Socio-Legal Analysis of Criminal Sentences: A Preliminary Study

G. Giura and **G. Giuffrida** and **C. Pennisi**
Dipartimento di Sociologia e Metodi delle Scienze Sociali
University of Catania, Italy

C. Zarba
Neodata Intelligence
Catania, Italy

Abstract

This paper discusses a research based on analyzing criminal sentences on criminal trials on organized crime activity in Sicily pronounced from 2000 through 2006.

Large criminal sentences related dataset collection activity in Italy is severely constrained for various reasons such as difficulty of data collection at the courthouses, unavailability of data in digital format, and classification criteria used in the public archives. Thus, in general, judicial statistics suffer from lack of reliability and informativeness.

The objective of this research is to analyze the text of criminal sentences in a revisable and verifiable way, so that information is extracted on the trial leading to the sentence, the socio-economic environment in which the relevant events occurred, and the differences between the various districts conducting the trials. The purpose is to elaborate a tool of automated analysis of the text of the sentences that is generalizable to other areas of jurisprudence, and, outside of jurisprudence, to other temporal and geographical contexts.

The 726 criminal sentences that have been converted into text files have been pronounced at all judicial levels in the four Sicilian districts for mafia-related crimes.

This research is relevant because, for the first time in Italy, we aim to empirically describe the juridical response to the phenomenon of organized crime, by using a large and extendable database of criminal sentences that can be analyzed with data mining techniques, rather than deriving general conclusions from a focused small set of sentences.

Introduction

There are two main sources of criminal, judicial, and court statistics in Italy: crime notifications of law enforcement authorities to judicial authorities and trial proceedings. Since not all crime notifications lead to a trial, the statistics from the former source are more numerous than the statistics from the latter source.

These statistics do not encompass all crimes, because not all crimes are reported to law authorities. This is particularly true for small crimes. However, for major crimes the discrepancy between unreported and reported crimes is considerably reduced. These major crimes are usually committed by organizations of people pursuing a specific activity (e.g., the illicit traffic of narcotic drugs) or the interests of the criminal organization itself. For these major crimes, a consistent amount of data is available.

Currently, these statistics are used in two ways. Local authorities use them as a tool to prevent and repress the criminal activity, while national authorities use them to assess the efficiency of all law offices involved. Unfortunately, these usages have the drawback of preventing a complete and accurate analysis of the social, political, and institutional aspects of the phenomenon of organized crime. The purpose of this research is to address this drawback by exploiting automatic knowledge extraction procedures based on text mining tools.

The research consists mainly of two phases. In the first phase data has been collected from the various courthouses in Sicily. They have been scanned through an OCR in order to convert them into a digital format analyzable through automatic techniques. Furthermore, social scientists have identified a *codebook* which is basically a collection of well thought variables to be devised from the text of each sentence.

The second phase, done in tight collaboration with computer science experts, consists of designing an automatic algorithm that gradually reconstructs the internal structure of the sentence and makes the extracted data available for querying and more sophisticated analysis such as Network Analysis. Thus, the extracted information has to be properly organized in a relational database for easy retrieval. The main challenge of the second phase is the construction of a *parser* that is able to understand the complex legal language of the sentences and extracts all relevant pieces of information from it.

Collection of criminal sentences

This research consists in the analysis of criminal sentences satisfying the following criteria: (1) the crimes are those mentioned by the Italian criminal procedure code, article 51, comma 3 bis—which basically encompasses all Mafia related crimes, (2) all sentences are final and irrevocable for at least one defendant, and (3) all sentences have been pronounced by Sicilian judicial authorities from January 1, 2000 to December 31, 2006.

In order to collect the criminal sentences required for this research, a preliminary interrogation to the Italian computer-

Authority	Standard	Abbr.	Plea	Total
GIP/GUP	23	135	118	276
Tribunale	122	9	7	138
C. d'ass.	90	10	59	161
C. d'app.	5	1	0	6
C. d'ass. d'app.	71	4	70	145
Total	311	159	254	726

Figure 1: Classification of the analyzed criminal sentences satisfying the criteria of the research. The rows indicate the degrees of judgments in the Italian judicial system (GIP/GUP, Tribunale, Corte d'assise, Corte d'appello, and Corte d'assise d'appello), while the columns indicate the proceedings formats (Standard; *Abbreviato*: a summary trial that omits the formal trial stage and reduces the penalty by one third; *Plea, applicazione pena su richiesta delle parti*: a kind of plea bargain where the accused and the prosecutor reach an agreement about the penalty—sentence to apply—without formally admitting guilt).

ized archive RE.GE[1] has been performed. This interrogation has been parameterized with the criteria of the research and produced a list of about 1,200 criminal sentences satisfying our query. Then a formal request to the applicable Sicilian courthouses has been made in order to gain physical access to the criminal sentences. After the authorizations were granted, we went physically to the various Sicilian districts and xeroxed the criminal sentences.

After we collected 1,147 sentences we performed an extensive manual data quality verification process. We then realized that due to misleading classification in the RE.GE. archive just 726 sentences satisfied all our criteria.

Afterwards, OCR technology was used in order to convert the paper copies into electronic files.

The entire process just described was extremely time consuming as, overally, it took more than two man-years. The collected sentences can be classified according to degree of judgment and proceedings format, as shown in Figure 1.

The codebook

While analyzing the collected criminal sentences, a *codecook* has been specified as a tabular tool to organize the results of the analysis. The codebook consists of a table that contains one row for each criminal sentence. The columns denote classificatory variables that describe the features of the criminal sentences that are important for the analysis (Bruschi 2005).

Variables in the codebook have been apriori defined by the social scientists with some analytical goals already in mind. Each sentence has been manually analyzed by the social scientists in order to fill in the codebook. Since the specification of the codebook has been gradually refined while the sentences were analyzed, many sentences had to be reanalyzed several times. While this manual approach is satisfactory to meet the designed goals, it definitively limits additional analysis based on different variables. Thus, an automatic process to harvest variables from free text becomes even more imperative in order to avoid long repetitive manual processes.

At the end of the analysis, the final codebook consisted in 44 variables belonging to 4 dimensions: *temporal*, *procedural*, *social*, and *environmental*.

The variables of the temporal dimension describe the durations of each phase of the trial process, starting from the registration to the RE.GE., including the pronouncements of the verdicts at each degree of judgement, and terminating with the final declaration of irrevocability of the sentence. Therefore, the total duration of the trial process can be ascertained.

The variables of the procedural dimension describe the legal events occurring during the temporal dimension. These events include custody measures, proceedings formats, contested crimes, modifications and integrations of contested crimes, recognition of extenuating circumstances, and the final verdict.

The variables of the social dimension describe the occupation or profession of the defendants, as well as their social and economic conditions.

The variables of the environmental dimension describe the geographic, political, institutional, and political aspects of the events discussed in the sentence. They also identify the economic sector that is harmed by the contested crimes, and report the official quantification of the economic cost suffered because of the contested crimes.

Merits and limits of the research

Traditional research (De Felice 2007; Mosconi and Padovan 2004; Mazzette 2006; Olgiati 1996; Raiteri 2009) focuses on the analysis of all official documents involving a limited number of criminal sentences. Our approach to this research domain, instead, considers only the final document—the text of the criminal sentence itself—for a large number of criminal sentences. Thus, this research has the merit to widen the scope and variability of the data analyzed.

Other research (Asmundo and Lisciandra 2008) has attempted to estimate the cost to society induced by the phenomenon of organized crime, but the method used for the estimation is not yet satisfactorily standardized. This research, instead, considers the official cost as reported in the criminal sentences. Thus, this research has the merit of accurately estimating at least the cost that has been officially ascertained.

The data collection necessary for our research faced many challenges. For instance, sentences are not generally available for research purposes. Furthermore, there is not a central archive for all sentences but they are scattered around various courthouses and each one of them may exhibit different policies to access their archive.

On top of this we need to consider the time required to xerox them and to scan them through an OCR. Not to mention the time spent to travel to the various courthouses and the time spent in the waiting rooms.

All together we collected 1,147 criminal sentences. They range from 2 to 3,268 pages for a total of more than 55,000 pages to be xeroxed and scanned. Also, as we already men-

[1]Registro Generale Notizie di Reato.

tioned, not all sentences were relevant for us due to misclassification in the RE.GE. archive. So we had to go thorugh each one of them to figure out manually whether or not it satisfied our criteria. This left us with a total of 726 criminal sentences. All this posed severe limits to this research and led to more than 30 months spent by the authors to collect all data in a proper form.

We believe that such challenges are the main reason why researchers so far have not approached the problem like we are doing. To the best of our knowledge, this is the first time in Italy (and perhaps in many other countries) someone is undertaking such research avenue.

Information Extraction

Information extraction is the process of extracting structured data from unstructured data such as the extraction of relational data from natural language documents (Sarawagi 2007). Typically, given a document written in natural language, there are four kinds of information that can be extracted: *entities*, *attributes*, *relations*, and *events*.

Entities can be individuals, things, dates, or measurements. Attributes are features associated to entities. For instance, an individual has attributes like birthdate, birthplace, profession, education title, telephone number, email address. Relations are associations between entities. Events are relations where time is of primary importance.

There are two main approaches to information extraction: *deep* and *shallow*. Deep information extraction is based on natural language processing. Information is extracted from the document by lexical analysis, semantic analysis, and interpretation of the discourse (Humphreys, Gaizauskas, and Azzam 1997). Deep information extraction is quite effective, but too slow computationally. Furthermore, the (manual) construction of the model necessary to carry out the interpretation of the discourse is complex and laborious.

Shallow information extraction does not aim at a human-like comprehension of the document, but aims only at the filling of the relational tables. This is done using a pipeline consisting of a finite number of finite state transducers (FSTs). A finite state transducer takes a sequential input and, if some conditions are verified, returns an output that depends on the input and on the internal state of the transducer (Brin 1998). Essentially, a finite state transducer performs a simple linguistic task. The idea is that a finite number of simple linguistic tasks is sufficient in order to fill the relational tables.

Automated analysis of criminal sentences

In this section we discuss the case of information extraction in the domain of criminal sentences.

We wish to analyze the corpus of text documents, extracting from the free text of the sentence all entities, attributes, relations, and events that can be useful for social studies.

Standard structure of an Italian criminal sentence

The standard structure of an Italian criminal sentence is described in the following. The sentence always starts with the denomination of the legal authority, and the wording "Repubblica Italiana", followed by "In nome del popolo italiano". The names of the members of the court follow. The first name we encounter is always that of the judge. The other names are the assistants. Then another section starts where the defendants are listed. By the law, each defendant has to be properly identified by his/her biographical data such as birthplace, birthday, etc. In the criminal sentence, the defendant name is always followed by the wording "nato a" (i.e., "born at"). The name of each defendant is followed by the name(s) of the defending lawyer(s). The name of each lawyer is preceded by the title "avv.". The first name that is not preceded by "avv." and not followed by "nato a" indicates the end of the defendant list, and this name is an involved part in the events discussed by the sentence (for instance, it could be an injured party or a witness). The verdict of the sentence is always preceded by the acronym "PQM" or "PTM". For first-degree sentences (GIP), each defendant is either condemned or absolved. In second-degree sentences, before the acronym PQM and after the defendant list, the first-degree sentence is described. Then, after the acronym PQM, it is explained how the first-degree sentence is modified.

The sentence analyzer

We have implemented a first version of an analyzer capable to perform this information extraction task. Given a sentence, our analyzer extracts the following entities: judge, assistants, defendants, lawyers, other people involved, crimes. Furthermore, our analyzer extracts, for each defendant, whether he/she is condemned or absolved. Finally, it extracts, for each defendant, the lawyer(s) that represent him/her.

The extraction is performed by means of a pipeline of finite state transducers that exploits the standard format of a criminal sentence. The finite state transducers we developed so far are the following.

`People-FTS`. This transducer uses a dictionary of Italian names in order to recognize individuals. If necessary, the user can extend the dictionary. An individual is considered as a sequence of at least two names, or as a capital letter followed by a point, a space, and a name.

`Defendants-FTS`. Each defendant is always accompanied by its birthplace and birthday. If an individual is followed by the wording "nato", then we assume that he/she is a defendant.

`Lawyers-FTS`. Lawyers are always preceded by their title, possibly abbreviated.

`Judge-FTS`. If at this point the first individual appearing in the text of the sentence is not a defendant, then it must be a judge. If the first individual is a defendant, then the information about the judge is unavailable.

`Assistants-FTS`. If the name of the judge has been extracted, then all individuals comprised between the judge and the first defendant must be assistants.

`Other-FTS`. At this point, all individuals that are not the judge, assistants, defendants, or lawyers, are categorized as "other people involved".

`Defendants-lawers-FTS`. This transducer associates each defendant to the list of lawyers that represent him/her.

`Crimes-FTS`. This transducer recognizes crimes disputed in the trial using regular expressions.

`Verdict-FTS`. This transducer attempts to deduce if a defendant has been condemned or absolved. This is done analyzing the text of the sentence following the acronym PQM or PTM, and looking for words such as "condanna" or "assolve" written before the name of the defendant.

Conclusion

The automatic comprehension of thousands of criminal sentences is clearly a long and complex work. The main problem is the disambiguation of different meanings for similar wordings of the Italian legal language. The project requires several refinements, each one originating from the errors done by automatic analyzer. Clearly, this is a laborious process that however diminishes with time, as the various FSTs become more and more accurate.

Another topic of research concerns the precision of the extracted information. In particular, one should define the tolerance of the errors in the extracted information. Some errors are absorbed by the subsequent statistical analyses; other are less acceptable. More attention should be dedicated to this topic.

The problem of the representation of the structured data has not yet been completely defined. The final representation will be a function of the various analysis tasks that one wishes to perform on the extracted tables. However, regardless of the final form of the structured data, the information extraction task currently performed is necessary.

A problem of this research is given by the frequent errors done by the OCR software. The collected criminal sentences contain not only typed text, but also some handwriting, and the OCR is unable to satisfactorily process the handwriting. Future progress of this research will therefore need to use more sophisticated OCR tools, specifically tailored to the kind of criminal sentences that need to be analyzed.

A further topic for future research is the creation of a social network from the corpus of collected sentences. The network should contain nodes for the actors, and edges that describe the social relationships between the actors. For instance, a "defendant" node should be connected with appropriately labelled edges to all "lawyer" nodes that represented the defendant in some trial.

Concluding, the problem is quite complex, and we will certainly encounter further challenges as this research progresses. Nonetheless, this project is very innovative and of great interest for both the computer science and social science communities.

References

Asmundo, A., and Lisciandra, M. 2008. I costi dell'illegalitá. In La Spina, A., ed., *Un tentativo di stima del costo delle estorsioni sulle imprese a livello regionale: il caso Sicilia*, 113–136. Il Mulino.

Brin, S. 1998. Extracting patterns and relations from the world wide web. In *International Workshop on Web and Databases*.

Bruschi, A. 2005. *Metodologia delle Scienze Sociali*. Laterza.

De Felice, D. 2007. *La costruzione istituzionale dell'interesse del minore*. Giuffré, Milano.

Humphreys, K.; Gaizauskas, R.; and Azzam, S. 1997. Event coreference for information extraction. In *Workshop on Operational Factors in Pratical, Robust Anaphora Resolution for Unrestricted Texts*.

Mazzette, A. 2006. *La criminalità in Sardegna. Reati, autori e incidenza sul territorio*. Centro di studi Urbani D.E.I.S. Unidata.

Mosconi, G., and Padovan, D. 2004. *Processo penale e costruzione sociale del delinquente*. Il Mulino, Milano.

Olgiati, V. 1996. Foreign courts: civil litigation in foreign legal cultures. In *Gessner V. Aldeshort*. Brook field, Vt., USA: Dartmouth.

Raiteri, M. 2009. Criminology and victimology: Are we teaching social sciences or not? In Sette, R., ed., *Cases on technologies for teaching criminology and victimology*. Hershey- New York, IG Global.

Sarawagi, S. 2007. Information extraction. *Foundations and Trends in Databases* 1(3):261–377.