Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1806-2563
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v39i3.29763

Semiodiscursive analysis of TV newscasts based on data mining and image processing

Felipe Leandro Andrade Conceição[1*], Flávio Luis Cardeal Pádua[2], Adriano César Machado Pereira[3], Guilherme Tavares de Assis[4], Giani David Silva[5] and Antonio Augusto Braighi Andrade[5]

*[1]Instituto de Engenharia e Tecnologia, Centro Universitário de Belo Horizonte, Av. Professor Mário Werneck, 1685, 30455-610, Belo Horizonte, Minas Gerais, Brazil. [2]Departamento de Computação, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [3]Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [4]Departamento de Computação, Universidade Federal de Ouro Preto, Ouro Preto, Minas Gerais, Brazil. [5]Departamento de Linguagem e Tecnologia, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte , Minas Gerais, Brazil. *Author for correspondence. E-mail: felipe.andrade@prof.unibh.br*

ABSTRACT. This work addresses the development of a novel computer-aided methodology for discourse analysis of TV newscasts. A TV newscast constitutes a particular type of discourse and has become a central part of the modern-day lives of millions of people. It is important to understand how this media content works and how it affects human life. To support the study of TV newscasts under the discourse analysis perspective, this work proposes a newscast structure to recover its main units and extract relevant data, named here as newscast discursive metadata (NDM). The NDM describes aspects, such as screen time and field size of newscasts' participants and themes addressed. Data mining and image analysis methods are used to extract and analyze the NDM of a dataset containing 41 editions of two Brazilian newscasts. The experimental results are promising, demonstrating the effectiveness of the proposed methodology.

Keywords: journalism, computing, discursive metadata, discourse analysis.

Análise semiodiscursiva de telejornais, baseada em mineração de dados e processamento de imagens

RESUMO. Este artigo aborda a análise semiodiscursiva de telejornais, baseada em técnicas de mineração de dados e processamento de imagens. Um telejornal constitui um tipo específico de discurso e desempenha papel central no quotidiano de milhões de pessoas. Objetivando-se apoiar o estudo de telejornais, sob a perspectiva da análise do discurso, este trabalho propõe uma estrutura para telejornais que permite recuperar suas principais unidades constituintes e extrair dados para sua análise. Estes dados são denominados neste trabalho de Metadados Discursivos de Telejornais (MDTs). Os MDTs descrevem aspectos como capital visual e plano fílmico dos participantes de telejornais, temáticas abordadas, entre outros. Técnicas de mineração de dados e processamento de imagens são utilizadas para extrair e analisar os MDTs associados a uma base contendo 41 edições de dois telejornais brasileiros. Os resultados experimentais são promissores, e demonstram a eficácia e aplicabilidade da abordagem proposta.

Palavras-chave: jornalismo, computação, metadados discursivos, análise do discurso.

Introduction

Study of TV newscasts is of great importance for media analysts in several domains, such as journalism, brand monitoring and law enforcement (Stegmeier, 2012; Van-Dijk, 2013). Because a TV newscast constitutes a particular type of discourse and a specific type of sociocultural practice (Van-Dijk, 2013), discourse analysis techniques (Charaudeau, 2002) have been applied to analyze the newscast structure at various levels of description, considering properties such as the overall topics addressed, schematic forms used and its stylistic and rhetorical dimensions (Cheng, 2012; Silva, 2008).

Traditionally, discourses have been analyzed without the support of computational tools, such as automated annotation software and information retrieval programs. However, with the fast development of computational linguistics, information retrieval and computer vision, novel methods have been frequently proposed to support discourse analysis, especially of multimedia content (e.g., TV newscasts) (Culpeper, Archer, & Davies, 2008; Baker, 2006).

As a step toward this goal, we present a corpus-based computational approach for discourse analysis of newscasts, which uses a specific newscast

structure to describe its main components by means of the here named Newscast Discursive Metadata (NDM), as well as techniques from image analysis and data mining domains. The NDM describes aspects such as screen time and field size of newscasts' participants and the theme addressed in each newscast component. The proposed approach was developed in partnership with the Brazilian TV channel Rede Minas in an attempt to provide media analysts with tools to assist their work, consisting in one of the components of an information system and created to support the discourse analysis of television programs (Pereira et al., 2015). As far as we know, it is the first methodology and approach for dealing with this demand for TV newscasts. Another contribution of this work was to develop a new methodology for registration and indexing of multimedia content, implemented in a Web information system.

Most computational studies of discourse have focused on written texts, such as the work of Biber and Jones (2005) and Marcu (2000), to cite just a few. In the work of Biber and Jones (2005), the authors use computational techniques based on a multidimensional analysis that combines corpus-linguistic with discourse-analytic perspectives to analyze the discourse patterns in a large corpus of biology research articles. Marcu (2000) explores the extent to which rhetorical structures can be automatically derived by means of surface-form-based algorithms. Conversely, a smaller group of computational studies on discourse have focused on spoken discourse or multimodal discourse (e.g., television broadcasts) (Rey, 2001; Passonneau & Litman, 1997; Al-surmi, 2012). Al-surmi (2012) adopts a corpus-based register analysis tool to investigate the extent to which soap operas, compared to sitcoms, reflect the linguistic representation of natural conversation. Rey (2001), in turn, performed a corpus-based study of dialogue spoken in the television series Star Trek, looking for differences between male and female language use. Passonneau and Litman (1997) proposes a method based on machine learning algorithms to automatically segment spontaneous, narrative monologues into units of discourse. The present work belongs to this last group of computational studies on discourse because it proposes a corpus-based approach for discourse analysis of newscasts, which are ultimately videos. Video data mining techniques are used to extract knowledge from the newscasts' editions to be studied under the discourse analysis perspective.

The remainder of this paper is organized as follows: The second section covers the material and methods proposed to support the discourse analysis of newscasts. Experimental results are presented in the third section, followed by the conclusions in the fourth section.

Material and methods

This section describes the proposed approach for discourse analysis of newscasts, which is divided in four steps, as illustrated in Figure 1. The first step consists of solving one of the main methodological problems of any corpus-based analysis of discourse structure, namely, the identification of the internal segments or units of the document to be analyzed, which are responsible for distinct communicative functions. Those discourse segments or, more specifically, their metadata are further used in the subsequent discourse analysis.

The second step comprises describing each discourse unit of the newscast by using the here-named Newscast Discursive Metadata (NDM). In the third step, data mining techniques are applied to extract knowledge from the newscasts' editions and detect patterns, which are finally evaluated in the fourth step under the discourse analysis perspective. The four steps of the proposed approach are described below.

TV Newscast structuring

A newscast is one of the most relevant programs within a television schedule and is considered in this work as a complex genre (Bakhtin, 1986), where several elements are organized according to a specific timeline. Newscasts are traditionally broken into familiar blocks (e.g., lead stories), whose structures are composed by content formats (Behnke & Miller, 1992), as well as compositional elements, such as an opening vignette and kicker.

Figure 1 illustrates a typical newscast edition, composed of a set of n blocks b_1,..., b_n. The newscast's structure is considered based on the content formats usually presented in the genre, disregarding the compositional elements. These content formats, understood here as discourse units, are represented in Figure 1 by a set of k possible distinct units u_1, ...u_k. Specifically, this work considers the eight content formats ($k = 8$) in the following:

1 - News Headlines: short summaries of stories that will follow in full in the newscast;

2 - Teaser: material promoting a story which 'teases' the viewer by hinting at, but not revealing, the real story that will be presented in the next block;

3 - Interview: a formal and structured conversation between a journalist and a source;

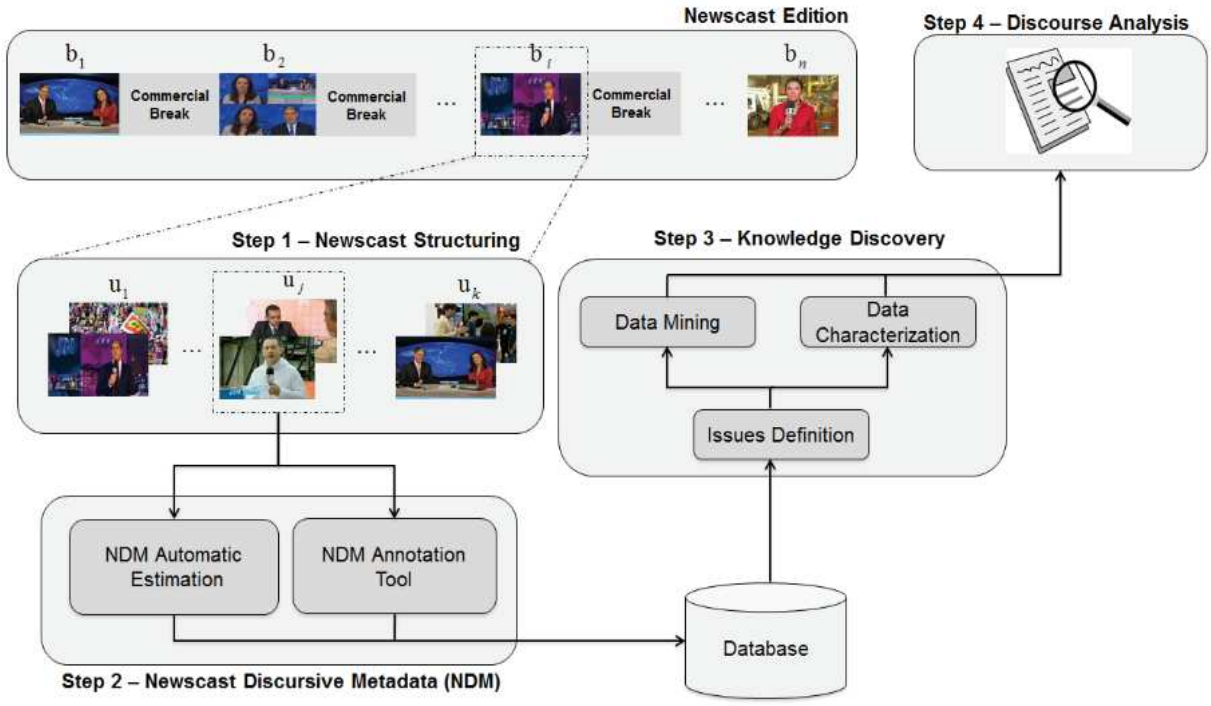

Figure 1. Overview of the proposed approach.

4 - News Report: the broadcast of a news sequence related to one or more themes;

5 - Float: a picture that is presented while the anchor is talking or interviewing a guest;

6 - Live Shot: a news story during which a reporter is live at a remote location;

7 - News Voiceover: a script read live by the anchor. In parallel, a video is shown;

8 - Standup: when a reporter speaks directly into the camera at the scene of the story.

The first step of the proposed approach is to identify those discourse units in each one of the newscasts blocks. Currently, this segmentation step is manually performed by a group of documentalists using specific tools of a multimedia information system that were specially created to support the discourse analysis of video recordings of television programs (Pereira et al., 2015). Those discourse units or, more specifically, their corresponding metadata are further used in the subsequent discourse analysis.

Newscasts Discursive Metadata (NDM)

At the heart of the proposed approach lies the concept of Newscast Discursive Metadata (NDM), which is used to describe the discourse units of a newscast's block. The NDM is determined during the second step and is essential to provide the means to describe, search and manage the videos of newscasts, ensuring maximum potential for their analysis.

The NDM is composed of two groups of metadata, namely, (i) a group that is automatically estimated and (ii) a group that is manually determined by documentalists using a specific annotation tool developed for this purpose. The metadata framework proposed by Pereira et al. (2015), based on the Dublin Core and MPEG-7 metadata schema, is used to store and manage the NDM. The two groups of metadata proposed are described in the next sections.

- Automatic Estimation. Two types of NDM are automatically estimated in this work: the screen time and the field size of each newscast's anchor and each newscast's reporter. To achieve this goal, image analysis techniques are proposed, as described in the following. Different from the current methodologies, those techniques are not prone to human error and do not place a significant demand on time or financial costs.

- Screen Time. The screen time of a newscast's participant is defined as his/her time of appearance during the newscast (Soulages, 2005). It may be considered as a strategy of the discursive staging and its estimation contributes to correlate the participant's discursive role with its relevance in the newscast. The automatic estimation of a participant's screen time is accomplished through a four-step methodology, as illustrated in Figure 2.

First, the participants' faces are detected in each frame of the discourse unit analyzed using the robust real-time algorithm proposed by Viola and Jones (2004). As a result, a list of faces with their corresponding frame labels is generated.

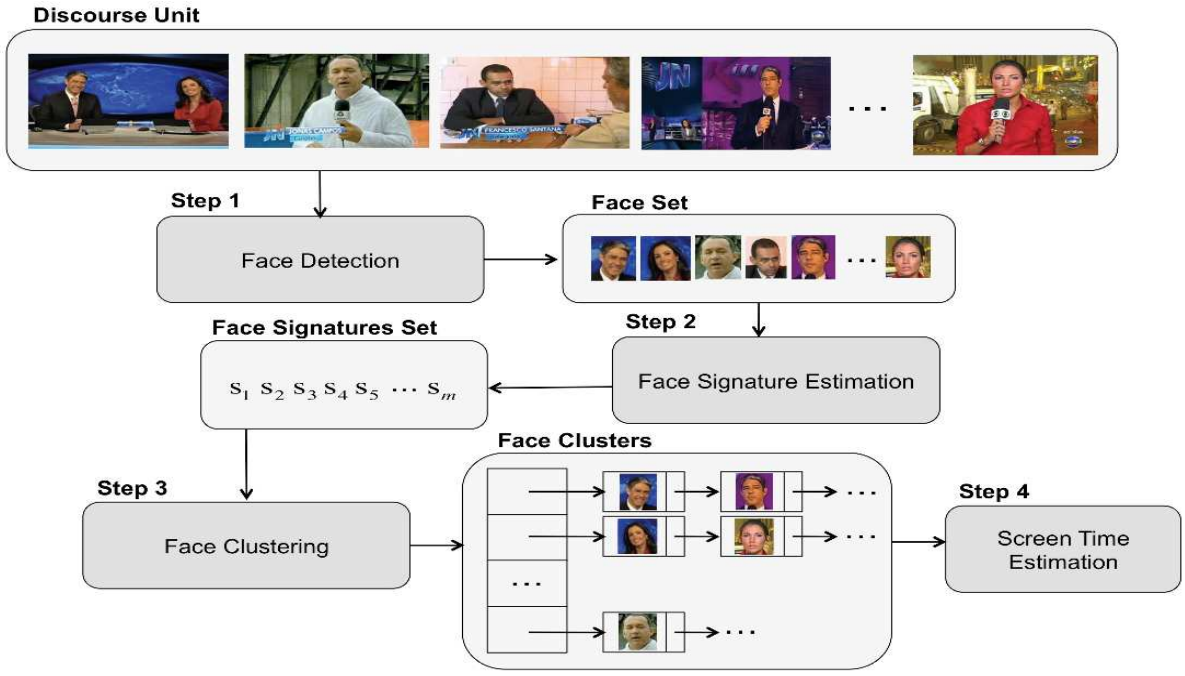

Figure 2. Methodology used to estimate the screen time of a newscast's participant.

The second step, in turn, is to estimate a visual signature for each detected face, by using color, shape and texture information. To compute this visual signature, we use the approach proposed by Souza, Pádua, Nunes, Assis, and Silva (2014), which obtains a visual signature containing 79 components representing each face image (54 refer to color, 18 refer to texture, and seven refer to shape positions).

The third step is essentially based on a clustering strategy to group the set of faces in such a way that faces in the same group are more similar to each other than to those in other groups. The fourth and final step consists of estimating the screen time for the newscast's participants of interest. The screen time of a participant in a specific discourse unit is directly estimated as the ratio between the number of his/her detected faces and the corresponding video's frame rate. The unit of measurement defined for screen time in this work is the second.

- Field Size. The field size refers to how much of a newscast's participant and his/her surrounding area is visible within the camera's field of view (Soulages, 2005) and is determined by two factors: the distance of the participant from the camera and the focal length of the lens used. This concept is usually applied in filmmaking and video production. Six types of field sizes are considered in this work, namely, Close-up (effect of intimacy), Medium Close-up (effect of personalization), Medium Shot (effect of sociability), American Shot (effect of sociability), Full Shot (effect of public space) and Long Shot (effect of public space). Those field sizes are illustrated in Figure 3.

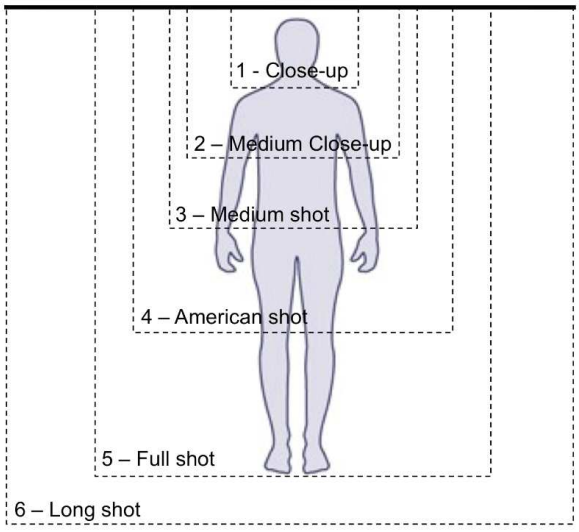

Figure 3. Basic types of field sizes for newscast participants.

Usually, the field size is only qualitatively defined (Soulages, 2005), as illustrated in Figure 3. In this scenario, to develop an automatic quantitative method to estimate a participant's field size at a given moment of the newscast, it was created a ground-truth. To achieve this goal, three discourse analysts classified 200 image samples belonging to three of the most popular Brazilian newscasts, namely, Jornal Nacional, Repórter Brasil and Jornal da Record. Each image sample was associated with

one of the six types of field sizes considered. Next, the ratio α between the face area and the complete area of the image plane was computed. Again, faces were detected using the method of Viola and Jones (2004). From the analysis of these ratio values, a set of ranges was proposed to determine the field size of a participant at a given instant, as shown in Table 1. The ranges estimated for the field sizes were successfully validated, achieving overall accuracy as high as 95%.

Table 1. Field sizes and corresponding ratios between face area and the complete area of the image plane.

Types of Field Size	Ratio (α)
Close-up	$(\alpha) > 0.40$
Medium Close-up	$0.28 < (\alpha) \leq 0.40$
Medium Shot	$0.22 < (\alpha) \leq 0.28$
American Shot	$0.19 < (\alpha) \leq 0.22$
Full Shot	$0.12 < (\alpha) \leq 0.19$
Long Shot	$(\alpha) \leq 0.12$

- NDM Annotation Tool. An annotation tool has been developed and incorporated into the multimedia information system proposed by Pereira et al. (2015) in order to describe the discourse units of a newscast's block by providing their corresponding NDM. In this case, NDM sets that must be manually provided have been proposed for all types of discourse units, as described in Table 2. Those NDM sets, as well as the screen time and field size of a participant, are jointly processed in the knowledge discovery step, when data mining techniques are applied to detect patterns and support the discourse analysis of the newscast's edition.

Knowledge discovery

Knowledge discovery is the process of characterizing, mining and processing data, aiming to extract relevant patterns in large data sets (Tan, Steinbach, & Kumar, 2005). Data characterization is critical because it allows identifying the real needs of end-users of a system. Conversely, the data mining phase is also responsible for finding patterns in the data, providing relevant information to users. As illustrated in Figure 1, a knowledge discovery process is performed during the third step of the proposed approach for discourse analysis of TV newscasts. In the following, data characterization and data mining techniques are applied in order to answer those issues.

- Issues Definition. The first step in the knowledge discovery process is the definition of the issues to be answered. Those issues are raised considering the NDM sets, as well as the information demands of media analysts regarding the investigation of TV newscasts. Table 3 presents the main issues considered by the proposed approach, which are especially related to the following aspects:

a) Themes addressed: the proposed solution can highlight the quantification of the frequency or the order of how news themes are presented in a specific period of time;

b) Discourse Units: a study of the temporal distribution of discourse units throughout the editions of a TV newscast allows to understand how the contents are reported;

c) Screen Time: allows the media analyst to correlate the discursive role of a specific participant with his/her relevance in the newscast;

d) Differences between TV newscasts: if common features allow the identification of patterns in newscasts and their characterization, the differences, in contrast, may support the definition of discursive identities to each newscast individually.

- Data Characterization. This step is performed in order to understand the users' needs (Tan et al., 2005). It works as a tool for statistical and quantitative analysis, which aims to answer the investigation issues. The characterized data can be visualized using statistical measures and graphical tools, such as histograms and probability distribution functions (PDF).

- Data Mining. Data mining is the computational process of discovering patterns in large datasets involving methods at the intersection of statistics, artificial intelligence and machine learning (Tan et al., 2005). Aiming to assist the analysis of the issues presented at the beginning of this section, three data mining techniques are used in the knowledge discovery process, specifically, (1) classification, (2) association rules and (3) sequence mining.

Discourse analysis

The final step of the approach is the task of analyzing the newscast discourse from the knowledge and patterns extracted in the knowledge discovery step. Such an analysis provides quantitative as well as qualitative alternatives to traditional methods of content analysis, allowing a systematic and interesting way to study this type of media (Colombo, 2004).

The knowledge extracted from the NDM sets contributes to understand the news-making process, especially the strategies used to represent the reality (Charaudeau, 2002). Moreover, by using those data, media analysts can map aspects, such as thematic organization, the themes addressed, the most frequent formats of the news-making process, shooting techniques and the representativeness of the anchor and reporter roles, among others. The discourse analysis approach (with the support of the

NDM sets) may contribute to the comprehension of newscasts as a genre by establishing comparative analysis between distinct editions of a specific newscast or between editions of newscasts of distinct television stations.

Results and discussion

This section presents and discusses the experimental results obtained. The experiments in this work are divided into three parts. The first and second parts present the results obtained from the data characterization and data mining techniques, respectively, which are used to extract and analyze the NDM associated to a dataset containing 41 editions of two popular Brazilian newscasts, namely, Jornal Nacional (22 editions, from Oct. 24 to Nov. 22, 2012) and Repórter Brasil (19 editions, 12 from Oct. 24 to Nov. 22 and 7 from Jul. 20 to Jul. 30, 2012), broadcast by the TV channels Rede Globo and Rede Minas, respectively. The data characterization and data mining techniques were applied to assist the analysis of the investigation issues in Table 3. The third part discusses the results obtained from the standpoint of discourse analysis.

Table 2. NDM sets for some newscast's discourse units: News Headlines, Teaser, Interview and Float.

Discourse Unit	NDM	Type	Description	Vocabulary
News Headlines Or Teaser	Slug	Text	Titles of headlines	Uncontrolled
	Time Slot	Time [mm:ss]	Time slot of this discourse unit	[0…60 : 0…60]
	Underlay	Boolean	Indicates if the presenters image is replaced by a picture when talking	*True*: replaced *False*: not replaced
Interview	Slug	Text	Title of interview	Uncontrolled
	Summary	Text	Interviews summary	Uncontrolled
	Interviewers	Text	Interviewers names	Uncontrolled
	Location	Text	Indicates if the interview is in a studio or outside	Studio, Outside
	Theme	Text	List of possible themes addressed	Politics, economy, police news, daily news, sport, behavior, education, culture, science and technology, weather, environment
	Time Slot	Time [mm:ss]	Time slot of this discourse unit	[0…60 : 0…60]
	Underlay	Boolean	Participants images are replaced by a picture when talking	*True*: replaced *False*: not replaced
Float	Slug	Text	Title assigned to the float	Uncontrolled
	Time Slot	Time [mm:ss]	Time slot of this discourse unit	[0…60 : 0…60]
	Summary	Text	Float's summary	Uncontrolled
	Location	Text	Float occurs nationally or internationally	National, International
	Theme	Text	List of possible themes addressed	Politics, economy, police news, daily news, sport, behavior, education, culture, science and technology, weather, environment

Table 3. Issues addressed by the proposed approach.

Issues	Dimension	What?	Objectives	Approach
Q_1	Themes	Identify the temporal distribution of themes	Realize the theme organization	Data Characterization
Q_2	Screen Time	Identify the hierarchies, sequences and usages of characters in the themes	Correlate participant's discursive role with its relevance	Data Characterization
Q_3	News Report	Identify the discursive roles of news report's participants	Realize the structural organization of news reports	Data Characterization
Q_4	News Report	Identify the temporal distribution of news reports in a period	Realize the structural organization of news reports	Data Characterization
Q_5	General	Identify ordering patterns of themes	Realize the theme hierarchy	Data Mining
Q_6	General	Identify matching rules between discourse units, themes and durations	Realize the relations between different discursive metadata	Data Mining
Q_7	General	Identify matching rules between field size, discursive role of each participant and themes	Realize the relations between different discursive metadata	Data Mining

Data characterization

The data characterization was used to analyze issues Q_1 to Q_4 in Table 3. By using graphical tools in this step, the discourse analyst may analyze the NDM associated to the dataset. Figure 4, for instance, allows for evaluating how each newscast distributes its themes throughout each edition (as a percent of the whole edition's time).

Figure 5, in turn, presents the temporal distribution (as a percent of the whole edition's time) of the main discourse units of each newscast individually. From Figure 4, one may note, for example, that the police news theme was much more frequently addressed than the others in the Jornal Nacional during the period analyzed (from October 24 to November 22, 2012).

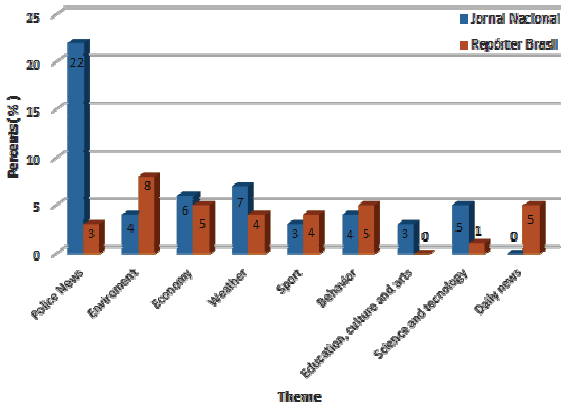

Figure 4. Temporal distribution of themes of two TV newscasts in Brazil: Jornal Nacional and Repórter Brasil.

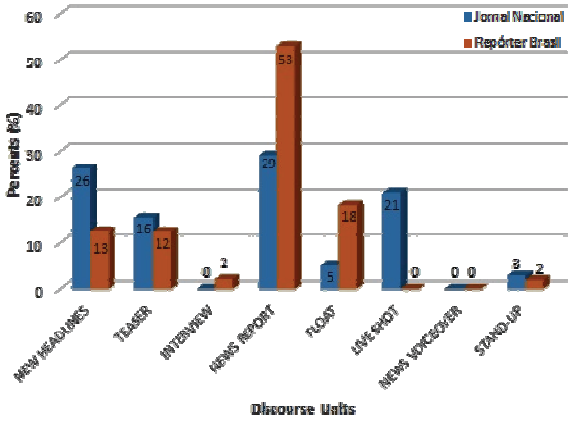

Figure 5. Temporal distribution of discourse units for each TV newscast, separately.

Data mining

The data mining techniques were used to extract association rules and sequences that enable the identification of patterns in the dataset. Those techniques were set with a minimum confidence value of 70%. Table 4 presents, for example, some new knowledge determined using this approach regarding issues Q_5 to Q_7 in Table 3.

To answer issue Q_5, the algorithm CSPADE (Mohammed, 2001) (Constraints Sequence Pattern Discovery using Equivalence classes) was used for mining sequences. This scenario evaluated the thematic hierarchy in the editions of both newscasts Jornal Nacional and Repórter Brasil. Through CSPADE, it was possible to identify, for example, that the themes police and political issues always occur in editions of both newscasts. When considering only the news, it was noted that in 95% of cases, the police issue always occurred sometime after the policy along each edition of the news.

Table 4. New Knowledge extracted in the data mining step.

Issues	New Knowledge
Q_5	In 100% of cases, the themes police and politics occurred in editions of both newscasts.
	In 100 % of cases, the police theme occurred after editing thematic policy be reported.
	In 95 % of cases, the police and political issues always occur in editions of both newscasts.
	In 95 % of cases, the police issue occurred after editing thematic policy be reported.
Q_6	In 54 % of cases, the thematic economy lasted less than one minute.
	In 90 % of the time, when the first issue was police, its duration was between 3 and 4 minutes.
	In 45 % of cases, the themes addressed in the second block lasted less than a minute.
	In 57 % of cases, the materials that approach the police issue occurred in the first block.
Q_7	In 100 % of cases, the first report was transmitted in the first block.
	In 95 % of the cases, the second report was transmitted in the first block.
	In 81 % of the cases, the third story was conveyed in the first block.
	In 57 % of the cases, the subject of police officers was reported in the first block.

Thus it is possible to identify a form of hierarchy between the two themes in question, with a final evaluation depending on other information such as time.

For issues Q_6 and Q_7, association rules were used to extract patterns of the dataset. An *a priori* algorithm was used with support of 20% for the extraction of association rules. This scenario obtained interesting results, confirming the existence of patterns between the news. In the next section, a discourse analysis of the results of the investigation issues is presented.

Discourse analysis

From the discourse analysis perspective, the data obtained from the survey reveals interesting information regarding the examination of television news. For example, the subject 'policy' is recurrent; it deserves the proposal of these programs, which exposes the dynamics of the executive, the legislature and the judiciary of the Repórter Brasil

attractions, because it is a program of a public broadcaster.

This statement emanates from the character of the contextual examination of Discourse Analysis. However, to the same extent, science is evidenced by detailed examination of the reports, which could indicate some positioning (sometimes partisan, sometimes as merely pro-government) in this communication vehicle - although the same can be done with the other news, whether with political themes or any of the others. The combination of "political" and "police" themes demonstrate a categorization of news, engendering prospects in semantic blocks. Such analysis should therefore be extended to verify if this has become a Brazilian scheme narrativization in newscasts. Already, sequencing (specifically the issue after the political police) can result from the predominance of political issues facing all others, as this entangles the responsibility for all social fronts.

The police theme, in turn, is notably higher in Jornal Nacional, which could represent a tendency of the vehicle to expose major ills of Brazil, in contrast to the intent of the government-controlled broadcaster that, although it should serve the interests of population, would not be interested by denouncing violence and just asks again. When analyzing the weather forecasting theme, we see almost double the attention in Jornal Nacional.

This happens because the TV channel Rede Globo invests in strategies to attract audiences with this entry; technological scenarios, with colors, animations, and a beautiful reporter indicating temperatures, are captivating the eyes of the viewer, rather than mere information strategies - this is fundamental to the operation of commercial media machinery (Charaudeau, 2002).

An important addendum, regarding the beauty of the weather reporters: concerning issue Q_2, which addresses the screen time, an increasing emergence of the presence of the presenters and reporters in enunciative news scenes is found in Brazil. This streamlines and humanizes programs, generating reciprocity with the viewing public - like the visual of capital holders (or managers), the information is, again, a targeted audience of apprehension, the assemblage of images with increasingly shorter takes and interviews with witnesses of the increasingly rapid events.

It is seen from the data obtained that sports stories have a place in the editions of television news - despite being smaller; these stories are usually the last block and represent the cosmos - a bit of stretching on the news, after all. News programs need to talk about what is important, that is information that is factual, emerging and impactful to ordinary life, at the end of which complementary information may be presented as a tactic to attract viewers.

To this end, news reports are the main mechanisms for the submission of information, occupying much of the display time for the news. That is, more than half of Repórter Brasil consists of reports, while Jornal Nacional devotes approximately one-third of its presentation to this type of input.

This distribution indicates a wider range of entries in the Rede Gobo program, stimulating more news. The program of the public broadcaster has more material, and more detailed material – contrasting the pace of private commercial networks of journalistic practice of broadcast television in Brazil, which otherwise may perhaps represent a limitation of teams and equipment or even, as expected, only an editorial choice.

Conclusion

This paper supports the work of researchers on Brazilian television systems, assists in preserving audiovisual memory, provides another important action to incorporate new services into the multimedia information system from the Center for Research Support on Television (CAPTE/CEFET-MG), provides a new methodology for registration and indexing of multimedia content, and implements a Web information system. In future work, we plan to validate our methodology in other datasets, especially from distinct newscasts, extend its application to other media genres and, finally, apply other types of promising computational approaches, such as machine learning and sentiment analysis.

Acknowledgements

The authors thank the support of Rede Minas, CNPq, FAPEMIG, CAPES and CEFET-MG.

References

Al-surmi, M. (2012). Authenticity and TV shows: a multidimensional analysis perspective. *TESOL Quarterly, 46*(1), 671-694.

Baker, P. (2006). Using corpora in discourse analysis. *Applied Linguistics, 28*(2), 327-330.

Bakhtin, M. M. (1986). The problem of speech genres. *Speech Genres and Other Late Essays, 17*(3), 60-102.

Behnke, R. R., & Miller, P. (1992). Viewer reactions to content and presentational format of television news. *Journalism & Mass Communication Quarterly, 69*(3), 659-665.

Biber, D., & Jones, J. (2005). Merging corpus linguistic and discourse analytic research goals: discourse units in biology research articles. *Corpus Ling. and Ling. Theory, 1*(2), 151-182.

Charaudeau, P. (2002). A communicative conception of discourse. *Discourse Studies, 4*(3), 301-318.

Cheng, F. (2012). Connection between news narrative discourse and ideology based on narrative perspective analysis of news probe. *Asian Social Science, 8*(1), 75-79.

Colombo, M. (2004). Theoretical perspectives in media-communication research: from linear to discursive models. *Forum: Qualitative Social Research, 5*(2), 1-14.

Culpeper, J., Archer, D., & Davies, M. (2008). Pragmatic annotation. In A. Ludeling, & M. Kytö (Eds.), *Corpus Linguistics* : an international handbook. (p. 613-642). (Handbooks of Linguistics and Communication Science). Berlin, DE: Mouton de Gruyter.

Marcu, D. (2000). The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics, 26*(3), 395-448.

Mohammed J. Z. (2001). SPADE: An Efficient Algorithm for mining frequent sequences. *Machine Learning, 42*(1), 1-2.

Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics, 23*(1), 103–139.

Pereira, M. H. R., Souza, C. L., Pádua, F. L. C., Silva, G., Assis, G., & Pereira, A. M. (2015). SAPTE: A multimedia information system to support the discourse analysis and information retrieval of television programs. *Multimedia Tools and Applications, 74*(1), 10923-10963.

Rey, J. M. (2001). Changing gender roles in popular culture: dialogue in Star Trek Episodes from 1966 to 1993. *Variation in English: Multi-dimensional studies, 1*(1), 138-156.

Silva, G. D. (2008). Análise Semiolinguística da identidade midiático-discursiva de telejornais brasileiros e franceses. *Glauks, 8*(1), 10-19.

Soulages, J. C. (2005). *Les mises en scène visuelles de l'information: Etude comparée, France, Espagne, Etats-Unis*. Paris, FR: Armand Colin.

Souza, C. L., Pádua, F. L. C., Nunes, C. F. G., Assis, G. T., & Silva, G. D. (2014). A unified approach to content-based indexing and retrieval of digital videos from television archives. *Artificial Intelligence Research, 3*(3), 49-61.

Stegmeier, J. (2012). Toward a computer-aided methodology for discourse analysis. *Stellenbosch Papers in Linguistics, 41*(1), 91-114.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA: Addison-Wesley Longman Publishing Co.

Van-Dijk, T. (2013). *News analysis*: case studies of international and national news in the press. (Routledge Communication Series). London, UK: Routledge.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision, 57*(2), 137-154.

Received on November 11, 2015.
Accepted on February 26, 2016.