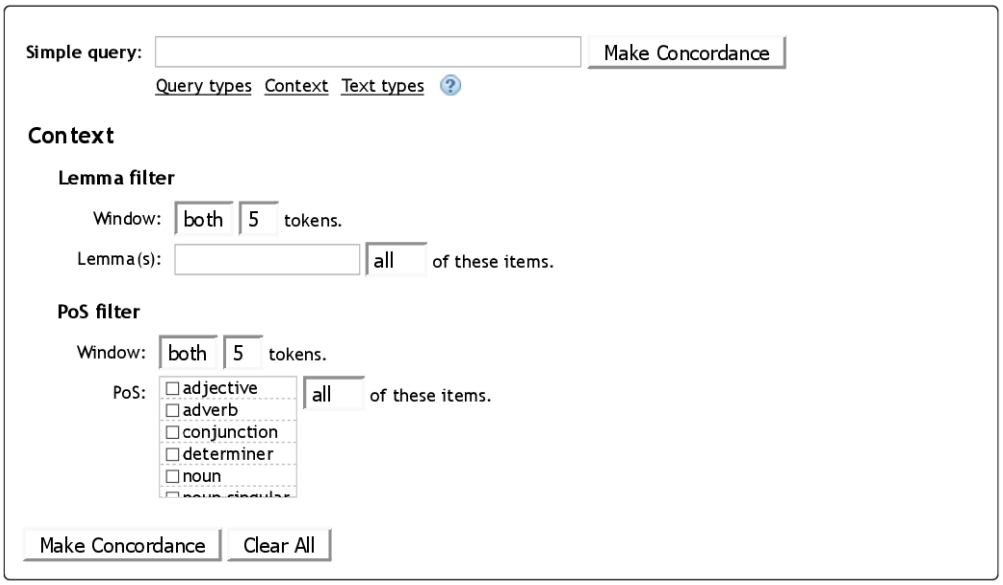


Figure 12 The Context options available in the concordance tool on Sketch Engine.

Moreover, if the corpus has been built in such a way that it includes metadata, researchers can also search for a given word in a specific 'place' of the corpus by using the Text type option (see Figure 13).

Figure 13 The Text Type option available on Sketch Engine in order to search specific 'places' of the corpus, if metadata have been encoded. The figure shows the possibility to limit a specific search to a given section encoded in the NTC.

When concordance lines have been generated, users can further investigate the various instances by sorting, filtering, sampling (by Context and Text types, for instance) or saving them (see Figure 14).

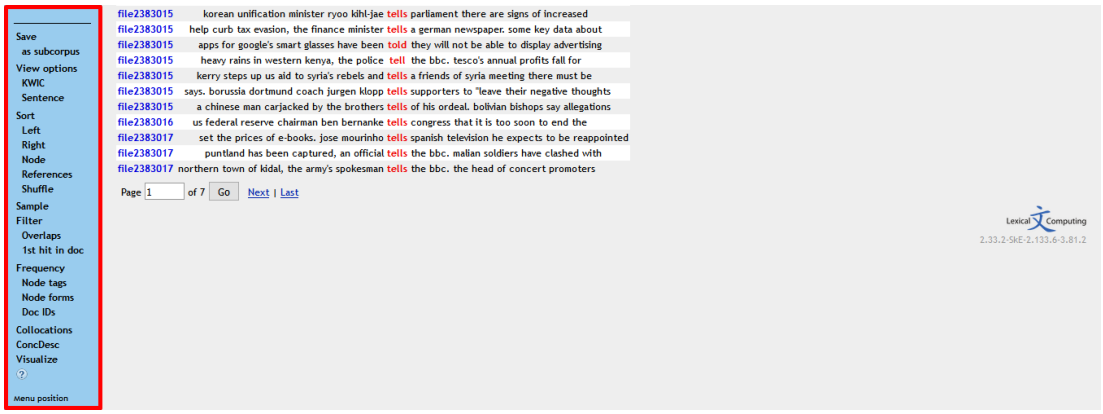


Figure 14 The menu on the left of the concordance lines (highlighted in the figure in red) can help researchers sort, filter or save the results of a particular search.

The menu on the left-hand side of the concordance lines can also allow researchers to access the Collocation desk, where collocations of the node word can be calculated by using different types of statistical formulae (see Figure 15).

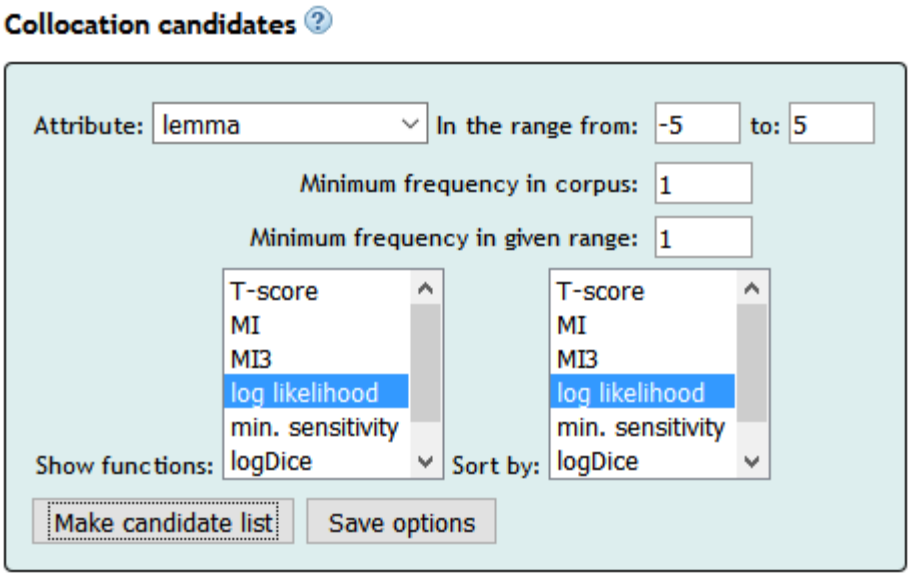


Figure 15 The collocation desk accessed through the left-side menu of the concordance lines on Sketch Engine, allowing researchers to calculate a candidate list of the words occurring with the node word according to specific statistical formulae.

Back to the concordance search results, if necessary, users can easily and immediately get more context by clicking on the individual hits (see Figure 16) or can look at the metadata associated with that particular concordance line by clicking on the ‘reference’ column on the left-hand side of the concordance line (see Figure 17).

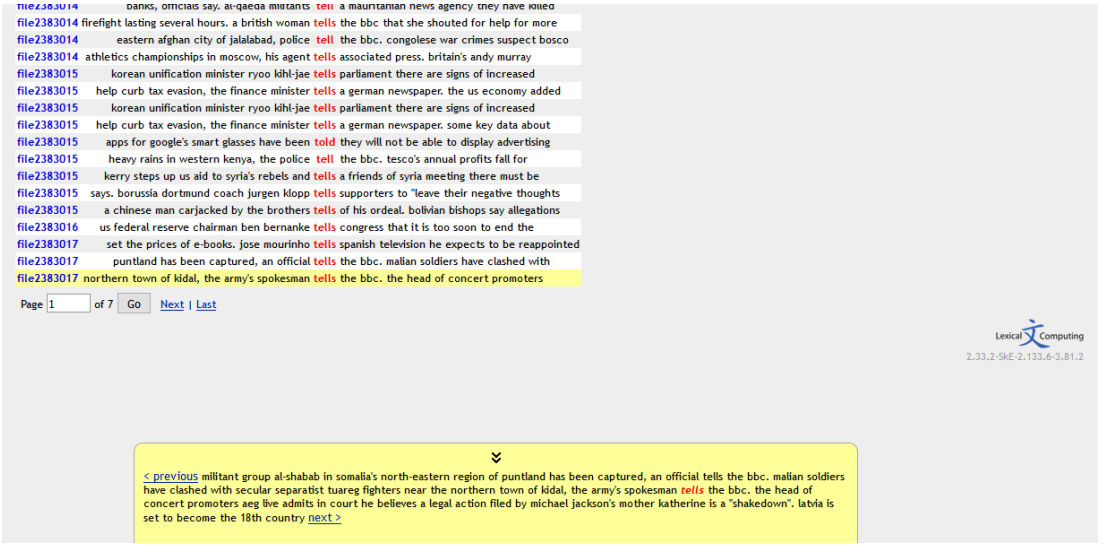


Figure 16 If researchers want to further investigate the linguistic context of a node word, they can do so by clicking on the individual hits and a window will appear at the bottom of the page displaying the paragraph where the node word occurred.

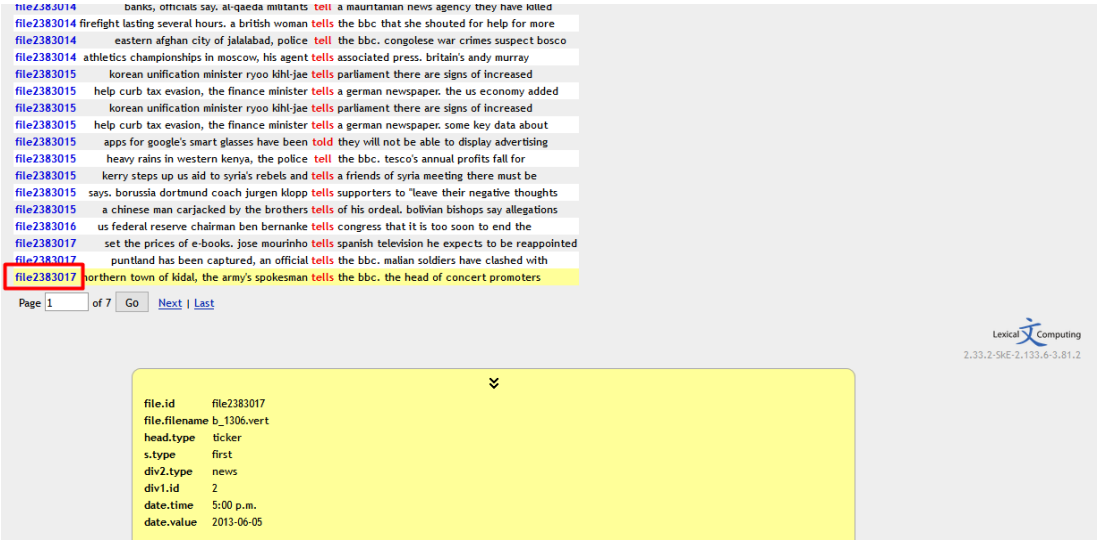


Figure 17 The reference column on the left of the concordance lines allows researchers to see the metadata linked to that particular occurrence of the node word.

Another important tool available on Sketch Engine is the Thesaurus (see Figure 18) that, given a certain lemma (whose POS-tag can be further specified), offers users a ranked list of the lemmas most similar, in terms of their grammatical and collocational behaviour, to the lemma entered (Kilgarriff *et al.* 2004).

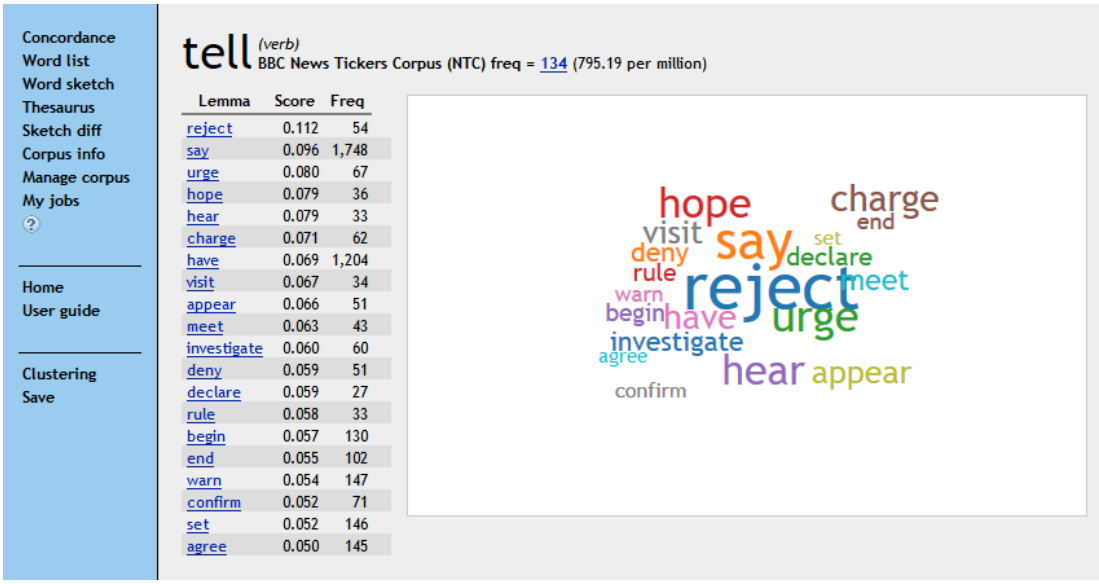


Figure 18 An example, created by using the NTC and by searching for the lemma 'tell', of the Thesaurus tool available on Sketch Engine.

Finally, but more importantly in the context of our investigation, another useful tool found on the Sketch Engine platform is the Word list feature (see Figure 19), which allows researchers to create a list of all the words, lemmas, tags, and other attributes from a corpus or a subcorpus. It further allows the creation of a keyword list, that is, a list of words, lemmas, tags, etc. that are significantly common or peculiar in the corpus/subcorpus under investigation in the comparison with a reference corpus (see Section 5.2).

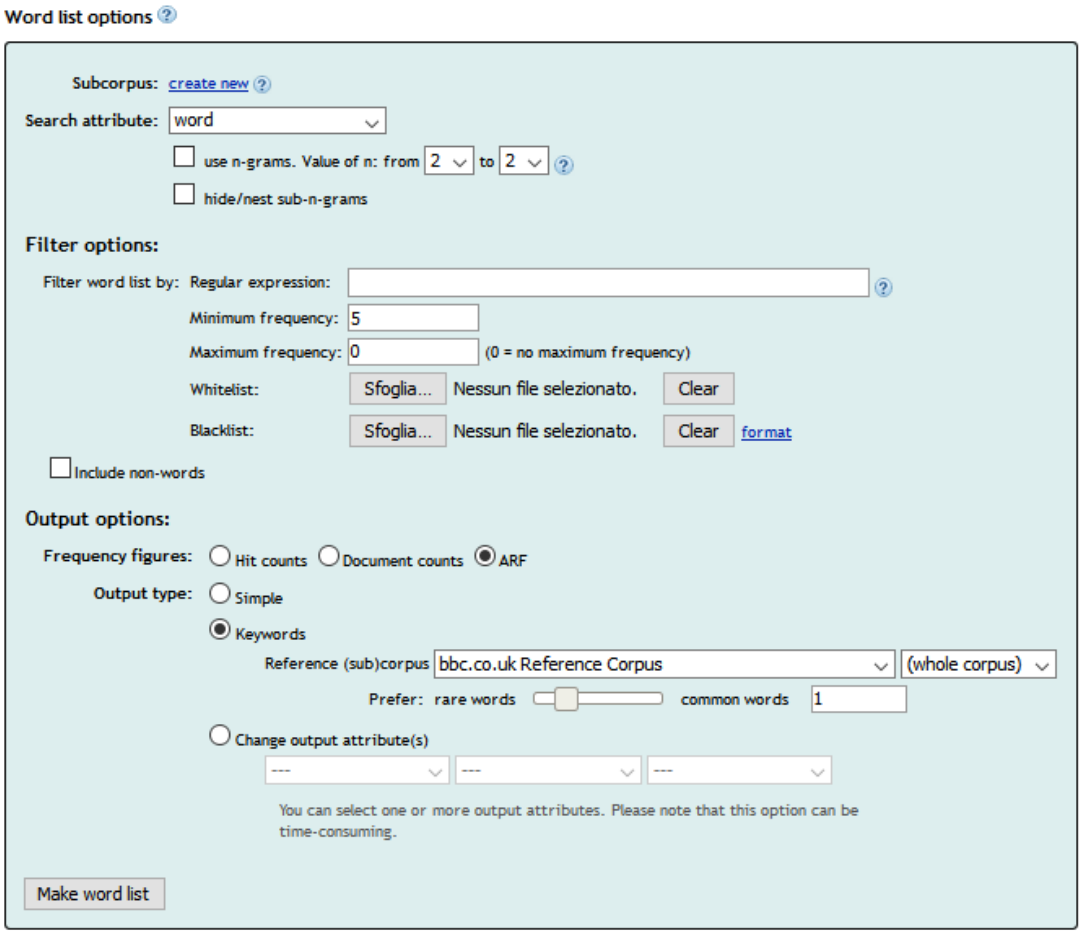


Figure 19 The Word list tool available on the Sketch Engine platform.

As we have seen in this brief introduction, the Sketch Engine service provides an extensive array of features that can be used in order to investigate a corpus. However, our selection as a main corpus linguistic tool in order to investigate our data is particularly linked to the fact that it also provides online access to a vast

number of invaluable corpora, which have been used, in our specific case, to check if a given trend discovered in our corpus was also occurring in similar genres. In particular, as we will see in Chapter 5, the use of the Siena-Bologna, Portsmouth corpus (from now on referred to as the SiBol/Port corpus) was particularly fruitful for our investigation.

Available on the Sketch Engine platform, the SiBol/Port corpus is a collection of up-market British print newspapers, consisting of 787,000 newspaper articles taken from *The Times*, *The Guardian*, *The Daily Telegraph*, *The Sunday Times*, and *The Sunday Telegraph* (amongst others) from the years 1993, 2005, and 2010. More specifically, the corpus as found on the Sketch Engine service, at the time of writing, consists of[61]:

- SiBol 93, which contains the entire output from 1993 of *The Guardian*, *The Times*, *The Sunday Times*, *The Telegraph*, and *The Sunday Telegraph*;
- SiBol 05, which comprises the entire output from 2005 of *The Guardian*, *The Times*, *The Sunday Times*, *The Telegraph*, *The Sunday Telegraph*, and *The Observer*;
- Port 2010, which contains the entire output from 2010 of *The Guardian*, *The Times*, and *The Telegraph*.

As previously said, such corpus was used in order to see if a specific phenomenon occurring in the NTC was also consistent in other news companies or if it was statistically significant in the corpus under investigation.

Despite the advantages so far highlighted in using the Sketch Engine online platform for corpus linguistic investigations, we have nonetheless also initially used WordSmith Tools (Scott 2014), specifically in order to calculate the type/token ratio (TTR) of the NTC. This measure was preliminary used in the case of our investigation in order to test the vocabulary richness of the NTC. Indeed, this statistical measure was firstly introduced by psychologists and researchers in

[61] Further information on the SiBol/Port corpus can be found online at `http://www.lilec-clb.it/?page_id=8`

communication in order to study the lexical richness of given texts (Biber 1988)[62]. The type/token ratio is calculated by simply dividing the number of types in a corpus by the number of tokens (and, if we want to express the result as a percentage, we can then multiply the result by 100). The closer the ratio is to 1 (or, in the case we have expressed it as a percentage, to 100), the more varied the vocabulary is.

Since Sketch Engine does offer the possibility to look at the total number of types and tokens in a specific corpus by accessing the 'Corpus info' in the left-hand menu of the Concordance tool, the use of WordSmith Tools could have been avoided. However, as Baker *et al.* (2006) highlight, this measure of lexical diversity should be used with caution, since "the larger the corpus or file is, the lower the type/token ratio will be" (Baker *et al.* 2006: 162). This is particularly due to the fact that, the bigger the corpus, the higher the repetition will be, for instance, of function words, thus, skewing the TTR. And this is the reason why the use of the raw TTR is advisable only when dealing with small corpora[63], while it should be avoided as a contrastive measure between corpora of different sizes when it comes to lexical richness (Scott 2015).

A possible solution to the problem of calculating the TTR when dealing with texts of different lengths was firstly introduced by Biber (1988), who advices to calculate it only on the basis of "the number of types in the first 400 words of each text, regardless of the total text length" (Biber 1988: 239). However, the author does not offer any particular reason for the cut-out point of 400 words.

Thus, a better solution is offered in the statistics menu of the WordList feature of WordSmith Tools, where a standardised type/token ratio (STTR) is automatically computed. As Scott (2015) explains, the STTR is calculated by every *n* words (by

[62] As Biber (1988) highlights, the type/token ratio was particularly used to study linguistic differences between spoken and written productions. In particular, Johnson *et al.* (1944) used the TTR in order to study differences between educational levels, telephone vs. ordinary conversations and, more specifically, differences in the spoken productions of schizophrenic patients compared to those of freshmen at the University of Iowa. Additionally, based on the hypothesis that language behaviour under increased drives becomes stereotypical, Osgood and Walker (1959) used the TTR in order to compare suicide notes to ordinary letters written to relatives, friends, etc. The statistical measure thus proved that "[s]uicide notes display greater stereotypy – the writer of a suicide note tends to use shorter, simpler words, his vocabulary is less diversified, he is more repetitious" (Osgood and Walker 1959: 66).

[63] As Baker (2006) points out, the raw type/token ratio can be particularly useful only "when looking at relatively small text files (say under 5,000 words)" (Baker 2006: 52).

default, *n* = 1,000, but it could be changed in the WordList Main Controller Settings) as WordList goes through each text file. In other words, the type/token ratio is initially computed for the first 1,000 running words and, then, calculated once more for the next 1,000, and so on to the end of the corpus. Once this operation is completed, an average type/token ratio based on this procedure is computed. The STTR thus calculated is finally displayed as a percentage, showing the amount of new types for every *n* (in the default case, 1,000) tokens.

The standardised type/token ratio calculated by following this procedure can thus be used in order to compare the lexical complexity or specificity of corpora of different sizes. Or, as Baker (2006) argues, "[a] low type/token ratio is likely to indicate that a relatively narrow range of subjects are being discussed, which can sometimes (but not always) suggest that the language being used is relatively simplistic" (Baker 2006: 52). However, as we have previously underlined when introducing the notion of specificity, a low type/token ratio can also be symptomatic of a certain formulaic language adopted in a given corpus, as we can see in the following figure (Figure 20), showing a comparison between the STTR computed for the NTC[64] and the two components of the bw_14, the bwh_14 (listing all the headlines collected in the bw_14) and the bwl_14 (listing all the lead paragraphs collected in the bw_14):

[64] In order to calculate the STTR of the NTC, we have decided to compute it, firstly, for the first round, second round, and new format of news tickers displayed from December 18, 2013 on, separately (see Chapter 5 for further details on the different formats of news tickers developed by and displayed on the BBC World News during the period of collection of our data). We have then proceeded to the calculation of the average percentage of the STTR computed for these three formats of news tickers. We have decided to follow this procedure because, since the second round of news tickers repeats some of the news stories displayed in the first round of tickers, this has allowed us to make sure that identical news stories did not skew the STTR.

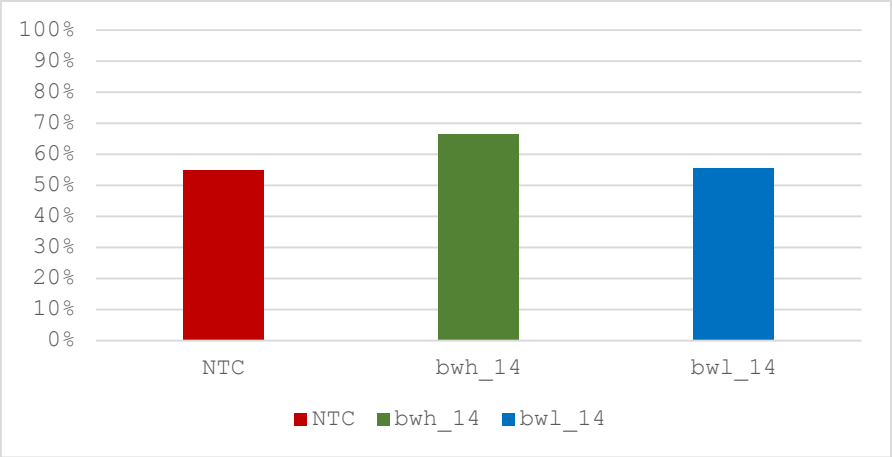


Figure 20 The STTR computed thanks to the use of WordSmith Tools (Scott 2014) of the NTC and the two components of the bw_14.

As we can see from Figure 20, the highest STTR is unsurprisingly displayed in the headline component of the bw_14 (66%). This is due to the fact that, since as a genre headlines are prone to the omission of functional/grammatical words (e.g., determiners, auxiliaries, etc.; Mardth 1980; Bell 1991; Isani 2011; Bednarek and Caple 2012a; Chovanec 2003, 2014), there is a certain tendency towards the use of lexical words, thus, reducing the repetition of certain items, something that is further enhanced by the fact that the bw_14 contains a wide variety of news stories focusing on different events (and, thus, lowering the possibility that given content words are repeated in the corpus).

As for the lead paragraph component of the bw_14, the lower STTR (56%), when compared to that of the headline component, is also unsurprising. Expanding on the information firstly provided in the headline, lead paragraphs are 'micro-stories', "packed with information and news appeal" (Bell 1991: 176), and enhancing the news values of newsworthiness, brevity, and clarity. Thus, while lead paragraphs do not show a tendency to the omission of functional/grammatical words, they do show a formulaic linguistic routine, orienting readers towards the news story by introducing the main actors, the where and the when. Indeed, while the main actors and other information providing the 'hook' of the news story (Cotter 2010) may vary, since they are circumstantially linked to the news story itself, the linguistic resources in introducing them may be repeated and, thus, this explains the lower

STTR when compared to headlines. However, the STTR of the bwl_14 is not that low to argue that there is truly a strict linguistic formulaic way of presenting news stories in lead paragraphs: it can merely be a 'symptom' in the professional practice of the BBC of some underlying linguistic patterns in presenting the news stories in lead paragraphs.

Finally, as for the NTC, the percentage computed on the WordList statistics section of WordSmith Tools displays an STTR similar to that of the bw_14 lead paragraph component (55%). Such a similarity can be a first insight into the nature of news tickers in the environment of the BBC World News. Indeed, from this preliminary observation offered by looking at the standardised type/token ratio, it would seem that news tickers, with regard to lexical complexity, are similar to lead paragraphs, thus, focusing on a wide range of topics, nonetheless employing given formulaic linguistic structures, that is, specific linguistic routines that will be explored more accurately in Chapter 5 of this contribution.

We must, however, promptly underline, as Baker (2006) points out, that "the type/token ratio merely gives only the briefest indication of lexical complexity or specificity" (Baker 2006: 52). In other words, arguing for the complexity or specificity of the vocabulary of a given corpus by only using a statistical measure such as the TTR (or the STTR) can be seen as staying inside and assuming that it is raining by simply looking outside and seeing a cloudy sky. In this sense, researchers need to go outside and get their 'feet wet' in order to corroborate the initial assumptions on the vocabulary richness of their data offered by the TTR. And, hence, further investigations are needed of the data collected in order to prove this claim right (or to prove it wrong).

So far, this chapter has focussed mainly on a step-by-step description of the procedure followed in collecting the corpus under investigation and the tools used in preparing the data to be analysed. However, as we have previously argued, corpus linguistic methodologies, in our view, are simply tools used to (semi)automatically investigate given phenomena in a corpus, and underline linguistic patterns that intuition alone may not identify. Thus, it is now necessary to implement this approach with the methodologies used in order to investigate our data. In this sense, we would like to clarify that, even though the approaches that have been introduced

in Chapter 3 may be referred to as a ‘theoretical framework’ used in analysing the data under investigation, we prefer to refer to them as methodologies, since we do believe in the freedom of researchers to ‘use’ given insights coming from different sources without ‘pledging their allegiance’ to a specific theoretical tradition.

In line with this view, in the following chapter, by combining corpus linguistic methodologies and the framework of analysis introduced in Chapter 3, we will proceed to our corpus-based genre investigation of news tickers.

5. Where the crawling things are

A corpus-based genre analysis of news tickers

As we have previously seen in Chapter 2, news tickers represent a relatively new genre in the context of TV news broadcasts and, while their evolution from their first live appearance in the 1920s has demonstrated their continuous re-shaping of the purposes they were meant to serve, once they migrated on TV news channels at the very beginning of the 21st Century, they have generally been used in order to offer viewers a daily summary of the major news stories covered by a specific TV news channel or to announce a breaking news story and its live development. While the tendency in the literature is to usually dismiss this genre by simply comparing it to other textual realisations found in the context of media discourse, in this Chapter we will try to offer an in-depth analysis of the genre in the context of the BBC World News channel.

Thus, we will firstly start with a brief description of the formal structure of the crawlers as displayed on this TV news channel. Then, thanks to corpus linguistic methodologies, we will focus on the linguistic patterns found in the NTC corpus, which can be ascribed to forms of hybridity and genre mixing in the context of news tickers. Finally, by focusing on the news values enhanced in the genre under investigation, we will try to highlight which strategies are generally used in order to present news stories in news tickers, thus, offering a summary of which values are particularly enhanced by editors at the BBC World News in presenting news stories in this format. This will allow us to further investigate the purposes at the very heart of the genre under investigation, thus, demystifying the core values that fuel the journalistic professional practice of the BBC when developing news tickers.

5.1 Formal structure of the BBC World News' crawlers

Before delving into the linguistic patterns highlighted in the NTC, we will briefly introduce the way news tickers are displayed on the BBC World News. The description of their formal structure is in line with the aims of this contribution, since from this brief account we will start to notice something that will be further enforced in the linguistic realisation of the genre.

From our year-long observation of the news tickers displayed on the BBC World News, we can start by saying that they are usually displayed 24/7 on the news network channel. They are absent only during commercial breaks but also when editors at the BBC want to focus viewers' attention on the video/audio contents presented on the news channel. For instance, during documentaries realised by the BBC, news tickers were generally silenced, thus, confirming what we have previously said about their information interference effect (see Section 2.2).

News tickers are presented as crawling at the bottom of the screen entering from the right-hand to the left-hand corner of the TV screen at an average speed of 11.74 millisecond per character on a video speed of 4 px/f[65]. Each news ticker is comprised of a number of characters that goes from a minimum of 5 characters to a maximum of 205 characters (on average, 84.48 characters per ticker). While each section under which news tickers were presented (see the next paragraphs) was displayed on a white background and written in dark-red capital letters[66], the actual news tickers were displayed on a dark-red background and written in white capital letters (see Figure 21).

[65] In order to measure the speed per each character displayed in the news tickers of the BBC World News, we have used the software Kinovea (available online at `http://www.kinovea.org`), a video software used by athletes in order to analyse their recorded performances. In particular, Kinovea allows users to measure distances and times manually or use semi-automated tracking to follow points and check live values or trajectories. This has allowed us to measure our recordings of news tickers on the BBC World News and, thus, calculate the average speed per character of the crawlers displayed on the news channel.

[66] The only exception is represented by the SPORT section of the news ticker routine that, while initially presented like all the other sections, from April 23, 2013 to December 18, 2013 was displayed on a bright-yellow background and written in black capital letters. This format implicitly and intertextually refers back to the BBC Sport news online platform, whose logo is identical in colour and format to the one found in the section on sport events found in news tickers. See Section 5.2.1 for further information on the overt strategies of promotion employed by the BBC in news tickers.

Figure 21 A screenshot taken from the April 30, 2013 recording of the news tickers displayed during the BBC World News' programme *GMT*.

However, in order to differentiate between every-day news tickers and breaking news tickers, the BBC introduced the latter through the use of the section label BREAKING, displayed on a white background and in bright-red capital letters, while the news tickers introducing the actual breaking news story were displayed on a bright-red background and written in white capital letters (see Figure 22).

Figure 22 A screenshot taken from December 6, 2013 of the news tickers displayed during the breaking news story linked to Nelson Mandela's death.

Moving from the typographic characteristics to their actual organisation, we must promptly say that, during the time-span of our data collection (from March 12, 2013 to April 11, 2014), we can distinguish between two types of textual organisation through which news tickers were displayed. Indeed, due to a major re-shaping of the TV news outlet of the BBC World News, the routine according to which news tickers were displayed up until December 17, 2013 was drastically changed. Thus, in our data, we can distinguish among two news ticker textual organisations.

In the time-span that goes from March 12, 2013 to December 17, 2013, a complete round of news tickers on the BBC World News comprised a quite complicated routine, which we have schematised in example (13). In this schematisation, each news section (i.e., HEADLINES, BUSINESS, MARKETS, CURRENCIES, and SPORT) and closing section (WEBSITE and CONTACT US) will be underlined in order to reproduce how they are highlighted in the BBC news tickers. The use of a symbol (i.e., ●), on the other hand, differentiates each news story from the others (the news stories in this schematisation are exemplified as the section where they are displayed), like in a typical round of news tickers:

(13) HEADLINES
- headline_1
- headline_2
- headline_3
- headline_4
- headline_5
- headline_6

BUSINESS
- business_1
- business_2
- business_3

MARKETS
- DOW
- NASDAQ
- FTSE 100
- DAX
- NIKKEI
- CAC
- HANG SENG
- SINGAPORE STI

CURRENCIES
- £:EURO
- £:$
- £:HKS
- EURO:£

- EURO:$
- $:YEN

SPORT

- sport_1
- sport_2

WEBSITE

MORE ON ALL THESE STORIES AT bbc.com/news

TWITTER

FOR LATEST FOLLOW US VIA @bbcworld AND @bbcbreaking

HEADLINES

- headline_1
- headline_2
- headline_6
- headline_7
- headline_8
- headline_9
- headline_10

BUSINESS

- business_1
- business_4
- business_5

SPORT

- sport_3
- sport_4

CONTACT US

HAVE YOUR SAY AT facebook/bbcworldnews

- WEBSITE: bbc.com/haveyoursay
- EMAIL: haveyoursay@bbc.co.uk
- SEND YOUR VIDEOS TO: whysvideo@bbc.co.uk
- FOR TERMS ON SENDING PICTURES AND VIDEOS: bbc.com/terms

As we can see from this summary of the typical organisation of the BBC news tickers, a complete round of tickers, up until December 17, 2013, was comprised of two parts.

In the first part, which ended with the closing section WEBSITE, five sections were differentiated. In the section HEADLINES, a number of six major news stories were displayed[67]. The HEADLINES section was followed by the BUSINESS section, where three major economic news stories were displayed. The peculiarity of this first part of a complete round of tickers is represented by the sections MARKETS and CURRENCIES, where stock indexes and the value of foreign exchange rates were presented. Finally, this first part ended with the SPORT section, where two major sport updates were shown.

[67] The term 'headlines' is used here to signal that the news stories displayed in this section were the ones introduced by the anchor at the beginning of the news programme. However, this section often featured other news stories that were not announced by the anchor.

Once the closing section WEBSITE scrolled, announcing the end of the first part, a second part was introduced, showing again some of the sections displayed in the first round of tickers. The section HEADLINES of the second round of tickers usually introduced seven major news stories. The first two news stories were the same first two displayed in the first round of tickers (in example (13), `headline_1` and `headline_2`), while the third news story was the last one presented in the section HEADLINES of the first round (in example (13), `headline_6`). These three news stories were followed by further four news updates that had not been previously presented. The section BUSINESS displayed as its first news story the same first one presented in the section BUSINESS of the first round of tickers (in example (13), `business_1`), followed by two new economic updates that had not been previously displayed (in example (13), `business_4` and `business_5`). Finally, the BUSINESS section was followed directly by the SPORT section, where two new sport updates were displayed (in example (13), `sport_3` and `sport_4`).

This quite complicated routine, however, was abandoned on December 18, 2013, when a simpler and much shorter single round of tickers was developed and displayed. Indeed, the MARKET and CURRENCIES sections were dropped and only the section HEADLINES and BUSINESS survived in this new format of news tickers, as we can see from example (14), where a schematisation of this new format can be seen:

(14) HEADLINES
- `healine_1`
- `healine_2`
- `healine_3`
- `healine_4`
- `healine_5`

BUSINESS
- `business_1`
- `business_2`
- `business_3`

MORE NEWS AT bbc.com/news
- facebook.com/bbcworldnews
- TWITTER @bbcworld AND @bbcbreaking

SEND YOUR PICTURES & VIDEOS yourpics@bbc.co.uk

As we can see, in the section HEADLINES, a number of five major news stories are displayed, followed directly by the BUSINESS section, which displays three major

economic news stories. At the end of the BUSINESS section, the message introduced by the section head MORE NEWS announced the end of the round of news tickers. The changes brought about in news tickers on December 18, 2013 are part of a wider change in the BBC World News graphic layout due to its implementation of HD technologies in the broadcasting of news (BBC launches five new HD channels 2013).

In this brief description of the way news tickers are textually organised on the BBC World News channel, we have seen that the presentation of the main news events reported in this genre are distributed according to a strict routine, which places given information in specific sections, thus, providing a further space on the screen of the TV news channel where news is provided to viewers. This textual space, therefore, represents what we can call an ancillary media platform realised at the same time as the audio and video contents are reproduced on the BBC World News. Hence, according to the parameters previously highlighted in defining a genre (see Section 3.2.2), we can rightfully define news tickers as a (sub-)genre found in the context of news production, given the adherence to a strict routine in presenting the news items found in this context.

While some of the elements highlighted in this paragraph will not be further investigated, some exceptions will be made since they are symptomatic of a general trend highlighted both in the textual organisation of news tickers and in their linguistic realisation by given cues. Thus, in the following paragraphs, we will concentrate, thanks to corpus linguistic methodologies, on the linguistic realisation of news tickers, starting with the keyword analysis performed in the comparison between the NTC and the bw_14.

5.2 Keyword analysis of the NTC

As we have previously seen in Section 3.2.1, Tribble's (1999) approach to genre analysis moves from the analysis of the keywords found in a given genre in order to define its stylistic information and, thus, its status as a genre. Therefore, textual patterns highlighted through a keyword analysis of a genre are used in order to define its peculiarities. In accordance with this view, the same procedure has been followed in the case of the NTC, whose keywords have been extracted by comparing it to the

bw_14 (the complete list of keywords extracted from this comparison can be found in Appendix 1). Keywords have been calculated by using the online corpus analysis tool Sketch Engine (see Section 4.3), and they have been calculated by searching lempos attributes, in order to display both the lexical form of the keyword and the part of speech attributed to it in the automatic POS-tagging process of the NTC.

A cut-off point of minimum frequency was also added to the computation of keywords. Indeed, since news tickers, as we have previously seen, up until December 17, 2013, were comprised of two different rounds, where some of the information provided in the first round of tickers were displayed once more in the second round, we have decided to impose a cut-off point of minimum frequency of five occurrences, thus, taking under consideration the fact that some elements are particularly frequent in news tickers not because they are statistically significant in the corpus under investigation but, rather, due to the textual realisation of the genre *per se*.

In order to further ensure that keywords were not due to their particular frequency in given 'places' of the text[68] or, conversely, that given rare phenomena in the reference corpus were to influence the keyword analysis of the NTC, even though they were also infrequent in the latter[69], we have decided to make use of the Average Reduced Frequency (ARF; Savický and Hlavácová 2002), a statistical measure available on Sketch Engine in the calculation of keywords. The ARF allows users to see if keywords are truly 'key' in the comparison between a target and a reference corpus. Indeed, as Kilgarriff (2009) argues, by using the ARF in the extraction of keywords from a corpus, users are able to discount "frequency for words with bursty distributions" (Kilgarriff 2009; see Katz (1996) on the concept of burstiness). Indeed, thanks to this statistical measure, "for a word with an even distribution across a

[68] Katz (1996) refers to this phenomenon as "document-level burstiness" (Katz 1996: 19), which is linked to the "multiple occurrences of a content word or phrase in a single-text document, which is contrasted with the fact that most other documents contain no instances of this word or phrase at all" (Katz 1996: 19).

[69] The same goes for rare phenomena occurring in the NTC and particularly frequent in the bw_14. Since they may have occurred in the NTC only once, they should not be considered as truly 'key' in the keyword analysis and, thus, the use of the Average Reduced Frequency, as we will see, has allowed us to discard these items that are not truly 'key', since they are not dispersed in the genre under investigation. Again, this may be linked to the concept of "burstiness" introduced by Katz (1996).

corpus, ARF will be equal to raw frequency, but for a word with a very bursty distribution, only occurring in a single short text, ARF will be a little over 1" (Kilgarriff 2009). In this way, when using a non-topic-specific corpus, as in the case of the NTC, this measure guarantees that the keywords highlighted are truly indicative of the conventionalised repertoire of the rhetoric of the genre under investigation. This precaution was used in order to ensure that, for instance, in the comparison between the NTC and the bw_14, news stories that were introduced only once in the NTC and/or were present in the NTC and not in the bw_14 were not highlighted in the keyword analysis, since the keyword linked to these specific cases are not indicative of the genre as a whole but, rather, of an absence in the bw_14. Thus, the ARF links together the concept of statistically significant items in corpus comparison and the statistically significant distribution of these items in the target corpus.

The keywords computed according to the methodology previously described have been listed in Appendix 1, where lempos have been organised according to their ARF score in the comparison between the NTC and the bw_14. Additionally, the ARF concerning the sole distribution of the lempos in the NTC and bw_14 has also been included. Finally, also the standardised (per million words) ARF of the sole distribution of the lempos in the NTC and bw_14 has been included.

If we look at the first 20 keywords listed in Appendix 1, we can see that news tickers are particularly concerned with international news, in line with the expectations of the BBC World News worldwide target audience.

A primary role, in the comparison with the bw_14, is given to the Eastern scenario, underlined by the significant frequencies of the lempos `china-n` and `australia-n`, amongst the others. When, however, the BBC World News in its news tickers covers local news stories, it usually focuses on their impact on a national scale, meaning that news stories regarding events in the UK become news only when they involve repercussions on a national scale (underlined by the lempos `england-n`, `Uk-n`, and `uk-n`).

The only exception is represented by the lempos `manchester-n`, but by looking at the concordances of this lempos, we have noticed that it significantly occurs in the SPORT section of news tickers, thus, highlighting the peculiar coverage

in news tickers of news stories linked to the Manchester United football club. However, amongst the first 20 keywords highlighted thanks to the methodology previously described, one in particular has caught our attention and we will further investigate it in Section 5.2.1.

The keyword analysis applied to the NTC in the comparison with the bw_14 has also been performed on the POS-tags automatically encoded in the target corpus. Appendix 2, thus, offers the results of this analysis which, as we will see, will be useful in the description of a particular phenomenon occurring in the NTC. Appendix 2 displays the key POS-tags in the comparison between the NTC and the bw_14, listed according to their ARF score. Additionally, the ARF concerning the sole distribution of each and every single tag in the NTC and bw_14 has also been included. Finally, also the standardised (per million words) ARF of the sole distribution of the tags in the NTC and bw_14 has been included.

As we can see from this further analysis of the NTC but, this time, from a grammatical point of view of the genre under investigation, foreign words (i.e., `[tag="FW"]`, with an ARF score of 4.3 in the comparison with the bw_14) play a significant role in news tickers, thus, further underlining how news stories in the news tickers displayed on the BBC World News focus on international news.

Symbols (i.e., `[tag="SYM"]` with an ARF score of 3.3 in the comparison with the bw_14) also seem to be particularly used in news tickers. This may be due to the fact that, given the need to present the most news stories by using the least amount of characters, symbols may reduce the 'space' occupied in presenting given news stories. However, this may also be indicative of the fact that the news value of facticity (see Section 3.4) is significantly enhanced in news tickers.

Adjectives and superlatives (i.e., `[tag="JJS"]` with an ARF score of 2.0 in the comparison with the bw_14) are also particularly used in the news tickers displayed on the BBC World News. This may be due, again, to reasons linked to the fact that, in the little space that news tickers are assigned to in order to present news stories, a specific stance towards given events must be conveyed in any which way. And, thus, through the use of adjectives and superlatives, the news network ensures that each news story is framed in a specific way.

As for the other POS-tags significantly occurring in the NTC when compared to the bw_14, we will further concentrate on them in Section 5.2.2, since they are indicative of a peculiar phenomenon in the genre under investigation, which seems to be bending specific conventions of traditional genres found in media discourse in order to achieve a peculiar professional purpose. Thus, in order to better highlight and investigate this particular phenomenon of genre bending, we have decided to dedicate a specific section in this contribution to this distinctive feature traced in the comparison between the NTC and the bw_14 (see Section 5.2.2).

However, in order to further corroborate the phenomenon highlighted from the tag keyword extraction in the comparison between the NTC and the bw_14, we have also performed a tag keyword analysis comparing the NTC with the two components of the bw_14: the bwh_14 (see Appendix 3), comprising only the headlines of the news stories collected in the bw_14, and the bwl_14 (see Appendix 4), comprising only the lead paragraphs of the news stories in the bw_14[70].

The keyword lists provided in Appendix 1, Appendix 2, Appendix 3, and Appendix 4, calculated according to the procedure previously described, allow us to highlight some fascinating insights on the nature of the genre under investigation. In particular, two phenomena can be traced in the textual realisation of news tickers. These phenomena can be seen as the evidence of two lexicogrammatical syndromes (Halliday 2005) in the genre of news tickers, highlighting their hybrid nature and the way they bend genre conventions of traditional genres found in the context of media discourse in order to achieve specific goals. Thus, in the following Sections, we are going to comment on the results of the keyword analysis, indicative of the previously highlighted tendencies found in news tickers.

[70] We have decided not to perform a lexical keyword analysis in the comparison of the NTC with the two components of the bw_14 and, thus, extracting the keyword only by comparing the NTC with the whole bw_14 since, in a preliminary investigation, the first five keywords extracted in the comparison between the NTC and the bwh_14 and the bwl_14 were the same as the ones extracted from the comparison between the NTC and the whole bw_14. This may be due to the fact that, from a lexical point of view, the lead paragraph elaborates and expands on the information initially provided in the headline and, thus, the same lexical items presented in the headline can be found repeated in the lead paragraph. Differences, however, in the presentation of the information provided in the headline and lead paragraph can be found at a deeper level, that is, by investigating the grammatical realisation of these items. And this is the reason why a differentiation in the calculation of the tag keyword extraction was performed.

5.2.1 Hybridity and the news tickers: Marketising the news

Hybridity has been defined as the "invasion of the integrity of one genre by another genre or genre convention, often leading to the creation of a hybrid form" (Bhatia 2004: 66). And, in the context of media discourse, Schiller (1986) notices how the "[…] boundaries between news, entertainment, public relations and advertising, always fluid historically, are now becoming almost invisible" (Schiller 1986: 21). In particular, as for the phenomena of hybridisation related to the advertisement of the news network products, Fairclough (1992) refers to these phenomena as processes of marketisation, "[…] whereby social domains and institutions […] come […] to be organised and conceptualized in terms of commodity production, distribution and consumption" (Fairclough 1992: 207).

These phenomena can be traced to Bhatia's (2012) view of interdiscursivity in genres, seen as "a function of appropriation of generic resources across professional genres, practices and cultures" (Bhatia 2010: 33). In the specific case of news tickers, given "the growing ability of viewers to avoid or ignore traditional commercials" (Elliott 2009), TV news networks have found in this genre a subtle way to market their products, "due to the ticker's location at the bottom of the screen, and its format, which does not interrupt programming" (Coffey and Clearly 2008: 896). Thus, in the context of American broadcasting companies, Coffey and Clearly (2008, 2011) have demonstrated this assumption and proposed that news tickers should be regarded as overt promotional agents, that is, as textual elements that openly advertise the news network itself or its programmes (Coffey and Clearly 2008), while no elements of covert promotion, that is, the subtle promotion of "the parent company's media properties" (Coffey and Clearly 2008: 897), were identified in the corpus of news tickers taken from CNN, MSNBC, and Fox News.

In the case of the NTC, overt promotional strategies can be firstly identified in the messages that are used to signal the end of a complete round of tickers:

(15) WEBSITE
MORE ON ALL THESE STORIES AT bbc.com/news
TWITTER
FOR LATEST FOLLOW US VIA @bbcworld AND @bbcbreaking

(16) CONTACT US
HAVE YOUR SAY AT facebook/bbcworldnews
WEBSITE: bbc.com/haveyoursay
EMAIL: haveyoursay@bbc.co.uk
SEND YOUR VIDEOS TO: whysvideo@bbc.co.uk
FOR TERMS ON SENDING PICTURES AND VIDEOS: bbc.com/terms

(17) MORE NEWS AT bbc.com/news
- facebook.com/bbcworldnews
- TWITTER @bbcworld AND @bbcbreaking

SEND YOUR PICTURES & VIDEOS yourpics@bbc.co.uk

All these messages, used in order to signal, in the case of example (15) and (16), the end of the first and second round of tickers respectively, and in the case of example (17), the end of the single round of ticker, can be seen as subtly implying a shift in news promotion, since the BBC does not seem to be treating its Web presence as an advertisement for the offline products (Deuze 2008) but, vice versa, as an extension and implementation of the offline contents. Indeed, the message that closes the first round of tickers (example (15)) is used as a sort of disclaimer, implicitly admitting that tickers do not offer a complete picture of the news stories reported in this textual space and, thus, instead of promoting the offline content itself, it guides viewers towards another platform, away from the TV screen and its 'talking heads'.

Additionally, in example (15), we can notice that, while the official website of the BBC is presented as offering an expansion of the news stories reported in news tickers (realised linguistically in the use of the adverb 'more'), the social network presence of the company is construed in a way that enhances the news value of timeliness (realised linguistically in the adjective 'latest'). This subtly acknowledges the fictitious recency value that TV news stories seem to convey, thus, recognising the limit of the offline environment and, once more, drawing viewers towards an online platform, where actual recency can be achieved in the presentation of news stories.

As for example (16), once more, we can notice a shift in news promotion but, in this case, it subtly plays on synthetic personalisation (Fairclough 1992), that is, "a compensatory tendency to give the impression of treating each of the people

'handled' *en masse* as an individual" (Fairclough 1992: 62). Synthetic personalisation is usually achieved thanks to the use of direct address of audience members with the pronoun 'you' and the use of imperative sentences. These strategies, particularly frequent in advertising discourse, are here used in order to elicit audience participation in the news making process. However, this is simply an illusion, since the power of news selection lies still in the hands of the BBC. Therefore, while, on the one hand, the news network seems to promote and cultivate a relationship with the viewers that shifts from a vertical to a horizontal one, on the other, data sent to the BBC create traffic on their online platforms, thus, turning them into product sales and advertising income for the news network. Thus, through synthetic personalisation, the BBC once more seems to direct its viewers away from its offline environment towards its online platforms.

Finally, as for example (17), which was introduced as a closing message in the new format of news tickers (see Section 5.1), it combines both the features highlighted for example (16) (i.e., in the use of synthetic personalisation) and for example (15) (i.e., in the adverb 'more', which can be seen as both a disclaimer and as a cue to direct viewers towards the online platform by way of offering more information on the news stories presented).

Overt promotions seem to be absent in other sections of news tickers, with an exception in the section dedicated to sport news stories, where live coverage of sport events on the BBC TV channels is advertised. This overt promotion strategy is highlighted in the keyword analysis in the lempos `live-j` (ARF score: 2.4), whose realisation can be seen in the following examples:

(18) LIVE TEXT REACTION AFTER LOTUS DRIVER KIMI RAIKKONEN WINS THE SEASON-OPENING AUSTRALIAN GRAND PRIX IN MELBOURNE

(19) PREVIEW FOLLOWED BY LIVE COVERAGE OF SUNDAY'S GAME BETWEEN ST MIRREN AND HEARTS IN THE SCOTTISH LEAGUE CUP

(20) PREVIEW FOLLOWED BY LIVE COVERAGE OF MONDAY'S GAME BETWEEN MANCHESTER UNITED AND ASTON VILLA IN THE PREMIER LEAGUE

(21) LIVE TEXT AND SMS COMMENTARY AS ENGLAND TAKE ON NEW ZEALAND IN THE THIRD ONE-DAY INTERNATIONAL AT TRENT BRIDGE

(22) LIVE TEXT AND RADIO COMMENTARY AS INDIA RACE PAST 100 WITHOUT LOSS AGAINST SOUTH AFRICA IN THE OPENING MATCH OF THE CHAMPIONS TROPHY

(23) LIVE CHAMPIONS TROPHY TEXT AND RADIO COMMENTARY FROM THE OVAL

(24) LIVE TEXT AND RADIO COMMENTARY AS PAKISTAN LOSE TWO WICKETS EARLY IN THEIR PURSUIT OF 235 TO BEAT SOUTH AFRICA IN THE CHAMPIONS TROPHY

Beside the messages displayed at the end of each round of tickers, these examples represent the only other overt promotional strategy used by the BBC in order to market their products. In particular, a various number of platforms come into play in this type of messages. For instance, example (18) refers to the BBC online website dedicated to live coverage of sport events[71], while example (21) refers to both the BBC online sport platform and the subscription option available to viewers in order to receive via SMS major updates on sport events. Finally, example (24) also refers to the possibility to follow live updates on the BBC radio station.

In line with these examples, we must also acknowledge the subtle promotion of the BBC Sport online platform in the way the label introducing the section dedicated to sport events in news tickers is represented. Indeed, as we have previously underlined, while initially presented like all the other sections, from April 23, 2013 to December 18, 2013, the SPORT section found in news tickers was displayed on a bright-yellow background and written in black capital letters, thus, implicitly and intertextually referring to the BBC Sport news online platform, whose logo is identical in colour and format to the one found in the section on sport events found in news tickers.

In conclusion, in all these cases linked to sport events, it is striking to notice the fact that the BBC itself, instead of keeping viewers on their offline coverage of these events and, thus, tuned in on the BBC World News channel, the news network seems to routinely stress the limitation of the offline coverage, thus, constantly suggesting to viewers to go elsewhere, moving from the offline environment to other online platforms. While subtly implying the limitations of the offline products offered to viewers, these online platforms are persistently presented and marketed in

[71] BBC Sport can be reached online at `http://www.bbc.com/sport`

news tickers as offering news stories that are more newsworthy since they can be considered as more recent.

Turning our attention to covert promotional strategies, Coffey and Clearly's (2011) definition seems not to take under consideration all those cases where overt promotional strategies are covertly achieved. In other words, in identifying overt promotional strategies in their corpus, Coffey and Clearly (2008, 2011) only focus their attention on those phenomena of marketisation realised in news tickers linked to messages overtly acknowledging the news network's desire to advertise its products (i.e., the message at the end of each round of tickers). Thus, overt promotional strategies are limited to these realisations, while covert strategies were only identified in those signals highlighting the news network's desire to sponsor their sister properties without disclosing organizational ties. Thus, for instance, covert promotional strategies are realised when MSNBC airs a positive news story on Microsoft, since the news network was founded thanks to the partnership between Microsoft and NBC; or when Fox News Channel airs a positive news story about MySpace, since they are both owned by News Corp (Coffey and Clearly 2011). However, as we have previously underlined, overt promotional strategies may also be realised covertly in news tickers.

This is particularly the case of the NTC, whose keyword analysis highlights something quite interesting in the comparison with the bw_14 from a marketisation point of view. Indeed, by looking at the very first keywords highlighted in this comparison (see Appendix 1), we can notice that the lempos `bbc-n` (ARF score: 389.3) displays a statistically significant occurrence in the NTC when compared to the bw_14. This is particularly interesting since the reference corpus used for this comparison has been compiled from the same news network organisation as the NTC and, thus, this keyword seems to be indicative of a particular strategy taking place in the news tickers displayed on the BBC World News.

In order to better understand this particular strategy, we offer the following examples randomly taken from the KWIC view of the lempos `bbc-n`:

(25) A LEADING INDIAN SPORTS LAWYER TELLS THE BBC THAT MORE ACTION SHOULD BE TAKEN TO STOP DOPING IN INDIAN ATHLETICS

(26) A BRITISH WOMAN TELLS THE BBC THAT SHE SHOUTED FOR HELP FOR MORE THAN AN HOUR BEFORE JUMPING FROM A HOTEL BALCONY IN INDIA FEARING A SEXUAL ASSAULT

(27) FIVE POLICE OFFICERS ARE KILLED AND 13 PEOPLE INJURED AS SUICIDE ATTACKERS STORM A POLICE STATION IN THE EASTERN AFGHAN CITY OF JALALABAD, POLICE TELL THE BBC

(28) THE LARGEST-EVER SURVEY OF SOCIAL CLASS, CONDUCTED BY THE BBC, SUGGESTS THERE ARE NOW SEVEN SOCIAL CLASSES IN THE UK

(29) TWO CHILDREN ARE TRAPPED AFTER A LANDSLIDE BURIED THEIR HOUSE FOLLOWING HEAVY RAINS IN WESTERN KENYA, THE POLICE TELL THE BBC

(30) THE US SECRETARY OF STATE JOHN KERRY CALLS THE AFGHAN LEADER TO DEFUSE TENSION OVER THE OPENING OF A TALIBAN OFFICE, AN AFGHAN OFFICIAL TELLS THE BBC

(31) ENGLAND CRICKET STAR MONTY PANESAR IS FINED BY POLICE FOR BEING DRUNK AND DISORDERLY AFTER URINATING ON BOUNCERS AT A BRIGHTON NIGHTCLUB, THE BBC UNDERSTANDS

(32) AMERICAN SPRINTER TYSON GAY'S POSITIVE TEST WAS FOR A BANNED STEROID THAT CARRIES A TWO-YEAR SUSPENSION, THE BBC LEARNS

(33) AT LEAST SIX PEOPLE HAVE BEEN KILLED AND 15 INJURED AFTER A SUSPECTED SUICIDE ATTACK AT A HOTEL IN SOMALIA'S CAPITAL, THE INTERIOR MINISTER TELLS THE BBC

As we can see from the examples provided above, we can notice that the keyword under investigation is prominently found in reporting clauses introducing indirect reported clauses, as we can see from Table 10 showing the collocates of the lempos `bbc-n`:

lempos	Cooccurrence count	Candidate count	log likelihood
tell-v	56	134	666.284
the-x	98	7,830	504.275
.-x	63	6,926	247.172
learn-v	13	17	172.279
,-x	34	4,938	104.779
that-i	14	320	91.007
sport-n	6	22	61.929
's-x	13	2,241	33.75
be-v	15	3,379	32.086
will-x	6	541	22.436
have-v	8	1,204	22.432
official-n	5	370	20.539
a-x	12	4,912	13.855
say-v	7	1,748	13.376
as-i	5	1,089	10.651
us-n	5	1,100	10.565
to-i	5	1,427	8.391

Table 10 List of the collocates of the lempos bbc-n.

In news discourse, reporting clauses are typically found in lead paragraphs and usually provide "[…] the evidence for believing or entertaining" the content of given news stories (Huddleston and Pullum 2002: 131). However, from a textual colligation point of view, we must also acknowledge a difference in the use of this particular strategy when compared to the reference corpus of lead paragraphs (i.e., bwl_14). Indeed, while in the bwl_14 these phrases usually occupy sentence-initial positions (followed by the direct or indirect speech of a given actor), in the NTC these phrases show a strong tendency for final-sentence positions, preceded by the indirect reported speech, which shows a tendency for sentence-initial positions. Indeed, as we can additionally see from Table 10, reporting clauses in the NTC are not preceded by the direct quotation of the source they have cited (no inverted commas are displayed as its collocates) but, rather, the source is indirectly acknowledged and its voice is thus accessed through the authority of the BBC, which guarantees the genuineness of the reported speech. Thus, all contents are mediated in news tickers through the authorial voice of the BBC.

This tendency is also confirmed if we look at the bigger picture, that is, if we take under consideration a bigger corpus, such as the Sibol/Port corpus available on Sketch Engine, as we can see from the following table (Table 11), where the first five collocates of reporting clauses found in the Sibol/Port corpus have been calculated[72]:

lempos	Cooccurrence count	Candidate count	log likelihood
"-x	1,370	3,527,500	8118.362
that-x	1,101	3,526,047	5949.581
.-x	1,635	15,378,037	5732.027
:-x	714	1,509,662	4351.906
the-x	1,238	19,859,933	2959.028

Table 11 List of the first five collocates of reporting clauses in the Sibol/Port corpus.

As we can see from Table 11, when reporting clauses occur in the Sibol/Port corpus, they strongly collocate with items indicating that an accessed voice is directly reported and, thus, it is not mediated through the voice of the newspaper. More specifically, if we look more closely at the elements hinting at the introduction of a direct quotation in the previous search, we can notice that they strongly collocate to the right of the node, as we can see from the following table (Table 12):

lempos	Raw frequency	
	3L	3R
"-x	454	1,093
that-x	16	1,089
:-x	4	710

Table 12 Elements hinting at the introduction of a direct quotation in the Sibol/Port corpus.

[72] The list of collocates has been compiled by searching the corpus for the following linguistic construction thanks to the CQL search option available on Sketch Engine: `[tag="N.*"] []{0,3} [lemma="tell"] [word="the"] [word="Guardian|Telegraph|Times"]`. The CQL search has been modelled on the most significantly occurring reporting clauses found in the NTC (see Table 10), thus, offering a comparison to the phenomenon under investigation. The collocates have been calculated in a span of three words to the right and the left of the key word in context.

Table 12, thus, confirms the tendency in the Sibol/Port corpus for reporting closes to occur in sentence-initial positions and, consequently, this further underlines on a big scale the peculiarity of the textual organisation of news tickers.

Indeed, if reporting clauses are generally in initial positions in order to foreground the accessed voices in the reported speech and background the actual process of collecting these voices by the media agency in the reporting clause, in news tickers, both actions have the same importance and, thus, they are both foregrounded. In other words, the news stories must be considered true and newsworthy given the authority of the BBC and, in this process, we thus have a subconscious representation in the viewers' mind of the BBC as a source of reliability and trustworthiness. In other words, in legitimising the newsworthiness and trustworthiness of the news story, news tickers seem to make a constant reference to what van Leeuwen's (2007) describes as forms of expert authority. Additionally, the BBC seems to be used as a membership categorisation device (Sacks 1972; see also Montgomery 2007). Indeed, according to Montgomery (2007), "[p]ersons in headlines are designated by expressions that refer to them not so much as particular individuals but as members of significant groups or institutions" (Montgomery 2007: 78). However, membership categorisation devices depend on the degree of popularity of the actors presented in headlines. If the person is well-known to the viewers/readers, then no membership categorisation device is needed. When, on the other hand, the person is not well-known, then membership categorisation devices are applied. In the reporting clauses where the phenomenon under investigation has been highlighted, the name of the reporter should have been displayed. This hypothesis is supported by the fact that, in the bw_14 reference corpus, we have noticed that the name of the reporter or the complete absence of the person or institution that has collected the voice of the actor is displayed in lieu of the news network. Thus, in the multi-platform authoring environment (Bivens 2014) of the BBC, when a content is migrated from one platform to the other, some transformations will occur. In news tickers, since it is impossible to add a byline with all the information on the reporter and, as previously seen, since crawlers represent a prolific environment for strategies of marketisation, the news network is displayed as the source of the news story.

Indeed, in evaluating the news through the use of the BBC's authority, this particular strategy also seems to further underline the marketisation features of news tickers. In fact, these textual patternings highlighted in the news tickers displayed on the BBC World News seems to distinctly enhance the news value of attribution (see Section 3.4), whereby socially validated authorities are allowed to see their voices represented in the media (Bednarek 2015). However, these voices are always controlled and accessed through the authority of the news network, whose continuous reference in news tickers makes it a prominent source thanks to which news stories are proofread and communicated to viewers. And since, in the comparison with the reference corpus, this phenomenon seems to be peculiar of news tickers, we can rightfully consider it as a specific textual realisation of the genre. Additionally, since this phenomenon can be seen as constantly 'putting on the spotlight' the news network as a source of reliability and trustworthiness, we can include this in the overt promotional strategies used in news tickers in order to covertly promote the news network and its products.

As a way of concluding this Section, we can say that the use of corpus linguistic methodologies has helped us underline a particular marketisation strategy realised in the news tickers displayed on the BBC World News. This strategy can be seen as a demonstration of the hybrid nature of news tickers, through which a specific private intention is textually realised in the professional environment of the BBC. We have also seen that this textual realisation can be highlighted thanks to the use of the notion of textual colligation, which has allowed us to compare the occurrence of the phenomenon under investigation in different corpora, thus, demonstrating the peculiarity underlined in the NTC. However, this analysis is strictly linked to a lexical approach to the analysis of the target corpus. Thus, in the following Section, we are going to concentrate on the grammatical status of the genre under investigation, realised thanks to the tag keyword extraction in the comparison between the NTC and the bw_14.

5.2.2 Bending genre conventions: The use of temporal deixis in news tickers

Focusing our attention on the tag keyword extraction compiled thanks to the comparison between the NTC and the bw_14, we can highlight another significant

phenomenon occurring in the textual realisation of the genre under investigation, which is indicative of a genre bending practice realised in news tickers.

As we have seen in Section 3.2.2, by bending socially accepted generic norms coming from traditional genres, users may create new genres or sub-genres that are, thus, shaped in order to achieve specific private intentions. But, before focusing our attention on the phenomenon highlighted from the tag keyword analysis, we would like to start from a quite simple observation that comes from the study of the media literature on news tickers.

Indeed, when searching for given references in the literature to the genre of news tickers, we have noticed that, when introduced in books or papers on media discourse, crawlers are generally compared to headlines, as we can see from the following quotation taken from Chovanec (2014), where he discusses the functional value of given segments in newspapers articles (Chovanec 2014: 61-62, emphasis added):

> It could be argued that the journalistic lead is inextricably linked to the body copy: the functional value of that segment rests in the fact that it is followed by the article proper. However, the lead can also operate as a stand-alone unit, existing only in conjunction with the headline – as is the case with some very brief news items. [...] In the absence of a body copy, the whole news text can be constituted merely by the combination of the headline and the lead. Similarly, with some genres of brief news items such as live tickers on TV, a mere *headline* can constitute a *complete and self-contained* news text.

Thus, as a genre, news tickers are usually compared to news headlines and, thus, they are associated with their function and rhetoric. In other words, news tickers are implicitly associated with the lexicogrammatical features of headlines[73].

[73] In line with Isani (2011), when referring to headlines, we use the term in order to point to its generic status, in contrast with the term 'headlinese', which refers to a sub-genre of headlines, whose function is to "persuade the reader to read the news story through use of linguistic manipulation and decontextualisation designed to resist comprehension and prod, lure and incite the reader into reading the following body copy" (Isani 2011: 8).

However, as we have previously seen in paragraph 5.2.1, news tickers seem more comparable to lead paragraphs, since tickers tend to bend some of their conventions in order to achieve given purposes (in the specific case of the corpora under investigation, we have seen how reporting clauses are specifically constructed in news tickers in order to achieve a particular marketisation purpose). Thus, what is the real nature of news tickers? And are they really, as Chovanec (2014) argues, complete and self-contained news texts?

In order to answer to these questions, we will once again use corpus linguistic methodologies. In particular, since we are now focusing on what Halliday (1985) defines as "the grammar of little texts" (Halliday 1985: 392), the keyword analysis will be performed on the POS-tags automatically encoded in the NTC and the bw_14 thanks to the Sketch Engine online platform.

The lexicogrammatical status of headlines has attracted the attention of various scholars in the field of media studies (Mardth 1980; Bell 1991; Isani 2011; Bednarek and Caple 2012a; Chovanec 2003, 2014), who have underlined how headlines can be generally seen as having their own distinctive grammar. Indeed, as Fairclough (1995b) argues, "headlines have distinctive syntactic properties, which make them a grammatical oddity" (Fairclough 1995b: 21). And as we can see from the following table (based on Chovanec 2014: 119-120), some of their most characteristic grammatical 'oddities' are listed in order to better understand their peculiar grammatical status:

GRAMMATICAL FEATURES OF HEADLINES		
Features	**Textual realisations**	**Examples from the bwh_14**
Shift of tenses	• Past/Present Perfect → Present • Future → non-finite verbal group	• CHINA DENOUNCES US-JAPAN SPEECHES • TORIES TO LAUNCH DEVOLUTION PLANS
	Ellipsis of the definite forms of the verb 'to be'	TURKEY CAPITAL HIT BY FRESH CLASHES
	Ellipsis of definite/indefinite articles and determiners	MAN DIES AFTER BEING HIT BY TRAIN
Non-finite passive constructions	Ellipsis of finite operators	MAN CHARGED WITH CAR PARK RAPE
Unattached nominals	Noun groups stating the mere existence of a phenomenon	BUSINESS BOOST IN QUEEN'S SPEECH
Condensed quotations	Replacement of verbs of speech with the colon preceded by a designation of the relevant news actor	RSPB: 'WE DID NOT KILL 22 BIRDS'
Avoidance of modalised statements	Lexical expression of modality (e.g. through the verb 'to face')	HARDAKER FACES 'HOMOPHOBIC' INQUIRY
Shift in modality from possibility towards certainty	From modally hedged utterances in the lead and the body copy to the presentation of the relevant proposition in the headline as factual and without hedging or modality	INTERCONNECTOR ENERGY SHORTAGE FEARS (followed by the lead paragraph: "Northern Ireland could face serious energy shortages if the north-south electricity interconnector project is not started soon")

Table 13 Grammatical features of headlines as summarised by Chovanec (2014: 119-120), with examples taken from the bwh_14.

Among the grammatical features exemplified in Table 13, we would like to focus our attention on the shift of tenses occurring in headlines (in particular, from past to present tenses), since the chronological orientation of news texts plays an important role in the representation of world entities and events. Indeed, as White (2003: 85-87) argues:

> The chronological orientation […] is indicative of a concern with, or focus upon, how processes unfold in real or fictional worlds […]. There is a clear focus, then, within narrative upon what Halliday terms the 'ideational', upon the representation of external-world entities and events and the logical relationships which hold between them.

And if the present simple can be seen as the default tense of headlines (Declerk 1991), this entails that, from an ideational point of view, it seems to convey a certain concern with an everlasting 'now' of the event, without necessarily locating it in time.

Focusing our attention on news tickers, as we can see from Appendix 2, they seem to be appropriating this feature of headlines. Indeed, in the comparison between the NTC and the bw_14, the ARF score (for `[tag="VVP"]` is 1.9 and for `[tag="VVZ"]` is 1.8) highlights that the present simple is one of the most frequently used grammatical features of news tickers. In particular, the comparison with the two sections of the reference corpus (see Appendix 3 and Appendix 4) confirms that news tickers are indeed borrowing this grammatical aspect from headlines, as we can further see from Table 14, where the log-likelihood calculated from the comparison between the NTC and the bw_14, the bwh_14, and the bwl_14 is displayed[74]:

	RAW FREQUENCY NTC	**RELATIVE FREQUENCY NTC**	**RAW FREQUENCY bw_14**	**RELATIVE FREQUENCY bw_14**	**OVERUSE (+) UNDERUSE (-)**	**LL**
`[tag="VVP|VVZ"]`	6,382	3.79	13,691	2.22	+	1158.35
	RAW FREQUENCY NTC	**RELATIVE FREQUENCY NTC**	**RAW FREQUENCY bwh_14**	**RELATIVE FREQUENCY bwh_14**	**OVERUSE (+) UNDERUSE (-)**	**LL**
`[tag="VVP|VVZ"]`	6,382	3.79	6,955	5.51	-	469.01
	RAW FREQUENCY NTC	**RELATIVE FREQUENCY NTC**	**RAW FREQUENCY bwl_14**	**RELATIVE FREQUENCY bwl_14**	**OVERUSE (+) UNDERUSE (-)**	**LL**
`[tag="VVP|VVZ"]`	6,382	3.79	6,335	1.29	+	3527.50

Table 14 Log-likelihood calculation computed in the comparison between the NTC and the bw_14, bwh_14, and bwl_14.

A possible explanation to this peculiarity occurring in the NTC may be found in the fact that both headlines and news tickers are bound to specific constraints on the

[74] The statistics offered in Table 14 have been calculated thanks to the online log-likelihood and effect size calculator developed by Rayson (2015) at Lancaster University. The calculator is freely available at `http://ucrel.lancs.ac.uk/llwizard.html`

number of characters to be used in presenting news stories, thus, creating this similarity. However, the number of characters used in news tickers seems not to prevent the appropriation of specific features borrowed from the rhetoric configuration of lead paragraphs (see Section 5.2.1). Thus, the explanation to this statistically frequent use of the present simple in news tickers is linked to something else.

Indeed, just like headlines, in appropriating the use of the present simple, news tickers seem to be attending the interpersonal function (Halliday 1985) by projecting the deictic centre of the news stories in the viewers' temporal space. Thus, assumed recency becomes one of the most prominent features of this genre. This is also demonstrated by the fact that, by searching the NTC for `[tag="RB"]`, which allows us to look at all the adverbs occurring in the corpus, very few occurrences of temporal adverbs are present[75]. This is the reason why the temporal deixis conveyed by the use of specific tenses plays a crucial role in the temporal anchorage of given events presented in news tickers. Thus, the prominence of the present simple in the NTC acquires a significant value, indicative of a specific purpose that the genre wants to achieve in presenting the news stories with no further temporal anchorage other than the present simple.

Indeed, going back to the lexicogrammatical status of headlines, the lack of a further specific temporal anchorage other than the present simple is part of a subtle strategy to keep readers reading the news stories they announce and retrieve this information from the lead paragraph or the body copy. Additionally, Chovanec (2014) provocatively argues that the present tense in headlines can be considered as non-deictic at all. Indeed, according to the author, the present tense used in headlines does not act referentially as it does not provide a temporal anchorage of the event but, rather, it works only in order to "heighten the current relevance of the event" (Chovanec 2014: 273), thus, showing a strong interpersonal orientation.

In the case of news tickers, on the other hand, the use of the present simple seems to implicitly encourage viewers to either keep on watching the TV newscast or

[75] The only noticeable use of a temporal anchorage is entailed by the occurrence in the corpus of the adverb 'now', but also in this case it occurs very infrequently, with only 41 raw occurrences. A list of all the adverbs occurring in the NTC can be found in Appendix 5.

go to the BBC website to retrieve the temporal specification that is not provided in this genre. However, as we have seen in the previous Section, since news tickers seem to direct viewers towards the BBC online platforms, the news value of timeliness enhanced by the choice of this particular tense seems to further confirm the previously underlined professional purpose, that is, drawing away viewers from the offline platform and lead them towards the online environment.

Thus, the lack of a specific temporal anchorage forces viewers to either assume that the news stories are actually recent or to use other platforms in order to retrieve this specification. By doing so, viewers "are drawn more actively into the process of communication because they have to reconstruct some of the missing grammatical information" (Chovanec, 2014: 188), increasing viewers' involvement in the multiplatform environment of the BBC.

In conclusion, going back to the questions introduced at the beginning of this Section on the nature of news tickers, we can now say that they do not represent complete and self-contained news texts or, in other words, they should not be considered as stand-alone news items but, rather, their lack in providing specific information is used in order to achieve particular purposes. Indeed, in the case of news articles, the 'inverted pyramid' structures news stories from the highest form of generalisation of the event in the headline to further specifications in the lead and body copy. This outlet of news articles is usually adopted by journalists in order to keep readers reading the news stories, since the generalisation and lack of information found in the headlines lead readers towards other parts of the article in order to retrieve the missing information. In news tickers, however, the absence of given information, such as the case of the temporal anchorage of the news story, is used in order to enhance the multifaceted news flow of the BBC environment, where different platforms are used in order to distribute news stories to the public. Each platform, however, has its own purpose and, in the specific case of news tickers, the linguistic peculiarities highlighted in this section seem to indicate that their main aim is to draw viewers away from themselves and, more importantly, towards other platforms. In other words, as a (sub-)genre found in TV news broadcasts, by bending specific genre conventions of traditional genres, news tickers play on their

incompleteness in order to favour other platforms[76]. Thus, while they can resemble news headlines in the way they strategically borrow some of their lexicogrammatical characteristics, they seem to be appropriating also some textual conventions of lead paragraphs in order to provide certain information. However, the lack of an in-depth summary of the news stories they present seems to contradict this last observation. And, thus, they should not be considered as either headlines or lead paragraphs, but as a mixed (sub-)genre that bends some of the conventions found in these two genres of news reporting in order to achieve their specific purpose.

As a further confirmation of some of the observations provided in the last Sections, we would like to conclude our investigation by taking under consideration another aspect that can shed light on the nature of the genre under consideration. Indeed, as we have previously argued, genres may also entail the values that a particular discourse community holds dear in its professional practice. Thus, by analysing the news values that are particularly frequent in the NTC, we can further define the genre under investigation. Additionally, the news values highlighted in the NTC further confirm some of the claim that specific patterns retrieved in the corpus have supported. Thus, in the following section, we will illustrate the methodology used in order to retrieve the pointers to newsworthiness in the NTC and what they entail in shaping the genre under consideration.

5.3 News values in news tickers

As we have previously argued in Section 3.4, by highlighting specific shared values in the context of professional practices, researchers may further investigate the private intentions that a given community of practice wants to achieve when resorting to these values. In particular, in the context of news industry, editors usually and routinely rely on a series of factors when deciding which and how specific events can be taken into the news. These factors are generally referred to as news values (Bell 1991; Bednarek and Caple 2012a, 2012b, 2014; Potts, Bednarek and Caple 2015) and, as we have previously seen, by highlighting the way specific

[76] This, again, may be seen in line with the observations made by Isani (2011) in the case of headlinese. Indeed, also with reference to news tickers, opacity (more specifically, temporal opacity) is strategically used in order to achieve a given purpose.

media genres regularly enhance given values, we can gain a first insight into the professional practice of news reporting in specific situations. Consequently, we can further investigate the specific private intentions that given genres want to achieve in discursively constructing newsworthiness.

Thus, in this Section, we will concentrate on the way news tickers routinely enhance given news values. This investigation has been again carried out by using corpus linguistic methodologies and, in particular, by following the methodology used by Bednarek and Caple (2014) and Potts, Bednarek and Caple (2015). While the purposes of their investigation are not specifically linked to the type of analysis that we are carrying out, we can nonetheless use the type of approach to the analysis of news values they have performed in order to further see if, from a newsworthiness point of view, the observations introduced in Section 5.2.1 and 5.2.2 are further confirmed or contradicted if we look at the genre under investigation from a different perspective.

In order to carry out our analysis of the news values that are routinely enhanced in the NTC, we have once again resorted to the Word list tool available on the Sketch Engine online platform. As for the computation of keywords in the NTC, the word list has been calculated by searching for lempos attributes. A cut-off point of minimum frequency of five occurrences has been imposed and, in order to further ensure that the selection of given items was not only due to their frequency, but also to their dispersion in the corpus under investigation, we have decided to make use of the ARF. This is something different from the methodology used by Bednarek and Caple (2014) and Potts, Bednarek and Caple (2015), since they only take under consideration as a parameter in the analysis of news values the raw frequency of given words in the corpora they set out to investigate, without considering the dispersion of the phenomena highlighted.

The word list thus computed was not, however, contrasted with a reference corpus, since the scope of this part of our analysis is not to contrastively highlight differences among genres in news discourse but, rather, see how given items in the NTC are specifically and routinely used in order to construe newsworthiness. Indeed, a comparison with the bw_14 may have altered the results, due to the nature of the genres *per se* and not to the way newsworthiness is construed in the two corpora.

Among the items in the word list computed in accordance with the methodology previously described, we have decided to focus only on the first 200 words in this list, since the statistics used in order to calculate them ensures that they do represent the most routinely used lexical items in the NTC. We have then proceeded to the manual analysis of the items extracted thanks to this methodology, attributing to each one a news value. In order to help us decide which value given words convey, we have used the Concordance tool, in order to better understand their context of occurrence, thus, helping us better understand the way they were strategically used in order to enhance a given value. However, we have decided to refer to them as 'potential pointers' to specific news values because, while the majority of the collocates in the KWIC pointed towards the realisation of a given news value, others did not corroborate this analysis. As such, when referring to given items as 'potential pointers' to specific news values we do not mean that they are always used in order to convey these values but, rather, that they discursively show a strong tendency towards the realisations of specific news values.

We must also underline that, in the analysis of the first 200 words in the word list computed according to the methodology previously described, some items have not been considered as enhancing a specific news value (this is particularly the case, for instance, of function words).

Finally, we must also acknowledge that, in the attribution of given news values to specific items found in the Word list, some words have been analysed as simultaneously construing more than just one news value (for instance, the adjective `late-j` (ARF score: 64.6), which was specifically used in the NTC in order to present the 'latest' results of given reports or elections, was both analysed as a potential pointer of Timeliness and Novelty).

While Appendix 6 shows all the results of our manual annotation of the first 200 words in the Word list, Figure 23 summarises the results of our investigation on the news values most frequently enhanced in the NTC.

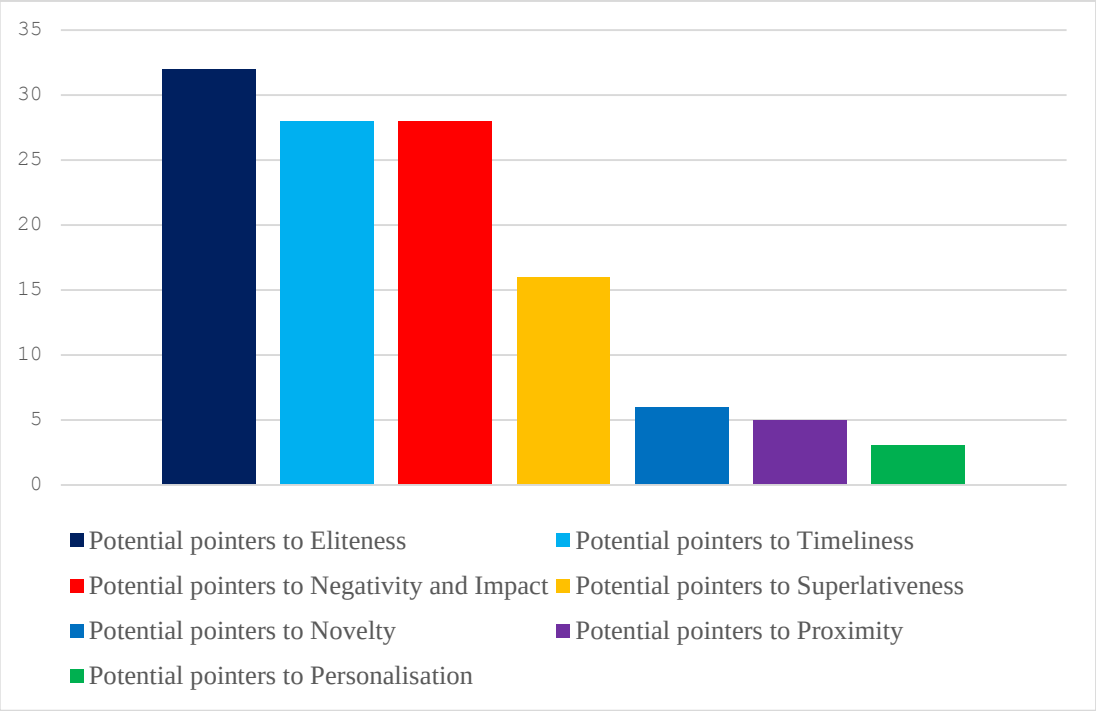


Figure 23 A summary of the news values most frequently enhanced in the NTC.

The results shown in Figure 23 seem to corroborate the observations that we have previously made in Section 5.2.1 and 5.2.2. In particular, as we can see, Eliteness is one of the most frequently enhanced news values in the NTC, proving our claim that socially validated authorities are routinely allowed to see their voices represented in news tickers. Additionally, in our manual attribution of given news values to specific items in the first 200 words in the Word list, we have included in the potential pointers to Eliteness also the verbs `say-v` (ARF score: 1180.1), `tell-v` (ARF score: 72.1), and `report-v` (ARF score: 68.7) since, as we have previously seen, they are particularly used in the NTC in order to access given elite sources but, more importantly, they are used in order to enhance the authority of the news network, which constantly controls and let these voices 'speak' in the genre of news tickers.

Figure 23 also supports our observation on the particular enhancement of the news value of Timeliness in the NTC. While, in Section 5.2.2, we have seen how this news value is enhanced grammatically, thus, revealing a specific purpose in the presentation of events in news tickers, this is also confirmed in the lexical analysis of the items found in the Word list of the NTC. Thus, the combination of

lexicogrammatical features found in the NTC demonstrates how this genre found in the professional environment of the BBC World News is strategically shaped in order to enhance recency, giving viewers an impression of a continuous and live commentary of the events.

Potential pointers to Negativity and Impact seem also to be significantly used in the NTC. However, while some of them can be specifically linked solely to the news value of Negativity (i.e., the lempos `kill-v` (ARF score: 275.2), `die-v` (ARF score: 119.4)), the majority of them seem to be used in order to highlight the effects of given events. In this way, the news value of Impact seems to be highlighting the current effects of a given event, thus, implicitly enhancing the Timeliness of the news story reported in news tickers.

As we have seen from this brief analysis of the news values enhanced in the genre of news tickers, by looking at the professional practice in the way given shared values are textually constructed in discourse, we can further see how given specificities in the genre are particularly highlighted. This analysis has, thus, allowed us to further investigate the genre of news tickers as found in the professional media context of the BBC World News, and see how given claims previously proven by using different methodologies are further confirmed by looking at the genre from the point of view of the way particular values are enhanced in the genre of news tickers.

Conclusion

As we have argued at the very beginning of this contribution, journalistic practices have been undergoing, in the last few years, radical changes due to the increasing pressure of new digital media on the professional practice. As our lives and social institutions are increasingly and constantly in flux, this ceaseless fluctuation of social practices has inevitable consequences on the genres and discourses created by social institutions. Thus, journalistic practices and genres "should be understood within the wider context of liquidity" (Bivens 2014: 77), as practices which try to incorporate in their routines and in their genres the liquidity of contemporary society. In such a fluid social context (Deuze 2008), genres are, therefore, increasingly becoming dynamic rhetorical configurations, whose conventions can be exploited to achieve new goals. Thus, mixed or hybrid forms are most frequently the results of these manipulations. Indeed, given the ever-changing social context where journalistic practices operate, they are constantly exploiting new forms of hybridity and genre-mixing in order to compete with new ways of delivering the news. This intensifying pressure on traditional media has given rise to a variety of hybrid and mixed-generic forms, among which we have focused our attention on a relatively new genre of TV news broadcasts, generally referred to as news tickers (or crawlers).

This genre, which made its first appearance on 9/11 in order to deal with the enormous amount of information coming from the American news agencies, has been adopted by various TV news channels and programmes in order to constantly deliver to viewers a summary of the major news stories of the day or to alert viewers of particular breaking news stories. However, as we have seen, during the years and given the increasing pressure on TV journalism to allure viewers, the genre of news tickers has been slowly appropriating certain generic conventions from other genres to serve this purpose.

Indeed, thanks to a corpus-based linguistic analysis, we have seen how the BBC World News uses its news tickers in order to promote itself and its products. The mixed nature of this genre is, first and foremost, proven by the merging of two functions traditionally belonging to the journalistic genres of headlines and lead paragraphs. Indeed, while headlines typically "function to frame the event,

summarize the story and attract readers", the lead paragraphs work on the information provided in the headline and "describe newsworthy aspects of the event (e.g. the who, the what, the where)" (Bednarek and Caple 2013: 96-97). News tickers, thus, by merging these two purposes thanks to given lexicogrammatical features coming from these two traditional genres, must at the same time catch viewers' attention and give viewers a point of view on the story. However, these two functions coexist with a constellation of other communicative purposes highlighted by structural patterns thanks to the use of corpus linguistic methodologies.

One of these communicative purposes, as we have seen, can be ascribed to what Meech (1999) defines as brandcasting, which refers to the vast array of corporate branding techniques that broadcasters use in order to project their brand identity. These branding techniques are highly frequent in the NTC corpus and, while some of them may be classified as overt promotional strategies (Clearly and Coffey 2008, 2011), others may be seen as achieving the same purpose more subtly. In these cases, the authority of the BBC is used in order to legitimise the newsworthiness of the news story found in the news ticker, thus conveying a subconscious representation in the viewers' mind of the BBC as a source of reliability and trustworthiness. Additionally, we have also highlighted how these strategies adhere to a quite strict textual colligational pattern (O'Donnell *et al.* 2012) in the textual organization of news tickers. Thus, these results highlight a peculiarity in the genre under investigation, which has been further proven by the comparison with a reference corpus collected from the same professional environment. The differences in the two media represented in the target corpus and the reference corpus underline how relevant brandcasting is for a TV genre such as that of news tickers, which has found a compromise between its communicative function to inform its viewers/readers and to subtly promote its brand identity.

Limitations and future steps

One of the main limitations in our investigation is represented by the fact that, by focusing on a specific genre in TV news broadcasts, all other genres have not been taken into consideration. Due to the nature of our study (i.e., defining the genre of news tickers), other graphic elements displayed on the BBC World News in the time-

span of the collection of our data have not been investigated. Additionally, the interaction between textual elements and the audio content reproduced on the BBC World News have not been analysed. However, from this investigation into the generic nature of news tickers, we are now developing a new research project which will investigate how news contents related to specific news stories are presented in different ways in the multiplatform media environment of the BBC and, more specifically, in the different genres found in these platforms. In particular, the purpose of this future research project based on a constructed week investigation of the journalistic practices of the BBC World News is that of analysing how news stories are firstly introduced in the TV newscasts' headlines. The news stories are then further investigated in the way they are presented in the different textual elements implemented on-screen. The audio content will also be investigated in order to see which elements are particularly enhanced in the textual elements found in the TV newscasts and what absences can be thus highlighted in the comparison with the audio content. Finally, two other media platforms are also taken into account in order to see how news stories migrate from one platform to the other and what differences can be highlighted in this migration. In particular, we will focus on the BBC website and the Twitter account of the BBC. Given the contrastive nature of this future development of our research, a more qualitative approach to the data will be adopted. In particular, as we have seen in this investigation, since news values can be used in order to see how and which elements of news stories are being highlighted in given genres, we will again focus our attention on them, in order to see if different platforms entail a different enhancement of given pointers to newsworthiness.

Appendix

Appendix 1

Keyword extraction from the comparison between the NTC and the bw_14

	NTC		bw_14		
lempos	**ARF**	**ARF/mill**	**ARF_ref**	**ARF_ref/mill**	**Score**
china-n	202.9	1204.1	0.0	0.0	1205.1
australia-n	105.8	627.9	0.0	0.0	628.9
england-n	97.1	576.4	0.0	0.0	577.4
Uk-n	96.1	570.0	0.0	0.0	571.0
syria-n	92.8	550.4	0.0	0.0	551.4
india-n	69.9	415.1	0.0	0.0	416.1
bbc-n	65.4	388.3	0.0	0.0	389.3
korea-n	59.1	350.6	0.0	0.0	351.6
japan-n	57.9	343.9	0.0	0.0	344.9
obama-n	57.5	341.0	0.0	0.0	342.0
russia-n	55.7	330.4	0.0	0.0	331.4
eu-n	55.2	327.8	0.0	0.0	328.8
france-n	48.3	286.6	0.0	0.0	287.6
uk-n	47.3	281.0	0.0	0.0	282.0
manchester-n	43.6	259.0	0.0	0.0	260.0
africa-n	40.1	238.2	0.0	0.0	239.2
barack-n	38.8	230.3	0.0	0.0	231.3
german-j	38.1	226.4	0.0	0.0	227.4
egypt-n	35.6	211.4	0.0	0.0	212.4
african-n	34.7	206.2	0.0	0.0	207.2
york-n	34.7	206.0	0.0	0.0	207.0
brazil-n	34.0	201.9	0.0	0.0	202.9
john-n	33.1	196.2	0.0	0.0	197.2
afghan-n	32.0	189.9	0.0	0.0	190.9
zealand-n	30.1	178.7	0.0	0.0	179.7
un-j	29.1	172.8	0.0	0.0	173.8
europe-n	28.3	168.0	0.0	0.0	169.0
britain-n	27.7	164.1	0.0	0.0	165.1
london-n	26.7	158.4	0.0	0.0	159.4
afghanistan-n	26.2	155.3	0.0	0.0	156.3
germany-n	26.2	155.2	0.0	0.0	156.2
morsi-n	26.1	155.1	0.0	0.0	156.1
andy-n	25.7	152.6	0.0	0.0	153.6
pakistan-n	24.4	144.6	0.0	0.0	145.6

liverpool-n	23.7	140.6	0.0	0.0	141.6
islamist-n	23.5	139.3	0.0	0.0	140.3
mohammed-n	23.5	139.2	0.0	0.0	140.2
spain-n	23.4	138.6	0.0	0.0	139.6
mexico-n	23.2	137.9	0.0	0.0	138.9
taliban-n	22.1	130.9	0.0	0.0	131.9
moscow-n	21.5	127.8	0.0	0.0	128.8
tv-n	20.6	122.5	0.0	0.0	123.5
murray-n	20.6	122.2	0.0	0.0	123.2
mps-n	20.4	120.9	0.0	0.0	121.9
google-n	20.1	119.3	0.0	0.0	120.3
mandela-n	19.8	117.3	0.0	0.0	118.3
iran-n	19.6	116.5	0.0	0.0	117.5
chelsea-n	19.6	116.2	0.0	0.0	117.2
david-n	19.2	114.0	0.0	0.0	115.0
facebook-n	19.1	113.5	0.0	0.0	114.5
nelson-n	18.8	111.7	0.0	0.0	112.7
pope-n	18.6	110.2	0.0	0.0	111.2
supreme-j	17.7	105.3	0.0	0.0	106.3
kerry-n	17.6	104.7	0.0	0.0	105.7
america-n	17.3	102.6	0.0	0.0	103.6
al-qaeda-n	17.1	101.4	0.0	0.0	102.4
barcelona-n	16.7	99.4	0.0	0.0	100.4
ireland-n	16.7	99.2	0.0	0.0	100.2
james-n	16.4	97.4	0.0	0.0	98.4
turkey-n	16.2	96.3	0.0	0.0	97.3
paris-n	16.1	95.6	0.0	0.0	96.6
tony-j	15.7	93.3	0.0	0.0	94.3
un-n	39.5	234.3	1.0	1.6	89.8
japan-j	14.8	87.8	0.0	0.0	88.8
sebastian-j	14.8	87.8	0.0	0.0	88.8
david-j	14.8	87.6	0.0	0.0	88.6
scotland-n	14.5	85.8	0.0	0.0	86.8
vettel-n	14.1	83.6	0.0	0.0	84.6
yen-n	14.0	82.8	0.0	0.0	83.8
senate-n	13.8	81.8	0.0	0.0	82.8
francis-n	13.7	81.6	0.0	0.0	82.6
cameron-n	13.6	80.8	0.0	0.0	81.8
lewis-n	13.4	79.8	0.0	0.0	80.8
washington-n	13.2	78.2	0.0	0.0	79.2
baghdad-n	13.1	78.0	0.0	0.0	79.0
munich-n	13.0	77.0	0.0	0.0	78.0
california-n	12.7	75.7	0.0	0.0	76.7

putin-n	12.4	73.5	0.0	0.0	74.5
sudan-n	12.3	72.8	0.0	0.0	73.8
texas-n	12.2	72.5	0.0	0.0	73.5
dutch-n	12.2	72.4	0.0	0.0	73.4
microsoft-n	12.0	71.4	0.0	0.0	72.4
martin-n	12.0	71.2	0.0	0.0	72.2
fifa-n	11.9	70.6	0.0	0.0	71.6
hong-n	11.9	70.4	0.0	0.0	71.4
nigeria-n	11.7	69.6	0.0	0.0	70.6
gulf-n	11.7	69.2	0.0	0.0	70.2
sir-n	11.7	69.2	0.0	0.0	70.2
angela-n	11.5	68.1	0.0	0.0	69.1
morgan-n	11.4	67.8	0.0	0.0	68.8
boeing-n	11.3	67.2	0.0	0.0	68.2
canada-n	11.3	67.2	0.0	0.0	68.2
merkel-n	11.1	66.2	0.0	0.0	67.2
damascus-n	11.1	66.0	0.0	0.0	67.0
kevin-n	11.1	65.8	0.0	0.0	66.8
oscar-n	11.1	65.6	0.0	0.0	66.6
samsung-j	10.8	63.9	0.0	0.0	64.9
fukushima-n	10.7	63.3	0.0	0.0	64.3
bp-n	10.6	62.7	0.0	0.0	63.7
dreamliner-n	10.5	62.0	0.0	0.0	63.0
sri-n	10.3	61.2	0.0	0.0	62.2
jp-n	10.3	61.2	0.0	0.0	62.2
hamilton-n	10.3	61.2	0.0	0.0	62.2
bangladesh-n	10.3	60.9	0.0	0.0	61.9
brotherhood-n	10.1	59.9	0.0	0.0	60.9
italy-n	10.0	59.3	0.0	0.0	60.3
madrid-n	9.9	58.6	0.0	0.0	59.6
august-j	9.9	58.5	0.0	0.0	59.5
barclays-n	9.8	58.3	0.0	0.0	59.3
indonesia-n	9.8	58.2	0.0	0.0	59.2
milan-n	9.7	57.7	0.0	0.0	58.7
amazon-n	9.6	57.1	0.0	0.0	58.1
tottenham-n	9.5	56.6	0.0	0.0	57.6
lebanon-n	9.5	56.4	0.0	0.0	57.4
colombia-n	9.4	55.8	0.0	0.0	56.8
shia-n	9.4	55.7	0.0	0.0	56.7
congo-n	9.3	55.4	0.0	0.0	56.4
istanbul-n	9.2	54.8	0.0	0.0	55.8
italy-a	9.2	54.5	0.0	0.0	55.5
republic-n	30.5	181.1	1.4	2.3	55.0

george-n	9.1	53.9	0.0	0.0	54.9
boston-n	8.9	53.0	0.0	0.0	54.0
mourinho-n	8.9	53.0	0.0	0.0	54.0
hollande-n	8.9	52.6	0.0	0.0	53.6
hsbc-n	8.9	52.6	0.0	0.0	53.6
ukraine-n	8.8	52.1	0.0	0.0	53.1
regulator-v	8.7	51.6	0.0	0.0	52.6
ben-n	8.7	51.6	0.0	0.0	52.6
cardiff-n	8.7	51.5	0.0	0.0	52.5
moyes-n	8.7	51.4	0.0	0.0	52.4
geneva-n	8.7	51.4	0.0	0.0	52.4
cia-n	8.6	50.9	0.0	0.0	51.9
vladimir-n	8.6	50.9	0.0	0.0	51.9
f-n	8.4	49.9	0.0	0.0	50.9
peter-n	8.3	49.4	0.0	0.0	50.4
beirut-n	8.2	48.5	0.0	0.0	49.5
july-a	8.1	48.2	0.0	0.0	49.2
qatar-n	8.1	48.1	0.0	0.0	49.1
paul-n	8.0	47.7	0.0	0.0	48.7
los-n	8.0	47.6	0.0	0.0	48.6
airbus-n	8.0	47.3	0.0	0.0	48.3
g-n	7.9	47.0	0.0	0.0	48.0
greece-n	7.9	47.0	0.0	0.0	48.0
cup-n	69.8	414.1	4.7	7.7	47.8
musharraf-n	7.9	46.6	0.0	0.0	47.6
Pervez-n	7.9	46.6	0.0	0.0	47.6
lanka-n	7.8	46.4	0.0	0.0	47.4
venezuela-n	7.8	46.4	0.0	0.0	47.4
cook-n	7.8	46.2	0.0	0.0	47.2
aston-n	7.8	46.1	0.0	0.0	47.1
novak-n	7.8	46.0	0.0	0.0	47.0
arctic-j	7.8	46.0	0.0	0.0	47.0
state-v	20.5	121.6	1.0	1.6	46.5
colorado-n	7.6	45.3	0.0	0.0	46.3
greenpeace-n	7.6	44.9	0.0	0.0	45.9
cyprus-n	7.5	44.7	0.0	0.0	45.7
chile-n	7.5	44.6	0.0	0.0	45.6
bayern-j	7.5	44.4	0.0	0.0	45.4
joe-n	7.5	44.3	0.0	0.0	45.3
abbott-n	7.5	44.2	0.0	0.0	45.2
erdogan-n	7.4	43.9	0.0	0.0	44.9
airway-n	7.4	43.8	0.0	0.0	44.8
mp-n	7.4	43.7	0.0	0.0	44.7

cuba-n	7.3	43.6	0.0	0.0	44.6
jose-n	7.3	43.4	0.0	0.0	44.4
nato-n	7.3	43.3	0.0	0.0	44.3
everton-n	7.3	43.2	0.0	0.0	44.2
uefa-n	7.3	43.2	0.0	0.0	44.2
san-n	7.3	43.1	0.0	0.0	44.1
premier-j	29.8	177.0	1.9	3.0	44.0
pistorius-n	7.2	42.9	0.0	0.0	43.9
costa-n	7.2	42.8	0.0	0.0	43.8
mumbai-n	7.2	42.7	0.0	0.0	43.7
singapore-n	7.2	42.7	0.0	0.0	43.7
shinzo-n	7.1	42.3	0.0	0.0	43.3
abe-n	7.1	42.3	0.0	0.0	43.3
mcilroy-n	7.1	42.2	0.0	0.0	43.2
icc-n	7.1	42.1	0.0	0.0	43.1
usain-n	7.1	41.9	0.0	0.0	42.9
nicolas-n	7.0	41.7	0.0	0.0	42.7
thailand-n	6.9	41.1	0.0	0.0	42.1
sony-n	6.9	40.8	0.0	0.0	41.8
borussia-n	6.9	40.7	0.0	0.0	41.7
angeles-n	6.9	40.7	0.0	0.0	41.7
catholic-j	6.8	40.5	0.0	0.0	41.5
hezbollah-n	6.8	40.4	0.0	0.0	41.4
yemen-n	6.8	40.3	0.0	0.0	41.3
florida-n	6.8	40.2	0.0	0.0	41.2
imf-n	6.7	40.0	0.0	0.0	41.0
francois-n	6.6	39.3	0.0	0.0	40.3
tayyip-n	6.6	39.1	0.0	0.0	40.1
recep-n	6.6	39.1	0.0	0.0	40.1
nokia-n	6.6	39.0	0.0	0.0	40.0
warner-n	6.5	38.6	0.0	0.0	39.6
mali-n	6.5	38.6	0.0	0.0	39.6
saudi-n	6.4	38.2	0.0	0.0	39.2
kabul-n	6.4	38.1	0.0	0.0	39.1
sao-n	6.4	37.7	0.0	0.0	38.7
paulo-n	6.4	37.7	0.0	0.0	38.7
kong-n	6.3	37.4	0.0	0.0	38.4
taliban-j	6.3	37.4	0.0	0.0	38.4
roger-n	6.3	37.3	0.0	0.0	38.3
netherlands-n	6.3	37.3	0.0	0.0	38.3
shanghai-v	6.3	37.3	0.0	0.0	38.3
marc-n	6.3	37.3	0.0	0.0	38.3
wikileaks-n	6.2	36.9	0.0	0.0	37.9

bangkok-n	6.2	36.7	0.0	0.0	37.7
ukrainian-n	6.2	36.6	0.0	0.0	37.6
bradley-n	6.1	36.4	0.0	0.0	37.4
smith-n	6.1	36.3	0.0	0.0	37.3
rory-j	6.1	36.2	0.0	0.0	37.2
iphone-n	6.1	36.0	0.0	0.0	37.0
juan-j	6.0	35.8	0.0	0.0	36.8
xilai-n	6.0	35.7	0.0	0.0	36.7
uganda-n	6.0	35.7	0.0	0.0	36.7
farc-n	6.0	35.6	0.0	0.0	36.6
hague-n	6.0	35.5	0.0	0.0	36.5
haram-n	6.0	35.5	0.0	0.0	36.5
boko-n	6.0	35.5	0.0	0.0	36.5
djokovic-n	6.0	35.4	0.0	0.0	36.4
silvio-n	6.0	35.3	0.0	0.0	36.3
berlusconi-n	6.0	35.3	0.0	0.0	36.3
dan-n	5.9	35.2	0.0	0.0	36.2
steve-n	5.8	34.7	0.0	0.0	35.7
alex-n	5.8	34.6	0.0	0.0	35.6
warren-n	5.8	34.5	0.0	0.0	35.5
gp-n	5.8	34.4	0.0	0.0	35.4
chris-n	5.8	34.3	0.0	0.0	35.3
rome-n	5.8	34.3	0.0	0.0	35.3
organization-n	5.7	34.0	0.0	0.0	35.0
san-j	5.7	33.6	0.0	0.0	34.6
rodgers-n	5.6	33.5	0.0	0.0	34.5
brendan-n	5.6	33.5	0.0	0.0	34.5
kenya-n	5.6	33.5	0.0	0.0	34.5
berlin-n	5.6	33.4	0.0	0.0	34.4
panama-n	5.6	33.4	0.0	0.0	34.4
portugal-n	5.6	33.3	0.0	0.0	34.3
zuma-n	5.6	33.1	0.0	0.0	34.1
assad-n	5.5	32.9	0.0	0.0	33.9
sunni-n	5.5	32.5	0.0	0.0	33.5
aleppo-n	5.5	32.5	0.0	0.0	33.5
jones-n	5.5	32.5	0.0	0.0	33.5
dna-n	5.4	32.0	0.0	0.0	33.0
st-j	5.3	31.7	0.0	0.0	32.7
tim-n	5.3	31.5	0.0	0.0	32.5
bailout-n	14.6	86.4	1.1	1.7	32.4
ohio-n	5.3	31.4	0.0	0.0	32.4
indian-administered-j	5.3	31.3	0.0	0.0	32.3
edward-n	5.3	31.2	0.0	0.0	32.2

bashar-n	5.2	31.1	0.0	0.0	32.1
al-assad-n	5.2	31.1	0.0	0.0	32.1
ki-moon-n	5.2	31.1	0.0	0.0	32.1
messi-n	5.2	31.1	0.0	0.0	32.1
alastair-n	5.2	31.0	0.0	0.0	32.0
beijing-n	5.2	31.0	0.0	0.0	32.0
tripoli-n	5.2	30.9	0.0	0.0	31.9
apple-n	22.8	135.1	2.0	3.3	31.9
fulham-n	5.2	30.7	0.0	0.0	31.7
maduro-j	5.2	30.6	0.0	0.0	31.6
detroit-n	5.1	30.3	0.0	0.0	31.3
arabia-n	5.1	30.3	0.0	0.0	31.3
suarez-n	5.1	30.2	0.0	0.0	31.2
inter-v	5.1	30.2	0.0	0.0	31.2
de-x	5.1	30.1	0.0	0.0	31.1
kim-n	5.1	30.0	0.0	0.0	31.0
heathrow-n	5.0	29.8	0.0	0.0	30.8
qantas-n	5.0	29.8	0.0	0.0	30.8
prix-n	16.6	98.8	1.5	2.5	28.8
Pakistani-n	22.4	132.9	2.3	3.8	28.2
anti-government-j	18.3	108.8	2.1	3.4	25.2
unite-v	43.9	260.8	6.3	10.2	23.4
blackberry-n	10.0	59.6	1.0	1.6	23.0
us-n	702.0	4166.0	118.9	192.6	21.5
ashes-n	25.0	148.6	3.8	6.2	20.8
ousted-j	16.8	99.7	2.4	3.9	20.5
formula-n	8.8	51.9	1.1	1.8	19.1
rudd-n	8.1	48.2	1.0	1.6	18.8
villa-n	8.1	47.8	1.0	1.6	18.6
Greek-n	8.0	47.6	1.0	1.6	18.5
eurozone-n	36.3	215.6	7.0	11.3	17.6
plus-c	7.3	43.1	1.0	1.6	16.8
congress-n	16.3	96.5	3.0	4.9	16.6
democratic-j	11.5	68.3	2.0	3.2	16.4
shutdown-n	7.0	41.3	1.0	1.6	16.2
mare-n	7.0	41.8	1.0	1.7	16.1
ham-n	6.8	40.4	1.0	1.6	15.8
pm-n	54.2	321.9	12.1	19.7	15.6
wale-n	25.1	149.1	5.4	8.7	15.5
twitter-n	16.6	98.6	3.4	5.5	15.3
ford-n	6.4	37.9	1.0	1.6	14.9
Brussel-n	6.3	37.7	1.0	1.6	14.8
bale-n	6.8	40.6	1.1	1.8	14.7

Iraqi-n	10.5	62.4	2.0	3.3	14.7
north-western-j	7.4	43.7	1.3	2.0	14.7
Cairo-n	18.5	110.1	4.1	6.6	14.7
garment-n	6.3	37.3	1.0	1.6	14.6
kashmir-n	8.0	47.3	1.6	2.6	13.6
ex-president-n	7.3	43.3	1.4	2.3	13.6
negotiator-n	5.8	34.7	1.0	1.6	13.5
Hassan-n	6.5	38.3	1.2	1.9	13.5
Snowden-n	12.3	72.9	2.8	4.6	13.2
monetary-j	9.0	53.5	2.0	3.2	13.0
frank-j	5.6	33.1	1.0	1.6	13.0
rover-n	5.2	31.1	1.0	1.6	12.3
bitcoin-n	5.2	31.1	1.0	1.6	12.3
Somali-n	8.5	50.2	2.0	3.2	12.2
outlook-n	8.7	51.6	2.1	3.3	12.2
south-j	129.3	767.1	39.7	64.4	11.8
high-level-j	5.7	33.8	1.3	2.1	11.2
Philippine-n	12.3	73.0	3.5	5.7	11.1
fa-n	8.7	51.5	2.4	3.8	10.9
downgrade-v	5.4	32.3	1.3	2.1	10.9
ambush-n	5.2	31.0	1.2	2.0	10.8
demonstrator-n	6.5	38.7	1.7	2.7	10.8
treasury-n	5.2	31.2	1.2	2.0	10.7
index-n	9.1	54.0	2.6	4.2	10.7
xi-n	5.3	31.2	1.3	2.0	10.6
grand-j	27.3	161.7	8.9	14.4	10.5
Brazilian-n	18.1	107.5	5.8	9.3	10.5
commission-n	23.9	142.0	7.9	12.7	10.4
telecoms-n	7.3	43.1	2.0	3.2	10.4
cross-n	7.2	42.7	2.0	3.2	10.4
p-n	5.9	35.1	1.5	2.5	10.3
electronics-n	14.7	87.4	4.8	7.7	10.2
battery-n	6.0	35.6	1.6	2.6	10.1
Syrian-j	48.9	290.1	17.1	27.7	10.1
Lebanese-j	9.2	54.7	2.8	4.5	10.0
recession-n	11.9	70.5	3.8	6.2	9.9
second-largest-j	7.8	46.5	2.4	3.8	9.9
surveillance-n	14.8	88.1	5.0	8.1	9.8
president-n	328.0	1946.2	125.4	203.1	9.5
Egyptian-j	34.6	205.4	12.7	20.6	9.5
league-n	67.8	402.6	25.6	41.5	9.5
Israeli-n	16.6	98.3	5.9	9.6	9.4
restart-v	6.6	39.0	2.0	3.3	9.3

stimulus-n	15.9	94.4	5.8	9.4	9.1
slowdown-n	10.6	62.9	3.8	6.1	9.0
carmaker-n	13.7	81.3	5.2	8.4	8.8
tycoon-n	5.6	33.4	1.8	3.0	8.7
protester-n	50.9	301.9	21.0	34.0	8.7
evasion-n	8.8	52.0	3.2	5.1	8.6
leaker-n	6.6	39.3	2.3	3.7	8.6
stock-n	26.6	157.6	11.0	17.9	8.4
pray-v	6.2	36.9	2.2	3.5	8.4
depose-v	5.9	35.1	2.1	3.3	8.4
bankruptcy-n	9.4	55.6	3.6	5.8	8.3
radioactive-j	5.3	31.4	1.8	3.0	8.2
o-n	9.9	58.8	3.9	6.3	8.2
royal-j	17.9	105.9	7.6	12.3	8.1
prince-n	6.7	39.8	2.5	4.1	8.0
embassy-n	12.8	75.8	5.3	8.6	8.0
chemical-j	9.5	56.5	3.8	6.2	8.0
north-west-j	5.6	33.0	2.0	3.3	7.9
intelligence-n	23.2	137.7	10.3	16.7	7.8
prime-j	72.7	431.6	33.7	54.6	7.8
amnesty-n	11.3	67.1	4.8	7.8	7.8
closely-watched-j	6.1	36.0	2.4	3.8	7.7
association-n	12.1	71.5	5.2	8.5	7.7
smartphone-n	18.1	107.6	8.2	13.2	7.6
import-n	11.1	66.1	4.9	8.0	7.5
Edward-n	10.7	63.7	4.8	7.7	7.4
peninsula-n	5.3	31.6	2.1	3.4	7.4
escalate-v	8.6	51.0	3.7	6.0	7.4
navy-n	8.6	51.1	3.8	6.2	7.2
recapture-v	5.2	30.6	2.1	3.4	7.2
chemical-n	25.9	153.7	12.7	20.6	7.2
spy-v	14.6	86.5	6.9	11.2	7.2
jailed-j	6.0	35.5	2.5	4.1	7.2
tension-n	23.8	141.4	11.7	18.9	7.2
conspire-v	7.3	43.5	3.2	5.2	7.1
consumer-n	19.3	114.2	9.4	15.2	7.1
march-n	24.6	146.1	12.2	19.8	7.1
representative-n	6.9	40.9	3.0	4.9	7.0
bribery-n	8.7	51.8	4.0	6.5	7.0
unemployment-n	18.9	112.3	9.3	15.1	7.0
chancellor-n	13.1	77.6	6.4	10.4	6.9
province-n	27.8	164.9	14.3	23.2	6.9
open-a	5.7	33.8	2.5	4.1	6.8

Turkish-j	25.9	153.5	13.4	21.6	6.8
comply-v	6.2	36.6	2.8	4.5	6.8
poor-n	5.0	29.8	2.2	3.6	6.7
euro-n	7.9	47.1	3.8	6.2	6.7
boycott-v	6.0	35.7	2.8	4.5	6.6
Japanese-j	29.7	176.5	15.9	25.8	6.6
mark-n	22.2	132.0	11.8	19.1	6.6
peace-n	31.9	189.2	17.2	27.9	6.6
earning-n	7.4	43.8	3.6	5.8	6.6
palace-n	10.3	60.9	5.3	8.5	6.5
deflation-n	5.4	31.9	2.5	4.1	6.5
ash-n	16.9	100.1	9.0	14.6	6.5
bolt-n	7.1	41.9	3.5	5.7	6.4
Vatican-n	6.4	37.8	3.1	5.0	6.4
constitution-n	7.5	44.8	3.8	6.1	6.4
motor-n	13.6	80.5	7.3	11.8	6.4
central-j	86.5	513.0	49.5	80.1	6.3
electoral-j	5.5	32.7	2.7	4.3	6.3
Korean-j	28.7	170.0	16.1	26.2	6.3
Peru-n	5.2	31.0	2.5	4.1	6.3
federal-j	34.3	203.6	19.5	31.6	6.3
interim-j	7.3	43.3	3.8	6.1	6.2
south-a	19.0	112.5	10.9	17.6	6.1
stable-j	7.0	41.5	3.7	6.0	6.1
nuclear-j	44.3	263.0	26.4	42.8	6.0
weak-j	14.0	83.4	8.0	13.0	6.0
disgrace-v	6.3	37.3	3.3	5.4	6.0
square-n	5.5	32.9	2.9	4.7	6.0
rescuer-n	7.9	47.0	4.4	7.1	6.0
man-v	8.1	47.8	4.4	7.2	6.0
interior-j	5.4	31.9	2.8	4.6	5.9
decline-v	6.2	36.6	3.4	5.5	5.8
ocean-n	5.3	31.2	2.8	4.6	5.8
typhoon-n	9.2	54.5	5.3	8.6	5.8
patent-n	6.8	40.4	3.8	6.2	5.7
oversee-v	7.3	43.1	4.1	6.7	5.7
resolution-n	5.3	31.4	2.9	4.7	5.7
earth-n	6.4	38.1	3.6	5.9	5.7
broad-j	5.9	35.2	3.3	5.4	5.7
economy-n	90.4	536.6	58.0	94.0	5.7
weapon-n	38.4	227.8	24.4	39.5	5.6
agency-n	52.5	311.7	33.7	54.7	5.6
bombing-n	16.1	95.3	10.1	16.3	5.6

zone-n	16.1	95.7	10.1	16.4	5.6
denounce-v	5.5	32.5	3.1	5.1	5.5
transaction-n	6.4	38.1	3.8	6.1	5.5
manufacturing-n	19.0	112.5	12.1	19.6	5.5
bull-n	12.5	74.3	7.9	12.8	5.5
Nazi-n	5.4	32.3	3.2	5.2	5.4
mainly-a	6.6	39.4	4.0	6.5	5.4
prosecutor-n	26.9	159.9	17.8	28.8	5.4
output-n	7.3	43.2	4.4	7.2	5.4
underline-v	6.6	39.2	4.0	6.5	5.4
ambassador-n	8.5	50.7	5.3	8.6	5.4
bank-n	147.1	872.9	100.2	162.2	5.4
tiger-n	9.0	53.6	5.7	9.2	5.3
disputed-j	11.8	70.3	7.7	12.4	5.3
bay-n	7.5	44.4	4.7	7.6	5.2
maker-n	34.9	206.9	23.8	38.6	5.2
February-n	9.3	54.9	6.0	9.7	5.2
plot-v	6.3	37.3	3.9	6.4	5.2
opposition-n	56.5	335.5	39.5	63.9	5.2
landmark-j	8.7	51.5	5.6	9.1	5.2
constitutional-j	6.2	37.0	3.9	6.4	5.1
moon-n	5.5	32.8	3.5	5.6	5.1
king-n	11.4	67.9	7.7	12.5	5.1
western-j	30.3	179.9	21.4	34.6	5.1
Iranian-j	10.7	63.6	7.3	11.8	5.0
arsenal-n	26.2	155.5	18.6	30.1	5.0
capital-n	96.2	571.0	69.8	113.1	5.0
bond-n	7.9	47.1	5.4	8.8	4.9
insurgent-n	8.4	49.6	5.7	9.3	4.9
summon-v	5.6	33.0	3.7	5.9	4.9
ruler-n	5.6	33.3	3.7	6.0	4.9
riot-n	14.7	87.3	10.5	17.0	4.9
consortium-n	5.4	32.0	3.5	5.7	4.9
trading-n	20.1	119.4	14.6	23.6	4.9
fugitive-j	8.7	51.5	6.1	9.8	4.8
all-time-j	6.9	41.2	4.8	7.7	4.8
euros-n	17.9	106.1	13.1	21.2	4.8
lord-n	6.8	40.2	4.7	7.6	4.8
credit-n	17.0	101.0	12.5	20.3	4.8
lawmaker-n	9.3	54.9	6.6	10.7	4.8
north-j	64.0	379.9	48.9	79.2	4.7
deficit-n	11.6	68.8	8.5	13.7	4.7
evacuation-n	6.8	40.2	4.8	7.7	4.7

terrorist-j	10.4	61.8	7.6	12.3	4.7
good-n	5.8	34.2	4.0	6.5	4.7
Australian-j	49.4	293.0	38.1	61.7	4.7
corruption-n	25.9	154.0	19.9	32.2	4.7
west-j	26.2	155.7	20.3	32.8	4.6
least-j	164.1	974.0	129.5	209.8	4.6
Muslim-n	27.4	162.4	21.2	34.4	4.6
accusation-n	8.4	50.0	6.3	10.2	4.6
shell-n	5.7	33.7	4.1	6.6	4.6
rating-n	9.4	55.7	7.1	11.6	4.5
executive-j	13.8	82.2	10.8	17.5	4.5
supporter-n	24.8	147.1	20.0	32.3	4.4
gain-n	9.2	54.6	7.2	11.6	4.4
wound-v	11.4	67.5	9.0	14.5	4.4
civilian-n	14.5	85.8	11.6	18.8	4.4
oust-v	9.4	55.6	7.4	11.9	4.4
resignation-n	8.3	49.3	6.5	10.5	4.4
further-a	7.9	46.8	6.2	10.0	4.3
detention-n	7.5	44.6	5.9	9.5	4.3
leak-v	18.2	107.8	15.0	24.2	4.3
queen-n	9.4	55.9	7.5	12.2	4.3
text-n	11.6	68.6	9.3	15.1	4.3
bomb-n	58.1	344.9	49.2	79.6	4.3
north-n	16.9	100.3	14.0	22.6	4.3
growth-n	62.3	369.8	53.0	85.8	4.3
foreign-j	55.4	329.1	47.3	76.7	4.3
billionaire-n	6.1	36.4	4.8	7.8	4.2
state-n	172.1	1021.3	148.6	240.7	4.2
revise-v	7.4	43.9	6.0	9.7	4.2
mission-n	20.1	119.3	17.0	27.6	4.2
forecast-n	19.5	115.8	16.6	26.8	4.2
two-day-j	5.7	33.5	4.5	7.2	4.2
indicate-v	10.9	64.8	9.1	14.7	4.2
build-up-n	7.6	44.9	6.2	10.0	4.2
airline-n	31.5	186.9	27.2	44.1	4.2
coalition-n	12.7	75.3	10.7	17.3	4.2
earthquake-n	8.1	48.1	6.7	10.8	4.2
Chinese-j	71.4	423.7	62.7	101.6	4.1
same-sex-n	7.7	45.5	6.5	10.5	4.1
quarter-n	48.6	288.1	43.4	70.2	4.1
Russian-j	48.6	288.3	43.6	70.6	4.0
Chinese-n	8.4	49.6	7.1	11.6	4.0
protest-n	74.5	442.1	68.0	110.1	4.0

magazine-n	8.9	52.8	7.7	12.5	4.0
Palestinian-n	7.5	44.7	6.5	10.5	4.0
coup-n	8.5	50.5	7.4	12.0	4.0
regain-v	8.4	49.8	7.3	11.8	4.0
spokesman-n	7.6	45.0	6.5	10.6	4.0
derail-v	6.5	38.6	5.6	9.0	4.0
trade-n	39.5	234.3	36.2	58.6	3.9
software-n	7.0	41.4	6.0	9.8	3.9
collect-v	8.3	49.4	7.3	11.8	3.9
envoy-n	6.7	40.0	5.8	9.4	3.9
lion-n	7.0	41.4	6.1	9.8	3.9
French-n	19.3	114.4	17.6	28.5	3.9
currency-n	17.7	105.3	16.2	26.2	3.9
activist-n	24.1	143.1	22.2	35.9	3.9
shortly-a	5.0	29.8	4.3	6.9	3.9
Thai-n	8.1	48.2	7.2	11.6	3.9
secretary-n	35.2	209.0	32.6	52.9	3.9
wood-n	8.4	49.7	7.4	12.0	3.9
mobile-j	29.8	177.1	27.7	44.9	3.9
northern-j	54.4	322.9	50.9	82.5	3.9
reserve-n	18.0	107.0	16.6	26.9	3.9
ministry-n	11.6	69.0	10.5	17.1	3.9
mislead-v	5.7	33.9	5.0	8.0	3.9
wing-n	6.5	38.5	5.7	9.3	3.9
left-wing-j	6.0	35.8	5.3	8.6	3.8
ten-n	10.0	59.3	9.1	14.7	3.8
apparently-a	9.7	57.8	8.8	14.3	3.8
jump-n	13.7	81.0	12.6	20.4	3.8
gunman-n	19.4	115.0	18.2	29.5	3.8
quarterly-j	6.7	39.7	6.0	9.7	3.8
export-n	13.2	78.5	12.3	19.9	3.8
official-n	220.4	1308.0	212.4	344.1	3.8
violation-n	6.4	37.7	5.7	9.2	3.8
strongly-a	7.1	42.3	6.4	10.4	3.8
lung-n	6.4	37.8	5.7	9.3	3.8
investor-n	17.7	104.9	16.7	27.0	3.8
shareholder-n	9.1	53.8	8.4	13.5	3.8
news-n	33.4	198.1	32.0	51.8	3.8
guard-n	12.0	71.5	11.3	18.3	3.8
file-v	10.5	62.4	9.8	15.9	3.8
talk-n	115.0	682.5	112.3	182.0	3.7
programme-n	41.9	248.8	41.0	66.5	3.7
explode-v	10.4	62.0	9.9	16.1	3.7

white-j	15.0	88.9	14.4	23.4	3.7
rebel-j	12.1	71.7	11.6	18.9	3.7
inflation-n	14.1	83.8	13.8	22.4	3.6
brand-n	8.1	48.0	7.8	12.7	3.6
amid-i	89.0	528.4	90.6	146.8	3.6
negotiation-n	10.5	62.1	10.3	16.6	3.6
affair-n	9.1	53.9	8.9	14.4	3.6
carrier-n	10.9	64.6	10.7	17.4	3.6
movement-n	10.2	60.5	10.0	16.2	3.6
million-x	38.9	231.0	39.8	64.4	3.5
clothing-n	6.0	35.7	5.8	9.4	3.5
wound-n	6.9	40.9	6.7	10.9	3.5
cite-v	6.0	35.5	5.8	9.4	3.5
electronic-j	8.4	49.6	8.3	13.4	3.5
temperature-n	7.4	44.0	7.3	11.9	3.5
retail-j	8.8	52.0	8.8	14.2	3.5
south-n	10.4	61.6	10.5	17.0	3.5
general-j	31.1	184.3	32.3	52.3	3.5
reform-n	14.8	87.6	15.1	24.5	3.5
stake-n	13.6	80.8	14.0	22.6	3.5
leak-n	7.1	42.0	7.1	11.4	3.5
master-n	5.9	35.1	5.8	9.5	3.4
explosion-n	19.9	117.8	20.7	33.5	3.4
finance-n	14.5	86.1	15.0	24.4	3.4
share-n	61.3	363.7	65.1	105.4	3.4
policeman-n	10.9	64.7	11.2	18.2	3.4
product-n	14.3	84.6	14.9	24.1	3.4
parliament-n	31.5	187.1	33.5	54.2	3.4
November-n	9.8	58.3	10.1	16.4	3.4
exchange-n	12.8	75.8	13.3	21.6	3.4
natural-j	6.5	38.7	6.6	10.7	3.4
pace-n	26.7	158.7	28.5	46.1	3.4
hostage-n	6.2	36.5	6.3	10.2	3.4
ethnic-j	6.7	39.6	6.9	11.1	3.4
uphold-v	6.1	35.9	6.2	10.0	3.4
Bangladesh-n	7.5	44.2	7.8	12.6	3.3
diplomatic-j	8.4	49.8	8.8	14.3	3.3
slightly-a	7.3	43.0	7.6	12.3	3.3
bn-n	110.2	654.1	122.4	198.3	3.3
outrage-n	5.3	31.2	5.4	8.8	3.3
army-n	47.6	282.3	52.9	85.7	3.3
political-j	45.7	271.2	50.9	82.4	3.3
troubled-j	6.9	40.7	7.3	11.8	3.3

low-n	8.6	50.7	9.2	15.0	3.2
recovery-n	29.3	173.7	32.7	53.0	3.2
debt-n	19.1	113.4	21.3	34.5	3.2
union-n	41.3	245.1	46.6	75.6	3.2
global-j	36.8	218.1	41.7	67.6	3.2
state-owned-j	5.4	32.2	5.8	9.4	3.2
newspaper-n	16.7	99.3	18.8	30.5	3.2
slow-j	8.0	47.6	8.8	14.3	3.2
unrest-n	7.9	47.1	8.8	14.2	3.2
dollar-n	9.2	54.9	10.3	16.7	3.2
lower-v	5.9	34.7	6.4	10.4	3.1
winter-n	10.4	61.7	11.7	19.0	3.1
preview-n	6.1	36.0	6.7	10.9	3.1
mosque-n	5.6	33.4	6.2	10.1	3.1
reportedly-a	18.5	110.0	21.6	35.0	3.1
refugee-n	12.2	72.5	14.1	22.9	3.1
troop-n	27.0	160.2	31.7	51.3	3.1
apparent-j	8.1	48.2	9.3	15.0	3.1
security-n	82.6	490.1	98.1	159.0	3.1
Somalia-n	9.4	55.9	10.9	17.6	3.1
cleric-n	6.7	40.0	7.8	12.6	3.0
justice-n	14.9	88.5	17.7	28.7	3.0
rule-n	42.9	254.3	51.8	84.0	3.0
suicide-n	28.0	166.3	33.8	54.7	3.0
criminal-j	18.4	109.0	22.0	35.6	3.0
settlement-n	9.1	54.0	10.7	17.3	3.0
minister-n	147.5	875.2	179.7	291.1	3.0
virtual-j	5.3	31.4	6.0	9.8	3.0
several-j	36.9	219.3	45.2	73.2	3.0
inmate-n	7.8	46.4	9.2	15.0	3.0
world-n	245.4	1456.5	303.1	491.0	3.0
detain-v	20.1	119.2	24.5	39.6	3.0
bomber-n	11.2	66.5	13.5	21.8	3.0
plant-n	24.7	146.9	30.4	49.2	2.9
European-j	84.2	499.7	104.4	169.2	2.9
resolve-v	10.1	59.6	12.2	19.7	2.9
economic-j	52.5	311.6	65.8	106.6	2.9
southern-j	39.0	231.2	48.9	79.2	2.9
chase-n	6.4	38.1	7.7	12.5	2.9
citizen-n	14.0	83.0	17.4	28.2	2.9
cell-n	6.3	37.4	7.7	12.4	2.9
supply-v	6.1	36.3	7.5	12.1	2.8
industrial-j	15.3	91.0	19.5	31.5	2.8

red-j	20.7	123.1	26.5	42.9	2.8
humanitarian-j	5.8	34.5	7.1	11.6	2.8
demand-n	22.4	132.6	28.7	46.5	2.8
luxury-n	5.6	33.3	6.9	11.2	2.8
devastating-j	5.7	33.9	7.1	11.4	2.8
landslide-n	6.2	36.7	7.7	12.5	2.8
oil-n	49.3	292.6	64.1	103.8	2.8
Asia-n	8.9	52.7	11.2	18.2	2.8
labour-n	10.5	62.4	13.4	21.8	2.8
clash-n	42.6	253.0	55.8	90.4	2.8
revenue-n	11.7	69.5	15.1	24.4	2.8
ship-n	21.3	126.5	27.8	45.0	2.8
technology-n	18.7	110.7	24.3	39.3	2.8
report-n	149.5	887.0	197.5	319.9	2.8
founder-n	10.9	64.9	14.1	22.9	2.8
force-n	83.4	495.1	110.8	179.4	2.7
diplomat-n	10.4	61.4	13.4	21.8	2.7
collapse-n	13.4	79.8	17.6	28.5	2.7
giant-n	36.9	219.0	49.0	79.4	2.7
aircraft-n	21.2	125.8	28.0	45.3	2.7
financial-j	27.6	164.1	36.9	59.8	2.7
lawsuit-n	8.4	50.1	11.0	17.8	2.7
corporation-n	6.2	36.8	8.0	12.9	2.7
recall-v	15.6	92.4	20.7	33.6	2.7
storm-v	6.2	36.8	8.0	13.0	2.7
kill-v	275.2	1633.4	373.2	604.6	2.7
island-n	32.8	194.8	44.2	71.7	2.7
commander-n	7.0	41.7	9.2	14.9	2.7
Asian-j	6.6	38.9	8.6	13.9	2.7
trial-n	68.4	405.6	93.4	151.2	2.7
file-n	5.8	34.5	7.6	12.3	2.7
deputy-j	12.5	74.0	16.8	27.2	2.7
control-v	12.7	75.4	17.1	27.7	2.7
banking-n	20.4	121.3	27.8	45.0	2.7
authority-n	51.1	303.5	70.3	113.8	2.7
rain-n	20.0	118.7	27.3	44.2	2.6
Beijing-n	5.9	35.1	7.8	12.6	2.6
analyst-n	7.2	42.5	9.6	15.5	2.6
overtake-v	5.7	33.9	7.6	12.3	2.6
rise-v	68.8	408.2	95.4	154.6	2.6
activity-n	21.6	128.3	29.8	48.2	2.6
sustain-v	5.6	33.2	7.4	12.0	2.6
profit-n	54.9	325.6	76.6	124.1	2.6

powerful-j	15.9	94.5	22.0	35.7	2.6
election-n	76.2	452.5	107.4	173.9	2.6
monitor-v	7.7	45.8	10.5	17.1	2.6
erupt-v	5.3	31.4	7.1	11.5	2.6
Venezuela-n	8.1	47.8	11.0	17.9	2.6
casualty-n	5.4	32.0	7.3	11.8	2.6
agree-v	79.5	471.5	113.1	183.1	2.6
separate-j	6.1	36.3	8.4	13.6	2.6
Kiev-n	5.4	32.2	7.4	12.1	2.5
check-v	5.6	33.4	7.8	12.6	2.5
market-n	50.9	302.3	73.2	118.6	2.5
Italian-j	27.0	160.5	38.7	62.8	2.5
commentary-n	7.4	44.0	10.4	16.8	2.5
its-d	475.2	2820.0	690.9	1119.2	2.5
voter-n	11.3	66.9	16.0	26.0	2.5
cocaine-n	5.3	31.2	7.3	11.8	2.5
govern-v	8.5	50.5	12.0	19.5	2.5
allegedly-a	14.1	83.7	20.2	32.8	2.5
spill-n	7.5	44.6	10.6	17.2	2.5
resume-v	18.1	107.4	26.3	42.6	2.5
sector-n	19.4	115.3	28.4	45.9	2.5
north-a	7.6	45.3	10.9	17.7	2.5
governor-n	15.0	89.0	21.9	35.5	2.5
suburb-n	5.6	33.0	7.9	12.8	2.5
commercial-j	6.3	37.4	9.0	14.6	2.5
relation-n	7.8	46.5	11.3	18.4	2.5
consecutive-j	8.1	48.3	11.8	19.1	2.5
tax-n	33.8	200.7	50.1	81.1	2.5
rebel-n	50.9	302.1	75.6	122.5	2.5
parliamentary-j	8.8	52.0	12.7	20.6	2.4
live-j	21.0	124.8	31.2	50.6	2.4
broadcaster-n	7.0	41.3	10.1	16.4	2.4
Swiss-j	8.0	47.7	11.8	19.1	2.4
military-j	59.5	353.0	89.4	144.9	2.4
swear-v	8.5	50.6	12.6	20.3	2.4
document-n	7.9	47.0	11.7	18.9	2.4
ice-n	6.3	37.3	9.2	14.9	2.4
Olympic-j	13.3	78.7	19.8	32.1	2.4
period-n	15.0	89.2	22.6	36.6	2.4
form-v	10.9	64.6	16.2	26.3	2.4
sharply-a	8.3	49.3	12.3	20.0	2.4
pole-n	7.0	41.4	10.3	16.7	2.4
slow-v	10.0	59.4	15.0	24.3	2.4

ruling-n	21.5	127.3	32.6	52.8	2.4
killing-n	14.6	86.6	22.1	35.8	2.4
restriction-n	7.8	46.5	11.7	19.0	2.4
withdraw-v	15.5	91.8	23.5	38.0	2.4
storm-n	20.3	120.2	30.9	50.0	2.4
wave-n	7.4	43.6	11.0	17.8	2.4
standard-n	10.7	63.7	16.3	26.4	2.4
forecast-v	6.6	39.3	9.9	16.1	2.4
act-n	10.5	62.3	16.0	25.9	2.4
armed-j	19.0	112.9	29.2	47.3	2.4
arm-n	14.8	87.7	22.6	36.7	2.4
district-n	5.7	33.8	8.5	13.8	2.4
broadcast-v	6.2	37.0	9.4	15.2	2.4
previous-j	7.4	44.2	11.3	18.3	2.3
department-n	11.1	65.8	17.1	27.7	2.3
camp-n	16.3	96.7	25.3	41.0	2.3
secret-j	13.3	78.9	20.6	33.4	2.3
effort-n	30.3	179.6	47.4	76.8	2.3
phone-n	37.1	220.4	58.4	94.6	2.3
wall-n	7.1	41.9	10.8	17.5	2.3
rebel-held-n	5.5	32.9	8.5	13.8	2.3
rate-n	44.3	262.8	70.7	114.6	2.3
scandal-n	12.3	72.8	19.4	31.4	2.3
plot-n	8.8	52.4	13.9	22.5	2.3
Indian-j	61.6	365.5	99.1	160.5	2.3
investment-n	23.8	141.0	38.0	61.6	2.3
hearing-n	9.9	58.7	15.7	25.4	2.3
official-j	26.4	156.9	42.4	68.7	2.3
say-v	1180.1	7003.0	1918.9	3108.5	2.3
punishment-n	5.7	33.8	8.9	14.5	2.2
retailer-n	14.8	87.9	23.8	38.5	2.2
model-n	5.7	33.9	9.0	14.5	2.2
material-n	7.2	42.6	11.4	18.4	2.2
survivor-n	9.1	53.8	14.5	23.4	2.2
trip-n	13.2	78.5	21.3	34.5	2.2
conviction-n	11.5	68.2	18.5	30.0	2.2
party-n	45.7	271.0	74.7	121.0	2.2
six-x	72.2	428.2	118.4	191.8	2.2
general-n	9.6	57.1	15.5	25.1	2.2
formally-a	8.3	49.3	13.4	21.6	2.2
presidential-j	23.6	140.2	38.7	62.7	2.2
American-n	22.9	136.1	37.6	60.9	2.2
release-n	21.5	127.6	35.2	57.1	2.2

marathon-n	6.8	40.1	10.9	17.6	2.2
decline-n	7.3	43.1	11.7	19.0	2.2
missile-n	8.3	49.3	13.5	21.8	2.2
politician-n	18.0	106.5	29.6	47.9	2.2
personnel-n	5.2	31.0	8.4	13.6	2.2
increasingly-a	6.5	38.7	10.6	17.1	2.2
march-v	6.5	38.6	10.5	17.1	2.2
summit-n	16.6	98.6	27.5	44.5	2.2
spark-v	20.5	121.5	34.0	55.2	2.2
sentence-v	21.9	129.7	36.4	58.9	2.2
initial-j	6.5	38.4	10.5	17.1	2.2
territory-n	10.2	60.4	16.8	27.2	2.2
chief-n	16.7	98.9	27.8	45.0	2.2
direct-j	5.6	33.5	9.2	14.9	2.2
employee-n	10.5	62.0	17.3	28.1	2.2
key-j	37.0	219.8	62.4	101.1	2.2
Libyan-j	7.3	43.5	12.1	19.6	2.2
drone-n	9.1	54.0	15.1	24.5	2.2
launch-n	8.5	50.3	14.1	22.8	2.2
speech-n	9.1	54.1	15.2	24.6	2.2
organisation-n	11.0	65.0	18.3	29.7	2.2
measure-n	21.0	124.9	35.5	57.5	2.2
although-i	5.7	33.9	9.4	15.3	2.1
Canadian-j	14.0	83.0	23.6	38.2	2.1
main-j	29.6	175.7	50.5	81.8	2.1
Xinjiang-n	5.5	32.5	9.1	14.7	2.1
operate-v	9.3	54.9	15.6	25.2	2.1
relate-v	8.4	49.7	14.1	22.8	2.1
curb-v	8.2	48.5	13.8	22.3	2.1
order-n	25.7	152.5	44.0	71.3	2.1
succeed-v	6.7	39.8	11.3	18.2	2.1
helicopter-n	13.6	80.6	23.2	37.5	2.1
policy-n	25.3	150.1	43.4	70.4	2.1
west-n	7.6	45.2	12.9	20.8	2.1
nation-n	33.9	201.3	58.6	94.9	2.1
crisis-n	38.4	228.1	66.5	107.7	2.1
drop-n	12.7	75.4	21.8	35.3	2.1
resign-v	20.1	119.3	34.7	56.2	2.1
survey-n	24.1	143.2	41.7	67.6	2.1
protection-n	10.6	63.1	18.2	29.5	2.1
fuel-n	9.2	54.7	15.8	25.6	2.1
court-n	131.4	779.8	229.9	372.4	2.1
enter-v	16.0	95.1	27.8	45.0	2.1

possible-j	17.6	104.5	30.6	49.6	2.1
attempt-n	30.9	183.3	54.0	87.5	2.1
climate-n	6.6	39.3	11.4	18.4	2.1
restrict-v	6.2	36.5	10.5	17.1	2.1
dead-a	24.4	144.9	42.8	69.4	2.1
price-n	36.4	215.8	64.1	103.8	2.1
real-j	16.0	94.8	28.0	45.3	2.1
capture-v	10.4	61.5	18.1	29.3	2.1
country-n	153.5	910.8	272.1	440.8	2.1
manufacturer-n	6.9	41.1	12.0	19.4	2.1
post-n	17.3	102.6	30.5	49.4	2.1
bill-n	21.9	130.2	39.0	63.1	2.0
source-n	10.3	61.1	18.2	29.5	2.0
hold-n	6.3	37.2	11.0	17.8	2.0
expect-v	70.3	417.2	126.7	205.2	2.0
company-n	83.8	497.4	152.0	246.3	2.0
computer-n	12.8	76.0	23.1	37.5	2.0
high-j	55.9	331.7	102.1	165.4	2.0
operator-n	9.6	57.2	17.4	28.2	2.0
rescue-n	18.3	108.8	33.4	54.1	2.0
trigger-v	5.6	33.2	10.0	16.2	2.0
Spanish-j	19.7	117.0	36.1	58.5	2.0
budget-n	17.8	105.8	32.6	52.9	2.0
crackdown-n	8.0	47.4	14.5	23.4	2.0
international-j	82.8	491.5	153.1	248.0	2.0
low-j	27.4	162.3	50.5	81.7	2.0
potentially-a	5.4	31.9	9.7	15.7	2.0
ease-v	19.6	116.3	36.2	58.6	2.0
rival-j	14.1	83.8	26.0	42.1	2.0
human-j	25.1	149.1	46.5	75.3	2.0
mass-j	8.5	50.2	15.5	25.1	2.0
datum-n	21.8	129.1	40.4	65.4	2.0
war-n	46.4	275.4	86.6	140.2	2.0
since-i	80.1	475.2	149.7	242.6	2.0
violent-j	9.8	58.2	18.1	29.3	2.0
people-n	341.6	2026.9	640.2	1037.0	2.0
unless-i	7.7	45.6	14.1	22.9	2.0
transfer-v	5.5	32.6	10.0	16.3	1.9
spread-v	6.1	36.3	11.2	18.2	1.9
Janeiro-n	7.9	46.7	14.5	23.5	1.9
giant-j	20.7	123.0	38.9	62.9	1.9
statement-n	8.4	50.1	15.7	25.4	1.9
high-profile-j	5.4	32.2	10.0	16.2	1.9

petition-n	5.3	31.4	9.8	15.8	1.9
fine-n	8.2	48.6	15.3	24.8	1.9
conclude-v	8.1	47.9	15.1	24.4	1.9
decade-n	23.0	136.7	43.6	70.7	1.9
fresh-j	22.2	131.9	42.1	68.3	1.9
progress-n	12.4	73.6	23.4	37.9	1.9
government-n	223.1	1324.0	426.0	690.0	1.9
militia-n	5.9	34.8	10.9	17.7	1.9
black-j	16.9	100.4	32.1	52.0	1.9
regulator-n	15.6	92.6	29.6	48.0	1.9
asset-n	5.7	33.7	10.6	17.2	1.9
abandon-v	10.2	60.7	19.4	31.5	1.9
plane-n	36.9	218.9	70.8	114.7	1.9
factory-n	21.2	126.0	40.7	65.9	1.9
detail-n	14.0	83.2	26.8	43.4	1.9
hundred-n	34.7	205.9	66.8	108.2	1.9
Rafael-n	8.9	52.7	16.9	27.4	1.9
fast-j	18.0	107.1	34.7	56.2	1.9
certain-j	6.8	40.1	12.8	20.8	1.9
militant-j	17.7	105.3	34.2	55.4	1.9
actress-n	8.9	53.1	17.1	27.7	1.9
heavy-j	18.0	106.7	34.7	56.3	1.9
agreement-n	17.4	103.1	33.6	54.5	1.9
law-n	41.6	247.0	81.2	131.5	1.9
mayor-n	9.3	55.1	17.9	29.0	1.9
marriage-n	10.8	63.8	20.9	33.8	1.9
suspend-v	25.9	153.6	50.6	82.0	1.9
collapse-v	8.7	51.4	16.9	27.3	1.9
defence-n	28.7	170.3	56.6	91.7	1.8
Thailand-n	5.7	34.0	11.1	18.0	1.8
halt-v	13.6	80.9	26.8	43.4	1.8
leader-n	142.0	842.4	283.0	458.4	1.8
responsible-j	7.5	44.2	14.6	23.7	1.8
remain-v	43.5	258.3	86.9	140.8	1.8
spending-n	12.9	76.8	25.7	41.7	1.8
demand-v	10.9	64.7	21.6	35.0	1.8
crucial-j	7.2	42.6	14.2	22.9	1.8
die-n	5.8	34.6	11.4	18.5	1.8
national-j	43.5	257.9	87.2	141.3	1.8
newly-a	6.7	39.6	13.2	21.3	1.8
researcher-n	16.5	97.6	32.9	53.3	1.8
funeral-n	9.7	57.3	19.2	31.2	1.8
right-n	34.4	204.4	69.4	112.5	1.8

limit-n	6.8	40.3	13.5	21.8	1.8
January-n	7.4	44.0	14.8	23.9	1.8
controversial-j	24.3	144.1	48.9	79.3	1.8
private-j	16.8	99.5	33.8	54.7	1.8
city-n	146.1	867.1	296.7	480.6	1.8
emerge-v	20.3	120.6	41.1	66.6	1.8
add-v	22.8	135.4	46.3	74.9	1.8
fighter-n	14.9	88.5	30.2	48.8	1.8
prison-n	40.0	237.3	81.4	131.9	1.8
migrant-n	10.9	64.6	22.0	35.7	1.8
within-i	15.1	89.9	30.8	49.9	1.8
request-n	7.8	46.5	15.8	25.6	1.8
aid-n	16.9	100.3	34.5	55.9	1.8
dispute-n	12.2	72.2	24.8	40.2	1.8
link-v	18.3	108.8	37.6	60.9	1.8
eastern-j	32.0	189.9	65.8	106.7	1.8
prisoner-n	15.8	93.6	32.4	52.4	1.8
October-n	8.5	50.5	17.4	28.1	1.8
gay-j	13.2	78.1	27.1	43.9	1.8
loss-n	32.1	190.6	66.5	107.7	1.8
cost-n	21.8	129.4	45.1	73.0	1.8
jail-n	29.3	173.6	60.6	98.2	1.8
production-n	9.7	57.8	20.0	32.4	1.8
fault-n	5.5	32.7	11.2	18.2	1.8
nine-x	21.5	127.5	44.6	72.2	1.8
equipment-n	5.1	30.4	10.5	16.9	1.8
buy-v	30.6	181.5	63.8	103.3	1.7
respond-v	5.8	34.7	12.0	19.4	1.7
transfer-n	12.8	75.7	26.5	42.9	1.7
angry-j	6.4	38.2	13.2	21.4	1.7
kidnap-v	12.3	72.7	25.5	41.3	1.7
poll-n	16.5	97.9	34.4	55.7	1.7
conference-n	6.7	39.7	13.8	22.4	1.7
dead-j	39.4	233.8	82.6	133.9	1.7
cabinet-n	12.1	71.7	25.2	40.8	1.7
ty-n	11.6	68.7	24.1	39.1	1.7
express-v	6.8	40.6	14.2	23.0	1.7
border-n	22.9	136.0	48.3	78.3	1.7
violence-n	28.0	166.1	59.1	95.7	1.7
candidate-n	13.7	81.5	28.9	46.8	1.7
abduct-v	10.8	64.0	22.6	36.6	1.7
dope-v	6.0	35.4	12.4	20.2	1.7
interest-n	22.3	132.3	47.2	76.5	1.7

huge-j	24.4	145.0	51.8	83.9	1.7
building-n	24.0	142.3	50.8	82.3	1.7
blame-v	17.9	106.0	37.8	61.3	1.7
gun-n	15.5	92.2	32.9	53.2	1.7
allow-v	42.2	250.6	89.8	145.5	1.7
vow-v	9.5	56.1	20.0	32.4	1.7
website-n	12.0	71.0	25.4	41.2	1.7
suspected-j	22.3	132.4	47.7	77.3	1.7
surge-n	9.2	54.4	19.5	31.5	1.7
rally-n	15.2	90.4	32.6	52.8	1.7
also-a	18.8	111.7	40.4	65.5	1.7
warn-v	83.4	495.2	180.1	291.8	1.7
fall-v	67.0	397.4	144.7	234.3	1.7
investigator-n	9.0	53.3	19.2	31.1	1.7
include-v	57.9	343.5	125.1	202.6	1.7
burn-v	6.4	38.0	13.6	22.0	1.7
response-n	12.3	72.8	26.5	42.9	1.7
gather-v	11.1	65.7	23.9	38.7	1.7
jump-v	7.4	44.2	16.0	25.9	1.7
African-j	19.2	113.7	41.6	67.4	1.7
deep-j	6.0	35.9	13.0	21.0	1.7
injure-v	51.8	307.1	113.0	183.0	1.7
deadline-n	6.2	36.9	13.4	21.7	1.7
continue-v	60.7	360.0	133.1	215.6	1.7
landing-n	5.5	32.4	11.8	19.0	1.7
religious-j	5.1	30.4	11.0	17.9	1.7
push-v	13.3	79.2	29.2	47.3	1.7
group-n	109.4	649.5	241.3	390.8	1.7
suspect-v	14.9	88.4	32.6	52.9	1.7
aim-v	26.8	158.8	58.9	95.4	1.7
value-n	9.6	57.0	21.0	34.1	1.7
purchase-n	5.2	31.1	11.4	18.4	1.7
sentence-n	20.2	119.8	44.5	72.1	1.7
alleged-j	30.1	178.6	66.5	107.7	1.7
card-n	6.6	39.4	14.5	23.5	1.6
reporter-n	5.8	34.5	12.7	20.5	1.6
spy-n	7.0	41.3	15.3	24.7	1.6
fix-v	7.6	45.3	16.9	27.3	1.6
play-off-n	6.3	37.6	13.9	22.6	1.6
boost-v	31.8	188.4	71.0	115.0	1.6
favourite-n	7.5	44.5	16.6	27.0	1.6
office-n	27.8	165.1	62.4	101.1	1.6
widespread-j	6.2	36.6	13.7	22.2	1.6

sale-n	66.0	391.9	148.9	241.2	1.6
executive-n	20.6	122.2	46.3	74.9	1.6
chief-j	51.1	303.1	115.2	186.7	1.6
outside-i	21.5	127.4	48.3	78.2	1.6
level-n	25.1	149.1	56.8	92.0	1.6
payment-n	12.1	71.9	27.3	44.3	1.6
issue-n	21.3	126.6	48.4	78.3	1.6
former-j	149.9	889.3	341.4	553.1	1.6
apologise-v	21.3	126.4	48.3	78.3	1.6
prepare-v	29.7	176.5	67.6	109.6	1.6
scientist-n	26.6	157.7	60.5	98.0	1.6
information-n	16.0	95.0	36.4	59.0	1.6
Tottenham-n	7.0	41.4	15.9	25.7	1.6
soon-a	8.8	52.4	20.1	32.6	1.6
as-i	713.3	4232.9	1647.9	2669.5	1.6
infection-n	5.7	33.9	13.0	21.1	1.6
vote-n	30.1	178.4	69.4	112.5	1.6
miner-n	5.6	33.4	12.8	20.8	1.6
fund-n	18.6	110.4	43.0	69.6	1.6
fraud-n	20.8	123.3	48.0	77.8	1.6

Appendix 2

Tag keyword extraction in the comparison between the NTC and the bw_14

	NTC		**bw_14**		
tag	**ARF**	**ARF/mill**	**ARF_ref**	**ARF_ref/mill**	**Score**
FW	14.0	82.9	11.4	18.4	4.3
SYM	175.6	1042.2	193.9	314.1	3.3
JJS	489.0	2901.8	873.9	1415.7	2.0
VVP	1320.4	7835.7	2574.5	4170.4	1.9
VVZ	2964.9	17594.8	6196.9	10038.6	1.8
PP	1605.7	9528.6	3456.0	5598.4	1.7
NNS	8623.9	51176.4	20336.9	32944.4	1.6
NN	29474.1	174906.7	69979.4	113361.7	1.5
VBP	496.7	2947.5	1278.2	2070.6	1.4
JJ	9088.5	53933.2	23759.9	38489.3	1.4
IN/that	186.0	1103.5	490.8	795.0	1.4
WDT	290.5	1723.7	782.5	1267.6	1.4
VBZ	866.4	5141.2	2359.6	3822.4	1.3
JJR	265.2	1573.8	770.0	1247.3	1.3
SENT	6516.0	38667.3	19296.6	31259.1	1.2

CC	1475.7	8757.5	4483.9	7263.5	1.2
POS	1452.3	8618.6	4478.4	7254.7	1.2
VVG	2110.8	12526.3	6589.5	10674.5	1.2
DT	10040.6	59583.3	31687.4	51331.4	1.2
VV	2285.3	13561.8	7844.0	12706.7	1.1
PP$	1181.0	7008.2	4059.1	6575.5	1.1
MD	687.3	4078.7	2369.3	3838.1	1.1
CD	2617.8	15534.6	9025.1	14619.9	1.1
IN	15509.3	92036.5	53547.0	86742.3	1.1
TO	1662.0	9862.9	5930.0	9606.2	1.0

Appendix 3

Tag keyword extraction in the comparison between the NTC and the bwh_14

	NTC		**bwh_14**		
tag	**ARF**	**ARF/mill**	**ARF_ref**	**ARF_ref/mill**	**Score**
VBN	189.4	1123.8	3.0	23.8	45.4
IN/that	186.0	1103.5	4.0	31.7	33.8
EX	29.3	173.8	1.0	7.9	19.6
SENT	6516.0	38667.3	315.0	2497.2	15.5
WP$	15.9	94.1	1.0	7.9	10.7
WDT	290.5	1723.7	22.0	174.4	9.8
DT	10040.6	59583.3	1021.0	8094.2	7.4
PP$	1181.0	7008.2	131.0	1038.5	6.7
VHP	216.9	1287.3	24.0	190.3	6.7
VBP	496.7	2947.5	63.0	499.4	5.9
VHZ	420.8	2497.4	57.0	451.9	5.5
VBG	108.7	644.9	18.0	142.7	4.5
VHG	6.3	37.3	1.0	7.9	4.3
PP	1605.7	9528.6	292.0	2314.9	4.1
VH	65.7	389.9	14.0	111.0	3.5
VBZ	866.4	5141.2	197.0	1561.8	3.3
SYM	175.6	1042.2	53.0	420.2	2.5
JJS	489.0	2901.8	150.0	1189.2	2.4
VHD	37.9	225.0	12.0	95.1	2.4
VBD	186.9	1109.4	65.0	515.3	2.2
CC	1475.7	8757.5	519.0	4114.5	2.1
FW	14.0	82.9	5.0	39.6	2.1
RBR	89.5	531.2	38.0	301.3	1.8
VVG	2110.8	12526.3	925.0	7333.2	1.7
RBS	40.1	238.0	21.0	166.5	1.4

IN	15509.3	92036.5	8275.0	65602.2	1.4
JJ	9088.5	53933.2	4854.0	38481.4	1.4
MD	687.3	4078.7	378.0	2996.7	1.4
VVN	2670.2	15845.7	1495.0	11852.0	1.3
JJR	265.2	1573.8	155.0	1228.8	1.3
WP	203.1	1205.3	125.0	991.0	1.2
VB	270.8	1607.1	167.0	1323.9	1.2
POS	1452.3	8618.6	924.0	7325.3	1.2
RB	1265.0	7506.6	851.0	6746.5	1.1
NN	29474.1	174906.7	20038.0	158856.5	1.1
NNS	8623.9	51176.4	5923.0	46956.1	1.1

Appendix 4

Tag keyword extraction in the comparison between the NTC and the bwl_14

	NTC		**bwl_14**		
tag	**ARF**	**ARF/mill**	**ARF_ref**	**ARF_ref/mill**	**Score**
FW	14.0	82.9	7.0	14.3	5.5
VVZ	2964.9	17594.8	2660.0	5415.8	3.2
SYM	175.6	1042.2	160.0	325.8	3.2
VVP	1320.4	7835.7	1250.0	2545.0	3.1
JJS	489.0	2901.8	769.0	1565.7	1.9
NN	29474.1	174906.7	50264.0	102337.9	1.7
NNS	8623.9	51176.4	15230.0	31008.4	1.7
PP	1605.7	9528.6	3233.0	6582.4	1.4
JJ	9088.5	53933.2	19247.0	39187.1	1.4
JJR	265.2	1573.8	644.0	1311.2	1.2
VBP	496.7	2947.5	1212.0	2467.6	1.2
VBZ	866.4	5141.2	2178.0	4434.4	1.2
POS	1452.3	8618.6	3667.0	7466.0	1.2
WDT	290.5	1723.7	759.0	1545.3	1.1
IN/that	186.0	1103.5	487.0	991.5	1.1
VV	2285.3	13561.8	6236.0	12696.6	1.1
VVG	2110.8	12526.3	5787.0	11782.4	1.1
CC	1475.7	8757.5	4115.0	8378.2	1.0
SENT	6516.0	38667.3	18769.0	38213.9	1.0
CD	2617.8	15534.6	7550.0	15371.9	1.0

Appendix 5

List of all the adverbs occurring in the NTC with their raw frequency

ADVERB	RAW FREQUENCY
not	246
ahead	77
back	74
ago	46
nearly	46
still	46
dead	42
now	41
almost	38
reportedly	37
south	37
about	36
up	36
again	33
also	33
as	33
allegedly	28
too	28
down	26
ever	26
only	26
there	25
just	24
very	21
yet	19
apparently	18
far	18
july	18
seriously	15
soon	15
away	14
even	14
formally	14
never	14
once	14
potentially	14
sharply	14
slightly	13
well	13
mainly	12

newly	12
north	12
so	12
strongly	12
forward	11
over	11
east	10
illegally	10
shortly	10
around	9
increasingly	9
much	9
open	9
secretly	9
first	8
officially	7
previously	7
probably	7
successfully	7
unexpectedly	7
along	6
already	6
angrily	6
badly	6
deeply	6
extremely	6
highly	6
long	6
no	6
politically	6
twice	6
unanimously	6
all	5
closely	5
critically	5
deliberately	5
due	5
enough	5
finally	5
home	5
mostly	5
narrowly	5

publicly	5
significantly	5
worldwide	5
alive	4
before	4
dangerously	4
hotly	4
immediately	4
in	4
inadvertently	4
little	4
many	4
n't	4
out	4
partly	4
prior	4
quickly	4
rather	4
respectively	4
safely	4
slowly	4
systematically	4
temporarily	4
together	4
widely	4
accidentally	3
aground	3
any	3
artificially	3
automatically	3
currently	3
early	3
fatally	3
forcibly	3
formerly	3
fully	3
globally	3
hugely	3
incredibly	3
indefinitely	3
initially	3
instead	3

largely	3
live	3
overnight	3
routinely	3
suddenly	3
sure	3
utterly	3
abroad	2
absolutely	2
accidently	2
apart	2
aside	2
behind	2
below	2
briefly	2
broad	2
categorically	2
close	2
completely	2
controversially	2
desperately	2
directly	2
double	2
dramatically	2
easily	2
elsewhere	2
eventually	2
everywhere	2
fine	2
genetically	2
heavily	2
ill	2
inappropriately	2
intentionally	2
lawfully	2
medically	2
mistakenly	2
modestly	2
needlessly	2
off	2
often	2
openly	2

overwhelmingly	2
precisely	2
predominantly	2
properly	2
provisionally	2
recklessly	2
repeatedly	2
seemingly	2
sexually	2
since	2
some	2
swiftly	2
totally	2
usually	2
voluntarily	2
west	2
2012	1
ablaze	1
abruptly	1
accurately	1
afterwards	1
alone	1
always	1
annually	1
anxiously	1
anywhere	1
apace	1
barely	1
belatedly	1
bis	1
brutally	1
by	1
centrally	1
competitively	1
daily	1
definitely	1
downstream	1
eastwards	1
effectively	1
entirely	1
excessively	1
falsely	1

fast	1
favourably	1
financially	1
flatly	1
forcefully	1
functionally	1
geographically	1
grossly	1
halfway	1
inside	1
internationally	1
last	1
late	1
left	1
legally	1
likely	1
low	1
meanwhile	1
mentally	1
merely	1
midwest	1
nearby	1
normally	1
on	1
overly	1
overseas	1
painfully	1
partially	1
particularly	1
past	1
peacefully	1
personally	1
physically	1
possibly	1
purely	1
purportedly	1
quietly	1
quite	1
racially	1
recently	1
remotely	1
smoothly	1

soundly	1
specially	1
strategically	1
though	1
unduly	1
unethically	1
unfairly	1
unlawfully	1
unwittingly	1
upright	1
upwards	1
urgently	1
vertically	1
weekly	1
wide	1

Appendix 6

Pointers to newsworthiness in the NTC

NEWS VALUES	lempos
Not clearly related to a specific news value: function words	the-x (ARF score: 5295.7); a-x (ARF score: 3259.5); in-i (ARF score: 3040.9); of-i (ARF score: 2851.9); to-x (ARF score: 1662); and-c (ARF score: 1293.5); for-i (ARF score: 1064.3); to-i (ARF score: 922.6); on-i (ARF score: 829.7); as-i (ARF score: 713.3); at-i (ARF score: 704.5); by-i (ARF score: 597.8); with-i (ARF score: 549.2); an-x (ARF score: 537.8); its-d (ARF score: 475.2); from-i (ARF score: 436.6); his-d (ARF score: 397.2); it-d (ARF score: 364.9); over-i (ARF score: 323.2); he-d (ARF score: 232.9); their-d (ARF score: 198.7); that-i (ARF score: 186); who-x (ARF score: 167.1); than-i (ARF score: 166.9); that-x (ARF score: 152.7); not-a (ARF score: 145.6); into-i (ARF score: 135.4); which-x (ARF score: 124.1); they-d (ARF score: 113.3); but-c (ARF score: 104.8); amid-i (ARF score: 89); out-x (ARF score: 84.5); up-x (ARF score: 81.9); between-i (ARF score: 58.1); him-d (ARF score: 56.6); her-d (ARF score: 54.3).
Not clearly related to a specific news value: nouns	us-n (ARF score: 702); world-n (ARF score: 245.4); china-n (ARF score: 202.9); country-n (ARF score: 153.5); city-n (ARF score: 146.1); talk-n (ARF score: 115); group-n (ARF score: 109.4); australia-n

	(ARF score: 105.8); capital-n (ARF score: 96.2); syria-n (ARF score: 92.8); economy-n (ARF score: 90.4); election-n (ARF score: 76.2); india-n (ARF score: 69.9); cup-n (ARF score: 69.8); part-n (ARF score: 69.2); trial-n (ARF score: 68.4); league-n (ARF score: 67.8); champion-n (ARF score: 67.1); sale-n (ARF score: 66); car-n (ARF score: 62.2); share-n (ARF score: 61.3); korea-n (ARF score: 59.1); test-n (ARF score: 58); japan-n (ARF score: 57.9); russia-n (ARF score: 55.7); home-n (ARF score: 54.8); plan-n (ARF score: 53.6); power-n (ARF score: 52.7); protester-n (ARF score: 50.9); rebel-n (ARF score: 50.9); oil-n (ARF score: 49.3).
Not clearly related to a specific news value: verbs	be-v (ARF score: 2215.6); have-v (ARF score: 772.8); take-v (ARF score: 115.1); make-v (ARF score: 98.5); could-x (ARF score: 74.7); go-v (ARF score: 70); may-x (ARF score: 68.1); use-v (ARF score: 63.7); accord-v (ARF score: 62.2); include-v (ARF score: 57.9); can-x (ARF score: 49); help-v (ARF score: 49).
Not clearly related to a specific news value: adjectives	south-j (ARF score: 129.3); international-j (ARF score: 82.8); Chinese-j (ARF score: 71.4); north-j (ARF score: 64); Indian-j (ARF score: 61.6); foreign-j (ARF score: 55.4); northern-j (ARF score: 54.4); economic-j (ARF score: 52.5); Australian-j (ARF score: 49.4); Syrian-j (ARF score: 48.9).
Potential pointers to Eliteness	say-v (ARF score: 1180.1); president-n (ARF score: 328); government-n (ARF score: 223.1); official-n (ARF score: 220.4); police-n (ARF score: 180.7); state-n (ARF score: 172.1); former-j (ARF score: 149.9); report-n (ARF score: 149.5); minister-n (ARF score: 147.5); bank-n (ARF score: 147.1); leader-n (ARF score: 142); court-n (ARF score: 131.4); central-j (ARF score: 86.5); European-j (ARF score: 84.2); company-n (ARF score: 83.8); force-n (ARF score: 83.4); international-j (ARF score: 82.8); security-n (ARF score: 82.6); prime-j (ARF score: 72.7); tell-v (ARF score: 72.1); medium-n (ARF score: 69.5); report-v (ARF score: 68.7); bbc-n (ARF score: 65.4); firm-n (ARF score: 62.6); military-j (ARF score: 59.5); obama-n (ARF score: 57.5); eu-n (ARF score: 55.2); pm-n (ARF score: 54.2); agency-n (ARF score: 52.5); authority-n (ARF score: 51.1); chief-j (ARF score: 51.1); market-n (ARF score: 50.9).

Potential pointers to Personalisation	people-n (ARF score: 341.6); man-n (ARF score: 81.7); child-n (ARF score: 52.5).
Potential pointers to Proximity	england-n (ARF score: 97.1); Uk-n (ARF score: 96.1); European-j (ARF score: 84.2); British-j (ARF score: 74.2); eu-n (ARF score: 55.2).
Potential pointers to Timeliness	after-i (ARF score: 678.8); will-x (ARF score: 335.6); year-n (ARF score: 302.3); day-n (ARF score: 125.5); month-n (ARF score: 105.2); last-j (ARF score: 82.7); since-i (ARF score: 80.1); agree-v (ARF score: 79.5); time-n (ARF score: 77); find-v (ARF score: 76.1); announce-v (ARF score: 73); expect-v (ARF score: 70.3); second-j (ARF score: 68.4); begin-v (ARF score: 68.4); week-n (ARF score: 67.9); hold-v (ARF score: 67.5); during-i (ARF score: 67.5); see-v (ARF score: 66.3); late-j (ARF score: 64.6); continue-v (ARF score: 60.7); become-v (ARF score: 60.7); end-v (ARF score: 57.2); call-v (ARF score: 57.2); face-v (ARF score: 53.3); next-j (ARF score: 53); this-x (ARF score: 52.6); give-v (ARF score: 50.7); reach-v (ARF score: 49.5).
Potential pointers to Novelty	new-j (ARF score: 288.4); first-j (ARF score: 226); time-n (ARF score: 77); find-v (ARF score: 76.1); announce-v (ARF score: 73); late-j (ARF score: 64.6).
Potential pointers to Superlativeness	two-x (ARF score: 214.8); least-j (ARF score: 164.1); more-j (ARF score: 141.3); one-x (ARF score: 122.1); bn-n (ARF score: 110.2); three-x (ARF score: 104.3); m-n (ARF score: 103.6); six-x (ARF score: 72.2); five-x (ARF score: 67); about-i (ARF score: 65); late-j (ARF score: 64.6); big-j (ARF score: 64.5); some-x (ARF score: 59.9); four-x (ARF score: 58.1); high-j (ARF score: 55.9); thousand-n (ARF score: 55).
Potential pointers to Negativity and Impact	kill-v (ARF score: 275.2); against-i (ARF score: 170.8); die-v (ARF score: 119.4); attack-n (ARF score: 100.9); win-v (ARF score: 100.5); deal-n (ARF score: 87.2); leave-v (ARF score: 86.5); warn-v (ARF score: 83.4); set-v (ARF score: 82.1); follow-v (ARF score: 75.9); death-n (ARF score: 74.8); protest-n (ARF score: 74.5); rise-v (ARF score: 68.8); fall-v (ARF score: 67); growth-n (ARF score: 62.3); hit-v (ARF score: 60.1); accuse-v (ARF score: 58.6); bomb-n (ARF score: 58.1); charge-n (ARF score: 57.1); lead-v (ARF score: 56.6); opposition-n (ARF score: 56.5); beat-v (ARF score: 55.2); profit-n (ARF score: 54.9); show-v (ARF score: 54.4); record-n

	(ARF score: 53.3); `despite-i` (ARF score: 52.2); `injure-v` (ARF score: 51.8); `final-j` (ARF score: 49.5).

References

Aarts, J. and Meijs, W. (eds) 1984. *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. Amsterdam: Rodopi.

Adolphs, S. 2006. *Introducing Electronic Text Analysis. A Practical Guide for Language and Literary Studies*. London & New York: Routledge.

Anthony, L. 2014. AntConc (3.4.0w) [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.laurenceanthony.net/

Aston, G. 2001. Text categories and corpus users: A response to David Lee. *Language Learning & Technology* 5 (3): 73–76.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London & New York: Continuum International Publishing Group.

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P. 2013. *Using Corpora to Analyze Gender*. London: Bloomsbury.

Baker, P., Hardie, A. and McEnery, T. 2006. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P., Gabrielatos, C. and McEnery, T. 2013. *Discourse analysis and media attitudes: the representation of Islam in the British press*. Cambridge: Cambridge University Press.

Bauman, Z. 2000. *Liquid modernity*. Cambridge: Polity Press.

Bauman, Z. 2005. *Liquid life*. Cambridge: Polity Press.

Bednarek, M. 2006. *Evaluation in Media Discourse. Analysis of a Newspaper Corpus*. London & New York: Continuum.

Bednarek, M. 2015 (in press). Voices and values in the news: News media talk, news values and attribution. *Discourse, Context & Media*. DOI: http://dx.doi.org/10.1016/j.dcm.2015.11.004

Bednarek, M. and Caple, H. 2012a. *News discourse*. London & New York: Bloomsbury.

Bednarek, M. and Caple, H. 2012b. 'Value added': Language, image and news values. *Discourse, context & media* 1 (2): 103–113.

Bednarek, M. and Caple, H. 2014. Why do news values matter? Towards a new methodological framework for analyzing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society* 20 (10): 1–24.

Bell, A. 1991. *The Language of News Media*. Oxford: Blackwell Publishers.

Berber-Sardinha, T. 2000. Comparing corpora with WordSmith Tools: How large must the reference corpus be?. In A. Kilgariff and T. Berber-Sardinha (eds) *Proceedings of the Workshop on Comparing Corpora. The Association for Computational Linguistics*: 7–13.

Berkenkotter, C. and Huckin, T.N. 1995. *Genre knowledge in disciplinary communication: Cognition/culture/power*. New Jersey: Lawrence Erlbaum Associates.

Bhatia, V.K. 1993. *Analysing Genre: Language Use in Professional Settings*. London: Longman (reprinted as a Pearson Education print on demand edition in 2013).

Bhatia, V.K. 1996. Methodological issues in Genre Analysis. *Hermes, Journal of Linguistics* 16: 39–59.

Bhatia, V.K. 2000. Genres in Conflict. In A. Trosborg (ed.) *Analysing Professional Genres*. Amsterdam & Philadelphia: John Benjamins, 147–161.

Bhatia, V.K. 2002. Applied genre analysis: a multi-perspective model. *Ibérica* 4: 3–19.

Bhatia, V.K. 2004. *Worlds of written discourse: A genre-based view*. London: Continuum International (reprinted in 2014 in London & New York: Bloomsbury).

Bhatia, V.K. 2007. Interdiscursivity in critical genre analysis. In A. Bonini, D.C. Figueiredo and F. Rauen (eds) *Proceedings from the 4th International Symposium on Genre Studies (SIGET)*. Tubarão: Unisul, 391–400. Available online at http://linguagem.unisul.br/paginas/ensino/pos/linguagem/eventos/cd/English/36i.pdf

Bhatia, V.K. 2008. Towards critical genre analysis. In V.K. Bhatia, J. Flowerdew and R.H. Jones (eds) *Advances in Discourse Studies*. London & New York: Routledge, 166–177.

Bhatia, V.K. 2012. Critical reflections on genre analysis. *Ibérica* 24: 17–28.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 1993a. The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities* 26: 331–345.

Biber, D. 1993b. Representativeness in corpus design. *Literary and linguistic computing* 8 (4): 243–257.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus Linguistics. Investigating Language Structures and Use*. Cambridge: Cambridge University Press.

Bivens, R. 2014. *Digital currents: How technology and the public are shaping TV news*. Toronto: University of Toronto Press.

Blain, C. 2002. *Television news crawls and their effects on memory of the verbal message*. Unpublished Master's Thesis. Manhattan, Kansas: Kansas State University.

Blain, C. and Meeds, R. 2004. *Effects of Television News Crawls on Viewers' Memory for Audio Information in Newscasts*. Unpublished manuscript. Manhattan, Kansas: Kansas State University.

Bonini, A. 2010. Critical Genre Analysis and professional practice: The case of public contests to select professors for Brazilian public universities. *Linguagem em (Dis)curso* 10 (3): 485–510. Available online at http://linguagem.unisul.br/paginas/ensino/pos/linguagem/linguagem-em-discurso/1003/100303.pdf

Bowker, L. and Pearson, J. 2002. *Working with specialized language: A practical guide to using corpora*. London & New York: Routledge.

Bull, A. 2010. *Multimedia Journalism: A Practical Guide*. London & New York: Routledge.

Burnard, L. 2005. Metadata for corpus work. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 30–46. Available online at http://ahds.ac.uk/linguistic-corpora/

Caple, H. and Bednarek, M. 2015. Rethinking news values: What a discursive approach can tell us about the construction of news discourse and news photography. *Journalism*: 1–22. Available at

http://jou.sagepub.com/content/early/2015/02/02/1464884914568078

Chovanec, J. 2003. The mixing of modes as a means of resolving the tension between involvement and detachment in news headlines. *Brno Studies in English* 29 (1): 52–66. Available online at http://www.phil.muni.cz/plonedata/wkaa/BSE/BSE_2003-29_Scan/BSE_29_05.pdf

Chovanec, J. 2014. *Pragmatics of Tense and Time in News: From canonical headlines to online news texts*. Amsterdam: John Benjamins Publishing Company.

Christ, O. 1994. A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX '94, 3rd Conference on Computational Lexicography and Text Research*, 23–32.

Christ, O. and Schulze, B.M. 1994. *The CQP User's Manual*. Stuttgart: Institute of Natural Languages, University of Stuttgart.

Christensen, L. and Johnson, B. 2000. *Educational research: Quantitative, qualitative, and mixed approaches* (5th ed., 2014). Los Angeles, London, New Delhi, Singapore & Washington: Sage.

Coffey, A.J. and Cleary, J. 2008. Valuing New Media Spaces: Are Cable Network News Crawls Cross-promotional Agents?. *Journalism & Mass Communication Quarterly* 85 (4): 894–912.

Coffey, A.J. and Cleary, J. 2011. Promotional practices of cable news networks: A comparative analysis of new and traditional spaces. *International Journal on Media Management* 13 (3): 161–176.

Cohen, S. and Young, J. (eds) 1973. *The manufacture of news. Deviance, social problems and the mass media*. London: Constable.

Conrad, S. and Biber, D. 2001. Multi-dimensional methodology and the dimensions of register variation in English. In S. Conrad and D. Biber (eds) *Variation in English: Multi-dimensional studies*. London: Longman, 13–42.

Cotter, C. 2010. *News Talk: Investigating the Language of Journalism*. Cambridge: Cambridge University Press.

Crisell, A. 1997. *An Introductory History of British Broadcasting* (2nd ed., 2002). London & New York: Routledge.

Davies, M. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25 (4): 447–464.

Danesi, M. 2002. *Understanding Media Semiotics*. London: Hodder Arnold.

Declerck, R. 1991. *Tense in English. Its Structure and Use in Discourse*. London & New York: Routledge.

Deese, W.B. 2011. Streaming Messages. In E.F. Provenzo *et al.* (eds) *Multiliteracies: Beyond Text and the Written Word*. Charlotte, NC: Information Age Publishing, 73–76.

Deuze, M. 2008. The changing context of news work: Liquid journalism and monitorial citizenship. *International Journal of Communication* 2: 848–865.

Diamond, E. 1980. *Good News, Bad News*. Cambridge, Massachusetts & London, England: The MIT Press.

Epstein, E.J. 1973. *News from nowhere*. New York: Vintage Books.

Epstein, E.J. 1975. *Between Fact and Fiction: The Problem of Journalism*. New York: Vintage Books.

Fairclough, N. 1989. *Language and Power*. London: Longman.

Fairclough, N. 1992. *Discourse and Social Change* (15th ed., 2013). Cambridge: Polity Press.

Fairclough, N. 1995a. *Critical Discourse Analysis: The critical study of language*. London & New York: Longman.

Fairclough, N. 1995b. *Media discourse*. London: Hodder Arnold.

Fairclough, N. 2003. *Analysing Discourse: Textual Analysis for Social Research*. London & New York: Routledge.

Fairclough, N. 2011. Discursive hybridity and social change in Critical Discourse Analysis. In S. Sarangi, V. Polese and G. Caliendo (eds) *Genre(s) on the move. Hybridization and discourse change in specialized communication*. Napoli: Edizioni Scientifiche Italiane, 11–26.

Fillmore, C. 1992. Corpus Linguistics or Computer-aided Armchair Linguistics. In J. Svartvik (ed.) *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 13–38.

Firth, J. 1957. *Papers in Linguistics*. Oxford: Oxford University Press.

Fishman, M. 1980. *Manufacturing the News*. Austin & London: University of Texas Press.

Flowerdew, L. 2004. The argument for using English specialized corpora to understand academic and professional language. In U. Connor and T.A. Upton (eds) *Discourse in the professions. Perspectives from corpus linguistics*. Amsterdam: John Benjamins, 11–33.

Flowerdew, L. 2008. Corpora and context in professional writing. In V.K. Bhatia, J. Flowerdew and R.H. Jones (eds) *Advances in Discourse Studies*. London & New York: Routledge, 115–127.

Francis, W.N. 1992. Language corpora B.C. In J. Svartvik (ed.) *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 17–32.

Galtung, J. and Ruge, M.H. 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus Crises in four Norwegian newspapers. *Journal of Peace Research* 2 (1): 64–90.

Gans, H.J. 1979. *Deciding What's News. A Study of* CBS Evening News, NBC Nightly News, Newsweek, *and* Time (10th ed., 2004). Evanston, Illinois: Northwestern University Press.

Garside, R., Leech, G. and Sampson, G. (eds) 1987. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Garside, R. and Smith, N. 1997. A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech and T. McEnery (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, 102–121.

Glasgow University Media Group 1976. *Bad News*. London: Routledge & Kegan Paul.

Glasgow University Media Group 1980. *More Bad News*. London: Routledge & Kegan Paul.

Glasgow University Media Group 1982. *Really Bad News*. London: Writers and Readers.

Glasgow University Media Group 1985. *War and Peace News*. Philadelphia: Open University Press.

Granger, S. 2003. The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37 (3): 538–546.

Gray, J. 2006. *Watching with* The Simpsons*: Television, parody, and intertextuality*. New York & London: Routledge.

Gries, S.T. 2011. Methodological and interdisciplinary stance in Corpus Linguistics. In G. Barnbrook, V. Viana and S. Zyngier (eds) *Perspectives on Corpus Linguistics*. Amsterdam & Philadelphia: John Benjamins, 81–98.

Gries, S.T. and Newman, J. 2013. Creating and using corpora. In R.J. Podesva and D. Sharma (eds) *Research methods in linguistics*. Cambridge: Cambridge University Press, 257–287.

Halliday, M.A.K. 1985. *An Introduction to Functional Grammar* (4th ed., 2014). Revised by C.M.I.M Matthiessen. London & New York: Routledge.

Halliday, M.A.K. 1991. Corpus studies and probabilistic grammar. In K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London & New York: Longman, 30–40.

Halliday, M.A.K. 1992. Language as system and language as instance: the corpus as a theoretical construct. In J. Svartvik (ed.) *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 61–77.

Halliday, M.A.K. 2005. *Computational and Quantitative Studies* [Volume 6 in the *Collected Works of M.A.K. Halliday*, edited by J. Webster]. London: Continuum.

Halliday, M.A.K. and Hasan, R. 1989. *Language, Context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halloran, J.D., Elliott, P. and Murdock, G. 1970. *Demonstrations and Communication: A Case Study*. Harmondsworth: Penguin Books.

Handford, M. 2010. What can a corpus tell us about specialist genres?. In A. O'Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*. London & New York: Routledge, 255–269.

Hardie, A. 2014. Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal* 38 (1): 73–103.

Hart, R.P. 1986. Of Genre, computers and the Reagan inaugural. In H.W. Simons and A.A. Aghazarian (eds) *Form, Genre and the Study of Political Discourse*. Columbia: University of South Carolina Press, 278–298.

Hartley, J. 1982. *Understanding News* (3rd ed., 1994). London and New York: Routledge.

Helios Software Solutions 2014. TextPad. The Text Editor for Windows (version 7.2.0) [software]. Preston, Lancashire: Helios Software Solutions.

Hoey, M. 2001. *Textual Interaction. An introduction to written discourse analysis*. London & New York: Routledge.

Hoey, M. 2005. *Lexical priming*. London & New York: Routledge.

Hoover, D.L., Culpeper, J. and O'Halloran, K. 2014. *Digital Literary Studies. Corpus Approaches to Poetry, Prose, and Drama*. New York & London: Routledge.

Huddleston, R. and Pullum, G.K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Isani, S. 2011. Of headlines & headlinese: Towards distinctive linguistic and pragmatic genericity. *ASp, la revue du GERAS* 60: 81–102. Accessed through *ASp* [Online] 60: 1–18, available at `http://asp.revues.org/2523`

Jindal, A., Tiwari, A. and Ghosh, H. 2011. Efficient and language independent news story segmentation for telecast news videos. *Multimedia (ISM), 2011 IEEE International Symposium*: 458–463.

Johnson, W., Fairbanks, H., Mann, M.B. and Chotlos, J.W. 1944. Studies in language behavior. *Psychological Monographs* 56 (2): 1–111.

Josephson, S. and Holmes, M.E. 2006. Clutter or content? How on-screen enhancements affect how TV viewers scan and what they learn. *Proceedings of the 2006 symposium on Eye tracking research & applications*: 155–162.

Kanoksilapatham, B. 2003. *A corpus-based investigation of scientific research articles: Linking move analysis with multidimensional analysis*. Unpublished Doctoral Dissertation. Washington, D.C.: Georgetown University.

Katz, S.M. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering* 2 (1): 15–59.

Keefe-Feldman, M. 2007. *The cable news ticker, viewer comprehension and information overload: less is more*. Unpublished Master's Thesis. Washington, DC: University of Georgetown.

Kilgarriff, A. 2009. Simple maths for keywords. In M. Mahlberg, V. González-Díaz and C. Smith (eds) *Proceedings of Corpus Linguistics Conference CL2009*, University of Liverpool, UK. Available online at http://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc

Kilgarriff, A. and Grefenstette, G. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics* 29 (3): 333–347.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. The Sketch Engine. In G. Williams and S. Vessier (eds) *Proceedings of the Eleventh EURALEX International Congress: EURALEX 2004*. Lorient: Université de Bretagne-Sud, 105–16.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubícek, M., Kovár, V., Michelfeit, J., Rychlý, P. and Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography* 1 (1): 7–36.

Knight, D. 2011. The future of multimodal corpora. *Revista Brasileira de Linguística Aplicada* 11 (2): 391–415.

Kress, G. 1987. Genre in a social theory of language: A reply to John Dixon. In I. Reid (ed.) *The place of genre in learning: Current debates*. Geelong, Australia: Deakin University Press, 35–45.

Kučera, H. and Francis, W. 1967. *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Kytö, M. 1996. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. Helsinki: University of Helsinki. Available online at http://clu.uni.no/icame/manuals/HC/INDEX.HTM

Lee, D.Y.W. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5 (3): 37–72. Available online at

http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1615&context=artspapers

Lee, D.Y.W. 2008. Corpora and discourse analysis. New ways of doing old things. In V.K. Bhatia, J. Flowerdew and R.H. Jones (eds) *Advances in Discourse Studies*. London: Routledge, 86–99.

Lee, D.Y.W. and Swales, J.M. 2006. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes* 25: 56–75.

Leech, G. 1991. The state of the art in Corpus Linguistics. In K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London & New York: Longman, 8–29.

Leech, G. 1993. Corpus annotation schemes. *Literary and linguistic computing* 8 (4): 275–281.

Leech, G. 2005. Adding Linguistic Annotation. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 17–29. Available online at http://ahds.ac.uk/linguistic-corpora/

Louw, B. 2000. Contextual prosodic theory: bringing semantic prosodies to life. In C. Heffer, H. Sauntson and G. Fox (eds) *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham, 48–94 (reprinted in *Texto!* 13 (1), available online at http://www.revue-texto.net/docannexe/file/124/louw_prosodie.pdf).

Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19 (2): 313–330.

Mardth, I. 1980. *Headlinese: On the Grammar of English Front Page Headlines*. Lund: CWK Gleerup.

Matsukawa, R., Miyata, Y. and Ueda, S. 2009. Information redundancy effect on watching TV news: Analysis of eye tracking data and examination of the contents. *Library and Information Science* 62: 193–205.

McEnery, T. and Gabrielatos, C. 2006. English Corpus Linguistics. In B. Aarts and A. McMahon (eds) *The Handbook of English Linguistics*. Oxford: Blackwell, 33–71.

McEnery, T. and Hardie, A. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. 1996. *Corpus Linguistics. An Introduction* (2nd ed., 2001). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. 2006. *Corpus Based Language Studies: An Advanced Resource Book*. London & New York: Routledge.

McIntyre, D. 2013. Corpora and literature. In C. Chappelle (ed.) *Encyclopedia of Applied Linguistics*. Oxford: Wiley Blackwell, 1–6.

McKendry, J. 2012. *One Times Square: A Century of Change at the Crossroads of the World*. Boston: David R. Godine Publisher.

Meech, P. 1999. Watch this space: the on-air marketing communications of UK television. *International Journal of Advertising* 18: 291–304.

Miller, C.R. 1994. Genre as Social Action. In A. Freedman and P. Medway (eds) *Genre and the New Rhetoric*. London: Taylor & Francis, 20–36 (originally printed in 1984 in *Quarterly Journal of Speech* 70: 151–167).

Montgomery, M. 2007. *The Discourse of Broadcast News: A Linguistic Approach*. New York: Routledge.

Nuance Communications 2013. Dragon NaturallySpeaking (Premium Edition 12.5) [software]. Burlington, MA: Nuance Communications.

O'Donnell, M.B., Scott, M., Mahlberg, M. and Hoey, M. 2012. Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory* 8 (1): 73–101.

O'Hagan, M. 2010. Japanese TV Entertainment: Framing Humour with Open Caption Telop. In D. Chiaro (ed.) *Translation, Humour and the Media (Volume 2)*. London & New York: Continuum, 72–88.

Osgood, C.E. and Walker, E.G. 1959. Motivation and language behavior: A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology* 59 (1): 58–67.

Partington, A. 2004. Corpora and discourse, a most congruous beast. In A. Partington, J. Morley and L. Haarman (eds) *Corpora and Discourse*. Bern: Peter Lang, 11–20.

Partington, A. 2008. The armchair and the machine: Corpus-Assisted Discourse Studies. In C.T. Torsello, K. Ackerley and E. Castello (eds) *Corpora for University Language Teachers*. Bern: Peter Lang, 95–118.

Partington, A. 2009. Evaluating evaluation and some concluding reflections on CADS. In J. Morley and P. Bayley (eds) *Corpus-Assisted Discourse Studies on the Iraq Conflict: Wording the War*. London & New York: Routledge, 261–303.

Partington, A. 2015. Corpus-Assisted Comparative Case Studies of Representations of the Arab World. In P. Baker and T. McEnery (eds) *Corpora and Discourse Studies. Integrating Discourse and Corpora*. Basingstoke: Palgrave, 220–243.

Partington, A., Duguid, A. and Taylor, C. 2013. *Patterns and Meanings in Discourse. Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam & Philadelphia: John Benjamins.

Phillips, M. 1989. *Lexical Structure of Text*. Birmingham: University of Birmingham.

Potts, A., Bednarek, M. and Caple, H. 2015. How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse & Communication* 9 (2): 149–172.

Rayson, P. 2015. *Log-likelihood and effect size calculator* [online software]. Lancaster: University Centre for Computer Corpus Research on Language (UCREL). Available online at `http://ucrel.lancs.ac.uk/llwizard.html`

Rau, C. 2010. *Dealing with the Media. A handbook for students, activists, community groups and anyone who can't afford a spin doctor*. Sydney: University of New South Wales Press.

Reppen, R. 2010. Building a corpus. What are the key considerations?. In A. O'Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*. London & New York: Routledge, 31–37.

Rodrigues, R., Veloso, A. and Mealha, Ó. 2012. A television news graphical layout analysis method using eye tracking. *Proceedings of the 16th International Conference on Information Visualisation (IV)*: 357–362.

Sacks, H. 1972. On the analyzability of stories by children. In J.J. Gumperz and D. Hymes (eds) *Directions in Sociolinguistics: The ethnography of communication*. New York: Rinehart and Winston, 325–345.

Savický, P. and Hlavácová, J. 2002. Measures of word commonness. *Journal of Quantitative Linguistics* 9 (3): 215–231.

Schiller, D. 1986. Transformation of news in the US information market. In P. Golding, G. Murdock and P. Schlesinger (eds) *Communicating Politics: Mass communications and the political process*. Leicester: University of Leicester Press, 19–36.

Scott, M. 2009. In Search of a Bad Reference Corpus. In D. Archer (ed.) *What's in a Word-list? Investigating word frequency and keyword extraction*. Oxford: Ashgate, 79–92.

Scott, M. 2014. WordSmith Tools (Version 6) [software]. Liverpool: Lexical Analysis Software.

Scott, M. 2015. *WordSmith Tools Manual* (Version 6). Liverpool: Lexical Analysis Software. Available online at `http://lexically.net/downloads/version6/HTML/index.html?getting_started.htm`

Scott, M. and Tribble, C. 2006. *Textual Patterns. Key words and corpus analysis in language education*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Sinclair, J. 2004. *Trust the Text. Language, corpus and discourse*. London & New York: Routledge.

Sinclair, J. 2005. Corpus and text – basic principles. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 1–16. Available online at `http://ahds.ac.uk/linguistic-corpora/`

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-assisted studies of language and culture* [Reprinted in 1998]. Oxford: Blackwell.

Swales, J.M. 1990. *Genre Analysis. English in academic and research settings* (13th ed., 2008). Cambridge: Cambridge University Press.

Swales, J.M. 2002. Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (ed.) *Academic Discourse*. London: Longman, 150–167.

Swales, J.M. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Syncro Soft 2015. Oxygen XML Editor (Version 17.1, Academic License) [software]. Craiova, Romania: Syncro Soft SRL.

Tabbert, U. 2015. *Crime and Corpus. The linguistic representation of crime in the press*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Taylor, C. 2009. Negative Politeness Forms and Impoliteness Functions in Institutional Discourse: A Corpus-assisted Approach. In B.L. Davies, M. Haugh and A.J. Merrison (eds) *Situated Politeness*. London & New York: Continuum, 209–231.

Taylor, C. 2011. Negative Politeness Forms and Impoliteness Functions in Institutional Discourse: A Corpus-assisted Approach. In B. Davies, M. Haugh and A. Merrison (eds) *Situated Politeness*. London: Continuum International Publishing, 209–231.

TechSmith Corporation 2011. Camtasia Studio (Version 7.1.1) [software]. Okemos, Michigan: TechSmith Corporation.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*: 2214–2218. Available online at http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

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

Tribble, C. 1999. *Writing difficult texts*. Unpublished Doctoral Dissertation. Lancaster: Lancaster University. Available online at http://www.ctribble.co.uk/text/Writing_Difficult_Texts.pdf

Tribble, C. 2002. Corpora and Corpus Analysis: New Windows on Academic Writing. In J. Flowerdew (ed.) *Academic Discourse*. Harlow: Longman, 131–149.

Tuchman, G. 1978. *Making News. A Study in the Construction of Reality*. New York & London: The Free Press.

van Dijk, T.A. 1988a. *News as discourse* [Reprinted in 2009]. London & New York: Routledge.

van Dijk, T.A. 1988b. *News analysis. Case studies of international and national news in the press* [Reprinted in 2011]. London & New York: Routledge.

van Dijk, T.A. 2008a. *Discourse and Context. A sociocognitive approach*. Cambridge: Cambridge University Press.

van Dijk, T.A. 2008b. *Discourse and Power*. New York: Palgrave Macmillan.

van Dijk, T.A. 2009a. *Society and discourse. How Social Contexts Influence Text and Talk*. Cambridge: Cambridge University Press.

van Leeuwen, T. 2005. *Introducing Social Semiotics*. London & New York: Routledge.

van Leeuwen, T. 2007. Legitimation in discourse and communication. *Discourse & Communication* 1 (1): 91–112.

van Leeuwen, T. 2008. *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford: Oxford University Press.

White, P.R.R. 1997. Death, Disruption and the Moral Order: The Narrative Impulse in Mass-Media Hard News Reporting. In F. Christie and J.R. Martin (eds) *Genres and Institutions: Social Processes in the Workplace and School*. London: Cassell, 101–33.

White, P.R.R. 2000. Media Objectivity and the Rhetoric of News Story Structure. In E. Ventola (ed.) *Discourse and Community: Doing Functional Linguistics*. Tübingen: Gunter Narr Verlag, 379–397.

White, P.R.R. 2003. News as history: Your daily gossip. In J.R. Martin and R. Wodak (eds) *Re/reading the past: Critical and functional perspectives on time and value*. Amsterdam & Philadelphia: John Benjamins Publishing Company, 61–89.

White, P.R.R. 2009. Media power and the rhetorical potential of the "hard news" report – attitudinal mechanisms in journalistic discourse. *Proceedings of the XXIX VAKKI Symposium*, University of Vaasa, Finland, Vaasa, Finland, 13–15 February. Available at http://www.vakki.net/publications/2009/VAKKI2009_White.pdf

Wodak, R. and Krzyżanowski, M. 2008. *Qualitative Discourse Analysis in the Social Sciences*. Basingstoke: Palgrave MacMillan.

Xiao, R.Z. 2008. Well-known and influential corpora. In A. Lüdeling and M. Kytö (eds) *Corpus Linguistics. An International Handbook (Volume 1)*. Berlin & New York: Walter de Gruyter, 383–457.

Xiao, R.Z. and McEnery, T. 2005. Two Approaches to Genre Analysis. Three Genres in Modern American English. *Journal of English Linguistics* 33 (1): 62–82.

Websites

BBC launches five new HD channels 2013 (December 9). *BBC News*. Retrieved October 20, 2014, from http://www.bbc.com/news/entertainment-arts-25298109

Elliott, S. 2009 (January 22). In 'Trust Me', a Fake Agency Really Promotes. *The New York Times*. Retrieved September 8, 2014, from http://www.nytimes.com/2009/01/22/business/media/22adco.html

Lower Third [Wikipedia entry] 2005 (March 26). Retrieved February 18, 2014, from http://en.wikipedia.org/wiki/Lower_third

Moore, F. 2001 (December 27). News crawl not just for bulletins anymore. *Pittsburgh Post-Gazette*. Retrieved September 9, 2014, from http://news.google.com/newspapers?id=liQxAAAAIBAJ&sjid=MnADAAAAIBAJ&pg=6570%2C3575355

News Tickers [Wikipedia entry] 2004 (September 8). Retrieved March 7, 2013, from http://en.wikipedia.org/wiki/News_ticker

Poniewozik, J. 2010. The Tick, Tick, Tick of the Times. *The Time*. Available online at http://content.time.com/time/specials/packages/article/0,28804,2032304_2032745_2032850,00.html

Porter, R. 2007 (January 22). Changing looks [Web log post]. Retrieved September 9, 2014, from http://www.bbc.co.uk/blogs/legacy/theeditors/2007/01/changing_looks.html

Sella, M. 2001 (December 9). The year in ideas: A to Z; the Crawl. *The New York Times*. Retrieved October 3, 2014, from http://www.nytimes.com/2001/12/09/magazine/the-year-in-ideas-a-to-z-the-crawl.html

Subtirelu, N. 2014 (March 10). Some data to support the gendered nature of "bossy". *Linguistic pulse*. Retrieved March 10, 2014, from http://linguisticpulse.com/2014/03/10/some-data-to-support-the-gendered-nature-of-bossy/

Ticker Tape [Wikipedia entry] 2004 (March 19). Retrieved January 6, 2014, from http://en.wikipedia.org/wiki/Ticker_tape

The truth about news tickers [Web log post] 2011 (March 9). Retrieved March 15, 2013, from http://runningheaders.wordpress.com