

Narrative as Data: Linguistic and Statistical Tools for the Quantitative Study of Historical Events*

ROBERTO FRANZOSI

This paper illustrates some linguistic and statistical tools that can be profitably used by historians and social historians in the study of events (such as strikes, demonstrations and other types of collective conflict). More specifically, the paper shows that “semantic grammars” provide rigorous tools for the collection of rich event narratives. Semantic grammars structure information around the “canonical form” of the language: noun phrase/verb phrase, or subject, action, object and their modifiers (e.g. time, space). The fact that semantic grammars can be easily implemented in a computer environment using relational database systems (RDBMS) makes feasible the practical application of such powerful coding schemes. The data that computer-based semantic grammars make available are richer, more flexible and more reliable than those delivered by more traditional content analysis methods. They are also very well suited for the application of new tools of data analysis such as network models. Both semantic grammars and network models are fundamentally concerned with actors and their actions, with agents and agency. As such, these linguistic and statistical tools should draw sociology closer to history, traditionally much more concerned with issues of agency. I illustrate the power of both the linguistic and statistical tools using data that I collected from some 15,000 newspaper articles on the 1919–1922 period of Italian history, a period characterized by widespread working-class mobilization (1919–1920, the “red years”) and fascist counter-mobilization (1921–1922, the “black years”).

INTRODUCTION

Historians have relied on narrative as both data and mode of analysis and communication. Few historians, however, have analyzed the narratives that form the basis of their empirical evidence (e.g. archival documents, police records, newspaper articles) in systematic ways through the application of content analysis techniques. Yet the use of content analysis in socio-

* Data collection for this project was supported, in part, by the National Science Foundation (grant SES-8511632). Data analysis and writing have benefited from another National Science Foundation grant (SBR-9411739). I am indebted to William Bainbridge and Andrea Kline for their help in transferring the NSF grant to the University of Oxford. I would also like to thank Larry Griffin for his comments on a previous draft and Mauro Giorgetti and PierPaolo Mudu for their help with data analysis.

historical research has proved fruitful.¹ Partly, historians' reliance on qualitative methods, but also the limitations of content analysis have made historians and social historians reluctant to "go from words to numbers" using narrative as data for quantitative analysis.

In this paper, I illustrate a linguistic approach to collecting text data that avoids the main shortcomings of traditional content analysis. The approach is based on intrinsic properties of narrative text as captured by a semantic grammar (or text grammar, or story grammar). A semantic grammar is *nothing but the simple semantic structure: Subject, Action, Object (SAO) and respective modifiers, such as time, space, etc.*² This approach to the systematic use of narrative texts should be particularly appealing to historians as:

- (a) it is best applicable to the study of historical events;
- (b) it is squarely centered on agents and their actions;
- (c) it preserves both the information and the narrative flow of the original text;
- (d) it does not rely on theoretically defined coding categories;
- (e) it yields far richer and more flexible historical records than traditional content analysis methods;
- (f) it can be easily implemented in a computer environment through commercially available relational database systems;
- (g) it allows researchers to go from words to numbers, i.e. to analyze qualitative narrative data in quantitative ways; in particular, the organizational structure of the narrative data in relational format centered on agents and actions mirrors the requirements of such statistical techniques as network models.

I will illustrate the power of the technique with reference to data that I collected, on the basis of such a grammar, from some 15,000 newspaper articles on protest events during the 1919–1922 period in Italy. I will show that the empirical evidence that the grammar produces could not be obtained in other ways.

A LINGUISTIC APPROACH TO CONTENT ANALYSIS: SEMANTIC GRAMMARS

In recent years, social scientists have recommended the use of linguistics to overcome methodological (e.g. reliability) and theoretical (e.g. lack of con-

1. See for a review, Roberto Franzosi, "The Press as a Source of Socio-Historical Data. Issues in the Methodology of Data Collection from Newspapers", *Historical Methods*, 20 (1987), pp. 5–16.

2. Roberto Franzosi, "From Words to Numbers: A Generalized and Linguistics-Based Coding Procedure for Collecting Event-Data from Newspapers" in Clifford Clogg (ed.), *Sociological Methodology*, vol. 19 (Oxford, 1989), pp. 263–298; idem, "From Words to Numbers: A Set Theory Framework for the Collection, Organization, and Analysis of Narrative Data", in Peter Marsden

sistent foundations) pitfalls inherent in traditional content analysis.³ In particular, semantic grammars have been shown to yield richer and more flexible data records than traditional coding schemes.⁴ Semantic grammars consist of a set of functionally defined (coding) categories that are compatible with a large variety of input text.⁵ Semantic grammars can capture the deep meaning of a unit of text within a restricted set of linguistic categories. The following provides an example of a simplified semantic grammar specifically developed for the study of protest events:

```

<semantic triplet> → {<subject>} {<action>} [{<object>}]
<subject>          → {<actor>} [{<actor modifiers>}]
<actor>           → crowd, mob, students, peasants, . . .
<action>          → {<action phrase>} [{<action modifiers>}]
<action phrase>   → riot, assemble, protest, rally, . . .
<object>          → {<actor>} {<inanimate object>} {<implicit targets>}
                  {<infinitive clauses>}
<inanimate object> → building, road, car, bomb, . . .
<implicit target> → <actor>
<infinitive clauses> → <semantic triplet>

```

where: the symbol → refers to a rewrite rule (or production), i.e. it “rewrites” the element to its left in terms of the element(s) to its right; the symbol <> indicates that the element enclosed within it is a non-terminal entity, i.e. it can be further rewritten in terms of other elements; terminal entities that cannot be further rewritten appear in the grammar without the angular brackets; the symbol {} indicates that the element enclosed within it can be repeated several times; the symbol [] indicates that the element enclosed within it is optional.

According to the grammar outlined above, the semantic triplet is made up of (rewritten as) the set subject/action/object. This set provides the basic template for structuring text using a semantic grammar. It corresponds to the basic linguistic structure noun phrase/verb phrase, otherwise called the “canonical form” of the language. A <subject> is made up of one or more <actor>, such as crowd, mob, etc., with possible <actor modifiers>. The type and number of modifiers varies with the type of substantive application. In collective action research, typical <actor modifiers> might be: the <proper

(ed.), *Sociological Methodology*, vol. 24 (Oxford, 1994), pp. 105–136; idem, *From Words to Numbers: Narrative as Data* (Cambridge, forthcoming).

3. E.g. John Markoff *et al.*, “Toward the Integration of Content Analysis and General Methodology”, in David R. Heise (ed.), *Sociological Methodology*, vol. 4 (Oxford, 1974); Franzosi, “From Words to Numbers: A Generalized and Linguistics-Based Coding Procedure”.

4. *Ibid.*

5. On semantic grammars as general content analysis schemes see *ibid.*; Roberto Franzosi, “Narrative Analysis, or . . . Why (and How) Sociologists Should Be Interested in Narrative”, in John Hagan (ed.), *Annual Review of Sociology*, vol. 24 (Palo Alto, 1998) and the bibliography cited there.

name of an individual>, the <number of actors>, the <type of actor> (e.g. male, female, armed, skilled), or the <organization> involved. Possible <action modifiers> could be: <negation> (e.g. do not), <type of action> (such as “general” or “wildcat” for a strike), <reason for action>, <outcome of action>, the <instrument of action> (e.g. “cannons” in military actions) or <time> and <space> of the action.

A semantic grammar not only provides a set of semantically defined categories of general applicability; it also specifies explicit relationships between categories (<actors> to <actions> and both to their respective <modifiers>). Furthermore, a grammar can specify different hierarchical levels of aggregation for the information. For example, we could aggregate semantic triplets into events and events into collective campaigns. Thus, a semantic grammar can organize text in both hierarchical and relational form.

To understand how a linguistic approach to content analysis works let’s take one of the events described by French historian Roland Mousnier in his *Peasant Uprisings in Seventeenth-Century France, Russia, and China*⁶ and code it within the categories of a semantic grammar.

At Bordeaux in May and June, 1635, a gabelle imposed on taverns and on wine set off a revolt of tavern keepers and makers of casks which drew in many craftsmen and day-laborers, with the complicity of the majority of the bourgeois, a large number of “persons of condition,” and officials of the parlement, who “regarded these rebels as their liberators.” The rebels “wanted to get control of Saint Julien’s Gate [. . .] and let in the peasants of the surrounding district, to help them. This was being loudly demanded by the peasants, who wanted to share in the plunder of the town, which they thought they had at their mercy.” This was at any rate the motive ascribed to them by the secretary to the duc d’Épernon, who wrote these lines; but among these peasants there were, along with gardeners and market gardeners, also vinegrowers and *laboueurs*, very seriously affected by the indirect taxes. A certain number of them managed to get into the town and take part in the revolt. The duc d’Épernon, governor of Guyenne, was unable to secure the help of the town militia or that of most of the neighboring gentry, nor was he able to raise soldiers in the usual way from his lands, among his own peasants, who hid themselves. He was reduced to a few gentlemen, a few “*honnêtes bourgeois*,” and, in the town, to a few of his former soldiers, with whose aid he either overcame the rebels or persuaded them to yield. Hardly had they been more or less quieted when “the madness extended to the peasants of the countryside. These, having, during one of the riots, managed to plunder some houses in the town, withdrew with the loot to their villages, and their neighbours were soon stirred up by their bad example to engage in plundering [. . .] In a moment, all the villagers dropped their ploughs and took up arms instead: after robbing the houses in the countryside they gathered in great numbers in all the suburbs and strove to enter the town itself. The poor of the town wanted them to enter and tried to open a way for them [. . .]” They assembled in the suburb of Saint-Surin, where they burned

6. Roland Mousnier, *Peasant Uprisings in Seventeenth-Century France, Russia, and China* (New York, 1972), pp. 45–46.

several houses. The duc d'Épernon mounted his horse and set forth at the head of forty or fifty gentlemen of his guard. This time he secured the support of a few companies of the town militia. The peasants scattered. The cavalry chased them and killed forty or fifty. "The rest of the peasants, who had been waiting to see whether these would succeed or not, behaved themselves properly that year [. . .]" The king then granted an amnesty.

Appendix A gives Mousnier's text coded within the semantic categories of the grammar, with events and semantic triplets chronologically ordered. As one can see, a semantic grammar provides a tight organizational structure for much of the information available in the original narrative. Only one question: why bother? What is the point of organizing text within the framework of a semantic grammar? What do we gain from that effort? After all, as far as we know, Mousnier did not rely on semantic grammars to catalogue and organize his narratives.

No doubt it is possible to tell masterful historical stories without story grammars and computers. But the more complex a story, the more difficult it is to keep the threads together. Consider the case where researchers are confronted not just with one event narrative, but with hundreds, perhaps thousands of such narratives. Clearly, the scale of those numbers make it imperative for researchers to find more systematic ways of cataloguing and retrieving the information contained in the narratives. It is out of that work that hopefully will come the evidence on patterns of historical behavior and of interactions among actors, on changes in routines and reasons. It is out of that work, in other words, that will come a shift of focus from the event to the conjuncture, from the single event to patterns of events.⁷

A grammar helps us in taking the first step in that direction (by sorting out and cataloguing information in appropriate ways), but only the first step. Because, surely, once we have all that information on thousands of events sorted out along the categories of a semantic grammar, the question remains: what are we going to do with it? To do something with "it" we need at least two things. First, we need to be able to store the information in the computer. Otherwise, we might as well keep the narrative in its original form without the paraphernalia of all those coding labels. Second, we need a way to relate coding categories to each other, for example actors to actions or both actors and actions to their characteristics, in order to trace meaningful patterns of interaction. Fortunately, the two things are closely interrelated. The logic or set theoretical underpinnings of a semantic grammar allow us to express in a formal mathematical notation the relationship between the coding categories of the grammar. Those same mathematical principles serve as the foundation of powerful organizational data models for computer storage and retrieval, namely relational database management systems (RDBMS).⁸

7. Fernand Braudel, *On History* (Chicago, 1980), pp. 25–54, 64–82.

8. Franzosi, "From Words to Numbers: A Set Theory Framework".

In summary, a linguistic approach to content analysis yields richer and more flexible text data than traditional coding schemes.⁹ First, coding under a semantic grammar preserves much of the original text input (if not syntactically, at least semantically). Second, all factual elements of discourse find an appropriate organizing label. Third, the various parts of discourse are interconnected in complex ways. Fourth, output coded under a semantic grammar has face validity: it must make sense to any competent reader of the language. This characteristic greatly increases data quality. Fifth, a semantic grammar provides a more natural way of organizing text than traditional coding schemes. Last but not least, a semantic approach to content analysis provides a coding scheme that is invariant across a large class of texts of particular interests to historians (namely, narrative texts). This scheme is deeply rooted in a body of linguistic theoretical knowledge, rather than the empirical *ad hoc* approach of much traditional content analysis. Finally, semantic grammars can be easily implemented in a computer environment via powerful RDBMS commercially available for both mainframe and PC platforms. It is the implementation of semantic grammars in a computer environment that makes their use practical, despite the complexity of their design. For all these reasons, whenever applicable, semantic grammars can achieve quality without sacrificing quantity.

FROM WORDS TO NUMBERS: THE LINCHPIN

No doubt having narrative information in a computer-based relational structure allows investigators to store and extract rich narrative information in general ways (e.g. asking which kind of actions peasants were more likely to engage in when they protested in seventeenth-century France). But the question remains: what are we going to do with that information? The answer to that question depends upon the ability to manipulate quantitatively the qualitative information available in a relational database.

I have shown¹⁰ that through counting of information in the different categories of the grammar one can easily construct variables and estimate statistical models of the type $y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$, despite the fact that the underlying data are basically words. The ability to manipulate information statistically gives researchers greater power to investigate relationships, to bring out patterns, to tease out structures that lie behind the surface of the myriad of historical facts. Historians, however (and an increasing number of social scientists), may not feel at ease with a

9. On these issues, see Roberto Franzosi, "Strategies for the Prevention, Detection and Correction of Measurement Error in Data Collected from Textual Sources", *Sociological Methods and Research*, 18 (1989–1990), pp. 442–471; idem, "Computer-Assisted Coding of Textual Data: An Application to Semantic Text Grammars", *Sociological Methods and Research*, 19 (1990–1991), pp. 224–256.

10. Franzosi, "From Words to Numbers: A Set Theory Framework"; see also idem, *From Words to Numbers*.

representation of social reality based on variables rather than social actors. Indeed, where are the actors gone in the representation $y = f(X_1, X_2, \dots, X_k)$? Where is agency? *These* are the fundamental concerns of historians. In Elton's words: "Despite attempts to deny this, it [quantitative history] can effectively operate only by suppressing the individual by reducing its subject matter to a collectivity of human data in which the facts of humanity have real difficulty in surviving."¹¹

There is another way to go "from words to numbers", however, that these historians and social scientists may find more appealing.¹² Again, a set theoretical framework provides the fine thread that goes from words to numbers. Consider the set of skeleton semantic triplets, SAO (without modifiers), that a project of data collection on historical events will yield. This set T of size T ($T = t_1, t_2, t_3, \dots, t_T$) will contain a set of subjects S of size S (with $S = s_1, s_2, s_3, \dots, s_S$), a set of actions A of size A (with $A = a_1, a_2, a_3, \dots, a_A$), and a set of objects O of size O (with $O = o_1, o_2, o_3, \dots, o_O$). While the set of subjects S is made up exclusively of social actors, the set of objects O contains both social actors and inanimate objects (e.g. <subject> peasants <action> plunder <object> houses; <subject> poor <action> burn <object> houses). Let us now select a subset O_1 from the set of objects O that will include only those objects that are social actors; then, let us join the set S of subjects and the subset O_1 of objects, to obtain the set of actors N of size N (with $N = n_1, n_2, n_3, \dots, n_N$). Some twenty-four actors appear in Mousnier's description of the events that took place in Bordeaux in 1635. Thus, N has size 24, with $N =$ authorities, tavern keepers, makers of casks, craftsmen, day-laborers, bourgeois, persons of condition, officials of the *parlement*, rebels, peasants, gardeners, market gardeners, vine-growers, *laboueurs*, governor, town militia, gentry, gentlemen, *honnêtes bourgeois*, soldiers, villagers, the poor, cavalry, king.¹³

Similarly, thirty-one distinct actions appear in Mousnier's account. The set of actions A is thus of size 31 with $A =$ are tacit accomplice, approve, assemble, burn, behave properly, chase, control, demand, enter, get into, gather, grant an amnesty, help, impose *gabelle*, join revolt, kill, let into, open a way, overcome, persuade, plunder, rebel, revolt, rob, scatter, strive, take part in revolt, take up arms, try, yield, want.

In the SAO representation, the action basically operates as a relation between actors – a relation which is both dichotomous and directional: any

11. G.R. Elton, "Two Kinds of History", in Robert William Fogel and G.R. Elton, *Which Road to the Past? Two Views of History* (New Haven, 1983), pp. 71–121, 118–119.

12. For an excellent introduction to network analysis, see Stanley Wasserman and Katherine Faust, *Social Network Analysis. Methods and Applications* (Cambridge, 1994), chs 3 and 4.

13. Many elements in N simply provide alternative descriptions of the same social actors. For example, the word "rebels" is first used in the narrative to encompass tavern keepers, makers of casks, craftsmen, day-laborers if not also bourgeois, persons of condition and officials of the *parlement*; similarly, villagers may include peasants, laborers, vinegrowers, etc.

member of the set of actors N , n_i (where $n_i \in N$) either relates to another member n_j or does not; furthermore, whether an actor occupies the role of subject or object makes a difference (consider the triplets: “cavalry kills peasants” versus “peasants kill cavalry”). Not all social actors present in the database will relate to all other actors. If a tie is present between n_i and n_j , then we say that the ordered pair $\langle n_i, n_j \rangle$ belongs to a special collection of pairs L . L , in other words, represents the set of triplets that contain a (human) object. As we have seen, not all triplets do (e.g. $\langle \text{subject} \rangle$ peasants $\langle \text{action} \rangle$ behave themselves properly).¹⁴ If an ordered pair $\langle n_i, n_j \rangle$ is in L , then the first actor in the pair relates to the second on the relation under consideration; we can write $n_i n_j$.¹⁵ Each relation has a corresponding set of arcs, L_r , containing L_r ordered pairs of actors as elements (the subscript r ranges from 1 to R , the total number of aggregated relations).

The trouble is that a large-scale data collection project is likely to yield hundreds, perhaps thousands, of different actions/relations, too many for easy manipulation. One could reduce the number of distinct actions, (1) by eliminating synonymous expressions (e.g. “join” and “take part in”; “plunder” and “engage in plundering”; “rebel”, “revolt”); (2) by aggregating the distinct actions in broader spheres of actions, such as “protest”, “violence”, “communication”, “movement”, “facilitation”, “authority”. Four basic spheres of action account for almost all basic actions found in Mousnier’s narrative: *protest* (assemble, gather, rebel, revolt, join revolt, take part in revolt, take up arms), *violence* (burn, chase, kill, overcome, plunder, rob, scatter), *movement* (enter, get into, gather, let into, open a way, scatter), *facilitation* (are tacit accomplice, approve, grant an amnesty, help, let into).¹⁶ If we exclude the

14. It is in connection with a network representation of a story grammar that the limits of a story grammar start becoming apparent. A story grammar captures in its structure the elements of a syntactically well-formed narrative. “Tavern keepers and makers of casks revolt” is certainly a syntactically properly formed clause. Like all intransitive verbs, the verb “revolt” does not necessarily require an object. Yet tavern keepers and makers of casks do revolt *against somebody*, namely against those who imposed the *gabelle* on taverns and wine (presumably, state authorities). When we start analyzing networks of social interactions, dropping the triplet “tavern keepers and cask makers revolt” from analysis because it contains no *explicit* object would bias the network. Thus, when using narratives as mirrors of social reality, a semantic grammar must capture not only the narrative elements explicitly present in the text but also those semantically implied by the narrative. Although it is possible to modify the rewrite rule for the $\langle \text{object} \rangle$ in the following way: $\langle \text{object} \rangle \langle \text{explicit object} \rangle \langle \text{implicit object} \rangle$ to correct for the problem, there are further problems with a story grammar that may ultimately require investigators to use different linguistic tools (e.g. semantic roles; see Franzosi, *From Words to Numbers*. In the analyses that follow I have used triplets with implied objects.

15. The arrow does not refer to the linguists’ rewrite rule but to the statisticians’ directed relation from n_i to n_j .

16. It is probably impossible to classify verbs into mutually exclusive categories. Some verbs belong to different spheres of action (“fuzzy meaning”; see Franzosi, *From Words to Numbers*). “Assemble” or “gather” are both actions of “protest” and “movement”. “Scatter” and “chase” are both actions of “violence” and “movement”.

Table 1. *Relations of protest, violence and facilitation from Mousnier's narrative*

L_1 Relation ₁ <i>Protest</i>	L_2 Relation ₂ <i>Violence</i>	L_3 Relation ₃ <i>Facilitation</i>
<tavern keepers, authorities> <makers of casks, authorities>	<peasants, town houses> <governor, rebels>	<bourgeois, rebels> <persons of condition, rebels>
<craftsmen, authorities>	<villagers, country houses>	<officials of the <i>parlement</i> , rebels>
<day-laborers, authorities> <gardeners, authorities>	<villagers, suburb houses> <governor, peasants>	<bourgeois, rebels> <persons of condition, rebels>
<gardeners, authorities>	<gentlemen, peasants>	<officials of the <i>parlement</i> , rebels>
<market gardeners, authorities> <vinegrowers, authorities> < <i>laboureurs</i> , authorities>	<town militia, peasants> <cavalry, peasants> <cavalry, peasants>	<rebels, peasants> <gentlemen, governor> < <i>honnêtes bourgeois</i> , governor>
<villagers, authorities> <villagers, authorities>		<soldiers, governor> <poor, villagers> <king, rebels>

sphere of action of movement which does not include social actors as objects, we can bring together in three sets of ordered pairs, L , the subject and object of all the triplets having actions that belong to the spheres of protest, violence, and facilitation and objects that are social actors (see Table 1).

The sets L_1 , L_2 , L_3 of all ordered pairs of actors and their relations on the spheres of action of protest, violence and solidarity provide the basic information on the networks of social interactions in 1635 Bordeaux. But the longer the lists, the more difficult it would be to grasp the patterns of interactions. A graphic representation of the information in the three sets L_1 , L_2 , L_3 would provide a better alternative. The elements, or ordered pairs of relating actors, in L_1 , L_2 and L_3 can be represented graphically by drawing a line from the first actor in the element to the second. In this representation, each actor is also known as *node*, and each directed line is known as *arc*. A given set of actors and their corresponding arcs is called *directed graph* and can be visually represented as a diagram where nodes are points in a two-dimensional space and arcs are directed arrows between points.

A large number of relations implies a large number of directed graphs – one graph per relation. But a large number of actors may imply graphs with poor readability – too many nodes, with too many arcs cluttering the graph. Again, aggregation may be required in order to reduce the number of actors. Thus, such actors as makers of casks, craftsmen, day-laborers (also referred

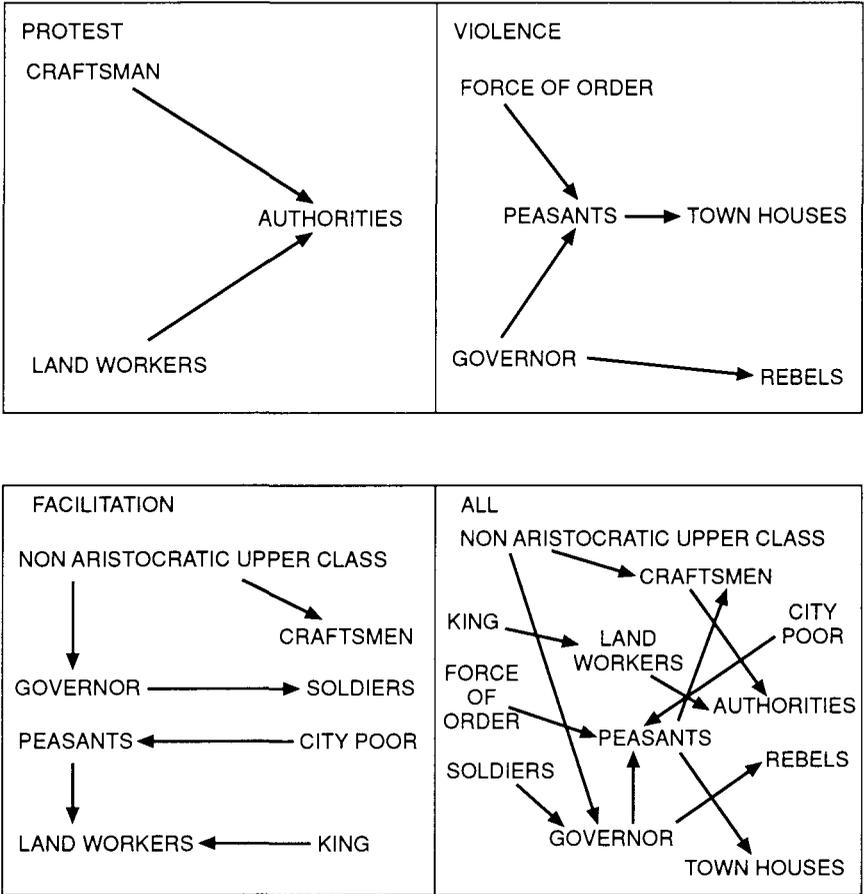


Figure 1. Network graphs of the social actors involved in Mousnier's narrative for selected spheres of action

to in the text as rebels) could all be classified as craftsmen; peasants, villagers, vinegrowers, gardeners, market gardeners, and *laboueurs* could be grouped together as agricultural workers; bourgeois, persons of condition, officials of the *parlement* and gentlemen as non-aristocratic upper class; cavalry, town militia and soldiers as forces of order. With the actors and actions properly aggregated, we can then place the arcs of the directed graphs on several figures (one for each relation), or on a single figure containing points representing the actors and arcs or lines for all relations, simultaneously (a *multivariate directed graph*). For Mousnier's data, this would yield the graphs of Figure 1.

The graph provides a clear picture of the social relations of seventeenth-century France and of the pattern of class alliances, of the coming together

of town and country, of sections of the upper classes and of the lower classes against an encroaching state. Yet does one really need to enroll the help of linguistics, set theory and network analysis to get that picture? No doubt an old-fashioned but much simpler reading of Mousnier's passage would do. But would that still do if you have hundreds or thousands of passages of this kind?

TOY EXAMPLES AND REAL DATA: THE TEST OF LARGE DATA SETS

In this section, I illustrate the results of a large historical project, based on data collected from some 15,000 newspaper articles using a semantic grammar on one of Italy's most troublesome periods: the 1919–1922 period. In this span of just four years, we go from the widespread labor mobilization of 1919 and 1920 leading up to the factory occupation movement of September 1920 (the *red years*), to the reaction of 1921 with its increasing levels of political violence and ending with the fascist take-over of power by Mussolini in the fall of 1922 (the *black years*). The plots of Figure 2 of the numbers of strikers and hours lost between 1879, when reliable strike data first became available, and 1922 leave no doubt about the uniqueness of the immediate post-war years.

Figure 3, based on monthly strike data, gives the tempo of mobilization during the red years, with peaks centered around May–July 1919 and January–June 1920. Then, in the fall of 1920 silence fell upon the Italian working class, with one last quick and sudden flurry of protest in June 1922.

What explains the temporal pattern of workers' mobilization during the *red years*? Did workers' mobilization abate because they obtained what they were after? Or was such an impressive movement harshly repressed? As we start asking questions, ever more come to mind. What was the movement all about? Who was involved? What were their actions?

We will search in vain for an answer to those questions on the basis of official strike statistics. Figures 2–4 go as far as those statistics take us. To find an answer to those questions we need to turn elsewhere. The methodological development illustrated in this paper finds its roots in the limitations of official strike statistics and, more generally, in the limitations of event counts in the quantitative study of historical processes.

Data collection for my 1919–1922 project was based on the Socialist newspaper *Il Lavoro*, published in Genoa. According to comparative analyses of some ninety newspapers published in Italy at the time, *Il Lavoro* (together with another socialist paper, *Avanti!*) provides by far the largest number of articles on social and industrial conflict and violence. The data in Table 2 on the frequency distribution of articles, disputes, triplets, actors and actions present in the database give a sense of the scale of the project.

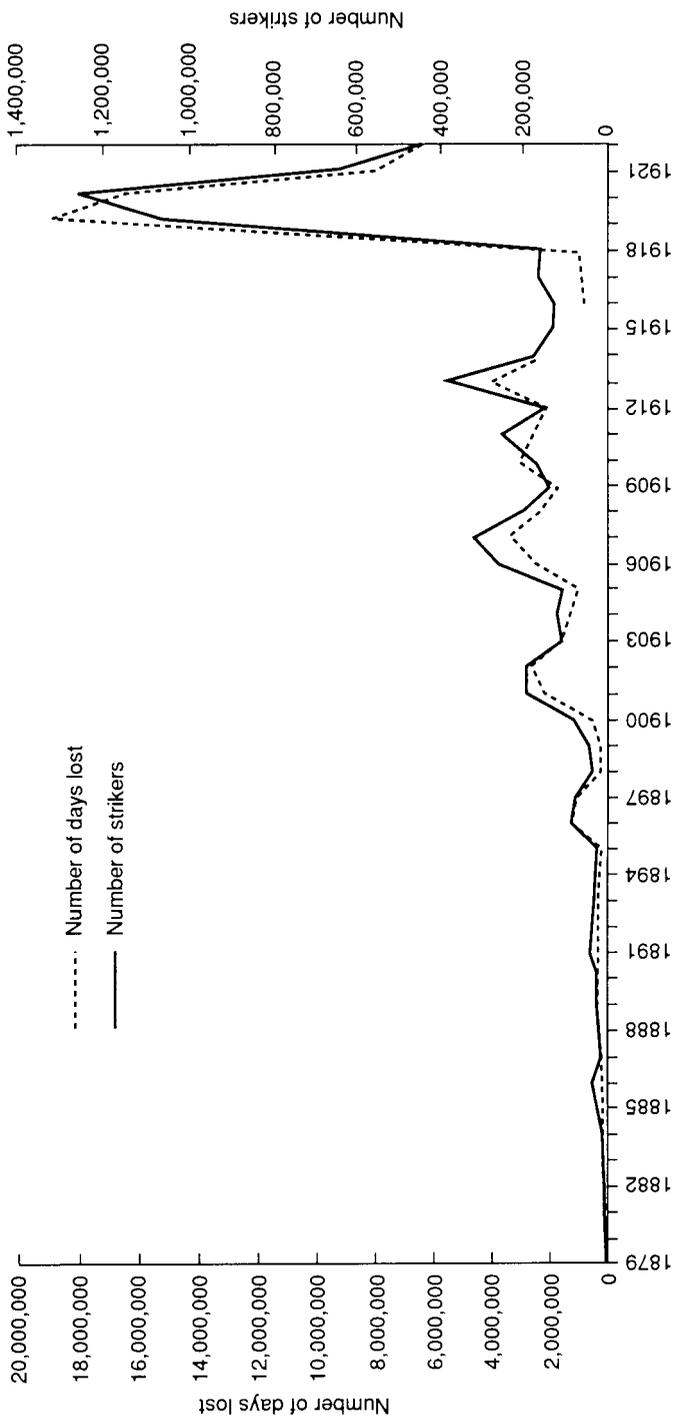


Figure 2. Yearly plots of the numbers of strikers and days lost (1879–1922)

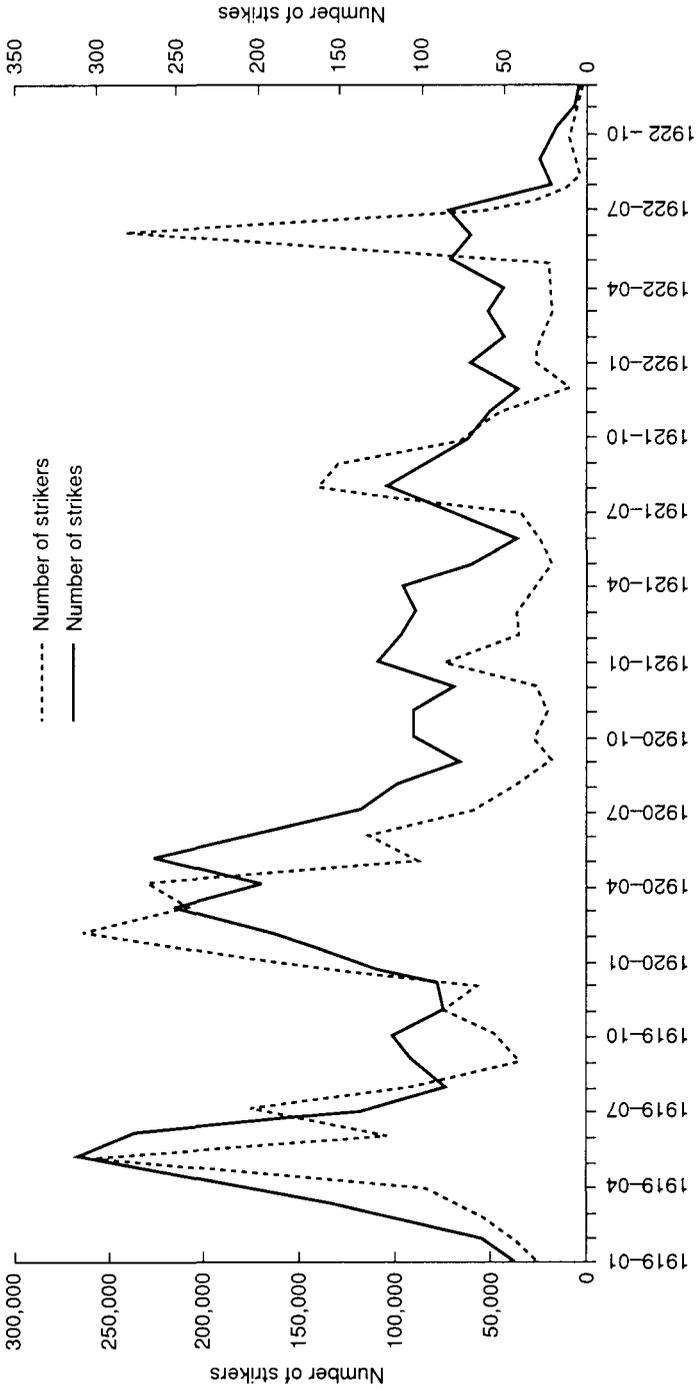


Figure 3. Monthly plots of the numbers of strikes and strikers (1919–1922)

Table 2. *Frequency distribution of articles, disputes, triplets, actors and actions in the database*

	1919	1920	1921	1922	Total
Number of articles	4,213	4,329	3,715	2,889	15,146
Number of disputes	597	622	1,132	926	3,277
Number of triplets	5,511	5,078	2,961	3,195	16,745
Number of <i>distinct</i> actors	498	412	355	367	1,052
Number of <i>distinct</i> actions	1,145	757	682	538	2,220

I have shown elsewhere¹⁷ that even simple exploratory analyses of the relational properties of these data, of “who does what and why”, can tell us a great deal about the 1919–1922 period in Italy. But the graphical representation of the network of social relations may tell us even more. For the sake of simplicity and for the purpose of illustration, I will focus here on the network graphs around the spheres of action of conflict and violence. The network graphs for the sphere of action of *conflict* (Figure 4) give a dramatic picture of the shift in the network of social relations from the “red years” to the “black years”, with a great deal of conflict in the economic arena during 1919 and 1920, as evidenced by the thickness of the line between workers and employers. In 1921, however, a distinct political arena of conflict emerges (“network cluster”), involving the fascists against the socialists and, especially, the unions. This confirms the historians’ accounts on early fascist tactics, aimed at smashing both the political and economic organizations of the working class. In fact, the total amount of political conflict borne out by the network graphs is grossly underestimated because most actions performed by the fascists are classified under *violence* (next network graph). Even the revival of conflict in the economic arena hides the basic fact that the nature of conflict shifted from offensive to defensive in the four-year period. Furthermore, while in the early years the total number of actions of conflict that workers carried out against employers is roughly equal to that of employers against workers (as represented by the double-pointed arrow between employers and workers in the network graphs for 1919 and 1920), in 1921 the employers were on the offensive (as represented by the thicker line going from employers to workers).

The graphs for *violence* (Figure 5) clearly bring home the point that things changed drastically between 1919–1920 and 1921–1922. The graphs for 1919, 1920 and 1921 all show a “star” shape, with one actor at the center of the network of interactions, a hub for relations where actors have no links to one another, except with the star. But, interestingly enough, while in 1919 and 1920

17. Roberto Franzosi, “Mobilization and Counter-Mobilization Processes: From the Red Years (1919–20) to the Black Years (1921–22) in Italy. A New Methodological Approach to the Study of Narrative Data”, in John Mohr and Roberto Franzosi (eds), *New Directions in Formalization and Historical Analysis*, special issue of *Theory and Society*, 26 (1997), pp. 2–3.

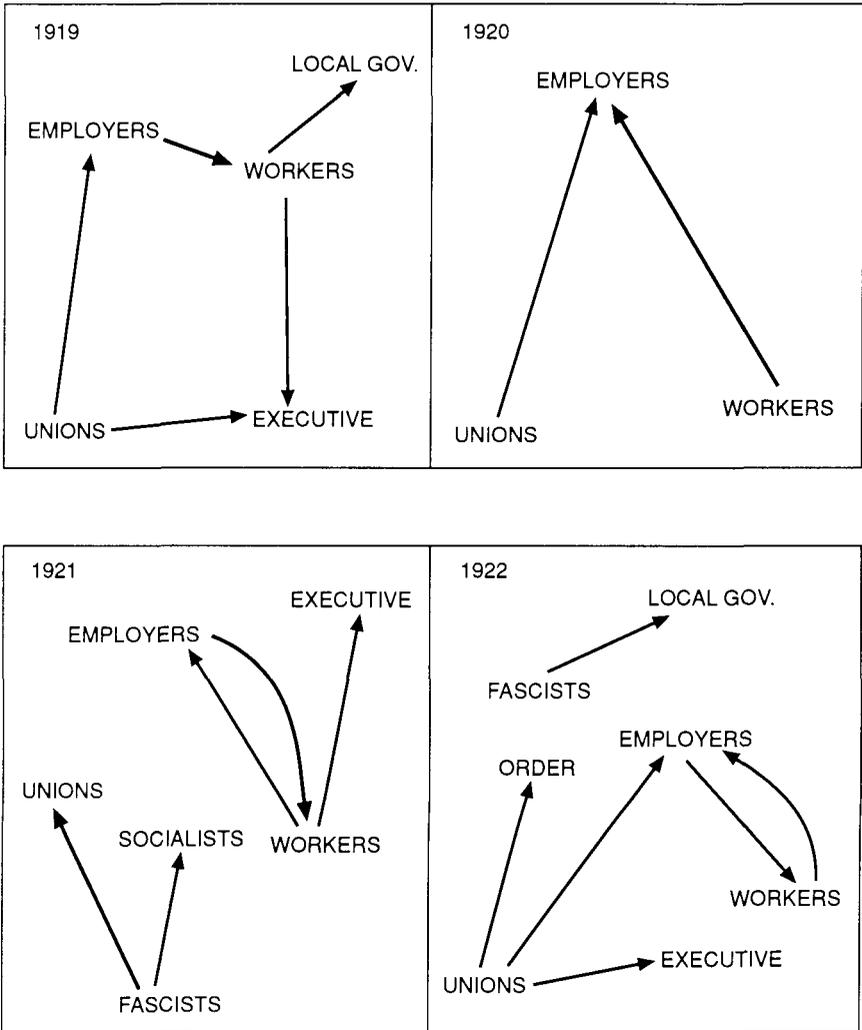


Figure 4. Network graphs of the social actors involved for the sphere of action on *conflict* (1919–1922)

the police and the army (“order”) occupy the center of the star, in later years the “fascists” take that position. The graphs also bring home another point, namely that most violence is actually performed by the police and the army, as a number of studies on protest movements and collective action have long underscored.¹⁸ The graph for 1919 highlights an important actor – the petty

18. E.g. Charles Tilly, “State-Incited Violence, 1900–1999”, *Political Power and Social Theory*, 9 (1995), pp. 161–225.

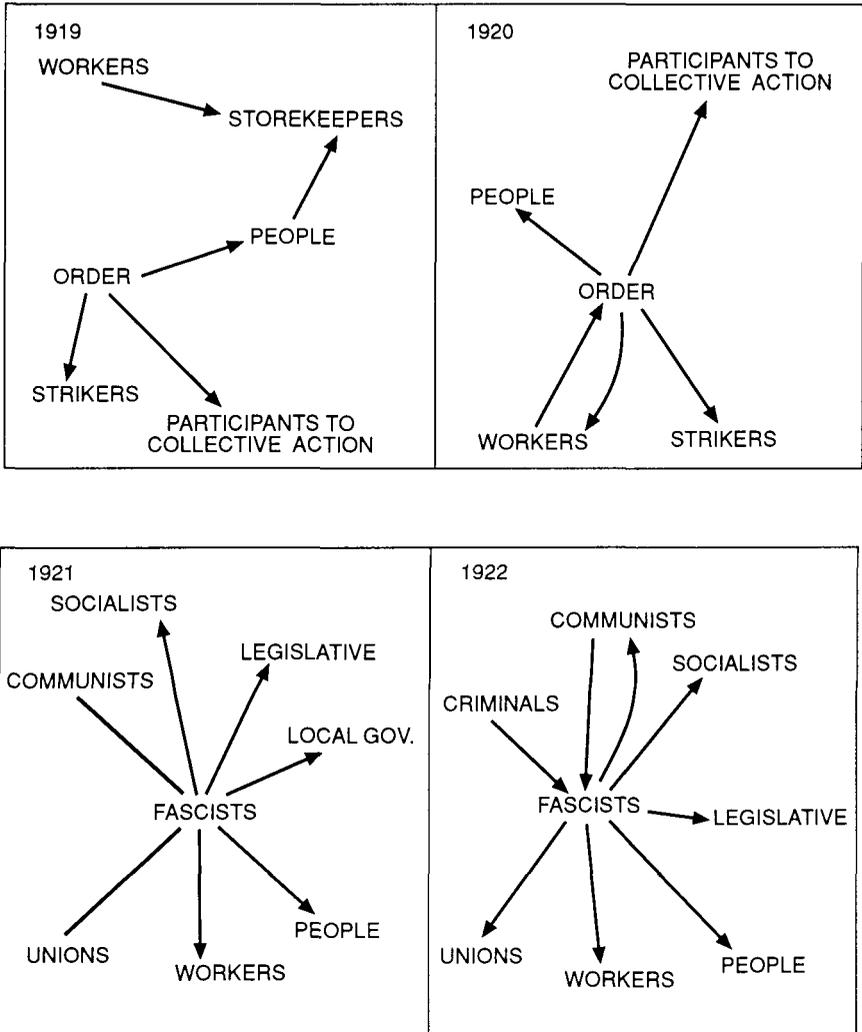


Figure 5. Network graphs of the social actors involved for the sphere of action of *violence* (1919–1922)

bourgeoisie (namely, shopkeepers) – the target of people’s and workers’ violence, particularly during the intense period of mobilization for the *caroviveri*. It is not surprising that the petty bourgeoisie embraced fascism a year later!

The double arrow between communists and fascists in the network graph for 1922 indicates that the communists were not mere recipients of fascist violence. They acted as agents in many violent actions against the fascists. That finding is not surprising. The Communist Party was born out of the

controversy between the reformist and the revolutionary wing of the Socialist Party after the failure of the 1920 factory occupation movement. From the beginning, the communists took the fascists head on, responding with violence to violence, and forming such organizations as the *arditi del popolo* to counter the fascists' *arditi d'Italia*.

I could continue with this exercise, faithfully describing the network graphs for different spheres of action (relations), each new graph adding more or less detail to, shedding a dimmer or brighter light on basically the same picture of a drastic change in the number and type of actors involved and in the forms of their interactions from the "red" to the "black" years. From a methodological viewpoint, the combination of a linguistic approach to content analysis (semantic grammars) and of network models applied to the qualitative data collected from narratives has thus proved to be a powerful tool of socio-historical research. From a substantive viewpoint, historians are likely to take issue with the validity of my account (what conclusions can one reach on the basis of one newspaper only, and a socialist paper at that?); but they are also likely to sympathize with the shift of focus in statistical representations from variables to actors and social action. Sociology, as well, has seen a resurgence of interest in issues of agency.¹⁹ Sociologists, however, will generally find the descriptive nature of the network graphs not to their taste. How can one test hypotheses on the basis of descriptive work? Indeed, the network graphs on the sphere of actions of conflict and violence presented here are not well suited to address empirically either the theoretical concerns of such theories of mobilization processes as resource mobilization, political opportunity structure and frame alignment,²⁰ or, for that matter, hypotheses drawn from broad theories of fascism based on class alliances.²¹ Notwithstanding, the network evidence does highlight specific spheres of conflict (in the economic, social and political arenas with specific sets of actors involved), specific points of tension in the social structure (from the obvious ones, such as workers and unions versus employers, to the less obvious ones, such as working-class women versus storekeepers). The identification of those points of tension and of

19. E.g. Mustafa Emirbayer and Jeff Goodwin, "Network Analysis, Culture, and the Problem of Agency", *American Journal of Sociology*, 99 (1994), pp. 1411–1454.

20. For an introduction, see McAdam *et al.*, *Comparative Perspectives on Social Movements*.

21. For example, Barrington Moore saw fascism as the political regime that results from the alliance of the landed upper class, the state and the industrial bourgeoisie (Barrington Moore, *Social Origins of Dictatorship and Democracy* (Boston, 1966), p. 436). Contrary to this view, for Salvatorelli, "Fascism represents the class struggle of the petty bourgeoisie, wedged between capital and proletariat": cited in Renzo DeFelice, *Le interpretazioni del Fascismo* (Bari, 1995 [1969]), p. 186. Seymour Martin Lipset echoed that view, decades later: "Fascism is basically a middle-class movement representing a protest against both capitalism *and* socialism, big business *and* big unions": Seymour Martin Lipset, *Political Man. The Social Bases of Politics* (Baltimore, 1981 [1959]), p. 131; emphasis in the original.

those arenas of conflict does reveal that certain types of alliances among social actors may be more likely than others.²²

CONCLUSIONS

The study of conflict has typically been based on event counts.²³ When I first conceived of a project on the 1919–1922 years, one thing was clear to me: to make any inroads in the understanding of those conflict-ridden years and the rise of Italian fascism I would have to abandon official strike statistics (and event counts, more generally). Those data have the advantage of being readily available; but the poverty of their information is such that nothing could be said, on the basis of those data, on what employers, the state or other social groups do during periods of high mobilization. The danger of using official strike counts is to produce a one-sided and distorted view of social conflict where only one of the social actors involved (workers) acts. Yet, if “it takes two to tango”, it takes at least two to fight.²⁴ Paradoxically, even for workers, reliance on strike counts would produce a distorted picture of working-class strategies. After all, strikes are but one form of protest tactic within wider repertoires.

Driven by this quest for a richer source of event data, I focused on the narratives of police reports and newspaper articles. I then set out on the task of developing an approach to content analysis where coding scheme design would not be dependent upon the investigator’s theoretical or substantive interests *but on the inherent linguistic properties of the text itself*. The end result of that quest was a “semantic grammar”, the simple linguistic structure subject-action-object and their respective modifiers (e.g. time, space, demands of action, or number, type and organization of subjects and objects).

To some extent, a semantic grammar reflects intrinsic properties of narrative text, rather than the investigator’s theoretical approach. After all, narrative texts, at their core, are characterized by actors performing some actions.²⁵ Furthermore, despite the complexity of semantic grammars as coding schemes, their mathematical underpinnings in set theory make them easily implementable in a computer environment using relational database systems. Finally, in the process of coding text, of extracting information from narrative accounts, of going from words to numbers, semantic gram-

22. Another approach to the analysis of social networks (the sociometric approach) is likely to appeal to the structural concerns of social scientists. Indeed, the blockmodels that this approach yields have enjoyed increasing popularity (see the review and critique of recent work in Emirbayer and Goodwin, “Network Analysis”).

23. E.g. counts of strikes, see Roberto Franzosi, *The Puzzle of Strikes. Class and State Strategies in Postwar Italy* (Cambridge, 1995).

24. On these issues, see *ibid*.

25. See Franzosi, “Narrative Analysis, or . . .”.

grams retain more of the rich texture of the original narrative than traditional content analysis techniques.

The SAO structure is not only invariant across purely narrative clauses but because of its relational properties, it also allows researchers to analyze the linguistic content of its categories with the help of statistics. In particular, the data that semantic grammars deliver are very well suited for the application of new tools of data analysis such as network models. Both semantic grammars and network models are fundamentally concerned with actors and their actions, with agents and agency. As such, these methodological tools should draw sociology closer to history, traditionally much more concerned with issues of agency. There may be a variety of ways of going from words to numbers, of analyzing qualitative data with quantitative tools. In this paper I have illustrated one thin thread that runs from words to numbers, from story grammars to network models via a common representation in set theory.

On the basis of a semantic grammar implemented in a computer environment, I coded information from some 15,000 newspaper articles on the 1919–1922 period in Italy. Even the handful of empirical analyses of the data presented in this paper as an illustration of the technique bear witness to the power of the technique. The application of network models to narrative data brings out clear historical patterns in the data, such as the stark shift from working-class mobilization in 1919–1920 to counter-mobilization in 1921–1922. No other available data could have brought out those same patterns to the extent that this technique has. Available official strike statistics, for instance, with their crude counts of just one type of event (strikes) would not have allowed us to tease out the broader repertoire of working-class actions or the variety of actors, with their specific claims and forms of actions, involved during the phase of mobilization (e.g. students, women, war veterans). Even less so would they have allowed us to map the counter-mobilization process, the quick spread of fascist violence. Official strike statistics simply do not provide information on those issues. Neither would qualitative historical studies greatly help. The growing number of city or regional monographs on the rise of fascism certainly helps to bring out the much needed detail of local situations. But it is hard to get a sense of patterns from these studies.

And yet, expressions that abound throughout my paper – “a semantic grammar preserves much of the original text input”; “semantic grammars can achieve quality without sacrificing quantity”; “coding scheme design would not be dependent upon the investigator’s theoretical or substantive interests *but on the inherent linguistic properties of the text itself*” – overlook the many obstacles and problems one finds along the way in going “from words to numbers”, even when using semantic grammars. Indeed, for all their power, these new tools for the collection and analysis of narrative data do have their shortcomings.²⁶

26. For an in-depth treatment of these issues, see Franzosi, *From Words to Numbers*.

A purely semantic and structural analysis of narrative misses out a great deal of information contained in a text.²⁷ Semantic grammars do not eliminate the “role of the reader”, i.e. the role of the coder in the interpretation of text. Narrative texts are hardly ever made up of purely narrative clauses only. Narrative and non-narrative clauses typically mix in a narrative text. Coding of non-narrative clauses within the framework of the grammar will pose serious problems to the coders. Even purely narrative clauses may not necessarily conform to the simple template subject-action-object of the grammar.²⁸ The greater the difference between the surface representation of these clauses and the deep structural representation of the grammar, the greater the interpretative role of the coder, in trying to squeeze one into the other (the simplest case being the linguistic operations required to convert passive forms to active). And semantic grammars have nothing to offer to ease the problem of aggregation. Once investigators have coded thousands of individual actions, how do they classify them, how do they aggregate them for analysis?²⁹

Reliance on narratives and semantic grammars may also tend to focus scholarly attention on particularly transformative events (such as the “red years” or the “black years”) and on descriptive modes of explanation. The costs of data collection on longer sample periods are responsible for the first drawback; and the very richness of the available information on “who did what” for the second. Paradoxically, the event data delivered by a semantic grammar from newspaper sources are both too rich and not rich enough. Alas, there may be no escape from a multiple-evidence, multiple-method approach to the solution of puzzles of historical nature.

APPENDIX A. MOUSNIER'S PASSAGE CODED WITHIN THE CATEGORIES OF A SEMANTIC GRAMMAR

<subject> authorities (<comment> agent inferred) <action> impose gabelle
(<space> Bordeaux <time> May 1635 <time> June 1635) <object> taverns
<object> wine <implicit target> tavern keepers <implicit target> makers of
casks <implicit target> vinegrowers
<subject> tavern keepers <subject> makers of casks <action> revolt <implicit
target> authorities
<subject> craftsmen (<number> many) <subject> day-laborers (<number>
many) <action> join revolt <object> rebels <implicit target> authorities
<subject> bourgeois (<number> majority) <subject> persons of condition

27. E.g. the examples discussed in Franzosi, “Narrative Analysis, or . . .”, or Larry Griffin, “Narrative, Event-Structure Analysis, and Causal Interpretation in Historical Sociology”, *American Journal of Sociology*, 98 (1993), pp. 1094–1133.

28. On these issues, see Franzosi, “Narrative Analysis, or . . .”.

29. For a brilliant critique see Aaron V. Cicourel, *Method and Measurement in Sociology* (New York, 1964), pp. 7–38; see also Griffin, “Narrative, Event-Structure Analysis, and Causal Interpretation”.

(<number> a large number) <subject> officials of the *parlement* <action> are tacit accomplice of <object> rebels <implicit target> authorities
 <subject> bourgeois (<number> majority) <subject> persons of condition (<number> a large number) <subject> officials of the *parlement* <action> approve (<comment> regarded as their liberators) <object> rebels
 <subject> rebels <action> want <object> infinitive clause
 <subject> rebels <action> to control <object> city gate (<proper name> Saint Julien)
 <subject> rebels <action> want <object> infinitive clause
 <subject> rebels <action> to let into (<space> the city) <object> peasants (<space> surrounding district)
 (<triplet relation> so that) <subject> peasants <action> help <object> rebels
 <subject> peasants <action> demand (<type> loudly) <object> infinitive clause
 <subject> peasants <action> to enter (<space> city <reason> share in the plunder of the town (<comment> motive ascribed to the peasants by the secretary to the duc d'Épernon, who wrote these lines))
 <subject> gardeners (<number> a certain number of) <subject> market gardeners (<number> a certain number of) <subject> vinegrowers (<number> a certain number of) <subject> *laboueurs* (<number> a certain number of) (<comment> very seriously affected by the indirect taxes) <action> get into (<space> the town) <action> take part in the revolt <implicit target> authorities
 <subject> peasants <action> plunder <object> houses (<number> some <space> in the town)
 <subject> town militia <subject> gentry (<space> neighboring <number> most) <action> do not help <object> governor (<space> Guyenne <proper name> duc d'Épernon)
 <subject> gentlemen (<number> few) <subject> *honnêtes bourgeois* (<number> few) <subject> soldiers (<space> town <number> few) <action> help <object> governor (<space> Guyenne <proper name> duc d'Épernon)
 <subject> governor (<space> Guyenne <proper name> duc d'Épernon) <action> overcomes <object> rebels
 <subject> governor (<space> Guyenne <proper name> duc d'Épernon) <action> persuades <object> peasants <object> infinitive clause
 <subject> peasants <action> to yield
 <subject> villagers (<number> all) <action> take up arms (<space> countryside) <implicit target> authorities
 <subject> villagers (<number> all) <action> rob <object> houses (<space> in the countryside)
 <subject> villagers (<number> in great numbers) <action> gather (<space> in all the suburbs) <implicit target> authorities
 <subject> villagers (<number> in great numbers) <action> strive <object> infinitive clause

<subject> villagers <action> to enter (<space> town)
 <subject> poor (<space> town) <action> want <object> infinitive clause
 <subject> villagers <action> to enter (<space> town)
 <subject> poor (<space> town) <action> try <object> infinitive clause
 <subject> poor (<space> town) <action> to open a way <object> villagers
 <subject> villagers <action> assemble (<space> in the suburb of Saint-Surin)
 <implicit target> authorities
 <subject> villagers <action> burn <object> houses (<number> several
 <space> in the suburb of Saint-Surin)
 <subject> governor (<space> Guyenne <proper name> duc d'Épernon)
 <subject> gentlemen of his guard (<number> forty or fifty) <subject> town
 militia (<number> few companies) <action> scatter <object> peasants
 <subject> cavalry <action> chase <object> peasants
 <subject> cavalry <action> kill <object> peasants (<number> forty or fifty)
 <subject> peasants (<number> all other) <action> behave themselves prop-
 erly (<duration> that year)
 <subject> king <action> grant an amnesty <object> rebels.

ANNOTATED BIBLIOGRAPHY

Braudel, Fernand, *On History* (Chicago, 1980).

A collection of essays of reflections upon history by one of the great historians of the twentieth century.

Cicourel, Aaron V., *Method and Measurement in Sociology* (New York, 1964).

A truly insightful book. The chapter on "Measurement and Mathematics" is brilliant. I would make it required reading in any graduate training.

DeFelice, Renzo, *Le interpretazioni del Fascismo* (Bari, 1995; 1st pub. 1969).

An overview of various interpretations of fascism. Social science models do not fare well in the eyes of this great historian of Italian fascism.

Elton, G.R., "Two Kinds of History", in Robert William Fogel and G.R.

Elton, *Which Road to the Past? Two Views of History* (New Haven, 1983), pp. 71–121.

A debate between two great historians – G.R. Elton, a traditional, narrative historian, and Robert Fogel, the "cliometrician". A "must" for any one interested in the terms of the debate between narrative history and "scientific" history.

Emirbayer, Mustafa and Jeff Goodwin, "Network Analysis, Culture, and the Problem of Agency", *American Journal of Sociology*, 99, 6 (1994), pp. 1411–1454.

An investigation into the recent explosion of network models and into the relationship between network models and agency.

Franzosi, Roberto, "The Press as a Source of Socio-Historical Data: Issues in the Methodology of Data Collection from Newspapers", *Historical Methods*, 20, 1 (1987), pp. 5–16.

An investigation into the validity of newspapers as sources of socio-historical data and a discussion of coding scheme design for the collection of those data.

- Franzosi, Roberto, "From Words to Numbers: A Generalized and Linguistics-Based Coding Procedure for Collecting Event-Data from Newspapers", in Clifford Clogg (ed.), *Sociological Methodology*, vol. 19 (Oxford, 1989), pp. 263–298.

The first step in the author's journey "from words to numbers". Traces the background of semantic grammars and develops a grammar most appropriate for the study of conflict events.

- Franzosi, Roberto, "Strategies for the Prevention, Detection and Correction of Measurement Error in Data Collected from Textual Sources", *Sociological Methods and Research*, 18, 4 (1990), pp. 442–471.

An investigation into the possible sources of errors when coding narrative data with the help of a computer-based semantic grammar.

- Franzosi, Roberto, "Computer-Assisted Coding of Textual Data: An Application to Semantic Text Grammars", *Sociological Methods and Research*, 19, 2 (1990), pp. 224–256.

The author highlights the characteristics that a software for the collection of narrative data on the basis of a grammar should have.

- Franzosi, Roberto, "From Words to Numbers: A Set Theory Framework for the Collection, Organization, and Analysis of Narrative Data", in Peter Marsden (ed.), *Sociological Methodology*, vol. 24 (Oxford, 1994), pp. 105–136.

The final step in going "from words to numbers". And set theory provides the link.

- Franzosi, Roberto, *The Puzzle of Strikes. Class and State Strategies in Postwar Italy* (Cambridge, 1995).

An attempt to solve the puzzle of Italian strikes in light of economic, institutional, organizational and political factors. Brings in a great deal of both quantitative and qualitative evidence. A scathing critique of mindless quantitative approaches to historical processes through "doing it".

- Franzosi, Roberto, "Mobilization and Counter-Mobilization Processes: From the Red Years (1919–20) to the Black Years (1921–22) in Italy. A New Methodological Approach to the Study of Narrative Data", in John Mohr and Roberto Franzosi (eds), *New Directions in Formalization and Historical Analysis*, a special issue of *Theory and Society*, 26, 2–3 (1997).

The first substantive results of the author's long journey "from words to numbers".

- Franzosi, Roberto, "Narrative Analysis, or . . . Why (and How) Sociologists Should Be Interested in Narrative", in John Hagan (ed.), *Annual Review of Sociology*, vol. 24 (Palo Alto, 1998).

An analysis of narrative focused on the linguists' contribution but with an eye to the sociologist. A very provocative piece.

- Franzosi, Roberto, *From Words to Numbers: Narrative as Data* (Cambridge, forthcoming).
- A long journey across time and a number of disciplines as the author tackles the problem of narrative as a source of data for quantitative analysis. An investigation into the methodological, epistemological and ontological aspects of this enterprise with much breaking of academic canons.
- Griffin, Larry, "Narrative, Event-Structure Analysis, and Causal Interpretation in Historical Sociology", *American Journal of Sociology*, 98, 5 (1993), pp. 1094–1133.
- A very detailed and careful investigation in the structure of a short narrative on the basis of structural properties of narrative and using ETHNO as an aid.
- Lipset, Seymour Martin. *Political Man. The Social Bases of Politics* (Baltimore, 1981; 1st pub. 1959).
- A classic that deals with many aspects of politics, including fascism.
- Markoff, John, *The Abolition of Feudalism: Peasants, Lords, and Legislators in the French Revolution* (University Park, PA, 1996).
- Twenty years in the making, this is a book that was well worth waiting for. Approaches issues of historical explanations with an incredible wealth of historical evidence (including content analysis of the *Cahiers de doléances*). Well done!
- Markoff, John, Gilbert Shapiro and Sasha Weitman, "Toward the Integration of Content Analysis and General Methodology", in David R. Heise (ed.), *Sociological Methodology* (Oxford, 1974), pp. 1–58.
- A plea for a linguistic approach to content analysis that delivers more than previous pleas and less than hoped for.
- McAdam, Doug, John D. McCarthy and Mayer N. Zald (eds), *Comparative Perspectives on Social Movements* (Cambridge, 1996).
- The latest reader in a "growth industry".
- Moore, Barrington, *Social Origins of Dictatorship and Democracy* (Boston, 1966).
- Many argue that modern historical sociology started from here.
- Mousnier, Roland, *Peasant Uprisings in Seventeenth-Century France, Russia, and China* (New York, 1972).
- A narrative account by a French historian of the peasants' rebellions of the seventeenth century.
- Tilly, Charles, "State-Induced Violence, 1900–1999", *Political Power and Social Theory*, 9 (1995), pp. 161–225.
- There are some advantages in having monopoly control over the means of violence. Not surprisingly, it is the state that is the most violent social actor.
- Wasserman, Stanley and Katherine Faust, *Social Network Analysis. Methods and Applications* (Cambridge, 1994).
- A clear and thorough exposition of network models. Successive chapters get progressively more difficult without mathematics.