Methods of digital access for legal language documentation

Paola Mariani*, Costanza Badii*

* Istituto di Teoria e Tecniche dell'Informazione Giuridica
Consiglio Nazionale delle Ricerche
Via Panciatichi n. 56/16, Florence, Italy
[mariani, badii]@ittig.cnr.it

Abstract

For many years the Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG) of the Consiglio Nazionale delle Ricerche has studied the evolution of legal language, creating databases for documentation and digital retrieval of law texts. The ITTIG is attending to document legal language through information technology in order to provide as wide an access as possible to its findings. The Institute has recently created an on-line digital database that includes the full text of the most important Italian laws (Codes and Constitutions) from the 16th to the 20th century. The ITTIG is also in the process of preparing another database made up of contexts from the original 10th to the 20th century legal sources.

1. The characteristics of legal language

Legal language is the product of centuries of semantic reworking of natural language. Through this process of re-elaboration legal language has become a technical terminology inserted within the structure of natural language (Scarpelli, 1969).

As a consequence of this it pertains to both society as a whole as well as being very much a technical discipline of one specific sector.

This is even more the case if one takes into consideration the variety of law sources (such as legislation, jurisprudence, legal science), each of which has a distinct set of recurring linguistic patterns. Similarly, there are a multiplicity of legal-documentary acts (such as laws, treaties, contracts, wills, etc.) that are all articulated through their own specific terminology (Mariani, 2002).

In particular, the specifically technical character of legal language is ordered via the distinguishing of:

- specific technical terms (e.g. "anatocismo", ["anatocism"] understood as compound interest) which tend to assume a univocal meaning and which are exclusive to each particular sector and do not occur outside it;
- redefinitions (e.g. "confusione", ["confusion"] understood as a way in which obligations are annulled) that give normal language terms meanings that are different to those usually associated with them;
- collateral technical terms (e.g. "escussione del teste" ["examination of witness"]) that can be defined as particular stereotypical expressions that, while not absolutely necessary for scientific rigor, are used for a series of technical reasons.

In all cases, while it is true that they have a specific and autonomous value single terms must necessarily be taken within their natural context.

In our opinion the context is fundamental because a single term assumes a different semantic meaning precisely in relation to the legal context in which it is used.

Legal language therefore reflects the characteristics of its object: if on the one hand the *scientia iuris* has a specific content and technical character on the other it is deeply embedded within social reality[1].

Law and language, in fact, have common characteristics in that they are both intrinsically linked to the social environment in which they are inserted (Fiorelli, 1993-97). They are systemic realities that are modeled and evolve through the needs of social reality (and by social reality the totality of the specific socio-economic-political elements that characterize a particular epoch and a particular historically and geographically determined environment is intended)[2].

It is precisely this nexus with society that makes law and language so similar in their most intrinsic characteristics.

It is important in fact to point out that legal language has a tendency to preserve itself and to repeat practices that have come to be codified through use or by law and that are expressed in terms that acquire a particular meaning so as to respect the certainty of law.

At the same time, however, the natural nexus with society requires legal language to define newly formed institutions and/or paradigms.

For example, the offense of illegally accessing a computing or telecommunications system was inserted into the Italian Penal Code, 1930 (which was drawn up in the Fascist period) in 1993 (law n. 547/93) because the development of new technologies had made such a law necessary. It was introduced as an offence against the person alongside traditional offences such as illegal entry or the violation of correspondence privacy.

[1]Conceptual terminology is defined as follows: "language" signifies the communicative tool, the institutionalized oral and written system with which verbal communication is articulated; "terminology" is the way in which a particular language is expressed; "lexicon" means the whole terms that are typically associated with a particular discipline.

[2]In this context it is worth quoting the German Jurist Georg Friedrich Puchta: "*Das Recht ist eine gemeinsame Überzeugung (...) Die Entstehung eines Rechtssatzes ist daher die Entstehung einer gemeinsamen Überzeugung, welche die Kraft in sich trägt, das, was sie als Recht erkennt, zur wirklichen Ausfürung zu bringen*". Puchta G. F., 1872, *Pandekten*, Leipzig-Barth edition.

As a consequence of this language must necessarily elaborate a terminology able to define institutions that are in the process of emerging.

Historical-semantic analysis deals with the interpretation of the meaning of legal concepts including through the study of the origin and the historical development of language with the help of advanced systems of digital documentation that allow the storage of a huge amount of information.

In this sense deepening our understanding of legal language is necessary if one considers the fact that it is precisely language that allows concepts to circulate and to be communicated, and that lies at the base of other kinds of knowledge development [3].

Linguistic analysis is in fact the precondition of all interpretative activity that each jurist cannot but tackle in the course of his or her work (Belvedere, 2000).

In this context paragraph 1, article 12 of the preamble of the Italian Civil Code states: "In applying the law no meaning can be attributed to its wording except that which was its intended meaning according to its context and the legislator's intention."

Attributing meaning to normative statements therefore entails the making of precise choices at the syntactic and semantic level that the jurist cannot avoid if he or she wants to pin-point the underlying *ratio* behind a particular law. At the same time this must be done within the context of certainty of law which remains one of the fundamental tenets of our legal system (Bobbio, 1994).

Today, in this kind of research, information technology is absolutely indispensable.

In fact, due to the existence of a huge number of legal documents and sources that need to be classified and retrieved (something that has always pushed traditional information retrieval systems to their limit), the legal field has been affected particularly positively by the changes brought about by the use of multimedial technology.

This has decisively changed the methodology of communicating and distributing legal resources.

At the linguistic level the modalities in which data is presented has changed. Analytical tools and procedures have changed and now allow the creation of exhaustive linguistic *corpora*, representative of own legal system.

Using information technology brings together elaboration and information retrieval techniques and offers new opportunities to the user. In an electronic archive all documents can be singled out in their actual physical shape seeing as their content is reproduced in image format.

This is even more the case in the restoration and safekeeping of the whole historical-linguistic archive which would otherwise be subject to progressive and inevitable decay.

Historical legal texts in particular are often in a poor state of conservation and require their digital restoration if their content is not to be lost (Badii, 2003).

This is particularly pressing at the European level where it is becoming ever more necessary to compare separate legal systems in order to find a common, supra-national legal platform.

Comparing the legal language of the different European states is, in fact, necessary in order to extract legal concepts that are not always rendered intelligible by literal translation, and particular attention should be paid to those languages[4] that have assumed a preponderant position in the global and European arena or to the specific realities of institutionalized bi- or multi-linguism (Rega, 2000)[5].

There is no permanent correspondence between words and concepts in all languages and this is the principle cause of translation problems and the risk of losing original conceptual terminological meaning (that is specific concepts often end up being "forced") (Sacco, 1994).

This is even more the case if one takes into account the influence that historical models have on a given legal system and consequently on its corresponding lexical *corpus*. As far as Italy is concerned, the reception of the French exegetical model first, and the pandectistic German one later has implied a kind of linguistic compliance in Italy to the legal categories developed in those countries that has had profound effects on the characteristics of the Italian legal system.

2. Digital documentation methodologies elaborated by the ITTIG

For years the ITTIG has been dedicating itself to legal language and to the study of information-access methodologies in this specific sector.

Their scope has been the analysis of the language of national law but they are nevertheless applicable to other European languages

What counts in fact is the study of techniques and methods of information-access that guarantee the retrieval of documents whatever their content, and that has the possibility of diversification with respect to the legal culture of other countries.

[3] In our opinion, alongside historical-semantic analysis, other ways of understanding legal language are also applicable: e.g. philosophical analysis, that concerns the philosophical implications of the discipline; structural analysis, developed in legislative drafting and legimatics, deals with the structure of laws and tries to tackle the chaotic nature of legislative production as well as to render the drafting of legal texts an automatic process.

[4] English has invaded all disciplinary sectors including law; as an example one can cite the recent atypical contract types that have become common in recent years (e.g. outsourcing, factoring, leasing, joint venture). These are described with untranslated English terms as they have become part of linguistic practice in all countries.

[5] We are referring to the bi-lingual experience of the Alto –Adige [South Tyrol] which attempts to deepen an understanding of the relationship between Italian and German in order to create a more uniform and comprehensive legal-linguistic platform. *Linguistica giuridica italiana e tedesca*, 2000, Sezione di Traduttologia (*Übersetzungswissenschaft*), Unipress, Padova, pp. 449-500.

This activity has led to the creation of on-line digital archive *Lingua Legislativa Italiana* (LLI)[6] [*Italian Legislative Language*]. This is a database made up of a *corpus* containing a vast number of legislative texts memorized as full-text.

Within this database searches can be conducted not only by word heading but also by all its possible grammatical variants.

The *corpus* contains circa 190 primary legislation texts in their first and official edition, such as codes, constitutions, consolidated law panning across a huge time-span (1539- to the present)[7].

These texts were chosen not only because of their importance in the history of Italian law but also because their authority was superior to other laws and because they have had a profound influence on Italian legal language.

LLI is therefore of great importance in the historical-legal-linguistic sector and constitutes a vital consultation tool for legal language via its most important texts.

The *Lessico Giuridico Italiano* (LGI) [*Italian Legal Lexicon*], which has recently been made available on-line[8], constitutes another access point to documentation on legal language experimented by the ITTIG. It complements the primary legislation of LLI with a variety of other legal sources.

This archive contains circa 2000 legal-historical documents published between the 10th and the 20th centuries and is made up of legislative, legal science and legal practice. The latter includes judicial sentences, notary letters, wills etc.[9]. Legislation in this archive concerns those aspects of ordinary and day to day law creation tied to the particular and contingent needs that, if on the one hand required immediate legislation (e.g. decrees, ordinary laws, proclamations, edicts) were nevertheless integrant the linguistic and conceptual framework envisaged by primary legislation.

Given the huge number of legal documents and the vast time-scale covered, it was necessary to make a careful selection of the sources to be included. Those that had had a significant impact on the history of law were chosen but temporal and geographical coverage was also of primary importance. Within the selected texts the parts, and indeed the words that have been of particular significance for legal language have also been highlighted. From this documentary base the Institute researchers have created circa 900,000 "source-cards". These are image format reproductions of the single selected documents.

LGI therefore contains the digital images derived from the selected sources as well as a whole series of bibliographical and lexical references. The result is direct access to the object-heading required within its documentary context[10]. The archive also contains non-Italian head-words that have been found in the Italian documents selected. These are prevalently Latin terms that crystallized in legal language, for example "*ab intestato*", "*mortis causa*", "*a latere*", but there are also terms in Greek, English, French, German, etc.[11].

The relevance of this archive is also made evident by the fact that the terms it covers appertain to various legal fields (civil, penal, commercial, canon etc.) and this makes it of interest to a wide variety of users. The geographical area covered does not just include Italy's present territorial boundaries but all those areas that in one way or another were impacted by Italian legal systems or language (for example Switzerland or Malta).

In every field it is possible to select the information required and access the digital "source-cards" with correlated bibliographical and lexical data. In fact, from the strictly operative point of view LGI is made up of the following information units on which the multiple querying methods have been structured:

- Entry;
- Language, if the user is interested in the non-Italian terms used in Italian sources;
- Date, if one is analyzing the historical development of a term from the lexical or semantic point of view, that is when the term first appeared or if it has fallen into disuse;
- Author or title of document;
- Lexical variants of the principle entry. This allows for an understanding of a term's development over time and on the basis of the geographical area in which it has been located;
- Legal locution/syntagm in which the single words are given meaning in their particular legal connutation (e.g. "responsabilità contrattuale", ["contractual liability"] "usucapione abbreviata" ["shortened usucapion"]). In fact it is often precisely within a syntagm that terms assume meanings relevant to law.

The latter two query modes were implemented later in order to give the user other information for search orientation.

As an example let us take the term "contingente" ["contingent"] as it has various meanings that are documented in different contexts.

The archive's consultation, in fact, allows not only to understand the term's use in its specific context but also the semantic value of single terms.

From an analysis of the "source-cards" present in the archive one can see that the term "contingente" has more than one meaning:

1. Masculine adjective referring to a particular and contemporary event;
2. Masculine substantive signifying "tax";

[6] www.ittig.cnr.it/BancheDatiGuide/lli.

[7] The number of texts will rise because new ones are continuously being added by the ITTIG researchers working on the project.

[8] http://w3.ittig.cnr.it/vocanet. The archive is currently being tested by ITTIG technicians and will in any case be corrected and updated.

[9] The first occurrence is in Placito di Capua (960). An Italian testimonial formula, one of the earliest examples of the Italian vulgar, is inserted within the main Latin text.

[10] Reproduction techniques have changed over the years. From magnetic perforation, which allowed the automatic alphabetical ordering of entries, to the scanner.

[11] Note the German term "*Rechtsgeschäft*", from which the Italian term "negozio giuridico", which is fundamental to the Italian legal system, is derived.

3. Masculine substantive signifying "a detachment of troops". In this third meaning the phrase "contingente di leva" ["conscript detachment"].

To give another example: the syntagm "common law". Consulting the archive we can see that in Italian legal science sources this expression is encountered in the period between 1874 and 1955.

From the operative point of view activity carried out allows for a multiplicity of results in terms of access to documentation, these are: the subdivision of the archive into the three fundamental legal sources (legislation, legal science and practice); the elaboration of single information units implemented with their lexical variants and legal syntagms; the correction and/or integration of data.

The elaboration of homogenous and coherent classificatory criteria for each type of entry (a fundamental requirement whenever the aim is to conserve the scientific quality of data) is particularly important given the size of the *corpus*.

Due to the fact that the archive was created by different people the elaboration of standard criteria was and still is necessary: rules envisaged must be rigorous but at the same time must be flexible enough to deal with changes and the evolution of language.

The positive results obtained so far have challenged the ITTIG to carry out another ambitious project: the *Indice ragionato della lingua giuridica* [*legal language subject Index*]

This is conceived as a kind of digital dictionary of legal language that provides documentation for a historical view-point via the evidence contained in the sources.

On the basis of the results obtained by LLI and LGI it is hoped that the Index will be created semi-automatically in order to rationalize efforts and workloads.

The first operation undertaken was the study of the editing reference path for each entry and of the software required for its creation.

The editing phase is divided into two distinct parts: an automatic one that allows the recouping from already created digital archives of the largest amount of data possible[12]; a part that requires a high level of historical-legal analytical in order to:

- Single out the meanings relative to each entry derived from the source references;
- Select the "source-cards" that documents corresponding meanings, with the bibliographical data[13].

The result would be unprecedented in the historical-legal-linguistic sector and its effects would be felt over a long period.

In Italy, in fact, up to now, apart from in particular sectors, a historical-legal dictionary able to document the evolution of legal language, does not exist.

[12]The documentary units to prepare automatically are e.g. the language, the date, lexical variants, etc.

[13]As far as legal practice is concerned this operation has been completed.

3. References

Badii, C., 2003. *Firenze e il (Neo-)Umanesimo: un prezioso spunto di riflessione nel Terzo Millennio*, review of the *Firenze e il (Neo-) Umanesimo; arte, cultura, comunicazione multimediale all'alba del Terzo Millennio*, international conference, in www.ittig.cnr.it.

Belvedere, A., 2000. *Pragmatica e semantica nell'art. 12 Preleggi*, in *Linguistica giuridica italiana e tedesca*, Unipress, Padua

Bobbio, N., 1994. *Scienza del diritto e analisi del linguaggio*, in *Il linguaggio del diritto*, Edizioni Universitarie di Lettere Economia e Diritto, Milan.

Fiorelli, P., 1995. *Gli archivi selettivi del Vocabolario giuridico italiano*, in Atti del Convegno Idg 1-3 dicembre 1993, *Verso un sistema esperto giuridico integrale. Esempi scelti dal diritto dell'ambiente e della salute*, Ciampi C, Socci F., Taddei Elmi G. (Eds), Cedam.

Fiorelli, P., 1993-97. *Premessa* all'*Indice della lingua legislativa italiana. Inventario lessicale dei cento maggiori testi di legge tra il 1723 e il 1973*, Mariani Biagini P. (Ed.), Florence, vol. I, pp. V-XII.

Mariani Biagini, P., 2002. *Informatica e lingua del diritto*, in *Lineamenti di Informatica giuridica. Teoria, metodi, applicazioni*, Nannucci R. (Ed.), ESI.

Mortara Garavelli, B., 2001. *Le parole e la giustizia*, Piccola Biblioteca Einaudi, Turin.

Mortara Garavelli, B., 2002. *Persistenza del latino nell'uso giuridico odierno*, in *L'Accademia della Crusca per Giovanni Nencioni*, Casa Editrice Le Lettere, Florence.

Palazzolo, N., 2002. *Informatica e storia del diritto*, in *Lineamenti di Informatica giuridica. Teoria, metodi, applicazioni*, Nannucci R. (Ed.), ESI.

Rega L., 2000. *Aspetti e problemi della traduzione delle formule di rito nell'ambito testuale italo-tedesco*, in *Linguistica giuridica italiana e tedesca*, Unipress, Padua.

Sacco, R., 1994. *La traduzione giuridica*, in *Il linguaggio del diritto*, Edizioni Universitarie di Lettere, Economia e Diritto, Milan.

Scarpelli, U., 1969. *Semantica giuridica*, in A. Azara / E. Eula, *Novissimo Digesto Italiano*, vol. XVI, Utet, Turin.