Technical Report on the SPIN Project 2022

Adeline Nazarenko, Francois Levy and Adam Wyner

April 2022

Contents

1 Introduction

This report reflects the SPIN annotation campaign conducted in March 2022 on the English version of the European General Data Protection Regulation (GDPR)[1] using the *Core Legal Annotation Language* (CLAL).

Such an annotation campaign is an annotation work that mobilizes several annotators and outputs a reference annotation (or gold annotation). The annotators have an annotation task to perform and must conform to the annotation instructions. They work in parallel and can make different proposals but adjudication allows them to converge towards consensual solutions. The reference annotation is the result obtained after annotation and adjudication.

1.1 Scientific goal

To navigate between the difficult analysis of statutory rules and needs of current legal practice, [1] proposed a coarse-grained and interpretation-neutral approach to annotating legal texts with semantic information, enabling semantically-based information retrieval capabilities.

This approach aims at enriching the legal texts with coarse-grained annotations describing the elementary provisions so that one legal practitioners can easily retrieve, for instance, all the obligations incumbent on a given actor, the exceptions to a given provision or the procedures to be followed to perform a particular act.

1.2 Experimentation

Based on a first experiment, [1] claimed that the CLAL language is a simple enough language so that 1) people familiar with legal sources can annotate the text without being professional lawyer or logician, and 2) the annotation can be deployed on a large scale.

The Swansea 2022 annotation campaign is a second, broader experiment that aims to 1) verify the initial findings, 2) define an annotation protocol for non-experts, 3) possibly adjust the language if certain elements prove difficult to handle and 4) provide a gold-standard annotation for a significant part of the GDPR.

For this experiment, 6 undergraduate and graduate law students were asked within one month to 1) get familiar with the CLAL language, the provided documentation and the annotation tool, 2) annotate a large portion of the GDPR and participate in the adjudication process and 3) give feedback on the overall process.

They were assisted by 3 staff members, who provided them with explanation and guidance and had the responsibility to deliver the final agreed-on reference annotation in the end.

[1] Adopted in 2016 and entered into force the 25th May 2018.

1.3 Findings

Due to time constraints, the new annotators could not be trained sufficiently for us to have reliable results in terms of annotation speed and quality. On the other hand, the experience was rich in lessons concerning the organisation of annotation campaigns and the training of annotators. The overall experiment confirms the relevance of the annotation language appeared, even if the difficulties encountered by the annotators show that annotation instructions should be improved. Finally, the adjudication work carried out by the staff provides a reference corpus.

1.4 Context

2 Task description

2.1 Corpus and corpus preparation

The corpus chosen for this annotation is the English version of the European General Data Protection Regulation (GDPR) of which the annotators were asked to annotate 64 of the 99 articles

The source text is given to the annotators is an XML document with a simplified structural markup.

In the original document, there are many structural tags as the XML markup gives the division into sections, articles, paragraphs, alineas, lists, some identifiers for some of these elements and the encoding of some special characters. However, that structural markup is simplified to allow annotators to navigate and find their way around the document more easily during the annotation. Only those structural annotations that are actually useful for the annotators, such as the division into paragraphs and articles, are kept. Once the semantic annotation is finished, the initial structural markup is restored. These two operations of simplification and restoration are done automatically and are hidden from the annotators.

It should be noted, in fact, that semantic markup is an independent layer of annotation on top of structural markup. This allows the two layers to be managed independently and even allows semantic annotation to be applied on top of another form of structural markup, with minimal adaptation cost.

To ease the annotation, an automatic pre-annotation has also been done, so that the annotators can focus on the truly semantic tasks for which they have a real added value. In particular, the fragment to annotate have been pre-annotated and associated with their identifiers so that the annotators can concentrate on the semantic type of the fragments and on its semantic roles.

2.2 Annotation language

The Core Legal Annotation Language (CLAL) is composed of

- a set of elementary provision types, among which some deontic types,

```
<PARAG IDENTIFIER="006.003">
  <NO.PARAG>3.</NO.PARAG>
  <ALINEA>
    <P>The basis for the processing referred to in point
    (c) and (e) of paragraph 1 shall be laid down by:</P>
    <LIST TYPE="alpha">
      <ITEM>
        <NP>
          <NO.P>(a)</NO.P>
          <TXT>Union law; or</TXT>
        </NP>
      </ITEM>
      <ITEM>
        <NP>
          <NO.P>(b)</NO.P>
          <TXT>Member State law to which the controller is
          subject.</TXT>
        </NP>
      </ITEM>
    </LIST>
  </ALINEA>
  <ALINEA>The purpose of the processing shall be determined
  in that legal basis or, as regards the processing referred
  to in point (e) of paragraph 1, shall be necessary for the
  performance of a task carried out in the public interest or
  in the exercise of official authority vested in the control—
  ler. That legal basis may contain specific provisions to adapt
  the application of rules of this Regulation, inter alia:
  the general conditions governing the lawfulness of processing
  by the controller; the types of data which are subject to the
  processing; the data subjects concerned; the entities to, and
  the purposes for which, the personal data may be disclosed;
  the purpose limitation; storage periods; and processing
  operations and processing procedures, including measures
  to ensure lawful and fair processing such as those for other
  specific processing situations as provided for in Chapter IX.
  The Union or the Member State law shall meet an objective of
  public interest and be proportionate to the legitimate aim
  pursued.</ALINEA>
</PARAG>
```

Figure 1: GDPR extract with the original structural markup.

```
<PARAG IDENTIFIER="006.003">
  <NO.PARAG>3.</NO.PARAG>
  <P>The basis for the processing referred to in point
    (c) and (e) of paragraph 1 shall be laid down by:</P>
      <LIST TYPE="alpha">
        <ITEM>
          a) Union law; or </ITEM>
        <ITEM>
          b) Member State law to which the controller is
          subject.
        </ITEM>
      </LIST>

  The purpose of the processing shall be determined
  in that legal basis or, as regards the processing referred
  to in point (e) of paragraph 1, shall be necessary for the
  performance of a task carried out in the public interest or
  in the exercise of official authority vested in the control—
  ler. That legal basis may contain specific provisions to adapt
  the application of rules of this Regulation, inter alia:
  the general conditions governing the lawfulness of processing
  by the controller; the types of data which are subject to the
  processing; the data subjects concerned; the entities to, and
  the purposes for which, the personal data may be disclosed;
  the purpose limitation; storage periods; and processing
  operations and processing procedures, including measures
  to ensure lawful and fair processing such as those for other
  specific processing situations as provided for in Chapter IX.
  The Union or the Member State law shall meet an objective of
  public interest and be proportionate to the legitimate aim
  pursued. </PARAG>
```

Figure 2: GDPR extract with the simplified structural markup.

```
<PARAG IDENTIFIER="006.002">
  <NO.PARAG>2.</NO.PARAG>
  <leg:FRAGMENT IDENTIFIER="006.002.001">
    Member States may maintain or introduce more specific
    provisions to adapt the application of the rules of this
    Regulation [...]
  </leg:FRAGMENT>
</PARAG>
```

Figure 3: Example of pre-annotated GDPR extract (Paragraph `006.002`.

Table 1: CLAL vocabulary: the terms (3^{rd} column) are encoded as XML elements (`UPPER CASE`) and XML attributes (`lower case`). Some element types can be subtyped (4^{th} column). "Fragment" is a term that is used to refer to the elementary provisions to annotate. Note that the CLAL XML elements are usually prefixed by `leg:` in XML documents (*e.g.* `<leg:RIGHT id=".."...>...</leg:RIGHT>`.

Fragments	Autonomous fragments	OBLIGATION	
		PROHIBITION	
		PERMISSION	
		RIGHT	
		POWER	executive ruling
		QUALITY	competence qualification responsability
		DEFINITION	
	Subordinate fragments	COMPLEMENT	
		EXCEPTION	
	Sub-fragments	EXCEPT	
Entities	Actors	PERSON	
		LEGAL_ENTITY	
	Concepts	CONCEPT	
Relations	Inter-fragment relations	rel	
		except	
	Roles	obj	
		bearer	
		target	

- a set of entity types, entities be being actors (persons and legal bodies) and concepts,
- a set of relational types, corresponding to inter-provision semantic relations,
- a set of relational types, corresponding to the roles that entities plays within provisions.

Table 1 presents the CLAL vocabulary

The language is described in the semantic guide, which lists all the language components and associate them with definitions and recommendations on how to discriminate one type from another.

The CLAL language is implemented in the form of an XML Schema implemented in the shema description language (`GDPR_SemanticSchema.xsd`). The provision and entity types are defined as XML elements. The relational types

are defined as XML attributes to be associated to XML elements. The schema specifies the grammar of the language: How to place the elements in relation to each other? Which attributes an element requires or admits? The schema has been designed to restrict the freedom of annotators, so that they can focus on the semantic choices and not the syntax of the annotation.

2.3 Annotation task

In theory, the semantic annotation of a legal source should consists in :

1. Segmenting it into elementary provisions,
2. Typing and characterising each of the resulting fragments,
3. Tagging the mentions of the entities that play a key role in the provisions and
4. Recording their unique identifiers into actors and concepts dictionaries

However, the 1^{st} and 4^{th} sub-tasks are not required from annotators. Segmentation (1) is performed at the pre-annotation stage: the provisions are segmented into sentences and each sentence is assimilated to an elementary provision (hereafter "fragment")[2]. Since the marking of entity mentions (3) is time consuming and has a lower priority than the annotation of the roles that entities play in fragments, the annotators are asked to focus on dictionary construction (4) and role filling (2). It is assumed that, in many cases, it should be possible to annotate entity mentions automatically, at a post-processing stage.

The annotators are therefore asked to focus on:

Typing and characterizing the fragments

- Read the textual content of the fragments together with the surrounding text,
- Check if the fragment contains a sub-fragment that expresses an exception and needs to be annotated; if necessary, frame it with the sub-fragment tags (`<leg:EXCEPT>` and `</leg:EXCEPT>`).
- Choose the appropriate type for the fragment ((*e.g.* `POWER`, for Fragment `006.002.001`, see Figure 4) and substitute it the neutral (`FRAGMENT`) type,
- Add the relevant semantic attributes (*e.g.* `bearer`, `target`, `rel`, `except`) to the fragment with the appropriate values ; some attributes are required as the `type` and `bearer` attributes in case of a `POWER` while others are optional,

Recording the entity identifiers into the dictionaries Add a new entry in the actor or concept dictionary for any entity to be referred to but

[2]

```
<PARAG IDENTIFIER="006.002">
  <NO.PARAG>2.</NO.PARAG>
  <leg:POWER IDENTIFIER="006.002.001" bearer="le_MS" type="ruling">
     Member States may maintain or introduce more specific
     provisions to adapt the application of the rules of this
     Regulation [...]
  </leg:POWER>
</PARAG>
```

Figure 4: Example of annotated GDPR extract (Paragraph `006.002`)

```
<?xml version="1.0" encoding="UTF-8"?>
<VOCAB xmlns:leg="http://www.lipn.univ-paris13.fr/rcln/legal">
  <leg:DICTIONARY>
    <leg:LEGAL_ENTITY_ENTRY id="le_MS">
      <LABEL lang="FR" value="Etat membre"/>
      <LABEL lang="EN" value="Member State"/>
    </leg:LEGAL_ENTITY_ENTRY>
  </leg:DICTIONARY>
</VOCAB>
```

Figure 5: Dictionary in a separate file, with a single entry, that of the "Member States" legal entity

which does not yet have an identifier. Actually, the key role that entity plays in a fragment should be recorded by associating the entity identifier as values of one of the fragment attribute. In the example of Figure 4, the `bearer` role is associated with the `le_MS` identifier that corresponds to the "Member States" entry in the actor dictionary of Figure 5.

3 Organisation of the annotation campaign

The SPIN annotation campaign involves several types of actors and is organised in a classical way as an alternation of annotation subtasks and adjudication steps, according to a schedule set up by the annotation managers. As expected, training annotators and allowing them to get used to the resources and tools is a critical step.

This section presents the overall organisation as it was planned initially. See Figure 7 for an overview. The following section explains how it had to be adapted in practice.

3.1 Actors and roles

The SPIN annotation campaign involves three types of actors:

```
<PARAG IDENTIFIER="006.003">
  <NO.PARAG>3.</NO.PARAG>
  <ALINEA>
    <leg:COMPLEMENT IDENTIFIER="006.003.001" type="precision" rel="
        006.001.001"><P>The basis for the processing referred to in point (c) and (e
        ) of paragraph 1 shall be laid down by:</P>
    <LIST TYPE="alpha">
      <ITEM>
        <NP>
          <NO.P>(a)</NO.P>
          <TXT>Union law; or</TXT>
        </NP>
      </ITEM>
      <ITEM>
        <NP>
          <NO.P>(b)</NO.P>
          <TXT>Member State law to which the controller is subject.</TXT>
        </NP>
      </ITEM>
    </LIST></leg:COMPLEMENT>
  </ALINEA>
  <ALINEA>
   <leg:COMPLEMENT IDENTIFIER="006.003.002" type="precision" rel="
        006.001.001">The purpose of the processing shall be determined in that legal
        basis or, as regards the processing referred to in point (e) of paragraph 1, shall
        be necessary for the performance of a task carried out in the public interest or
        in the exercise of official authority vested in the controller.</
        leg:COMPLEMENT> <leg:POWER IDENTIFIER="006.003.003" bearer="
        le_EU le_MS" type="ruling">That legal basis may contain specific provisions to
         adapt the application of rules of this Regulation, inter alia: the general
        conditions governing the lawfulness of processing by the controller; the types of
        data which are subject to the processing; the data subjects concerned; the
        entities to, and the purposes for which, the personal data may be disclosed; the
        purpose limitation; storage periods; and processing operations and processing
        procedures, including measures to ensure lawful and fair processing such as
        those for other specific processing situations as provided for in Chapter IX.</
        leg:POWER> <leg:OBLIGATION IDENTIFIER="006.003.004" bearer="le_EU
        le_MS">The Union or the Member State law shall meet an objective of public
        interest and be proportionate to the legitimate aim pursued.</
        leg:OBLIGATION> </ALINEA>
</PARAG>
```

Figure 6: Annotated GDPR extract, with the structural markup restored

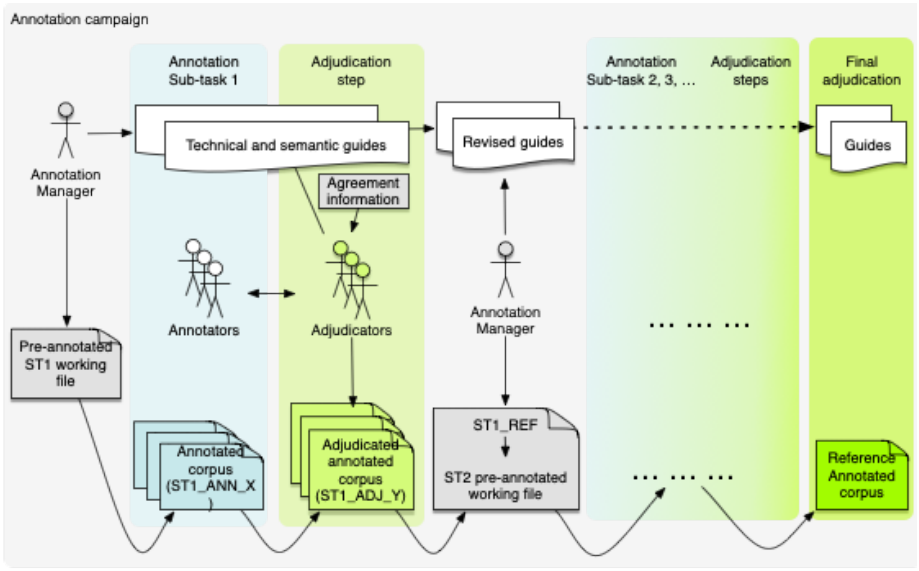

Figure 7: Organisation of the annotation campaign.

- The annotation managers who organize and supervise the campaign,
- The annotators who annotate the source text (undergraduate students, working 15 hours per week during 1 month) and
- The adjudicators who propose a reference annotation from the potentially different proposals made by the different annotators (graduate students working 6 hours a week during 1 month).

3.2 Phasing of the campaign

The annotation campaign is designed as an alternation of annotation subtasks and adjudication steps, according to a schedule set up by the annotation managers.

3.2.1 Coordination of actors

Each cycle consists in a sequence of annotation phase, adjudication step and final revision:

Annotation phase The annotators are expected to work in parallel and independently of each other. The work plan for each subtask is defined by the managers who provide the annotators with an XML version of the GDPR in which a certain amount of fragments were pre-annotated, together with the dictionaries. The annotators have to annotate that version of the

GDPR (Typing and characterizing the fragments and Recording new entity identifiers into the dictionaries, see Section 2.3) and to deliver the files enriched with semantic annotations and new entity entries.

Adjudication step The adjudicators collect the outputs of the annotators and compare them. A collective meeting is organised by the managers so that adjudicators can discuss together and with the annotators. They have to address the annotators' disagreements and more generally to check their annotations. Issues are discussed but the adjudicators have the responsibility to solve them. In the end, each adjudicator outputs an adjudicated file, which is the merge of the annotators' proposals.

Final revision The managers then perform a sort of post-adjudication of the adjudicators proposals. Little or no disagreement is expected at that step, since the adjudicators are expected to have reached a consensus on each controversial issue. However, the managers perform a final check of the result before delivering a unique reference version of the annotation.

The source text is thus annotated incrementally, each cycle providing the annotation for an additional set of articles. The campaign ends either when the scheduled time has elapsed or when everything that should be annotated has been annotated. We then have a partial or complete reference annotation.

3.2.2 Time scheduling

Due to time constraints, each campaign cycle is organized over 1 week as a precise sequence of phases:

Preparation phase: The managers determine the new work to be done (which articles to annotate) and upload the work files for the annotators. Every thing has to be ready on Monday morning.

Subtask sart-up meeting: On Monday morning, all the participants meet for the presentation by the managers of the new subtask workplan, the feedback of the managers on the previous cycle and the resolution of remaining issues if any.

Annotation phase: From Monday to Thursday, the annotators enrich the working files with annotations. They are expected to upload the result of their work on the shared folder on Thursday at noon.

Adjudication meeting: On Friday morning, the adjudicators and the annotators meet to discuss the annotators' annotation choices and their potential disagreements. The presence of the managers is optional.

Adjudication phase: The adjudicators have to upload on Friday evening the result of their work (one adjudication file per adjudicator).

Finalizing phase: The managers resolve the potential remaining adjudicators' disagreement and produce a single reference file before the next campaign cycle begins.

3.3 Resources

Several resources are provided by the managers to support the work of annotators and adjudicators:

Documentation

Annotation semantic guide [3] This is a long document of approximately 50 pages that presents the goal of the annotation, the basics of XML, the CLAL annotation language together with definitions, explanations, examples and recommendations on how to use the annotation tags.

A large part of guide est organized as a dictionary, so that the reader has direct access to the information[4] related to a certain CLAL vocabulary element, *e.g.* `PROHIBITION`.

It is important to note that the guide contains numerous examples of GDPR provisions

Annotation technical guide [5] This is a shorter document (approx. 20 pages) that explains what is expected from the annotators and adjudicators and how to perform the tasks assigned to them in practice. The technical guide starts with the annotation of a short example, it presents the overall organisation of the campaign and the distribution of tasks, the tools, files and folder to work with, as well as a description of the annotation and adjudication elementary tasks.

Oxygen installation instructions [6] A separate guide helps installing Oxygen, the recommended annotation tool (see bellow), and set up its parameters

XML Files

XML Working files The XML working files are the files that are modified by the annotators and the adjudicators. They correspond to the annotated versions of the GDPR and to the dictionaries of actors and concepts. Different versions of these files exist during the campaign: parallel versions output by annotators and adjudicators, incremental versions following the organisation of the annotation campaign. This calls for a strict file naming protocol[7].

XML schemas Various XML schemas are provided to the annotators and adjudicators to enable the validation of their annotations, the XML files output by them. The schemas are organised in various

[3] See `GDPR Annotation Semantic Guide.pdf`

[4] The sections corresponding to vocabulary elements are all structured in the same way, with definition, syntax, annotation recommendations and examples subsections.

[5] See `GDPR Annotation Technical Guide.pdf`

[6] See `Oxygen Installation Instructions.pdf`

[7] Described in the technical guide.

layers to make the semantic annotation as separate as possible to the structural annotation.

The overall schema organisation is rather complex but the annotators and adjudicators simply have to download those files. They do not have to modify them or even to look at them during the campaign.

The most interesting schema is the core one that define the elements and attributes of the CLAL annotation language[8].

Tools

Annotation tool An annotation tool is a tool that helps the annotators to enrich the source XML text with semantic annotations.

It it is possible to read and modify an XML file using a plain text editor, provided that the editor does not add hidden formatting characters. However, the task is nevertheless much easier using a specialized editor which has many helping features.

The tool recommended for the campaign is Oxygen XML Editor (Oxygen for short[9]), which automatically reads the schema associated to an XML file, and uses this schema to 1) suggest at writing time which labels and attributes may be used at the current cursor position, 2) check on the fly if the XML file presently conforms to its schema and warn the user of any detected error, 3) ease modifications: *e.g.* a change in the label of the opening tag automatically triggers the corresponding change in the closing tag. Oxygen also provides facilities such as a summary tree view of the text, a quick search tool and folding/unfolding flags which help navigation in the text.

As indicated above, an installation instructions guide helps the campaign actors to install and tune the editor for their specific needs.

Adjudication tool An adjudication tool is a tool that aligns the annotations provided by different annotators, points out disagreements among annotators and provides edition facilities to help the adjudicator to produce a consensual annotation out of the source ones.

During the SPIN campaign, the adjudicators do not have any specific adjudication tool. They are invited to use Oxygen editor as the annotators, where they can upload the various annotation files, visualize them in parallel and edit a new consensual version out of it. The adjudicators also have a report on the points of disagreement the various annotators outputs.

Shared folder To allow for collaborative and incremental work, a shared folder is set up. It contains all the pieces of information that the actors need to manage the different steps of the annotation campaign. It is organised in different subtask folders, each one with 4 sub-folders

[8] (`GDPR_SemanticSchema.xsd`)
[9] `https://www.oxygenxml.com`

for the source files (the sub-task working files and schemas), the outputs of annotators, the outputs of adjudicators and the final reference files (annotated file and updated dictionaries).

All campaign participants have access to that shared folder to download the resources they need to work and upload their results that can thus be passed on to their successors, therefore from the managers to the annotators, from the annotators to the adjudicators and from those back to the managers in charge of the reference.

Communication tools The entire campaign being conducted remotely, it is essential to set up communication channels between participants. Meetings are organised by videoconference , with recordings and transcripts being made available at the end of the meetings for those who are absent or want to return to certain points of discussion. A chat is also open during the meetings to exchange technical details or precise information (*e.g.* definitions, examples) among participants. In addition to that, remote discussion groups are organised so that participants can address and resolve issues between meetings.

3.4 Assessment and training of the participants

The training of participants is recognized as a critical step in annotation campaigns [2].

Due to time constraints (the campaign lasts only 4 weeks), the training of the participants is done in two ways, during the assessment tutorial and through progressive increase in load.

3.4.1 Assessment

A few days before the start of the campaign, potential participants complete a preparatory and assessment tutorial to check that they are able to participate in the campaign and to familiarize themselves with the annotation work. The adjudicators perform the same exercise as the annotators in order to master the technique and philosophy of annotation.

The assessment lasts 1 week and involves all participants.

- The campaign starts with a kick-off meeting where the managers set the context: they explain the benefit of the semantic annotation of legal source in terms of search for legal practitioners but also the importance of the annotation campaign for the evaluation of the proposed annotation language and its deployment. They also explain the work expected from the annotators and adjudicators, present the working organisation and introduce the resources that are made available to the participants. Note that the participants are expected to be familiar with the GDPR.

- The annotators and adjudicators then have two days to get acquainted with the campaign resource and organisation. In particular, they are

expected to browse the semantic guide to get familiar with the CLAL annotation language and to know where to look for precise information when they need details.

- An intermediate meeting is then organized to answer their questions and set up the assessment tutorial.
- The annotators and adjudicators have again two days to perform the assessment tutorial[10], *i.e.*:
  - Download all the work material,
  - Install the Oxygen annotation tool, following the Oxygen installation instructions.
  - Perform a guided annotation test by following the instructions in the tutorial for annotating a fragment using the Oxygen annotation tool.
  - Perform a semantic annotation test by choosing the appropriate tags for 3 additional fragments.
  - Upload the result on the share folder.
- In the end, the managers check the annotations delivered by the participants. Those who successfully install and use Oxygen for annotating the identified fragments and who are willing to engage in intense annotation work are recruited for the annotation campaign. [11]

3.5 Progressive increase in work load

Even if the participants get acquainted with the technical organisation of the campaign during the assessment tutorial, it takes time to get accounted to CLAL annotation language, not because the language itself is complex but because annotators face of wide variety of provisions types and wordings. Some of the fragments are typical of a given semantic type (*e.g.* `OBLIGATION`) and are quite easy to annotate while others represent borderline cases that deserve some reflection and can lead to disagreements between annotators.

To help annotators and adjudicators to get familiar with the CLAL language and its use for annotation, the campaign is organised in short cycles with intermediate phases of adjudication and discussion (see Section 3.2).

In addition, the managers plan the subtasks in a progressive way, with a small number of fragments to annotate in the first cycle, but an increasing number of fragments in the following cycles, knowing that the annotators have a fixed time to annotate per week (15 hours). This progression also takes into account the number of disagreements which should decrease as the annotators become more trained.

The progression is not planned in advance. The objective is to determine the volume of annotation that trained annotators can perform in a given time without losing annotation quality.

[10] See `GDPR Annotation Assessment tutorial.pdf`

[11]

4 Annotation campaign and results

This section describes the campaign as it actually unfolded and how it deviated from the original plan.

The core part of the campaign lasted for 1 month, in 2022 March. This does not include the preparation of the campaign, organisation setup, the assessment phase, nor the analysis of the results and report writing. The overall experiment actually covers a period of 7 months.

4.1 Participants

Annotation managers The managers have been involved well in advance for the preparation of the campaign (resources and organisation), during the campaign to supervise the annotators' and adjudicators' work and to output reference annotations, as well as after the campaign to analyse the results.

Annotators 3 undergraduate law students worked in parallel 15 hours per week on the annotation sub-tasks, including the participation to the meetings; they also participated in the kick-off meeting and assessment tutorial before hand.

- Mohamed Felaya
- Atif Ismail
- Berivan Sonmez

Adjudicators 3 graduate law students were recruited as adjudicators and worked 6 hours a week during 1 month.

- Adepeju Adesanwo
- Joseph Anim
- Omolola Majolagbe

4.2 Campaign actual organisation

As indicated above, the campaign has been organized in cycles. It consisted in 4 cycles, each one corresponding to a specific annotation subtask, with progressive workload (see Table 2) to accommodate for the training of the participants and an expected increasing annotation quality.

The annotation subtasks were carried out as planned (Section 4.3). However the outputs of annotators proved difficult to exploit and the adjudication steps could not be proceed as planned (Section 4.4). In order to get the expected reference annotation at the end of the campaign, an expert annotation has therefore been conducted in parallel (Section 4.5) and the role of the adjudicators evolved.

Table 2: Annotation subtasks

Subtask	Number of fragments to annotate
1	50
2	87
3	105
4	110

Table 3: Annotators' output annotations

Subtask	Number of annotated fragments	
	Target	Annotators' result (avg)
1	50	35
2	87	82
3	104	103
4	108	100

4.3 Annotators' results

As indicated above, the campaign has been organized in cycles. It consisted in 4 cycles, each one corresponding to a specific annotation subtask, with progressive workload to accommodate for the training of the participants and an annotation quality which was expected to increase over the subtasks.

4.3.1 Workload progress

Table 3 shows the actual campaign progression.

The target column shows how the annotators were expected to progress: after the training phase, they were expected be able to annotate approx. 100 fragments in 10 hours of actual annotation work (roughly, 10 fragment per hour).

However the 3^{rd} column of Table 3 shows that the annotators failed to reach that goal. Several factors interplay to explain this result. Their training took longer than expected. Probably, they did not dedicate enough time to annotation, even if the time indications they declared are partial and probably questionable. Also, after training had progressed (during the third subtask), the next step processed many passages which had to see with relations between different powers and would have needed a specific training.

Table 4 , which gives more detailed indications of the different annotators' work, actually shows very significant differences between annotators and, to a lesser degree, from between the subtasks.

These results should be analysed with caution because some of the annotators met administrative and health issues that hindered their work. It must also be noted that some groups of fragments were harder to annotate than oth-

Table 4: Annotators' time indications and results in terms of number of annotated fragments.

Subtask		Ann. 1		Ann. 2		Ann. 3	
#	Target	Hours	# of ann.	Hours	# of ann.	Hours	# of ann.
1	50	1h30	8	10h30	49	6h	49
2	87	6h	79	14h	83	7h	85
3	104	2h50	104	16h	102	8h	103
4	108	5h20	103	16h30	93	5h30	103

Table 5: Annotators' correctly or erroneously annotated fragments (with the reference file as golden standard. Percentages of correct annotations are computed w.r.t. the target, i.e. the total number of fragments of the task, including unanswered.)

Subtask	Ann. 1		Ann. 2		Ann. 3	
	Correct	error	Correct	error	Correct	error
1	6 (12%)	2	27 (54%)	22	27 (54%)	22
2	58 (66.6%)	21	67 (77%)	16	59 (67.8%)	26
3	62 (59.6%)	42	63 (60.5%)	39	59 (56.7%)	44
4	53 (49%)	50	55 (50.9%)	38	63 (58.3%)	40

ers (see the disagreement figures in table 8). However, Table 4 confirms that the training of the annotators took longer than expected and that meta-legal provisions[12] need a supplementary training.

4.3.2 Error analysis

The report on annotators disagreement is not significant. First because the annotators did not annotate the same fragments or even the same number of fragments. Then because, the annotators who were not enough trained, were often inconsistent in their annotations.

The evolution over time of the ratio of correct annotations is nevertheless interesting (table 5): this ratio reaches between 2/3rd and 3/4th on step 2, but begins to decrease at step 3. A detailed reading of the answers shows that errors increase in the section about codes of conduct, which deals very indirectly with what is permitted or obligatory to the controller.

A confusions table has been produced for autonomous vs dependant classification. It shows that they have a success rate of 84.8% or, otherwise said, a F-measure of 0.77 for dependency recognition.

A confusions table has also been produced for categories (table 7). It shows that exceptions, obligations and rights are better recognised than others. For

[12] We call *meta-legal* provisions which organise the way institutions apply the law

Table 6: annotator's Dependancy confusion matrix

$ref^{proposed}$	autonomous	dependent	total
Autonomous	573	66	639
dependent	75	247	322
total	648	313	961
F-measure	0,890	0,778	

obligations (the bigest group), 40% of false positive are in fact powers where not enough care has been given to actors and content (typically "the supervisory authority must decide if processing is authorised"). Note that one third of powers are so misclassified into obligations, as are one third of procedures and 26% of prohibitions. These figures include the training period, and the number of annotations asked may have increased a tendency to a quick and superficial reading at the expense of a semantic one.

Table 7: annotators' annotations confusion matrix.

proposed / *ref*	ac	aq	ar	ci	cpre	cpro	ct	cv	d	e	no ann	o	pe	po	pr	r	Total
ac																	
aq		3	2						1								6
ar			4									2					6
ci			2	16	3					2	6	2			2		27
cpre		1	1	7	38	8	22	2			9	21	12	2			114
cpro				2	7	42		8		1	5	24	7				91
ct					10	1	18				1						29
cv					3		1	14									18
d									3								3
e					2					40	2		1				43
o	1		5	2	9	13	7			2	25	289				1	329
pe		1		1	4						6		19	3	2		30
po	2		5		4	2		6		1	8	47	30	45			142
pr			3		8	1	1	1	1	4	14	20	7		30		76
r											1		9			38	47
Total	3	5	22	28	88	67	49	31	5	50	77	405	85	50	34	39	961
F-measure		0,55	0,29	0,58	0,38	0,53	0,46	0,57	0,75	0,86		0,79	0,33	0,47	0,55	0,88	

Abbreviations

leg:ATTRIBUTION.competency	ac	leg:DEFINITION	d
leg:ATTRIBUTION.quality	aq	leg:EXCEPTION	e
leg:ATTRIBUTION.responsibility	ar	leg:OBLIGATION	o
leg:COMPLEMENT.impact	ci	leg:PERMISSION	pe
leg:COMPLEMENT.precision	cpre	leg:POWER	po
leg:COMPLEMENT.procedure	cpro	leg:PROHIBITION	pr
leg:COMPLEMENT.text_specification	ct	leg:RIGHT	r
leg:COMPLEMENT.validity	cv	not annotated	no ann

The 77 items in the "no ann" column are not included in the Total column nor in the 961 items in the (Total, Total) cell. The F-measure for one category is computed as recognizing this single category against all others, so it gives a hint of how this category is distinguished from the rest. Cohen's Kappa coefficient on the set of 15 used categories is 0.539 and Krippendorff's alpha coefficient of agreement is 0.551

Table 8: Experts' time indications and number of fragments in agreement with the adjudication.

Subtask	# of ann.	Exp. 1		Exp. 2	
		hours	agr.	hours	agr.
1	50		50(100%)		50 (100%)
2	87		79 (90.8%)		82 (94.3%)
3	104		95 (91.3%)		93 (89.4%)
4	108		88 (81.5%)		88 (81.5%)

Subtask 1 reused an existing reference file, so the agreement rate is not meaningful.

4.4 Adjudicators' results

Due to the poor quality of annotators' output, the adjudicators could not play their role. Some actually tried to revise annotators' annotation but this meant more or less redoing all the annotation, and they did not have the time for that.

Consequently, the adjudicators' expertise has been exploited in a different way, for discussing adjudication issues raised by annotators' and experts' annotations rather than to produce adjudicated annotated files.

In practice, the final adjudication has been made by the experts after discussion with adjudicators.

4.5 Experts' results

As the output of annotators quickly proved not to be exploitable for an adjudication, the two campaign managers in charge of outputting the reference acted as expert annotators, and produced an annotation in parallel with the annotators. They compared their annotations with each other and with those provided by adjudicators when any, discussed the remaining issues during the adjudication meeting and performed the final adjudication in order to output the reference annotation.

This expert annotation was not planned, but it was carried out to compensate for the difficulties encountered by the annotators first and then by the adjudicators, and to provide a reference annotation at the end of the campaign.

4.5.1 Overall results

Table 8 show the time spent by experts and their correctness scores for each annotation subtask.

Given that the expert annotators are already trained, this table shows the expected speed and quality of annotation, given the current state of the language and guidance. It also shows that Subtask 4 seems more difficult to achieve than the previous ones. Actually, independently of the figures, the experts themselves reported more difficult passages to annotate in that subtask.

4.5.2 Detailed qualitative results

A first analysis of expert's agreement measures their success in classifying fragments as dependant or not. Figures in table 9 show that they have a (cumulative) success of 559 over 594 (94.1%). Otherwise seen, the F-measure of recognizing dependant fragments among all fragments is 0.91 (0.956 if seen as recognizing autonomous fragments). Training seems to have clarified the notion, while some marginal uncertainty remains.

A more accurate view can be found in the expert's dependency confusion matrix (table 10). First, the global success rate is 525 over 594 (88.4%) and Krippendorf's alpha coefficient is 0.731, which reflects some maturity and also remaining issues (see 4.6). Among the 69 errors, 26 concern misclassified complements of type precision (38% while these complements are only 12% of the total) - that is the category which has the worse recognition F-measure.

Table 9: Experts' dependency confusion matrix

$ref^{proposed}$	autonomous	dependent	total
autonomous	380	12	392
dependent	23	179	202
total	403	191	594

Table 10: Experts' annotations confusion matrix.

proposed / *ref*	ac	aq	ar	ci	cpre	cpro	ct	cv	d	e	o	pe	po	pr	r	total
ac																
aq		4														4
ar			2													2
ci				15		1										16
cpre			1	4	46	9					9	1	2			72
cpro					2	48					4	1	3			58
ct					3		16				1					20
cv								12								12
d									2							2
e										23				1		24
o					1	4					199		5		1	210
pe												12				12
po	1				1	1				3	4		77	1	2	90
pr					1	1								39	1	42
r															30	30
total	1	4	3	19	54	64	16	12	2	26	217	14	87	41	34	594
F-measure		1	.8	.86	.73	.79	.89	1	1	.92	.93	.92	.87	.94	.94	

Abreviations

leg:ATTRIBUTION.competency	ac	leg:DEFINITION	d
leg:ATTRIBUTION.quality	aq	leg:EXCEPTION	e
leg:ATTRIBUTION.responsability	ar	leg:OBLIGATION	o
leg:COMPLEMENT.impact	ci	leg:PERMISSION	pe
leg:COMPLEMENT.precision	cpre	leg:POWER	po
leg:COMPLEMENT.procedure	cpro	leg:PROHIBITION	pr
leg:COMPLEMENT.text_specification	ct	leg:RIGHT	r
leg:COMPLEMENT.validity	cv		

The F-measure for one category is computed as recognizing this single category against all others, so it gives a hint of how this category is distinguished from the rest. Cohen's Kappa on the set of 15 categories is 0.857 and Krippendorff's alpha coefficient of agreement is 0.731

4.6 Issues raised during adjudication meetings

Discussions during the adjudication meetings focused on various issues that were raised either by the adjudicators themselves based on the annotators' outputs or by the managers' feedback, including the experts' own disagreements.

4.6.1 Issues

These issues highlight the topics that are not well understood in the semantic guidelines by the participants or those that raise multiple interpretation.

- Technical xml constraints:
  - Lists are made of a head sentence and several items. From the semantic point of view, they are an unit – one single fragment marked by a semantic tag. From the layering point ov view, the head sentence is sometimes part of a whole unit (tagged ¡P¿ for paragraph) which precedes the items. In this case, layering and semantic units overlap, which is not accepted by XML. We had to adopt a special coding for that, and the coding is rather tricky for lawyers.
  - There are– rare – cases where the list appears in an EXCEPT subfragment (which is by definition included inside a fragment), since both the enclosing fragment and the EXCEPT overlap. The trick is still worse in this case, and difficult to apply by lawyers

  Interns were not asked to enter such technicalities. They have been applied *a posteriori* by the staff, thanks to schema conformance checking which immediately indicates their location.

- Which fragment relationships should be annotated ? Relationships result from various clues which are more or less explicit. Annotating the most obvious of them would increase the annotator's effort and overload the user with information he already has. It has been necessary to give criteria of which relationships are worthy to be annotated.

- How to choose the deontic value of a fragment ? In particular, what is the difference between `PERMISSION` and `RIGHT` or between `PERMISSION` and `POWER` ? How does a negation and `PERMISSION` , `OBLIGATION`, `PROHIBITION` interact (according to which one is in the scope of the other).

- How to determine whether a fragment is autonomous or subordinate? The question reveals a difficulty to analyse if a fragment has a semantic value of its own or if it essentially complements the semantic value of another one. We tried to clarify and give criteria of what 'complementing the semantic value' means.

- Subordinate fragments include `EXCEPTION` and `COMPLEMENT`. Participants have had difficulties to choose between the five subtypes of `COMPLEMENT`, which try to organize how legal provisions are related.

- The difference between a responsibility and an obligation is difficult, since a responsible one is committed to realize what he is responsible for. More, liability may follow in both cases[13]

- It may happen that mandatory roles point to an entity which is difficult to represent as an actor. Three abstract values have been introduced to cope with different cases: `UNDEFINED`, `UNKNOWN` and `ALL`. The choice between these values requires a precise analysis of the reason of vagueness.

These issues showed the necessity to improve the annotators' training, to revise the semantic guide and/or to provide them with an additional FAQ[14] document.

In the midst of the campaign that was launched, the choice was made to discuss these issues during the adjudication meetings and to begin drafting a FAQ document. Eventually, most of these elements will be integrated into the annotation semantic guide.

4.6.2 Relevant categories to look at

The most controversial categories or category ambiguities appear to be the following:

`complement` **elements** have in general a weak rate of recognition by annotators. The worse of them is the one of type `precision`: one third of them are marked as another complement (mainly a `text specification`), and another third as an autonomous fragment (mainly as an obligation, then as a pemission). Note that experts have far better scores on these categories, so we incline to improve first explanations and examples.

`attibution` **elements of type** `responsibility` are widely over-estimated. There is in several legal traditions, particularly in the English one, a distinction between responsibility, accountability and liability, which is not so clear cut in French. As CLAL makes semantic annotations, types of `ATTIBUTION` need to be improved according this line.

Somme sentences have been difficult to annotate, due to a difficulty to estimate their legal effect. This is particularly the case of general framing affirmations (e.g. "This Regulation protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data."). They are presently marked as `COMPLEMENT` of type `precision` related to the whole text or a chapter, a section. This solution has to be evaluated, in particular with respect to future annotation of recitals, which have the same kind of framing value.

[13] Liability and Responsibility have the same french translation, which did not helped clarifying the point

[14] Frequently asked questions.

More specific questions

- Add a category to handle penalties ; mind that in some texts (e.g; the Scottish smoking Act), the provisions make reference to the penalties, whereas the penalty statements of the GDPR refers to the provisions. A `REPARATION` element has been now added, with `repar` relations to the provisions and optional `penalty` relation that refers to the `PENALTY` statements. There may also be `PENALTY OR FEE` subfragments within `REPARATION` statements.
- Add a `scope` subtype in `COMPLEMENT` to handle provisions which limit the scope of a related passage according to various criteria (territorial, kind of actor, criteria on facts, etc.). Note that the `validity` subtype already copes with temporal scope.
- Fragments creating a legal entity ("The European Data Protection Board (the Board) is hereby established as a body of the Union") are presently marked as `ATTRIBUTION` of type `quality`. Should they justify of a specific type or even category ?
- Some vocabulary questions are pending : first the name `legal entity` for actors in charge of enforcing the law is not a good choice, and should be replaced with `legal body` ; second, should we keep the UNLISTED entity in the langauge or consider it as a caveat for annotators?

5 Discussion

5.1 Error analysis

Notwithstanding the technicalities of XML, mastering it at the level needed to annotate has not been a problem. The main source of errors is a weak assimilation of language, leaving too much importance to superficial lexical cues at the expense of semantics.

Two reasons mainly explain the point. At first, the training time has been underestimated: it had to remain compatible with the total duration of the internship (4 weeks), but that leaves a very short period for training. At the opposite, the guide was very detailled, involving for each category a definition, xml syntax, detailled examples from the GDPR and comments. So interns had to make a significant effort in order to acquire a global view of the annotation language. To that must be added that the translation of the vocabulary from french to english produced in some cases unclear definitions, which did not help.

On the adjudicators' side, the training difficulty has been augmented by a very short working time available per week (5 hours), which was too short to cope with annotations of a poor quality. More, alternating annotation and adjudication phases in the same week caused hard time constraints for interns, whatever their role.

5.2 Interns' feedback

5.3 Recommendations

5.3.1 Campaign organisation

The fist recommendation is to leave more training time. Few hours are needed to get familiar with the XML editing environment. Acquiring a sufficient familiarity with the annotation language as a whole to be able to return efficiently to the documentation when a difficulty is met needs more time, because a superficial reading of some pages is not enough.

The work will also be easier if the campaign is spread, allowing either to alternate annotation and adjudication phases, or to stack them with less tight time constraints. Annotators need some time between adjudication and a new annotation phase, so that adjudication can play its role to provide a feedback for the training of annotators. For trained annotators, adjudication could also be performed as a moderation step.

5.3.2 Documentation

A short and simple version of the annotation guide with simplified examples has to be provided, for allowing untrained annotators to easily get a global view of the language. A FAQ has been built to help then clarifying the most frequent difficulties ; it can be continued as needed. Some definitions will also be revised.

5.3.3 Language

Some modifications of the language are needed, at first the responsibility / liability distinction. The creation of new legal bodies must also be integrated, either as a new category, or an extension of an existing one. What kind of texts can be the object of a text specification is also to be more precisely specified.

6 Conclusion

The internship has been the first experience in other people than the creators of the language annotating a text with CLAL . Annotators were undergraduate law students. It proved that technicalities of the annotation are not an obstacle. On the other hand, work remains to be done to improve the training of annotators, including the time schedule of the work and the guidance given. Some adjustments of the language are also in project.

Acknowledgment

We thank Swansea Law University which funded the internships.

7 additional notes

31 Oct to change. See notes by Adeline.

Different versions of the paper and wrt the XSD. Homogenise across versions wrt the following:

- fragment → statement (query language and working xsd)
- quality → attribution (done)
- responsibility type to → liability. The latter seems rather more general and not specifically associated to violation and reparation. Issues with the English and French meanings of terms. Issues with the term. A person is ascribed liability or have liability; the person is held responsible or accountable for the liability (violation and reparation) should something go wrong. Seems responsibility is not needed.
- legal entity → legal body

Prior notes.

A confusions table has been produced for autonomous vs dependant classification by all interns. It shows that they have a success rate of 85.2% or, otherwise said, a F-measure of 0.78 for dependency recognition.

Table 11: intern's Dependancy confusion matrix

$ref^{proposed}$	autonomous	dependent	total
autonomous	995	127	1122
dependent	123	448	571
total	1118	575	1693

A category confusions table has also been produced (table 12). It shows that recognition of deontic values suffer some uncertainty: obligations are over-estimated (696 while they are 575) as permissions (139 while they are 49), but prohibitions (62 while they are 138) and powers (97 while they are 258) are strongly under-estimated. For instance, more than one third of powers are miss-classified as obligations. The number of annotations asked may have increased a tendency to a quick and superficial reading at the expense of a semantic one.

Table 12: interns' annotations confusion matrix.

ref \ *proposed*	ac	aq	ar	ci	cpre	cpro	ct	cv	d	e	o	pe	po	pr	r	total
ac																
aq		4	3						2							9
ar			7								3					10
ci			5	29	8					4	4			3		53
cpre		1	4	13	75	16	29	5			30	23	3			199
cpro				6	15	73		13		2	39	10				158
ct				2	21	2	28									53
cv					4		3	27								34
d									5							5
e					4					69		1				74
o	2		9	9	17	23	10			3	499	1			2	575
pe		1		2	6							31	6	3		49
po	2		8	1	6	6		11		2	90	47	85			258
pr			4	1	16	1	1	2	2	9	31	11	3	56	1	138
r				1								15			62	78
total	4	6	40	64	172	121	71	58	9	89	696	139	97	62	65	1693
F-measure		0,533	0,280	0,496	0,404	0,523	0,452	0,587		0,847	0,785	0,330	0,479	0,560	0,867	

Abbreviations

leg:ATTRIBUTION.competency	ac	leg:DEFINITION	d
leg:ATTRIBUTION.quality	aq	leg:EXCEPTION	e
leg:ATTRIBUTION.responsibility	ar	leg:OBLIGATION	o
leg:COMPLEMENT.impact	ci	leg:PERMISSION	pe
leg:COMPLEMENT.precision	cpre	leg:POWER	po
leg:COMPLEMENT.procedure	cpro	leg:PROHIBITION	pr
leg:COMPLEMENT.text_specification	ct	leg:RIGHT	r
leg:COMPLEMENT.validity	cv		

The F-measure for one category is computed as recognizing this single category against all others, so it gives a hint of how this category is distinguished from the rest.

References

[1] Nazarenko A, Lévy F, Wyner A. A Pragmatic Approach to Semantic Annotation for Search of Legal Texts - An Experiment on GDPR. In: JURIX. vol. 346 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2021. p. 23–32.

[2] Fort K. Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects. Wiley-ISTE; 2016. Available from: https://hal.inria.fr/hal-01324322.