ISSN 2448-1165
Campo Grande | MS
Vol. 24 | Nº 47 | 2020

Papéis
REVISTA DO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS DE LINGUAGENS - UFMS

Football and web: lexical analysis of a genre through time

Futebol na rede: análise lexical de um gênero através do tempo

Rodrigo Esteves de Lima-Lopes[1]

Resumo: Este artigo tem por objetivo estudar as escolhas lexicais em um conjunto de notícias publicadas pelo jornal britânico The Guardian. Busca-se observar se houve transformação nas escolhas lexicais em notícias de esportes nos artigos publicados durante a realização das últimas cinco Copas do Mundo. As bases teóricas estão na Linguística Sistêmico-Funcional (Halliday, 1978), na análise multimodal (KRESS; VAN LEEUWEN 2001, 2008), em especial em estudos que exploram a transformação dos modos de linguagem em gêneros noticiosos propiciada pela microinformática (HIIPPALA, 2017; KNOX 2007) e a metodológica na Linguística do Corpus (BAKER 2006; CAMERON; PANOVIĆ, 2014A). Esta pesquisa é motivada pelo fato de existirem diversos trabalhos na área de multimodalidade que exploram tal transformação, todavia, poucos são as pesquisas que observam se a mudança no ambiente tecnológico do jornal também afetou as escolhas lexicais em seus gêneros. Os dados foram coletados da API (Application Program Interface) do jornal, tendo como base as datas de realização das Copas do Mundo entre 2002 e 2018. Os corpora são representativos do gênero notícias de esporte dentro do contexto do jornal estudado, uma vez que todas as matérias no período fazem parte dos corpora. A análise e a coleta de dados foram realizadas por meio de programas escritos na linguagem R. Foram processadas listas de palavras por relevância estatística, dendrogramas de clusters e redes de colocados em cada um dos grupos de textos. Exemplos foram levantados por meios de

[1] Lecturer in the Department of Applied Linguistics at the Institute of Language Studies of the State University of Campinas (DLA/IEL/UNICAMP) – Campinas – São Paulo - Brazil. Email: rll307@unicamp.br

concordanciadores, também utilizando a linguagem R. Os resultados mostram uma grande consistência entre os corpora das cinco Copas, com espaço para alguma variação contextual que não influencia na definição do gênero. Tal resultado pode ser interpretado a partir da definição de propósito comunicativo, que parece ser determinante no que tange a regularidade das escolhas lexicais. Em outras palavras, a transposição para a internet não causou modificação nas escolhas lexicais, em contraste com questões de organização imagética e espacial.

Palavras-chave: Notícias esportivas, Jornal, Gênero, Análise do Léxico, Linguística do Corpus

Abstract*: This article studies the lexical choices in a set of news published by the UK newspaper the Guardian. It aims to analyse whether there was a change in the lexical choices in sports news published during the last five World Cups. The theoretical bases lie on the Systemic-Functional Linguistics (Halliday, 1978) and Multimodal Analysis (KRESS; VAN LEEUWEN 2001, 2008), especially in studies that explore the transformation of language modes in news genres caused by the digital revolution (HIIPPALA, 2017; KNOX 2007) and methodological on Corpus Linguistics (BAKER 2006; CAMERON; PANOVIĆ, 2014a) and its application to discourse analysis. There are several research examples in the area of multimodality that explore this change; nevertheless, few studies are analysing whether the change in the newspaper's technological environment has affected the lexical choices. The data were collected from the Guardian's API (Application Program Interface), scraping World Cup sports articles from 2002 to 2018. The corpora are representative of the sports news genre within the studied context since all the articles in the period were collected. The analysis and data collection were performed using programs written in the R language. Wordlists, dendrograms clusters, collocate networks and concordances were processed in each text group also using R. The results show consistency amongst the five Cups, with space for some contextual variation that does not influence the definition of the genre. Such a consistency could be a result of the Communicative Purpose, which seems to be a decisive influence regarding lexical choices. In other words, the transposition to the internet did not change the lexical landscape, in contrast to new uses of image and spatial organisation in the Guardian's sports news.*

Keywords: *Sport news, Newspaper, Genre, Lexical Analysis, Corpus Linguistics*

Introduction

This article aims to discuss the lexical choices in a set of English-language sports news published by The Guardian between 2002 and 2018. The primary motivation of this research is the fact that most studies related to genre change after the popularization of the internet are not centred on lexical choices, but on other equally important factors, such as the relationship between language and visuals or changes on the text displaying.

In the field of Multimodality, Kress and van Leeuwen (2001, 2008) and Baldry and Thibault (2006) discuss how the printed page has changed into the digital environment. Their main emphasis is on the different affordances and meaning-making possibilities these new media offer. Kress and van Leeuwen (2001) observe how elements related to design, production and distribution can contribute to constructing such forms of meaning. The authors are mainly concerned with discussing how such processes are socially constituted, forming blueprints of influence on meaning-making. Their research was quite influential and was the basis for a whole new research field. Bateman and colleagues (2014) take their model as a reference to perform a systematic study of the visual constitution of texts in print and digital contexts, as they use such elements as the defining elements of the newspaper as a digital genre. According to the authors, the visual organization of a genre could be defined in terms of the regularity of these visual elements (BATEMAN; DELIN; HENSCHEL, 2014). Lima-Lopes (2015), uses the theory of multimodality to discuss how typographic regularities can be mapped out. His results show typographical choices might be not only a way of defining a genre but also as a way of identifying different media supports. The author compares newspapers and magazines to show that each media has a typical typographic identity.

Baldry and Thibauld (2006) discuss how different printing technologies can contribute to the construction of meaning. The authors compare two editions of *The Capital of Karls Marx* to show how typographic choices are conditioned by technology and how such technology contributes to the constitution of the written page. The authors also discuss how the visual and informational organization of a group of webpages for children. The organization of information on the front page of newspapers is the topic of a study by Ribeiro and Souza (2018). The authors build on the work of Kress and van Leeuwen (1998) and discuss how the processes of transposing newspapers from computers to mobile devices can affect the processes of reading and understanding news. Their results question some critical elements of the theory, especially the notions of Theme vs Rheme, due to the meaning potentials made possible by the affordances of each device. A similar application of this theory is made by Carvalho and Magalhães (2009) when analysing the front page of newspapers in the state of Minas Gerais, Brazil.

The authors start from the concept of compositional meaning, as defined by Kress (1995), in order to analyse how newspapers balance the text/image relationship.

Kong (2006) studies the relationship between text and image in Hong Kong newspaper headlines. His results reveal that text and image establish a relationship very similar to the one between two propositions. That is, for Kong (2006) texts and images establish relations such as extension, projection and complementation, and behave as co-dependents in the production of meaning. Hiippala (2017) discusses multimodality issues published by Longform journalism sites. Their results converge with previous research on surveying Longform's defining aspects as a multimodal genre in terms of its content and layout (HIIPPALA, 2017). Knox (2016, 2009, 2007) reflects on image and text relations in online journals, both from an Australian and cross-cultural perspective. His results show that hard news tends to have a very text-to-image relationship similar to an online image gallery, as it seems to be organized to tell a story. Such galleries do not seem to function merely for illustrative purposes, but they also seem to serve the discursive intentions of the newspaper. The author noted the consistency in visual-verbal design of news across longer time scales and different cultures tend to be a result of technical demands, mainly due to the automation of updating systems.

In the field of communication, some research is centred on the role that newspapers occupy in the digital context. Amongst the main themes, I can highlight the emergence of alternative media, which have grown in the last few years as they have placed themselves as fact-checkers of mainstream media (HOLT; USTAD FIGENSCHOU; FRISCHLICH, 2019). A relationship that could not exist without the struggle for the establishment of boundaries for its mainstream media power. Ali and colleagues (2019) reflect on the small-market newspapers. Their results show that such publications tend to take longer to surrender to new technological processes as they are more tied to their print distribution for financial reasons. Another essential aspect of this researches is the emergence of digital volunteer networks, which appear as an emerging form of journalistic cover (CHERNOBROV, 2018). According to Chernobrov (2018), this type of coverage can be characterized as a new form of citizen journalism in

contexts in which various social actors take for themselves the functions of traditional media.

Despite their importance, such studies do not discuss whether the migration of newspapers to the internet has affected lexical choices. This paper seeks to fill this part of this gap by analysing sports news on the World Cup.

The methodological basis of this paper is Corpus Linguistics (henceforth CL) (KENNEDY, 1998), which can be defined as an area of linguistic research that is characterized by extensive computer use and the careful collection of large amounts of linguistic data. A corpus could be described as a collection of texts structured to fulfil a particular research objective (BIBER; CONRAD; REPPEN, 1998). Among its main features, it is possible to point out (ESIMAJE; HUNSTON, 2019):

- A corpus consists of naturally occurring data, whether spoken or written.
- Texts are selected to represent a specific language, register or context.
- Its size is relatively large, so it would be difficult, if not impossible, to study it without computer software.

Baker (2006) and Cameron and Panović (2014b) discuss the possible relationship between Corpus Linguistics and discourse analysis. First of all, it would be related to the support (or not) of our initial impressions, since a survey employing computational tools would lead to the observation of patterns of use, which would not be observed otherwise. As Cameron and Panović (2014b, p. 13) put it, the use of corpus allows the researcher to come up with answers and results that are non-obvious. In the context of this paper, the main objective of CL would be to enable the search for linguistic patterns that can help to clarify several questions, including those related to different linguistic theory and language usage in a specific context. In this paper, I use Corpus Linguistics as an instrument for discourse analysis in the context of Systemic-Functional Linguistics (henceforth SFL), specifically genre analysis. This is because CL can be an essential tool for eliciting discursive patterns within a specific context (CAMERON; PANOVIĆ, 2014a). This assumes a definition of discourse as

knowledge (KRESS; VAN LEEUWEN, 2001) and as a social practice in which language has a determining role (CAMERON; PANOVIĆ, 2014a).

The quality of the answers offered by CL seems not to be only a direct result of the number of texts that can be processed, but especially in the patterns it allows us to see. As Scott (2012) puts it, computers do not do the job for the analyst; it is up to her/him to interpret the data offered by the various programs. Moreover, important decisions such as the nature of the tools used, what type of processing is performed, besides the size and organization of the corpus, are subjective decisions that influence the results. Tribble (2010) states that programs only perform queries that lack an interpretative eye, which includes conclusions about their significance and importance, as it characterizes the approach as both qualitative and quantitative (BIBER; CONRAD; REPPEN, 1998).

CL is the primary methodological basis for this research because it shares a common origin with the SFL (STUBBS, 1996). Therefore, they both create the epistemological space that this research is located. As a consequence, they share several common features. First of all, both characterise linguistics as an applied social science, since CL and LSF are concerned with producing research that goes beyond mere linguistic description (CHAMBERS, 2010) by approaching the social issues inherent in textual production (STEINER, 2018).

These are approaches that have similar concerns regarding the way data are obtained and analysed. Real data analysis, both quantitatively and qualitatively, is an essential aspect for CL and SFL (BEAUGRANDE, 2002), inasmuch as it is a response to a common practice of linguistic analysis which is based solely on intuitive data. In this sense, both CL and SFL study data collected in the real world, seeking to observe patterns within the various contexts of language practice. This means that the analysis does not bring any *a priori* hypotheses, nor have its examples chosen to satisfy the analyst's theoretical assumptions. As a result, the study reflects the different contexts of language usage. Nevertheless, when it comes to CL, one can establish hypotheses whenever handling either corpus-driven or corpus-based investigations.

The assumption that form and meaning are inseparable elements (STUBBS, 1996) can be seen as a consequence of their empirical nature. In the

case of CL, the construction of meaning is clearly the result of the interaction of a word with its linguistic environment: meaning is constituted by the collocation and association of words in mappable phraseologies in each linguistic and social context (BIBER; CONRAD; REPPEN, 1998). In LSF, this such a relationship occurs through the instantiation of functions within each Registry (or situation context). Such instantiations would lead to the building of patterns that would manifest themselves in terms of choices and stand as more or less likely within each communicative situation (BEAUGRANDE, 1993; HALLIDAY; MATTHIESSEN, 2014).

Finally, both could be defined as post-Saussurean approaches (HASAN, 2014). It means they move against some of the canons of linguistics, which traditionally follows the concept of the sign as defined by Saussure in his *Cours de Linguistique générale*, (HASAN, 2014; KRESS, 1993; STUBBS, 1996). Such a change of perspective was motivated by the idea that all levels of instantiation, structure and meaning, are inseparable. The meaning is motivated both contextually and by the relationship amongst the different lexicogrammatical choices, what makes actual linguistic instantiation the only parameter for sign-making.

This study is related to the research developed by the so-called Sydney school. This school originates from the LSF studies developed by Halliday concept of Register and Hasan's notion of genre (HALLIDAY; HASAN, 1991) and it is later developed the works of Martin and Rose (2008) and Eggins and Martin (1997). According to Moran and Herrington (2005), this approach is developed to use genre as a principle of social inclusion for Australian natives, who do not have English as their mother tongue, and for immigrants with difficulty of social insertion. As a result, it had a significant influence on primary education contexts in that country. The research of Sydney school relies mainly on the search for lexicogrammatical and structural regularities that make the genre a stable textual production. Genres tend to consist of a certain number of stages, each with a specific objective within the process of textual instantiation. These stages have a relatively stable order and can be classified according to their function within the text (LIMA-LOPES, 2006). Thus, there is a two-way relationship between language and the situation, as a genre is a system of choices that mediates

human experience. In this system, lexical and potential organisational structures are core elements.

The corpora of this research are a set of sports reports collected from the British newspaper *The Guardian*. Data were collected using software (see methodology section) to answer the following question: *Have the lexical choices used to narrate the history of the World Cup changed as the internet gained importance in society?*

Processing and collecting data

Data from this survey were collected through data scraping using R, a statistical software that can be used for data collection and processing. The *GuardianR* package (Bastos and Puschmann 2018) was the primary tool for data collection. It allows data scraping using Guardians Newspaper's API (Application programming interface). An API is a communication protocol between a server and a client which makes it possible to build applications for many ways of data exchange.

Data were collected from the following criteria:

- Only sports articles published in the months of the World Cup were considered;
- Genres that did not fit the classification of **sports articles** were disregarded;
  - Amongst these are quizzes, letters from readers and chronicles;
- No photo or video galleries were considered;
  - The focus of this work was the written material.

Since Guardian's API allows the total scraping of the articles in the given period for research, all its articles referring to the World Cups studied were collected. Such a fact makes the corpus 100% representative of the articles written during the World Cups period from the Gaurdian. The data collection featured articles written during six Football World Cups, from 2002 through 2018. It was a limitation of the API, which allows researchers to scrape data from 1999.

In the case of this research, such a limitation made it impossible for us to collect data on prior World Cups.

The data was saved as a *data frame* and processed through a script written in R computer language. Such processing aimed at cleaning the data from possible HTML/XML format, deleting information such as headings and inopportune tagging. R was also a tool for text processing clusters, data cleansing and concordancing. All data were processed using two main packages for text analysis:

- **Quanteda**:[2] Responsible for concordancing, calculating a list of n-grams and obtaining general statistics of the corpus, as well as word network representation (BENOIT *et al.*, 2018);
- **TM**:[3] Responsible for processing the cleaning of the text and calculating the hierarchical cluster representation (FEINERER; HORNIK, 2019).
- **Tidytext**:[4] Responsible for organizing the data and calculating the significance of words (SILGE; ROBINSON, 2017).

The significance of the words in the corpus was calculated using a Term Frequency (TF)/Inverse document frequency (IDF). The TF/IDF approach measures the rate of a term in a document and then it weights down the importance of more frequent words and scales up the rarer ones (RAJARAMAN; ULLMAN, 2011; WU *et al.*, 2008). The idea is to decrease the importance of commonly occurring words, such as articles and prepositions, and increase content words, such as nouns and verbs. This is very helpful in terms of determining the aboutness of the texts since grammatical words tend to score highly. In a simplified way, such measurements are calculated using the formula in figure 1.

2 https://quanteda.io
3 https://cran.r-project.org/package=tm
4 https://cran.r-project.org/web/packages/tidytext/index.html

Figure 1: TF and TDF calculation

$$TF = \frac{n_{documents\ T\ is\ in}}{n_{documents\ terms}}$$

$$IDF = log_e\left(\frac{n_{Total\ document\ number}}{n_{Documents\ t\ in\ it}}\right)$$

The cluster calculation was performed using the Ward (1963) *Hierarchical Grouping Method.* Ward (1963) developed a method for creating hierarchical and mutually exclusive groups, in which members tend to be maximally similar concerning a given characteristic. The calculation took into account the co-occurrence of words. It might help to understand how the main topics in the comments were defined in terms of their strongest collocates.

Table 1 brings the corpus in numbers. As one can see, corpora was just under 4 million words (tokens) and just over 1 million 8000 different words (types) in size. The number of words has grown regularly over the years, although the number of documents is relatively stable. This appears to show that despite what common sense believes, the use of the internet as the primary newspaper platform does not seem to decrease the size of articles. On the contrary, they seem to be getting larger.

Table 1: The corpus in numbers

World Cup	Types	Tokens	Documents
2002	269.262	553.831	**779**
2006	288.276	563.196	**928**
2010	387.387	890.042	**987**
2014	489.938	1.212.400	**1.194**
2018	426.965	1.092.614	**832**
Total	**1.861.828**	**3.814.083**	**4.720**

Final Algorithm for data analysis

This study followed the protocol I present below:

1. Scrape the sports news, following criteria below:

- The data span of the article collection would be the days each World Cup was taking place
- The articles should be in the sports section

2. Clean data to process the articles:

- Clean special characters
- Clean numbers
- Clean punctuation
- Clean HTML and XML tags

3. Writing the script for data processing

- Creating a list of stop-words to delete sensitive terms that could identify video participants
- Calculating the TF/TDF frequency tables
- Creating a wordlist
- Generate network and cluster representation
- Create TDF and TDM matrixes for representation

4. Analysing each topic via concordances and collocates

- The words of each group of texts were studied via concordances and a list of collocates. The criteria for selecting such words was intuitive, based on my personal interpretation of the wordlist and of the clusters that resulted from data processing.

5. All the analysis was comparative

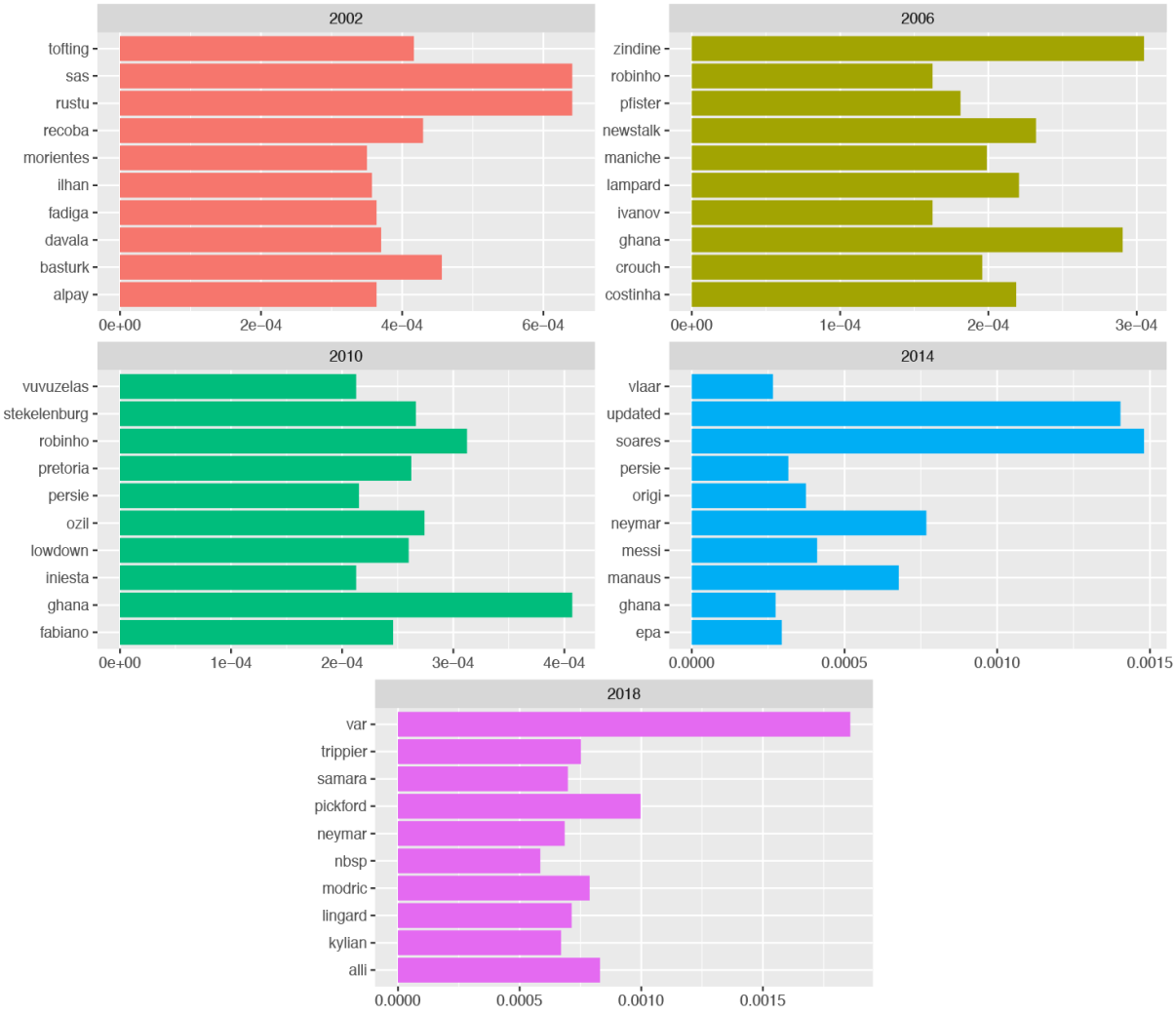
Data analysis

Figure 2 shows the results of wordlists comparing articles published during the five World Cups. The words in this list were processed to decrease the importance of grammatical words.

The most frequent words are the names of players or words that would relate to the context of each World Cup. The players in the list are usually those

who have some prominence in the competition, such as Turkish player Hasan Gökhan Şaş in 2002 (ex 01), French player Zinédine Zidane in 2006 (ex 02), Robinho in 2010 (ex 03), Suarez in 2014 (ex 04) and Neymar Jr. in 2018 (ex 05).

Figure 2: Most frequent words in the corpus



Source: Data

(**ex01**) (…) after half time to cancel out **hasan sas's** goal just before the interval seconds (…)[2002]

(**ex02**) (…) benefit of the doubt a win for **zidane** and will bring the french out (…)[2006]

(**ex03**) (…) despite wright s possible arrival brazilian funboy **robinho** still doesn t want to return to (…)[2010]

(**ex04**) (…) suarez must remain a footballing pariah luis suarez's appeal against his four month ban (…)[2014]

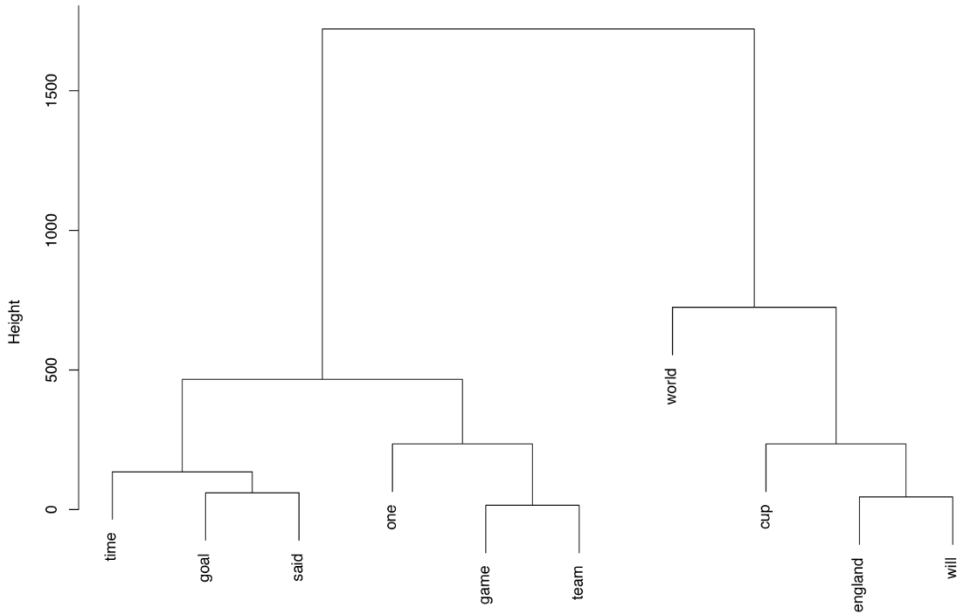
(**ex05**) (…) he gallops down the left and **neymar** and gabriel jesus up in support (…)[2018]

The reasons that lead to this emphasis are related to the both to a performance of the athletes and as a criticism made by the newspaper. In some cases, the team's names also occur in the most frequently used word list. Among these examples is Ghana (in 2006, 2010 and 2014).

(**ex06**) (…) who may well be blowing **vuvuzelas** and soaking up the occasion rather than (…)

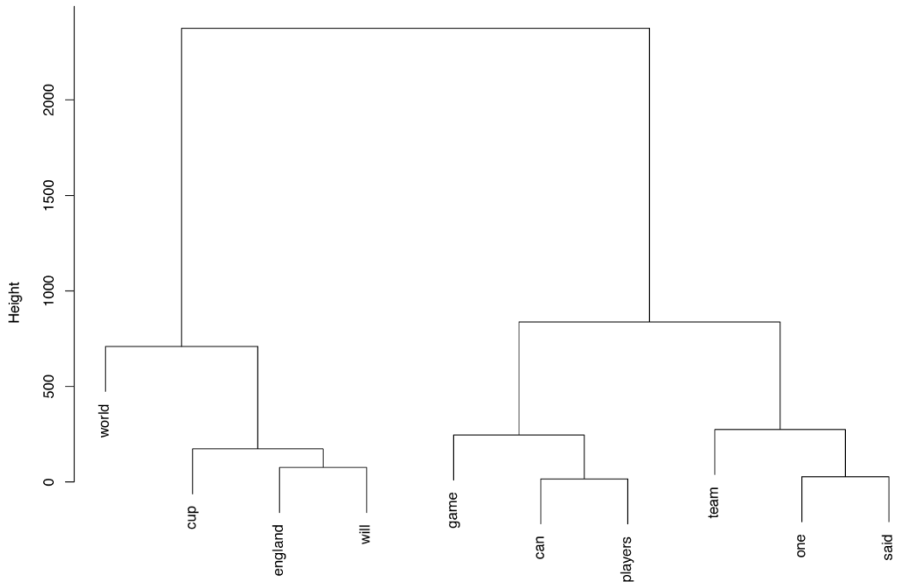
(**ex07**) the judicial lottery of **var** has given luddism a good name (…)

Figure 3: Clusters Dendrogram 2002 World Cup



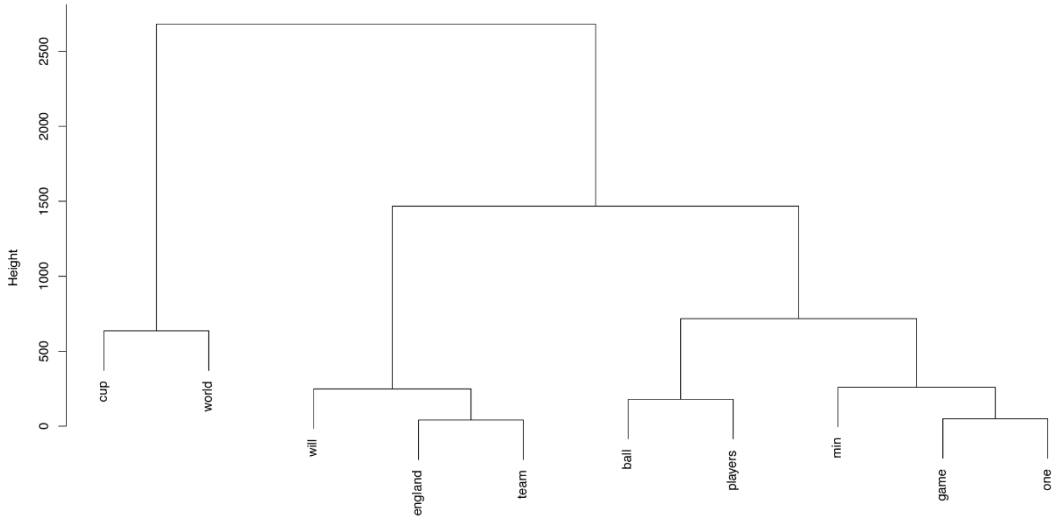
Source: Data

Figure 4: Clusters Dendrogram 2006 World Cup



Source: Data

Figure 5: Clusters Dendrogram 2010 World Cup

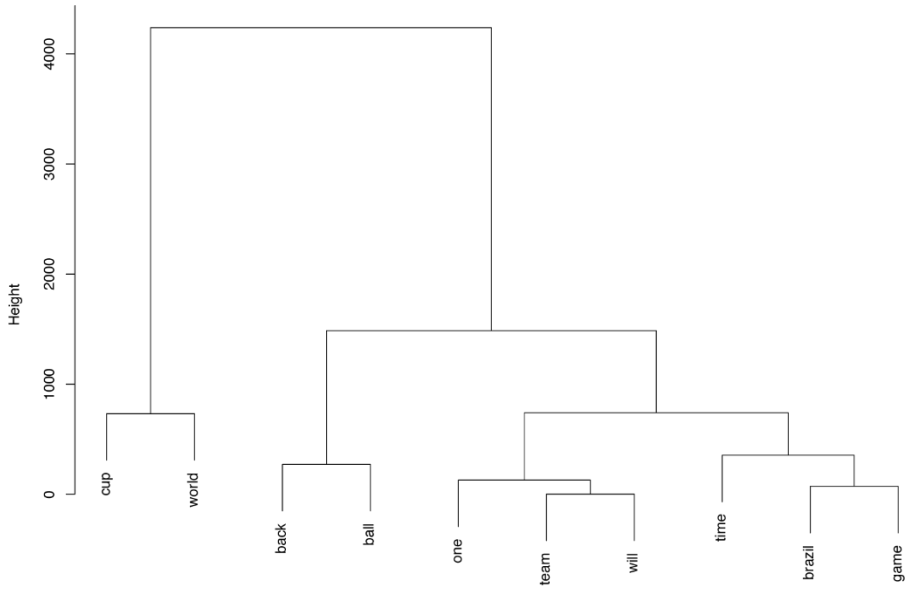


Source: Data

Contextual references are related to specific facts that become news and reverberate in the press. They are not present in all tournaments and seem to represent themes that, for some reason, were relevant to the competition. Some examples would be *Vuvuzelas* in 2010 (ex 06) and the *Video assistant referee* (VAR) in 2018. Although it is not directly related to the competition, the name *Newstalk* frequently occurred in 2006. It represents an independent Irish radio

station that has its website cited in the articles as a way to promote the availability of live audio streaming.

Figure 6: Clusters Dendrogram 2014 World Cup



Source: Data

If we look at each of the competitions, we may imply that their lexical profile is very similar. Except for the terms VAR and Vuvuzela (ex 06 and 07), players occupy almost all the words in the lists every year, which might mean that lexical choices do not vary over the years.

To examine if the results would continue to be similar following another statistical parameter, I performed a cluster calculation using the Ward (1963) model. As I said earlier, clusters were calculated for each newsgroup as in figures 3 through 7.

One more time, there seems to be some similarity amongst the articles. Lexical items appear to be related to the tournament name *World Cup* or typical elements of the game such as *time*, *goal*, *ball*, *players* and *game*.

(**ex08**) (…) Argentina far more cautious and often cynical play they knocked out Brazil but even then seemed to be playing for extra time and penalties [2002]

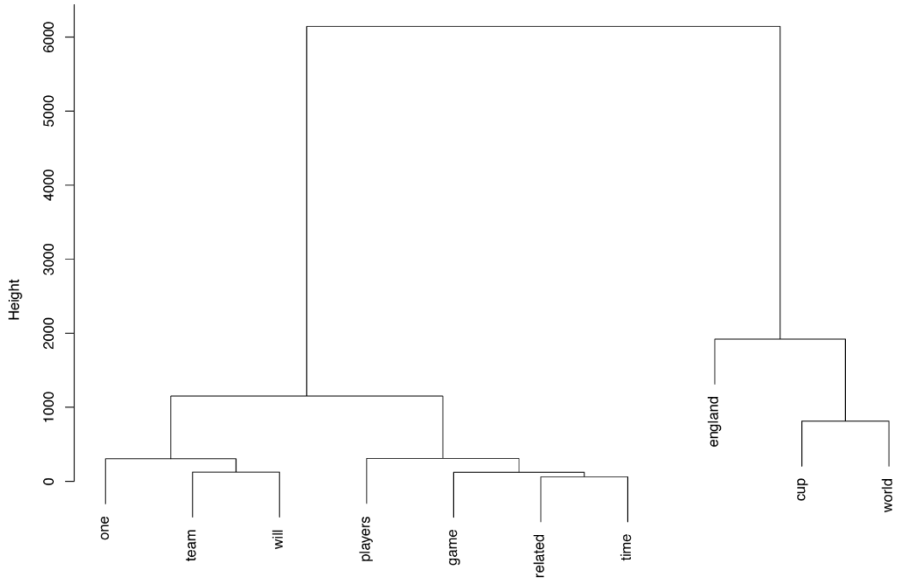
(**ex09**) (…) Yesterday morning disappointed and disillusioned by Mexico extra-time defeat by Argentina [2006] (…)

(**ex10**) (…) Drama in the first period of extra time as tumbles in the box under Heitinga [2010] (…)

(**ex11**) (…) just four of the eight matches had extra time [2014] (…)

(**ex12**) (…) Alain Giresse who slots in France equaliser to make it in extra time play [2018] (…)

Figure 7: Clusters Dendrogram 2018 World Cup

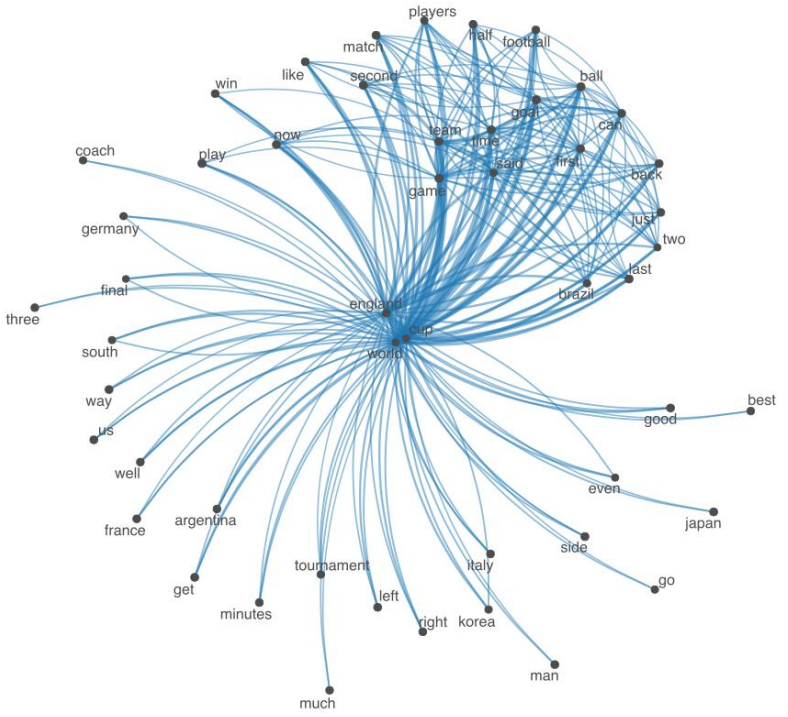


Source: Data

Some regularity of this usage can be seen in examples 08 to 12. In these examples we have the use of the word *time* which appears in all articles following the same usage pattern (*extra time*). In football terms, it would represent the need for a time extension due to lack of goals. The word *team* brings similar results, as it represents the set of players from a particular country (examples 13 and 14).

England is a direct reference to the country's team in which the newspaper is published, and it is centred on the evaluation of the team's results. **England** is strongly related to the modal **will** (ex 17) in 2002, 2006 and 2010. The performance of this country is also evaluated, but with one important difference exists, since it is appraised is in terms of the future actions it will have to take in the tournament.

Figure 8: Network of terms 2002 World Cup



Source: Data

(**ex13**) (…) no one will be surprised to learn that Holland are the **team** most guilty of throwing their toys from the [2006] (…)

(**ex14**) If there was period when Gareth Southgate team began to lose [2018] (…)

(**ex15**) (…) uruguayan official who disallowed that **england** goal so while [2010] (…)

(**ex16**) (…) stars of brazil like pele conceding that **england** might very well win since though [2002] (…)

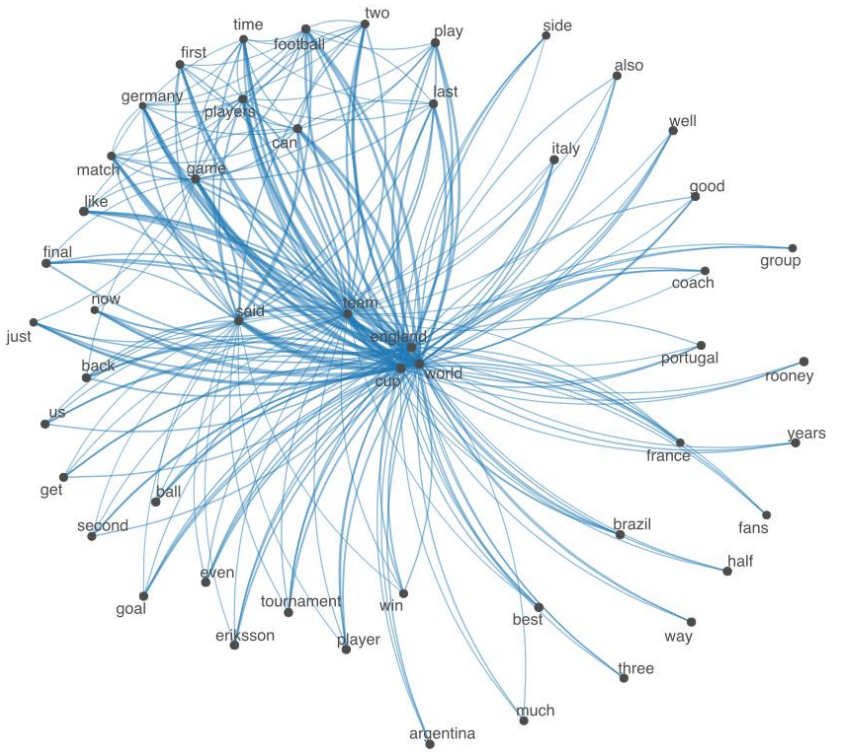
The only cup in which *England* does not appear on a Dendrogram is in 2014, which was in Brazil. Although some *England* team assessment is present in the data, it seems to give way to the country that hosted the competition. To my mind, two reasons might explain such a change. Firstly, it is important to note that the 2014 World Cup was highlighted by the political activism in Brazil, since several protests occurred during its realization (ex 19). In the second, Brazil had its most significant defeat in World Cups, losing the semi-final to Germany by 7 x 1 (ex. 19).

(**ex17**) (…) i'm not sure **england will** have to simply wait on [2006] (…)

(**ex18**) (…) the upcoming presidential election **brazil** supporter [2014] (…)

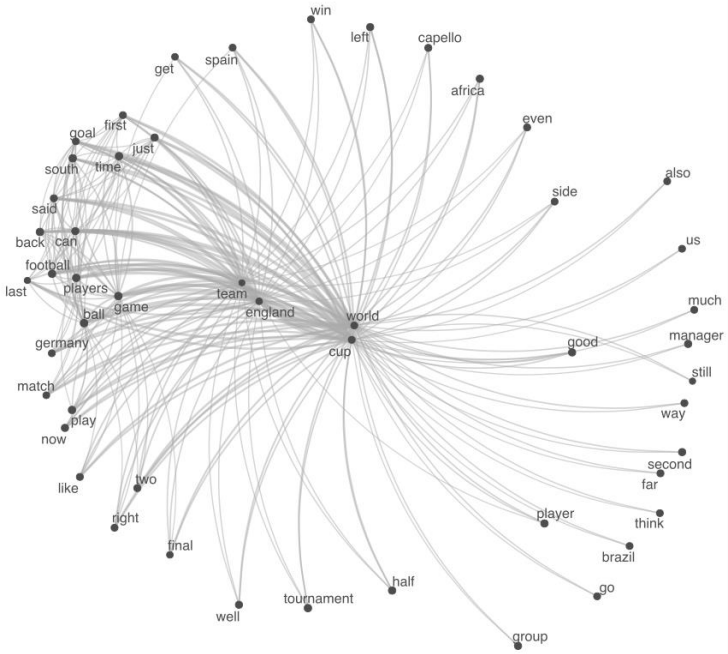
(**ex19**) (…) be covering tonight's semi-final between **brazil** and germany on a minute by minutely [2014] (…)

Figure 9: Network of terms 2006 World Cup



Source: Data

Figure 10: Network of terms 2010 World Cup



Source: Data

In all dendrograms, there seems to be room for some terms that are present for some contextual matter. Some examples would be *goal* in 2002, *players* in 2006 and 2018, *ball* in 2010 and *back* in 2014 (ex 20 to 24). The search for patterns did not return relevant results in these cases. It is possible, then, that the calculation used brought a degree of randomness in the choice of words, something that should be discussed in future research.

(**ex20**) (…) three draws but if a Roger Milla **goal** had not been ruled out for offside [2002] (…)

(**ex21**) (…) former communications director told me once the **players** arrive in the brazil [2006] (…)

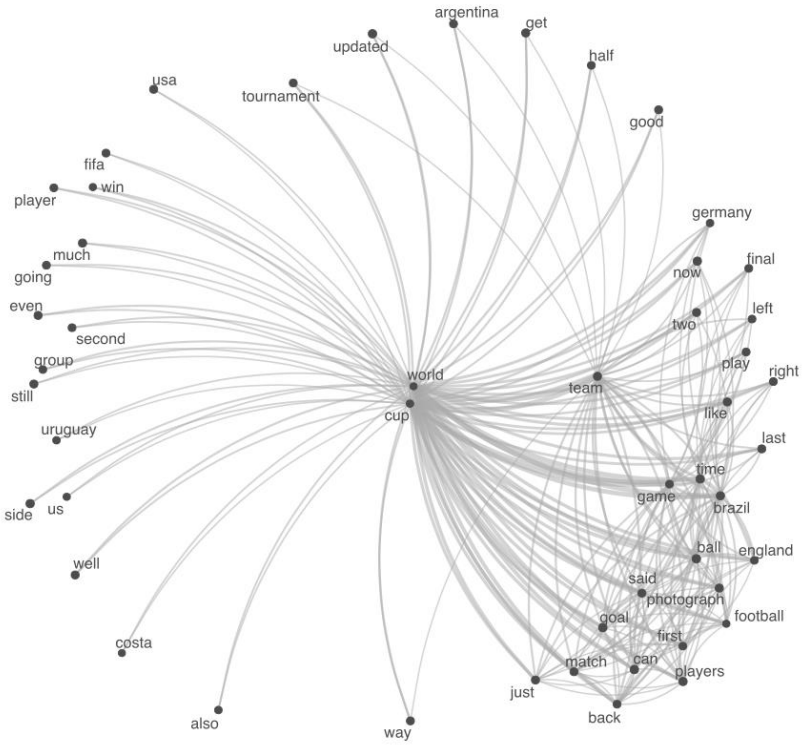
(**ex22**) (…) team that really wants the **ball** you deserve to suffer the fact [2010] (…)

(**ex23**) (…) behind centre back and full **back** and flick the ball into the bottom [2014] (…)

(**ex24**) (…) matches featuring Argentina they had two **players** sent off in the final [2018] (…)

Finally, to observe the similarities would persist in a refined collocation analysis, a co-collocation network was processed in each research corpora. As already mentioned earlier, this calculation takes into account a co-occurrence matrix of words to observe if there are strong ties amongst them. The network representation is in figures 8 through 12.

Figure 11: Network of terms 2014 World Cup



Source: Data

Figures 8 through 12 show, one more time, a consistency amongst the articles of the five world cups studied. *England* along with *World Cup* are central words in all networks. The five are made up of two well-defined areas. In the first, a series of grammatical words, names, and team names seem to be linked with *England* and *World Cup* without, however, connecting to each other. One of the implications of this type of structure is that, despite the existence some news topics, such lexical items always seem to be related either to the tournament or to the country of origin of the newspaper (ex 25 and 26).

(**ex25**) (…) the trophy two weeks on sunday **brazil italy** and england as previous winners will all [2002] (…)

(**ex26**) (…) revelling in possession and england can frustrate **portugal** that is why it seems that sven [2006] (…)

Some words connect to **England** and **World Cup**, besides forming a small network themselves. Such words represent the game and its story, as is the case with **half**, **game** and **couch**. A relationship is established within different teams, especially in the first three cups. All these teams also relate to **England**, which is central in the network.

(**ex27**) (…) each **coach** can dispute one call per **half** using instant slow motion replay it is [2002] (…)

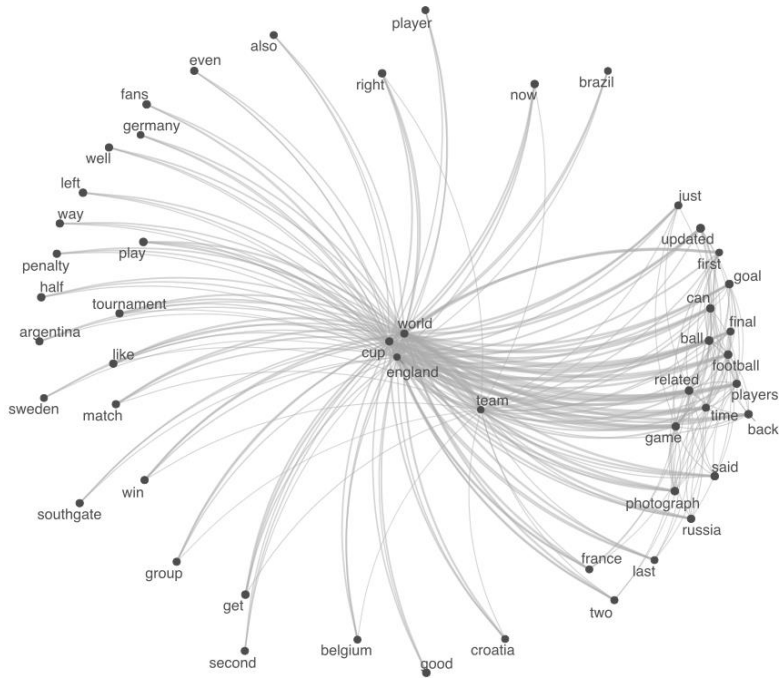
(**ex28**) (…) the venue for england third group **match** against sweden on tuesday night [2006] (…)

(**ex29**) (…) **spain** deserved to win against **germany** they were the better side [2010] (…)

(**ex30**) (…) saturday game against sweden will meet **russia** or croatia for place in the final [2014] (…)

(**ex31**) (…) as germany crushed **brazil** germany kept going [2018] (…)

Figure 12: Network of terms 2018 World Cup



Source: Data

The last two cups have a greater emphasis on the host country: **Brazil** and **Russia**. But again, **England** is part of the highlighted words. In addition to predictable use as a circumstance of location, these national teams are regularly cited and evaluated for their results and performance (ex. 30 and 31).

Final Remarks

The theoretical foundations were the Systemic-Functional Linguistics and Corpus Linguistics, mainly due to their common origin. Both see language as a relationship of choice, deeply related to the context.

This article aimed to study five corpora of sports news published by the British newspaper *The Guardian* during the last five World Cups. I sought to understand if the lexical choices made in the news would have any variation over the years since there are few studies out of a multimodal approach that make such a comparison. Traditionally in LSF, studies on the changes of genres are centred on a multimodal perspective. They seek, in most cases, to reflect on the fundamental transformation in the way the text is represented on the printed page, reflecting how its interaction with images and new forms of representation. Besides, the focus of such studies, as previously discussed, is also on the new forms of meaning-making brought about by the so-called 'digital revolution'.

This study took another path. The main concern was to analyse whether a change in media support has influenced lexical choices over time. The choice for journalistic texts was based on the essential analytical fortune for these texts within the context of LSF and multimodality, as discussed in section 0.

The results seem to show that there is consistency amongst the nature of the lexical choices made through time. Figure 2 showed the consistency of elements that contextually represent the players and some specific characteristics of some World Cups, as the video referee in 2018. Figures between 3 and 7 showed the importance that England and its team has for the newspaper. It shows that an essential part of Guardian's speech aims at analysing English football, the country in which the newspaper is. Figures 8 to

12, on the other hand, seem to show another type of lexical consistency, in which elements of the competition, as well as structural elements of football, appear.

Such consistency can be explained as a result of the football as a sport and the communicative purpose of the texts. In the last five World Cups, football has had very few changes in terms of its structure. Perhaps the most impactful was the video assistant referee, which was observed in the results representing the 2018 World Cup (see figures 2 and 12). Therefore, it would be possible that this stability of the sport reflects in the lexical choices related to it. Second, the communicative purpose of the texts also seems to reflect such stability; the little transformation in sport leads to make the choices related to the genre more stable.

These results show that the process of digitalising the journalistic text does not seem to influence the lexical level, at least in the context of the Guardian and its news about the World Cup. This is in no way a rejection of the importance that multimodal analysis has had in the context of genre studies, but rather an acknowledgement that some of the new modes of language may not have such an influence on other strata of analysis.

This has an important impact in the context of LSF since one of the assumptions of the theory is that the transformations in a stratum of meaning should have an impact on the others. In the Guardian World Cup articles, it does not seem to be the case due to the stability of the textual purpose and the football itself. Although the context interferes in such choices, which seem to be motivated by tournament-specific events, the overall results show that the digitization of Guardian have not had impact on the way this news is written. Such results are essential for a better understanding of the influence of digital platforms on journalistic writing processes. The most frequent lexicon is related to specific elements of the sport, such as the name of the players, integral parts of football, participating countries, in addition to some evaluation of the performances of players and teams.

In summary, the contributions of this research are:

- The digitalisation of Guardian Newspaper seems not to have an impact on lexical choices:

- Lexical choices seem to be consistent through the years
- Some contextual variation might be observed due to the influence of the Context of Situation
- The few changes of football as a sport might have contributed to the lexical consistency
- The communicative purpose of such genre has not changed during the last five World Cups, contributing to the lexical consistency.
- The change of the multimodal landscape seems to have a weak impact in terms of lexical choices in these texts.

Analysis and scraping of data tools were used to process the almost four million words that make up the corpora. It is clear to me that the procedures shown hare may have some influence on the results. Given that newspapers are multiple textual colonies (HOEY, 1986), it is expected that different sections would have different textual and linguistic configurations. Moreover, it should be emphasised that printing is a discourse domain deeply related to its culture of origin. As an outcome, the results discussed here may not be valid for other cultural contexts. However, they could be a basis for comparison between other media contexts, since my methodology can be replicated or applied to studies in different settings. Suggestions for future studies would include replicating the methodology developed here in broader research contexts, as well as in other genres and media. Thus, some future studies should compare these results with other media and languages and compare different newspapers sections in order to verify whether other contexts of situation might show different results.

References

ALI, C. *et al.* The Digital Life of Small Market Newspapers: Results from a multi-method study. **Digital Journalism**, v. 7, n. 7, p. 886–909, 9 ago. 2019. DOI 10.1080/21670811.2018.1513810.

BAKER, P. **Using corpora in discourse analysis**. London; New York: Continuum, 2006.

BALDRY, A.; THIBAULT, P. J. **Multimodal transcription and text analysis**. London; Oakville, CT: Equinox Pub, 2006.

BATEMAN, J.; DELIN, J.; HENSCHEL, R. Mapping the multimodal genres of traditional and electronic newspapers. **New directions in the analysis of multimodal discourse**. Oxon: Routledge, 2014. p. 147–172.

BEAUGRANDE, R. Descriptive linguistics at the millennium: Corpus data as authentic language. **Journal of Language and Linguistics**, v. 1, n. 2, p. 91–131, 2002.

BEAUGRANDE, R. Register in discourse studies: A concept in search of a theory. *In*: GHADESSY, M. (org.) **Register Analysis. Theory and Practice. Londres, Nueva York: Pinter Publishers**. New York: Pinter Publishers, 1993. p. 7–25.

BENOIT, K. *et al.* quanteda: An R package for the quantitative analysis of textual data. **Journal of Open Source Software**, v. 3, n. 30, p. 774, 6 out. 2018. DOI 10.21105/joss.00774.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus linguistics: investigating language structure and use**. Cambridge; New York: Cambridge University Press, 1998.

CAMERON, D.; PANOVIĆ, I. Discourse and Discourse Analysis. *In*: CAMERON, D.; PANOVIĆ, I. (org.) **Working with Written Discourse**. London: SAGE Publications, Ltd, 2014a. p. 3–14.

CAMERON, D.; PANOVIĆ, I. Working with Written Discourse in Social Research. *In*: CAMERON, D.; PANOVIĆ, I. (org.) **Working with Written Discourse**. London: SAGE Publications, Ltd, 2014b. p. 153–168.

CARVALHO, F. F.; MAGALHÃES, C. Mídia Impressa e Multimodalidade: Os Significados Composicionais na Primeira Página de Jornais Mineiros. **Revista da Anpoll**, v. 2, n. 27, 2009. DOI 10.18309/anp.v2i27.143. Disponível em: https://anpoll.emnuvens.com.br/revista/article/view/143. Acesso em: 3 nov. 2019.

CHAMBERS, A. What is data-driven learning. *In*: O'KEEFFE, A.; MCCARTHY, M. (orgs.) **The Routledge handbook of corpus linguistics**. London: Routledge, 2010. p. 345–358.

CHERNOBROV, D. Digital Volunteer Networks and Humanitarian Crisis Reporting. **Digital Journalism**, v. 6, n. 7, p. 928–944, 9 ago. 2018. DOI 10.1080/21670811.2018.1462666.

EGGINS, S.; MARTIN, J. R. Genres and Registers of Discourse. *In*: VAN DIJK, T. A. (org.) **Discourse as Structure and Process: Discourse Studies: A Multidisciplinary Introduction**. London: SAGE Publications Ltd, 1997. p. 230–256. DOI 10.4135/9781446221884.n9.

ESIMAJE, A. U.; HUNSTON, S. What is corpus linguistics? *In*: ESIMAJE, A. U.; GUT, U.; ANTIA, B. E. (orgs.) **Studies in Corpus Linguistics**. Amsterdam: John Benjamins Publishing Company, 2019. v. 88. p. 8–35. DOI 10.1075/scl.88.02esi. Disponível em: https://benjamins.com/catalog/scl.88.02esi. Acesso em: 11 fev. 2019.

FEINERER, I.; HORNIK, K. **tm - Text Mining Package**. Vienna: [s.n.], 2019. Disponível em: http://tm.r-forge.r-project.org/. Acesso em: 20 mar. 2020.

HALLIDAY, M. A. K.; HASAN, R. **Language, context and text: aspects of language in a social-semiotic perspective**. Oxford: Oxford University Press, 1991.

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. **Halliday's introduction to functional grammar**. Fourth Edition ed. Milton Park, Abingdon, Oxon: Routledge, 2014.

HASAN, R. Towards a paradigmatic description of context: systems, metafunctions, and semantics. **Functional Linguistics**, v. 1, n. 9, p. 1–54, 2014. DOI 10.1186/s40554-014-0009-y.

HIIPPALA, T. The Multimodality of Digital Longform Journalism. **Digital Journalism**, v. 5, n. 4, p. 420–442, 21 abr. 2017. DOI 10.1080/21670811.2016.1169197.

HOEY, M. The Discourse Colony: A Preliminary Study of a Neglected Discourse Type. *In*: COULTHARD, M. (org). **Talking about Text**. Birmingham: University of Birmingham, 1986. p. 1–25.

HOLT, K.; USTAD FIGENSCHOU, T.; FRISCHLICH, L. Key Dimensions of Alternative News Media. **Digital Journalism**, v. 7, n. 7, p. 860–869, 9 ago. 2019. DOI 10.1080/21670811.2019.1625715.

KENNEDY, G. **An introduction to corpus linguistics**. London: Longman, 1998.

KNOX, J. S. Online newspapers: Structure and layout. *In*: JEWITT, C. (org.) **The Routledge Handbook of Multimodal Analysis**. 2 edition ed. London New York: Routledge, 2016. p. 60–75.

KNOX, J. S. Punctuating the home page: image as language in an online newspaper. **Discourse & Communication**, v. 3, n. 2, p. 145–172, maio 2009. DOI 10.1177/1750481309102450.

KNOX, J. S. Visual-verbal communication on online newspaper home pages. **Visual Communication**, v. 6, n. 1, p. 19–53, 2007. DOI 10.1177/1470357207071464.

KONG, K. A taxonomy of the discourse relations between words and visuals. **Information Design Journal**, v. 14, n. 3, p. 207–230, 2006.

KRESS, G. Against Arbitrariness: The Social Production of the Sign as a Foundational Issue in Critical Discourse Analysis. **Discourse & Society**, v. 4, n. 2, p. 169–191, 1993. DOI 10.1177/0957926593004002003.

KRESS, G. Representational resources and the production of subjectivity: Questions for the theoretical development of Critical Discourse Analysis in a multicultural society. *In*: CALDAS-COULTHARD, C. R.; COULTHARD, M. (orgs.) **Texts and practices**. New York: Routledge, 1995. p. 24–40.

KRESS, G.; VAN LEEUWEN, T. Front pages:(The critical) analysis of newspaper layout. *In*: BELL, A.; GARRET, P. (orgs.) **Approaches to media discourse**. Hoboken: Blackwell Publishing, 1998. v. 186.

KRESS, G.; VAN LEEUWEN, T. **Multimodal discourse: the modes and media of contemporary communication**. London: New York: Arnold; Oxford University Press, 2001.

KRESS, G.; VAN LEEUWEN, T. **Reading images: the grammar of visual design**. 2. ed ed. London: Routledge, 2008.

LIMA-LOPES, R. E. de. A estrutura genérica em cartas de venda. **Trabalhos em Linguística Aplicada**, v. 45, n. 2, p. 293–309, 2006. DOI 10.1590/S0103-18132006000200009.

LIMA-LOPES, R. E. de. Explorando o Significado Tipográfico em Gêneros Escritos: Potencialidades e Regularidades. *In*: LIMA-LOPES, R. E. de; FISCHER, C.; APARECIDA GAZOTTI VALLIM, M. (Org.) **Perspectivas em Línguas para Fins Específicos: Festschrift para Rosinda Ramos**. Campinas: Pontes, 2015. p. 103–140.

MARTIN, J. R.; ROSE, D. **Genre relations: mapping culture**. London/Oakville, CT: Equinox Pub, 2008.

MORAN, C.; HERRINGTON, A. (Org.) **Genre Across The Curriculum**. Logan, Utah: Utah State University Press, 2005.

RAJARAMAN, A.; ULLMAN, J. D. Data Mining. **Mining of Massive Datasets**. Cambridge: Cambridge University Press, 2011.

RIBEIRO, A. E.; DE SOUZA, L. M. Capas de jornal e multimodalidade em dispositivos móveis: questões de layout e leitura. **Polifonia**, v. 24, n. 35/2, p. 89–104, 2018.

SCOTT, M. What can corpus software do? *In*: O'KEEFFE, A.; MCCARTHY, M. (org). **The Routledge handbook of corpus linguistics**. Routledge handbooks in applied linguistics. New York: Routledge, 2012. p. 136–152.

SILGE, J.; ROBINSON, D. **Text mining with R: a tidy approach**. Beijing/Boston: O'Reilly, 2017.

STEINER, E. A tribute to M.A.K. Halliday. **Lingua**, v. 216, p. 1–9, 2018. DOI 10.1016/j.lingua.2018.10.009.

STUBBS, M. British Traditions in Text Analysis: Firth, Halliday and Sinclair. **Text and corpus analysis**. London: Blackwell, 1996. p. 23–50.

TRIBBLE, C. What are concordances and how are they used? *In*: O'KEEFFE, A.; MCCARTHY, M. (org). **The Routledge Handbook of Corpus Linguistics**. [s.l.]: Routledge, 2010. DOI 10.4324/9780203856949.ch13.

WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association**, v. 58, n. 301, p. 236–244, mar. 1963. DOI 10.1080/01621459.1963.10500845.

WU, H. C. *et al.* Interpreting TF-IDF term weights as making relevance decisions. **ACM Transactions on Information Systems**, v. 26, n. 3, p. 1–37, 1 jun. 2008. DOI 10.1145/1361684.1361686.

Recebido em: 10-11-2019

Aprovado em: 25-03-2020