Bringing PC-ACE to perfection

Roberto Franzosi, Department of Sociology and Linguistics Program,
Emory University, Atlanta, USA rfranzo@emory.edu

Fabio Cunial, Helsinki Institute for Information Technology (HIIT),
Department of Computer Science, University of Helsinki, cunial@cs.helsinki.fi

PC-ACE (*Program for Computer-Assisted Coding of Events*) is a research software developed to carry out a particular type of textual analysis used in sociology: *Quantitative Narrative Analysis* (QNA, see Franzosi 2010, 2012; Franzosi, De Fazio and Vicari, 2012). While there are several, *commercial* software programs available to perform textual analysis for the social scientists (e.g., Atlas.ti, NVivo, MaxQda), none give users the analytical power and control that PC-ACE does (see Franzosi, Doyle, McClelland, Putnam Rankin, and Vicari, 2012). This project is about bringing the current version of PC-ACE to perfection, so that it can be widely distributed to social scientists.

Quantitative Narrative Analysis in a nutshell

Quantitative Narrative Analysis (QNA) is a social science methodology that Franzosi has developed for the collection and analysis of narrative texts on the basis of a computer-assisted story grammar, where: (1) narrative is a type of text genre that tells a story, where someone does something, pro or against someone else; (2) a story grammar captures the basic elements of a story, the Who (and their attributes, such as name, gender, race, age, etc.) and the What (and their circumstances, such as When and Where, Why and How); basically, a story grammar is nothing but the 5 Ws + H of journalism expressed in terms of formal rewrite rules; (3) the relational nature of a story grammar makes it possible to implement a story grammar in a relational database management system: PC-ACE is such a system, designed for the collection, organization, and storage of *large bodies of narrative data* in a computer.

Perhaps the best way to understand the power of QNA/PC-ACE (note that no PC-ACE, no QNA!) is looking at the website of Franzosi's project on Georgia lynchings (1875-1930) (http://dev.emorydisc.org/galyn: all functionality is there but not yet the web design). The website was developed as a Digital Humanities Scholarship project at Emory University (Atlanta, USA), funded by the Mellon Foundation, and it is entirely based on data collected and extracted in PC-ACE: over 1,300 original newspaper articles and over 7,000 *semantic triplets*, i.e. basic narrative sentences of 5 Ws+H organized sequentially in story form. No current textual analysis approach can produce such data, and automatic data mining algorithms are not there yet to make possible the realiable automatic extraction of narrative information from texts (see, Sudhahar et al., 2012). Franzosi is now working on a similar Digital Humanities project on the

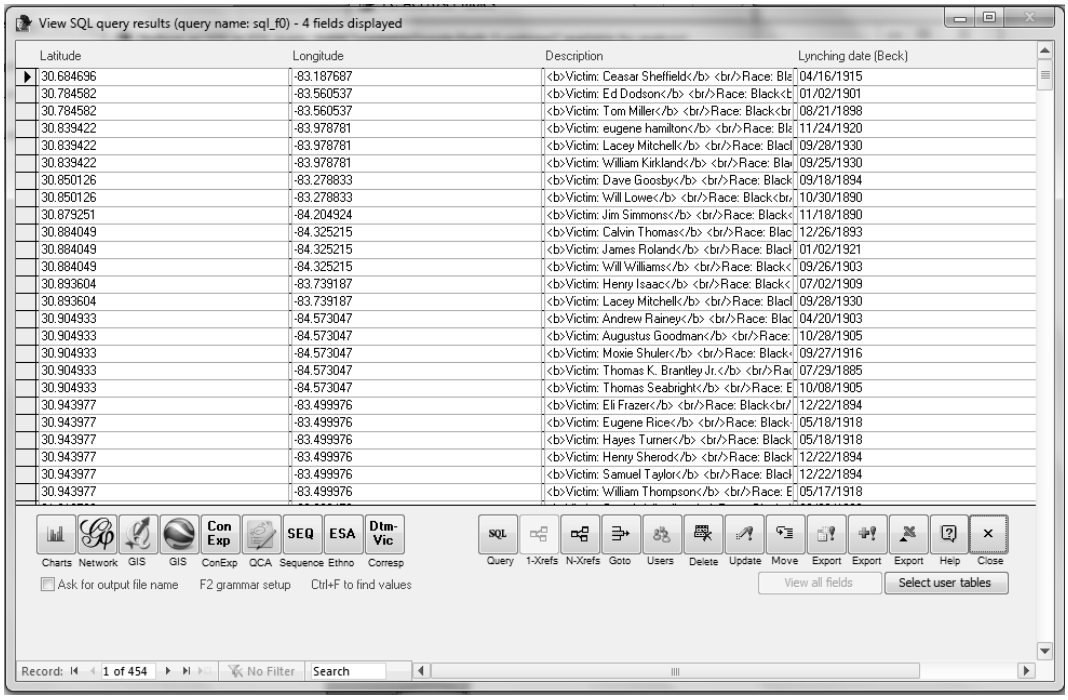
rise of Italian fascism (1919-1922), based on some 50,000 newspaper articles, yielding some 250,000 semantic triplets and leading to fascinating network graphs and GIS maps of the social relations and geography of violence.

Impact of the project

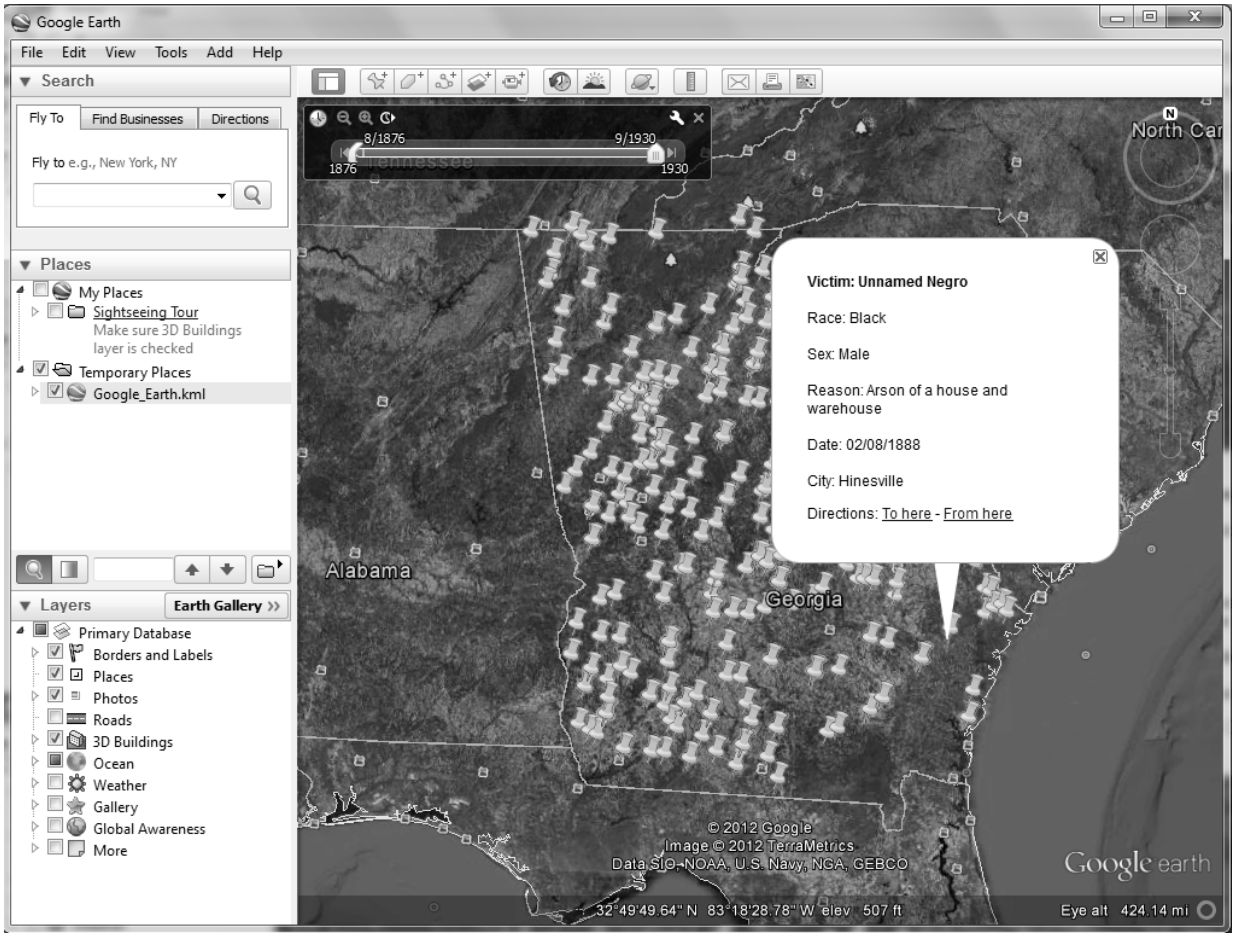
Besides Franzosi, PC-ACE has been used by Robert Biggert (Assumption College), Gianluca De Fazio (Emory University), Sophie Doyle (Oxford University), Andrew Junker (University of Chicago), René Karpantschof (University of Copenhagen), Marco Calaresu (European University Institute), Stefania Vicari (University of Leicester). Other users could be using PC-ACE without explicitly contacting the developers: some two hundred users are registered on the PC-ACE website. The potential user base for PC-ACE, however, is very large, for several reasons:

1. narrative provides a huge class of texts (Franzosi, 1998);
2. PC-ACE is distributed as freeware (open source with the next release): all competitive programs are commercial (e.g., Atlas.ti, NVivo, MaxQda);
3. no other competitive program allows to carry out Quantitative Narrative Analysis (Franzosi, Doyle, McClelland, Putnam Rankin, and Vicari, 2012).

Performing advanced textual analysis in general, and QNA in particular, requires an unusual set of skills: from linguistics to computer science, from sociology or political science to GIS tools, networks, statistics. Unfortunately, those social scientists who know words do not know numbers, and those who know numbers do not know words. PC-ACE is the only textual analysis tool that interfaces with well-known data visualization and data analysis software. The View Query form provides a unique broad platform of generalized data handling of words (see Figure).



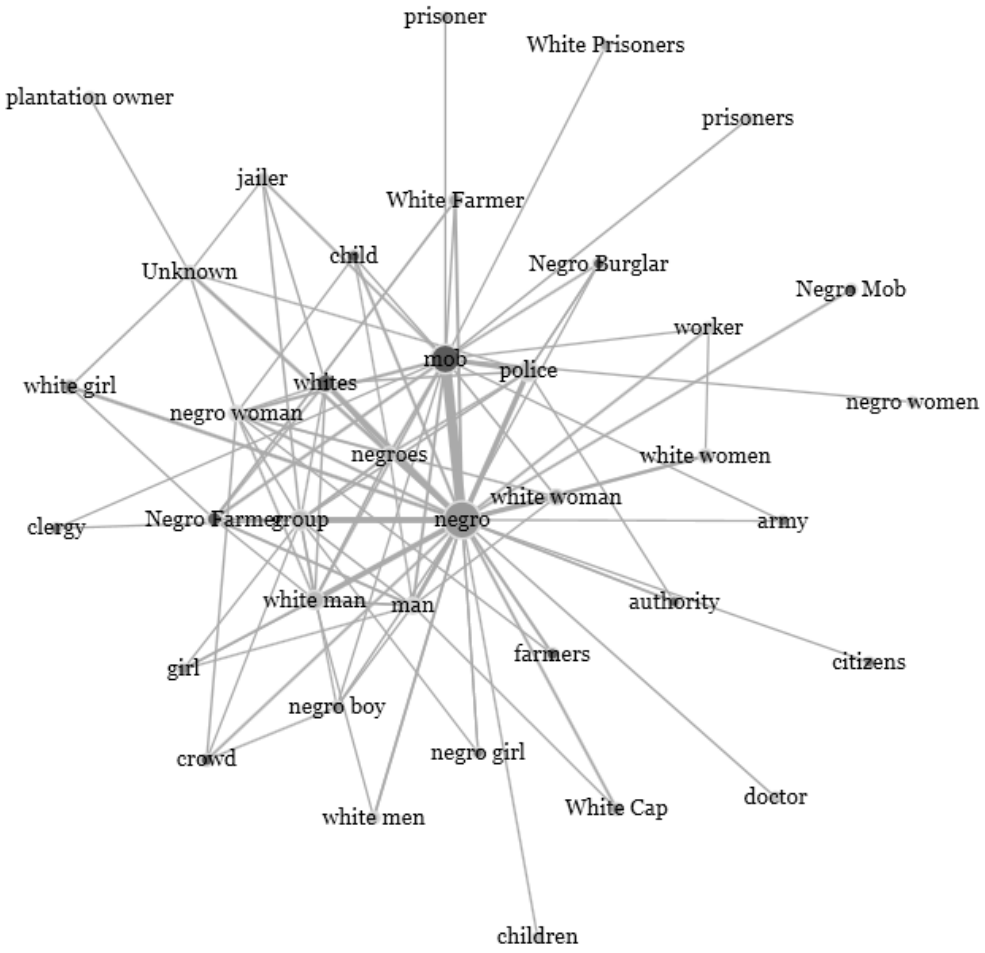
The View Query form makes available a number of buttons that allow users to run automatic data visualization and data analysis routines: from Excel charts to network graphs in Gephi or Ucinet, from GIS maps in QGIS or Google Earth to sequence analysis, qualitative comparative analysis, and correspondence analysis. Thus, clicking on the Google Earth button with the SQL results displayed above we would obtain automatically the following map of Georgia lynchings. Similarly, if the View Query form displayed information on vertices and edges for relations of violence, clicking on the network button would display the corresponding network graph (see Figures).



Objectives

Few annoying bugs and a fundamental lack of easy-to-use data handling tools have so far stood in the way of a wider diffusion of PC-ACE. This project aims at fixing all standing bugs and at giving users more flexible tools for checking and managing their data.

1. **Setup module** - The module for the setup of the story grammar allows PC-ACE users to handle only one type of document (e.g. a newspaper article) and crossreference this document to the objects setup for the story grammar of data collection. If users wanted to

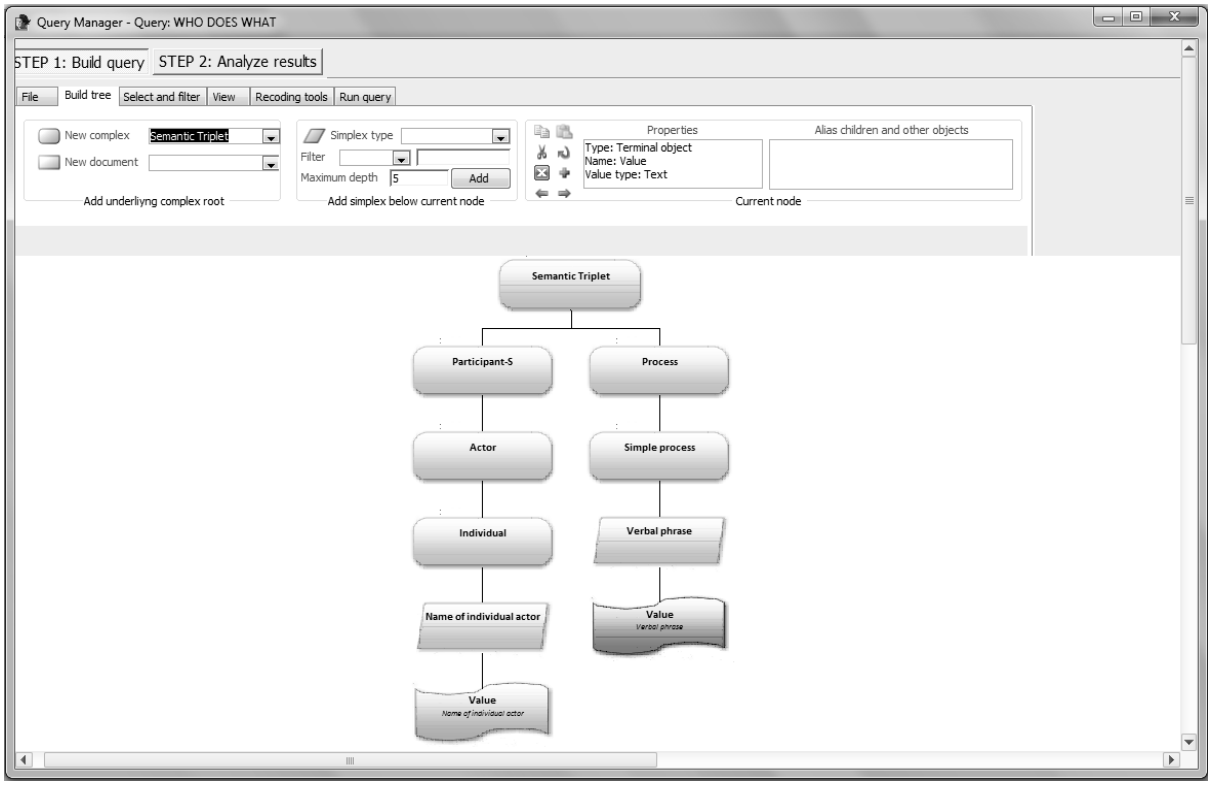


collect data from *different* types of sources (e.g., newspaper articles *and* archival documents *and* books), this would not be possible in the current design.

2. **Data entry module -** Many users perform text coding in Excel, in a rectangular matrix for what should be a relational problem. They would welcome the opportunity to work in PC-ACE and take advantage of all the built-in functions of data checking, querying, visualization, and analysis. PC-ACE does have many functions of importing Excel data but not a generalized import function that would allow unfamiliar users to have PC-ACE automatically build the grammar as it imports the data. Many users feel that learning the language of a story grammar is a daunting task. We need a generalized routine for importing from an Excel spreadsheet in a way that the story grammar corresponding to the Excel data is automatically set up by the import routine. That would require parameterizing several routines of data import from Excel files already available in PC-ACE. We also need to extend on-line data validity checks in the data entry forms. Data entry in PC-ACE is typically done using the same words found in the original documents. Verbs such as “kill”, “wound”, or “kick” are coded as such under an object “verb phrase”.

That, in a typical project, results in hundreds of different verbs. For the purpose of data analysis, these disaggregated items need to be aggregated. Routines need to be introduced that would allow users to code aggregate items bringing up automatically a distribution of previous values.

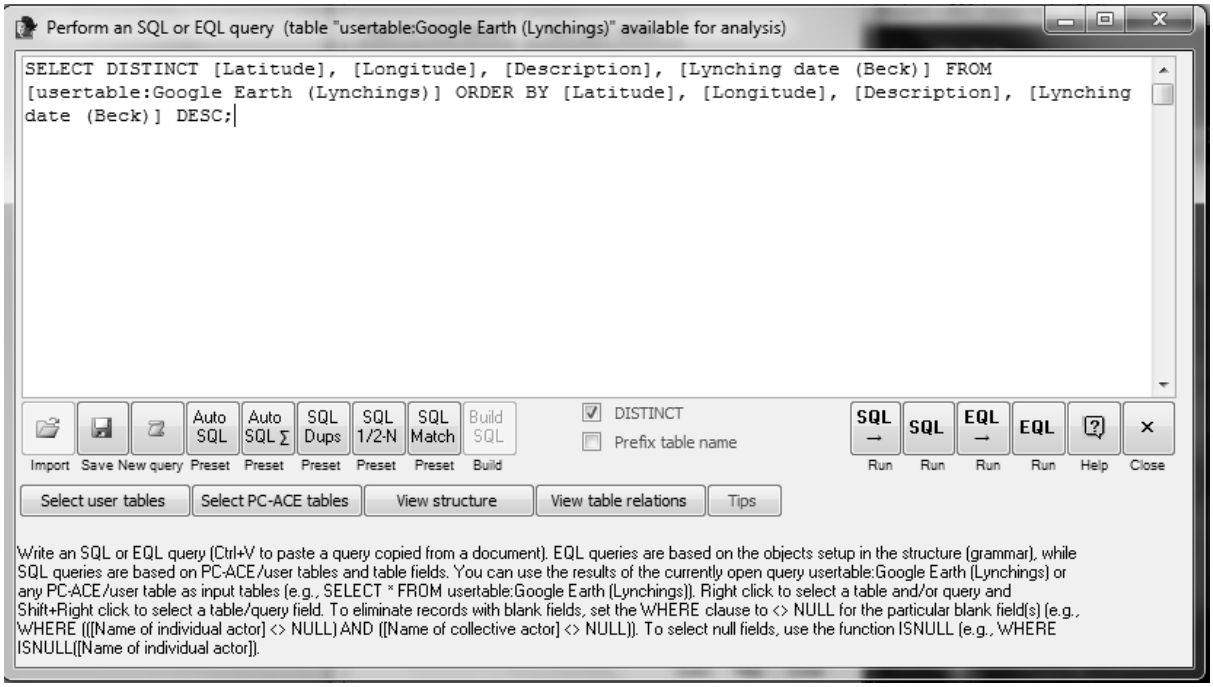
3. **Query module -** The GUI-based Query Manager (QM) first introduced in PC-ACE in 2009 provides a powerful tool for users to carry out SQL queries on the basis of objects



they setup in the grammar without any knowledge of the underlying table structure, unknown to the end users (see Figure).

Unfortunately, both the QM and the underlying query language upon which it relies to translate queries based on user-defined grammar objects to SQL tables, have limitations which make the tools less than ideal. For instance, the QM only allows AND conditions for specific sub-branches of a tree; we need to introduce OR conditions. We also need to extend the GUI to include usertable and not just PC-ACE tables – handling of both usertable and PC-ACE tables requires an SQL query that the user must write, another daunting task for the average social scientist, although the SQL/EQL form offers many tools for the automatic creation of SQL queries (see Figure). Finally, the EQL language behind the Query Manager has some bugs, leading to faulty query results: these needs to be found and fixed.

4. **Data cleaning module -** Users need more flexible tools of data manipulation in the process of data cleaning. This would require generalized routines for the copying, cutting,



pasting, moving, deleting objects as revealed by a query. Several routines of this kind are available but, developed at different times and for different purposes, they vary in programming standards. What is needed is to parameterize all these routines eliminating redundancies. Furthermore, there is a need for a generalized routine for checking documents/objects crossreferences and for fixing faulty documents/objects crossreferences, in case errors are introduced (by PC-ACE crashes, for instance).

Advantages for the student

1. Impact: develop good code that other people will use, and which will be part of a larger project. This could go on the student's CV.
2. Receive a frameable "Certificate of Participation" signed by the Chair of the Sociology department at Emory University and by Dr. Franzosi. Emory University is one of America's top 20 universities.
3. Talented students will have the possibility of continuing the collaboration with the development of PC-ACE past the work required by this project.
4. Not just software development: have fun hacking with existing code, experimenting with it and improving it.

Tentative work organization

Regular weekly Skype meetings with Dr Franzosi will be the main form of task organization. Dr Franzosi knows the program extremely well. Fabio Cunial has also collaborated to the development of PC-ACE since 2006: if necessary, students working on the project can rely locally on his expertise and advice. Most tasks in this project are highly modular: for example,

the work on the query module and on the EQL language behind it is completely decoupled from the work on other modules. However, making some key changes to PC-ACE requires all students to work together, due to nontrivial long-range dependencies in legacy code.

Some technical facts

PC-ACE is developed on top of a relational database with 46 tables, 136 GUI forms, and nearly 100,000 lines of procedural and declarative code. The main architecture dates back to 2003. For historical reasons, the software has been developed on top of Microsoft Access 2010 using a peculiar mix of VBA and SQL (maybe it's the most complex software ever developed on Access :-). It is currently housed at Emory University and available for free download (www.pc-ace.com). We are planning to release it as open source soon.

References

Franzosi, Roberto. 2012. "On Quantitative Narrative Analysis." In: pp. 75-98, James A. Holstein and Jaber F. Gubrium (eds.), *Varieties of Narrative Analysis*. Thousand Oaks, CA: Sage.

Franzosi, Roberto, Gianluca De Fazio and Stefania Vicari. 2012. "Ways of Measuring Agency and Action: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875-1930)." In: pp. 1-41, Tim Liao (ed.), *Sociological Methodology*, Vol. 42, No. 1. DOI 10.1177/0081175012462370

Franzosi, Roberto, Sophie Doyle, Laura McClelland, Caddie Putnam Rankin, Stefania Vicari. 2012. "Quantitative Narrative Analysis. Software Options Compared: PC-ACE and CAQDAS (ATLAS.ti, MaxQda, and NVivo)." *Quality & Quantity*. DOI 10.1007/s11135-012-9714-3.

Franzosi, Roberto. 2010. *Quantitative Narrative Analysis* (Quantitative Applications in the Social Sciences). Beverly Hills, CA: Sage.

Franzosi, Roberto. 1998. "Narrative Analysis – Why (And How) Sociologists Should Be Interested in Narrative." In: pp. 517–54, John Hagan (ed.), *The Annual Review of Sociology*, Palo Alto: Annual Reviews.

Sudhahar, Saatviga. 2012. "Automating Quantitative Narrative Analysis of News Data." Joint author with Roberto Franzosi, Gianluca De Fazio, and Nello Cristianini. Under review at *Natural Language Engineering* 10/28/2012.