METHODOLOGY FOR BUILDING A TEXT-STRUCTURE ORIENTED LEGAL CORPUS

Cristina ONESTI, PhD
Dipartimento di Scienze Letterarie e Filologiche, University of Turin
Via Sant'Ottavio, 20 - 10124 Torino, Italy
cristina.onesti@unito.it

Abstract: The present study aims at showing the methodology for building a legal corpus with a special attention paid to the internal structure of legal documents and juridical texts. Built at the University of Turin, the Corpus *Jus Jurium* tries to cover the entire legal universe current in contemporary Italy, whose life is represented from their first conception in the parlamentary discussion, to their codification in normative rules, to their application in judgements.
The Corpus *Jus Jurium* will be lemmatized, POS-tagged and have added a textual markup, casting some light on the still neglected textual expression of legal and juridical texts, which is decisive for every national legal system.
Elaborating texts with such peculiarites implies a long amount of manual work. However, the final result can be an extremely useful resource for translators looking for idioms, collocations or terminological elements in specific parts of texts and for forensic linguists, providing them with an extensive repository of well-structured data and with fine-grained querying opportunities, whether at the morphosyntactic or lexical or textual level.

COSTRUIRE UN CORPUS GIURIDICO: OSSERVAZIONI METODOLOGICHE PER L'ANALISI DELLE STRUTTURE TESTUALI

Astratto: Il contributo presenta la procedura di creazione di un corpus giuridico che mira a rappresentare l'intero universo legale attualmente corrente in Italia, illustrandone in particolare la metodologia utilizzata per il markup testuale dei dati.
Il Corpus *Jus Jurium*, in fase di realizzazione presso l'Università di Torino, vuole superare le caratteristiche di un tradizionale database giuridico grazie a finalità precipuamente linguistiche: il corpus, infatti, è in corso di lemmatizzazione, etichettatura per parti del discorso e prevede un robusto markup testuale e diplomatico. Tra le sue finalità, in particolare, è proprio quella di poter interrogare in modo "ricco" i documenti, intersecando la loro definizione diplomatica con il loro assetto linguistico e testuale. Tale strumento costituirà auspicabilmente un'utile risorsa anche per traduttori giuridici e linguisti forensi.

METODOLOGIA TWORZENIA KORPUSU TEKSTÓW PRAWNICZYCH UWZGLĘDNIAJĄCEGO STRUKTURĘ TEKSTU

Abstrakt: Praca ma na celu pokazanie metodologii konstruowania korpusów z uwzględnieniem wewnętrznej struktury dokumentów sformułowanych w języku prawnym i prawniczym. Autorka analizuje prace nad korpusem *Jus Jurium* tworzonym na Uniwersytecie w Turynie, który z założenia ma zawierać wszelkie teksty prawne i prawnicze tworzone współcześnie we Włoszech. W przypadku takich korpusów pojawia się konieczność lematyzacji, wprowadzania tagów i innych znaczników tekstu. Należy tutaj podkreślić, że taki korpus będzie stanowił niezmiernie przydatne narzędzie dla tłumaczy poszukujących kolokacji, idiomów, związków frazeologicznych czy terminów.

1. Introduction and objectives

Linguistic corpora are fundamental for analyzing juridical and legal texts as repositories of authentic language data, especially considering that law is, above all, a matter of language, as Cortelazzo (1997, 39) has clearly pointed out: *"Il diritto non si serve della lingua, ma è fatto di lingua"*, i. e. law does not use the language, but it is 'a fact of language'.

The present paper aims at showing the construction of an Italian legal corpus, the *Corpus Jus Jurium*, with its different realization phases and its peculiar attention to the internal and textual structure of legal documents and juridical texts.

This contribution will be divided into a) a first introductory part; b) a section concerning the collection of data; c) a more specific paragraph on markup procedures; d) an exemplification of the described textual markup; e) possible applications to the fields of legal translation and legal linguistics; f) desirable future developments of the methodology.

Started at the University of Turin thanks to an idea of Manuel Barbera and to his tireless fervour lavished in the data collection, the *Corpus Jus Jurium* aims at covering the entire legal universe current in contemporary Italy, which is particularly rich with examples compared to other systems because of the Italian high productivity of laws. We can say that the life of Italian laws will be represented in the final output of the corpus from their first conception in the parliamentary discussion to their codification in normative rules, to their application in judgements – as already shown in Barbera and Onesti 2009.

The corpus will consist of three subcorpora, namely:

- a regulative section, consisting of the Italian Constitution and Codes, laws and decrees;
- a judicial section, consisting of the judgements produced by Italian Courts of law of different degrees;
- a parliamentary section, consisting of the minutes or reports taken down in shorthand during the sessions of Chambers and Commissions.

The purpose of the project is not the creation of a juridical database (Italian scholars have tools like that already, see for example http://www.epicentro.iss.it/biblio/data_giu.asp). The goals of *Jus Jurium* are mainly linguistic. With its innovative conception, the corpus will be POS-tagged and it will have a robust textual markup. Beyond the possibility of retrieving part-of-speech information, a special attention is given to the internal structure of legal texts (see sections 4-5).

The Turin research group has begun with those texts freely obtainable according to Italian law (see in particular law n. 633, 22 April 1941: "Protezione del diritto d'autore e di altri diritti connessi al suo esercizio", i.e. 'Protection of copyright and other rights connected to its application', with special instructions for official State and amministrative acts. Article 5 in particular states that «Le disposizioni di questa legge non si applicano ai testi degli atti ufficiali dello stato e delle amministrazioni pubbliche, sia italiane che straniere», thus giving scholars the right to collect juridical and legal texts freely).

2. Brief project history

The first working phase was launched at the University of Turin, Italy, within the framework of a FIRB national project devoted to "Italian language in the variety of texts: incidence of diachronic, textual and diaphasic variation in the annotation and querying of general and specialistic corpora"[2]. Working on different textual varieties and facing a variegated series of tagging and markup difficulties turned out to be a highly effective test-bed, which led to the first release of the homepage www.corpora.unito.it with its totally free, available and querable texts on the Internet. The web portal is nowadays one of the most meaningful free resources of corpus linguistics in Italy.
The legal section of the project took advantage of the financial support awarded to Cristina Onesti within the national *CNR Promozione Ricerca Giovani* 2005[3].

3. Data collection

Some brief remarks about data selection and collection follow.
In the choice of texts, two criteria were highly important: the definition of being *contemporary* and the definition of being *representative*.
The first concept was curiously more problematic than expected, due to the "not historical" nature of legal texts: a royal decree as well, if never repealed, has to be considered still in force in the Italian legislation. Since it belongs to regulation in force, it is relevant and therefore "contemporary".
Beyond this aspect *de jure*, we have also considered a requirement *de facto* for our corpus materials: their presence online in more websites, also non-institutional ones, but assuring an authentic presence in the current "use" – with a sort of "natural balancing".
In this respect we have paid attention to the two following directions:
(a) representativeness and normative importance (an acknowledged importance hierarchy exists between Italian regulations sources);
(b) the online availability of texts.

Compared to sector-based analysis (see for example the abundant literature about Italian judgements: Bellucci 2005a, Ondelli 2004), the *Corpus Jus Jurium* collects a wider range of textual types. *Jus Jurium*, in fact, will be properly a body of distinct subcorpora, whose general contents and articulation will be the ones which were already displayed in section 1.

4. Markup for a text-structure oriented resource

[2] Original title: "L'italiano nella varietà dei testi. L'incidenza della variazione diacronica, testuale e diafasica nell'annotazione e interrogazione di corpora generali e settoriali", FIRB - RBAU014XCF 2001, coordinated by Carla Marello.
[3] We are grateful to CNR for supporting the project "Tipologie testuali e parti di articolazione del testo nei documenti giuridici: l'uso di corpora per la traduzione giuridica" within a national initiative promoting young researchers.

The corpus presents a lemmatization and POS-tagging which are possible thanks to the Corpus Query Processor (CQP system) elaborated by the IMS Stuttgart (see Christ et al. 1999 and Heid 2007 for a more detailed explanation).

The legal corpus can be defined as 'text-structure oriented', having a further peculiarity consisting of a manual textual markup. Our analysis has indeed cast some light in particular on the still neglected *textual* expression of legal and juridical texts, having the research group undertaking a careful analysis of articulation levels in many different text types. These levels are often not considered within linguistics and legal translation.

Textual distinctions are examined, marking different typical parts of text constituting the proper form of a specific legal text genre.

For this purpose the project has drawn on categories from the branch of *diplomatics* as well, defining diplomatics as the "science of diplomas, or of ancient writings, literary and public documents, letters, decrees, charters, codicils, etc., which has for its object to decipher old writings, to ascertain their authenticity, their date, signatures, etc." (as with the basic definition in en.wikipedia.org), along the same lines as Tessier 1930 and Tessier 1952.

Methodologically, the corpus will therefore enable a more meaningful kind of querying of legal and juridical texts in comparison to a traditional database or to already existing corpora, making use of such preliminary theoretical analyses starting from notions of textual linguistics and diplomatics, as outlined in Onesti (i.p.) as well.

Recent important literature and systematization works of parts of text (Cortelazzo 1997, Mortara Garavelli 2001, Sabatini 1990) have confirmed a high interest in the topic. However, they remarked only to a smaller extent on the importance of textual structures rigidity, deep-rooted in the cultures of the juridical and administrative systems of every country and therefore an evident source of difficulties for specialist translators.

5. Exemplification of textual markup

In order to show the concrete result of such diplomatic systematization work, here follows an extensive markupped example, as submitted to the TreeTagger tool[4].

It is a judgement by the Court of Cassazione (the highest court in the Italian judicial system), more precisely the *Sentenza n. 12070 del 01/07/2004*:[5]

<body>

[4] See the already mentioned Christ et al. 1999 and Heid 2007 for computational details.

[5] In the following examples every existent tag is showed: for a full comprehension, please note that <rfL> refers to every legal reference in the text (with details about number, article, year, etc.); for proper nouns <anth>, <topn>, <ent> indicate respectively the names of people, places, companies/institutions; while <date> refers to the occurring dating.

<neretto>09/08/2004 8.21.57 - L' omessa o incompleta indicazione della commissione tributaria provinciale competente non comporta la nullità del ricorso. […]**</neretto>**
<massima>Nonostante l'articolo <rfL_dl n="546" art="19" data="1992-??-??">19</rfL_dl> del decreto legislativo 546/1992 preveda che […] si deve dedurre la volontà del legislatore, espressa sia pure implicitamente anche nella importante sede dello Statuto, di non assegnare a questo vizio la capacità di produrre la nullità dell'atto.**</massima>**
<testo><titL>[…]</titL>
<intestazione>CASSAZIONE CIVILE, Sezione V, Sentenza n. 12070 del 01/07/2004
REPUBBLICA ITALIANA
IN NOME DEL POPOLO ITALIANO
LA CORTE SUPREMA DI CASSAZIONE SEZIONE TRIBUTARIA
Composta dagli Ill.mi Sigg.ri
[…] **</intestazione>**
<svolgimento>Svolgimento del processo […]**</svolgimento>**
<motivi>Motivi della decisione. Con il primo motivo la ricorrente ha dedotto violazione degli articoli […]**</motivi>**
<dispositivo>P. Q. M. Rigetta il ricorso. Compensa le spese.**</dispositivo>**
<data> Così deciso in <topn>Roma</topn>, nella <ent>Camera di consiglio della Sezione Tributaria</ent>, il <date_2003-09-09>9 settembre 2003</date>.**</data>**
<sottoscrizione>[omissis]**</sottoscrizione>**
<deposito>Depositato in Cancelleria il <date_2004-07-01>1 luglio 2004</date>**</deposito>**
</testo>
</body>

Example 1: Exemplification of the textual markup in a sentence of *Jus Jurium* corpus.

Instead of presenting the characteristics of every textual type represented in the corpus, it seemed to me more useful to face a specific case functioning as an exemplification. Moreover, Italian judgements show a fixed structure of markup. The *sentenza* represents the "most complete" type and it is therefore representative of many text types present in the corpus (see also some interesting remarks in Bellucci 2005b, chapter 4).
It is possible to summarize its macrostructure in the following table, differently visualizing the same text taken from the Court of *Cassazione*:

<body>	<body>
<neretto>__</neretto>	<neretto>09/08/2004 8.21.57 - L' omessa o incompleta indicazione della commissione tributaria provinciale competente non comporta la nullità del ricorso. CASSAZIONE CIVILE, Sezione V, Sentenza n. 12070 del 01/07/2004</neretto>
<massima>__</massima>	<massima>Nonostante l'articolo <rfL_dl n="546" art="19" data="1992-??-??">19</rfL_dl> del decreto legislativo 546/1992 preveda che il ricorrente debba indicare la commissione tributaria provinciale competente per territorio, a giudizio della Cassazione siffatto errore od omissione configura una semplice irregolarità che non inficia la validità del ricorso. Occorre fare riferimento, secondo la Cassazione, ad un contesto normativo processuale (generale, e speciale tributario) che, nel suo complesso, pone l'onere di individuare l'organo

	giurisdizionale esclusivamente su chi presenta un ricorso, onere che è autonomo e che prescinde dall'osservanza di eventuali obblighi di specificazione posti a carico di altri soggetti. Una conferma per questa impostazione deriva dal fatto che la legge n. 212/200 (meglio conosciuta come lo Statuto del contribuente), dopo avere ribadito all'articolo 7, comma secondo, <rfL_l n="212" data="2000-??-??" art="7" comma="2" lettera="d)">lettera d)</rfL_l> la necessità che nell'atto venga indicato l'organo giurisdizionale cui potere ricorrere, non ha poi previsto alcuna sanzione nel caso di omessa o incompleta indicazione. L'art. <rfL_dlgs n="32" art="6" data="2001-??-??">6 D.Lgs. n. 32/2001</rfL_dlgs>, che ha dato attuazione allo Statuto, allorchè ha modificato l'art. <rfL_dlgs n="507" art="10" data="1993-??-??">10 D.Lgs. n. 507/1993</rfL_dlgs> in tema di imposta di pubblicità, ha privilegiato il profilo della necessità della motivazione dell'atto in tutte le sue implicazioni, ma non ha preso in alcuna considerazione il profilo della omessa o incompleta indicazione dell'organo giurisdizionale cui ricorrere. Da tutto ciò si deve dedurre la volontà del legislatore, espressa sia pure implicitamente anche nella importante sede dello Statuto, di non assegnare a questo vizio la capacità di produrre la nullità dell'atto.</massima>
<testo>	<testo>
<titL>__</titL>	<titL>[omissis]</titL>
<intestazione>	<intestazione>CASSAZIONE CIVILE, Sezione V, Sentenza n. 12070 del 01/07/2004
	REPUBBLICA ITALIANA IN NOME DEL POPOLO ITALIANO LA CORTE SUPREMA DI CASSAZIONE SEZIONE TRIBUTARIA Composta dagli Ill.mi Sigg.ri Magistrati: Dott. <anth>FAVARA Ugo</anth> - Presidente -
	Dott. <anth>PAPA Enrico</anth> - Consigliere - Dott. <anth>ODDO Massimo</anth> - Consigliere -
	Dott. <anth>FALCONE Giuseppe</anth> - rel. Consigliere - Dott. <anth>DEL CORE Sergio</anth> - Consigliere - ha pronunciato la seguente:
	sentenza
	sul ricorso proposto da:
	<ent>GFP GRAFICA FOTO PUBBLICITÀ S.P.A.</ent>, in persona del legale rappresentante <anth>Lucillo Azzano</anth>, elettivamente domiciliato in <topn>ROMA</topn> VIA G. B. TIEPOLO 21, presso lo studio dell'avvocato <anth>GIORGIO ALABRESE</anth>, difeso dall'avvocato <anth>LUCIANO FALOMO</anth>, giusta procura a margine;
</intestazione>	</intestazione>
<svolgimento>__</svolgimento>	<svolgimento>Svolgimento del processo

	#___# La <ent>Grafica Foto Pubblicità s.p.a.</ent> ha impugnato l'avviso di accertamento relativo alla imposta sulla pubblicità per l'anno <date_1997-??-??>1997</date> emesso dalla <ent>Ica s.r.l.</ent>, Concessionaria del <ent>Comune di <topn>Azzano</topn> Decimo</ent> per la gestione del servizio di tale tributo. #___# La società ha dedotto la nullità dell'atto perché emanato in violazione dell'articolo 19, <rfL_dlgs n="546" art="19" comma="2" data="1992"-??-??> comma 2^</rfL_dlgs> del D.Lgs. n. 546/1992, in quanto non contenente l'indicazione del Giudice competente per territorio cui ricorrere, e delle modalita' di impugnazione. Nel merito, poi ha sostenuto che l'acronimo <ent>GFP</ent>, utilizzato sulle pareti esterne dello stabilimento, non poteva costituire un presupposto per l'applicazione della imposta sulla pubblicità. #___# La Commissione di primo grado ha respinto il ricorso e la sentenza è stata confermata dalla <ent>Commissione Regionale</ent> che: a) ha ritenuto valido il provvedimento impugnato sul presupposto che nello stesso risultava indicata con un timbro la sua possibilità di impugnazione entro 60 giorni, con ricorso dinanzi alla <ent>Commissione Tributaria Provinciale</ent>, con le modalità previste dal D.Lgs. n. <rfL_dlgs n="546" data="1992-??-??">546/92</rfL_dlgs>; aa) ha ritenuto che sussistevano i presupposti per tassare le scritte apposte sulla parete esterna dello stabilimento in quanto costituenti un veicolo pubblicitario. Ha proposto ricorso la società grafica. Ha resistito la <ent>Ica s.r.l.</ent>. </svolgimento>
<motivi>__</motivi>	<motivi>Motivi della decisione. Con il primo motivo la ricorrente ha dedotto violazione degli articoli <rfL_dlgs n="546" art="19" data="1992-??-??>19</rfL_dlgs> e <rfL_dlgs n="546" art="20" data="1992-??-??">20 D.Lgs. n. 546/92</rfL_dlgs> perché la <ent>Commissione Regionale</ent>: b) avrebbe errato a rigettare l'eccezione di nullità dell'atto di accertamento derivante dalla mancata indicazione in esso delle modalità di impugnazione; bb) avrebbe anche omesso di motivare la sua decisione. In particolare, ha rilevato come l'avviso non contenesse l'indicazione di quale <ent>Commissione Tributaria Provinciale</ent> fosse competente per territorio, ne' l'indicazione delle forme da osservare nella proposizione del ricorso. #___# La resistente ha sostenuto l'infondatezza del motivo sul presupposto che nell'avviso è sufficiente indicare che il ricorso deve essere proposto "alla <ent>Commissione Tributaria

	Provinciale</ent> Competente", senza la individuazione, volta per volta, del Giudice territorialmente competente, e che una contraria conclusione sarebbe gravosa per il riscossore che spesso (come nel caso di specie) riscuote tributi sparsi su tutto il territorio dello Stato. Ha poi aggiunto che l'indicazione che il ricorso va fatto secondo le modalità del D.Lgs. n. <rfL_dlgs n="546" data="1992-??-??">546/92</rfL_dlgs> è sufficiente per ritenere assolto l'obbligo previsto dalla norma. […] #___# Con il terzo motivo la ricorrente ha dedotto gli stessi vizi indicati nel secondo motivo per sostenere che il marchio impresso sulle pareti serve unicamente ad identificare il luogo ove si svolge l'attività ma non funge da collettore della clientela perché si rivolge a soggetti già preventivamente orientati verso quella determinata azienda. #___# Ritiene la Corte che la doglianza è inammissibile poiché si tratta di una questione di fatto per quanto detto testé. #___# Sussistono giusti motivi per compensare le spese di questo giudizio tra le parti. </motivi>
<dispositivo>__</dispositivo>	<dispositivo>P. Q. M. Rigetta il ricorso. Compensa le spese.</dispositivo>
<data>__</data>	<data>Così deciso in <topn>Roma</topn>, nella <ent>Camera di consiglio della Sezione Tributaria</ent>, il <date_2003-09-09>9 settembre 2003</date> .</data>
<sottoscrizione>__</sottoscrizione>	<sottoscrizione>[omissis]</sottoscrizione>
<deposito>__</deposito>	<deposito>Depositato in Cancelleria il <date_2004-07-01>1 luglio 2004</date></deposito>
</testo>	</testo>
</body>	</body>

Table 3: Visualization of the internal articulation of a sentence of *Cassazione*.

Note that the tag <testo> we have highlighted in the first table (in bold font) begins after the "neretto" and the "massima", which are two typical and unavoidable parts of Italian judgements, and it closes at the end of the deed, after *data, sottoscrizione* and *deposito* (see below).

An Italian "massima" corresponds to the diplomatic document summary, with its synthesis of all the relevant points (both judicially and factually relevant) of the document. In other cases, there is only the "neretto" at the beginning of the judgement, acting as a sort of extended title. The term "neretto" cannot easily be translated since it is rooted in the Italian juridical system. Both notions are official only in judgements of the Court of *Cassazione*.

Within this 'text', common but suitable tags were then attributed in order to mark the internal discourse articulation:

- <titL> refers to the "title" of a deed;

- <intestazione> can be assimilated to the heading of a text. It is the part of the protocol of a judgement, which bears the heading "Repubblica Italiana", "in nome del popolo italiano [in the name of the Italian people]" and "the reference to the judge who has passed the judgement" as established by the Code of Civil Procedure, article 132;
- <svolgimento> points out the main conduct of the trial, marking the "account of the development of the trial" (Code of Civil Procedure, article 132). It normally corresponds to the part of the text which begins with "Fatto", "Ritenuto in fatto", "Svolgimento del processo" or similar. In some cases this section is integrated with the following one, in which case they are marked as "fatto e diritto [fact and law]". If the two parties are integrated one into the other, we can in any case clearly distinguish and markup them;
- <motivi> contains "the concise account of the development of the trial and of the pleas of fact and law of the decision" (Code of Civil Procedure, article 132). Such reasons normally correspond to the part of the text which begins with "Diritto", "Motivi della decisione", "Considerato in diritto" or similar. We have not found any case of absence of the reasons in the material examined so far. Therefore, in the case of absence it will be marked as [omissis];
- <dispositivo> is the actual content of the judgement: what the judgement orders (Code of Civil Procedure, article 132, p. 5). It normally corresponds to the part of the text which begins with "PQM" or "per questi motivi";
- <data> indicates the date of the deliberation, which always follows the final order (again according to the rules contained in the Code of Civil Procedure, article 132, p. 5);
- <sottoscrizione> means the signing of the relevant and involved authorities;
- <deposito> is meant as a deposit and official recording: in a very few cases, at the bottom of the text, we can find the information regarding the 'deposito' of a judgement for a given date, although it is clearly compulsory (as we can read from the above articles of the Code of Civil Procedure).

Such terminology follows a praxis coming from the Italian diplomatic tradition and it can provide a renewed tool for corpus linguists.
We aim at analyzing the corpus with separate queries for different recurrent text parts, which is an aspect that is still not studied in the Italian linguistic research panorama - at least not in a comprehensive way.

6. Possible applications: legal translation and legal linguistics

Although the results of legal drafting are already remarkable in some academic case studies (Biagioli et al. 1995 for example, and their group of legimatics in Florence, at the ITTIG - Istituto di Teoria e Tecniche dell'Informazione Giuridica), the application of such an analysis to legal translation was insufficient.
Legal translation has made use mainly of recommendations and techniques at a terminological or syntactical level. However, textual structure influences considerably the drafting of a legal document - and therefore every subsequent translation of it.

Thanks to a text-structure oriented legal corpus like *Jus Jurium*, translators will have the consciousness of working within the boundaries of a specific part of text.
They will have a special markup of the beginning and ending of text sections. Translators will have the opportunity to create appropriate strategies concerning them, also supported by rightful "expectations" as regards the preceding or following language fragments.

In addition, the distinction of textual parts permits a focused collection of fixed formulas which are present in the textual parts themselves. For the translator, a collection of technical terms and phrases deep-rooted in the cultures of the juridical and administrative systems of every country is indisputedly useful, and much more so if the tool is able to provide detailed information of the specific parts of text where those expressions, idioms or collocations occur.

Elaborating markupped texts with such peculiarities implies a long period of manual work, since most of these distinctions cannot be done by a machine (awareness of different legal text genres and of compulsory or optional textual parts is needed). However, the final result can be a very useful resource for translators looking for idioms, collocations or terminological elements in a specific fragment of legal texts.

A large amount of linguistic data can be much more useful for legal linguists – see also Coulthard 1994 – providing them with an extensive repository of well-structured data and with fine-grained querying opportunities, whether at the morphosyntactic or lexical or textual level, contributing to the branch of research a real corpus (see definition in Barbera, Corino and Onesti 2007) and not a mere database.
The final interface for querying the tool will be developed having a look at the different needs of translators and linguists in order to browse the data in as user-friendly a way as possible. The user interface which is already exploited within the other corpora of the project is totally free and open to scholars and users alike.
Lemmatized and POS-tagged, in the future all subcorpora will be querable separately or all together by means of the CQP system.

7. Concluding remarks and future perspectives

The main objective of the project is to make the corpus available and free on the Internet in 2012. A first draft of the homepage is already available at:
http://www.bmanuel.org/projects/ju-HOME.html.

Hopefully it will be possible to create in the future another two subcorpora as well:

- *Sectio Didactica*: a fourth subcorpus devoted to law teaching (obviously bound to the possibility of copyright acquisition of some law manual books);
- *Sectio Communis*: devoted to the way "common people" talk of rights and law (exported from the corpus of Italian newsgroups also available on the portal www.corpora.unito.it[6], a repository for messages posted from many users in different locations, functionally similar to discussion forums) selecting newsgroups messages concerning law.

[6] See in particular http://www.bmanuel.org/projects/ng-HOME.html.

Considering that the research group of *Jus Jurium* has collected until now only texts in the Italian language, a promising perspective for the future would be certainly related to the possibility of applying the same methodology interlinguistically: Roffinella 2010 has already started a feasibility study on English sentences. Multilingual text-structure oriented legal corpora could become important, in particular, for translators' training, both for resources in the target language and for analyzing comparable data. Nonetheless, the collection of data in other languages implies a considerable amount of work, which is still not officially scheduled into the project.

Bibliography

Barbera, Manuel, Elisa Corino, Cristina Onesti. 2007. Che cos'è un corpus? Per una definizione più rigorosa di corpus, token, markup. In *Corpora e linguistica in rete,* eds. Manuel Barbera, Elisa Corino and Cristina Onesti, 25-88. Perugia: Guerra Edizioni.

Barbera, Manuel, Cristina Onesti. 2009. Scheda progetto di ricerca Corpus Jus Jurium. In *Progetto JURA. Manuale per la formazione dei docenti e dei traduttori che operano nell'ambito del linguaggio giuridico italiano-tedesco*, eds. Pierangela Diadori, 349-350. Perugia: Guerra Edizioni.

Bellucci, Patrizia. 2005a. La redazione delle sentenze: una responsabilità linguistica elevata. *Diritto & Formazione* 5, 3: 448-66.

Bellucci, Patrizia. 2005b. *A onor del vero. Fondamenti di linguistica giudiziaria,* Torino: UTET Libreria.

Bhatia, Vijay K. 1987. Language of the Law. *Language teaching: the international abstracting journal for language teachers and applied linguists* 20(3): 227-34.

Biagioli, Carlo, Mercatali Pietro, Sartor Giovanni, eds. 1995. *Legimatica: informatica per legiferare,* Napoli: Edizioni scientifiche italiane.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann, Esther König. 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP). User's Manual*, Stuttgart: Institut für maschinelle Sprachverarbeitung (CQP V2.2). http://www.ims.uni-stuttgart.de/projekte/CorpusWorknbench/CQPUserManual/PDF/cqpman.pdf.

Cortelazzo, Michele. 1997. Lingua e diritto in Italia: il punto di vista dei linguisti. In *La lingua del diritto. Difficoltà traduttive, applicazioni didattiche*, ed. Leo Schena, 35-49 (Atti del primo convegno internazionale, Università Bocconi, Milano, 5-6 ottobre 1995), Roma: CISU.

Coulthard, Malcolm. 1994. On the Use of Corpora in the Analysis of Forensic Texts. *Forensic Linguistics. The International Journal of Speech Language and Law* I/1: 27-43.

Heid, Ulrich. 2007. Il Corpus WorkBench come strumento per la linguistica dei corpora. Principi ed applicazioni. In *Corpora e linguistica in rete*, ed. Manuel Barbera, Elisa Corino, Cristina Onesti, 89-108. Perugia: Guerra Edizioni.

Mortara Garavelli, Bice. 2001. *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*. Torino: Giulio Einaudi Editore.

Ondelli, Stefano. 2004. *Il genere testuale della sentenza penale in Italia: l'impiego dell'indicativo tra performatività e narrazione*, PhD diss., University of Padova.

Ondelli, Stefano. 2007. *La lingua del diritto. Proposta di una classificazione di una varietà dell'italiano.* Roma: Aracne.

Onesti, Cristina. In press. 意大利语法律语料库的建设 [Construction of an Italian legal corpus]. *Journal of Guangdong University of Foreign Studies* (translated by Ge Yunfeng | 葛云峰).

Roffinella, Paola. 2010. *Analisi comparativa della struttura di sentenze italiane e inglesi volta alla creazione di un corpus giuridico inglese*, Master thesis, University of Turin, Faculty of Foreign Languages, unpublished.

Rovere, Giovanni. 2005. *Capitoli di linguistica giuridica. Ricerche su corpora elettronici*. Alessandria: Edizioni Dell'Orso.

Sabatini, Francesco. 1990. Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi. In *Corso di studi superiori legislativi (1988-89),* ed. Mario D'Antonio, 675-724. Padova: Cedam.

Tessier, Georges. 1930. Leçon d'ouverture du cours de diplomatique de l'École des chartes. *Bibliothèque de l'École des chartes* XCI : 241-63.

Tessier, Georges. 1952. *La diplomatique*. "Que sais-je?" 536. Paris: PUF.

Visconti, Jacqueline. 2000a. La traduzione del testo giuridico. Problemi e prospettive di ricerca. *Terminology and Translation. A Journal of the Language Services of the European Institutions* 2000/2: 38-66.

Visconti, Jacqueline. 2002. Un corpus comparativo di testi legali. Considerazioni di linguistica forense. In *La parola al testo. Scritti per Bice Mortara Garavelli*, eds. Gian Luigi Beccaria and Carla Marello, 481-497. Alessandria: Edizioni dell'Orso.

Visconti, Jacqueline. 2007. A textual approach to legal drafting and translation. In *Approaching the Multilanguage Complexity of European Law: Methodologies in comparison*, eds. Gian Maria Ajani et al., 107-132. Firenze: European Press Academic Publishing.

Zuanelli Elisabetta. 2000. Macrostruttura pragmatica e modelli di interazione nel testo normativo. In *Linguistica giuridica italiana e tedesca / Rechtslinguistik des Deutschen und Italienischen*, ed. Daniela Veronesi, 85-99. Padova: Unipress.