Linguistic Features and Newsworthiness: An Analysis of News style

Maria Pia di Buono, Jan Šnajder

University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{mariapia.dibuono,jan.snajder}@fer.hr

Abstract

English. In this paper, we present a preliminary study on the style of headlines in order to evaluate the correlation between linguistic features and newsworthiness. Our hypothesis is that each particular linguistic form or stylistic variation can be motivated by the purpose of encoding a certain newsworthiness value. To discover the correlations between newsworthiness and linguistic features, we perform an analysis on the basis of characteristics considered indicative of a shared communicative function and of discriminating factors for headlines.

Italiano. *Questo contributo descrive uno studio preliminare sullo stile dei titoli nelle notizie, al fine di valutare la correlazione tra gli aspetti linguistici e il valore delle notizie. La nostra ipotesi è che ogni particolare forma linguistica o variazione stilistica possa essere motivata dall'obiettivo di codificare un certo valore di notiziabilità. Al fine di analizzare la correlazione tra il valore delle notizie e gli aspetti linguistici, effettuiamo un'analisi sulla base delle caratteristiche considerate indicative di una funzione comunicativa condivisa e di fattori discriminanti per i titoli.*

1 Introduction

Newsworthiness refers to a set of criteria by means of which quantity and type of events are selected in order to produce news (Wolf and de Figueiredo, 1987). That is to say, 'news is not simply that which happens, but that which can be regarded and presented as newsworthy' (Fowler, 2013). Galtung and Ruge (1965) identify a list of factors that an event should satisfy to become news; in other words, the likelihood of an event being considered newsworthy increases with the number of factors it complies with.

The newsworthiness factors reflect a set of values and provide a certain representation of the world (Fowler, 2013). This representation and the corresponding values are constructed and encoded in the language used in the news. For this reason, each particular linguistic form or stylistic variation can be motivated by the purpose of representing a certain value. According to Labov's axiom (1972), style ranges along a single dimension, namely the attention paid to speech. Bell (1984) refutes this axiom, stating that style can be considered also as a response to other factors. These factors constitute a new dimension of stylistic variation, that, in headlines, might be related to the necessity of reflecting newsworthy factors, and meeting two needs: attracting users attention and summarizing contents (Ifantidou, 2009).

This paper aims to provide a preliminary analysis of the linguistic features in news headlines and how these relate to specific newsworthiness categories. The analysis rests on the hypothesis that each particular linguistic form or stylistic variation can be motivated by the purpose of encoding a certain newsworthy value. The remainder of the paper is structured as follows. In Section 2, we describe the related work on stylistic analysis of news and headlines. In Section 3, we describe the data set and the classification scheme we use. In Section 4 we introduce our methodology together with the analysis we perform, while in Section 5 we discuss the results. Section 6 concludes the paper.

2 Related Work

Several works, based on sociolinguistic and discourse analysis frameworks, have investigated stylistic features and linguistic variations in both newspapers and headlines, on the basis of different parameters and aspects (Develotte and Rechniewski, 2001; Pajunen, 2008). The large amount

of existing contributions to the field is justified by the social implications of news media communication and its language.

A considerable amount of research has analyzed the language of news media from a broader prospective (Bell, 1991; Matheson, 2000; Cotter, 2010; Conboy, 2013; Fowler, 2013; Van Dijk, 2013). Generally speaking, these works emphasize the influence of news language on our perception of the world, due to the fact that news media operate a selection of events and narrative, and use the language to project those.

Another strand of research focuses on specific linguistic aspects in journalistic style. For instance, Tannenbaum and Brewer (1965) analyze the syntactic structure across different news content areas, while Schneider (2000) analyzes the textual structures in British headlines, revising the traditional distinction among verbal and nominal headlines.

3 Data Description

In our work, we adopt the data set proposed for SemEval-2007 task 14 (Strapparava and Mihalcea, 2007), which is a corpus formed by 1250 headlines, extracted from major newspapers and news web sites such as New York Times, CNN, BBC News, and Google News search engine. Originally, SemEval-2007 task 14 data set has been developed for emotion classification and annotated with emotion labels. Relevant for the purpose of the present work is the annotation of this dataset by di Buono et al. (2017), who provided additional newsworthiness labels ("news values"), using the scheme proposed by Harcup and O'Neill (2016). Harcup and O'Neill proposed a set of 15 values, corresponding to a set of requirements that news stories have to satisfy to be selected for publishing. They claimed that these criteria are related also to practical considerations, e.g., the availability of resources and time, and to a mix of other influences, e.g., who is selecting news, for whom, in what medium and by what means (and available resources), that can cause fluctuations within the suggested hierarchy. Di Buono et al. report that two out of 15 news value labels (Audio-visuals, News organization's agenda) were difficult to annotate out of context even for trained annotators, while two (Exclusivity, Relevance) were not well-represented in the data. Their final dataset thus contains 11 labels.

Table 1 lists the news value labels, their counts in the data set, and the inter-annotator agreement measured in terms of (adjudicated) kappa-score, as reported by Di Buono et al.

News value	Count	IAA
Bad news	85	0.74
Celebrity	82	0.76
Conflict	86	0.56
Drama	178	0.66
Entertainment	351	0.84
Follow-up	29	0.45
Good news	65	0.56
Magnitude	45	0.37
Shareability	130	0.34
Surprise	43	0.41
Power elite	166	0.72

Table 1: News values labels, their counts, and the inter-annotator agreement in terms of kappa-score.

4 Linguistic Features

Our methodology to define the stylistic variations related to newsworthiness categories relies on a descriptive analysis of different features, i.e., syntactic, lexical and compositional features.

We extracted these using Coh-Metrix,[1] a computational tool that provides a wide range of language and discourse metrics (Graesser et al., 2004; McNamara et al., 2014). Coh-Metrix has been developed on the basis of cognitive models in discourse psychology to detect both coherence and cohesion in texts. According to Louwerse (2004), "coherence refers to the representational relationships of a text in the mind of a reader whereas cohesion refers to the textual indications that coherent texts are built upon." Coh-Metrix describes coherence and cohesion by means of more than one hundred linguistic features, based on a multilevel framework, i.e., words, syntax, the situation model, the discourse genre, and rhetorical structure (Dowell et al., 2016).

We ran Coh-Metrix analysis on headlines from our dataset, grouped according to the 11 newsworthiness labels. We then analyzed these results manually and decided to adopt a subset of Coh-Metrix indices, which, according to our initial hypothesis, we consider to be discriminating factors for newsworthiness, i.e., *text easibility principal component* and *word information* indices. Being representative of linguistic characteristics and syntax context, such features are suitable to represent stylistic variations and, therefore, the underlied news value.

[1] http://cohmetrix.com

News value	PCSYNz	PCCNCz
Bad news	2.274	3.669
Celebrity	2.239	1.933
Conflict	1.834	1.992
Drama	2.502	2.194
Entertainment	1.923	1.788
Follow-up	1.27	2.646
Good news	1.741	3.122
Magnitude	2.585	2.873
Shareability	2.057	1.678
Surprise	1.451	3.253
Power elite	2.671	0.896

Table 2: Z-scores for PC Syntactic simplicity (PCSYNz) and PC Word concreteness (PCCNCz).

5 Analysis and Results

In our preliminary analysis, we consider two main types of linguistic features: *text easability* and *word information scores*.

5.1 Text Easability Features

Coh-Metrix text easibility indices ("Text easability principal component scores") are designed to measure text ease that goes beyond traditional readability metrics. We focused specifically on two indices related to the syntactic simplicity (PCSYNz) and word concreteness (PCCNCz) (Table 2).

The syntactic simplicity is evaluated on the basis of the number of words and the complexity of syntactic structures of sentences. As far as the syntactic simplicity is concerned, the variability among the categories is not so high, nevertheless, we can distinguish two groups. The first group, with a higher PCSYNz, consists of headlines labeled with the 'Power elite', 'Bad news', 'Shareability', 'Drama', 'Magnitude', and 'Celebrity' news values. Higher scores here indicate that the sentence presents more words and uses complex syntatic structures, as exemplifed by the following headlines from this group:

(1a) *China says rich countries should take lead on global warming* (Power elite)

(1b) *Iraqi suicide attack kills two US troops as militants fight purge* (Bad news)

(1c) *Second opinion: girl or boy? as fertility technology advances, so does an ethical debate* (Shareability)

(1d) *Damaged Japanese whaling ship may resume hunting off Antarctica* (Drama)

(1e) *Ready to eat chicken breasts recalled due to suspected listeria* (Magnitude)

(1f) *Jackass' star marries childhood friend The secrets people reveal* (Celebrity)

The second group consists of headlines labled with 'Entertainment', 'Surprise', 'Follow up', 'Good news', and 'Conflict', which received lower PCSYNz scores, and are thus of less syntactic complexity. Examples of headlines form this group are as follows:

(2a) *Action games improve eyesight* (Entertainment)

(2b) *Breast cancer drug promises hope* (Good news)

(2c) *Merkel: Stop Iran* (Conflict)

The second index, word concreteness, differentiates three groups of headlines: (i) 'Power elite', 'Entertainment', 'Shareability', 'Celebrity', and 'Conflict', all with a low z-score; (ii) 'Follow up', 'Drama' and 'Magnitude', with a medium z-score; and (iii) 'Bad news', 'Surprise' and 'Good news' with a high z-score. The following headlines exemplify each of the three groups:

(1a) *Action intensity boosts vision* (Shareability)

(2a) *Ex-suspect slams anti-terror laws* (Drama)

(3a) *Ancient coin shows Cleopatra was no beauty* (Surprise)

The word concreteness index measures the concreteness level of content words. Thus, news values with lower scores are characterized by a higer number of abstract words and, for this reason, may be less easy to understand without an appropriate context. Our analysis thus suggests that 'Bad news', 'Surprise', and 'Good news' headlines are typically refering to more concrete events and entities than the other categories of news values.

5.2 Word Information

This Coh-Metrix index refers to information about syntactic categories and function words, evaluated in the sentence context. To visualize the relations among newsworthiness and word information, we performed a hierarchical cluster analysis. We first represent each headline as a vector of ten word incidence scores (the number of words of a specific part-speech per 1000 words): incidence scores

for nouns, verbs, adjectives, adverbs, personal pronouns, pronouns in first, second, and third person, separately for singular and plural. We then use hierarchical agglomerative clustering with complete linkage and one minus Pearsons correlation coefficient as the distance measure to obtain the clusters.

Fig. 1 shows the resulting dendrogram. We can identify three groups of news values on the basis of their syntactic structures.

The first group consists of only news values that can be defined positive contents/sentiments, namely 'Good news', 'Entertainment', and 'Shareability'. This group is characterized by a quite high incidence of adjective, low incidence of first person singular and third person plural pronouns. Furthermore, this group presents the highest incidence of second person pronouns. As in the samples below:

(1a) *Feeding your brain: new benefits found in chocolate* (Good news)

(1b) *Free Will: Now you have it, now you don't* (Entertainment)

(1c) *Nap your way to a successful career* (Shareability)

The second group consists of 'Celebrity', 'Power elite', and 'Drama'. This group presents low incidence of adjective and adverbs. The most incident pronouns are the first person plural and the third person singular.

(2a) *Beyonce new SI bikini cover girl* (Celebrity)

(2b) *Bush vows cooperation on health care* (Power elite)

(2c) *Collision on icy road kills 7* (Drama)

The third group consists of two subsets, the first one formed by 'Surprise' and 'Magnitude', and the second subset formed by 'Follow up', 'Bad news', and 'Conflict'. 'Surprise' and 'Magnitude' form a different subset due to the presence of the highest score within all categories for the adjective incidence and a low incidence of pronouns. For instance:

(3a) *In the world of life-saving drugs, a growing epidemic of deadly fakes* (Surprise)

(3b) *Flu Vaccine Appears Safe for Young Children* (Magnitude)

The second subset is formed by negative contents/sentiment, characterized by the lowest incidence of adverbs and pronouns:

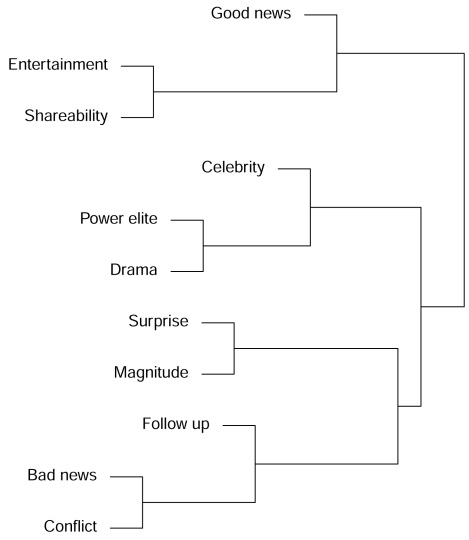

Figure 1: Dendrogram of the 11 newsworthiness categories based on the headline word information features.

(3c) *Eight years for Damilola killers* (Follow up)

(3d) *Bomb kills 18 on military bus in Iran* (Bad news)

(3e) *Venezuela, Iran fight U.S. dominance* (Conflict).

6 Conclusions and Future work

We described a preliminary study for on style of headlines in order to evaluate the correlation among syntactic features and newsworthiness. Our hypothesis is that each particular linguistic form or stylistic variation can be motivated by the purpose of encoding a certain newsworthy value. We performed a linguistic analysis to discover the correlations among newsworthiness and some stylistic features, on the basis of characteristics considered indicative of a shared communicative function and discriminating factors for headlines.

This preliminary analysis opens up a number of interesting research directions. One is the study of other stylistic variations of headlines, besides the ones examined in this paper. Another research direction is the comparison between style in headlines and full-text stories. It would also be interesting to analyze how communicative functions in headlines correlate with the events described in the pertaining text. We intend to pursue some of this work in the near future.

Acknowledgments

This work has been funded by the Unity Through Knowledge Fund of the Croatian Science Foundation, under the grant 19/15: EVEnt Retrieval Based on semantically Enriched Structures for Interactive user Tasks (EVERBEST).

References

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(02):145–204.

Allan Bell. 1991. *The Language of News Media*. Language in society. Blackwell.

Martin Conboy. 2013. *The language of the news*. Routledge.

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

Christine Develotte and Elizabeth Rechniewski. 2001. Discourse analysis of newspaper headlines: a methodological framework for research into national representations. *Web Journal of French Media Studies*, 4(1).

Maria Pia di Buono, Jan Šnajder, Bojana Dalbelo Bašić, Goran Glavaš, Martin Tutek, and Natasa Milic-Frayling. 2017. Predicting news values from headline text and emotions. In *Proceedings of Natural Language Processing Meets Journalism Workshop (EMNLP 2017)*, page to appear.

Nia M. Dowell, Arthur C. Graesser, and Zhiqiang Cai. 2016. Language and discourse analysis with coh-metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3):72–95.

Roger Fowler. 2013. *Language in the News: Discourse and Ideology in the Press*. Routledge.

Johan Galtung and Mari Holmboe Ruge. 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of peace research*, 2(1):64–90.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Tony Harcup and Deirdre O'Neill. 2016. What is news? news values revisited (again). *Journalism Studies*, pages 1–19.

Elly Ifantidou. 2009. Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics*, 41(4):699–720.

William Labov. 1972. *Sociolinguistic patterns*. Number 4. University of Pennsylvania Press.

Max M Louwerse. 2004. Un modelo conciso de cohesión en el texto y coherencia en la comprensión. *Revista signos*, 37(56):41–58.

Donald Matheson. 2000. The birth of news discourse: Changes in news language in British newspapers, 1880-1930. *Media, Culture & Society*, 22(5):557–573.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Juhani Pajunen. 2008. Linguistic analysis of newspaper discourse in theory and practice.

Kristina Schneider. 2000. The emergence and development of headlines in British newspapers. *English Media Texts, Past and Present: Language and Textual Structure*, 80:45.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Percy H. Tannenbaum and Richard K. Brewer. 1965. Consistency of syntactic structure as a factor in journalistic style. *Journalism Quarterly*, 42(2):273–275.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Mauro Wolf and Maria Jorge Vilar de Figueiredo. 1987. *Teorias da comunicação*. Presença.