ATLANTIS
Journal of the Spanish Association of Anglo-American Studies
41.2 (December 2019): 63-84
ISSN 0210-6124 | e-ISSN 1989-6840
DOI: http://doi.org/10.28914/Atlantis-2019-41.2.03

Mark-up and Annotation in the *Corpus of Historical English Law Reports* (CHELAR): Potential for Historical Genre Analysis

Paula Rodríguez-Puente[a], Cristina Blanco-García[b] and Iván Tamaredo[c]
Universidad de Oviedo[a], Universidade de Santiago de Compostela[b], Universidade de Vigo[c]
rodriguezppaula@uniovi.es, cristina.blanco@usc.es

Adding annotation and mark-up to linguistic corpora has become a standard practice in corpus building over the past few decades as a way to facilitate data extraction and at the same time guarantee that new corpora are compatible with existing and future tools. The purpose of this article is twofold. First, we provide an overview of the main forms of annotation and mark-up available to the research community and how they have been applied to the *Corpus of Historical English Law Reports 1535-1999* (CHELAR), a specialized corpus consisting of law reports or records of judicial decisions. Second, we give an account of preliminary research based on the annotated versions of CHELAR, which so far has been primarily aimed at identifying the distinctive linguistic characteristics of law reports, as well as at investigating how the language of law reports has evolved over a time span of almost five centuries. Our article illustrates the multiple advantages of applying a simple annotation schema to a corpus and how this can enhance the potential of a corpus for historical genre analysis.

Keywords: corpus annotation; corpus mark-up; law reports; TEI-XML; legal English

. . .

Etiquetado y anotación lingüística en el *Corpus of Historical English Law Reports* (CHELAR): potencial para el análisis textual desde la perspectiva histórica

El etiquetado y anotación lingüística de un corpus se ha convertido en una práctica generalizada en las últimas décadas, con el objetivo de facilitar la extracción de datos del propio corpus, así como garantizar que los corpus nuevos son compatibles con otros ya existentes o creados con posterioridad. El presente artículo tiene una doble finalidad. Por una parte, proporcionamos un análisis de las principales formas de etiquetado y anotación que están a nuestra disposición. Tomando como ejemplo el *Corpus of Historical English Law Reports 1535-1999* (CHELAR), un corpus especializado de decisiones judiciales, ilustraremos el modo en que estos etiquetados pueden implementarse de forma sencilla para su posterior aprovechamiento. En segunda instancia, presentamos un resumen de los trabajos de investigación que se han llevado a cabo hasta la fecha con CHELAR, y que se han centrado principalmente en identificar las características lingüísticas significativas de las decisiones judiciales, así como en investigar la evolución del lenguaje de estos documentos durante casi cinco siglos. Nuestro trabajo ilustra las múltiples ventajas de la implementación de un sistema de anotación simplificado y las mejoras que este supone para el análisis textual desde la perspectiva histórica.

Palabras clave: anotación de corpus; etiquetado de corpus; decisiones judiciales; TEI-XML; inglés legal

1. Introduction

Adding annotation and mark-up to linguistic corpora has become standard practice in corpus building over the past few decades as a way to facilitate data extraction and at the same time guarantee that new corpora are compatible with existing and future tools. In this article we provide an overview of the main forms of annotation and mark-up available to the research community and how they have been applied to the *Corpus of Historical English Law Reports 1535-1999* (CHELAR) (Rodríguez-Puente et al. 2016, 2018). Although they are closely related concepts, annotation here refers to the encoding of linguistic features, whereas mark-up is used for the code system that contains information about the text itself.

CHELAR is a specialized corpus consisting of law reports or records of judicial decisions (Fanego et al. 2017). The corpus was made available in 2016 as plain text and with part-of-speech (POS) annotation, and was continued in 2018 by CHELAR v.2, an extensively revised and enhanced second version of the corpus that has served as the basis for the creation of its TEI-XML edition. Although multiple forms of annotation are readily available to the research community, we have limited the annotation and mark-up of CHELAR to POS and TEI-XML for two main reasons: (1) they facilitate the extraction from the corpus of the data required to satisfy the needs of our initial research questions, some of which are briefly expounded in section 5 below; and (2), we wanted to avoid the addition of excessive tags which would be both time-consuming and might turn the corpus into a less readable and user-friendly tool.

In section 2 we describe the main types of linguistic annotation that have been applied to linguistic corpora over the past few decades together with some of the advantages and drawbacks of implementing them. Section 3 is concerned with corpus mark-up, focusing specifically on the TEI-XML schema and its applications in corpus linguistics. In section 4, we provide an overview of CHELAR and how the annotation and mark-up systems described previously have been applied to it. The article concludes with an overview of preliminary research based on CHELAR (section 5), and with some conclusions and suggestions for future work (section 6).

2. Corpus Annotation

A prototypical linguistic corpus is made up of a set of machine-readable (written or spoken) texts that aim to be representative of a particular language, dialect, register, medium of production or even speaker, and which comprise data produced in real-life communicative situations (Gries and Berez 2017, 380). Texts should also be selected to represent a balanced sample of speakers, registers, mediums and/or varieties, that is, the amount of material for one particular subsample—for example, spoken language—should reflect the importance of this category in the population. Besides including transcripts of original unprocessed language, most corpora nowadays contain information regarding the phonological, lexical, grammatical, semantic or structural features of texts. The process whereby all this information is made explicit is called *annotation* and, even though annotating a corpus can be very time-consuming, the outcome of this process is a powerful tool for linguistic research that provides a corpus with "added value" (Leech 2005, section 1)—information can be easily retrieved, either manually or automatically, thus greatly facilitating the task of linguists when conducting their research, and the same set of annotations can be used by many different researchers with diverse interests and aims. In spite of this, corpus annotation also has detractors who argue that it produces cluttered corpora that impose a particular linguistic analysis and make corpora less accessible, updateable and expandable (see, for example, Hunston 2002, 91-95). Indeed, corpus annotation can only be useful if annotations are unambiguous as to their meanings and transparent, that is, with easily interpreted tags that are short enough to facilitate the task of reading an annotated text. As such, corpus builders may decide to include an untagged or raw version of the corpus to satisfy the needs of those who argue against any type of annotation.

The types of linguistic information that corpora can be annotated for are varied, since the motivations behind the compilation of corpora can also be extremely diverse. The following types of annotation are commonly added to corpora (Leech 2005, section 2; McEnery et al. 2006, 29-45; Gries and Berez 2017, 383-92):

- Lemmatization: the process of marking each word in the corpus for its lemma or base form stripping away inflectional morphology
- Part-of-speech annotation (POS): identification of the word class—noun, verb, preposition, etc.—of each word in the corpus
- Syntactic annotation: segmentation of the corpus into phrasal and clausal units
- Semantic annotation: addition of information about the specific sense in which words are used, their semantic category, issues such as the aspect, modality, polarity and factuality of constructions or, less frequently, metaphorical and metonymical phenomena
- Discourse annotation: signaling of anaphoric relations between elements in the corpus or the way in which information is presented and structured
- Stylistic annotation: identification of how speech and thought are represented in a text by means of direct discourse, indirect discourse or free indirect discourse

Additionally, spoken corpora can be annotated for phonetic, phonological and prosodic information and even pragmatic features such as illocutionary force or contextual information of the speech situation. Other more specialized kinds of annotation include gestures in multimodal corpora, signs in sign language corpora, translations of original texts into other languages in parallel corpora and errors in learner corpora.

Leech (2005, section 4) puts forward a series of recommendations for corpus annotators. First, and despite the fact that annotations are indeed useful for many types of linguistic enquiries, the annotated version of the corpus should always be accompanied by the original unprocessed texts for those users whose research goals do not require, or could even be hindered by, the presence of annotations. Second, corpus compilers should provide comprehensive documentation including, among other things, the annotation and coding schemes used, the quality of annotations and information regarding where, when, by whom and by means of which tools the annotation process was carried out. Third, the set of categories on which the annotations are based should be sufficiently general to be compatible with most linguistic approaches and theories. Finally, newly compiled corpora should comply with de facto standards of linguistic annotation, that is, coding schemes, formats and practices in widespread use and with sufficient acceptance and recognition in the research community.

3. Corpus Mark-up

Annotation is but one part of corpus construction, which can contribute to the creation of powerful, flexible and accessible corpus resources. Corpus mark-up is closely related to corpus annotation but rather consists in providing a corpus with some degree of standardization to guarantee that it is compatible with both existing and future tools and that it can be extensively exploited by the research community. As defined by Tony McEnery et al., corpus mark-up "is a system of standard codes inserted into a document stored in electronic form to provide information *about* the text itself and govern formatting, printing or other processing" (2006, 22; italics in the original). In this section we describe one of the most widely used and recommended languages in encoding corpora, eXtensible Markup Language (XML) (Bray et al. 2008), along with its predecessors.

The first attempts at marking up corpus files followed COCOA references, which consisted of a set of attributes and values enclosed in angle brackets that were quite limited in terms of the number of features they could encode (McEnery and Wilson 2001, 34-35; McEnery et al. 2006, 23). More recently, some more ambitious mark-up schemes have emerged (for an overview, see McEnery et al. 2006, 23), but the Standard Generalized Markup Language (SGML) is considered to be the first fully systematized and flexible approach to encoding not only structural mark-up but also text-level and corpus-level metadata as well as analytic annotations (Hardie 2014, 74). The best-known application of SGML for corpus encoding is the Text Encoding Initiative (TEI), sponsored by three major academic associations—the Association for Computational Linguistics, the Association for Literary and Linguistic Computing and the Association for Computers and Humanities—the first edition of which was published in the 1990s (Sperberg-McQueen and Burnard 1990, iii-iv). The TEI Guidelines, aimed at facilitating data exchange by standardizing the mark-up of information stored electronically, have been published periodically ever since, the most recent version being the *TEI P5 Guidelines*, published in 2013, although the online edition is constantly being updated.

Even though SGML was a great step in marking up corpus files based on the use of angle brackets to delimit the tags from the actual corpus text, it was far from ideal. As a result, in the late 1990s XML came to be favored over SGML for most text-encoding purposes, and has now become the most widespread form of annotation. It enables structural annotations by means of tags, or elements, which represent information by virtue of their names and the attribute/value pairs associated with them (Carletta et al. 2004, 455).

In corpus linguistics the more user-friendly TEI-XML (Text Encoding Initiative-eXtensible Markup Language) mark-up language, used for storing and transporting data based on its inherent structure, has become the standard system adopted in digitally-based humanities for research on both historical and present-day English. It is, for instance, the system employed in *A Representative Corpus of Historical English Registers* 3.2, the *British National Corpus*, the *Helsinki Corpus*, the *Old Bailey Corpus*, the *Coruña Corpus of Early Scientific Writing* and the *Late Modern English Medical Texts 1700-1800* (Taavitsainen et al. 2014), among others. Elements in a body of data are marked with customizable tags that can be further defined using attributes (Gries and Berez 2017, 393). XML documents must adhere to proper syntax and are designed to be machine- and human-readable; additionally, they can be easily converted into other formats, such as databases (Carletta et al. 2004, 455-57; Gries and Berez 2017, 393-94).

In spite of the multiple benefits of adding TEI-XML mark-up to a corpus, it must be acknowledged that the accepted standards may be found "top-heavy" and "over-engineered" (Hardie 2014, 77) for three main reasons: (1) the complexity of the TEI standards; (2) the over-weighty level of mark-up suggested by those standards; and (3) the degree of technical knowledge required both to use and apply the TEI standards and to handle the software employed for that purpose (Hardie 2014, 77-78). Thus, Andrew Hardie suggests the use of a "modest" (2014, 82-103) XML schema for corpora in order to avoid becoming entangled with all the technical details of XML or the full weight of TEI encoding.

4. An Example of Modest Annotation and Mark-up: The *Corpus of Historical English Law Reports*

CHELAR is a specialized corpus consisting of law reports, records of judicial decisions that can be used as precedent in subsequent cases. It contains approximately half a million words and is structured into nine periods of fifty years each, running from 1535 to 1999 (Rodríguez-Puente 2011; Fanego et al. 2017).[1] Following Geoffrey Leech's assertion that "adding annotation to a corpus is giving 'added value'" to it (2005, section 1) and in order to take advantage of the multiple and varied forms of annotation at the

[1] The first subperiod is slightly larger, running from 1535 to 1599.

disposal of corpus builders, the initial version of the corpus (2016) was released in two different formats: (1) as raw text files, and (2) with part-of-speech (POS) annotation. The latter was implemented by means of CLAWS C7 (Garside 1987), with accuracy rates that ranged between 95.5% and 98.5% (Fanego et al. 2017, 66-69). The latest version of the corpus, CHELAR v.2 (2018), has been extensively revised and enhanced and differs from its predecessor in several respects. First, the word count is slightly higher in the second version, mostly due to the correction of typos encountered during the revision of the initial release as well as to the addition of further textual material in the case of some files. Second, the extralinguistic data of the texts has been enhanced by including information on who the reporters of the cases were and the exact date of the publication of the reprints from which the CHELAR texts were obtained. The revised texts of CHELAR v.2 have also served as the basis for the creation of the TEI-XML version of the corpus.

The TEI-XML edition of CHELAR follows the *TEI P5 Guidelines* developed by the Text Encoding Initiative Consortium (Bray et al. 2008). These were implemented by means of the software *Oxygen XML Editor*, which facilitated the insertion of the various tags and helped ensure that the texts in the corpus were well-formed and suited the XML standards. Although the annotation possibilities of the TEI-XML schema are infinite, priority was given to a type of annotation that satisfied the needs of the texts, yet at the same time facilitated a varied range of corpus analyses. In general, a modest XML tagging was advocated for, that is, "modesty at the level of corpus markup" and "modesty of scope arising from the avoidance of all advanced aspects of XML" (Hardie 2014, 80). Thus, the annotation system is rather minimalist in that it only includes some renditional and structural features, editorial corrections and conceptual characteristics, but avoids typical applications of heavyweight standards which are of little interest to corpus linguistics, such as, for example, peculiarities of the original appearance of the texts (Hardie 2014, 75). In the remainder of this section, we provide an account of how the TEI-XML schema was applied in CHELAR, the specific tags and attributes used and the problem-solving process followed during the annotation of the corpus.

4.1. The TEI Header

Being TEI-conformant, in every single CHELAR file the body of the text(s) is preceded by a TEI header that carries information about the text itself, its sources and its encoding. Although a TEI header can be a very large and complex element, the CHELAR headers have been simplified so as to include only the most relevant information, encoded in three major parts: the *file description* (fileDesc), the *encoding description* (encodingDesc) and the *profile description* (profileDesc).

The file description contains a full bibliographic account of every file in the corpus and is further subdivided into smaller sections following the schema below:

```
<fileDesc>
 <titleStmt>
      </titleStmt>
      <editionStmt>
      </editionStmt>
      <publicationStmt>
      </publicationStmt>
      <sourceDesc>
      </sourceDesc>
      <extent>
      </extent>
</fileDesc>
```

The *title statement* (titleStmt) carries information about the title, the author and publication date of the text. The *edition statement* describes the names of the corpus compilers and provides the full reference of the corpus and the source from which the texts were extracted, together with acknowledgment of the funding projects. Next, the *publication statement* (publicationStmt) gives the name of the file where the text was stored, information about the authorship of the corpus, its availability and date of release. The subsection *source description* (sourceDesc) includes a detailed account of the content of the particular text to which the header refers, namely its full title and reference, the court where the case reported was judged, the name of the case, the judges—only when this information could be retrieved—and the date(s) on which the case was heard. Finally, the last subitem of the file description refers to the *extent*, that is, the length of the text in number of words.

The encoding description documents the relationship between the electronic text and the source(s) from which it was obtained and is likewise subdivided into smaller sections structured as follows:

```
<encodingDesc>
      <projectDesc>
      </projectDesc>
      <samplingDecl>
      </samplingDecl>
      <editorialDecl>
      </editorialDecl>
</encodingDesc>
```

The *project description* (projectDesc) summarizes the different stages through which the compilation process went and the dates of completion of each of these phases. The *sampling declaration* (sampligDecl) gives details of the postediting process of the samples selected for the corpus. In the CHELAR headers it refers to the noninclusion of footnotes (Rodríguez-Puente 2011, 111), the preservation of blank lines and spaces, as well as the correction of unclear or blurred punctuation (110). The last item in this section is the *editorial declaration* (editorialDecl), mostly concerned with clarifying how lettering size, special characters, quotation marks, text-alignment, indentation and hyphenation have been treated in the corpus documents.

The profile description is the last item of the TEI header and accounts for nonbibliographic aspects of the text. In CHELAR, this section is restricted to information about the language variety (British English) and the domain of the texts: Law Reports for those texts produced from 1865 onwards and English Reports for those produced before 1865 (Fanego et al. 2017, 56-60). Additionally, TEI headers may also include non-TEI metadata (xenoData) and a revision description (revisionDesc) summarizing the revision history of a file. These, however, are not part of the CHELAR headers.

4.2. The TEI Body

The item <body> contains the whole body of a single unitary text. In the case of CHELAR, the body of the text comprises one or more law report(s) produced in the same year. The texts are preceded by an XML ID that provides a unique identifier for the corresponding file. Our initial goal was to include samples of approximately 2,500 words in each of the files, but in many cases one single report would not suffice to reach that target, especially when dealing with the oldest reports (Fanego et al. 2017, 65). In cases where more than one law report had to be included in the XML file, a further identifier was added to mark the division of the different texts. This appears as `<div type="TEXT" n="1">`, where the attribute-value pair at the end of the tag (`n="1"`) refers to the number of the text in that file. In the earliest corpus subperiods, some files consist of up to twenty different texts, meaning that this label has to be repeated twenty times and the value of the number changes with every new text. Following these identifiers, the text itself is preceded by the tag <head>, which indicates the abbreviated title of the particular text (e.g., *Regina v. Woolin*).

As we did not wish to overload the texts with unnecessary mark-up, but rather advocated for a minimalist tagging system, the focus of our mark-up system was on renditional features, structural features, editorial corrections and conceptual features. Renditional features in CHELAR include italics, used for multiple purposes in law reports (Rodríguez-Puente 2011, 108-10), and superscript numbers. In both these cases, we employed the tag <hi>, normally used to mark graphically distinct words or phrases, followed by the appropriate attribute (rend) and its value—either “italics” or “superscript”— as in (1) and (2).

```
(1) <hi rend="italics">the joint account of the above parties</hi>
(2) 23<hi rend="superscript">1/2</hi>
```

Structural features in CHELAR encompass titles (<head>), paragraphs (<p>), line breaks (<lb\>), page breaks (<pb>), as well as the text divisions mentioned earlier in this section with regard to those files which contain more than one report. For the sake of simplification, minor units such as sentences, utterances, words and the like were not marked in the corpus texts. Sample (3) shows an example of how paragraphs, line breaks and page breaks have been annotated.

(3)

```
<p>
 <lb/>APPEAL from <name type="person">Brandon J.</name>
 <lb/>By a writ dated <date when="18-06-1968">June 18, 1968</
date>, the plaintiffs, La Plata Cereal Co. S.A.,
 <lb/>of <place>Buenos Aires</place> ("shippers"), André &amp;
Cie. S.A. ("Swiss André"), of
 <lb/><place>Lausanne</place>, Comptoir Commercial André Cie.,
of <place>Marseilles</place> ("French
 <lb/>André") and Sorveglianza, of <place>Rome</place>
(collectively described as owners of
 <lb/>the <foreign lang="sp">cargo</foreign> lately laden on
board the vessel the <hi rend="italics">Annefield</hi>), began an
action
 <lb/><foreign lang="la">in rem</foreign> against the
vessel owned by a Liberian company, Asimarfield <choice
n="hyphenation"><orig>Ship-<lb break="no"/>ping</
orig><reg>Shipping
 <lb/></reg></choice> Corporation ("the shipowners") for damages
for breach of contract
  [...]
</p>
[...]
<pb n="171"/>
```

As can be seen in (3), unlike most XML tags, those for line breaks are self-closing—indicated by a bar before the closing tag—and are introduced at the beginning of every new line. The last tag in the example corresponds to the start of a new page in the original document; the attribute-value pair of the tag indicates the number of the new page. The extract in (3) also contains an example of a word that was split over two lines by means of a hyphen in the original source text (*Shipping*). Three tags were necessary to mark words like these. First, <choice>, which groups a number of alternative encodings for the same point in a text, specified by the attribute value "hyphenation"; then <orig>, indicating the original reading of the word (*Ship-ping*); and, finally, <reg>, providing the regularized form. The latter tag was also necessary for the regularization of original page numbers. In CHELAR those reports produced before 1865 are reprints of original versions of reports (Fanego et al. 2017, 56-64). The tag <pb> in our texts refers to the page breaks in the new edited versions of the reports, whereas original page breaks appear in the body of the text (Rodríguez-Puente 2011, 112), occasionally splitting a word in two (e.g., re-[123]-nounce). These split words had to be regularized by placing the page number after the originally split word (see example 4), in order to facilitate corpus searches, as well as the addition of further annotation, either manually or automatically (e.g., POS).

```
(4) <reg orig=”re-[123]-nounce”>renounce [123]</reg>
```

Blank spaces and text omissions have likewise been marked in the XML files by means of the tag <gap>. These are quite frequent in the CHELAR texts, especially when anonymizing proper names, as in (5), or when the printing of the original source text was unclear, as in (6). As can be seen in examples (5) and (6), the specific attributes between quotation marks help identify the reason for the gap in the source text.

```
(5) <lb/>of his manor of <gap reason=”blank space”/> in the county aforesaid
(6) <lb/>47, 48; <gap reason=”unclear”/>
```

Similarly, typos in the original documents were corrected in the XML files using the tag <sic>, a very useful method of making clear that the error was already in the source text, rather than being committed in the transcription or manipulation of texts during the process of corpus building (see example 7).

```
(7) <sic corr=”does”>doe</sic>
```

Another important structural feature marked in the CHELAR files concerns the use of quotations and the transcription of direct speech. The most modern reports typically start with a short summary of the case and then proceed with a direct narration of the facts, that is, a record of the actual words of the participants in the case. Short-hand notation, stenographic transcription and tape recording have enabled reporters to gradually provide more accurate reproductions of what actually happened in court, so that these sections of direct speech, though recorded in writing, are meant to represent real conversations between the parties involved, between the parties and the judge(s), between the judges themselves—if more than one was present—between any of the former and the witnesses during questioning, and also the judges’ monologues produced in the process of presenting their arguments, their reasoning, the judgment they arrive at and the way they do it (Bhatia 1993, 119). These sections of direct speech, however, are far from constituting faithful representations of real speech and must have been subject to much editorial intervention and "correction" on the part of the judges themselves before publication (Mitchell 2015, 38-40), since they do not include features typical of spoken language, such as slips of the tongue, false starts, hesitations or bad language and insults, which are frequently present in other legal records, such as trial proceedings (Kytö and Walker 2003, 225; Widlitzki and Huber 2016). However, they constitute the most oral sections of the reports in CHELAR, which contrast with other primarily prescriptive sections that are plagued with cross-references to other precedent cases and other legal documents. Linguistically, the sections of direct speech in the law reports in CHELAR have already been shown to differ from the more prescriptive sections and from legal documents such as statutes, Acts of Parliament and proclamations as far as the use of personal pronouns is concerned (Rodríguez-Puente 2019). The most oral sections in law reports tend to be characterized by an extensive use of first and second person pronouns, which contrasts with the typical detached, impersonal style generally adopted in legal documents and confers on them a greater degree of subjectivity, involvement and interpersonality.

The TEI-XML system allows for various methods of marking direct speech. In the *Old Bailey Corpus*, for example, the selected tag for direct speech was <speech>, normally used for an individual speech in a performance text or a passage presented as such in a prose or verse text. Given that the so-called direct speech sections in CHELAR do not constitute real speech events but rather edited reproductions of them, we decided to mark those sections with the tag <q>, which contains material distinguished from the surrounding text by using quotation marks or a similar method to represent, among other possible features, direct speech. Quotation marks are not used for these exchanges in the reports of the CHELAR files; rather, these sections of direct speech are introduced by the name(s) of the person(s) speaking, as in example (8), which in XML would be rendered as shown in example (9).

(8)
J. A. Plowman Q.C. and *D. A. Thomas* for the defendant.
Jenkins L.J. held that the boundaries on the filed plan were only
general boundaries...
[...]
D. A. Thomas following. No reliance can be placed on the
transfer plan. It is only a rough plan and the measurements
on it are inaccurate.
[...]

(9)

```
<lb/><name type=”person”><hi rend=”italics”>J. A. Plowman Q.C.</
hi></name> and <name type=”person”><hi rend=”italics”>D. A.
Thomas</hi></name> for the defendant.
    <lb/><q><name type=”person”>Jenkins L.J.</name> held that
the boundaries on the filed plan were only
    <lb/>general boundaries
[...]
<lb/><name type=”person”><hi rend=”italics”>D. A. Thomas</hi></
name> following. <q>No reliance can be placed on the
    <lb/>transfer plan. It is only a rough plan and the
measurements
    <lb/>on it are inaccurate.</q>
[...]
```

In contrast, in order to mark those phrases or passages attributed to some agency external to the text marked by quotation marks in the source texts, we employed the tag <quote>. In our corpus, this tag normally reproduces quotes from books, laws, statutes, other reports, wills, letters, contracts or quotes of someone else's words, among other sources. Example (10) shows a quotation (<quote>) from a contract produced within direct speech (<q>).

(10)

```
<lb/><q>On the true construction of this covenant the tenant has
the
<lb/>right to use the demised premises for a restaurant,
<quote>”including
<lb/>the sale of tobacco, confectionery and other similar
produce,”</quote>
<lb/>and has an alternative right to use the whole premises as
<quote>”offices
<lb/>and showrooms for such business to be first approved by the
<lb/>landlord.”</quote><q>
```

As far as conceptual features are concerned, we restricted tags to indicate names of persons and cases, place names and foreign words. For persons, we used the tag <name> with the corresponding attribute value "person", as shown in example (11). The tag <name> was also used for cases, this time with the attribute value "case", as shown in example (12). Place names, including names of cities, counties, villages, states, manors, streets, etc. were labelled as <place>, as in example (13).

```
(11) <name type=”person”>George II</name>
(12) <name type=”case”>Law Society v. United Service Bureau Ltd. 1 K.B. 343</name>
(13) [...] a dwelling house known as St. Leonard’s, <place>The Street</place>, <place>Staple</place>, <place>Canterbury</place>, <place>Kent</place> [...]
```

For dates, we employed the tag <date> followed by the attribute "when" and the value of the date after the equal symbol with the schema dd-mm-yy. When information about day, month or year was not retrievable from the text, we completed the value of the attribute with the word *none* (see example 14).

```
(14) <date when=”18-04-none”>18th of April</date>
```

Finally, foreign words were marked as <foreign> further specified by the attribute value of the corresponding language of origin—"la" for Latin, "sp" for Spanish, "fr" for French, "AN" for Anglo-Norman, "ger" for German and "it" for Italian—as shown in (15).

```
(15)
<foreign lang=”la”>donatio mortis causa</foreign>
<foreign lang=”sp”>flotilla</foreign>
<foreign lang=”fr”>mesne</foreign>
<foreign lang=”AN”>succour</foreign>
<foreign lang=”ger”>Europaische</foreign>
<foreign lang=”it”>ditto</foreign>
```

The language of origin was ascertained by means of the *Oxford English Dictionary* (*OED*), though not without difficulties. Deciding which words to include as foreign was a difficult task because it implied making decisions as to the degree of integration of those words into the language. Additionally, being a diachronic corpus, what might be regarded as foreign in the sixteenth century might not be considered foreign today. In order to be as consistent as possible, when a word was marked as foreign it was labelled as such throughout the whole corpus. As far as the degree of integration is concerned, the classification of words and expressions as foreign was mostly based on spelling: when words maintained the spelling of their language of origin in their English form, they were labelled as foreign. Thus, *memorandum*, *lieu* and *flotilla*, for example, were labelled as Latin, French and Spanish respectively, although all three words have been well integrated in the English language for a long time. Even though CHELAR is relatively small in size for lexical analysis, the <foreign> tags might be particularly useful for the analysis of certain aspects, such as binomials, that is, coordinations of a native term or a well-integrated loan word and its foreign (near-)synonym—e.g., "bargain and sale" (Nevalainen 1999, 363)—so typical of legal language, as well as for investigation on the extent of use of specialist terms and expressions derived from Latin and Norman French, also common in legal documents.

5. Preliminary Research Based on Chelar

Interest in legal English has grown exponentially over the last decades, as the result of developments in (applied) linguistics and social sciences generally (Bhatia 1987, 227). Law-related fields of language study are manifold (for an overview, see Fanego and Rodríguez-Puente 2019, 2-5) and the creation of annotated diachronic corpora like CHELAR can help broaden the research scope of linguists by providing, as in our case, a balanced sample of a specific type of legal document extending over five centuries. Teresa Fanego et al. (2017, 69-72) outlined a number of broad research trajectories that can be explored by means of CHELAR. In the remainder of this section, we provide an account of how some of those trajectories have been pursued to date.

The law comprises a wide variety of activities, all of which must be recorded in written form, so that "legal documents are classified under a very large number of text types" (Görlach 1999, 145) with different structural, formal and linguistic features. As a text type, law reports can be characterized as "hybrid" in nature (Šarčević 2000, 11), because they fulfil both prescriptive (regulatory) and descriptive (informative) functions—for these labels, see Tiersma (1999, 139-41), Šarčević (2000, 11-12) and Williams (2007, 28-29), among others—and therefore their linguistic features and structural elements differ from those of other legal documents which are purely prescriptive and regulatory. Initial research with CHELAR has thus been primarily aimed at identifying differences between law reports and other legal and formal genres, as well as at investigating how the language of law reports has evolved over a time-span of almost five centuries.

In this connection, Paula Rodríguez-Puente's (2019) study of personal pronouns in CHELAR provides strong evidence that the language of law reports is indeed distinctive, as already mentioned in Section 4.2 above. Due to its overall function, the law tends to adopt an impersonal style, steering clear of markers of subjectivity and (inter)personal involvement, such as first and second person pronouns (Tiersma 1999, 67-68; Sancho-Guinda et al. 2014, 13). Yet first and second person pronouns are relatively common in law reports, especially in those sections that portray transcriptions of dialogues and monologues recorded in the trials (see section 4 above), and also as a resource through which judges assert their claim to speak as an authority (Rodríguez-Puente 2019, 186). In fact, most first-person pronouns in law reports correspond to the nominative singular form (*I*), a feature that makes this type of legal document stand out from other documents more prescriptive in character, such as Acts of Parliament, proclamations and statutes, where first-person singular pronouns are rare and first-person plural pronouns are mostly used as majestic plurals (royal *we*), that is, referring to a single person holding a high office (2019, 182-88).

The findings in Rodríguez-Puente (2019) also challenge the common assumption that legal written texts are resistant to change (Tiersma 1999, 135) and "outside the 'ravages of time'" (Görlach 1999, 145), since substantial variation can be found when observing developments in the language of law reports from a diachronic perspective. As far as personal pronoun usage is concerned, law reports become more involved over time, with the most recent reports displaying statistically significant higher rates of first-person pronouns than the oldest reports sampled in CHELAR. Evidence for increasing rates of personal involvement in law reports over time is also supported by ongoing research on the active/passive alternation in this type of document (Rodríguez-Puente 2018a). Passives are typically associated with formal academic prose, as markers of a detached, impersonal style (Biber 1988, 228; Biber et al. 1999, 476; Seoane 2006a, 2006b, 2013), so that, as a formal written type of discourse, legal English is known to make extensive use of passive structures (Hiltunen 1990, 76-77; Tiersma 1999, 74-77; Williams 2004; Williams 2007, 35-36). However, although, as expected, passive clauses outnumber actives in CHELAR, from the nineteenth century onwards passives display a tendency to decrease in frequency, particularly from the second half of the twentieth century, probably as a response to the so-called Plain Language Movement, which recommends the avoidance of passive verb forms whenever possible (Seoane and Williams 2006, 124; Williams 2007, 177; Williams 2013).

The size and structure of CHELAR also render it suitable to address the analysis of the trend towards the increasing use of colloquial linguistic features that can be observed in written genres over the past two centuries, especially in the last decades of the twentieth century (see, among others, Biber and Finegan 1989, 1997; Mair 1997, 2006; Atkinson 1999; Hundt and Mair 1999; Leech et al. 2009; Rühlemann and Hilpert 2017). A recent study by Douglas Biber and Bethany Gray (2019) has examined this issue based on the law reports in CHELAR, and has compared the development of this genre over the period 1700 to 1999 to developments in three other registers: science research articles, newspaper articles and fiction. Newspaper articles and fictional texts are representative of popular written registers that are agile in the sense that they adopt colloquial innovations from spoken discourse. In contrast, law reports appear as a "relatively uptight register" (Biber and Grey 2019, 166) because, although they have adopted some colloquial innovations, their frequencies remain low. Linguistic features traditionally associated with literate discourse, such as nominalizations, attributive adjectives, relative clauses and noun complement clauses, have increased in law reports over time, according to Biber and Gray's findings.

Nominalizations were explored in greater depth, also on the basis of CHELAR, in Rodríguez Puente (2018b). They are a feature generally taken as characteristic of legal (Bhatia 1993; Tiersma 1999; Williams 2004; Mattiello 2010) as well as scientific genres (Halliday and Martin 1993; Atkinson 1999; Banks 2005; Tyrkkö and Hiltunen 2009), both of which are representative of formal written discourse intended for a specific audience. Stanisław Goźdź-Roszkowski argues that "the dense use of nominalization [in legal writing] can be attributed to the highly specialist and informational nature of this text variety" (2011: 21), because nominalizations, like passives and prepositions, have the function of conveying highly abstract—as opposed to situated—information, and thus bear a positive load on Biber's Dimension 3, "Explicit vs. Situation-Dependent Reference" (1988).

Rodríguez-Puente's (2018b) analysis showed that, whereas the rates of nominalizations remain relatively similar between 1535 and 1799, their frequency increases significantly from 1800 onwards, peaking in the second half of the twentieth century. This is in agreement with the results of Biber and Gray's study mentioned earlier (2019), but is unexpected considering that excessive use of nominalization goes against the recommendations of the Plain Language Movement (Williams 2007, 177). However, given the considerable reduction in the use of the passive voice described by Rodríguez-Puente (2018a), the increased rates of nominalization in law reports seems to respond to the use of a different strategy to portray an impersonal kind of discourse, legal English being primarily "nouny" rather than "verby" (Williams 2013, 354).

In sum, initial research with CHELAR has already shown that law reports are a distinctive text type as far as their linguistic features and structure are concerned, both synchronically and diachronically. These preliminary studies are, however, far from providing a thorough account of the language typical of these documents. The research possibilities of an annotated tool like CHELAR are infinite and further research should aim at investigating the language of law reports from a broader variety of perspectives and trajectories.

6. Conclusions and Further Research

This article has put forward the manifold advantages of corpus annotation and mark-up. Multiple and complex forms of annotation are at our disposal nowadays, but an excess of superimposed tags on a raw corpus text can confound rather than facilitate the aims of a researcher. However, when applied sensibly and with a modest scope, the research possibilities of an initially raw corpus can be improved substantially. This was our primary goal when compiling CHELAR. Although relatively small in size, CHELAR can be employed for a wide range of research topics thanks in part to the enhancement provided by a simple annotation schema. It must be acknowledged that once the TEI-XML schema has been implemented in a corpus, adding other types of annotation becomes relatively easy, and that probably the CHELAR texts could be further enhanced by combining annotation and mark-up features, for example, embedding the POS annotated texts within the TEI-XML schema. Although we do not rule out this possibility in the future, for the time being we prefer to keep the different types of annotation separate so as not to end up producing a cluttered corpus, thereby complying with our initial idea of applying a modest tagging system. As things stand, research based on CHELAR to date has consistently shown the solid potential uses and applications of the corpus for both diachronic and synchronic research.[2]

Works Cited

Aarts, Bas et al., eds. 2013. *The Verb Phrase in English: Investigating Recent Language Change with Corpora*. Cambridge: Cambridge UP.

Atkinson, Dwight. 1999. *Scientific Discourse in Sociohistorical Context. The Philosophical Transactions of the Royal Society of London, 1675-1975*. Mahwah, NJ: Lawrence Erlbaum.

Banks, David. 2005. "On the Historical Origins of Nominalized Process in Scientific Text." *English for Specific Purposes* 24 (3): 347-57.

Bhatia, Vijay K. 1987. "Language of the Law." *Language Teaching* 20 (4): 227-34.

—. 1993. *Analysing Genre: Language Use in Professional Settings*. Harlow: Pearson Education.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge UP.

Biber, Douglas and Edward Finegan. 1989. "Drift and the Evolution of English Style: A History of Three Genres." *Language* 65 (3): 487-517.

—. 1997. "Diachronic Relations among Speech-Based and Written Registers in English." In Nevalainen and Kahlas-Tarkka 1997, 253-75.

Biber, Douglas et al. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas and Bethany Gray. 2019. "Are Law Reports an 'Agile' or an 'Uptight' Register? Tracking Patterns of Historical Change in the Use of Colloquial and Complexity Features." In Fanego and Rodríguez-Puente 2019, 149-69.

Bray, Tim et al., eds. 2008. *Extensible Markup Language (XML) 1.0*. 5th edition. W3C Recommendation 26 November 2008. [Accessed online on July 10, 2019].

Breeze, Ruth, Maurizio Gotti and Carmen Sancho-Guinda, eds. 2014. *Interpersonality in Legal Genres*. Bern: Peter Lang.

Burnard Lou and Syd Bauman, eds. 2013. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Charlottesville, VI: Text Encoding Initiative Consortium.

Carletta, Jean et al. 2004. "A Generic Approach to Software Support for Linguistic Annotation Using XML." In Sampson and McCarthy 2004, 449-59.

Dalton-Puffer, Christiane et al., eds. 2006. *Syntax, Style and Grammatical Norms: English from 1500-2000*. Bern: Peter Lang.

[2] We would like to thank Professor Teresa Fanego (University of Santiago de Compostela) for helpful discussion and valuable feedback on earlier versions of this article. We are grateful to the Spanish Ministry of Science, Innovation and Universities for their generous financial support for the research project "Constructionalization in Contemporary and Earlier English: Cognitive, Variationist and Discourse-Pragmatic Perspectives" (FFI2017-86884-P) and to the Regional Government of Galicia for grants ED431B 2017/12 and ED431D 2017/09.

Fanego, Teresa et al. 2017. "The *Corpus of Historical English Law Reports 1535-1999* (CHELAR): A Resource for Analysing the Development of English Legal Discourse." *ICAME Journal* 41: 53-82.

Fanego, Teresa and Paula Rodríguez-Puente. 2019. "'Why May not that Be the Skull of a Lawyer?' English Legal Discourse Past and Present." In Fanego and Rodríguez-Puente 2019, 1-21.

Fanego, Teresa and Paula Rodríguez-Puente, eds. 2019. *Corpus-Based Research on Variation in English Legal Discourse*. Amsterdam and Philadelphia: John Benjamins.

Garside, Roger. 1987. "The CLAWS Word-Tagging System." In Garside, Leech and Sampson 1987, 30-41.

Garside Roger, Geoffrey Leech and Geoffrey Sampson, eds. 1987. *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.

Goźdź-Roszkowski, Stanisław. 2011. *Patterns of Linguistic Variation in American Legal English. A Corpus-Based Study*. Bern: Peter Lang.

Görlach, Manfred. 1999. *Nineteenth-Century England: An Introduction*. Cambridge: Cambridge UP.

Gries, Stefan and Andrea Berez. 2017. "Linguistic Annotation in/for Corpus Linguistics." In Ide and Pustejovsky 2017, 379-409.

Halliday, M. A. K. and James R. Martin. 1993. *Writing Science: Literary and Discursive Power*. London: Falmer Press.

Hardie, Andrew. 2014. "Modest XML for Corpora: Not a Standard, but a Suggestion." *ICAME Journal* 38: 73-103.

Hiltunen, Risto. 1990. *Chapters on Legal English: Aspects Past and Present of the Language of the Law*. Helsinki: Suomalainen Tiedeakatemia.

Hundt, Marianne and Christian Mair. 1999. "'Agile' and 'Uptight' Genres: The Corpus-Based Approach to Language Change in Progress." *International Journal of Corpus Linguistics* 4 (2): 221-42.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge UP.

Ide, Nancy and James Pustejovsky, eds. 2017. *Handbook of Linguistic Annotation*. Berlin and New York: Springer.

Jucker, Andreas, Daniel Schreier and Marianne Hundt, eds. 2009. *Corpora: Pragmatics and Discourse*. Amsterdam and New York: Rodopi.

Kytö, Merja and Terry Walker. 2003. "The Linguistic Study of Early Modern English Speech-Related Texts: How 'Bad' Can 'Bad' Data Be?" *Journal of English Linguistics* 31 (3): 221-48.

Lass, Roger, ed. 1999. *The Cambridge History of the English Language*. Vol. 3, *1476-1776*. Cambridge: Cambridge UP.

Leech, Geoffrey. 2005. "Adding Linguistic Annotation." In Wynne 2005, 17-29.

LEECH, Geoffrey et al. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge UP.

LJUNG, Magnus, ed. 1997. *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*. Amsterdam: Rodopi.

LÓPEZ-COUSO, María José, et al. eds. 2016. *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora*. Leiden: Brill.

MAGRATH, Paul, ed. 2015. *The Law Reports 1865-2015*. London: The Incorporated Council of Law Reporting for England and Wales.

MAIR, Christian. 1997. "The Corpus-Based Approach to Language Change in Progress." In Ljung 1997, 195-209.

MAIR, Christian. 2006. *Twentieth-Century English: History, Variation and Standardization*. Cambridge: Cambridge UP.

MATTIELLO, Elisa. 2010. "Nominalization in English and Italian Normative Legal Texts." *ESP Across Cultures* 7: 129-46.

MCENERY, Tony and Anita Wilson. 2001. *Corpus Linguistics*. 2nd ed. Edinburgh: Edinburgh UP.

MCENERY, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

MITCHELL, Paul. 2015. "Between Speech and Writing." In Magrath 2015, 37-46.

NEVALAINEN, Terttu. 1999. "Early Modern English Lexis and Semantics." In Lass 1999, 332-459.

NEVALAINEN, Terttu and Leena Kahlas-Tarkka, eds. 1997. *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique.

OXFORD ENGLISH DICTIONARY ONLINE. 2019. [Accessed online on June 17, 2019].

RODRÍGUEZ-PUENTE, Paula. 2011. "Introducing the *Corpus of Historical English Law Reports*: Structure and Compilation Techniques." *Revista de Lenguas para Fines Específicos* 17: 99-120.

—. 2018a. "On the Active/Passive Alternation in Law Reports." Paper presented at the 42nd AEDEAN Conference, Córdoba, April 2018.

—. 2018b. "Frequency and Productivity of Nominalizations in Law Reports: A Diachronic Perspective." Paper presented at the 20th International Conference of English Historical Linguistics, Edinburgh, August 2018.

—. 2019. "Interpersonality in Legal Written Discourse. A Diachronic Analysis of Personal Pronouns in Law Reports, 1535 to Present." In Fanego and Rodríguez-Puente 2019, 171-99.

RODRÍGUEZ-PUENTE, Paula et al., comps. 2016. *Corpus of Historical English Law Reports 1535-1999 (CHELAR)*, v.1. Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, University of Santiago de Compostela.

RODRÍGUEZ-PUENTE, Paula et al., comps. 2018. *Corpus of Historical English Law Reports 1535-1999 (CHELAR)*, v.2. Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, University of Santiago de Compostela.

RÜHLEMANN, Christoph and Martin Hilpert. 2017. "Colloquialization in Journalistic Writing." *Journal of Historical Pragmatics* 18 (1): 101-35.

SAMPSON, Geoffrey and Diana McCarthy, eds. 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London and New York: Continuum.

SANCHO-GUINDA, Carmen, Maurizio Gotti and Ruth Breeze. 2014. "Framing Interpersonality in Law Contexts." In Breeze, Gotti and Sancho-Guinda 2014, 9-35.

ŠARČEVIĆ, Susan. 2000. *New Approach to Legal Translation*. The Hague: Kluwer Law International.

SEOANE, Elena. 2006a. "Information Structure and Word Order Change: The Passive as an Information Rearranging Strategy in the History of English." In van Kemenade and Los 2006, 360-91.

—. 2006b. "Changing Styles: On the Recent Evolution of Scientific British and American English." In Dalton-Puffer et al. 2006, 191-221.

—. 2013. "On the Conventionalisation of the Passive Voice in Late Modern English Scientific Discourse." *Journal of Historical Pragmatics* 14 (1): 70-99.

SEOANE, Elena and Christopher Williams. 2006. "Questions of Style: Legal Drafting Manuals and Scientific Style Manuals in Contemporary English." *Linguistica e Filologia* 22: 115-37.

SPERBERG-MCQUEEN, Michael and Lou Bernard, eds. 1990. *TEI P1: Guidelines for the Encoding and Interchange of Machine-Readable Texts*. Chicago, IL and Oxford: Text Encoding Initiative.

TAAVITSAINEN, Irma, et al. 2014. "'Late Modern English Medical Texts 1700-1800': A Corpus for Analysing Eighteenth-Century Medical English." *ICAME Journal* 38: 137-53.

TIERSMA, Peter M. 1999. *Legal Language*. Chicago, IL: The U. of Chicago P.

TYRKKÖ, Jukka and Turo Hiltunen. 2009. "Frequency of Nominalization in Early Modern English Medical Writing." In Jucker, Schreier and Hundt 2009, 297-320.

VAN KEMENADE, Ans and Bettelou Los, eds. 2006. *The Handbook of the History of English*. Oxford: Blackwell.

WIDLITZKI, Bianca and Magnus Huber. 2016. "Taboo Language and Swearing in Eighteenth Century and Nineteenth Century English: A Diachronic Study Based on the 'Old Bailey Corpus'." In López-Couso et al. 2016, 313-36.

WILLIAMS, Christopher. 2004. "Pragmatic and Cross-Cultural Considerations in Translating Verbal Constructions in Prescriptive Legal Texts in English and Italian." *Textus* 17 (1), 217-46.

—. 2007. *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts*. 2nd ed. Bern: Peter Lang.

—. 2013. "Changes in the Verb Phrase Legislative Language in English." In Aarts et al. 2013, 353-71.

Wynne, Martin, ed. 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow.

Received 1 May 2019 Accepted 27 June 2019

Paula Rodríguez-Puente is Assistant Professor (tenure track) of English language and linguistics at the University of Oviedo. Her research interests include English historical linguistics and corpus linguistics. She has published widely in international peer-reviewed journals and edited volumes for John Benjamins, Peter Lang and Cambridge Scholars. Her monograph on the history of phrasal verbs was published by Cambridge University Press.

Address: Departamento de Filología Inglesa, Francesa y Alemana. Facultad de Filosofía y Letras, Campus El Milán. Universidad de Oviedo. Calle Amparo Pedregal s/n. 33011 Oviedo, Spain. Tel.: +34 985104570.

Cristina Blanco-García holds an MA in English Language and Literature from the University of Santiago de Compostela. She is currently working on her PhD on ephemeral adverbial subordinators in the history of English, with a special focus on concessives, conditionals and causals. Her research interests also include linguistic variation and change, grammaticalization processes and corpus linguistics.

Address: Departamento de Filología Inglesa y Alemana. Facultad de Filología. Universidade de Santiago de Compostela. Avda. de Castelao, s/n. 15782 Santiago de Compostela, Spain. Tel. +34 981563100.

Iván Tamaredo, formerly a predoctoral researcher at the University of Santiago de Compostela, is currently a Lecturer in English at the University of Vigo. His research interests include varieties of English and linguistic complexity. He has presented papers at international conferences and published articles in peer-reviewed journals such as *English World-Wide*, *English Language and Linguistics* and *ICAME Journal*.