JMLR: Workshop and Conference Proceedings 17 (2011) 63–71 2nd Workshop on Applications of Pattern Analysis

Automating Quantitative Narrative Analysis of News Data

Saatviga Sudhahar SAATVIGA.SUDHAHAR@BRISTOL.AC.UK
Intelligent Systems Laboratory, University of Bristol, UK

Roberto Franzosi RFRANZO@EMORY.EDU
Department of Sociology/ Program in Linguistics, Emory University, Atlanta, USA

Nello Cristianini NELLO.CRISTIANINI@BRISTOL.AC.UK
Intelligent Systems Laboratory, University of Bristol, UK

Editor: Tom Diethe, José L. Balcázar, John Shawe-Taylor, and Cristina Tîrnăucă

Abstract

We present a working system for large scale quantitative narrative analysis (QNA) of news corpora, which includes various recent ideas from text mining and pattern analysis in order to solve a problem arising in computational social sciences. The task is that of identifying the key actors in a body of news, and the actions they perform, so that further analysis can be carried out. This step is normally performed by hand and is very labour intensive. We then characterise the actors by: studying their position in the overall network of actors and actions; studying the time series associated with some of their properties; generating scatter plots describing the subject/object bias of each actor; and investigating the types of actions each actor is most associated with. The system is demonstrated on a set of 100,000 articles about crime appeared on the New York Times between 1987 and 2007. As an example, we find that Men were most commonly responsible for crimes against the person, while Women and Children were most often victims of those crimes.

Keywords: network analysis, computational social science, story grammars, semantic triplets, text mining

1. Introduction

News media content has been widely used in the social sciences to study socio-historical events (e.g., Franzosi (1987); Earl et al. (2004)). An event according to social scientists is an action performed by human beings that can be summed up by a verb or a name of action (Franzosi (2010)). Linguistically, an event can be expressed in the form of a semantic triplet Subject-Verb-Object (SVO) which consists of a subject as an actor, the action performed by the subject and the object of the action (on narrative, see Franzosi (1998)). This structure is referred to as story grammar. Computer-assisted story grammars have been used in social science research to analyse narrative text (e.g., Franzosis quantitative narrative analysis or QNA, Franzosi (2010)). The advantage of QNA over other forms of coding is that it analyses social reality in terms of actors rather than variables; it combines quality and quantity; it is based on a rigorous linguistic theory of narrative. Its disadvantage is that it is very labour intensive. All parsing of narrative texts into the constitutive objects of a story grammar (coding) must be done by humans, even in a computer-assisted environment. One of the motivations of the study presented here is to automate the process

of coding applying statistical approaches and methods to news narratives.

This paper describes an application that automatically extracts Subject-Verb-Object triplets out of Crime data from the New York Times corpus and then uses this information to carry out a number of tasks aimed at capturing the role played by key actors in the narration of the news.

Our approach builds on an idea presented in Rusu et al. (2008, 2007) and Dali and Fortuna (2008), for purposes of text summarisation and related tasks, and later developed also for the generation of event templates (Trampus and Mladenic (2011)) in information extraction. In this approach, Subject-Verb-Object (SVO) triplets are extracted from text by means of a parser, and then used to generate a semantic graph that captures narrative structures and relations contained in a text.

We developed that idea in various ways, in order to address the specific needs of researchers in QNA, with the goal of creating a working system that performs some key tasks needed in that application area, and that can scale to very large corpora. In doing so, we also introduced several new concepts. By weighting the actors, we can identify the players most identified with a given domain (eg: crime); by analysing the centrality of the actors, we can identify the most influential characters in the news narrative; by classifying the types of actions (eg crimes against person) we can further analyse the roles different actors play in crime (eg: perpetrator vs. victim); by analysing the time series of the actors centrality, we can identify important changes in its narrative role (as done by hand in (Franzosi (2010)) where the emergence of Italian fascism was investigated in the same way).

All this has been implemented in a very scalable way and has been developed in direct response to specific needs in social science investigations. We demonstrate the system here by analysing 100,000 articles about crime from the New York Times, but our goal is to integrate this approach into the larger NOAM infrastructure described in (Flaounas et al. (2011)).

2. System Pipeline and Experiments

The generation of key actors and actions from data consists of a series of sequential steps forming a pipeline (Figure 1). Co-reference resolution is defined as the identification of

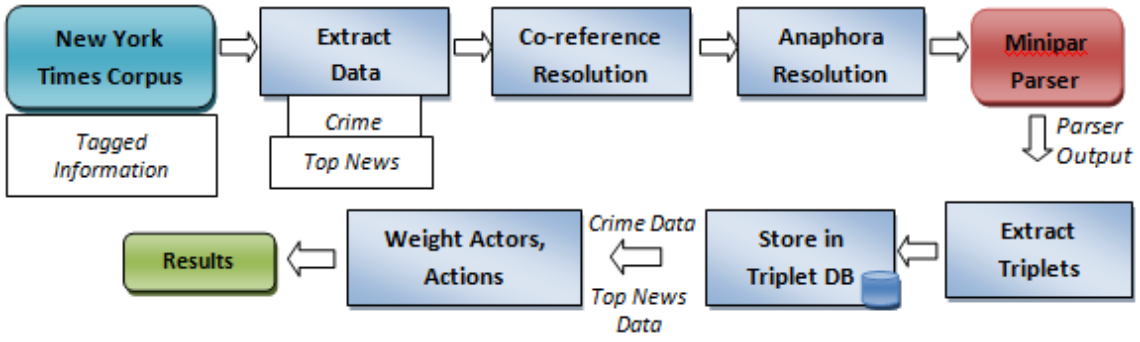

Figure 1: System Pipeline

surface terms (words within the document) that refer to the same entity (Leskovec et al. (2004)). In linguistics, anaphora is an instance of an expression referring to another. We considered the pronouns I, he, she, him, her and it for anaphora resolution. The ANNIE plugin in GATE (General Architecture for Text Engineering) distribution was used to perform co-reference resolution and a jape grammar was used to identify the named entities associated with the pronouns.

Our system uses a simple parser Minipar to extract triplets from a huge set of articles spanning over 20 years. An evaluation with the Susanne corpus shows that Minipar is able to achieve about 89% precision and 79% recall (Lin (1998)). In order to find the subject element of the triplet, we looked for nodes in the parser output having the grammatical relation s. The actions were found by identifying the nodes with grammatical relation i and the objects were the nodes having grammatical relation obj. An extracted triplet would look like,

Police (Subject)-Arrest (Verb)-Victim (Object)

Triplets were stored in the database for each year with the frequency of the subject, object and action in the data. Our application identified the key actors and actions in Crime for every year from 1987 -2007 by weighting the frequencies of each actor in crime against its frequencies in a background topic Top News (280,000 articles) in the New York Times corpus. The weights were calculated as follows,

$$Weight\ w_{Actor/Action} = \frac{f_{Actor/Action}\ Crime}{f_{Actor/Action}\ Top\ News\ World} \tag{1}$$

We define as key actors of a topic those actors whose weight exceeds a given threshold. In this way we could eliminate actors who more often appear in top news which means they are more generic. Our application can use any background topic to compute relative weights for actors/actions.

3. Results

3.1. Key Actors and Actions

Figure 2 shows the key subjects, objects and actions in Crime data in 2002 ranked according to their weights. When examined carefully we see that the application not only picks up the key players and actions described in crime data but also exposes a critical crime story that occurred during that year. Sexual abuse scandal in Boston archdiocese was a major chapter in the crime news in early 2002. Actors like Diocese, Detectives, Archdiocese, Cardinals, Bishops and actions such as Molest, Plead and Abuse reveal that.

4. Analysis of Triplet Networks

Network analysis and its directed graphs provide an ideal tool to map the network of social actors involved in an event and their reciprocal roles, with a homologous relationship between story grammars and network nodes and relations (Franzosi (2010)). In order to create networks we filtered only the triplets that contained the top 300 key subjects, objects and actions for each year in the following cases.

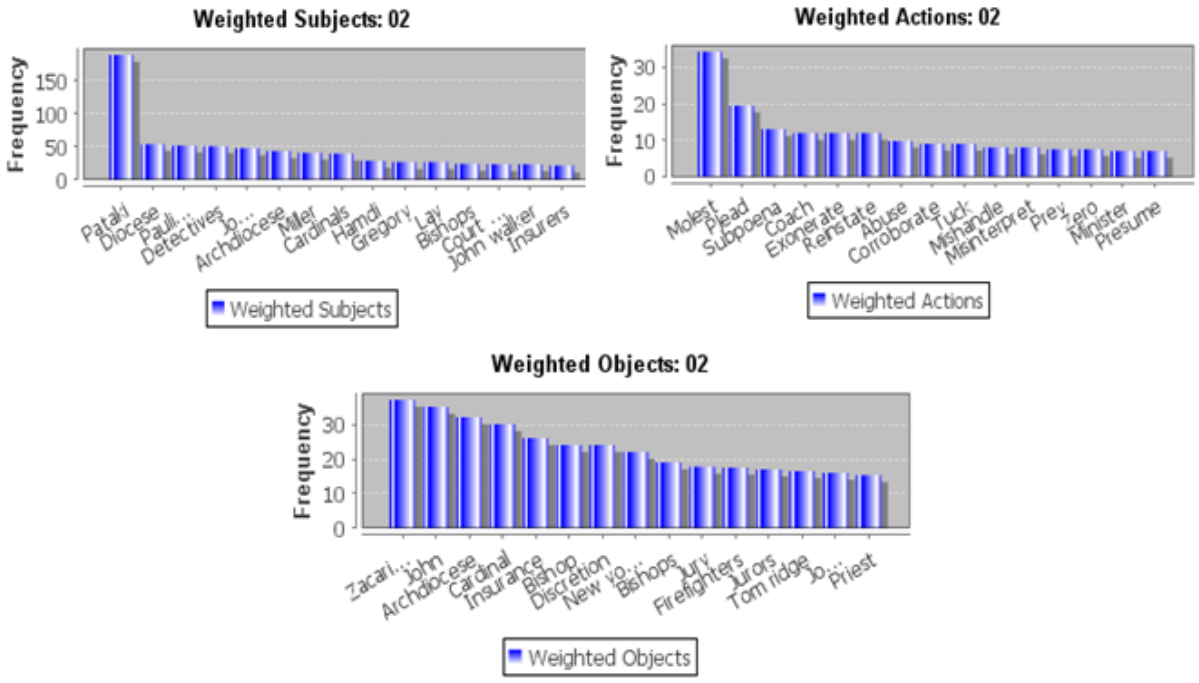

Figure 2: Key Actors and Actions in Crime Data in 2002

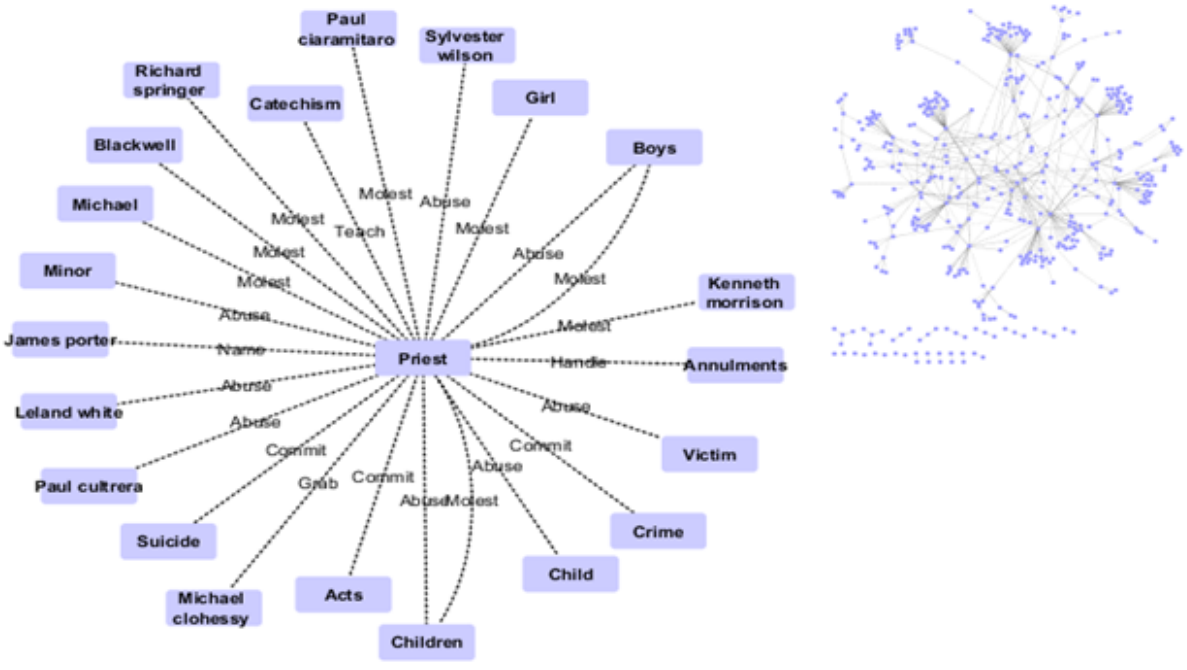

Figure 3: Triplet Network for Crime highlighting the Subject "Priest"

Case 1: Key Subject - Key Action - Object
Case 2: Subject - Key Action - Key Object

In this approach the networks were compact and also illustrated interesting information. We also tried creating networks out of the most frequent triplets obtained for each year but we found that they were less interesting compared to the triplets which contained weighted actors and actions. The major reason for this was that the most frequent triplets did not pick up the interesting actions. We generated directed networks using Cytoscape which is a general platform for complex network analysis and visualization. The networks had subjects and objects as nodes and the verbs linking them as edges. Figure 3 illustrates the triplet network for year 2002 on the right and highlights the interactions particularly between the subject "Priest" and other objects in the network.

4.1. Measures of Importance

In QNA it is essential to identify the central actors in the news narrative. In order to identify them we ranked all actors according to various network centrality measures like betweenness centrality, indegree and outdegree. Link analysis algorithms like HITS (Hyperlink-Induced Topic Search also known as hubs and authorities) and page rank were also used. Cytoscape and the Gephi tool-kit were used for performing the computations.

Table 1 shows the top 10 ranked actors for each network measure computed in Crime data for 2002. It shows actors like Priest, Archdiocese, Prosecutors, Schwartz and Zacari and noun phrases such as Law and Cases have been most central in the data, reflecting the leading crime story of that year in the US.

BC	**IND**	**OUTD**	**Hub**	**Authority**	**PageRank**
Law	Cases	Priest	Law	Cases	Cases
Archdiocese	Case	Judge	Archdiocese	Case	Court
Complaint	Letter	Law	Priests	Letter	Lawsuit
Suit	Allegations	Prosecutors	Suit	Questions	Anyone
Jurors	Boys	Jury	Abuse	Allegations	Nothing
Prosecutors	One	Lawyers	Firm	One	One
Diocese	Questions	Priests	Bishop	Law	Properties
Priests	Accusations	Archdiocese	Scandal	Suit	Play
Lawyers	Children	Church	Complaint	Nothing	Sorts
City	Law	Department	Diocese	Boys	Dying

Table 1: Top 10 ranked actors according to Network Centrality measures for Crime data in 2002

4.2. Classification of Actions

In QNA it is also common to investigate separately different spheres of interaction between actors (eg: communication, aggression, etc). We considered analysing the roles different actors play in crime by classifying actions into different types such as Crime against Person and Crime against Property. We acquired all English verbs from VerbNet and tagged them with these types using multiple sources. For each type we filtered triplets containing actions related to the type and created networks to analyse their properties. The top 10 ranked subjects and objects involved in crime in 2002 against person and against property are shown below.

Its interesting to see that Men are most commonly responsible for crimes against the person, while Women and Children are most often victims of those crimes.

Crime against Person		**Crime against Property**	
Subject	Objects	Subjects	Objects
Priest	People	Man	Money
Man	Boy	Police	Bank
Troops	Child	Soldiers	Records
Reyes	Girl	Winona Ryder	Millions
Geoghan	Man	Priest	Weapons
Shanley	Woman	People	Wallet
Forces	Jogger	Jason Bogle	Trade Secret
Police	Victim	Investigators	Steven Seagal
United States	Minors	Employee	Most
Others	Me	Agents	Man

Table 2: Top 10 ranked subjects and objects in crime against person and against property in 2002

4.3. Timeseries Analysis

To detect changes of roles of actors in crime over the 20 years we performed a time series analysis for each key actor. We discovered that network measures like outdegree and hub picked up the most central and interesting actors out of the data. Hence we used them and the frequency count of each actor to perform the time series analysis.

Figure 4 shows the time series graphs for “Priest plotted against its frequency, outdegree and hub values and actions “molest, “plead and “abuse plotted against their frequencies in 20 years. It clearly demonstrates that there has been a peak in all these measures during 2002 during when the news stated a lot about the involvement of the priest and archdiocese in the Boston sexual scandal.

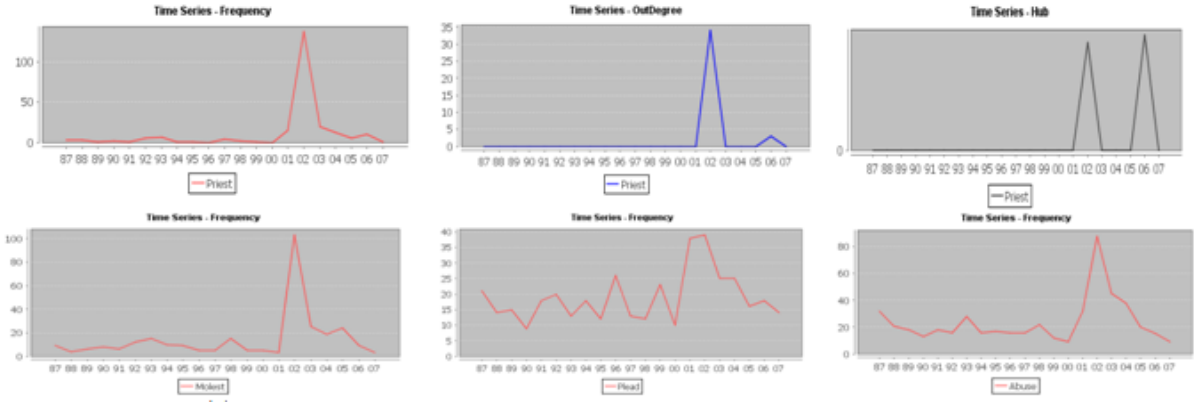

Figure 4: Time series graphs for actor "Priest and actions "molest, "plead and "abuse

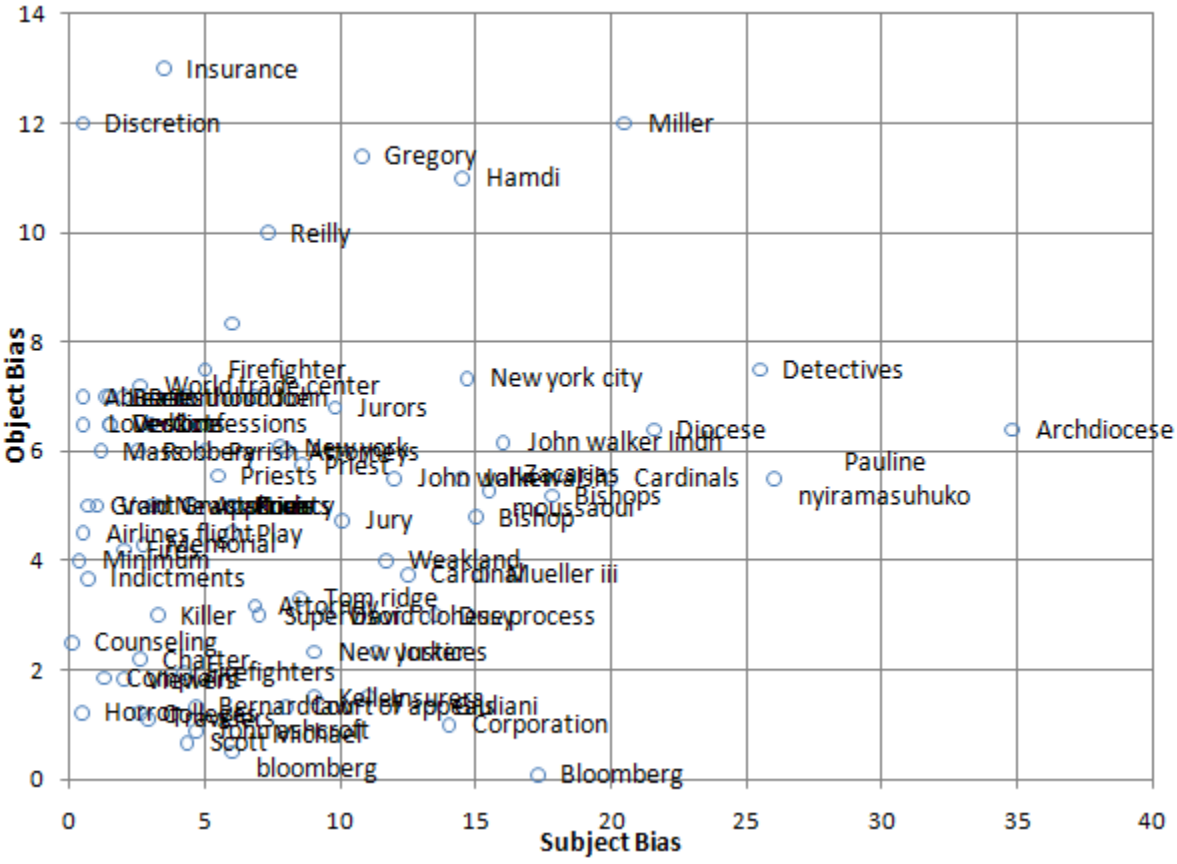

Figure 5: Scatter Plot showing the subject, object bias in data for year 2002. For ease of visualisation we removed NY Governor Pataki from set, as it had a very high subject bias.

5. Subject/Object Bias of Actors

The Subject/Object bias of an actor reveals the role it plays in the news narrative: that is its tendency to be portrayed as an active or passive element in the story. In order to find this bias we performed a study on the key actors that were found each year.

$$Subject\ Bias\ of\ Actor(K) = \frac{f_K\ (Subject, Crime)}{f_K\ (Subject, Top\ News) + f_K\ (Object, Top\ News)} \quad (2)$$

For each key actor K the subject bias was computed through dividing its subject frequency in crime by its subject frequency in top news plus its object frequency in top news. By doing this we could find out how subjective an actor is in Crime with respect to a more general topic like Top News. The same was done for computing the object bias. Figure 5 below illustrates the subject/object bias of actors in crime for year 2002 in a 2-dimensional scatter plot. The plot shows that Detectives, Dioceses, Archdiocese and Cardinals tend to be more often subjects of actions, while Discretion, Abuse and Priests being more often objects.

6. Conclusion and Future Work

We have demonstrated a scalable system for the automated narrative analysis of news corpora, a task traditionally labour intensive because performed by hand. This builds on various recent contributions from the field of Pattern Analysis, such as [Trampus and Mladenic (2011)], augmenting them with multiple analysis tools that respond to the needs of social sciences investigations.

Possible sources of error in the system could be co-reference or pronoun resolution; or other steps related to parsing. While allowing us to scale up to very large sizes, Minipar has its own limitations since it cannot parse sentences more than 1024 characters long. On the other hand, we observe that this length exceeds the size of typical sentences in the news. Future work will involve both a validation of the performance, and a deployment of the system to even larger analysis tasks. A validation of the quality of the story-triplets extracted can only be possible by comparing with human generated ones, and we intend this to be our next step. The system we are developing can directly feed into existing tools, such as PC-ACE (Program for Computer-Assisted Coding of Events, Franzosi (2010)) that are used for the analysis of news narratives, and for this reason we believe it can contribute towards the development of large-scale data-driven social sciences.

Acknowledgments

Saatviga Sudhahar is supported by the University of Bristol Postgraduate Research Scholarship Grant. Nello Cristianini is supported by a Royal Society Wolfson Merit Award, and the EU Project Complacs. Members of the Intelligent Systems Laboratory are supported by the 'Pascal2' Network of Excellence.

References

Lorand Dali and Blaz Fortuna. Triplet extraction from sentences using svm. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD), 2008*, Las Vegas, USA, 2008.

Earl, Jennifer, Andrew Martin, John D. McCarthy, and Sarah A. Soule. The use of newspaper data in the study of collective action. *Annual Review of Sociology*, 30:65–80, 2004.

I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, T.D. Bie, and N. Cristianini. Noam: News outlets analysis and monitoring system. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, Athens, Greece, 2011.

Roberto Franzosi. The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers. *Historical Methods*, 20:5–16, 1987.

Roberto Franzosi. Narrative as data. linguistic and statisitcal tools for the quantitative study of historical events. *New methods in Historical Sociology/Social History. Special issue of International Review of Social History*, 43:81–104, 1998.

Roberto Franzosi. Quantitative narrative analysis. *Sage Publications Inc, Quantitative Applications in the Social Sciences*, 162:200, 2010.

J. Leskovec, M. Grobelnik, and N. Milic-Frayling. Learning sub-structures of document semantic graphs for document summarization. In *Proc. of the 7th International Multi-Conference Information Society IS 2004*, pages 18–25, Ljubljana, Slovenia, 2004.

Dekang Lin. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*, Granada, Spain, 1998.

D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic. Triplet extraction from sentences. In *Proceedings of the 10th International Multiconference Information Society - IS 2007*, pages 218–222, Ljubljana, Slovenia, 2007.

D. Rusu, B. Fortuna, M. Grobelnik, and D Mladenic. Semantic graphs derived from triplets with application in document summarization. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD), 2008*, Las Vegas, USA, 2008.

Mitja Trampus and Dunja Mladenic. Learning event patterns from text. *Informatica*, 35, 2011.