ARTICLES
ARTICULES

AUTOMATED NARRATIVES AND JOURNALISTIC TEXT GENERATION:

The lead organization structure Translated into code

MÁRCIO CARNEIRO DOS SANTOS
Universidade Federal do Maranhão, Brazil

ABSTRACT - It describes the experiment of building a software capable of generating leads and newspaper titles in an automated fashion from information obtained from the Internet. The theoretical possibility Lage already provided by the end of last century is based on relatively rigid and simple structure of this type of story construction, which facilitates the representation or translation of its syntax in terms of instructions that the computer can execute. The paper also discusses the relationship between society, technique and technology, making a brief history of the introduction of digital solutions in newsrooms and their impacts. The development was done with the Python programming language and NLTK- Natural Language Toolkit library - and used the results of the Brazilian Soccer Championship 2013 published on an internet portal as a data source.
Keywords: Automated Narratives. Online journalism. Python. Artificial intelligence. NLTK.

NARRATIVAS AUTOMATIZADAS E A GERAÇÃO DE TEXTOS JORNALÍSTICOS: a estrutura de organização do lead traduzida em código

RESUMO - Descreve-se o experimento de construção de um software capaz de gerar leads e títulos jornalísticos de forma automatizada a partir de informações obtidas na internet. A possibilidade teórica já prevista por Lage no final do século passado baseia-se na estrutura simples e relativamente rígida desse tipo de construção narrativa, o que facilita a representação ou tradução da sua sintaxe em termos de instruções que os computadores podem executar. Discutem-se também as relações entre sociedade, técnica e tecnologia, fazendo um breve histórico sobre a introdução das soluções digitais nas redações jornalísticas e seus impactos. O desenvolvimento foi feito com a linguagem de programação Python e a biblioteca NLTK- *Natural Language Toolkit* – e usou os resultados do Campeonato Brasileiro de Futebol de 2013 publicados em portal da internet como fonte de dados.
Palavras-chave: Narrativas Automatizadas. Jornalismo online. Python. Inteligência artificial. NLTK.

NARRATIVAS AUTOMATIZADAS Y LA GENERACIÓN DEL TEXTO PERIODÍSTICO: La estructura de la organización del lead traducida a código

RESUMEN - Se describe la experiencia de la construcción de un software capaz de generar leads potenciales y títulos de periódicos de forma automatizada de la información obtenida a través de Internet. La posibilidad teórica Lage ya previsto para el final del siglo pasado se basa en la estructura relativamente rígida y simple de este tipo de construcción de la historia, lo que facilita la representación o la traducción de su sintaxis en función de las instrucciones que la computadora puede ejecutar. El documento también analiza la relación entre la sociedad, la técnica y la tecnología, haciendo una breve historia de la introducción de soluciones digitales en las salas de redacción y sus impactos. El desarrollo se hizo con el lenguaje de programación Python y biblioteca NLTK y utiliza los resultados del Campeonato de Fútbol de Brasil 2013 publicados en el portal de Internet como fuente de datos.
Palabras clave: Narrativas automatizado. El periodismo en línea. Python. Inteligencia artificial. NLTK.

1. Introduction – the fear and fascination of machines

Even though it is quite simplistic, the dual relationship between man and technology still exists today. Whether Promethean or Faustian (RÜDIGER, 2007), apocalyptic or integrated (ECO, 2006), cyber-illuminist or neo-Luddite[1], fields such as the Philosophy of Technology have invested a lot of effort in discussing this issue, starting with the concept of technique.

If the origins of technique lie in the past, then the concept of technology came much later. Lemos (2002) teaches us that technology is a modern technique, far from what could have been imagined in antiquity and free of its ties with divinity. Based on reason and scientific development, Newtonian physics, Cartesian mathematics, and empiricism, it is the technique that transforms nature into a "freely obtained object" (LEMOS, 2002, p. 45).

According to Rüdiger (2007, p. 175) "technique is, essentially, a mediation of the learning process of human life under certain social conditions". Technology is:

> the operational knowledge we designate by the term technique as it articulates to the form of knowing which we call science, through mediation from machines and potentially all passive areas of automation, as defined by the era it was created in; Modernity (RÜDIGER, 2007, p. 186).

According to Heidegger, technique is a form of existence in the world for man that, since Modernity, has taken an aggressive

stance towards nature; it is now subject to human understanding and a constant linear progress which cannot be stopped. For many writers, such as Sennett (2009), this is like the opening of Pandora's Box, the goddess of invention sent to Earth by Zeus. For the Greeks, this also represented Man's culture of producing which in itself could be harmful.

The great global conflicts of the first half of the twentieth century, Nazism, the Cold War, and the fear of nuclear threats are materializations of the Greeks' worst fears in a world that theoretically should be more evolved by the simple fact that technology exists.

Thinking about the relations between society and technology resulted in new fields emerging like *Science and Technology Studies* (STS). Thinkers such as Castells (1999) and Feenberg (2002) have dedicated themselves to formulating a scenario compatible with the challenges of studying such a clearly complex and multifaceted relation.

In his critique of the simplistic views held on the role of technology in contemporary times, Feenberg first proposes a mapping of the commonly presented positions, then tries to include issues such as democracy, power, and liberty as factors which are also important to consider in STS discussions.

Feenberg's cartography (2002) of modern societies has technology occupying a special place among the sources of power in social surroundings. He states that the political decisions which define much of the aspects of our daily lives are influenced by those who control the technical systems whether they be large corporations, the military or professional group associations like physicists, engineers, and more recently, software developers.

In making this determination, he refers to the thoughts of Marx in the nineteenth century who criticized the idea that the economy could be governed solely by extra-political factors; for example, the law of supply and demand. Similarly, to think about the role of technology without evaluating the diverse relations it establishes with society could result in a diminished view of the problem.

Along the same lines as the Marxist criticism of an economy governed by a natural and inexorable order, Feenberg (2010) views the rationality of technology through the assessment that its origin and development occur in a human's world, which is why they are also influenced by him.

> Technical creation involves an interaction between reason and experience. Knowledge of nature is necessary in order to make a working device. This is the element of technical activity we think of as rational. Yet the equipment needs to work in a social world, and the lessons of experience in that world influence the design (FEENBERG, 2010, p.17)[2].

It is a wide-ranging discussion in the field of philosophy, and cinema has translated this fear and fascination of machines over the decades into a wide array of films where technological solutions are represented by robots, automatons, machines, and even sophisticated computer programs. There are ships controlled by automated beings who rebel against humans, such as the computer HALL 9000 in the film "*2001-A Space Odyssey*" by Stanley Kubrick (1968). There are robots that carry out the extinction of humans in "*The Terminator*" by James Cameron (1984), and even robots that enslave humanity in a digital world created solely for the purpose of using them as energy sources, as depicted in "*The Matrix*" by the Wachowski brothers (1999).

In the TV series *Star Trek: The next generation* and the feature-length film (*Star Trek – First Contact*, Jonathan Frakes, 1996), one of the worst alien threats ever faced was the Borgs; a race of hybrid, biological, machine-like beings who very quickly took over the areas they invaded. They would insert implants into their victims, converting them into borgs, thereby integrating them to the central command. They acted like a colony of insects but on a much larger scale.

Yet this fascination with machines goes back much further than the almost obligatory dependence we currently have on cell phones, *smartphones*, *tablets*, and a myriad of other *gadgets* which we cannot do without anymore.

Reporting on automatons was restricted in ancient times and the Middle Ages; the eighteenth century is considered as its golden age. In an excerpt from Devaux (1964), he describes one such automaton that can still be seen in Paris today; Roentgen's "Tympanum Player", a musical doll which is supposed to be a depiction of Marie-Antoinette[3].

> In a room in the Palace of Versailles, among all the dresses and court gowns, the master automatist Roentgen presents Louis XVI with another work of art. The Tympanum Player, her divinely-shaped doll body in a low-cut corset and silk-broidered dress, triggers a certain interest and amazes with her precision and gracefulness. A short, lively aria comes to life in a flurry of

> ivory hammers; a whole century draws from this elegant and dry music. And when a young girl with her pouf hairstyle, turns her head around to greet them, the resemblance between them brings the room to whispers...a more touching and perfect mechanism than the Registry or Musician of Jaquet-Droz or the Duck from Vaucanson and his Flutists, the tympanum player faithfully echoes what we believe, the First Age of Automatism (DEVAUX, 1964, p. 7).

Figure 1- Roentgen's Tympanum Player, restored in 1864 by Robert Houdain.

Source: Lutice Créations ([200-]).

2. Journalism and Technology

Since the beginning, journalism has been connected to some form of technology. The printing press introduced by Gutenberg and its development is one of the major factors that has driven the expansion of journalism.

Much later, in the twentieth century, the arrival of networks, the Internet, and computers in newsrooms started a cycle of major changes still currently going on. Some people, such as Soria (2014), describe it as a *tsunami*, referring to the huge positive and negative impact that the digitalization of much of the production process has caused.

Machado (2003), while describing the beginning of the change, tells us that in order to make sense of what was happening, two positions were established. The first one we will call instrumentalist. It understood the fact that computers were just one more tool at the journalists' disposal; additional instruments

used to perform their jobs, just as the telegraph, the typewriter, and telex had been before.

The second one is the arrival of digital, which represented an even more extensive change.

> The lack of clarity on the consequences for journalism in disseminating in digital format makes it difficult to fully understand the particularities of online journalism, the changes to professional profiles, the structure of journalism companies and the roles that users have taken in producing content (MACHADO, 2003, p. 2).

Bradshaw and Rohumaa (2011) trace a brief history of the beginning of western online journalism and name the 1986 British newspaper *Today* as being a pioneer in using digital technology for production, they also point to the *Daily Telegraph* as one of the first newspapers to have an online version[4] on the then still fairly unknown Internet in 1994.

Nowadays, information is continuously flowing through digital newsrooms. It is through this information that we build our stories, stories that sometimes are just a few words in the *latest news* column or big reports on journalism sites such as the award-winning *Snow Fall* (BRANCH, ([200-]) in the *New York Times*.

However, the changes in technology and their impact cannot be evaluated on their own because social and economic factors appear to also make up the interlacing pathways within the complex contemporary media.

Haak, Parks and Castells (2012) present a range of trends when thinking about the future in a digitally interconnected age. They state that the new technological possibilities did not generate a crisis in journalism (which continues to have a fundamental social role) but they did help towards understanding that business models, which large media companies use, need to be reviewed or updated.

They also explain that the essential activities in journalism of observing important facts and asking the right questions to the right people, trying to understand observations and responses within a certain context, and, ultimately, explaining these results to others are synthesized into data collection, interpretation, and narration. In their opinion, this essential nucleus of doing journalism has not changed but been reshaped and expanded to include new technologies.

In detailing these premises, Haak, Parks and Castells (2012) list new tools and practices which would be the main trends in the future

of journalism which they propose to describe. They are: *networked journalism*, *crowdsourcing* and *user-generated content*, *data mining*, *analysis*, *visualization* and *mapping*, *visual journalism*, *point of view journalism*, *automated journalism*, and *global journalism*.

Analyzing all of these trends is beyond the scope of this text. Our focus is on the penultimate item on the list; automated journalism (AJ). In short, automated journalism can be explained by the fact that, nowadays, part of the journalistic content published is not written by humans anymore, but by machines. They use software, tools and solutions taken from simple lists of words (shown in the experiment we will present shortly) to complex models of artificial intelligence.

Regardless of what method is used to realize such a task, the fact that text and journalist are so radically disconnected appears to us to represent more problematic issues in technology yet at the same time it is interesting, precisely because it is the technology that operates with the whole imagination described in the introduction.

We would also like to point out that this is a very recent theme within Journalism studies and, as we understand it, should not be mistaken for Digital Journalism in Databases (DJDB) (BARBOSA, 2007, 2008, 2009, 2011; FIDALGO, 2004, 2007; MACHADO, 2006; RAMOS, 2011a, 2011b) which has already been researched by various writers, organizing conclusions based on databases "as defining structure and organization, as well as the composition and presentation of journalism content" (BARBOSA, TORRES, 2013, p.154).

Even within the vast spectrum of functionalities covered by DJDB, the section describes automation as:

> Inherent to the use of databases for storing, structuring, organizing, and presenting information. It allows for flexibility in verification processes, formatting file content as well as statistical dynamics or News Recommendation System (NRS), among others. There are three types of automation: partial, procedural (intermediate level), and total. (BARBOSA, TORRES, 2013, p.154-155).

The first noticeable difference is that databases constitute a specific kind of software, like electronic spreadsheets, word processors, and software for developing and presenting. The algorithms of artificial intelligence (AI), which support automated narratives even though they operate by connecting to or accessing databases, belong to a different

category and should not be confused with others, mainly because of the logic of specific procedures which they operate on.

Simply put, if the body of database work is to establish and commute relations between data which can then be adapted to multiple output forms, then AI software basically faces the problem of representing real world processes inside a computing environment. AI software learns and performs new functions by processing sets of data received as input, which is why they are more complex and rich in terms of what they can offer.

This complexity is also evident in the type of manipulation that databases and AI solutions propose. Database functions for processing digital journalism operate on a macro level; setting up a page dedicated to all the information on a particular football team using pre-existing content and metadata[5] connected to them. The AI algorithms in automated journalism operate on a micro level; they make up the actual text, indicating a new barrier within computing resources in journalism.

Now that the news clippings and the main lines of differentiation between DJDB and automated journalism have been defined, this paper turns its attention to:

a - understanding how this is done and by whom;

b - evaluating, on an exploratory capacity, the kinds of impact that are noticeable through analysis of the first academic works done on this object;

c - continuing on the experimental line under which we operate to learn more about this trend by developing a concept to replicate it on a simple level.

3. Automated Narratives – *Narrattive Science and Automated Insight*

Morozov's (2012) suggestive title "*A robot stole my Pulitzer!*" is a report on the first artificial intelligence companies. One such company, *Narrative Science*[6], was in the business of generating news stories. The company produced automated journalistic content and sold it to news websites, mainly in the areas of sports and finance. A good part of the information is provided by numbers and relations between measurable quantities like the rate of the dollar or the result of a football game.

Figure 2 – Printed article on automated journalism

HOME / FUTURE TENSE : WHAT'S TO COME?

Article from future tense
ASU | NEW AMERICA | SLATE

A Robot Stole My Pulitzer!

How automated journalism and loss of reading privacy may hurt civil discourse.

By Evgeny Morozov | Posted Monday, March 19, 2012, at 7:11 AM ET

Automated journalism like that produced by Narrative Science could perhaps save media jobs, but it can also hurt civil discourse
William Gottlieb/Library of Congress.

Can technology be autonomous? Does it lead a life of its own and operate independently of human guidance? From the French theologian Jacques Ellul to the Unabomber, this used to be widely accepted. Today, however, most historians and sociologists of technology dismiss it as naive and inaccurate.

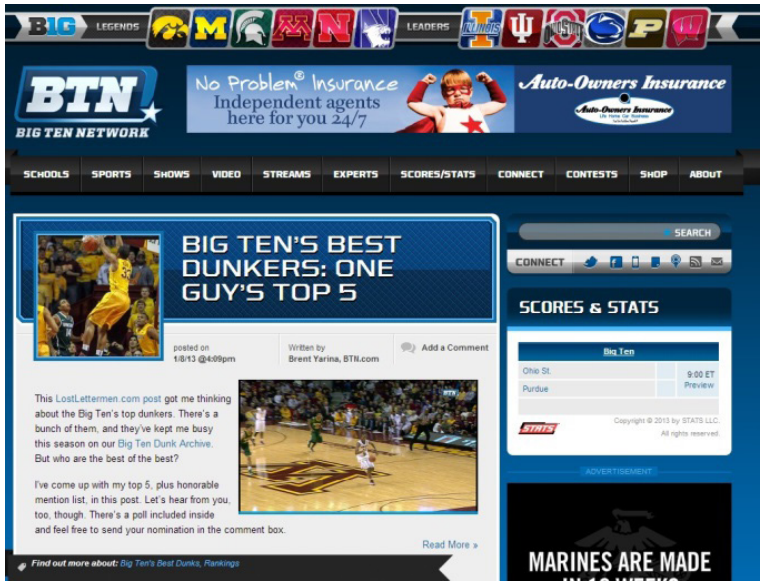
Yet the world of modern finance is increasingly dependent on automated trading, with sophisticated computer algorithms finding and exploiting pricing irregularities that are invisible to ordinary traders.

Meanwhile, *Forbes*—one of financial journalism's most venerable institutions—now employs a company called Narrative Science to automatically generate online articles about what to expect from upcoming corporate earnings statements. Just feed it some statistics and, within seconds, the clever software produces highly readable stories. Or, as *Forbes* puts it, "*Narrative Science*, through its proprietary artificial intelligence platform, transforms data into stories and insights."

Source: Morozov (2012).

Narrative Science (NS) got its start from a research project called "*Stats Monkey*" developed by students and professors of Computer Science and Journalism at *Northwestern University* through InfoLab. It basically wrote summaries of the results of American baseball games. In 2010, the company changed its name and shortly after patented the artificial intelligence platform *Quill*.

Figure 3 - Big Ten website specializing in sports and NS customers.



Source: Big Ten Network (2014).

Automated Insights (AI) is another company which provides automated journalistic content to various clients. It started developing under the name *StatSheet* in 2008 and was funded by a support agency in the state of North Carolina, USA. In 2014, according to its official site (AUTOMATED INSIGHTS, 2013), there were more than 300 million texts written automatically, from company reports to news reports.

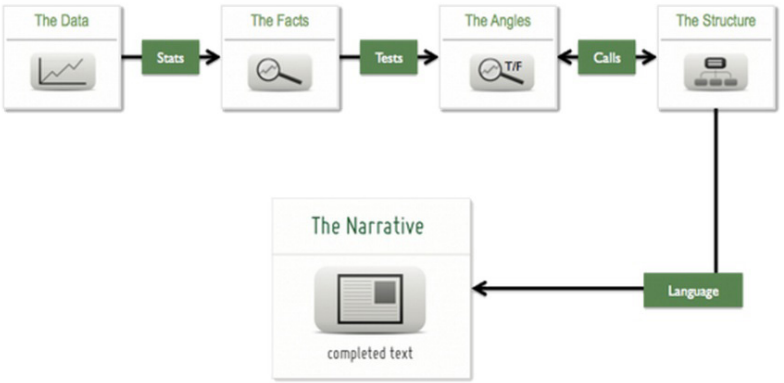
Figure 4 – Examples of AI content published on mobile devices

Source: Automated Insights (2013).

Arce (2009) had already evaluated the possibility of discourse automation and even included Lage's ideas (1997) on the issue; however, he only did so in theory and a non-experimental nature.

Coppin (2010, p.24) clarifies that one of the main questions in artificial intelligence is the representation of reality that the computer program will use: "for a computer to solve a real-world problem, it first needs an internal medium with which to represent the real world. Once it has this internal representation, the computer then becomes capable of solving problems".

Figure 5 – Transformation process from raw data to narratives on NS



Source: Narrative Science (2010).

In terms of journalistic content, the companies cited earlier in this paper started to produce *leads* as a basic way to present a defined internal structure which is then easily translated into a sequence of instructions which a machine will perform.

Among the new journalistic practices focusing on data, Carlson (2014) considers automated journalism to have the greatest disruptive power due to its limited human intervention, basically restricted to choices during program code development. We should also mention that this minimal human participation could occur with little or no professional journalist input since the main solutions in the market originate from private artificial intelligence companies. These companies use an industrial ownership model to protect their development processes and teams as well as to register patents.

In his research, which evaluates reactions published by

journalists on the use of *Narrative Science* services in editorial offices, Carlson writes that the emergence of AJ brings a series of questions about the future of journalism as a labour activity and the traditional standards that make up its content, even the actual identity and authority of journalism as a social function is questioned.

Clerwall (2014) approaches the question of software-generated content using an experimental approach. Journalistic texts, written by both humans and machines, are handed out to a group of readers without revealing whether the texts were written by humans or machines. He assesses the readers' perceptions of the quality, credibility, and objectivity of the narratives presented in the material.

Even though it is only an exploratory study, a few results merit attention. The texts generated by algorithms were defined as detailed and dull, yet highlighted for their objectivity. The study results also showed there was not a very clear distinction between which news was generated by journalists and which was generated by computers.

Dalen (2012) focuses his analysis on the abilities required for one to work in journalism and in news marketing. A little different from the view other professionals have on automated narratives; Dalen puts together a framework of positive and negative points.

The journalists that were consulted highlighted the analytical capability, the personalization, and the creativity as important points for features like factuality, objectivity, simplification and speed, these last few are closer to the scope of AJ. Even with journalists pointing this out, the positive point noted that automation of part of the narratives provides more time for investigating and further exploring material of greater importance.

Not as focused on automated journalism, the work from Lewis and Usher (2014) proposes to look at the possibilities of a meeting between journalism professionals and software developers from the study case of the initiative of global network *Hacks/Hackers*. Using the concept of trade zones, researchers analyzed the possibility of these two groups committing to and cooperating with each other, discussing the implications, challenges, and opportunities which could arise from such a partnership.

Much closer to a negative stance on automation, Latar (2014) describes the new logic of extracting data from large digital repositories as a trend willing to translate the complexity of social systems from the bits of information which we generate through

interactions and engagements on social media platforms and mobile devices; emerging trends in digital communication.

He associates this to the birth of a new science called *social physics*. Using atoms as an analogy, it is the microscopic parts of material that when studied, allow for conclusions to be drawn about what they make up. Within this context, Latar evaluates the emergence of robotic journalism as a by-product of this trend, based on the automated extraction of information from large silos of data and, using software, converting this knowledge into narratives for reading with no human intervention occurring in the production.

Considering the costs involved in traditional procedures of generating news, Latar warns of journalists becoming obsolete at the expense of consolidating software engineers and database managers becoming the most important employees within media companies.

> In parallel to the emergence of the new field of *social physics*, narration (the art of telling stories) is also becoming a scientific endeavour employing artificial intelligence algorithms to make use of the vast body of knowledge within linguistics and the study of natural languages. AI algorithms are being composed that can convert facts into readable stories in a fraction of a second[7]. (LATAR, 2014, p.65)

4. EXPERIMENTAL MODEL USING FOOTBALL RESULTS

To build our automated narrative experiment, we proposed to develop a programming code capable of writing small texts on the results of the 2013 Brazilian Football Championship. We used the programming language *Python*[8] as we believe it to be easier for amateur programmers like journalists and communication professionals to learn[9].

The *Python* language allows one to use various pre-developed programming models with specific purposes, making it easier to reach solutions by using a combination of functions with pre-existing codes. The *Natural Language Toolkit* library[10] (NLTK) which we used in this project is one such example and incorporates a large number of resources for word processing.

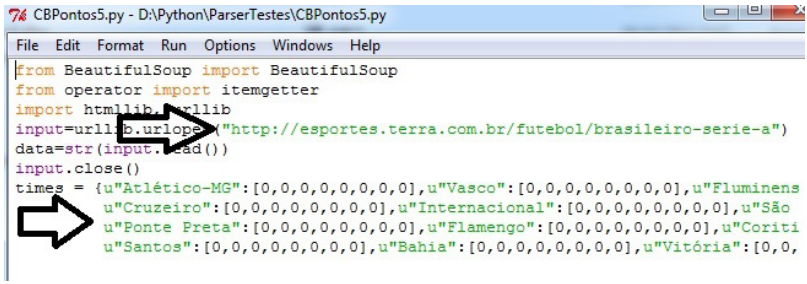
The model for the problem followed this sequence: first, obtain game results and complementary information like game site and the number of rounds, then register this information in a simple file structure which could be referred to afterwards to build material, then translate the rules of the tournament in terms of variables and

relations so that the syntax of regulations could guide continuity of text elements, and generate phrases out of the data results collected from the games.

We then move to a solution that, using a specific internet address where the data would be available, could automatically read all the initial information more quickly. For the tests we selected the *Terra* website sports pages which published the results from every round and updated the tournament bracket (PORTAL TERRA, 2014)[11]. The bracket was used as a tool to validate the software's calculations since it totalled the metrics that tournament rules generated like number of games, points won, goals scored, goals against, goal totals and approval rate.

With the data collection strategy in place, we then made the part of the code which saved these elements and associated them to each team. We used a *Python* structure called "*dictionary*" to do this. Each element, called a key, is given various values where each one represents information generated from the results of the game.

Figure 6 – Part of the code showing the data extraction address and the teams in the dictionary's key structure; initially all fields were blank.

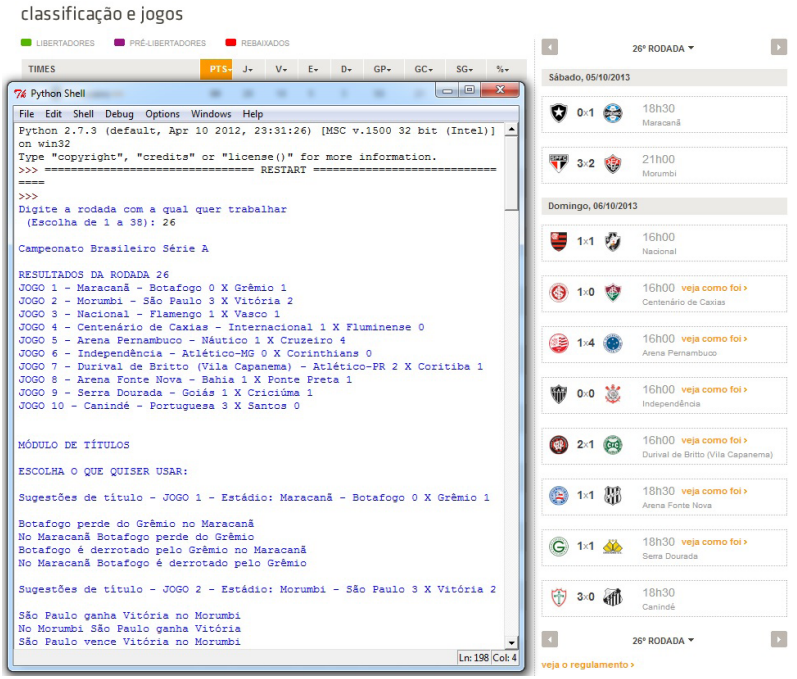


Source: Elaborated by author.

Once having started the code, the user only has to choose the number of rounds he wishes to look at. The software collects the results of all the rounds, finds and selects his, and then goes on to register the results and accumulate them in the dictionary. It is interesting to note that only game results are taken off the internet site. The software uses these results and applies the tournament rules to calculate the other values associated to the team. For example, when collecting the results of any given game, the software compares

the number of goals from both teams, if one has more than the other, that team wins the game and, consequently, the register of points earned goes up by three. The loser does not have anything summed up in the register and in the case of a tie, one point is awarded to each team indicating the points earned in a tie.

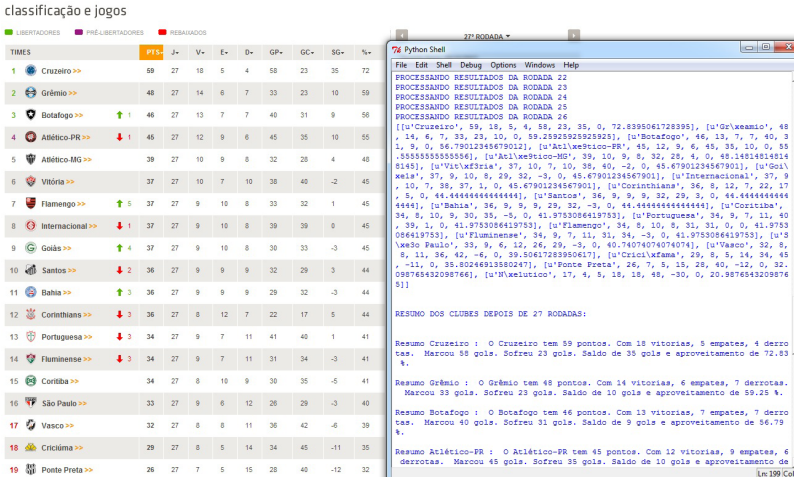
Figure 7 – Screen comparing the results from the website page and the screen generated by the program. The registered data is shown first, and the title suggestions based on the results after.



Source: Elaborated by author.

As the software registers the rounds of games, it is also updating all the additional parameters listed above which are representations of the actual tournament rules, including a set of data in the dictionary which will be used to calculate more information like a team's ranking within the bracket, the number of points each team has and its approval (calculated by dividing the total points earned by the total points available). These numbers allow the software to write more informative texts.

Figure 8 – Software screen showing the updated dictionary structure and a short summary of the positioning of the team in the championship based on registered elements

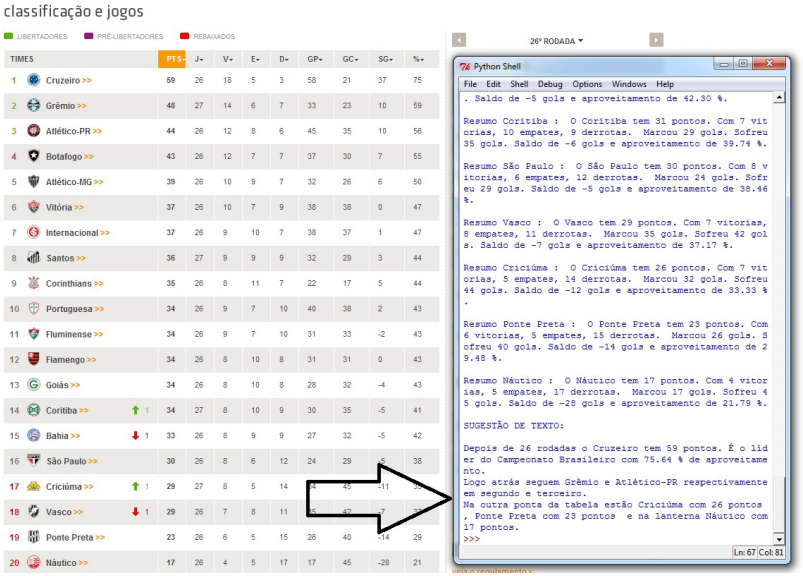


Source: Elaborated by author.

On a more complex level are leads, written with general information about one of the rounds of the championship. Basically, using the summary calculated from each team's individual situation, the software writes a text indicating the leaders' point totals and the last-place teams in the tournament, aspects which are normally reported in this kind of news report. Building this content, even though a bit more complicated, also comes from the idea of connecting smaller units of information using lists of common words and expressions in these types of texts.

By way of illustration, we can define a previous structure where some elements, like names of teams and their metrics, can be imagined as spaces to be filled out by whoever is occupying those positions in a particular round. The idea of dynamic files or, files that alter themselves over time, can be used here.

Figure 9 – Software screen with what would be the lead written from information read about a particular round of the championship



Source: Elaborated by author

Conclusions

Even though the experiment was just an exploratory one, it indicates the real possibility, and not just a theory, of producing some kinds of automated journalistic structures.

It was clear that content based on numerical information and easily translated mathematical expressions is more easily reproduced within a more restricted syntax like the one that can be extracted from the rules in championship sports.

The same thing we did with the Brazilian Championship results could be applied towards building something similar in order to generate small texts on the exchange rates or stock market values, weather forecasts for cities or regions, or other types of informative content that are built using a structure that repeats itself with small variations.

The ability to collect and process information in large quantities and varieties seems to point towards the potential for

using this kind of solution, particularly for online journalism and large internet sites that need to update their content quickly.

Theoretically speaking, it is important to remember that Database Journalism (DBJ) and Automated Journalism (AJ) are not the same and should not be mistaken for each other. They operate on different logics even though both are inserted in the larger evolutionary process of journalistic production routines through their use of technological resources.

The fact that its implementation uses algorithms of different categories is one of the reasons for this. While DBJ uses databases and a combinatory ability which results in outputs established through previously defined relations in its construction, the AJ mainly uses artificial intelligence algorithms capable of calculating new relations and literally learning as they are used and process the data made available to them.

Another notable difference is the level of granularity. The DBJ operates on a macro level, connecting entire news stories together from relative metadata, for example a personalized page for fans of a certain football club. The AJ, on the other hand, operates on a micro level, using words and a more basic construction of meaning via its planning and its syntactic and semantic relations.

The ramifications of this type of technology in the market cannot yet be evaluated. It is also worth highlighting that even the more complex artificial intelligence solutions are still far from replicating the subtleties and complexities of a good journalist, especially in a language like Portuguese. This poses difficulties, even today, for other software such as voice recognition and translation to reach high levels of efficiency.

One of the more recent studies on the theme evaluates the current situation of the AJ and states:

> The use of algorithms in recent years to automatically generate news from structured data has shaken up the journalism industry—especially since the *Associated Press*, one of the world's largest and most well-established news organizations, started to automate the production of its quarterly corporate earnings reports. Once developed, algorithms not only create thousands of news stories for a particular topic, but they can do it more quickly, cheaply, and potentially with fewer errors than any human journalist. Not surprisingly, this development has fueled journalists' fears that automated content will eventually eliminate newsroom jobs, while at the same time researchers and professionals see the technology's potential to improve the quality of news[12]. (GRAEFE, 2016, p. 4)

On the other hand, this uncertainty within the profession and the indiscriminate replication of releases and content from sources, justified by the simple fact of time pressure and the need to constantly update are risks to journalists since simple operations based in common structures have much more chance to be replicated automatically.

Further development of the work, investigative journalism, the extraction of complex relations from inter-related data and creation of info graphs and alternative forms of viewing information appear to us to be good examples of how human activity is still essential towards quality journalism. Improved curriculums and training programs in the area will also play an important role in the impact of new technology.

These premises which we have worked on over the last three years have been confirmed by the study *Tow Center* (GRAEFE, 2016) which aligns some of the main findings and consequences connected to the growth of automated journalism.

For example, the works shows how large news corporations are adopting AJ, basically guided by the growing availability of structured data (an important condition for working algorithms) and by the media companies' goal to reduce costs and increase the amount of available content. As previously discussed, this study is to determine the growth potential of AJ for using software to produce recurring topics more quickly, on a larger scale, and "potentially with less mistakes than a human journalist" (GRAEFE, 2016, p.5).

Another interesting aspect is the potential AJ has for creating news on demand out of user-questions on certain issues in order to achieve a greater level of personalization like generating content from the same set of data in different languages and angulations.

Among its limitations, Graefe (2016) highlights the fact that the software is based on data and inferences which could be subject to distortions and errors, therefore compromising the content and raising additional questions about who is responsible for the content as it continues to fall on the journalists or editors' shoulders. The transparency of the process, the information the user is given on how the algorithms operate is also under question. Graefe also reminds us that automated solutions cannot perform some essential work tasks such as explain a new phenomenon (which has not had any previous data collected on it), ask questions, and establish causality.

In relation to journalists, the study focuses on strengthening the human-machine relationship within newsrooms and suggests that they should focus on tasks which algorithms have difficulty performing, such as extensive analyses, interviews and investigative reporting. For society on a whole, the study points out that the excess of news content generated by AJ will make people's work more difficult when trying to find content which is more relevant to them.

If the phrase "resistance is futile"[13] is directly related to the relationship between humans and technology throughout the history of society, then in journalism a creative and well written text could guarantee us a peaceful coexistence with automated solutions valuable for repeating processes and low level of execution.

The information available on establishing a new genre of journalism is still inconclusive even though terms like "automated journalism" or "robotic" are being used more frequently. A specific form of narration is evidenced here, one based on a series of structural data, on the possibility of inference and semantic relations through the heavy use of large amounts of information devoid of human action and, according to authors such as Clerwall (2014) and Gaefer (2016), resulting in more objective and reliable texts.

It is interesting to note that the software we develop feeds on information intrinsic to the event or context to which it is directed as well as on internal relations established within. Nowadays, they deal only with simple questions but in the future...maybe they will be capable of identifying more complicated situations, and do so through the evolution of technology, like the development of a natural language; the *NLTK* solution we used in our experiment being one such example.

It appears that the automation of journalists which do not use complex human actions to perform its activities and practice the profession is much more harmful than the generation of journalistic texts via software. This seems to be the great problem we will have to face whether we are skeptical of, fearful of, or fascinated by technology.

*This paper was translated by Lee Sharp.

NOTES

1 The idea of cyber-illuminism is an extremely positive, somewhat naive, view of the relation between technology and humankind, normally represented by its innovative features that generate transformations capable of creating a more just and better world. Luddism supposedly gets its name from Ned Ludd, a weaver who led a movement preaching the destruction of machines in English weaving shops because they eliminated work positions. Some authors claim Ludd was a character created by a labour movement at the time in order to more easily spread campaign messages against the automation of the textile production at the beginning of the Industrial Revolution.

2 "Technical creation involves interaction between reason and experience. Knowledge of nature is required to make a working device. This is the element of technical activity we think of as rational. But the device must function in a social world, and the lessons of experience in that world influence design." – Our translation.

3 The history of the doll is detailed in the documentary "*L'Androïde de Marie-Antoinette*", available at: http://www.youtube.com/watch?v=pSxWmJLAaEg .

4 Mielnickzuk (2001) is telling us about the phases of digital journalism. He calls the first one transpositive because the printed content was just copied to the Internet without any major alterations.

5 Metadata is data that provides information about other data. Information such as the author of a text, the date the text was written, or the registers of all its versions. Like classification or *tags* showing which editor it belongs to, they are examples of metadata that are normally added to journalistic material via managing software and publishing content which is commonplace nowadays in newsrooms.

6 http://narrativescience.com/

7 "In parallel to the emergence of the new field of social physics, narration (the art of telling stories) is also becoming a scientific endeavor employing artificial intelligence algorithms to make use of the vast body of knowledge within linguistics and the study of natural languages. AI algorithms are being composed that can convert facts into readable stories in a fraction of a second." Our translation. The

Natural Language Toolkit (NLTK) library we used in the experiment described in this article is an example of a natural language software.

8 <www.python.org>.

9 Programming and journalism projects have been developed in the field of Investigative Journalism, aimed at extracting and processing data large quantities of data and using the information to build info graphs and narratives in digital journalism.<http://gijn.org/>.

10 <www.nltk.org>.

11 The current bracket address is<http://esportes.terra.com.br/futebol/brasileiro-serie-a/tabela>.

12 In recent years, the use of algorithms to automatically generate news from structured data has shaken up the journalism industry—most especially since the *Associated Press*, one of the world's largest and most well-established news organizations, has started to automate the production of its quarterly corporate earnings reports. Once developed, not only can algorithms create thousands of news stories for a particular topic, they also do it more quickly, cheaply, and potentially with fewer errors than any human journalist. Unsurprisingly, then, this development has fueled journalists' fears that automated content production will eventually eliminate newsroom jobs, while at the same time scholars and practitioners see the technology's potential to improve news quality (our translation).

13 "*Resistance is futile*"; phrase repeated by the Borgs to their victims in the series *Star Trek* (our translation).

REFERENCES

ARCE, Tacyana. O lead automatizado: uma possibilidade de tratamento da informação para o jornalismo impresso diário. **Revista Exacta**, Belo Horizonte, v. 2, n. 3, 2009.

AUTOMATED INSIGHTS. 2013. **Site Internet**. Available at: <www.automatedinsights.com>. Access on: 10 Jan. 2013.

BARBOSA, Suzana; TORRES, Victor. O paradigma 'Jornalismo Digital em

Base de Dados': modos de narrar, formatos e visualização para conteúdos. Galaxia (São Paulo, Online), n. 25, p. 152-164, June 2013.

BARBOSA, Suzana. Jornalismo em ambientes dinâmicos: perspectivas, tendências e desafios para a criação de conteúdos em tempos de convergência. In: Actas III Congreso Internacional de Ciberperiodismo y Web 2.0. Bilbao, Espanha: Universidad del País Vasco, 2011.

_____. Modelo JDBD e o ciberjornalismo de quarta geração. In: FLORES VIVAR, J. M.; RAMÍREZ, F. E. (Ed.). **Periodismo Web 2.0**. Madrid: Editorial Fragua, 2009. p. 271-283.

_____. Modelo Jornalismo Digital em Base de Dados (JDBD) em Interação com a Convergência Jornalística. In: Textual & Visual Media. Revista de la Sociedad Española de Periodística. vol. 1, Madrid, 2008, p. 87-106.

_____. Jornalismo Digital em Base de Dados (JDBD) – um paradigma para produtos jornalísticos digitais dinâmicos. (PhD Dissertation). PósCOM/UFBA, 2007. Available at: <http://migre.me/aTuYN>. Access on: 4 Feb. 2012.

BIG TEN NETWORK. 2014. **Portal Internet**. Available at: <www.btn.com>. Access on: 12 Apr. 2014.

BRADSHAW, Paul; ROHUMAA, Liisa. **The online journalism handbook**: skills to survive and thrive in the digital age. Essex: Pearson Education, 2011.

BRANCH, John. Snow Fall: the avalanche at Tunnel Creeak. **The New York Times,** New York, (200-]). Available at: <http://www.nytimes.com/projects/2012/snow-fall/?forceredirect=yes#/?part=tunnel-creek>. Access on: 2 June. 2014.

CARLSON, Matt. The Robotic Reporter: automated journalism and the redefinicion of labor, compositional forms and journalistic authority. In: LEWIS, Seth C. (Ed.). **Digital Journalism**. Vol. 3, nº 3. New York: Taylor&Francis Online, 2014, p. 416-431.

CASTELLS, Manuel. **A sociedade em rede**. São Paulo: Paz e Terra, 1999**.**

CLERWALL, Christer. Enter the Robot Journalist: User's perception of automated content. In: **Journalism Practice**. Special Issue – Future of Journalism in an age of digital media and economic uncertainty. Volume 8, Issue 5. New York: Taylor&Francis Online, 2014.

COPPIN, Ben. **Inteligência artificial**. Rio de Janeiro: LTC, 2010.

DALEN, Arjen. The Algorithms Behind the Headlines: How machine-written news redefines the core skills of human journalists. **Journalism Practice**. Volume 6, Issue 5-6. New York: Routledge, 2012, p. 648-658

DEVAUX, Pierre. **Autómatos, automatismo e automatização**. Tradução Luis Borges Coelho. Lisboa: Editorial Gleba, 1964.

ELLUL, Jacques. **A técnica e o desafio do século**. Rio de Janeiro: Paz e Terra, 1968.

FEENBERG, Andrew. E-book. **Transforming technology**: a critical theory revisited. New York: Oxford University Press, 2002.

_____. E-book. **Between reason and experience**. Essays in technology and modernity. Cambridge, MA: Mit Press, 2010.

FIDALGO, Antônio. A resolução semântica no jornalismo online. In: BARBOSA, S. (Ed.). J**ornalismo digital de terceira geração**. Covilhã, PT: LivrosLabCOM, 2007. p. 93-102.

_____. Do poliedro à esfera: os campos de classificação. A resolução semântica no jornalismo online. In: Anais II Encontro Nacional da SBPJor. Salvador-BA/Brasil, 2004.

GRAEFE, Andreas. **Guide to Automated Journalism**. Tow Center for Digital Journalism. Janeiro, 2016. Available at: https://www.gitbook.com/book/towcenter/guide-to-automated-journalism/details. Access on: 20 Jan. 2016.

HAAK, Bregtje; PARKS, Michael; CASTELLS, Manuel. The Future of Journalism: Networked Journalism. In: **Internacional Journal of Communication**. Vol. 6. 2012.

LAGE, Nilson. O lead clássico como base para a automação do discurso informativo. In: CONGRESSO BRASILEIRO DE PESQUISADORES DA COMUNICAÇÃO INTERCOM, 20., 1997, Santos. **Anais**... Santos, SP. 1997.
LATAR, Noam. The Robot Journalism in the Age of Social Physics: The end of human journalism? In: **The New World of Transitioned Media.** Springer, 2015.

LEMOS, André. **Cibercultura**: tecnologia e vida social na cultura contemporânea. 4. ed. Porto Alegre: Sulina, 2002.

LEWIS, Seth; USHER, Nikki. Code, Collaboration and The Future of Journalism: A case study of the Hacks/Hackers global network. In: **Digital Journalism**. Routledge Online, 2014.

LUTICE CRÉATIONS. **Site Internet**, Paris, [2000-]. Available at: <http://www.automates-boites-musique.com/>. Access on: 7 Apr. 2014.

KNIGHT, Megan; COOK, Clare. **Social media for journalists**: principles e practice. Londres: Sage, 2013.

MACHADO, Elias**. O ciberespaço como fonte para os jornalistas**. Salvador: Calandra, 2003.

_____. Elias. O Jornalismo Digital em Base de Dados. Florianópolis: Calandra, 2006.

MIELNICZUK, Luciana. **Características e implicações do jornalismo na web**. 2001. Available at: <http://200.18.45.42/professores/chmoraes/comunicacao-digital/13-2001_mielniczuk_caracteristicasimplicacoes.pdf>. Access on: 8 Sep. 2010.

MOROZOV, Evgeny. **A robot stoled my Pulitzer!**: future tense. 2012. Available at: <http://www.slate.com/articles/technology/future_tense/2012/03/narrative_science_robot_journalists_customized_news_and_the_danger_to_civil_discourse_.html>. Access on: 11 Apr. 2014.

NARRATIVE SCIENCE. 2010. **Site Internet**. Available at: <www.narrativescience.com>. Access on: 10 Jan. 2013.

PORTAL TERRA. Esporte. 2014. **Portal Internet**. Available at: <http://esportes.terra.com.br/futebol/brasileiro-serie-a>. Access on: 31 May 2014.

RAMOS, Daniela. **Formato: condição para a escrita do Jornalismo Digital em Bases de Dados.** Uma contribuição da semiótica da cultura. (PhD Dissertation). ECA/USP, 2011. Available at: <http://migre.me/aTvzX>. Access on: 30 April 2015.

RÜDIGER, Francisco. **Introdução às teorias da cibercultura**: tecnocracia, humanismo e crítica no pensamento contemporâneo. 2. ed. Porto Alegre: Sulina, 2007.

SANTOS, M. Journalism and the Internet of Things: can raw data change everything, again? In: **International Conference on Integrated Journalism Education, Research and Innovation (Resumos).** Integrated Journalism in Europe – IJE, 2015, Barcelona, pag. 59. Available at: http://eventia.estaticos.econgres.es/2015_IJIE/BookProgramme.pdf>. Access on: 07 Jan. 2015.

SENNETT, R. **O artífice**. Rio de Janeiro: Record, 2009.

SORIA, Carlos. **Convergência de mídias**. 2014. Palestra apresentada ao Seminário sobre Integração Multimídia, São Luís, 2014.

Marcio Carneiro dos Santos is journalist and doctor by the Intelligence Technologies and Digital Design program at PUC-SP, with postdoctoral fellow at UnB, within the line of research in Theory and Communication Technologies. Master in Communications from the UAM-SP and an MBA in Marketing from FGV-Rio-ISAN. He is currently associate professor of the Department of Social Communication at UFMA in Journalism of Digital Networks area and coordinates the Media Convergence Lab - LABCOM / UFMA. Has published works in areas as Digital TV, Network Theory, Social Network Analysis, Philosophy of Technology and Innovation in Journalism. Fellow of CNPq DT-II Productivity in Technological Development and Innovative Extension.

RECEIVED ON:12/12/2015 | APPROVED ON: 27/01/2016