Legal Corpora: an overview[1]

Gianluca Pontrandolfo

Università di Trieste

Abstract

The present paper is mainly addressed to researchers and/or translators who are daily confronted with the legal domain in different languages and are willing to approach legal language through 'real-life' examples, to paraphrase McEnery & Wilson's classical definition of corpus linguistics (2001: 2). With no claim of being exhaustive, the study has been devised as a practical guide, a tentative survey of the available corpora for legal language.

Emphasis has been placed on three main areas, namely, England and Wales, Spain and Italy, for being the focus of study of an ongoing PhD research project. However, reference has also been made to legal corpora and subcorpora available outside these countries, in Europe as well as in the rest of the world.

Primarily conceived as a classical PhD 'review' – the crucial step in every research study involving a state of the art analysis –, it can be viewed also as a preliminary map for those who are taking their first steps into the fascinating world of corpus linguistics. The practical approach is evident from the schematic method adopted: the tables and the final Appendix are meant to be useful tools for rapid consultation or comparison among the copious legal corpora listed in the paper.

[1] The author wishes to thank prof. Helena Lozano Miralles (Università di Trieste) for supervising his PhD research project, Stanisław Goźdź-Roszkowski (Uniwersytet of Łódzki) for his insightful comments and constructive criticism and prof. Jane Kellett (Università di Trieste) for kindly reviewing the whole paper.

1. Introduction

> Corpus linguistics has been widely claimed to be a powerful instrument for the study of linguistic frequency in and across a variety of discourses. The use of computerized corpora has further made it possible for linguists to undertake automatic analyses of lexico-grammatical and, to some extent, discoursal features of texts. In the last few years these corpus-based studies have become so popular that one rarely finds a textual study without the use of computerized corpora (Bhatia *et al.* 2004: 203).

The generalisation made by the authors in the introductory quotation to this paper was definitely true in 2004, but even more so in 2012, when corpus-based studies have become a fundamental trend in the study of legal language. Compared to the invention of the microscope and the telescope, which suddenly allowed scientists to observe things that had never seen before (Stubbs 1996: 231-232), the use of electronic corpora in language as well as in legal translation and interpreting studies has become a mainstream methodology (Biel 2010a).

The potential of corpus linguistics as a methodology for researching legal language and translation (e.g. Biel 2010a, Goźdź-Roszkowski 2011), and as a tool in translator training (e.g. Monzó 2008, Biel 2010b) is nowadays unquestionable. Whether we conceive it as a methodology or as a discipline – the controversy has not been ironed out yet (see Tognini-Bonelli 2001: 1-2, McEnery *et al.* 2006: 7-8) – the introduction of electronic corpora has represented a watershed in many branches of linguistics and it is still displaying its potential.

The present paper is primarily addressed to researchers and/or translators who are daily confronted with the legal domain in different languages and are willing to approach legal language based on examples of 'real life' use, to paraphrase McEnery & Wilson's classical definition of corpus linguistics (2001: 2). It is mainly conceived as a practical guide, a tentative survey of the available corpora for the study of legal language. As Xiao (2008: 383) points out, there are thousands of corpora in the world, but most of them are created for specific research projects and are not publicly available. This makes the task arduous and this is the reason why the present overview has no claim of being exhaustive.[2]

The paper stems from an ongoing PhD research project aiming at analysing qualitatively and quantitatively legal – to be more exact, judicial – phraseology in English, Spanish and Italian criminal judgments. Its main objective is providing legal translators with a multifunctional tool having a positive impact on the translation process, as well as on the quality of their texts. The present study is deeply rooted in this project and is part of it by being a synthesis of the state of the art of a significant number of existing legal corpora. Reviewing the criteria used to compile such corpora will be a fundamental step towards a refining of the methodology that will be adopted to build a specialised corpus of criminal judgments, specifically designed to address the PhD research objectives.

[2] See Xiao (2008) for a comprehensive survey on well-known and influential corpora, and the URLs to web pages containing useful lists of available corpora all around the world.

Since the paper is placed within the bounds of the ongoing PhD thesis, priority has been given to those corpora including the languages of the study, namely, English, Spanish and Italian and emphasis has been placed on corpora dealing with Criminal Law as a subject area and criminal judgment as a major genre. This also limits the scope of the survey that otherwise would have been too wide to be tackled in a short essay.

Legal corpora and subcorpora mentioned in the present paper have been grouped according to the area where the project was launched and not according to their primary uses (cf. Xiao 2008: 383). The resulting sections are the following ones: England and Wales (§ 2.1), Spain (§ 2.2), Italy (§ 2.3), European Union (§ 2.4) and rest of the world (§ 2.5). After the conclusions (§ 3) and the references, an Appendix gathers useful information on the website addresses discussed or hinted at throughout the paper, and some valuable web pages including lists of corpora.

2. Legal corpora: a tentative survey

The following sections describe the main features of some influential corpora for the study of legal language. A selected number of parameters chosen for being prototypical in corpus building will be identified. For the most important corpora in each area, especially for the English, Spanish and Italian ones, a table is provided containing crucial information on the corpora, in particular: name (if applicable); institution or university sponsoring it, together with the leading researchers; types of corpus (cf. Laviosa 2010, Zanettin 2012); languages included; dimension (in terms of number of tokens); text typologies/genres included; time span; purposes (in particular research vs. training); availability to the public; notes (a final section containing additional information on the corpus structure, such as, if it is annotated or not).[3]

2.1 England and Wales

English was certainly the forerunner in corpus research (Xiao 2008: 383) which explains the high number of corpora including it as main language. However, the analysis of the existing legal corpora developed in Great Britain, and especially in England and Wales, revealed a different picture.

[3] Abbreviations and symbols used in the table: BrE = British English; AmE = American English; w = words; Mw = million of words, # = number of; $ = purchasable; N/A = not available. The Note section is omitted when corpus annotation has not been performed at all.

2.1.1 Cambridge Corpus of Legal English

The *Cambridge Corpus of Legal English* is a subcorpus of a huge multi-billion corpus built by Cambridge University Press, named *Cambridge English Corpus* (CEC), formerly *Cambridge International Corpus* (cf. Xiao 2008: 410, 429) containing both text corpus and spoken corpus data.

Name	Cambridge Corpus of Legal English
Institution/University Leading researcher(s)	Cambridge University Press
Type of corpus	Monolingual
Languages	EN (BrE, AmE)
Dimension (#tokens)	20Mw
Text-types/Genres	books, journals, newspaper articles relating to the law and legal processes
Time span	1993-
Purposes	Research
Availability	No ($)

2.1.2 HOLJ

The *House of Lords Judgments Corpus* (HOLJ) is an interesting project developed at the University of Edinburgh with the primary objective of studying the rhetorical sections of a selection of judgments delivered by the House of Lords with the final aim of obtaining an automatic summarisation (see Grover *et al.* 2004).

Name	HOLJ Corpus
Institution/University Leading researcher(s)	University of Edinburgh B. Hachey – C. Grover
Type of corpus	Monolingual
Languages	EN
Dimension (#tokens)	2,887,037w
Text-types/Genres	188 HL judgments
Time span	2001-2003
Purposes	Research (primary aim: automatic summarisation)
Availability	Yes

2.1.3 Proceedings of the Old Bailey

The *Proceedings of the Old Bailey* (London's Central Criminal Court) is a fascinating example of diachronic corpus for the study of historical judicial language of criminal trials.

Name	Proceedings of the Old Bailey
Institution/University Leading researcher(s)	Open University (C. Emsley), University of Hertfordshire (T. Hitchcock) and University of Sheffield (R. Shoemaker).
Type of corpus	Monolingual (diachronic)
Languages	EN
Dimension (#tokens)	127Mw
Text-types/Genres	197,745 criminal trials
Time span	1674-1913
Purposes	Research
Availability	Yes

2.2 Spain

Spain holds the record of the highest number of legal corpora developed in the last few years. In the following tables, a detailed description of the most important projects launched at national level is provided.

2.2.1 JUD-GENTT

Name	JUD-GENTT
Institution/University Leading researcher(s)	Universidad Jaume I (Castellón) A. Borja Albi (coord.)
Type of corpus	Multilingual, comparable and parallel
Languages	EN-ES-DE-FR
Dimension (#tokens)	N/A
Text-types/Genres	Different kinds of texts produced as part of the criminal proceedings in England, Spain, Germany and France. Textual genres: N/A.
Time span	N/A
Purposes	Research, Translator Training
Availability	No

JUD-GENTT is an ongoing research project developed within the GENTT project (*Textual Genres for Translation*), that aims at building a multilingual (EN-ES-DE-FR) comparable corpus of textual genres (Law, Medicine and other technical fields) to provide a sort of encyclopedia of specialised texts for translation. JUD-GENTT, a new project coordinated by Anabel Borja Albi (University of Jaume I, Castellón), is an action-research project whose aim is to improve the socio-professional conditions of legal translators and their productive processes. It is a multilingual comparable and parallel corpus gathering different kinds of texts produced as part of the criminal proceedings in the different legal systems.

2.2.2 CORPUS

<table>
<tr><td>Name</td><td>CORPUS - (Corpus tècnic del IULA)</td></tr>
<tr><td>Institution/University
Leading researcher(s)</td><td>Universitat Pompeu Fabra
M. T. Cabré (Leading Researcher)
J. Vivaldi (coord.)</td></tr>
<tr><td>Type of corpus</td><td>Multilingual, comparable and parallel</td></tr>
<tr><td>Languages</td><td>CA-ES-EN-FR-DE</td></tr>
<tr><td>Dimension (#tokens)</td><td>
<u>Comparable</u> corpus:
Composition per number of <i>tokens</i> (in thousands):
<table>
<tr><th>Area</th><th>CA</th><th>ES</th><th>EN</th><th>FR</th><th>DE</th><th>Tot.</th></tr>
<tr><td><b>L</b></td><td><b>1684</b></td><td><b>2086</b></td><td><b>432</b></td><td><b>44</b></td><td><b>16</b></td><td><b>4262</b></td></tr>
<tr><td>Ec</td><td>1821</td><td>1091</td><td>275</td><td>78</td><td>27</td><td>3292</td></tr>
<tr><td>En</td><td>1506</td><td>1083</td><td>600</td><td>230</td><td>429</td><td>3848</td></tr>
<tr><td>M</td><td>2625</td><td>4375</td><td>1701</td><td>27</td><td>198</td><td>8926</td></tr>
<tr><td>CS</td><td>654</td><td>1227</td><td>339</td><td>194</td><td>83</td><td>2497</td></tr>
<tr><td>Tot.</td><td>8290</td><td>9862</td><td>3347</td><td>573</td><td>753</td><td>22825</td></tr>
</table>
[LAW: 4.26 Mw]
Composition per number of <i>documents</i>:
<table>
<tr><th>Area</th><th>CA</th><th>ES</th><th>EN</th><th>FR</th><th>DE</th><th>Tot.</th></tr>
<tr><td><b>L</b></td><td><b>153</b></td><td><b>124</b></td><td><b>65</b></td><td><b>10</b></td><td><b>60</b></td><td><b>412</b></td></tr>
<tr><td>Ec</td><td>81</td><td>47</td><td>18</td><td>8</td><td>1</td><td>155</td></tr>
<tr><td>En</td><td>78</td><td>55</td><td>86</td><td>22</td><td>61</td><td>302</td></tr>
<tr><td>M</td><td>236</td><td>401</td><td>284</td><td>3</td><td>27</td><td>951</td></tr>
<tr><td>CS</td><td>39</td><td>67</td><td>27</td><td>6</td><td>8</td><td>147</td></tr>
<tr><td>Tot.</td><td>587</td><td>694</td><td>480</td><td>49</td><td>157</td><td>1967</td></tr>
</table>
<u>Parallel</u> corpus:
<table>
<tr><th>Area</th><th colspan="2">CA-ES</th><th colspan="2">CA-EN</th><th colspan="2">ES-EN</th></tr>
<tr><th></th><th>Docs.</th><th>Words</th><th>Docs.</th><th>Words</th><th>Docs.</th><th>Words</th></tr>
<tr><td><b>D</b></td><td><b>64</b></td><td><b>485</b></td><td><b>1</b></td><td><b>12</b></td><td><b>2</b></td><td><b>57</b></td></tr>
<tr><td>E</td><td>21</td><td>600</td><td>10</td><td>253</td><td>13</td><td>283</td></tr>
<tr><td>MA</td><td>12</td><td>256</td><td>12</td><td>230</td><td>13</td><td>144</td></tr>
<tr><td>M</td><td>5</td><td>129</td><td>1</td><td>39</td><td>102</td><td>809</td></tr>
<tr><td>I</td><td>1</td><td>28</td><td>-</td><td>-</td><td>22</td><td>292</td></tr>
<tr><td>Tot.</td><td>103</td><td>1498</td><td>24</td><td>534</td><td>152</td><td>1585</td></tr>
</table>
http://www.iula.upf.edu/corpus/estates.htm [24/11/2012]
</td></tr>
<tr><td>Text-types/Genres</td><td><u>Legal Subcorpus</u>:[1] Legislative texts; Professional practice texts; Judicial texts; Theoretical texts (e.g. manuals); Instrumental texts (e.g. dictionaries).</td></tr>
<tr><td>Time span</td><td>1993-</td></tr>
<tr><td>Purposes</td><td>Research, Training</td></tr>
<tr><td>Availability</td><td>No</td></tr>
<tr><td>Notes</td><td>The corpus is annotated and marked up following the SGML standards and the guidelines of the Corpus Encoding Standard (CES) of the EAGLES[2] initiative.</td></tr>
</table>

[1] Subject Areas (http://www.iula.upf.educorpus/acdreca.htm, 24/11/2012): *Private Law*: Civil Law, Commercial Law, Labour Law, Criminal Law, Canon Law; *Public Law*: Constitutional Law, Administrative Law, Financial and Tax Law, International and Public Law; *Legal Theory*.

[2] http://www.ilc.cnr.it/EAGLES96/browse.html (24/11/2012).

The project *CORPUS (Multilingual Specialised Textual Corpus*, sometimes referred to as *Technical Corpus*), developed by the *Institute for Applied Linguistics* of the University Pompeu Fabra of Barcelona (IULA) collects a multilingual and comparable corpus of different domains: Law (L), Economics (Ec), Environment (En), Medicine (M), Computer Science (CS). It is used both for research (neologism detection, linguistic variation, syntactic analysis, etc.) and training purposes. For the purposes of the present survey, it is interesting because it contains a large subcorpus of legal language.

2.2.3 CLUVI

Name	CLUVI Corpus Lingüístico da Universidade de Vigo
Institution/University Leading researcher(s)	Universidade de Vigo G. X. Gómez, A. Simões
Type of corpus	Multilingual, parallel
Languages	EN-FR-ES-PT-DE-GL-EU-CA
Dimension (#tokens)	Tot. CLUVI: 27,541,023w LEGA (GL-ES): 6,582,415w LEGE-BI, Legebiduna, (EU-ES): 2,384,053w
Text-types/Genres	LEGA: legislative texts (leyes orgánicas, real decretos, regulamentos, diarios oficiales, etc.) LEGE-BI: Boletín Oficial de Gipuzkoa 1998-2001, Boletín Oficial del Territorio Histórico de Álava 1992-1994
Time span	1978 (Spanish Constitution)-
Purposes	Research
Availability	Yes

The *Linguistic Corpus of the University of Vigo* (*CLUVI*) is a parallel open corpus of specialised registers (fiction, computing, journalism, legal and administrative fields, etc.), totaling more than 27 million words of running texts (see Xiao 2008: 434-435). Two of its eight subcorpora are entirely dedicated to legal language, namely LEGA and LEGE-BI.

2.2.4 Other

In this section other corpora for the study of legal language developed in Spain will be mentioned. They are not included in the main sections either because they are not full-blown corpora or they are built by single researchers, often PhD students working on their theses.

The University of Valencia has built up the *GENTEXT-N corpus*, within the research group Gender, Language and Sexual (In)Equality. It is a bilingual (ES-EN) comparable corpus of almost 35 million words extracted from press articles

(*The Times, The Guardian, El País, El Mundo*) dealing with legal actions to cope with sexual (in)equality in Spain and Great Britain.

Another interesting project is *GARALEX* (University of the Basque Country), a web platform for the study of legal language, developed following a corpus-based methodology.

The *Corpus de Procesos Penales* (*CPP*) is a monolingual (ES) corpus of criminal trials built by Raquel Taranilla (University of Barcelona) of 98,943 words that collects 10 criminal trials held in Barcelona between 2009 and 2010. Its primary aim was the study of narrative elements in judicial discourse (cf. Taranilla 2011).[4]

The *British Law Report Corpus* (*BLaRC*) is another interesting corpus built by María José Marín Pérez (University of Murcia) for lexical and terminological purposes. It is a monolingual (EN) corpus of 8.8 million words extracted from law reports issued by five jurisdictions: Commonwealth, United Kingdom, England and Wales, Northern Ireland and Scotland.

Finally, Bianca Vitalaru (University of Alcalá) has also developed a trilingual (ES-EN-RU) ontological glossary for the study of criminal law language, based on a large corpus of legal documents.

2.3 Italy

As far as Italy is concerned, a growing interest in legal language has been recorded in recent years. Since the pioneer BoLC, a number of other corpora for the study of Italian legal language have been built, both from the academic and professional communities.

2.3.1 BoLC

The *Bononia Legal Corpus* (*BoLC*) is the most representative bilingual (EN-IT) corpus of legal language developed in Italy. It is an interdisciplinary project which started in 1997 at the University of Bologna as a 'corpus-driven research project' (Rossini Favretti *et al.* 2001: 14). The subcorpora of Italian and English legal languages are taken to represent two different legal systems, in particular the differences between the *civil law* and the *common law* systems.

[4] Another interesting project in which Taranilla was involved was the *Report on Written Language*, issued by the Studies on Academic and Professional Discourse Research Group (EDAP), leaded by Estrella Montolío Durán (University of Barcelona). As part of *Report of the Commission for the Modernization of Spanish Legal Language*, sponsored by the Spanish Ministry of Justice, a huge corpus of judicial documents was collected for the study and simplification of judicial language. Information available also at: http://www.mjusticia.gob.es/cs/Satellite/es/1288775399001/MuestraInformacion.html (24/11/2012).

Name	BoLC Bononia Legal Corpus
Institution/University Leading researcher(s)	Università di Bologna R. Rossini Favretti (Leading Researcher), F. Tamburini, E. Martelli [J. Sinclair]
Type of corpus	Bilingual, comparable
Languages	EN, IT
Dimension (#tokens)	Subcorpus EN: 21Mw Subcorpus IT: 33.5Mw
Text-types/Genres	EN: Acts of Parliament, Chancery Division, Court of Appeal, Family Division, House of Lords, Privy Council, Queen's Bench Division, Statutory Instruments IT: Costituzione, Codice Civile, Codice Penale, Codice di Procedura Civile, Codice di Procedura Penale, Decreti Legislativi, Leggi Costituzionali, Leggi Ordinarie, Sentenze Penali Corte di Cassazione, Sentenze Civili Corte di Cassazione, Sentenze e Ordinanze della Consulta
Time span	1968-1995
Purposes	Research
Availability	No
Notes	Pilot corpus (see Rossini Favretti *et al.* 2001: 15-16): Bilingual parallel corpus of EU documents (1995-1996) 2,232 directives EN: 6.5Mw 1,798 direttive IT: 5.8Mw 4,472 judgments EN: 13.7Mw 4,471 sentenze IT: 12.3Mw

2.3.2 CORIS/CODIS

Name	CORIS/CODIS CORIS (Corpus di Riferimento dell'Italiano Scritto) CODIS (Corpus Dinamico dell'Italiano Scritto)
Institution/University Leading researcher(s)	Università di Bologna R. Rossini Favretti
Type of corpus	Monolingual
Languages	IT
Dimension (#tokens)	CORIS: 130Mw CODIS: 100Mw
Text-types/Genres	PRESS: 38% - FICTION: 25% ACADEMIC PROSE: 12% LEGAL AND ADMINISTRATIVE PROSE: 10% [books, journals, legal and administrative documents] MISCELLANEA: 10% EPHEMERA: 5%
Time span	CORIS 1980-2010, CODIS 1980-2000
Purposes	Research
Availability	Yes
Notes	Both corpora were annotated by F. Tamburini.

The *Corpus di Riferimento dell'Italiano Scritto* (*CORIS*) and the *Corpus Dinamico dell'Italiano Scritto* (*CODIS*) are two different structures of the same reference corpus developed at the University of Bologna by Rossini Favretti's team. The project started in 1998 with the purpose of creating a representative and sizeable general reference corpus of written Italian – following the *Brown Corpus* model (see Xiao 2008: 395-397) – which would be easily accessible and user-friendly. Compared with CORIS (100 million words, plus 30 million words of monitor corpus), CODIS (100 million words) has a dynamic structure allowing researchers to exclude or include different subcorpora for specific analyses (Rossini Favretti *et al.* 2002). It has a subcorpus of legal language, totaling 10 million words.

2.3.3 CADIS

<table>
<tr><td>Name</td><td>CADIS
Corpus of Academic English</td></tr>
<tr><td>Institution/University
Leading researcher(s)</td><td>Università degli Studi di Bergamo
M. Gotti</td></tr>
<tr><td>Type of corpus</td><td>Bilingual, comparable</td></tr>
<tr><td>Languages</td><td>EN, IT</td></tr>
<tr><td>Dimension (#tokens)</td><td>2,761 academic texts (12Mw)</td></tr>
<tr><td>Text-types/Genres</td><td>Disciplinary areas:
- Applied Linguistics (AL)
- Economics (E)
- Law (L)
- Medicine (M)

Textual genres:
- Research articles (RA)
- Abstracts (A)
- Book reviews (B)
- Editorials (E)

Composition of the <i>Law subcorpus</i>:
<table>
<tr><td>Law</td><td>RA</td><td>A</td><td>B</td><td>E</td></tr>
<tr><td>1980-1999 EN</td><td>50</td><td>50</td><td>50</td><td>8</td></tr>
<tr><td>1980-1999 IT</td><td>14</td><td>14</td><td>-</td><td>2</td></tr>
<tr><td>2000-2011 EN</td><td>94</td><td>94</td><td>100</td><td>121</td></tr>
<tr><td>2000-2011 IT</td><td>50</td><td>23</td><td>12</td><td>4</td></tr>
<tr><td>Tot.</td><td>208</td><td>187</td><td>162</td><td>136</td></tr>
</table>
http://dinamico.unibg.it/cerlis/public/CADIS_Corpus.pdf [24/11/2012]</td></tr>
<tr><td>Time span</td><td>1980-1999 + 2000-2011</td></tr>
<tr><td>Purposes</td><td>Research</td></tr>
<tr><td>Availability</td><td>No</td></tr>
</table>

The *Corpus of Academic English* (*CADIS*) is a research project funded by the Italian Ministry of Research and developed at the University of Bergamo under the scientific direction of Maurizio Gotti. The corpus lies at the heart of a scientific project aimed at analysing identity traits in academic discourse (Gotti 2010). It is composed of a major English subcorpus and a smaller one in Italian for comparative purposes. CADIS represents four main disciplinary areas: Applied Linguistics (AL), Economics (E), Law (L) and Medicine (M). For each disciplinary area, four different textual genres have been considered: abstracts (A), book reviews (B), editorials (E), research articles (RA).

The comparability of the corpus stems not only from its bilingual structure, its disciplinary areas and its genres, but also from the historical period. CADIS can be queried also diachronically, since texts are subdivided into two main time spans (1980-1999; 2000-2011). It is interesting for the purposes of the present survey because of its legal subcorpus.

2.3.4 Other

An interesting project developed at the University for Foreigners of Perugia by Stefania Spina is the *Perugia Corpus* (*PEC*), a reference corpus of contemporary Italian which gathers both oral and written texts (25Mw) distributed among 10 textual genres. It contains a legal subcorpus (1.1 Mw) made up of administrative texts (laws, regulations, European legislation). Another corpus developed by the same University is the Academic Italian Corpus (*AIC*), totaling 1Mw, which contains a legal academic subcorpus (330,000 w).

Although it is not a corpus comparable to those aforementioned, it is worth hinting at *Testi Amministrativi Chiari e Semplici* (*TACS*), a project coordinated by Michele Cortelazzo (University of Padua). It is a monolingual corpus of original Italian administrative texts produced by a number of administrative bodies (municipalities, regions, provinces, universities, ministries) and its 'translation'/ rewriting in a simplified language in the wake of the simplification of legalese and legal administrative language.

2.4 European Union

It goes without saying that the European Union holds the record of the largest – freely available – parallel corpora for the study of EU languages, including the legal domain.

The *JRC-Acquis* is a multilingual parallel corpus available in 23 languages which gathers, in its latest release (3.0), more than a billion words (1,055,583,954). It is an important tool to study the acquis communautaire, that is, the total body of EU law applicable in the EU Member States. The corpus comprises selected legislative texts written between the 1950s and now (treaties and laws, declarations and resolutions, international agreements on EU affairs and the judgments given by the Court of Justice).

Another corpus based on the same acquis communautaire is the *DGT Multilingual Translation Memory of the Acquis Communautaire* (*DGT-TM*), totaling 6,226,855 translation units in its latest release (2012).

A recent corpus is the *DGT-Acquis*, a family of several multilingual parallel corpora extracted from the *Official Journal of the European Union*, consisting of documents from the middle of 2004 to the end of 2011 in up to 23 languages.The corpus is aligned according to paragraphs and has 253 language combinations, totaling 3.54 million files.

Lastly, the *European Parliament Proceedings Parallel Corpus* 1996-2011 (*EUROPARL*) is a multilingual parallel corpus containing more than 60 million words per language based on the EP proceedings.

2.5 Rest of the world

In this final part of the section, attention will be focused on legal corpora built in countries different from England and Wales, Spain and Italy, entirely or partially dedicated to the study of legal or judicial language.

As for the former, that is, corpora exclusively dedicated to legal language, it is worth mentioning the *American Law Corpus* (*ALC*) compiled by Goźdź-Roszkowski (University of Łódz), which collects more than 5.5 million words extracted from seven legal genres typical of American culture and education (Goźdź-Roszkowski 2011: 27-30): academic journals, briefs, contracts, legislation, opinions, professional articles and textbooks. One of the main aims of the corpus is studying linguistic patterns and phraseology across these legal genres.

Another corpus for the study of American judicial language is the *USCC corpus*, built by Davide Mazzi (University of Modena e Reggio Emilia), made up of 67 opinions (658,154 words) delivered by the US Supreme Court, with the primary aim of studying judicial argumentation (see e.g. Mazzi 2010).

The *Case Law Corpus* developed in the *Centre for Computers and Law* (Erasmus University, Rotterdam) by van Noortwijk and De Mulder is a monolingual corpus gathering 3,073 judicial decisions (16.5 million words) delivered both by civil and criminal UK jurisdictions and courts.

The *Polish Law Corpus* is a monolingual corpus (PL) of 4 million words, built by Łucja Biel (University of Gdansk) which includes 211 codes and major legal acts related to contract, company, civil and criminal law (Biel 2010a). One of the main objectives of the author is describing nominal, verbal and adjectival collocations of legal terms within the context of an ongoing project aimed at compiling the *Dictionary of Polish Legal Collocations for Translators*.

As far as national – mostly monolingual – corpora are concerned, almost every national corpus has a subcorpus of legal language: the National Corpus of Polish (*NKJP, I-PAN corpus* in Xiao 2008: 387), the most representative corpus of Polish (5% of its 200 million words is taken from legal documents); the *CNC corpus* (CZ) (legal subcorpus: 0.82% of the SYN2000 subcorpus, totaling 100 million words); the *HNC corpus* (EL) has a subcorpus of legal documents among its 47 million words; the *SNK corpus* (SK) with its 719 million words has a legal subcorpus; the *MCLC corpus* (ZH) has its subcorpus of legal texts; etc. Although it is not a national corpus, the *INL 38 Million Corpus 1996* (NL) has a 12.9 million legal subcorpus.

Among the monolingual corpora for the study of the English language, including its legal domain, there are: the diachronic *Helsinki Corpus of English Texts* (University of Helsinki, Matti Rissanen and Ossi Ihalainen) which contains an entire section made up of common law texts; the *International Corpus of English*

(*ICE*), compiled by Josef Schmied's team (University of Hong Kong), which has a legal section of oral documents (legal presentations, 10,000 tokens; cross-examinations, 10,000 tokens); the *Academic Corpus* (Victoria University of Wellington) which contains 72 legal texts (874,723 tokens).

As far as multilingual corpora are concerned, it is worth mentioning the *Corpus Multilíngüe para Ensino e Tradução* (*COMET*), a bilingual (EN-PTbr) comparable corpus built at the University of São Paulo (Stella Esther Ortweiler Tagnin) which has 1 million words of legal language in its *CorTec* subcorpus (Commercial Law). Another interesting project is the *Hong Kong Bilingual Corpus of Legal and Documentary Texts* (EN: 300,000 tokens; ZH: 500,000 characters), compiled by Xu Xunfeng (Hong Kong PolyU); the *Hong Kong Parallel Text*, which has a legal subcorpus (Hong Kong Laws, EN: 8,396,243, ZH: 14,868,621 characters); the *ENPC corpus* (English-Norwegian Parallel Corpus) and the *ESPC corpus* (English-Swedish Parallel Corpus) both containing legal subsections.

Obviously, these are only some of the legal corpora available worldwide. Mention has been made of those consulted by the author of the present paper in an effort to shape his own PhD specialised corpus.

3. Conclusion

The brief survey which has been carried out in this paper has shown that, despite a national and international interest for the study of legal language through corpus linguistics tools, there is only a small number of real, systematic, multilingual corpora for its study, in a contrastive perspective, especially if compared with the huge number of corpora of general language identified by Xiao in his 2008 study.

Legal corpora represent a promising tool in legal linguistics, as they can be exploited in innumerable applications, such as terminology, phraseology, syntax, textual structures, genre analysis, etc.

Taking stock of the analysis, a number of considerations are required: there are dozens of corpora made up of exclusively legal and judicial documents; most of them are monolingual or, if not, comparable; few of them adopt a contrastive, cross-linguistic perspective. More interesting for the purpose of the ongoing PhD research project is that, with the exception of few scholars (e.g. Biel, Goźdź-Roszkowski, Mazzi), legal phraseology has not been studied systematically, either by linguists or translation scholars, with a corpus-based or -driven methodology.

The ongoing PhD project is conceived as a first, tentative step towards filling that gap.

References

Bhatia V. K., Langton N.M. & Lung J. (2004) "Legal discourse: opportunities and threats for corpus linguistics", in *Discourse in the Professions. Perspectives from Corpus Linguistics*. Ed. by U. Connor & T. A. Upton, Amsterdam/Philadelphia, John Benjamins pp. 203-231.

Biel Ł. (2010a) "Corpus-based studies of legal language for translation purposes: methodological and practical potential", in *Reconceptualizing LSP*. Ed. by C. Heine & J. Engberg. Outline Proceedings of the XVII European LSP Symposium 2009, Aarhus 2010. Also available at: http://www.asb.dk/fileadmin/www.asb.dk/isek/biel.pdf (last accessed on 27 November 2012).

Biel Ł. (2010b) "The textual fit of legal translations: focus on collocations in translator training", in *Teaching Translation and Interpreting: Challenges and Practices*. Ed. by Ł. Bogucki, Newcastle upon Tyne, Cambridge Scholars Publishing, pp. 25-39.

Gotti M. (2010) "CADIS - A corpus for the analysis of identity traits in academic discourse", in *Fachsprachen in der weltweiten Kommunikation: Akten des 16. Europäischen Fachsprachensymposiums Hamburg 2007*. Hg. von C. W. von Hahn & C. Vertan, Frankfurt am Main, Peter Lang, pp. 421-430.

Goźdź-Roszkowski S. (2011) *Patterns of Linguistic Variation in American Legal English. A Corpus-based Study*, Frankfurt am Main, Peter Lang.

Grover C., Hachey B. & Hughson I. (2004) "The HOLJ corpus: supporting summarisation of legal texts", in *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora* (LINC-04), Geneva, Switzerland. Also available at: http://www.ltg.ed.ac.uk/SUM/PUBS/linc04-final.pdf (last accessed on 27 November 2012).

Laviosa S. (2010) "Corpora", in *Handbook of Translation Studies*. Vol 1. Ed. by Y. Gambier & L. V. Doorslaer, Amsterdam/ Philadelphia, John Benjamins, pp. 80-86.

Mazzi D. (2010) "This argument fails for two reasons... A linguistic analysis of judicial evaluation strategies in US Supreme Court judgments", *International Journal for the Semiotics of Law*, 23:4, pp. 373-385.

McEnery T., Xiao R. & Tono Y. (2006) *Corpus-based Language Studies: An Advanced Resource Book*, London/New York, Routledge.

McEnery T. & Wilson A. (2001) *Corpus Linguistics*, 2nd edition, Edinburgh, Edinburgh University Press.

Monzó E. (2008) "Corpus-based activities in legal translation training", *The Interpreter and Translator Trainer*, 2:2, Manchester/ Kinderhook, St. Jerome Publishing, pp. 221-251.

Rossini Favretti R., Tamburini F. & Martelli E. (2001) "Words from Bononia Legal Corpus", *International Journal of Corpus Linguistics,* 6 (Special Issue), pp. 13-34. Also available at: http://corpora.dslo.unibo.it/People/Tamburini/Pubs/TCML_2007.pdf (last accessed on 27 November 2012).

Rossini Favretti R., Tamburini F. & De Santis C. (2002) "CORIS/CODIS: A corpus of written Italian: a defined and dynamic model", in *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Ed. by A. Wilson, P. Rayson & T. McEnery, Munich, Lincom-Europa. Available at: http://corpora.dslo.unibo.it/People/Tamburini/Pubs/CL2001.pdf (last accessed on 27 November 2012).

Stubbs M. (1996) *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*, Oxford, Blackwell.

Taranilla G. R. (2011) *La configuración narrativa en el proceso penal. Un análisis discursivo basado en corpus*. Tesis doctoral. Universitat de Barcelona. Departament de Filologia Hispànica. Also available at: http://www.tdx.cat/handle/10803/48717 (last accessed on 27 November 2012).

Tognini-Bonelli E. (2001) *Corpus Linguistics at Work*, Amsterdam/Philadelphia, John Benjamins.

Zanettin F. (2012) *Translation-driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*, Manchester/Kinderhook, St. Jerome Publishing.

Xiao R. (2008) "Well-known and influential corpora", in *Corpus Linguistics. An International Handbook*. Vol. 1. Ed. by A. Lüdeling & M. Kyto, Berlin, Mouton de Gruyter, pp. 383-457.

Appendix

URLs [last accessed on 25 November 2012]

Corpus	*URL*
Academic Corpus:	http://www.victoria.ac.nz/lals/resources/academicwordlist/information/corpus
AIC	http://elearning.unistrapg.it/corpora/aic.html
ALC	N/A
BLaRC	N/A
BoLC	http://dslo.unibo.it/bolc_eng.html
CADIS	http://dinamico.unibg.it/cerlis/page.aspx?p=245
Case Law Corpus	N/A
CCLE	http://www.cambridge.org/gb/elt/catalogue/subject/item2701617/Cambridge-English-Corpus/?site_locale=en_GB
CLUVI	http://sli.uvigo.es/CLUVI/index_en.html
CNC	http://ucnk.ff.cuni.cz/english/index.php
CODIS	http://dslo.unibo.it/CODIS/ [http://corpora.ficlit.unibo.it/]
COMET	http://www.fflch.usp.br/dlm/comet/
CORIS	http://dslo.unibo.it/TCORIS/ [http://corpora.ficlit.unibo.it/]
CORPUS	http://www.iula.upf.edu/corpus/corpuses.htm
COSPE	N/A
CPP	N/A
DGT-ACQUIS	http://ipsc.jrc.ec.europa.eu/index.php?id=783
DGT-TM	http://ipsc.jrc.ec.europa.eu/index.php?id=197
ENPC	http://www.hf.uio.no/ilos/english/services/omc/enpc/
ESPC	http://www.sol.lu.se/engelska/corpus/corpus/espc.html
GARALEX	http://www.ehu.es/ehusfera/garalex/
GENTEXT-N	N/A
Helsinki Corpus of English Texts	http://icame.uib.no/hc/
HNC	http://hnc.ilsp.gr/en/
HOLJ	http://www.ltg.ed.ac.uk/SUM/CORPUS/index.html
Hong Kong Bilingual Corpus of Legal and Documentary Texts	http://langbank.engl.polyu.edu.hk/corpus/bili_legal.html
Hong Kong Parallel Text	http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2004T08
ICE	http://ice-corpora.net/ice/index.htm
INL	http://listserv.brown.edu/archives/cgi-bin/wa?A2=ind9608&L=TEI-L&P=3060
JRC-Acquis	http://langtech.jrc.it/JRC-Acquis.html
JUD-GENTT	N/A
MCLC	http://www.clr.org.en/retrieval
NKJP	http://nkjp.pl/index.php?page=0&lang=1
PEC	http://perugiacorpus.unistrapg.it/composizione.html
Polish Law Corpus	N/A
Proceedings of the Old Bailey	http://www.oldbaileyonline.org/
SNK	http://korpus.juls.savba.sk/stats_en.html
TACS	http://www.maldura.unipd.it/buro/tacs.html
USSC	N/A

Selected web pages containing updated lists of existing corpora:

David Lee: http://www.uow.edu.au/-dlee/CBLLinks.htm
Manuel Barbera: http://www.bmanuel.org/clr/index.html
Richard Xiao: http://www.lancs.ac.uk/fass/projects/corpus/cbls/corpora.asp