**Relations in Texts (and How to Get at Them)**

***Roberto Franzosi***

Thank you for having me here, although … I am going to take you further away from concerns with open-ended questions. For me, being here today is homecoming, because I conceived this whole research agenda of going "from words to numbers," of quantifying narrative information, at the University of Michigan. It was in 1981-82, an academic year I spent here as a post-doc. The irony is that I was a post-doc in the Perry School building, at the Center for Research on Social Organization (CRSO). Charles Tilly was the director. There were several people at CRSO – from Bill Gamson to Geoff Paige, who is still here, Bob Thomas and many graduate students – using newspapers and content analysis to study historical processes. Just like when I was a professor at Oxford where any reference to Cambridge was as "the other place," ISR was "the other place" for Perry School people, the place that housed the best of all quantitative scholars and survey researchers. Today, I asked the person setting up my computer for internet access, if David Featherman was still on this floor (David and I were colleagues at the University of Wisconsin and wanted to say hello). He said no, that David was now in the Perry building. I said: "the Perry building?" "Yes, ISR bought and restored the Perry building."

What I am going to talk about is similar to what Carl Roberts talked about before: relations in text. There are various types of relations, relations between words, relations between concepts. Kathleen Carley does some of this work, relating concepts in a text. What images do we put together, do we associate with each other? The relations I have been interested in are relations between social actors through their actions. To put it in context, I arrived at the University of Michigan having finished a Ph.D. at Johns Hopkins. I had just finished a very quantitative dissertation on Italian, postwar strikes. In graduate school, I had done nothing but mathematics and statistics (including 2 years of computer science prior to that). But, sadly, I had come to the end of my dissertation – although good enough to get me job offers from Columbia and Wisconsin – thinking that there was more to strikes than econometric and spectral coefficiencients. Sad conclusion, after years of investment in cutting-edge statistical methods! I came to CRSO with the thought that there were social actors behind those coefficients, but these actors had been lost behind the coefficients. Tilly and all these other people at CRSO at the Perry building were using newspapers as sources of social historical data on who does what, and they were using content analysis as a technique to extract this information. At the time, content analysis did not allow you to establish relations between different types of information. It just allowed you to say whether workers, employers, the police are mentioned in a text, whether there are strikes, demonstrations, arrests, or layoffs. Connections, for instance betweeen actors and their actions, were difficult to make. But, I was looking for an approach that would allow me to make those connections. I thought I could crack the problem in no time. Over 25 years later … I am still working on it.

I decided to apply my newly learned methodological approach to the study of the rise of Italian Fascism, to the period of Italian history between 1919 and 1922. Again, to give you some context for the project, take a look at the monthly plot of the number of strikers between 1879 and 1922. Notice the tremendous spike in 1919-1920 (only surpassed by the 1969 "hot autumn"). Other countries in Europe experienced similar trends, like the famous 1926 general strike in England. Yet, a closer look at the 1919-1922 period will reveal two distinct periods: the first two years, called the "red years," are characterized by high working class activism. The period ended in September 1920 with a widespread factory occupation movement. The next two years are known as the "black years," because of the fascist counter mobilization ending in October 1922 with Mussolini's takeover of power.

A typical newspaper narrative of conflict of the period reads something like this: "Fascists arrive on a truck on the night of 5/7/1921 in Bissone di S. Cristina at the pub of Mr. Prati. There, they seize 5 workers and take them outside. They beat up the 5 workers with retorted cowhide with lead inside." Ouch! The question is: What can you do with this type of text as a social scientist? What I came up with, by sheer brute force approach, is what linguists and cognitive psychologists have called "story grammar" – only later did I discover this; at the time, I didn't know any linguistics at all. Basically, a "story grammar" is made up of actors and their actions, or the structure subject-verb-object, where in

narrative, subject and verb are typically actors and actions. For each of these three categories (I call the SVO construct "semantic triplet", the basic skeleton narrative sentence), you can have such modifiers as the number and organizational affiliations of subjects (e.g., 10,000 GM workers belonging to the UAW) and the time and space of action (e.g., a strike, today, in Detroit). Time and space are known as the fundamental categories or narrative. Without time, in particular, there is no narrative (as stressed by the way children's and fairy tales start: "Once upon a time …" or "Once upon a time, in a faraway land …"). You can take a story grammar and use rewrite rules to establish stringent relations between the objects that make up the grammar. A rewrite rule (formally expressed by a right-pointing arrow →) says that an object on the left hand side (typically surrounded by angular brackets, e.g., <semantic triplet>) is rewritten as a function of the objects on the right hand side. Thus, a <semantic triplet> is rewritten as one or more {<subject>} (curly brackets {} read as one or more), one or more {<action>} and, optionally, one or more [{<object>}] (where the square brackets [] read as optional). Whether an object is required depends upon the type of verb you have, transitive or transitive (a transitive verb, a verb that allows the subject to "transit" to the object, requires an object, while an intransitive verb has no object). The story grammar that I used on the Italian fascism project is pages and pages of rewrite rules. It mimics the complexity of natural language. I am using a similarly complex story grammar for a current project on lynching of African Americans in Jim Crow South, between 1875 and 1930. The set of rewrite rules constitutes a relational structure where every object in the grammar is related to every other object in that grammar. Relations between certain objects can also be hierarchical, as shown in this slide, where one or more semantic triplets make up an event, and one or more events make up a larger unit of conflict, dispute/macro-event.

I developed computer software to carry out this type of research (PC-ACE, Program for Computer-Assisted Coding of Events, www.pc-ace.com). The unfortunate thing is that in the 1980's there was no software that allowed you to do this type of research. I started working with Morrow CPM machines. Long story. But even today there is no software that allows you to do what I call quantitative narrative analysis. You cannot do this type of work in ATLAS.ti, NVivo, or MaxQda. This type of software does not allow you to make connections between objects easily. As a result, over the years, I have spent a great deal of time doing programming work.

This slide shows my data. Basically words, something not normally recognized as data. Indeed, this is what killed my tenure prospects at Wisconsin-Madison. At least some colleagues were asking: "What's Franzosi gonna do with thousands of words in the computer?" They didn't believe that something could come out of it. I spent two years on unemployment. Well, in the end, I still got to be here today. And unemployment was like a glorified sabbatical. That is the truth. Anyway … These are my data. In black, are the objects of the story grammar (the coding categories, if you will). In red, are the words that I took from the story in the newspaper article. You can see the advantages of this type of coding. First, your data are exactly what you started with; you didn't aggregate. You just put different textual elements in appropriate categories. Second, this type of coding is far more natural than any other type of aggregate coding because it reflects the way we understand stories: who was involved? What did they do? When did it happen? Where? It is a more natural way of parsing a text. Instead of parsing a text and putting textual elements in aggregate, theoretically defined categories, you put them in the "natural" categories of narrative (namely, who, what, when, where, and why).

I used a story grammar and PC-ACE to code information on the rise of Italian fascism (1919-1922) from 3 different newspapers – 2 socialist newspapers (*Il Lavoro* and *Avanti!*) and one fascist newspaper (*Popolo d'Italia*). From the acknowledgments in this slide, you can see that several grants have gone into this (besides over 25 years of work). You can also see that I finished coding *Il Lavoro* in 1990 and the other two newspapers in 2006 and 2008. Take a look at the number of semantic triplets for *Il Lavoro* and *Avanti!* and *Popolo d'Italia*. Nearly 7 times as many triplets for about the same number of articles for the most recently coded newspapers. That is the result of the PC-ACE software I used in the 1990's versus the current release. It was like Galileo's 30X telescope, versus the 2-to-6X that his competitors had: you can see so much more with a 30X! Anyway … we coded some 60,000 articles from the three papers. It was a huge amount of work. I just finished about a month ago coding the third newspaper.

Now, for the question: “What’s Franzosi gonna do with thousands of words in the computer?” Remember! There were doubts in 1989 as to whether anything could come out of all these words in the computer.Remember also that the kind of approach I was looking for in order to analyze the data was not regression (although I have shown that you can use regression models to analyze this type of data). I was interested in techniques of data analysis that would allow me to exploit the fundamental features or narrative: time, space, actors and actions. This slide shows a GIS “hot-points” map. What is a hot-points map? In this particular map, actions of conflict by working class actors in 1919-1920 and violent actions by Fascist actors in 1921-1922 are plotted together; hot points are the points where the two overlaped more extensively than in other points (they go from yellow, to orange, and to red, the hottest points). If there is any truth to the so-called “Red Menace Hypothesis” that the rise of fascism was a reaction to socialism, you would expect that in the same place where there was working class activism you would later find fascist activism. Well, there are several red and orange points here and there in the map, lending some credence to the Red Menace Hypothesis. When I first saw this map, I was shocked. You travel for 25 years collecting data without knowing what you would see in the end. In an ironic twist, that night, I was caught in the tornado that hit Atlanta. Tree branches, garbage cans, windows were flying in the air, hitting my car left and right. I thought to myself: “within the framework of a tragic-romantic personality like mine, it would be only appropriate if I were killed in the day that I made a major discovery.” But that’s not how it ended.

The reason I was so shocked when I saw the hot-points map was that the red and orange points on the map are precisely the points that historians of local situations have studied: Puglia, Campania, Tuscany, Emilia Romagna. I knew to expect hot points there, and indeed I did. But at the same time, if you look at the scattering of the black points in the next map, you can see that the Fascists were not active in hot points areas only, but they hit everywhere, throughout the country. So, there is certainly truth to the Red Menace Hypothesis; but there is also truth to the fact that fascist violence just spread like a fan throughout the country.

GIS maps rely on time, space, actor and action, the fundamanetal categories of narrative. But another technique is similarly based on fundamental categories of narrative: network analysis. Network graphs map relationships among social actors, linking one actor to another in a particular sphere of action. The graphs of this slide depict the network around violent actions. Several distinct verbs have been aggregated into this category of violence (e.g., kick, beat up, kill, shoot). In network graphs, the thickness of lines measures approximately the number of actions behind each line (approximately, because if you made thickness truly proportional, some lines would cover the entire screen). The arrow measures the direction of the violence (e.g., from fascists to socialists). Notice how there are two centers of violence. One center is fascists’ violent actions toward everybody else; the other center of violence is the police. If the historians’ division of the 1919-1922 period into “red years” and “black years” has any empirical foundation, we should expect a different type of social relations in the two periods. Look at the network graph for the red years of 1919-1920. This is a typical “star” network. The police are at the center, sometimes the target of violence, but, mostly, the subject of violence – a well known finding in social movement research. The police are violent mostly toward the left: the socialists, the protestors, people in the crowd. Now, look at the network graph for the black years. The fascists are now the perpetrators of violence on the same actors as the police (socialists, workers, communists, demonstrators). However, this is not a star network because there are two centers of violence, arther than one. Notice how thin the line going from the police to the fascists is. The Italian state had a terrible responsibility in letting the fascists run undisturbed. The government thought it could use the fascists to curbe working-class militancy, and then control the fascists but … in the end, it didn’t go as planned and it all got out of hand.

This graph coincides with the factory occupation movement of September 1920, a nearly revolutionary situation that lasted one month. The workers across Italy occupied all the factories. Lenin was telling the Italian socialist leadership: “this is it! You’ve got it! Take over!” Prime Minister Giolitti played a card that was a political masterpiece (at least, in the short run; but it ultimately cost Italy 20 years of dictatorship). Giolitti thought the socialists would act “by the book,” as written in the *Communist Manifesto*, the revultion being the result of “the arm struggle with the state,” a frontal clash with the army and the police. “If I don’t send in the police and the army, they won’t know what to do.” That,

indeed, turned out to be the case. A month after the factory occupation movement started, the workers returned the factories to the employers. Notice how there is absolutely no violence by the police during this period. The police did not intervene. But look at the next network graph for the last quarter of 1920 (October through December): right after the factory occupation movement, the working class came under fire from both the fascists and the police. Counter-mobilization had started in earnest.

May 15, 1921. National elections in Italy. As you can see from this network graph, the fascists really stepped up the level of violence. From then on, they did not stop until 1922, when they took power. The only time they stopped, or at least slowed down, was in August 1921; as every tourist who has been on vacation in Italy in August knows, Italy shuts down for vacation in that month and it seems that even the fascists went on vacation in August 1922!

What does this tell us? First, the results show that there is an answer to the question "What's Franzosi gonna do with thousands of words in the computer?" and that the answer to the question has historical and substantive import. Methodologically, the approach to content analysis that I have illustrated (a computerized "story grammar") has both pros and cons. On the side of pros, coding is based on a more natural linguistic approach to text, as opposed to the theoretically-defined and abstract coding categories of traditional content analysis. As a result, coded data has greater reliability. Since coded output preserves the narrative flavor of the input text, linguistic rules of semantic coherence can be applied to data quality: coded output must make sense to any "competent user of the language." Finally, by preserving the narrative sructure of the input text, a story grammar approach to coding delivers data that are nearly hypothesis-free: all narrative information is coded within the categories of a story grammar, regardless of pre-specified theoretical hypotheses. On the negative side, a story grammar approach to text works well with narrative texts only. And like traditional content analysis, the approach is very labor intensive.

To conclude with a bit of advertising … I have published extensively on this linguistics approach to content analysis, but I have a book titled *Quantitative Narrative Analysis* to be published in 2009 by Sage (one of those books with a green cover in the popular series Quantitative Application in the Social Sciences).