Annotating the *Corpus of Historical English Law Reports 1535-1999*

Cristina Blanco-García
Universidade de Santiago de Compostela
cristina.blanco@usc.es

1. Introduction

The *Corpus of Historical English Law Reports 1535-1999* (CHELAR) is a specialised diachronic corpus of legal English containing approximately 500,000 words (for a detailed overview of the corpus, see Rodríguez-Puente 2011; Fanego *et al.* 2017).

2. Novelties in CHELAR

- First diachronic corpus of law reports, *i.e.*, records of judicial decisions which are "cited by lawyers and judges for their use as precedent in subsequent cases" (*Encyclopedia Britannica Online* s.v. *law report*).
- Broad timespan, covering nearly five centuries (1535-1999).
- It compares favourably with the legal component in *ARCHER 3.2. –A Representative Corpus of Historical English Registers, version 3.2.*, also consisting of law reports (for details, see López-Couso and Méndez-Naya 2012)

3. Annotation: POS

The corpus contains part-of-speech mark-up done by means of CLAWS-7 tagger (Constituent Likelihood Automatic Word-tagging System; see Garside 1987). Tagging was not a totally automatic process, as the programme does not recognise non-ASCII characters (see Table 1). Tags were tested for accuracy (see Table 2 for details).

List of special characters in CHELAR not recognised by CLAWS-7	Adaptations made for POS tagging
£	pounds
á, â, à, ä	same vowel without diacritics
æ	ae
œ	oe
’ (curved apostrophe)	' (straight apostrophe)
½, ¾, etc.	1/2, 3/4, etc.
°	degrees
§	Section
° (for ordinal numbers), e.g. 13°	*th*, e.g. 13th

Table 1: List of characters not recognised by CLAWS-7 and adaptations made for POS tagging

Subperiod	Accuracy
1535-99	95.5%
1600-49	95.7%
1650-99	96.5%
1700-49	97.9%
1750-99	98.5%
1800-49	97.3%
1850-99	96.2%
1900-49	96.7%
1950-99	97.6%

Table 2: Degree of accuracy of CLAWS-7 in the different subperiods of CHELAR in 1,000-word samples for each subperiod

4. Annotation: XML

TEI XML encoding is done using the XML editor Oxygen, which allows for a relatively rapid and error-free editing of TEI texts. Although the annotation possibilities are infinite, we advocate for a modest XML tagging, focusing on the particular structure and content of the law reports. See Figure 1 for an overview of the tags selected.

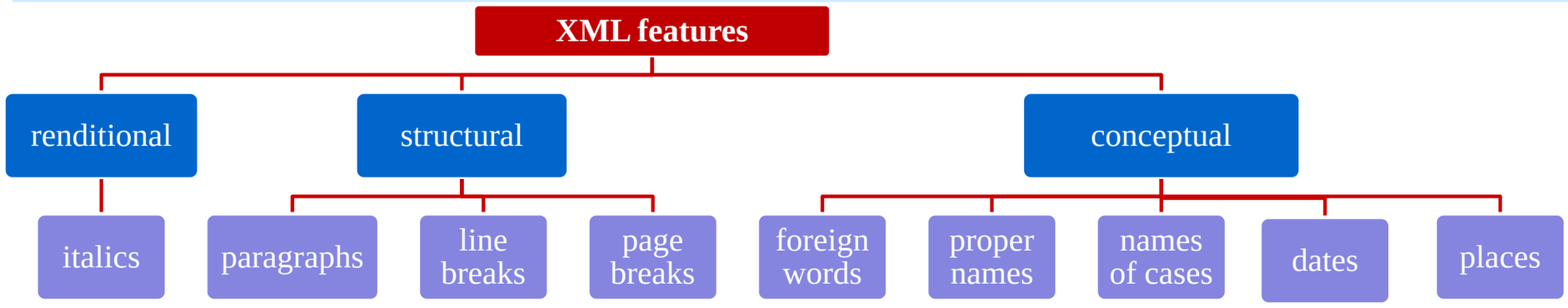

Figure 1: List of tags used in CHELAR

5. Prospects

At present CHELAR is available at request in its plain and POS-annotated versions. The XML-annotated version of the corpus is expected to be finished by the first semester of 2018.

6. Acknowlegments

The corpus annotation tasks have been done jointly with Iván Tamaredo and Daniela Pettersson-Traba, closely supervised by Teresa Fanego and Paula Rodríguez-Puente. For generous financial support, we are grateful to the European Regional Development Fund and the Spanish Ministry of Economy and Competitiveness (grants HUM2007-60706, FFI2014-52188-P, FFI2014-51873-REDT, BES-2012-05755 and BES-2015-071233).