Design and compilation of a legal English corpus based on UK law reports: the process of making decisions

María José Marín Pérez
Universidad de Murcia

Camino Rea Rizzo
Universidad de Murcia

ABSTRACT

The scarceness of reliable specific teaching materials and corpora within the field of legal English has led us, as lecturers of this ESP variety, to engage into corpus design. The available corpora existing do not satisfy our needs as we intend to establish the core vocabulary of this ESP branch, so we have opted for creating our own, BLaRC: British Law Report Corpus. The selected genre are law reports (judicial decisions) as they stand at the very basis of common law systems (uncodified) where the existing jurisprudence plays a determining role. The purpose of this paper is thus to show and justify the decisions we have made in its process of compilation and design.

KEYWORDS: *legal corpus, ESP, common law, representativeness, corpus size, word target.*

RESUMEN

La escasez de materiales y corpus específicos fiables dentro del área del inglés jurídico nos ha llevado, como profesoras de esta variedad de inglés específico, a involucrarnos en el diseño de un corpus. Los corpus específicos existentes no satisfacen nuestras necesidades pues nuestro objetivo es establecer el vocabulario básico de esta variedad, por este motivo hemos optado por crear nuestro propio corpus, BLaRC: British Law Report Corpus*. El género seleccionado son los* law reports *(decisiones judiciales) ya que son una pieza fundamental en los sistemas legales* common law *(no codificados) en los que la jurisprudencia existente juega un papel esencial. El propósito de este artículo es por tanto mostrar y justificar las decisiones que se han tomado en su proceso de diseño y compilación.*

PALABRAS CLAVE: *corpus legal, ESP,* common law*, representatividad, tamaño del corpus, word target*.

1. INTRODUCTION

It is commonly agreed that the amount of available teaching materials in English for Specific Purposes is considerably scarce in most fields (Rea, 2010). This derives into a clear methodological void which must be filled, thus resorting to specific corpora becomes a valuable method for ESP professionals. As McEnery and Wilson affirm: "[...] such corpora can be used to provide many kinds of domain-specific material for language learning" (1996: 121).

One of the main obstacles we encountered as lecturers of legal English was deciding on what method and particularly what materials to select in order to teach this specially obscure variety of English. As stated above, resorting to specific corpora was an interesting option, however, to our knowledge, the amount of written legal corpora is also reduced and designing our own became an urge. We thus engaged into ESP corpus design and decided to create the British Law Report Corpus (BLaRC): a legal English corpus of law reports (judicial decisions) that could act as a reliable source of specific vocabulary which could be employed to create new materials, as well as information for further linguistic analysis.

The purpose of this paper is thus to present the process of design and compilation of BLaRC according to Corpus Linguistics standards as stated in Wynne (2005) for general corpora and its adaptation to specific corpora (Rea, 2010). First, we will focus on the state of the art by looking into the legal corpora available; next, we will justify the reasons that lead to the selection of this legal genre. The mode of the texts and the organization of the corpus into different categories will also be explained to finish with some final remarks on further corpus applications and future research.

2. STATE OF THE ART

By using the term *state of the art* we are actually referring to the amount of available legal corpora we have found and their main goals and characteristics. As stated above, the amount of such corpora is scarce and the purpose they were created with, in most cases, did not satisfy our need for a specific legal corpus we could employ to identify the core vocabulary of law reports (a key genre within legal English) as well as to carry out further linguistic analyses.

The first corpus worth mentioning is the BoLC since this is probably the most comprehensive legal corpus existing due to its selection of texts from varied genres and topics. It is a multilingual comparable Italian-English corpus which aims at "representing the two different legal systems, in particular the differences between the civil law and the common law systems"[19].

However, the rest of corpora we found were either too small to act as a normative reference for us, or inaccessible. They either focused on aspects of the language we are

19 This quotation has been taken from the corpus website.

not interested in, or were conceived as parallel corpora with translational or comparative purpose.

The JRC-Acquis Corpus is one of them. It is a multilingual parallel corpus which includes European Union legislative texts affecting all Members States in twenty-two different languages.

The CorTec corpus is a scientific-technical parallel one divided into four sections, one of them deals with commercial law and includes agreements and contracts in English and Brazilian Portuguese.

As for the HOLJ corpus, it is a monolingual synchronic one comprising 188 judgments of the House of Lords from 2001 to 2003, its aim is to define a set of rhetorical role labels.

To finish, the Cambridge International Corpus, owned by Cambridge University Press, has a legal corpus section of twenty million words. It is not accessible or commercialised.

There also exist legal sections or materials within some of the best known general British English corpora like BNC or COBUILD, but they could not serve our purpose either as they are non-specific.

3. LAW REPORTS: A LEGAL GENRE AT THE CORE OF COMMON LAW SYSTEMS

Establishing the *sampling frame,* that is, "the entire population of texts from which we [would] take our samples" (McEnery and Wilson, 1996: 78), was our first objective, and law reports were selected due to the pivotal role they play in the UK judicial system as well as in any other common law countries. Following Sinclair: "the contents of the corpus should be selected (...) according to their communicative function in the community in which they arise" (Wynne, 2005: 5).

If representativeness is crucial for the design of any corpus (Biber, 1993; McEnery and Wilson, 1996; Sánchez, Cantos, Sarmiento and Simón, 1995; Sinclair, 1991; Wynne, 2005), narrowing the boundaries of our object of study became a must as we soon realised how legal language is intertwined with everyday language, how it is present both in the public and private fields, and consequently how the vastness of this ESP branch could not be covered or managed as a whole in a project of this nature. Therefore, we decided to focus on one of the most relevant legal genres in this ESP variety: law reports.

The United Kingdom belongs to the realm of common law, as opposed to civil or continental law which is the judicial system working in most Western European countries. In purely common law systems, the acts passed at their parliaments have gained greater importance being most often cited in case decisions. However, case law stands at the very basis of common law systems which rely on the principle of binding precedent to work, that is to say, a case judged at a higher court must be cited and applied whenever it is similar to the one being heard in its essence (the *ratio dicendi*). Another fact that makes

law reports an outstanding genre in common law legal systems is that they not only cover all the branches of law, but also touch upon other public and private law genres.

Due to the widespread use of information technologies, there is a tendency towards digitalising these texts and storing them in online databases, so case citation has recently become an easier task that used to take ages for legal practitioners to become fully informed about the existing jurisprudence in the past.

There are different ways of accessing cases online, most of them are restricted. However, the British and Irish Legal Information Institute (BAILII.org) has created a completely free and comprehensive online database where more than 200,000 authentic texts are available. It is supported by a number of sponsors like the Inns of Court (barristers' professional associations), law faculties like Cambridge, Oxford, Glasgow, law firms and other prestigious institutions, hence its importance and recognition by professionals. This is precisely why we have decided to use it as our main source to obtain the legal texts that form BLaRC.

4. LEXICAL VARIATION AND CORPUS STRUCTURE

As well as abiding by hierarchical criteria when organizing the corpus, one of the first elements that conditioned our choice was the way that legal vocabulary varies according to the system where it is used. This is so because of the laws and regulations that organise the countries which the UK is divided into. The judicial systems of Northern Ireland, Scotland, England and Wales do not solely depend on UK institutions, but rather have their own autonomous systems and structure. But for the Supreme Court (in general terms) and the UK Tribunal Service (except for some cases), each country is fully independent as regards its judicial system.

This being so, we decided to structure BLaRC into five main categories depending on the jurisdictions of their judicial systems, that is, the geographical scope of their courts and tribunals:

1. Commonwealth countries
2. United Kingdom
3. England and Wales
4. Northern Ireland
5. Scotland

5. TIME SPAN COVERED BY BLARC

BLaRC is a specific synchronic monolingual corpus of judicial decisions made in the UK. Following Pearson: "(…) a specific corpus compiled for terminological studies, [should

include texts] (…) delivered in the last 10 years prior to the date of compilation" (1998: 51), this is why we decided to compile the texts produced at UK courts and tribunals from 2008 to 2010 as we expect to finish compiling them before the end of 2011.

Moreover, due to the changes that the structure of these courts has experienced due to the recent modifications of the law that regulates it, we considered that, if the structure of the corpus responds to the structure of UK courts and tribunals because of thematic and hierarchical reasons -as it will be shown below-, it should adjust to the latest modifications it has experimented following the new regulations, hence the time span covered. We are specifically referring to the *Constitutional Reform Act, 2005* and the *Tribunals, Courts and Enforcement Act, 2007*.

6. TEXT MODE

Oral texts were disregarded given that obtaining oral samples of legal language that reflected the arguments, facts, and decisions dealt with at court, would have implied having access to courtrooms and permission to record the trial sessions, a certainly complicated objective for Spanish researchers merely interested in linguistic data. Therefore, BLaRC only covers the written mode and in raw text format, as the corpus is not intended to be tagged.

Regarding the texts themselves, they are authentic transcriptions of judicial decisions whose structure may vary depending on the nature of the case and the hierarchical position of the court where it was heard. They are full texts in digital format from BAILII gathered randomly within the time span established.

The average size of the texts is 2,000 to 2,500 words, although there is great variation. They have all been produced by British judges and reflect their decisions about the cases in question as well as the facts, arguments, prior decisions made at other courts, and any other kind of information relevant to the case.

7. SIZE AND REPRESENTATIVENESS: KEY ELEMENTS IN CORPUS DESIGN

Representativeness is central to corpus design, as shown above, and the size of a corpus may determine whether it is representative of the variety of language it aims at covering, or simply an illustrative sample of it with no predictive value.

Authors do not agree regarding the recommended size for a specific corpus. Whereas Pearson (1998) proposes a million words as a reasonable number, Sinclair (1991) believes that corpora must be as large as possible. On the other hand, Kennedy (1998) does not think that a big corpus necessarily represents the language better than a small one.

Taking these arguments into consideration as well as the availability of legal texts and their high numbers (16,612 texts in total from 2008 to 2010), we established our word target. Also the relevance of law reports in the judicial system coupled with the great

amount of topics covered by these texts, was determining when we had to make the first decisions on the size of our corpus.

As a consequence, we established that, although this is a specific corpus based just on one legal genre, the target should be 6,000,000 words (approximately), six times as big as Pearson proposes, essentially because of the easy access to already digitalized texts in either .rtf or .pdf format and, naturally, all the principles behind corpus compilation.

8. THE LEXICAL COMPREHENSIVENESS OF LAW REPORTS

Law reports should not only be paid special attention within ESP because of their essential function in common law systems, but also because of their vast topic coverage. This corpus has been organised according to the source where the texts originated, that is, what court or tribunal cases were heard at and decided on.

Tribunals and courts are specialized in a given branch of law: criminal law, family law, commercial law, intellectual law, etc., and law reports touch upon one and every branch of both the private and public fields. Judges are in charge of judging cases by both interpreting the law itself (the statutes passed at the parliament), and fundamentally taking into consideration the existing precedents. Therefore their judgments, as reflected on law reports, pertain to all the fields of law.

9. BLaRC STRUCTURE AND DISTRIBUTIONAL CRITERIA

McEnery and Wilson highlight the importance of justifying the categorisation of any corpus when citing Biber: "Biber ... emphasises the advantage of determining beforehand the hierarchical structure (or strata) of the population, that is, defining what different genres, channels and so on it is made up of" (1996: 79), this is why we believed it was essential to do so.

To begin with, our corpus retains the current UK tribunal and court structure as reflected on BAILII due to several reasons, the first one being the relevance of the hierarchy of courts and tribunals in the UK legal system. The principle of binding precedent, which the British judicial system revolves around, establishes that any decision made at a higher court or tribunal will set binding precedent as long as the case is similar to the one under examination, as stated above.

Secondly, if we maintain this structure, the texts will be grouped according to the field of law they belong to, so they will be similar in lexical terms, and comparing results by studying the categories separately will be easier and respond to a thematic criterion we consider fundamental as far as our further objective is concerned, that of establishing the core vocabulary of the genre.

To finish with the enumeration of the criteria that have conditioned the organization of the corpus, we would like to refer to the distribution of the population in the UK. As it

is shown in the UK official census 2011, elaborated by the Office of National Statistics, it appears that almost 90% of the population of the whole territory is concentrated in England and Wales while Northern Ireland only has about 3 % and Scotland 9%. Although we have not mathematically distributed the number of texts and word targets per category and subcategory depending on these figures, we did take them into account in order to reinforce the representativeness of the texts obtained from English and Welsh sources that amounted to approximately 55% of the total[20].

10. Final remarks

This paper has aimed at presenting all the stages followed in the design and ongoing compilation of a new corpus of legal English which may satisfy the linguistic needs of ESP students. As we firmly believe that the quality of the results deriving from corpus analysis depends crucially on the rigorous establishment of the corpus, we have closely observed the principles governing corpus compilation and tried to apply them to the design, collection and projection of BLaRC.

Taking advantage of the law reports made available on digital databases on the internet, we have access to a vast amount of naturally occurring samples of the language used by the judges that explain an order in any type of cases, and therefore, covering all possible issues reaching the court. Corpus Linguistics' techniques permit dealing with such amount of authentic samples and process them in such a way that we could obtain worthy and reliable results from the analysis of the language from several approaches.

An essential tool for corpus analysis is the computing programme selected for its processing. WordSmith.5 will be used to look into the samples through quantitative and qualitative analyses, first by assessing BLaRC's basic computational characteristics (types, tokens, type-token ratio, frequency lists, etc.) and second, by adopting a corpus comparison approach which enables to gain a deeper insight into legal English.

Even though BLaRC has been envisaged to serve multiple purposes in the long term, since the potential applications of a corpus are manifold, our overriding objective consists in identifying the essential vocabulary in legal English for the ease of teaching and learning. Moreover, we aim at filling in a gap for discipline-based lexical repertoires which may guide materials writers, assist ESP practitioners and notably meet students' specific needs (Nation, 2001; Hyland and Tse, 2007; Read, 2007; Rea, 2008). The framework of our future research is set by the long tradition of developing word lists (Coxhead, 2000; Nation, 1990; Thorndike and Lorge, 1944; West, 1953) for teaching and learning English as a second language.

20 The structure and distribution of the word targets per section and subsection will be exemplified in Appendix 1.

APPENDIX 1

This table exemplifies the structure of BLaRC divided into five main categories which are subdivided according the court and tribunal structure in each of them respectively. The UK court and tribunal section (number two in the general structure) comprises twenty-two subcategories, the distribution of the word targets in each of them has been made according to the number of texts available with respect to the total and also with respect to the 6m overall word target of the corpus. We have kept the numeric order main categories have been assigned in the general structure of the corpus.

Table A1

2. UK courts and tribunals

Court/ Tribunal	Available Texts	% Of Total	Word Target
2.1.Supreme Court	117	0,71%	42,600
2.2. House of Lords	74	0,45%	27,000
2.3. Upper Tribunal (Administrative Appeals Chamber)	550	3,31%	198,600
2.4.Upper Tribunal (Tax and Chancery)	44	0,27%	16,200
2.5.Upper Tribunal (Immigration and Asylum Chamber)	59	0,36%	21,600
2.6. Upper Tribunal (Lands Chamber)	135	0,82%	49,200
2.7. First Tier General Regulatory Chamber	124	0,75%	45,000
2.8. First-tier Tribunal (Health Education and Social Care Chamber)	139	0,84%	50,400
2.9. First-tier Tribunal (Tax)	865	5,21%	312,600
2.10. Competition Appeals Tribunal	100	0,61%	36,600
2.11. Nominet UK Dispute Resolution Service	370	2,23%	133,800
2.12. Special Immigrations Appeals Commission	24	0,15%	9,000
2.13. Employment Appeal Tribunal	971	5,85%	315,000
2.14. Financial Services and Markets Tribunal	16	0,1%	6,000
2.15. Asylum and Immigration Tribunal	141	0,85%	51,000
2.16. Information Tribunal including the National Security Appeals Panel	130	0,79%	47,400
2.17. Special Commissioners of Income Tax	80	0,49%	29,400
2.18. Social Security and Child Support Commissioners	219	1,32%	79,200
2.19. VAT & Duties Tribunals (Customs)	20	0,12%	7,200
2.20. VAT & Duties Tribunals (Excise)	92	0,56%	33,600
2.21. VAT & Duties Tribunals (Insurance Premium Tax)	1	0,01%	600
2.22. VAT & Duties Tribunals (Landfill Tax)	2	0,02%	1200

REFERENCES

ALCARAZ VARÓ, E. (1994). *El inglés jurídico: textos y documentos.* Madrid: Ariel Derecho.

ALCARAZ VARÓ, E. (2000). *El inglés profesional y académico.* Madrid: Alianza Editorial.

BHATIA, V. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

BHATIA, V. (2004). Applied genre analysis: a multi-perspective model. *Iberica 4,* pp 3-19.

BIBER, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing,* 8 (4).

BIBER, CONRAD AND REPPEN. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: C.U.P.

BoLC (Bononia Legal Corpus). Disponible en http://corpora.dslo.unibo.it/bolc_eng.html

CONSITUTIONAL REFORM ACT, 2005. Disponible en:
http://www.legislation.gov.uk/ukpga/2005/4/contents

DUDLEY-EVANS, T. AND ST JOHN, M. (1998). *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.

HUTCHISON, T. AND WATERS, A. (1998). *English for Specific Purposes*. Cambridge University Press.

HYLAND, K. & P. TSE (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41*(2), 235-253.

KENNEDY, G. (1998). *An introduction to corpus linguistics*. New York: Longman.

KENNEDY, G. Y BOLITHO, R. (1984). *English for specific purposes*. London: Mcmillan.

MALEY, Y. (1987). The Language of Legislation. *Language and Society, 16.*

MCENERY, T. AND WILSON, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MELLINKOFF, D. (1963). *The Language of the Law.* Boston: Little, Brown & Co.

NATION, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

ORTS LLOPIS, M.A. (2006). *Aproximación al discurso jurídico en inglés: las pólizas de seguro marítimo de Lloyd's.* Madrid: Edisofer.

PEARSON, J. (1998). *Terms in Context*. Amsterdam: John Benjamins Publishing Company.

REA, C. (2008). *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico*. Tesis doctoral. Disponible en http://www.tesisenred.net/TDR-0611109-134048/index_cs.html

REA, C. (2010). Getting on with Corpus Compilation: from Theory to Practice. *ESP World, 1* (27), Volume 9.

READ, J. (2007). Second Language Vocabulary Assessment: Current Practices and New Directions. *International Journal of English Studies*, *7* (2). Universidad de Murcia.

ROSSINI, R ET AL. (2001). Words from the Bononia Legal Corpus. *International Journal of Corpus Linguistics,* Vol. 6 (special issue), 13-34

SÁNCHEZ, A., CANTOS, P., SARMIENTO R., SIMÓN, J. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

SINCLAIR, J. (1991). *Corpus, Concordance and Collocation*. Oxford: Oxford University.

THORNDIKE, E.L. AND LORGE, I. (1944). *The teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

TIERSMA, P. (1999). *Legal Language.* Chicago: The University of Chicago Press.

TRIBUNALS, COURTS AND ENFORCEMENT ACT, 2007. Disponible en http://www.legislation.gov.uk/ukpga/2007/15/contents

WYNNE, M. (Ed.) (2005).*Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: ASDS Literature, Languages and Linguistics.