



École Polytechnique Fédérale de Lausanne

De-identification of Free Text in Swiss-French Electronic Health Records:  
From Internal Annotation Crowdsourcing to Personal Data Detection

by Valentin Oliver Loftsson  
Master's Thesis

Approved by the Examining Committee:

Olivier Verscheure  
Thesis Advisor

Jean-Louis Raisaro  
Thesis Supervisor

He Xu  
Thesis Supervisor

March 18, 2022

Knowledge is as wings to man's life,  
and a ladder for his ascent. . .  
—*Bahá'u'lláh*

Dedicated to my parents

# Acknowledgments

I want to express my immense gratitude to my supervisors at CHUV, He Xu and Jean-Louis Raisaro, for their constant support and guidance. Thanks also to Olivier Verscheure for being my supervising professor at EPFL and providing helpful feedback during my project. Thank you to my colleagues at CHUV for supporting my project by taking part in the annotation contest. Special thanks to Bayrem Kaabachi for helping me with the pre-campaign and for translating the annotation guidelines into French. Thanks to all the people who developed software and tools that I used in my project, especially the authors of pygamma-agreement and Prodigy. I have learned so much at EPFL and would like to give special thanks to all my teachers there for sharing their knowledge and enthusiasm. Thank you to all family members and friends who have supported me. Mom and dad, thank you for always supporting me in my education. Thank you, Atlas, my son, for always bringing joy and laughter to my heart. Finally, I could not have done this without the loving support, encouragement, and sacrifices of my wife, Karen—you are a true heroine for going on this journey with me and I am so lucky to have had you by my side this entire time.

*Copenhagen, March 18, 2022*

Valentin Oliver Loftsson

# Abstract

## Background

Clinical texts are a valuable resource for clinical applications and research. However, we need to remove the patient’s private information from a text to be able to share it. This process is called de-identification. Available de-identification solutions for French clinical texts are not many. Most of them are rule-based, tailored to specific corpora, and only target a subset of types of protected health information of interest at the Lausanne University Hospital (CHUV). Moreover, we cannot use external annotated corpora for incompatibility and privacy reasons.

## Objective

This work aims to build an annotated corpus and develop automated methods to detect personal data in clinical texts at CHUV. We also investigate the impact of rule-based pre-annotations on human annotation and present a novel stratified quality-prioritized sampling method.

## Methods

We defined categories of protected health information (PHI) based on the regulations in Switzerland and at CHUV. We created a rule-based system for automated PHI detection as a baseline for model comparison. Then, we designed an annotation contest to collect annotations. During the contest, we investigated the benefit of pre-annotating annotation tasks with the rule-based system for annotation quality and throughput. The quality of annotation was measured using the gamma  $\gamma$  method, which simultaneously computes inter-annotator agreement and determines the best alignment of multi-annotated samples. We built an annotated corpus from the annotations collected in the contest and then trained a bi-LSTM model for detecting PHI. Three variants of this model—including a hybrid model with a rule-based component—were evaluated and compared with the baseline rule-based system. We designed a custom sampling method to split the corpus in a stratified way while prioritizing high-quality samples for the test and validation subsets.

## Results

We defined 25 PHI categories organized into several supercategories. We conducted an annotation contest with 15 annotators divided into four teams. From this contest, we collected 5,454 samples in total, the average throughput was 7.9 tasks/min., the average annotator quality index was 91.8/100.0, and the mean gamma  $\gamma$  inter-annotator agreement was 0.86/1.00.

The rule-based system achieved a 0.85 macro averaged F1-score when evaluated on the entire reference corpus thus produced. Using our rule-based system for pre-annotation significantly improved annotation throughput and quality. Comparison of models for PHI detection revealed that the rule-based system outperformed other models, achieving macro averaged 0.94 precision and 0.88 recall on the test set. The hybrid model outperformed the pure bi-LSTM models.

## Conclusion

This work provides a foundation for the PHI detection system for clinical texts at CHUV. The annotated corpus can be used later for model fine-tuning and risk evaluation for text de-identification. We observed a high increase in the performance of annotators by providing good quality pre-annotations. The annotation strategy we present provides insights on how to organize annotation crowdsourcing with limited resources.

# Contents

<b>Acknowledgments</b>	<b>1</b>
<b>Abstract</b>	<b>2</b>
<b>1 Introduction</b>	<b>5</b>
<b>2 Preliminaries</b>	<b>9</b>
2.1 Electronic Health Records at CHUV . . . . .	9
2.2 Data Pre-processing . . . . .	9
2.3 Defining PHI Categories . . . . .	10
<b>3 Internal Annotation Crowdsourcing</b>	<b>15</b>
3.1 Contest Preparation . . . . .	15
3.1.1 Recruitment . . . . .	15
3.1.2 Annotation Guidelines . . . . .	15
3.1.3 Prodigy Annotation Tool . . . . .	16
3.1.4 Annotator Training . . . . .	17
3.2 Contest Management: Agile Annotation . . . . .	17
3.2.1 Maintaining Motivation . . . . .	19
3.2.2 Annotation Analysis and Annotator Feedback . . . . .	20
3.2.3 Scoreboard . . . . .	21
3.3 Computing Agreement and Building an Annotated Corpus . . . . .	21
<b>4 Rule-Based PHI Detection</b>	<b>28</b>
4.1 Building a Rule-based System . . . . .	29
4.2 Evaluating the Impact of Rule-Based Pre-Annotation on Human Annotation . .	34
<b>5 Advanced PHI Detection</b>	<b>39</b>
5.1 Stratified and Quality-prioritized Multi-label Sampling . . . . .	39
5.2 Bi-LSTM Model for PHI Detection . . . . .	44
5.3 Model Evaluation and Comparison . . . . .	45
<b>6 Conclusion and Future Work</b>	<b>48</b>
<b>Bibliography</b>	<b>50</b>
<b>Glossary</b>	<b>53</b>
<b>A Annotation Guidelines</b>	<b>55</b>

# Chapter 1

## Introduction

Electronic Health Records (EHRs) are electronically stored patient health data. They consist of *structured* records, such as patient meta data stored in a database, and *unstructured* data, such as clinical notes written by a physician, biomedical images, and audio recordings. These data are usually confidential by law and cannot be shared externally without authorization. In this work we consider the automatic de-identification of unstructured textual EHRs and will henceforth refer to these records as *clinical texts*.

De-identification of clinical texts is the process of *detecting* and *replacing* protected health information (PHI) until the risk of revealing the patient's identity is estimated to be acceptably low. The concept is also sometimes referred to as *anonymization*.

## Previous Work

Annotated corpus development is the first step towards de-identification. General corpus development has been studied by many researchers. Voorman and Gut [19] proposed *agile corpus creation*, an iterative approach to the annotation process. They present a reorganization of traditional linear and separate phases of design, annotation, and analysis, in recognition of the potential sources of errors during corpus creation. Fort [2] wrote a comprehensive book about collaborative annotation. She lays great emphasis on the sound preparation of annotation campaigns, including the writing of annotation guidelines and the training of annotators. She also stresses the benefit of holding a pre-campaign to evaluate and finalize the list of target categories and to refine the annotation guidelines. Importance is likewise given to regular feedback cycles during annotation campaigns to build consensus and improve consistency of annotations.

Research into reference corpus development for French clinical texts is generally lacking and has until now mostly focused on public data sets. Therefore, in our work we decided to focus heavily on this subject. An interesting study by Grouin and Névél [5] examines this subject on private French clinical records. They pre-annotated the texts before introducing them to human annotators. They compared two systems for pre-annotation, one is rule-based

and the other uses conditional random fields. They found that pre-annotations obtained with the rule-based system were more accurate but required more revision time by the annotators. They also found that very small amount of their own data was needed to train a statistical model that could outperform systems trained on corpora from a different medical specialty and from a different hospital. Their conclusion highlights the ineffectiveness of applying external models in a new context.

When an annotated corpus has been built, systems for detecting PHI can be built and evaluated. Much more work has been done on English corpora compared to French corpora. Advanced ensemble-based methods have been proposed by Kim et al. [8], [9] that have achieved state-of-the-art performance on the public 2006 and 2014 i2b2 clinical text corpora. They proposed three different ensemble-based approaches that combine a diverse set of deep learning, shallow learning, and rule-based models, and showed that they can successfully integrate predictions from individual models and offer better generalization across different corpora.

Studies on this subject for French clinical texts have not advanced at the same rate. Recently, Bourdois et al. [1] compared rule-based, transformer-based, and hybrid systems on a set of manually annotated documents from the University Hospital of Bordeaux in France. They found that hybrid systems showed the best performance since they combine expertise of different learners. They used a rule-based system to create an annotated corpus for model training, and manually annotated 3,000 clinical documents for evaluating the model.

Studies have been conducted in Swiss hospitals, too. Gaudet-Blavignac et al. [4] and Foufi et al. [3] developed rule-based systems for the Geneva University Hospitals. They observed good results while noting the natural disadvantages of a rule-based approach. Interestingly, another much older study conducted at the same university in 2000 by Ruch et al. used a semantic lexicon for the same task.

## Motivation

Clinical texts contain rich information that cannot be found in other EHRs. Therefore, they are important for clinical applications and medical research, such as quality measurement and improvement, public health, and epidemiology. To enable trustworthy privacy-preserving secondary usage of this information, the clinical texts need to be de-identified. Manual de-identification is time-consuming and expensive so there is a need to automate the work with more sustainable methods. Existing solutions for de-identification of French clinical texts are not many and most of them are rule-based which limits their portability and applicability to other contexts. Moreover, these solutions tend to be tailored to the data or the context they are concerned with. Therefore, a solution needs to be developed for CHUV clinical texts that also has the potential to be used in other hospitals and institutes.



An annotated corpus is a prerequisite for the development of an automated de-identification system but no annotated corpus existed at the beginning of this project. External French corpora cannot be used because the categories of personal data in CHUV clinical texts are different. Annotating CHUV clinical texts is, therefore, of utmost importance. However, annotations cannot be publicly crowdsourced because the data is private. Consequently, we needed to learn how to mobilize employees within the organization to participate in annotation efforts.

## Objectives

The central objective of this work is to develop a solid foundation for an automated de-identification system for clinical texts at CHUV. The entire timeline of the de-identification project can be seen in Figure 1.1, including tasks that will be covered in later work. In this work, we introduce an efficient infrastructure for annotated corpus development. The infrastructure can be divided into two parts: internal crowdsourcing of annotations through annotation contests, and quality evaluation and alignment of multi-annotated inputs. We also study the impact of rule-based pre-annotations on human annotation throughput and quality. Furthermore, we present a novel stratified quality-prioritized sampling method that is used to split the corpus into subsets for the development of machine learning models. Finally, we develop and evaluate automated rule-based and state-of-the-art machine learning methods to detect personal data in clinical texts.

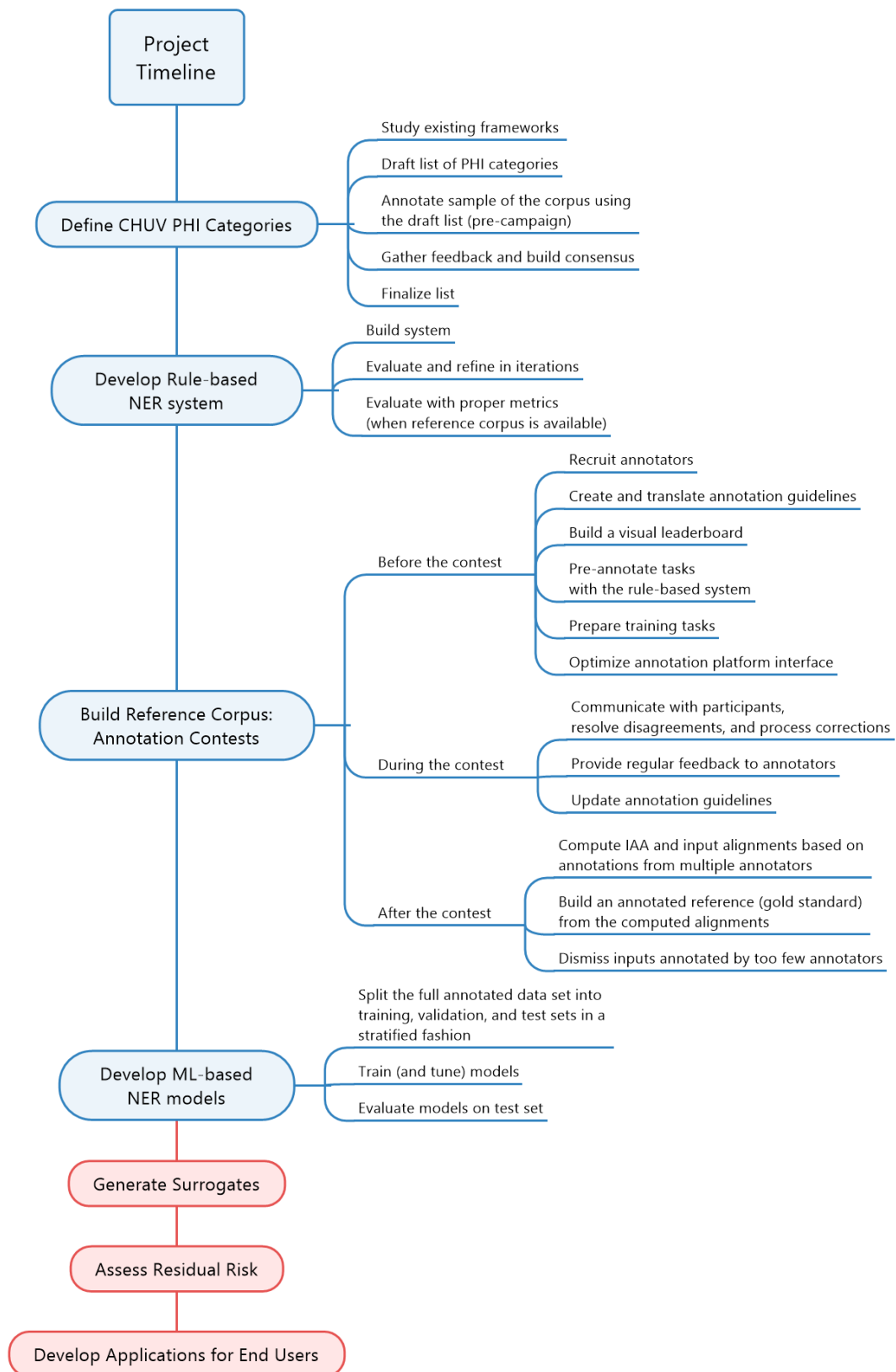


Figure 1.1: The CHUV de-identification project timeline. The current project scope includes the blue blocks, broken down into more detailed tasks. The red blocks represent elements that will be part of future work.

# Chapter 2

## Preliminaries

### 2.1 Electronic Health Records at CHUV

The EHRs are stored in a secure data warehouse for research purposes (DWH-RC) at CHUV. Data security and confidentiality are taken seriously and only relevant employees have direct access to the data warehouse. Clinical texts are stored in XML format, enabling the association of various structured information to the record, such as ID, first and last names, home address, blood type, etc. The text content is rendered in HTML format, embedded as CDATA (XML Character Data) inside the XML file.

There are many different types of clinical texts at CHUV, for example, consultation letters, discharge letters, and laboratory tests. In this work, we used the same number of samples for all types of texts.

### 2.2 Data Pre-processing

We took the following pre-processing steps for each text document:

1. HTML inside each CDATA node is extracted using XPath.
2. HTML is transformed to plain text with the following steps:
  - (a) The HTML is parsed into a tree structure with a tool to facilitate navigation, modification, and data extraction.
  - (b) The root node is enclosed in a `<div>` element since sometimes the CDATA only contains plain text and no HTML tags.
  - (c) New lines are prepended and appended to all instances of certain HTML elements (`<br>`, `<li>`, `<p>`, `<tr>`) that are typically displayed in that way due to their semantics. The new lines are required to correctly detect paragraph marks in the text.
  - (d) Text content inside table data cell elements (`<td>`) is padded with whitespace since often the content is unspaced and collides with nearby cells.

3. The resulting text is then cleaned with the following steps:
  - (a) The text is converted to the UTF-8 character set and Unicode normalized (NFKC).
  - (b) Unnecessary or excess characters and whitespace are removed.
  - (c) Full stop characters surrounded with digits or capital letters are replaced with a non-punctuation character, to ensure correct tokenization and avoid confusion with a proper full stop. We chose # for this purpose. For example, the date 18.03.2022 is transformed into 18#03#2022.
  - (d) Words connected with a dash are split into separate tokens when the first part ends with a digit and the second part starts with a capital letter. This is done to ensure the two parts are treated as separate tokens by the tokenizer. For example, 1015-Lausanne is transformed into 1015 Lausanne.
4. The processed text is split into several chunks at each new line and then fed into the annotation tool for manual annotation. We decided to break the text into chunks for two reasons. First, the annotators are likely to spend a long time reading and annotating a long piece of text, thus impairing their performance. Second, since a clinical text contains a lot of personally identifiable information about the patient, breaking the document into chunks and presenting them in random order can help reduce the privacy risk.

## 2.3 Defining PHI Categories

Some countries and associations have created standards and frameworks for health data privacy protection. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) [14], a U.S. federal statute, provides a legal framework to which studies involving health data privacy commonly refer. HIPAA defines a list of 18 categories of *Protected Health Information* (PHI) and proposes an approach to de-identification that implies removing these data from clinical texts. Legally, HIPAA only applies in the U.S., so it mainly serves as a primary example for foreign projects in health data privacy. Therefore, the HIPAA list of categories was not used directly in this project but mainly as an initial reference.

In Switzerland, the Federal Act on Data Protection of 1992 (FADP) [16] regulates data privacy and the rights of persons when their data is processed. Just like the European Union’s General Data Protection Regulation (GDPR), the FADP takes a more risk-driven approach towards data de-identification than HIPAA. It emphasizes that the data subject should no longer be identifiable in a de-identified text, and the risk of re-identification should be assessed by an expert using statistical approaches. The HIPAA protocol provides a “safe harbor” method for de-identification of medical documents that implies the removal of the 18 pre-defined categories. However, it is possible to do this in compliance with the law without any actual knowledge residual information can identify an individual. In that sense, the HIPAA approach falls short

of the European and Swiss requirements. The GDPR/FADP approach is more realistic and flexible in that it does not impose a fixed list of PHI and provides general methodological guidance instead. Indeed, the FADP defines *personal data* as any information relating to a person that directly identifies the person and information that allows *indirect* identification by additional information.

Based on the HIPAA and GDPR/FADP proposals, we determined a list of direct and indirect identifiers pertinent to the clinical texts at CHUV. Table 2.1 displays the final list of categories organized into several supercategories. Note that the clinical notes contain CHUV-specific identifiers, namely all the ID and CHUV categories listed in the table. Below is a description of each of the supercategories.

- AUTRES  
Fallback category
- CHUV  
CHUV-specific entities
- CONTACT  
Contact information
- DÉMOGRAPHIE  
Demography-related
- EMPLACEMENT  
Location-related
- ID  
Personal identifiers
- NOM  
Names of people
- ORGANISATION  
Name or abbreviation of any organization
- PERSONNES  
Words referring to people and relationships
- TEMPORAL  
Time-related information

The list presented in Table 2.1 was not conceived simply on the first try. The first draft was shorter and contained fewer categories with no supercategories. It was first written in English and later translated into French.

As discussed by Fort [2], there should be a consensus about the categories before they are applied. Therefore, we tested the draft in a pre-campaign involving a couple of volunteers. The volunteers were given simple guidelines and asked to annotate samples from the corpus with the provided list of categories. This process revealed ambiguity and disagreements among the annotators. Thus we needed to revise the names and organization of existing categories and introduce new ones.

Figure 2.1 shows category frequencies in the annotated corpus, revealing a considerable imbalance in the prevalence of categories. Some categories were later archived when we discovered they practically are non-existent in the texts. For example, annotators found no instances of account numbers, IBANs, or health insurance numbers.

We followed the approach of Stubbs and Uzuner [17] and produced granular PHI categories to maximize the utility of the data for research on any subsets of the categories. For instance, we have two **NOM** categories, **PATIENT\_E** is for the names of patients, and **PERSONNEL\_MÉDICAL** is for the names of medical staff. Fine-grained categories can also help the surrogate generation process, by either treating specific categories uniformly when appropriate (for instance, **ID** categories mainly consist of digits) or ensuring appropriate surrogates are generated for each subcategory (for instance, **EMPLACEMENT** categories are distinct from each other, and it helps to distinguish canton codes from other location elements).

We also provide the **AUTRES** category as a fallback for any information that cannot be assigned to any PHI and that may be considered personal or could be used, either directly or indirectly, to identify the patient. Phrases annotated with this category may later be studied to determine whether they warrant a new category.

Table 2.1: Categories of PHI in EHRs at CHUV.

Supercategory	Category
AUTRES	AUTRES
CHUV	BÂTIMENT_CHAMBRE_OU_LIT STRUCTURE RÉFÉRENCE
CONTACT	EMAIL FAX TÉLÉPHONE URL
DÉMOGRAPHIE	ÂGE ÉTAT_CIVIL NATIONALITÉ PROFESSION
EMPLACEMENT	CODE_CANTON CODE_POSTAL EMPLACEMENT_GÉOGRAPHIQUE NUMÉRO_HABITATION PAYS RUE
ID	IPP NUMÉRO_BON NUMÉRO_SÉJOUR
NOM	PATIENT_E PERSONNEL_MÉDICAL
ORGANISATION	ORGANISATION
PERSONNES	LIEN_DE_PARENTÉ
TEMPORAL	DATE TEMPS

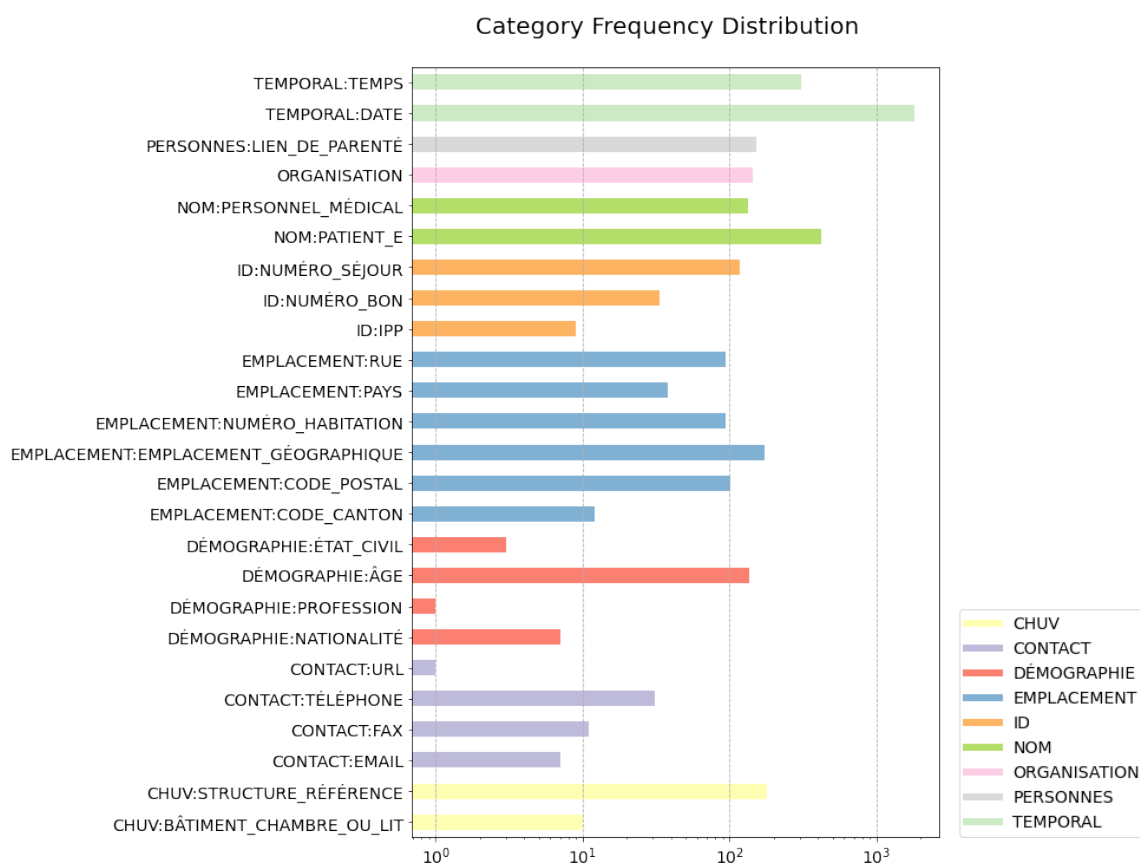


Figure 2.1: Category frequencies in the reference corpus built from the first annotation contest.



## Chapter 3

# Internal Annotation Crowdsourcing

This chapter presents our methods to build an annotated corpus by crowdsourcing annotations internally within the CHUV through annotation contests. We will regularly cite Karën Fort’s book, *Collaborative Annotation for Reliable Natural Language Processing* [2], which provides some well-grounded principles and strategies for holding annotation campaigns. We will also present how multi-annotated inputs were converted to a reference corpus using the gamma method, introduced by Mathet et al. [11]

### 3.1 Contest Preparation

As discussed by Fort, the preparation of annotation campaigns is often rushed or overlooked, and guidelines for annotators might be poor or even entirely missing. Improper preparation often results in lower-quality annotations or unmotivated annotators. Consequently, we considered Fort’s suggestions for the design of our contest. Due to data privacy and lack of resources, we had to carry out internal crowdsourcing, posing some serious challenges while at the same time providing an invaluable opportunity to learn about how to mobilize a group of individuals within an organization to contribute to the building of an annotated text corpus.

#### 3.1.1 Recruitment

We recruited annotators from within the Data Science Group of the IT Department. The nature of the task does not require any medical expertise or other skills besides basic French literacy, so we could invite everyone from the Data Science Group to participate. We held one contest over three weeks during this project, with 15 registered participants.

#### 3.1.2 Annotation Guidelines

Before starting the annotation contest it is important to write a first version of the annotation guidelines. The guidelines are then used by the annotators as a reference manual, that helps to

decide what to annotate (which words) and how (with which category). As Fort discusses, this should not be disregarded, as the quality of the contest depends on it.

The first version of our guidelines was written around the same time as consensus was being reached on the list of PHI categories, in December 2021. The document has been revised 6 times, each revision leading to a new edition. The latest English version is attached in Appendix A. It was translated from English into French, and made available in markdown, HTML, and PDF formats. Figure 3.1 shows snapshots of the cover page and the table of contents. The following is a list of some of its features:

- The categories are color coded the same way as in the annotation tool for visual aid.
- An interactive table of contents is displayed at the beginning of the document (on the first page before the cover page) to facilitate quick navigation.
- A short user guide for the annotation tool is included.
- Categories are listed in alphabetical order. We explain the scope of each category and provide examples of which words should or should not be annotated, clarifying the underlying logics when necessary as suggested by Fort. We present only what is essential and avoid being overly rigid or exhaustive to allow for an appropriate level of interpretation.

### 3.1.3 Prodigy Annotation Tool

In this work we used Prodigy [13], an efficient annotation tool for AI, machine learning, and NLP. The importance of good annotation tools is often overlooked in the data science community and Prodigy undoubtedly played a primal role in the success of our work.

The user interface (see Figure 3.2) is highly customizable and also available in French. Figure 3.3 shows how we were able to customize the list of labels with our own color scheme, associating a color to each supercategory. We were also able to customize keyboard shortcuts for the labels.

Prodigy provides many different pipelines, each suited for a specific task. For this project we used `ner.manual`, a fully manual named entity recognition pipeline. It suited our task because the goal was to build an annotated corpus from scratch, without any existing resources.

Prodigy uses the concept of an annotation *task* to refer to a collection of annotations associated to a given input by a single annotator. Henceforth, we will use this term without explanation and distinguish it from an *input*.

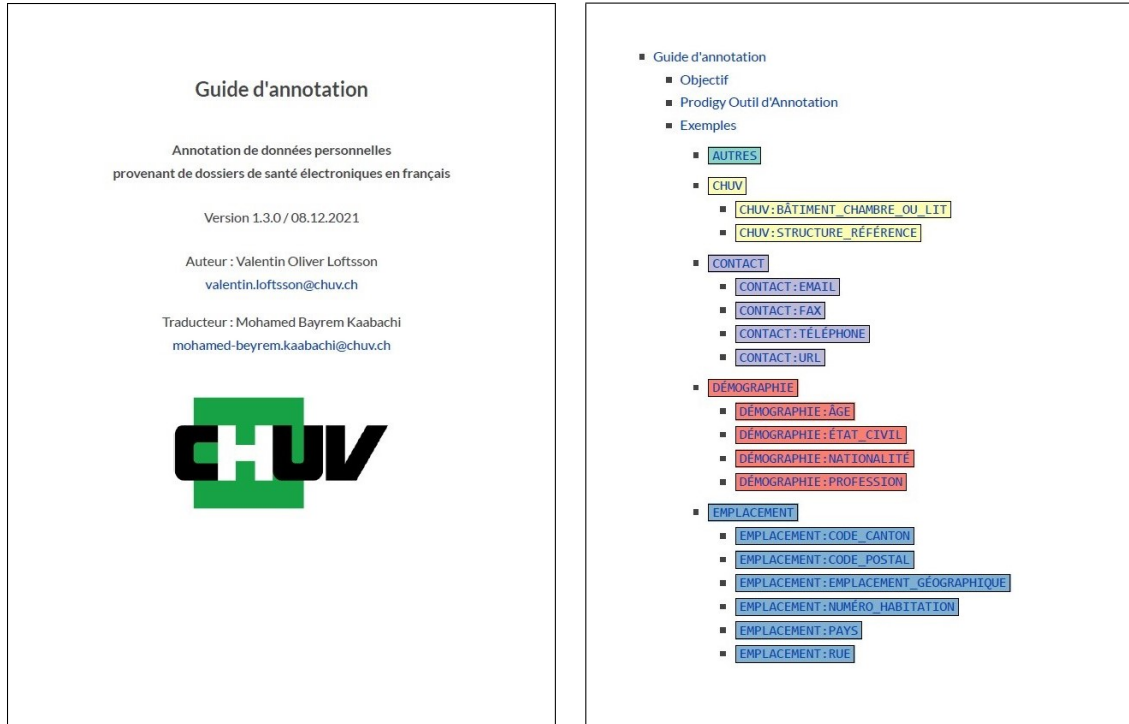


Figure 3.1: Annotation guidelines. The full document is attached in Appendix A.

### 3.1.4 Annotator Training

According to Fort [2], training the annotators is now widely considered vital for annotation quality. We prepared around 20 training tasks and held group training sessions for a few consecutive days before kicking off the contest. Participants were required to attend at least one session. Each session started with reading the annotation guidelines, during which any questions were welcome and discussed. The sessions, especially the first ones, exposed imprecisions in the annotation guidelines and thus provided opportunities for modification. We also explained how to use the Prodigy annotation tool. Afterward, participants went through the training tasks in Prodigy on their own. We instructed them to refer to the guidelines when in doubt.

## 3.2 Contest Management: Agile Annotation

We based our approach to the management of the contest on the idea of *agile annotation* introduced by Voorman and Gut [19]. Based on an analogy with the agile approach often used in software development, agile annotation implies an iterative process of updating annotation guidelines, annotation, and analysis throughout the entire annotation campaign (see Figure 3.4).

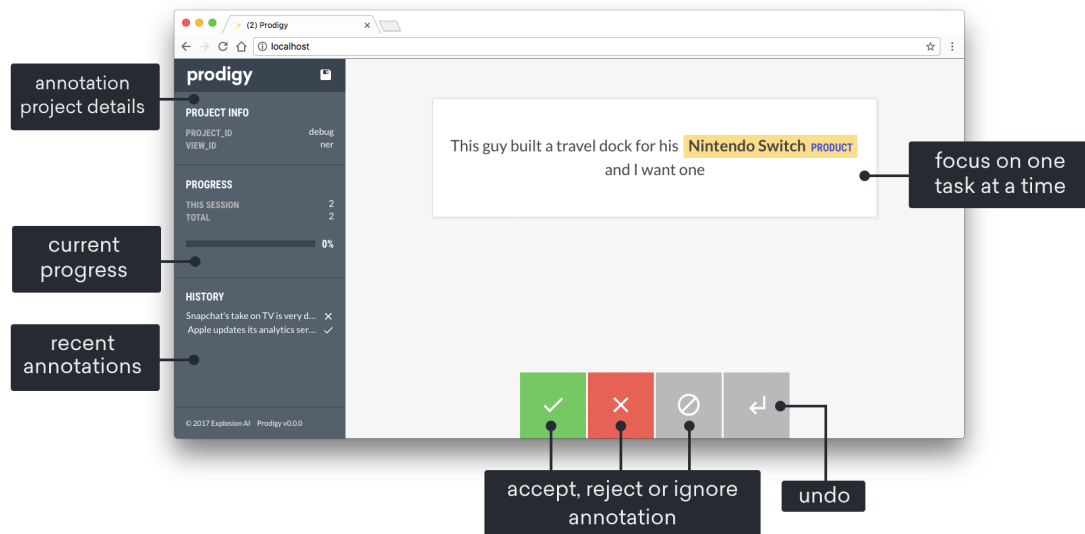


Figure 3.2: The Prodigy User Interface (source: [www.prodi.gy](http://www.prodi.gy)).

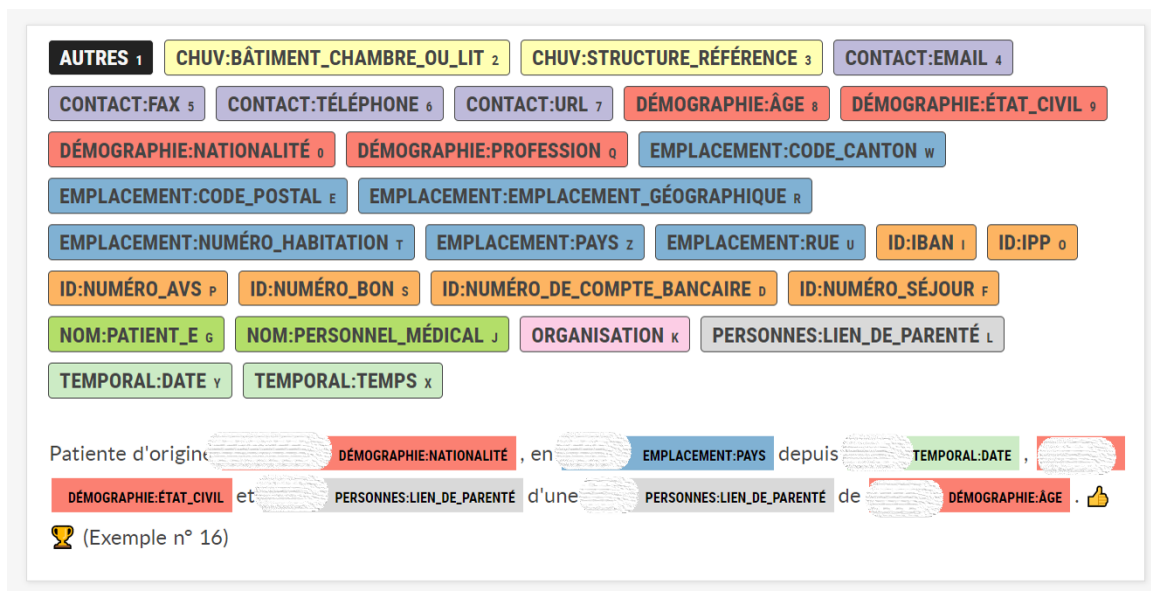


Figure 3.3: An annotation task in the Prodigy annotation tool. At the top is the color-coded list of labels in alphabetical order. The bottom right corner of each label shows the keyboard shortcut. Annotated personal data are erased in the image for privacy reasons.

### 3.2.1 Maintaining Motivation

We took several measures in the design of the contest to maintain motivation:

1. Participants are divided into teams that compete with each other. Even though annotators on the same team are not allowed to collaborate, annotators are motivated to do their best for their team.
2. Prizes are awarded to the best-performing individuals and teams.
3. Communication between annotators and contest managers is facilitated with a live chat room. As explained by Fort [2], this is a form of collaboration and fosters motivation. Therewith, annotators have a space where they can discuss ambiguities or bugs, encourage one another, and directly receive answers to questions.
4. Annotators receive regular feedback on their performance and rank via e-mail. They can also view their own and their team's rank in an online scoreboard.
5. Annotators can submit comments to contest managers when they believe the stated reference in their feedback file is incorrect and they have been wrongly penalized.

Maintaining enthusiasm towards the annotation work did not come without challenges. Annotators participated during working hours, taking time off from their daily work to annotate. Of course, the management of the Data Science Group was involved in the organization of the contest, and employees were encouraged to participate and not expected to deliver the same amount of work as usual during the contest. To further foster a collaborative spirit, we set a goal of 2,000 tasks per annotator and asked that each annotator make an effort towards the collective goal. We also estimated that each annotator would have to spend 20-30 minutes per day on average to reach this goal. At the end of the contest, the annotators had completed

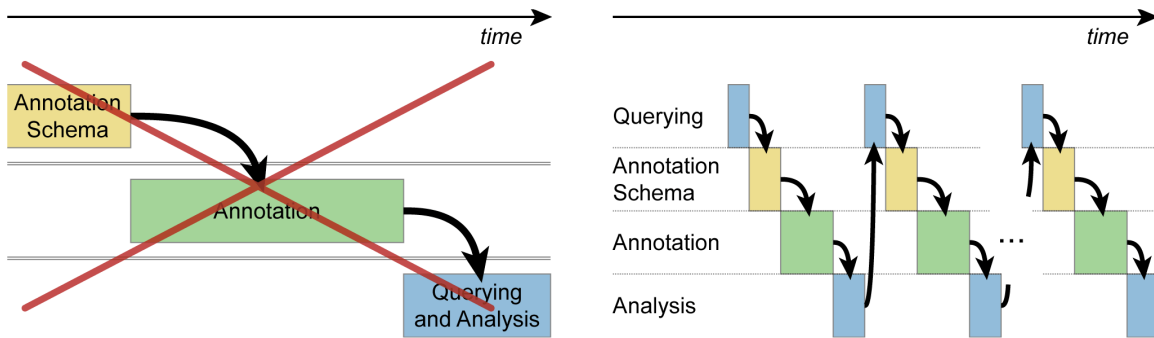


Figure 3.4: Traditional annotation stages (on the left) and cycles of agile annotation (on the right). Figure taken from Fort [2], p.19.

$\sim 14,000$  tasks in total and each annotator contributed  $\sim 880$  tasks on average towards the goal (around 44%). The average annotation throughput was 7.9 tasks/min.

### 3.2.2 Annotation Analysis and Annotator Feedback

Annotations were processed and analyzed every 2-3 days during the contest. The processing can be roughly divided into the following steps:

1. The raw data export from Prodigy is read and inserted into an easy-to-use data structure.
2. Metadata is extracted and validated. For example, for each task, we extract the annotator name and team from the session id which has the format `annotator_team`. The information is then checked against a registration sheet.
3. Ignored or rejected tasks are discarded.
4. Task durations are estimated based on timestamps returned from Prodigy. When the duration of a task is uncomputable, it is set to `NA`.
5. Duplicate tasks are discarded. This had to be handled due to a bug in Prodigy.
6. The inter-annotator agreement is computed and unitary alignments are computed from annotated units. The reference unit is computed as described in section 3.3 (readers are also referred to that section for the definitions of *unit* and *unitary alignment*).
7. Even with very detailed guidelines, the annotators will still disagree sometimes. Disagreements with the previously computed reference are recorded per annotator including associated penalties. Annotators get a single penalty if they disagree with either the reference category or segment.
8. Reference units associated with the same input sample are joined into a single record.

The resulting processed data structures were then used to update the scoreboard (see subsection 3.2.3) and auto-generate individualized feedback that was sent to every annotator privately. Figure 3.5 shows an example header of the feedback report and Figure 3.6 shows examples of disagreements recorded in the report. The list of disagreements is sorted in most-recent-first order for convenience. The annotators were encouraged to review the list for their improvement. They were also asked to take note of any potential mistakes in the reference since they have the option of reporting such mistakes to contest managers. Contest managers carefully review such reports and can subsequently record corrections in a dedicated file. This file is taken into account during the processing described above.

Moreover, such continuous evaluation enables contest managers to identify problems quickly and to minimize their impact by communicating constructive feedback to annotators, either individually or collectively. Also, ill-defined or ill-understood categories and rules can be resolved immediately by updating the annotation guidelines, and thus annotators waste less time on ambiguous examples.

### 3.2.3 Scoreboard

We built a simple scoreboard website with the Vue.js framework and Google Charts. The scoreboard displays individual and team statistics and ranks. The overall rank is a weighted sum of 4 factors: the number of tasks (20%); the number of annotations (20%); throughput (20%); and quality index (40%).

The quality index is computed based on the number of disagreements with the reference annotation. The reference is the majority vote unless a correction has been recorded (see section 3.3). Examples without a reference are not taken into consideration. The quality index for a given annotator is calculated as the following ratio:

$$\begin{aligned} \text{total}_{\#} &= \text{disagreements}_{\#} + \text{agreements}_{\#} \\ \text{index} &= \frac{\text{total}_{\#} - \text{disagreements}_{\#}}{\text{total}_{\#}} \end{aligned}$$

The quality index for each annotator is computed separately for segmentation and categorization. The average is then reported. The average index of contest participants is 91.8/100.0, meaning that annotators agreed on average 91.8% of the time with given reference.

Moreover, we set a goal for the number of tasks per annotator. The summed collaborative progress towards the overall goal is displayed at the top of the scoreboard. Figure 3.7 shows snapshots of the scoreboard.

## 3.3 Computing Agreement and Building an Annotated Corpus

We used the gamma ( $\gamma$ ) metric of Mathet et al. [11] as a measure of inter-annotator agreement. Like Krippendorff’s alpha ( $\alpha$ ), and in contrast with Cohen’s kappa ( $\kappa$ ), the gamma ( $\gamma$ ) is computed from observed and expected *disagreements*, instead of agreements. But the gamma is not merely an agreement measure. Before explaining the gamma method, let us introduce a few key concepts:

- An annotated *unit* consists of a *category* and a *segment* consisting of start and end boundaries, each corresponding to a position in the input.
- A *continuum* is a set of units produced by a given set of annotators which are attached to a timeline that represents the thing that is being annotated, for instance, text, audio, video (see Figure 3.8a).
- An *alignment* of a continuum is a set of *unitary alignments* such that each unit of every annotator in the alignment belongs to one and only one of its unitary alignments. In other words, the alignment is a partition of the complete set of units and each set in the partition is called a unitary alignment (see Figure 3.8b).

The gamma method concerns the joint task of unit locating and unit labeling. It is holistic in that the computation of the agreement (gamma) and the selection of the best alignment are interdependent and computed in a unified process. It considers all possible alignments and selects the one that minimizes the overall discrepancy, or *disorder*. This disorder is computed by considering the full continuum.

With the gamma method, we can hit two birds with one stone, since it provides us with two things: an inter-annotator agreement measure and the so-called *best alignment*. We used Titeux and Riad’s `pygamma-agreement` implementation of the gamma method, written in Python [18]. We report an average gamma measure of 0.86/1.00 for annotations produced in the contest. The measure becomes 0.95/1.00 when empty continuums are taken into account, assuming they have the maximum agreement value 1.00 since annotators perfectly agree that nothing should be annotated in those continuums.

Below, we list the steps taken to build an annotated reference for a given input text using the gamma method.

1. Collect all annotations of the input (possibly from multiple annotators).
2. Create a unit for each annotation.
3. Create a continuum from the set of units. Add all annotators to the continuum who annotated the input, including those who did not annotate anything in the text.
4. Keep separate collections for empty continuums and those annotated by only a single annotator, since the gamma method cannot be applied to them.
5. Determine the best alignment and the associated gamma value for every non-empty multi-annotated continuum.
6. Process the unitary alignments of the best alignment, and determine for each unitary alignment



- (a) segment and category frequencies (counts).
  - (b) the majority segment and category.
  - (c) the reference (gold standard) segment and category. The reference is set as the majority vote by default. A correction recorded by an annotation manager takes precedence regardless of whether a majority vote was reached or not.
  - (d) disagreements and associated penalties.
7. If the reference exists for *both* the segment and the category (and the reference category is not the empty category) we join them together in a unit and add it to the set of gold standard units for this input. We associate metadata to the input such as the list of annotators, segment and category counts, and the gamma value, which will help us later evaluate the quality of the sample.

Annotation Contest Feedback  
Auto-Generated Report  
13.03.2022 10:54:11

Annotator: linguine  
Team: gnocchi

Number of tasks: 1029  
Tasks per minute: 6.1  
Quality index: 96.8%

Rank

-----  
[20%] Number of tasks: 5/17  
[20%] Number of annotations: 5/17  
[20%] Tasks per minute: 10/17  
[40%] Quality index: 4/17  
-----

Total rank: 4/17

Number of disagreements: 60

~ Category Disagreement Distribution ~

This shows the number of times you disagreed with the reference annotation in each category.

Note: "None" denotes the blank category, which means you did not annotate while the reference does.

None: 17  
ORGANISATION: 3  
PERSONNES:LIEN\_DE\_PARENTÉ: 3  
CHUV:STRUCTURE RÉFÉRENCE: 3  
DÉMOGRAPHIE:ÂGE: 3  
TEMPORAL:TEMPS: 2  
CHUV:BÂTIMENT\_CHAMBRE\_OU\_LIT: 1

Below is a summary of your disagreements with the reference annotation, when it is available.

Figure 3.5: Feedback Document Header.

<p>Text span: «soins continus de médecine»</p> <p>~~~~~ TEXT ~~~~~</p> <p>Reference: None ✓</p> <p>(3/5): None ✓ (2/5): «soins continus de médecine» ✗</p> <p>linguine: «soins continus de médecine» ✗</p> <p>~~~~~ CATEGORY ~~~~~</p> <p>Reference: None ✓</p> <p>(3/5): None ✓ (2/5): CHUV:STRUCTURE_RÉFÉRENCE ✗</p> <p>linguine: CHUV:STRUCTURE_RÉFÉRENCE ✗</p>	<p>Text span: «12h»</p> <p>~~~~~ TEXT ~~~~~</p> <p>Reference: «12h» ✓</p> <p>(2/4): «12h» ✓ (2/4): None ✗</p> <p>linguine: «12h» ✓</p> <p>~~~~~ CATEGORY ~~~~~</p> <p>Reference: TEMPORAL:TEMPS ✓</p> <p>(2/4): TEMPORAL:TEMPS ✓ (2/4): None ✗</p> <p>linguine: TEMPORAL:TEMPS ✓</p>
--	--

(a) Example 1: The majority vote is that no annotation should be made. The annotator is penalized for not siding with the majority.

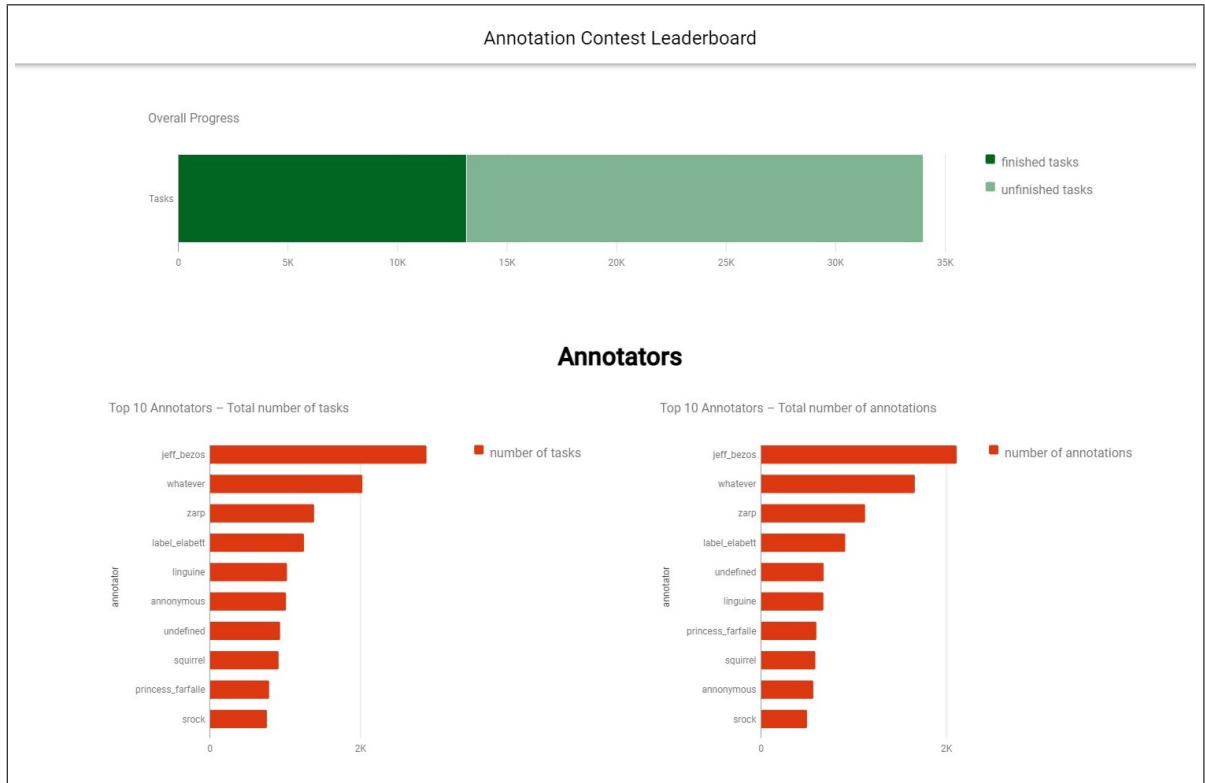
(b) Example 2: No majority is reached. However, a correction was submitted for both the segment and the category. Therefore, the annotator gets a point, and the others who annotated incorrectly are penalized.

<p>Text span: «lit A»</p> <p>~~~~~ TEXT ~~~~~</p> <p>No reference available</p> <p>(3/8): None (3/8): «lit A» (2/8): «A»</p> <p>linguine: «A»</p> <p>~~~~~ CATEGORY ~~~~~</p> <p>Reference: CHUV:BÂTIMENT_CHAMBRE_OU_LIT ✓</p> <p>(5/8): CHUV:BÂTIMENT_CHAMBRE_OU_LIT ✓ (3/8): None ✗</p> <p>linguine: CHUV:BÂTIMENT_CHAMBRE_OU_LIT ✓</p>	<p>Text span: «02:00»</p> <p>~~~~~ TEXT ~~~~~</p> <p>No reference available</p> <p>(2/4): None (2/4): «02:00»</p> <p>linguine: «02:00»</p> <p>~~~~~ CATEGORY ~~~~~</p> <p>No reference available</p> <p>(2/4): None (2/4): TEMPORAL:TEMPS</p> <p>linguine: TEMPORAL:TEMPS</p>
---	---

(c) Example 3: No majority is reached for the segment and no annotator is penalized. However, the category has a majority reference and this particular annotator gets a point for siding with the majority.

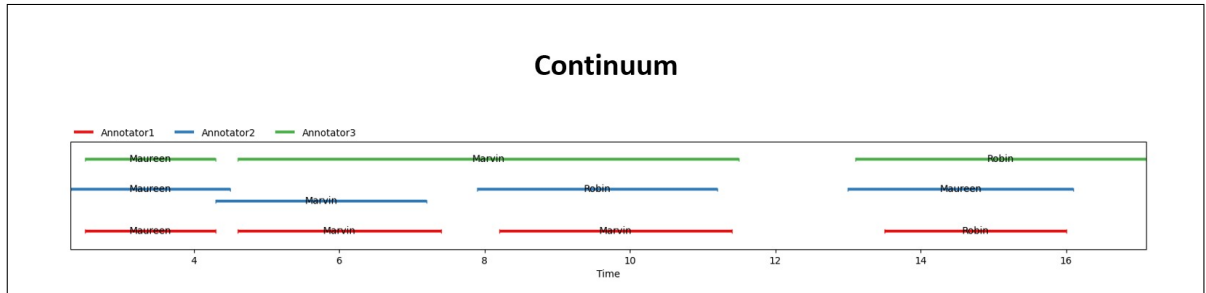
(d) Example 4: Majority is not reached and no correction was submitted, so the reference is not available. No points or penalties are given.

Figure 3.6: Examples of Annotator Feedback.

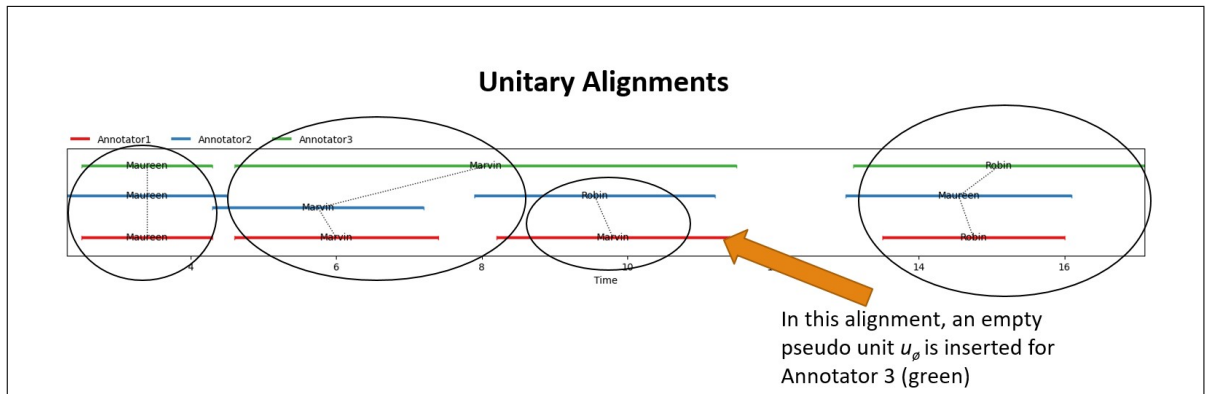


total rank	annotator	team	number of tasks	number of annotations	tasks per minute	quality index
1	jeff_bezos	farfalle	2,881	2,113	11.2	94.7
2	zarp	tagliatelle	1,389	1,126	9.4	95.4
3	linguine	gnocchi	1,029	677	6.1	96.5
4	whatever	spaghetti	2,030	1,662	11.3	94.5
5	princess_farfalle	spaghetti	792	601	7.4	96.2
6	label_elabett	farfalle	1,254	912	13.0	91.9
7	gogi	spaghetti	370	287	10.2	96.1
8	alexandre	tagliatelle	356	306	5.9	95.3
9	random	spaghetti	100	74	9.7	94.6
10	srock	farfalle	762	500	5.9	94

Figure 3.7: Scoreboard.



(a) A multi-annotated *continuum*. Each row represents the annotated units of a given annotator, from the start to the end of the text.



(b) Example of an *alignment* of annotated units on a continuum. The alignment consists of disjoint *unitary alignments*, encircled in the figure. Each unitary alignment groups at most one unit of each annotator.

Figure 3.8: Some basic concepts underlying the gamma  $\gamma$  metric.

## Chapter 4

# Rule-Based PHI Detection

The rule-based strategy is typically the first approach taken to build named entity recognition systems. Some studies have been done on rule-based detection of PHI in French clinical texts. Gaudet-Blavignac et al. [4] did this on Swiss-French clinical texts at the Geneva University Hospital. They achieved a noteworthy 0.93 total recall and 0.99 total precision on a random sample of 1000 letters. The output of a rule-based system is usually convenient to interpret since they tend to be fully deterministic. However, this comes at a price because this means they are more rigid and cannot handle the noise inherent in natural language as well as their machine learning counterparts. Gaudet-Blavignac et al. noted that errors in the text, such as spelling mistakes, affected their system’s performance.

Typically, rules are good at detecting structured entities like IDs and dates but not as good at detecting unstructured entities like locations. For instance, the lowest recall reported by Gaudet-Blavignac et al. was 0.79, on the Locations category. Low recall rates for location-related categories were similarly reported by Grouin and Zweigenbaum [6].

Moreover, as rules become more complex the system becomes more difficult to maintain. The complex and over-comprehensive rules are difficult to generalize for use in another context. Although Gaudet-Blavignac et al. achieved good results with their rule-based system, it cannot be used in the CHUV context.

Despite the drawbacks, a rule-based system can still be exploited. If it is reasonably accurate, it can be used to pre-annotate text samples before presenting them to human annotators. In section 4.2 we show that human annotation throughput and quality are significantly improved with the help of rule-based pre-annotation. Furthermore, we use the system as a baseline that machine learning models can compare with, and demonstrate the benefits it brings to use the rule-based approach in parallel with a machine learning model. Our results are presented in chapter 5.

In this chapter, we present the rule-based system we developed for detecting PHI categories in CHUV clinical texts. We also present experimental results on the impact of rule-based pre-annotations on human annotation.

## 4.1 Building a Rule-based System

Our rule-based system was built with spaCy [7], a natural language processing library written in Python. It was developed through repeated cycles of modification and manual evaluation. The manual evaluation was done by visualizing the output of the system with displacy, spaCy’s visualization suite.

The system’s capabilities were extended by allowing categories to depend on each other, which was conveniently made possible with spaCy’s API. In practice, this means that rules for particular categories refer to the presence (or absence) of other categories. For instance, house numbers appear after street names and we can exploit this knowledge by only tagging something that looks like a house number if a street name comes before it. This means that the processing order of categories matters as shown in Figure 4.1. The figure also reveals where the order of categories matters to avoid overwriting correct annotations with incorrect ones. The following list describes how the rule-based system operates for each category. Lookup tables are case insensitive unless otherwise stated:

- **CHUV**

BÂTIMENT\_CHAMBRE\_OU\_LIT is detected using regular expressions and trigger words like *Salle*. STRUCTURE RÉFÉRENCE uses a lookup table of acronyms and full names of departments and units harvested from CHUV’s website and the data warehouse. It also relies on trigger phrases like *Département des...* and *Unité de l’...*

- **CONTACT**

EMAIL and URL rely on spaCy’s internal regular expressions. TÉLÉPHONE uses regular expressions and trigger words like *Tél* and *Téléphone*. FAX uses the trigger word *Fax* with a simple regular expression that overwrites any previously tagged telephone numbers.

- **DÉMOGRAPHIE**

ÂGE relies heavily on trigger words to capture phrases that refer to a person’s age, like *patiente de 2 ans et trois mois* and *un nouveau-né d’une semaine et trois jours*. The rules are embedded in complex regular expressions with look-around patterns. ÉTAT\_CIVIL uses a small lookup table with words like *célibataire* and *marié*. NATIONALITÉ relies on a lookup table, harvested from a website with French words for nationality. PROFESSION does not have any rules and is ignored.

- **EMPLACEMENT**

CODE\_CANTON uses a simple case sensitive lookup table containing all the canton codes. It relies on the presence of a preceding word that is tagged with EMPLACEMENT\_GÉOGRAPHIQUE, for example, *Lausanne VD*. CODE\_POSTAL is identified as any number of digits preceding a EMPLACEMENT\_GÉOGRAPHIQUE term. EMPLACEMENT\_GÉOGRAPHIQUE uses a lookup table containing names of Swiss cantons, districts, municipalities, and localities from the public records of the Swiss Federal Statistics Office, dating back to the year 2000. NUMÉRO\_HABITATION uses a regular expression built by studying Swiss house number formats in data from the Federal Statistics Office and the Swiss Post. It relies on finding words that match the regular expression and were previously tagged with RUE. PAYS relies on a lookup table, harvested from a website with French

country names. RUE uses complex combinations of lookup tables, spaCy patterns, and regular expressions, developed by studying the corpus as well as data from the Federal Statistics Office and the Swiss Post. The method is greedy and also captures house numbers to provide hints for overwriting when NUMÉRO\_HABITATION is processed.

- ID

IPP does not have a fixed format and is detected with a regular expression using the trigger word *IPP*. NUMÉRO\_BON relies on the trigger phrase *No bon demande*. NUMÉRO\_SÉJOUR has a specific format that is encoded in a regular expression.

- NOM

PATIENT\_E and PERSONNEL\_MÉDICAL are detected with similar methods. One method exploits honorifics like *Dr.* and *Madame*. Another method recursively finds names by checking if an already tagged name is followed or preceded by a word that looks like a name and is not already tagged (for example, it should not be in lower case or a punctuation character).

- ORGANISATION

ORGANISATION detects medical institutions with trigger phrases like *Centre médical de...*

- PERSONNES

LIEN\_DE\_PARENTÉ uses a lookup table of words like *mère* and *frère* in singular and plural forms.

- TEMPORAL

DATE uses multiple thoroughly tested patterns and regular expressions to detect date elements in various formats like *2022*, *18.03.2022*, *mars 2022*, *vendredi*, etc. TEMPS uses a couple of simple regular expressions.

The system was evaluated on the full reference corpus built from the annotation contest. By *full* we mean that all samples regardless of the agreement value or the number of annotators were included (see section 5.1). We estimated that "low" quality samples should be acceptably scarce since the average annotator quality index and gamma  $\gamma$  agreement are reasonably good (see subsection 3.2.3 and section 3.3). Inputs annotated by a single annotator with a low quality index ( $< 85/100$ ) were discarded. The system achieved macro average 0.88 precision and 0.83 recall when evaluated on this corpus. Detailed results are presented in Table 4.1. Since we have imbalanced categories the macro average better represents the system's performance. The confusion matrix is shown in Figure 4.2.



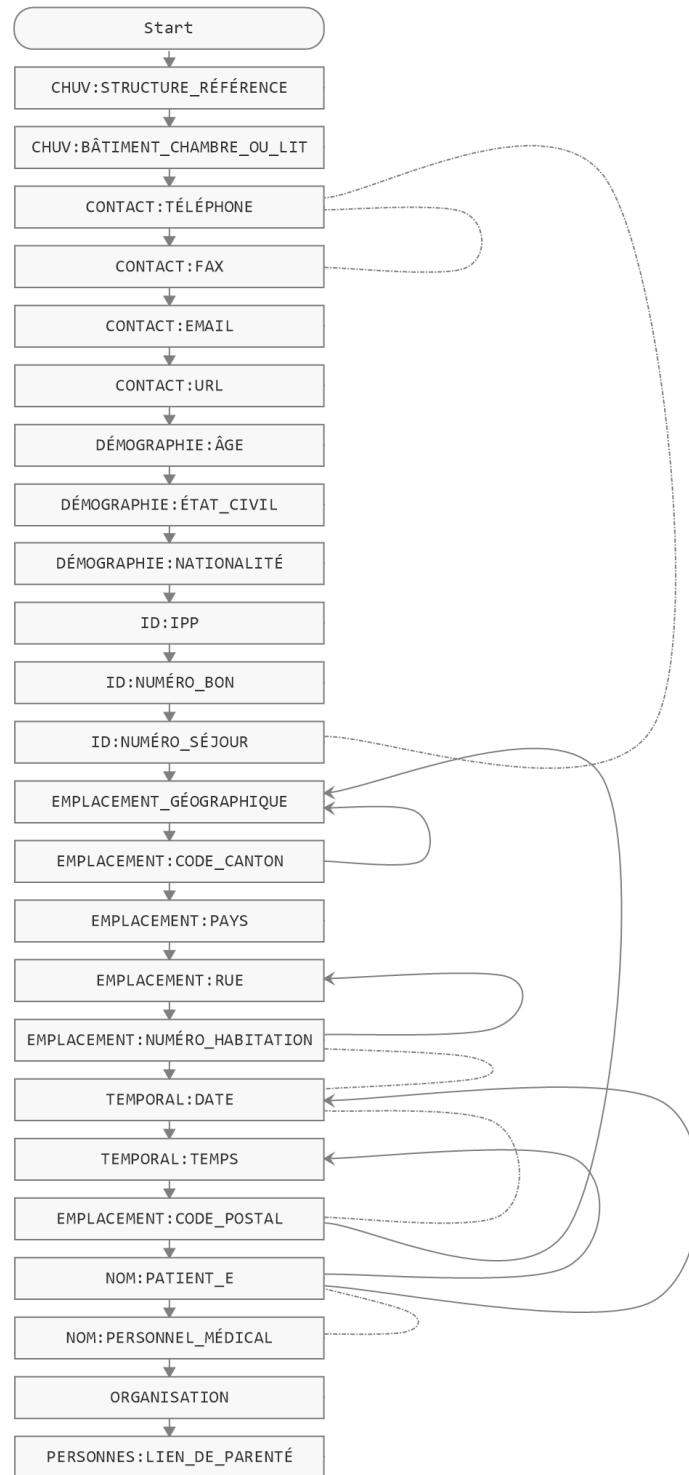


Figure 4.1: The rule-based system processing order of categories. Dependencies are illustrated with arrows. A solid arrow going from one category to another signifies that for identifying instances of the former the system relies on identifying instances of the latter. A dotted line between two categories signifies that the order of the categories is crucial to avoid overwriting.

Table 4.1: Evaluation metrics obtained from applying the rule-based system on the entire reference corpus built from annotations collected in the contest.

	support	precision	recall	F1-score
<b>CHUV</b>				
BÂTIMENT_CHAMBRE_OU_LIT	10	0.88	0.70	0.78
STRUCTURE RÉFÉRENCE	178	0.65	0.41	0.50
<b>CONTACT</b>				
EMAIL	7	1.00	0.86	0.92
FAX	11	1.00	0.91	0.95
TÉLÉPHONE	31	0.79	0.87	0.83
URL	1	1.00	1.00	1.00
<b>DÉMOGRAPHIE</b>				
NATIONALITÉ	7	0.67	0.86	0.75
PROFESSION	1	*0.00	0.00	*0.00
ÂGE	136	0.97	0.65	0.78
ÉTAT_CIVIL	3	0.75	1.00	0.86
<b>EMPLACEMENT</b>				
CODE_CANTON	12	1.00	0.92	0.96
CODE_POSTAL	101	1.00	0.96	0.98
EMPLACEMENT_GÉOGRAPHIQUE	172	0.85	0.88	0.87
NUMÉRO_HABITATION	94	0.93	0.96	0.94
PAYS	38	1.00	0.89	0.94
RUE	94	0.95	0.96	0.95
<b>ID</b>				
IPP	9	1.00	1.00	1.00
NUMÉRO_BON	33	1.00	1.00	1.00
NUMÉRO_SÉJOUR	116	0.99	1.00	1.00
<b>NOM</b>				
PATIENT_E	417	0.97	0.68	0.80
PERSONNEL_MÉDICAL	133	0.93	0.94	0.94
<b>ORGANISATION</b>				
ORGANISATION	144	0.92	0.68	0.78
<b>PERSONNES</b>				
LIEN_DE_PARENTÉ	153	0.84	0.86	0.85
<b>TEMPORAL</b>				
DATE	1818	0.99	0.97	0.98
TEMPS	307	0.98	0.89	0.93
<b>total support</b>				
	4026			
<b>micro average</b>		0.95	0.88	0.91
<b>macro average</b>		0.88	0.83	0.85
<b>weighted average</b>		0.95	0.88	0.91

\* Evaluated as undefined (zero division) but set to zero.

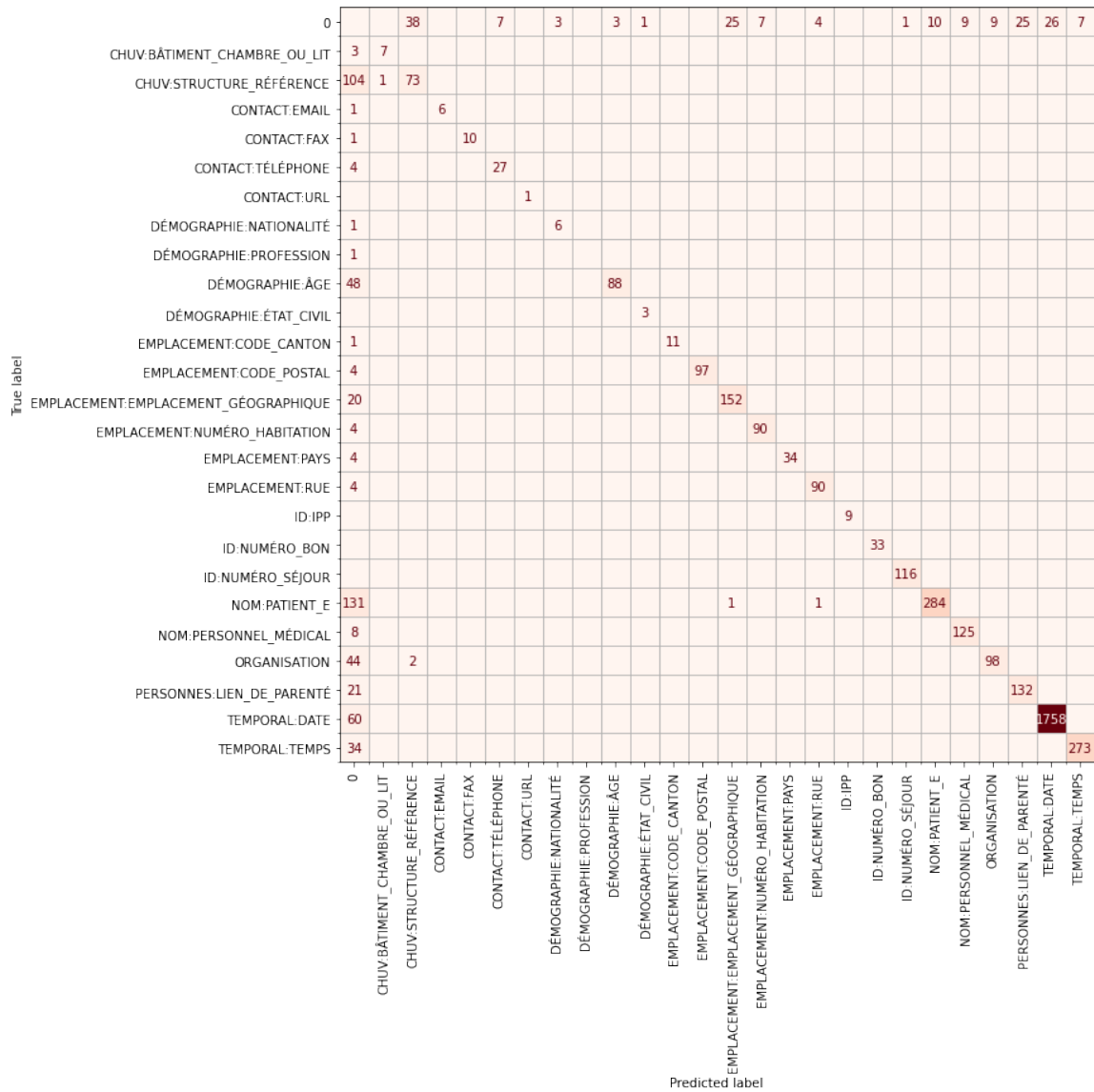


Figure 4.2: Confusion matrix obtained when the rule-based system is applied on the entire reference corpus. "0" denotes an empty annotation. The figure reveals that the rule-based system rarely confuses one category with another.

## 4.2 Evaluating the Impact of Rule-Based Pre-Annotation on Human Annotation

Rule-based systems have been used to pre-annotate text samples to support human annotation. This has been done by Grouin and Névél [5] on clinical notes in French. They concluded that better pre-annotations increase the quality of the reference corpus but require more revision time. Lingren et al. [10] studied dictionary-based pre-annotations on an English corpus and the impact on annotation time. They concluded that time savings were indeed significant. In our work, we also studied the impact of rule-based pre-annotation on human annotation throughput and quality. In addition, we propose an experiment to assess the level of precision rule-based systems need to have to start helping human annotation.

Assuming our rule-based model performed reasonably well—which was later confirmed and the results can be viewed in Table 4.1—we sought to answer the following questions:

1. Do rule-based pre-annotations improve human annotation throughput and quality?
2. How precise do rule-based pre-annotations need to be to stop impairing and start improving human annotation throughput and quality?

We studied the first question and hypothesized that both throughput and quality improve with pre-annotations. Table 4.2 shows the experimental design. We split the 4 teams of annotators into 2 groups, 2 teams per group. The 2 teams in the same group got the same tasks in the same order. However, one team (the experimental team) got tasks with pre-annotations, while the other team (the control team) got tasks without pre-annotations. We manipulated the recall rate of the two experimental teams using retention probabilities given in Table 4.2. Figure 4.3 shows examples of a task pre-annotated with different retention probabilities.

Following the annotation contest, we sampled task duration values and penalty indicators to test our hypothesis in the following way. We recall that an annotator is given a penalty for disagreeing with the reference unit. A penalty is given separately for disagreement with the segment and disagreement with the category. Readers are referred to section 3.3 for the definitions of *unit* and *segment*.

- We sampled task durations to test the hypothesis that pre-annotations improve annotation throughput. We followed these steps:
  1. Duration of each task is computed from timestamps stored by Prodigy.
  2. Duration values exceeding 5 minutes are ignored (assuming the annotator took a break).

3. The first task is ignored since it is not possible to compute its duration.
  4. We performed a t-test for each group. We treated the value sets corresponding to 2 teams of the same group as 2 independent samples for the test.
- We sampled task penalty indicator values (1 if there is a penalty, 0 otherwise) to test the hypothesis that pre-annotations improve annotation segmentation and categorization quality. We followed these steps:
    1. We only consider alignments annotated by an even number of annotators in each of the two teams (but with at least 2 annotators in each team).
    2. For each unitary alignment, we compute the average quality score among annotators of each team.
    3. We performed a t-test for each group. We treated the value sets corresponding to 2 teams of the same group as 2 independent samples for the test.

Table 4.3 shows the average duration values and penalties for all the teams. Interestingly, differences in the average task duration between experimental and control teams for the two groups are similar. This could be explained by the fact that team **1-Experimental** needs to spend more time on finding missing annotations while team **2-Experimental** spends this same time on reviewing the pre-annotations. However, differences between penalties are more significant for group 2 compared with group 1. This indicates that the quality improvement for team **2-Experimental** with 100% pre-annotations was significantly greater than for team **1-Experimental** with only 50% pre-annotations. Table 4.4 shows the P-values of the t-tests performed for each group. The P-values strongly support our hypothesis given a 95% confidence level.

Due to time constraints, we were not able to conduct a second annotation contest to test our hypothesis concerning the second question. Table 4.5 shows the proposed experimental design. The idea is to split the annotators into 8 teams of 3-4 annotators (or 6 teams if annotators are too few). As before, we have 4 (or 3) groups with 2 teams; one team is the experimental team and the other is the control team. For each experimental team, we apply the rule-based system on the annotation tasks and we perturb each detected unit with a given probability. This has the effect of reducing the precision of pre-annotations, as opposed to reducing the recall as in the first experiment. Annotations in tasks given to control teams are not perturbed. Figure 4.4 shows an example of a task that has been pre-annotated with different perturbation probabilities.

We sample task duration values and penalties to test our hypothesis in the same manner as before. We perturb the units in the following way:

1. We filter pre-annotations with a fixed retention probability (for instance, 0.8) for all teams to reduce recall evenly among the teams. The aim is to generate missing units to co-occur with our perturbed units.
2. For each of the remaining units, we perturb it with a given probability (as listed in Table 4.5). No annotations are perturbed for control teams.
3. In the event of a perturbation occurring, we select 1 of the following 3 actions with equal probability:
  - Perturb the category by picking another one that is "similar". We can assume categories within the same supercategory are similar. Also, we can find categories that are often confused with one another by analyzing the annotator disagreements. This was done and a few pairs were identified, for example, (TEMPORAL:DATE, DÉMOGRAPHIE:ÂGE) and (EMPLACEMENT:PAYS, DÉMOGRAPHIE:NATIONALITÉ).
  - Perturb the bounds of the segment (left, right, or both with equal probability). The category is perturbed instead if the bounds cannot be shifted.
  - Perturb both the category and the bounds.

Table 4.2: The experimental setting for research question 1: *Do rule-based pre-annotations improve human annotation throughput and quality?* The numbers are retention probabilities (the probability of retaining a pre-annotated unit).

Group	1	2
Experimental	0.5	1.0
Control	0	0

Table 4.3: Average task duration and penalty ratios for each team.

Group	Team	Duration	Segmentation penalty ratio	Categorization penalty ratio
1	Experimental	6.0	0.063	0.059
	Control	9.2	0.090	0.084
2	Experimental	6.1	0.034	0.036
	Control	9.1	0.131	0.131

Table 4.4: P-values of one-tailed T-tests supporting the hypothesis that pre-annotations improve annotation throughput and quality.

Group	1	2
Throughput	1.61e−15	2.67e−11
Segmentation quality	0.0227	4.48e−11
Categorization quality	0.0292	7.81e−10

Table 4.5: The experimental setting for research question 2: *How precise do rule-based pre-annotations need to be to stop impairing and start improving human annotation throughput and quality?* The numbers are probabilities for perturbing a pre-annotated unit. A fixed retention probability is applied for all groups.

Group	1	2	3	4
Experimental (E)	0.2	0.4	0.6	0.8
Control (C)	0	0	0	0

GOLDBERG JOSEPH, 02#01#1980, N° de séjour: 1190253765

(a) Retention probability: 0.0 (control team)

GOLDBERG NOM:PATIENT\_E JOSEPH, 02#01#1980 TEMPORAL:DATE ,  
N° de séjour: 1190253765 ID:NUMÉRO\_SÉJOUR

(b) Retention probability: 0.5

GOLDBERG NOM:PATIENT\_E JOSEPH NOM:PATIENT\_E ,  
02#01#1980 TEMPORAL:DATE , N° de séjour: 1190253765 ID:NUMÉRO\_SÉJOUR

(c) Retention probability: 1.0

Figure 4.3: Example of retaining pre-annotations with a given probability.

CT thoracique au CHUV ORGANISATION le 18#03#2022 TEMPORAL:DATE à 14h30  
TEMPORAL:TEMPS (avec copie du rapport chez le médecin traitant).

Patient de 60 ans DÉMOGRAPHIE:ÂGE, connu pour un tabagisme actif et un diabète de type 2,

(a) Perturbation probability: 0.0 (control team)

CT thoracique au CHUV ORGANISATION le 18#03#2022 DÉMOGRAPHIE:ÂGE à 14h30  
TEMPORAL:TEMPS (avec copie du rapport chez le médecin traitant).

Patient de 60 ans DÉMOGRAPHIE:ÉTAT\_CIVIL, connu pour un tabagisme actif et un diabète de type 2

(b) Perturbation probability: 0.4

CT thoracique au CHUV le ORGANISATION 18#03#2022 à TEMPORAL:DATE 14h30 (  
TEMPORAL:DATE avec copie du rapport chez le médecin traitant).

Patient de 60 ans, DÉMOGRAPHIE:ÂGE connu pour un tabagisme actif et un diabète de type 2,

(c) Perturbation probability: 0.8

Figure 4.4: Example of perturbing pre-annotations with a given probability.



## Chapter 5

# Advanced PHI Detection

In this section, we present a method for splitting the corpus into subsets for training, validation, and testing. Thereafter, we present and evaluate deep learning and hybrid approaches to detect PHI categories in the clinical texts at CHUV.

### 5.1 Stratified and Quality-prioritized Multi-label Sampling

In classification tasks addressed using deep learners, the corpus is typically split into training, validation, and test sets. The validation set is used for validating the learner at the end of each epoch, to determine if training should be put to an early stop when the model starts to overfit the training set. The test set is held out until after training the model to evaluate the model's performance.

The corpus is typically split in a *stratified* fashion such that the proportion of examples of each category in each subset is approximately equal to that in the full corpus. Though random sampling can be fine if the corpus is large enough, it can be especially problematic when classes are imbalanced or when there are very few examples of some categories, such as in our case. Randomly sampling inputs can lead to producing subsets with zero examples for rare labels. This raises an issue in the calculation of evaluation metrics. Moreover, stratification is commonly recognized to improve training and inference in terms of bias and variance. [15]

In single-label classification, it is straightforward to do stratified sampling by directly grouping the input samples based on their labels. However, the PHI inference task we are considering here is a multi-label classification task because each input can be associated with an arbitrary number and combination of units. We recall that each unit is composed of a segment and a category (see section 3.3). So what we need to do is *stratified multi-label sampling*, i.e. we want to have an equal representation of every category across the three subsets. This is not straightforward to do. Sechidis et al. [15] presented two approaches, each with its own merits based on the characteristics of the data. They concluded that each of the two approaches is consistently better than random sampling.

Our annotated corpus consists of 5,454 samples of varying quality. The test and validation

sets should ideally include the best samples for credible evaluation. Therefore, we are not only concerned with stratified sampling but also prioritizing high-quality samples over low-quality ones for the test and validation sets. We have two quality indicators. First, we have the gamma annotator agreement value for each sample. Second, we have the number of annotators for each sample. In our corpus, the number of annotators ranges between 1-8 as Figure 5.1 illustrates. Here we propose a new method we call *stratified and quality-prioritized multi-label sampling* that attempts to accomplish both of these goals at once. We describe the algorithm below:

1. Compute category frequencies (counts) for each sample.
2. Compute the combined category counts for the full corpus.
3. Discretize gamma values into bins of fixed size. For example, size 0.05 by rounding to the nearest 0.05.
4. Sort the samples by quality, first by the discretized gamma value and then by the number of annotators.
5. Split the corpus into empty and non-empty sample pools. Empty samples are those with zero annotations.
6. Do the following, first for the test set and then for the validation set:
  - (a) Initialize a counter with zero for all categories and create an empty list for collecting samples to add to the set.
  - (b) Compute the expected number of samples and the expected (target) number of instances for each category.
  - (c) Iterate over samples in the non-empty sample pool, starting from the sample with the highest quality. Do the following for each sample:
    - i. Update the counter with the sample category counts. If the resulting new counter has category counts that are all lower or equal to their respective target count, add the sample to the list and remove it from the pool. Else, remove the sample counts from the counter and continue.
    - ii. Stop the iteration if the count is equal to or higher than the target count for every category.
  - (d) Complete the list with samples from the empty sample pool (in high-quality first order) until the expected number of samples is reached. Remove the selected samples from the pool.
7. Use the remaining samples in both pools for the training set.

Table 5.1: Quality indicators for the three subsets.

Subset	Average gamma value	Average number of annotators
Test	0.96	6.1
Validation	0.97	3.8
Training	0.83	1.5

Figure 5.2 compares our method with the random strategy and the ideal target distribution. We only sampled categories addressed in section 5.3 and excluded the rarest ones. We remark that our method produces subsets where the proportion of categories is closer to the target distribution than subsets produced by random sampling. Table 5.1 shows average quality indicators for the three subsets. The test set is of the best quality and the validation set is second-best, providing evidence as to how well we achieved our second goal of quality-prioritizing.

As observed from Figure 5.2 significant gaps still exist between the stratified count and desired target count of rare categories such as `EMPLACEMENT:PAYS`. To improve this, our method can be adjusted by examining rare categories in priority as Sechidis et al. did with their iterative algorithm. They examine each label separately starting from the rarest category. The motivation is that if rare labels are not processed first, then they may be distributed in an undesired way and the current distribution cannot be repaired subsequently. With frequent categories, the algorithm has the chance, later on, to modify the current distribution towards the target.

## Input Examples Grouped by Number of Annotators

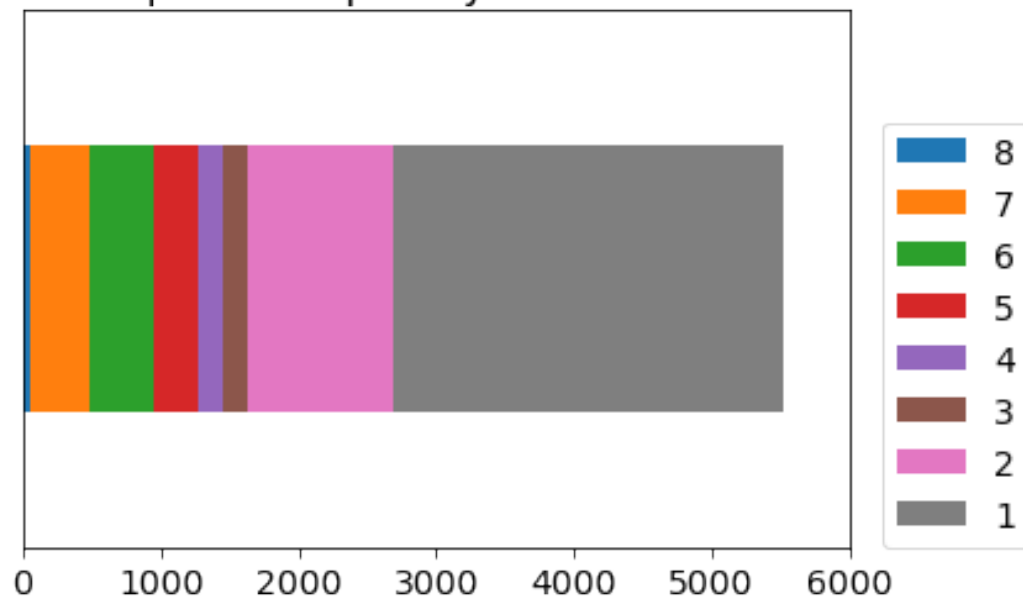


Figure 5.1: Inputs grouped by the number of annotators.

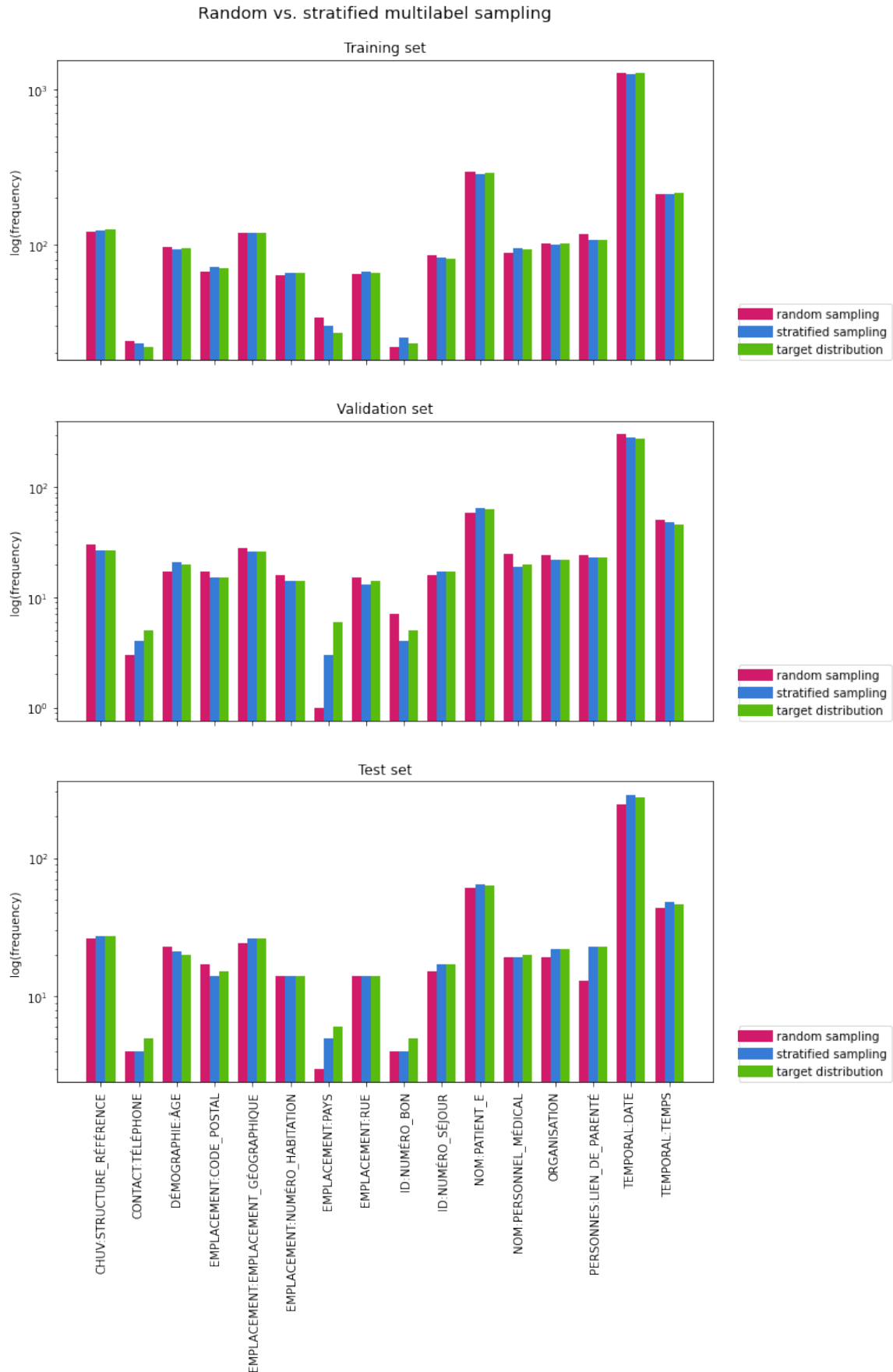


Figure 5.2: Random vs. stratified sampling. Note that the y-axes are log-scaled.

## 5.2 Bi-LSTM Model for PHI Detection

We developed a bidirectional LSTM model for PHI detection using Pytorch [12]. Figure 5.3 shows the architecture of our model. Dropout is used between layers to reduce overfitting. Pretrained embeddings were not used. All model variants were trained with the following fixed configuration:

- Embedding dimension: 128
- Hidden dimension: 64
- Dropout probability: 0.5
- Learning rate: 0.001
- Batch size: 32

We used the IOB2 tagging scheme to produce the target sequences which were mapped to indexes before feeding them to the model. Tokens were similarly mapped to indexes. All digits were replaced with 0 resulting in vocabulary size reduction by  $\sim 2,500$  and improved model performance. For example, replacing digits like this causes different time elements in the same format to be mapped to the same token: 12h00 and 23h59 are both mapped to 00h00. The resulting vocabulary size is  $\sim 11,000$ .

The model outputs log probabilities over the tagset and the tag with the highest probability is selected. In future work, other tags with high probabilities can be analyzed to understand which categories are ambiguous to the model.

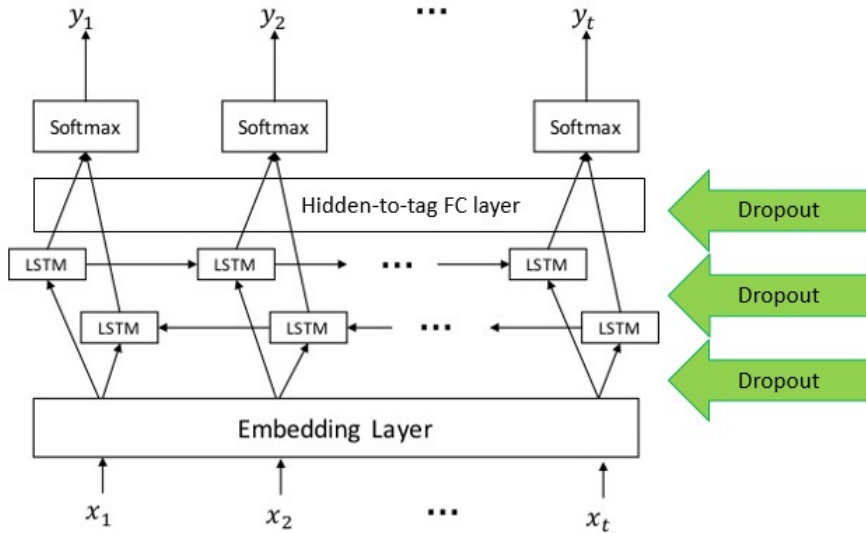


Figure 5.3: Bidirectional LSTM model architecture.

### 5.3 Model Evaluation and Comparison

We compared three bi-LSTM model variants and the rule-based system:

- **rulebased** is the baseline rule-based system.
- **lstm1** is a 1-layer bi-LSTM model.
- **lstm2** is a 2-layer bi-LSTM model.
- **hybrid** is a hybrid model composed of two components: a rule-based component and a bi-LSTM component with 1 layer. The rule-based component addresses categories on which the system achieved an F1-score higher than 0.9 on the corpus, as can be gleaned from Table 4.1 (with the exception of **TEMPORAL:TEMPS**). The bi-LSTM component addresses the other categories. Note that because this component addresses fewer categories, it has a smaller tagset than **lstm1** and **lstm2**.

We used a 70-15-15 train-validation-test split. Categories with fewer than 20 examples in the entire corpus were excluded (see the support column in Table 4.1).

Table 5.2 displays test set evaluation metrics. Results show that **rulebased** outperforms the other models both in terms of precision and recall with a macro average of 0.94 and 0.88, respectively. Interestingly, **lstm1** outperforms **lstm2**. This might be because 2 layers of LSTM contain too many parameters for the small size of the corpus, resulting in overfitting. Moreover, we observe that **hybrid** outperforms both **lstm1** and **lstm2**, showing the promise a hybrid arrangement has for improving PHI detection. Figure 5.4 shows the confusion matrix for **hybrid**.

Table 5.2: Evaluation results of individual PHI detection methods on the test set.

	support	precision rulebased	lstm1	lstm2	hybrid	recall rulebased	lstm1	lstm2	hybrid
CHUV									
STRUCTURE RÉFÉRENCE	27	<b>0.79</b>	0.65	0.67	0.67	0.41	0.48	0.15	<b>0.52</b>
CONTACT									
TÉLÉPHONE	4	0.60	0.75	0.50	<b>1.00</b>	<b>0.75</b>	<b>0.75</b>	0.25	<b>0.75</b>
DÉMOGRAPHIE									
ÂGE	21	<b>0.90</b>	0.61	0.72	0.56	0.43	0.52	<b>0.62</b>	0.48
EMPLACEMENT									
CODE_POSTAL	14	<b>1.00</b>	0.93	<b>1.00</b>	<b>*1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>*1.00</b>
EMPLACEMENT_GÉOGRAPHIQUE	26	0.96	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.81	0.62	0.77
NUMÉRO_HABITATION	14	<b>0.88</b>	0.87	0.87	<b>*0.88</b>	<b>1.00</b>	0.93	0.93	<b>*1.00</b>
PAYS	5	<b>1.00</b>	<b>1.00</b>	†0.00	<b>*1.00</b>	<b>1.00</b>	0.20	0.00	<b>*1.00</b>
RUE	14	0.93	<b>1.00</b>	<b>1.00</b>	<b>*0.93</b>	<b>1.00</b>	0.93	<b>1.00</b>	<b>*1.00</b>
ID									
NUMÉRO_BON	4	<b>1.00</b>	<b>1.00</b>	0.80	<b>*1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>*1.00</b>
NUMÉRO_SÉJOUR	17	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>*1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>*1.00</b>
NOM									
PATIENT_E	65	<b>1.00</b>	0.97	0.93	0.89	0.60	0.54	<b>0.66</b>	0.62
PERSONNEL_MÉDICAL	19	<b>1.00</b>	<b>1.00</b>	0.83	<b>*1.00</b>	<b>1.00</b>	0.47	0.53	<b>*1.00</b>
ORGANISATION									
ORGANISATION	22	<b>1.00</b>	0.79	0.68	0.91	<b>0.95</b>	0.86	0.86	0.91
PERSONNES									
LIEN_DE_PARENTÉ	23	0.95	<b>1.00</b>	0.69	0.95	<b>0.91</b>	0.83	0.78	0.83
TEMPORAL									
DATE	285	<b>1.00</b>	0.99	0.99	<b>*1.00</b>	0.98	0.98	<b>0.99</b>	<b>*0.98</b>
TEMPS	48	<b>1.00</b>	0.98	0.96	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
<b>total support</b>	608								
<b>micro average</b>		<b>0.98</b>	0.95	0.93	0.95	<b>0.90</b>	0.85	0.85	0.89
<b>macro average</b>		<b>0.94</b>	0.91	0.79	0.92	<b>0.88</b>	0.77	0.71	0.87
<b>weighted average</b>		<b>0.98</b>	0.95	0.91	0.95	<b>0.90</b>	0.85	0.85	0.89

\* Category was tagged with the rule-based component of the hybrid.

† Evaluated as undefined (zero division) but set to zero.



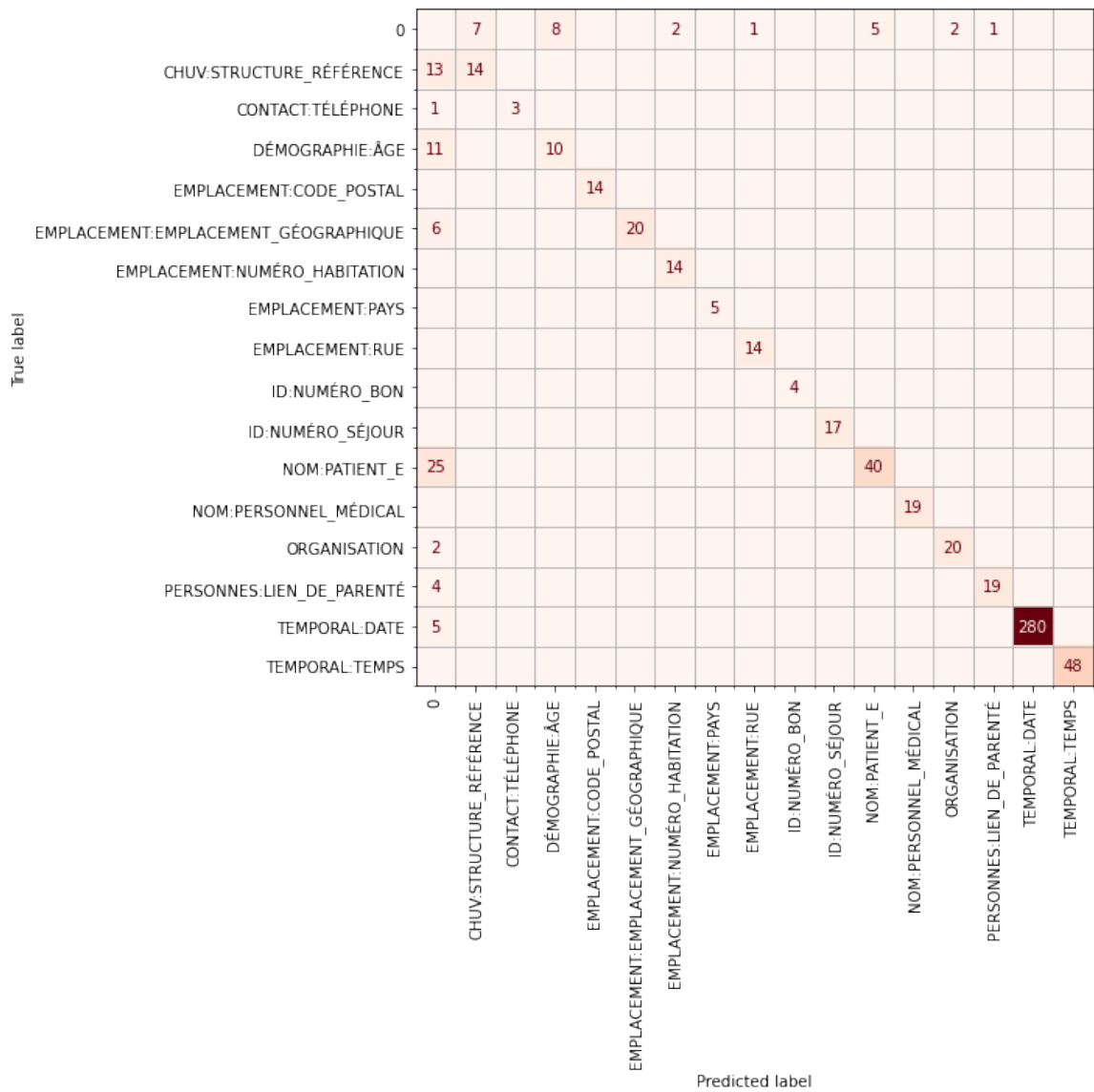


Figure 5.4: Confusion matrix obtained from applying the hybrid model on the test data set. "0" denotes an empty annotation. The figure reveals that the hybrid approach never misclassifies one category as another on this data set.

## Chapter 6

# Conclusion and Future Work

This work develops a solid foundation for the de-identification project at CHUV and introduces various strategies and findings relating to reference corpus development and PHI detection.

We determined the categories of PHI found in clinical texts at CHUV and grouped them into supercategories. Some of the categories are specific to CHUV, while others are more general and can provide inspiration for other projects.

We collected the first set of annotations by internal crowdsourcing within the organization. This was done by holding an annotation contest. The contest was carefully organized and involved several essential elements: easy-to-use annotation guidelines; an annotation tool with an optimized user interface, including customized keyboard shortcuts and color-coded labels; training sessions with practical examples; a live chat room to facilitate discussions and motivation; regular feedback via e-mail and an option for annotators to submit comments; a scoreboard with a collective progress-bar; and prizes. This foundation will be reused and elaborated in the future to hold more annotation contests and to expand the reference corpus. The strategies and methods we employed can also be replicated in other contexts, although their expression will surely be different in practice.

After annotations had been collected we built an annotated corpus by using the gamma  $\gamma$  method. The method computes the inter-annotator agreement and determines the best alignment of annotations from multiple annotators. The gold standard for each sample was built by a majority vote based on the best alignment. This enabled the computation of a quality index for each annotator based on the proportion of disagreements with the gold standard. Future work may involve developing an algorithm that calculates the gold standard by weighing annotated units with the annotator index instead of taking the majority vote, thus taking into account differences in annotator performance.

We built an inter-category-dependent rule-based model for PHI detection. It was evaluated on the entire reference corpus achieving macro averaged 0.88 precision and 0.83 recall. The impact of rule-based pre-annotations was studied and we reached the conclusion that pre-annotations improve human annotation throughput and quality. However, the following question remains to be answered: *How precise do rule-based pre-annotations need to be to stop impairing*

*and start improving human annotation throughput and quality?* This question will be addressed in a future annotation contest as discussed in section 4.2.

In section 5.1 we presented a new multi-label sampling method based on the gamma  $\gamma$  metric that achieves two goals: 1) stratified sampling to ensure an equal proportion of examples for each category across subsets, and 2) quality-prioritization to ensure test and validation samples are of the highest quality possible.

We developed a bi-LSTM model for PHI detection. Two variants of this model were compared with the rule-based system, one with a single layer and the other with two layers. Our rule-based system outperformed both models. We also built a hybrid model which did better than both bi-LSTM variants, thereby demonstrating the potential benefits of intertwining a rule-based approach with a recurrent network. However, the size of the corpus is yet too small to draw any conclusions since more training data will likely boost the performance of the bi-LSTM. Moreover, we used fixed hyper-parameters and did not tune them to improve performance. In addition, using pre-trained embeddings might boost performance even more. Future work may also involve active learning as a strategy to reduce annotation requirements (thereby saving financial resources) and potentially boost model performance, through interactive cycles of human annotation and model training.

Future work will improve the elements covered in this thesis (blue blocks in Figure 1.1) and start the remaining tasks in the project timeline: surrogate generation, re-identification risk assessment, and developing applications for end-users (red blocks in Figure 1.1). It is hoped that this work may be exploited cross-institutionally in other Swiss hospitals and inspire similar endeavors in French-speaking medical institutions around the globe.

# Bibliography

- [1] Loick Bourdois, Marta Avalos, Gabrielle Chenais, Frantz Thiessard, Philippe Revel, Cédric Gil-Jardiné, and Emmanuel Lagarde. “De-identification of Emergency Medical Records in French: Survey and Comparison of State-of-the-Art Automated Systems”. en. In: *Florida Artificial Intelligence Research Society* 34.1 (May 2021). DOI: 10.32473/flairs.v34i1.128480. URL: <https://hal.inria.fr/hal-03241384> (visited on 10/26/2021).
- [2] Karën Fort. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Ed. by Patrick Paroubek. Wiley-ISTE, July 2016. URL: <https://hal.archives-ouvertes.fr/hal-01324322> (visited on 11/09/2021).
- [3] Vasiliki Foufi, Christophe Gaudet-Blavignac, Raphaël Chevrier, I, and Christian Lovis. “De-Identification of Medical Narrative Data”. In: *The Practice of Patient Centered Care: Empowering and Engaging Patients in the Digital Era* (2017). Publisher: IOS Press, pp. 23–27. DOI: 10.3233/978-1-61499-824-2-23. URL: <https://ebooks.iospress.nl/doi/10.3233/978-1-61499-824-2-23> (visited on 10/24/2021).
- [4] Christophe Gaudet-Blavignac, Vasiliki Foufi, Eric Wehrli, and Christian Lovis. “De-identification of French medical narratives”. en. In: *Swiss Medical Informatics* 34.00 (Sept. 2018). Publisher: EMH Media. DOI: 10.4414/smi.34.00417. URL: <https://medical-informatics.ch/article/doi/smi.34.00417> (visited on 10/25/2021).
- [5] Cyril Grouin and Aurélie Névél. “De-identification of clinical notes in French: towards a protocol for reference corpus development”. eng. In: *Journal of Biomedical Informatics* 50 (Aug. 2014), pp. 151–161. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2013.12.014.
- [6] Cyril Grouin and Pierre Zweigenbaum. “Automatic de-identification of French clinical records: comparison of rule-based and machine-learning approaches”. eng. In: *Studies in Health Technology and Informatics* 192 (2013), pp. 476–480. ISSN: 1879-8365.
- [7] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. “spaCy: Industrial-strength Natural Language Processing in Python”. In: (2020). DOI: 10.5281/zenodo.1212303.
- [8] Youngjun Kim, Paul Heider, and Stéphane Meystre. “Ensemble-based Methods to Improve De-identification of Electronic Health Record Narratives”. In: *AMIA Annual Symposium Proceedings* 2018 (Dec. 2018), pp. 663–672. ISSN: 1942-597X. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371277/> (visited on 10/25/2021).

- [9] Youngjun Kim, Paul M Heider, and Stéphane M Meystre. “Comparative Study of Various Approaches for Ensemble-based De-identification of Electronic Health Record Narratives”. en. In: (), p. 10.
- [10] Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. “Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements”. In: *Journal of the American Medical Informatics Association : JAMIA* 21.3 (May 2014), pp. 406–413. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2013-001837. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3994857/> (visited on 11/04/2021).
- [11] Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. “The Unified and Holistic Method Gamma ( $\gamma$ ) for Inter-Annotator Agreement Measure and Alignment”. In: *Computational Linguistics* 41.3 (Sept. 2015), pp. 437–479. DOI: 10.1162/COLI\_a\_00227. URL: <https://aclanthology.org/J15-3003> (visited on 11/09/2021).
- [12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [13] *Prodigy*. URL: <https://prodi.gy/>.
- [14] Office for Civil Rights (OCR). *Health Information Privacy*. en. Text. Last Modified: 2021-08-16T16:08:51-0400. June 2021. URL: <https://www.hhs.gov/hipaa/index.html> (visited on 10/22/2021).
- [15] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. “On the Stratification of Multi-label Data”. en. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgianis. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 145–158. ISBN: 978-3-642-23808-6. DOI: 10.1007/978-3-642-23808-6\_10.
- [16] *SR 235.1 - Federal Act of 19 June 1992 on Data Protection (FADP)*. URL: [https://www.fedlex.admin.ch/eli/cc/1993/1945\\_1945\\_1945/en](https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en) (visited on 03/11/2022).
- [17] Amber Stubbs and Ozlem Uzuner. “Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth Corpus”. In: *Journal of biomedical informatics* 58.Suppl (Dec. 2015), S20–S29. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2015.07.020. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4978170/> (visited on 10/25/2021).

- [18] Hadrien Titeux and Rachid Riad. “pygamma-agreement: Gamma  $\gamma$  measure for inter/intra-annotator agreement in Python”. en. In: *Journal of Open Source Software* 6.62 (June 2021), p. 2989. ISSN: 2475-9066. DOI: 10.21105/joss.02989. URL: <https://joss.theoj.org/papers/10.21105/joss.02989> (visited on 11/19/2021).
- [19] Holger Voormann and Ulrike Gut. “Agile corpus creation”. In: 2008. DOI: 10.1515/CLLT.2008.010.

# Glossary

**AI** Artificial Intelligence. 16

**API** Application Programming Interface. 29

**bi-LSTM** Bidirectional LSTM (see definition of LSTM). 2, 3, 45, 49

**CDATA** The term CDATA, meaning character data, is used for distinct, but related, purposes in the markup languages SGML and XML. The term indicates that a certain portion of the document is general character data, rather than non-character data or character data with a more specific, limited structure. 9

**CHUV** Centre hospitalier universitaire vaudois (Lausanne University Hospital). 2, 3, 6–9, 11, 13, 15, 28, 39, 48

**EHR** Electronic Health Record. 5, 6, 9, 13

**FADP** The Federal Act on Data Protection of June 19, 1992 is the Swiss Data Protection Act. It has been revised several times, most recently on March 1, 2019. A new Data Protection Act (DPA) was passed by the Swiss parliament on September 25, 2020 and is expected to come into force in 2022. It has been revised to adopt a large variety of requirements from the General Data Protection Regulation (GDPR). 10, 11

**GDPR** The General Data Protection Regulation 2016/679 is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area. 10, 11

**HIPAA** The Health Insurance Portability and Accountability Act of 1996 is a United States federal statute enacted by the 104th United States Congress and signed into law by President Bill Clinton on August 21, 1996. 10, 11

**HTML** The HyperText Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. 9, 16

**IOB2** The IOB format (short for inside, outside, beginning) is a common tagging format for tagging tokens in a chunking task in computational linguistics (ex. named-entity recognition). The IOB2 format is the same as the IOB format except that the B- tag is used in the beginning of every chunk (i.e. all chunks start with the B- tag). 44

- LSTM** Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). 44, 45, 53
- NFKC** One of four unicode normalization forms. NF (“Normalization Form”) + KC (“Compatibility Decomposition, followed by Canonical Composition”). 10
- NLP** Natural Language Processing. 16
- PDF** Portable Document Format (PDF), standardized as ISO 32000, is a file format developed by Adobe in 1992 to present documents. 16
- PHI** Protected Health Information. 2, 3, 5, 6, 10–13, 16, 28, 39, 44–46, 48, 49
- XML** Extensible Markup Language is a markup language and file format for storing, transmitting, and reconstructing arbitrary data. 9
- XPath** XPath is an expression language designed to support the query or transformation of XML documents. 9



## Appendix A

# Annotation Guidelines

The annotation guidelines described in subsection 3.1.2 are attached in English (see next page).

- Annotation Guidelines

- Purpose
- Prodigy Annotation Tool
- Examples

- CHUV

- CHUV:BÂTIMENT\_CHAMBRE\_OU\_LIT
- CHUV:STRUCTURE RÉFÉRENCE

- CONTACT

- CONTACT:FAX
- CONTACT:TÉLÉPHONE

- DÉMOGRAPHIE

- DÉMOGRAPHIE:ÂGE
- DÉMOGRAPHIE:ÉTAT\_CIVIL
- DÉMOGRAPHIE:NATIONALITÉ
- DÉMOGRAPHIE:PROFESSION

- EMPLACEMENT

- EMPLACEMENT:CODE\_CANTON
- EMPLACEMENT:CODE\_POSTAL
- EMPLACEMENT:EMPLACEMENT\_GÉOGRAPHIQUE
- EMPLACEMENT:NUMÉRO\_HABITATION
- EMPLACEMENT:PAYS
- EMPLACEMENT:RUE

- NOM

- NOM:PATIENT\_E
- NOM:PERSONNEL\_MÉDICAL

- ORGANISATION

- PERSONNES
  - PERSONNES:LIEN\_DE\_PARENTÉ
- TEMPORAL
  - TEMPORAL:DATE
  - TEMPORAL:TEMPS

# Annotation Guidelines

## Personal Data Annotation of French Electronic Health Records

Version 2.0.0 / 08.02.2022

Author: Valentin Oliver Loftsson

[valentin.loftsson@chuv.ch](mailto:valentin.loftsson@chuv.ch)



# Purpose

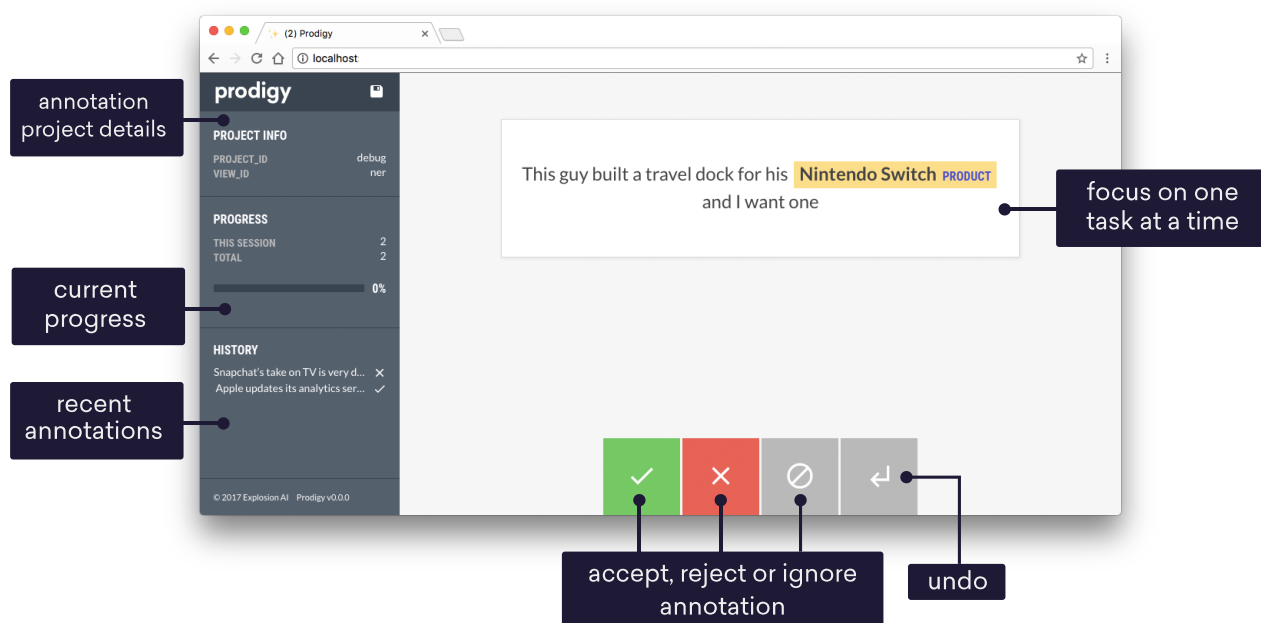
This document contains guidelines for annotating personal data in French electronic health records. An annotated corpus is an essential requirement for the development of methods to de-identify (a.k.a. anonymize) health records for privacy-preserving secondary usage, such as research, quality measurement and improvement, health and epidemiology. It should be noted here that these guidelines will *evolve* based on feedback and constant evaluation during annotation campaigns.

# Prodigy Annotation Tool

*Please read the below instructions carefully before you begin*



We will use the Prodigy annotation tool to accomplish our goal.

Open the [Prodigy demo](#) and try it!



Because Prodigy knows where words start and end, you don't have to select the exact letters – you just need to roughly hit the word to select it. To submit an annotation, click the green **ACCEPT** button or press the **A** key on your keyboard. If you don't know the answer or you are unsure about something, you can click **IGNORE** or press **SPACE**. Once you hit accept or ignore, the example will move into your outbox. You can go back through the annotations in your outbox by clicking the **UNDO** arrow, or pressing **BACKSPACE** or **DEL**. As you annotate, batches of examples will be cleared from your outbox and sent back to the Prodigy server to be saved in the database. You can also hit the save button in the top-left corner or press **CTRL+S** to clear your outbox

and manually save. It is especially important that you save before closing the browser tab. You are encouraged to save at a regular interval. Note that keyboard shortcuts are available for the categories and they are visible in the bottom right corner of each label in the green box.

This document is accessible in the Prodigy interface by clicking the  in the top left corner or with the  key on your keyboard.

# Examples

*Please read this section carefully before you begin*

⚠ To make it easier for statistical models to interpret dates and some other entities, we replaced `.` with `#` in most of the text. For example, the date `01.01.2021` will be presented as `01#01#2021` in the tasks. Therefore, the hashtag `#` should be interpreted as the period `.` in the below examples.

⚠ When you encounter incorrect word-splitting due to missing whitespace, you may annotate the resulting word. Examples:

- `M#Mustafa`
- `Tél:+41 78 333 22 11`
- `DrA. Chollet`
- `Dr J. Aschwanden/`
- `CH#DE BONMONT`
- `A#Ojanguren`
- `F#Abboretti`



## CHUV

Entities in this category are CHUV-specific

### CHUV : BÂTIMENT\_CHAMBRE\_OU\_LIT

Any building, floor of a building, room or bed number at CHUV

- Salle: BH 07 / 508
- Rendez-vous le 25#01#2019 à 14h au BH07
- Bloc opératoire dentiste CHUV -> BH11
- BH#11#632

### CHUV : STRUCTURE RÉFÉRENCE

Name or abbreviation of any unit (center, department, service, section, etc.) within CHUV

⚠ When in doubt, consult the [Ref-structure site](#)

⚠ Do not annotate the article la/le/l'

- son arrivée dans le service de cardiologie
- date de son transfert dans l'unité des Soins intensifs
- le patient quitte le service de chirurgie cardiaque
- Département femme-mère-enfant
- dans notre Centre des Maladies Cérébro-Vasculaires

⚠ Keep a watchful eye on abbreviations.

- NAT 18#02#20 - 21#02#20 --> SIP 21#02#20 - 06#03#20`

⚠ Terms like "service de" or "département de" before an abbreviation should be annotated too.

- Le patient susnommé a séjourné dans notre service de CHVH

## CONTACT

Entities in this category are contact information

### CONTACT : FAX

⚠ The word "Fax" should not be annotated

- Fax: +41 012 345 67 89

### CONTACT : TÉLÉPHONE

- (+354) 012 34 56
- 012#345#67#89

⚠ The word "Tél" or "Téléphone" should not be annotated

- Tél: +41 012 345 67 89
- Tél: (+41) 012#345#67#89

## DÉMOGRAPHIE

Entities in this category relate to demography

### DÉMOGRAPHIE : ÂGE

Any person's age.

⚠ The article **d'/de** should not be annotated

▼ Examples:

- patient de 63 ans
- patiente de treize ans
- enfant d'une semaine
- enfant d'une semaine et trois jours
- enfant de trois semaines
- enfant d'un an
- enfant de deux mois
- enfant d'un an et deux mois
- nourrisson de 40 jours de vie

⚠ Past or future ages of people should also be annotated

- pas de symptomes depuis l'age de 18 ans

⚠ Time of pregnancy or gestation period of an infant should **not** be annotated:

- Femme qui se présente à 40 SA 3/7
- Nouveau-né prématuré à 35 SA
- Patiente de 29 ans, 1G OP, transférée le 06/09/2019 à 32 6/7 SA

⚠ In the case when ages of multiple individuals is mentioned together, then please annotate them together as a whole if a key word like **ans** is only mentioned at the end. Examples:

### Correct

- Elle a deux demi-soeurs du côté paternel de **13 et 15 ans**
- Elle a deux demi-soeurs du côté paternel de **13 ans** et **15 ans**

### Incorrect

- Elle a deux demi-soeurs du côté paternel de **13** et **15 ans**

## DÉMOGRAPHIE : ÉTAT\_CIVIL

Any person's civil status.

### ▼ Examples:

- **célibataire**
- **divorcé**
- **divorcée**
- **marié**
- **mariée**
- **partenariat enregistré dissout**
- **séparé**
- **séparée**
- **veuf**
- **veuve**

## DÉMOGRAPHIE : NATIONALITÉ

Any person's nationality.

⚠ Should not be confused with **EMPLACEMENT : PAYS**. For example, "Elle habite en Suisse".

⚠ Sometimes, the name of a country is used when stating someone's nationality. In such cases the country should be annotated with **EMPLACEMENT : PAYS**.

### **DÉMOGRAPHIE : PROFESSION**

The profession of the patient, or that of a relative or any other person *excluding* medical staff.

## EMPLACEMENT

Entities in this category relate to locations

### EMPLACEMENT:CODE\_CANTON

Each canton of Switzerland has a 2 character identifier.

- Renens **VD**
- Fribourg **FR**

▼ Full list:

- **AG**
- **AI**
- **AR**
- **BE**
- **BL**
- **BS**
- **FR**
- **GE**
- **GL**
- **GR**
- **JU**
- **LU**
- **NE**
- **NW**
- **OW**
- **SG**
- **SH**
- **SO**

- SZ
- TG
- TI
- UR
- VD
- VS
- ZG
- ZH

⚠ Unlike canton names, canton codes should **not** be labeled as

**EMPLACEMENT : EMBLEMMENT\_GÉOGRAPHIQUE**

**EMPLACEMENT : CODE\_POSTAL**

Postal code in any country

- 1009 Lausanne
- 8003 Zürich
- 221 Hafnarfjörður, Iceland

**EMPLACEMENT : EMBLEMMENT\_GÉOGRAPHIQUE**

Names of geographical units smaller than a country fall under this category.

Examples: states, areas, cantons, districts, municipalities, cities, towns, villages, farms, ...

⚠ Postal codes do not fall under this category (see

**EMPLACEMENT : CODE\_POSTAL** )

⚠ Street names do not fall under this category (see **EMPLACEMENT : RUE** )

⚠ Please annotate geographical units **anywhere** in the world



⚠ When a geographical unit is used as a reference to an organization, it should be annotated as a geographical unit:

- Patient est suiv par les antalgistes à **Rennaz** ...

#### **EMPLACEMENT : NUMÉRO\_HABITATION**

The part of a street address that designates the house or building number.

- Avenue des Champs-Élysées **28b**

#### **EMPLACEMENT : PAYS**

Name of a country.

⚠ Should not be confused with **DÉMOGRAPHIE : NATIONALITÉ** For example "Elle est suisse".

⚠ Do not annotate the article **la/le/l'**

- Il est passé par l' **Algérie** puis le **Maroc** avant d'arriver en **Espagne** puis la **Suisse**
- patient de 28 ans, en **CH** depuis aout 2015
- patient de 28 ans, en **Suisse** depuis aout 2015

⚠ Sometimes, the name of a country is used when stating someone's nationality. In such cases the name of the country should be annotated with **EMPLACEMENT : PAYS**.

- patient de 28 ans, originaire de **Sierra Leone**

#### **EMPLACEMENT : RUE**

A street name *excluding* the house number

- **Avenue des Champs-Élysées** 28b

When you encounter incorrect word-splitting due to missing whitespace, you may annotate the resulting word:

- CH#DE BONMONT (CH.DE BONMONT)

## NOM

### Person names

⚠ All names, including first names, should be annotated

⚠ The honorific should not be annotated

⚠ Each name should be annotated individually *except* in the following cases:

1. Multiple names connected with a dash should be annotated as a whole
2. A particle preceding a name should be annotated with the name
3. Initials that abbreviate names should not be annotated on their own but annotated with an adjacent name.

## NOM: PATIENT\_E

### Name of a patient

- Madame **Pauline**

Each name should be annotated individually:

- Monsieur **Valentin** **Oliver** **Loftsson**

...except when a dash connects the names:

- Mme **Marie-Laure** **Christiansen**

Preceding particles should be labelled with the following name:

- **DEL BAGNO** **GIAN** **MARIO**
- **D'ETERNOD** **PAULINE**

Initials should not be labelled on their own:

- M. Valentin O. Loftsson

When you encounter incorrect word-splitting due to missing whitespace, you may annotate the resulting word:

- M#Mustafa (M.Mustafa)
- F#Abboretti (F.Abboretti)

## NOM: PERSONNEL\_MÉDICAL

Name of any medical staff (not only doctors)

- Dr. Beyrem
- Instrumentistes: M. Bayrem / Mme Xu

Each name should be annotated individually:

- Dresse He Xu

...except when a dash connects the names:

- Professeur Jean-Louis Raisaro

Preceding particles should be labelled with the following name:

- Dr d'Angelo
- Dre. de la Garma

Initials should not be labelled on their own:

- Pr. J.-L. Raisaro
- Dre Rochat N.

When you encounter incorrect word-splitting due to missing whitespace, you may annotate the resulting word:

- DrA. Chollet
- Dr#Mustafa (Dr. Mustafa)

- Dr J. Aschwanden/

## ORGANISATION

Name or abbreviation of any organization, including medical institutions or nursing homes

- Le patient travaille comme caissier chez Migros
- CHUV
- Centre hospitalier universitaire vaudois
- VIDY-MED
- Hôpital de Nyon

Another example is when an address of a patient is a retirement home:

- EMS LES DRIADES, 1400 YVERDON-LES-BAINS

When a geographical unit is used as a reference to an organization, it should be annotated as a geographical unit. See example in

EMPLACEMENT : EMBLEMATIQUE\_GÉOGRAPHIQUE.

## PERSONNES

Entities that have to do with people

### PERSONNES: LIEN\_DE\_PARENTÉ

Family relationship terms.

⚠ Corresponding plural terms should also be annotated

▼ Examples:

- belle-fille
- belle-mère
- beau-fils
- beau-père
- conjoint
- conjointe
- cousin
- cousine
- demi-frère
- demi-soeur
- demi-sœur
- enfant
- épouse
- époux
- femme
- fille
- fils
- frère
- garçon

- grand-mère
- grand-père
- maman
- mari
- neveu
- nièce
- mère
- oncle
- papa
- parent
- parente
- père
- soeur
- sœur
- tante

⚠ Since the terms **enfant**, **femme**, **fille**, and **garçon** can also mean woman, girl, and boy, respectively, it should be noted that these words must only be annotated when they refer to family (child, wife, daughter, son, resp.):

### Correct examples

- Le patient est venu avec sa femme.
- Monsieur Rodrigo a deux filles
- la patiente est venue avec son enfant
- ils ont 3 enfants

### Incorrect examples

- Enfant qui se présente avec...
- Le patient est une femme.
- garçon de 3 ans
- elle est une fille



## TEMPORAL

Entities in this category relate to time

### TEMPORAL : DATE

Any element of date or span of dates in a year, including dates, weekdays, months, years, holidays, special days or events.

⚠ Each date should be annotated individually

- 26#12#2021
- 5 fevrier au 4 mars
- en 2021
- En décembre, il a été opéré

*except* if the date is an interval within the same month, for example:

- 1-3 janvier
- 3-4#12#2020

⚠ Weekdays, special events or holidays also fall under this category.

- le lundi prochain
- le patient est venu le jour de Noël
- le patient est arrivé le jour de la fête nationale suisse

⚠ Weekdays should be annotated separately

- le mardi 6 avril
- lundi 22/03/2015

*except* if the weekday is followed by a single number referring to a date

- le jeudi 26 puis cranioplastie le vendredi 27#01

## TEMPORAL : TEMPS

### Time of the day

⚠ Only references to exact time of the day or exact time intervals should be annotated and not time durations that do not reference any time of the day

### Correct examples

- 9h30
- 11:00
- sept heures et demie

⚠ Each time should be labeled on its own:

- 13:00 - 15:00

⚠ Times that can be interpreted as approximate times (such as medication times) should still be annotated:

- Dafalgan 1g 1x/h [20:00]

### Incorrect examples

- après 48h
- il a attendu 2 heures
- l'opération a duré 20 heures