CRI : A Competent Reader Imitator for detecting binomial names in an historical corpus

Clément Morand, Olivier Ridoux

CRI: a Competent Reader Imitator for detecting binomial names in an historical corpus

Clément Morand* , Olivier Ridoux**
* *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique*
** *IRISA - University of Rennes*

1. Introduction

Many collections of historical documents are publicly available as scanned files in online databases (Ehrmann et al., 2021). These collections continue to be of interest, not necessarily for their *prima facie* content, which is often plainly obsolete, but for what they reveal about the thinking of each era and its evolution. This kind of inquiry belongs to the digital humanities domain and requires specific tools to be effective (Burdick et al., 2012).

The French popular science magazine *La Nature* (Tissandier, 1873 1962) (abbreviated here as LN) is a particularly rich source for this kind of inquiry on account of its longevity (about 90 years) and its broad spectrum of domains. In particular, its publication period covers many technical and scientific revolutions (Kuhn, 1962) that helped shape the modern world (Smil, 2021). For instance, it is true for Life Science with the developments of genetics, the "Modern synthesis of evolution theory", plate tectonic, etc., but it is also true in physics, chemistry, astronomy and technology. Unfortunately, exploring the archive of a magazine such as LN is a complex task in that the indexing, in the case it exists, has been performed for the primary intentions of contemporary readers rather than for indirect analyses performed *a posteriori* by future researchers.

In the case of LN, examples of such analyses might try to answer questions such as "What was the reception of Marie Curie in the public?", "Given a selection of articles of interest, find similar articles", or "Was the possible climatic effect of burning fossil fuel presented to the public?". With respect to the last question, direct analyses would have trouble identifying Svante Arrhenius's work on this topic, which was published in LN in the early 1890s. Indeed, Arrhenius and his contem-

poraries did not use the term carbon dioxide, they called it carbonic acid. A direct query like "carbon dioxide" would fail to find the answer.

To deal with such difficulties and to answer such a variety of questions, we propose to use a multi-faceted querying approach (Sacco & Tzitzikas, 2009) which favors indirect querying. The center of this approach is to use an *article* $\times$ *attributes* matrix, where the attributes are computed by various processes, which we refer to collectively as *semantic annotation*. Based on semantic annotation, corpus exploration could be done by *association of ideas*, following attribute patterns. Semantic annotation includes lexical analysis of words in titles or figure legends, statistical analysis of word distribution, and detection of *named-entities*.

Classical categories of named entities like people and locations provide valuable attributes. In the context of *La Nature*, less frequently studied entities like biological species in the Linnean taxonomy are of particular interest, as the use of scientific names is a distinctive trait of scientific literature, including popular science magazines. Scientific names for biological species are called *binomial names*.

Binomial names are a named entity category that is much less frequently studied than those of persons, places, and institutions. Although this category has already received some attention in the domains of microbiology (Nédellec et al., 2006), biodiversity studies (Mozzherin et al., 2017), or biomedical studies (Akella et al., 2012), the situation of a corpus like *La Nature* differs significantly from previous studies in several aspects.

— The longevity of LN's publication period implies *diachronicity*, i.e., that the corpus does not follow the same linguistic norms from beginning to end. Diachronic corpora have been shown to impede the recognition of named entities (Ehrmann et al., 2021). Previous works on recognizing binomial names focus on publications from short time periods, whereas the publication of LN spans almost a century. The Linnean taxonomy changed a lot during this period and was revolutionized by modern evolution synthesis that mixes Darwinian evolution and genetics.

— LN's broad spectrum of domains implies *heterogeneity*. Heterogeneous corpora have also been shown to be a challenge for named entity recognition (Nadeau & Sekine, 2007, Ehrmann et al., 2021). Whereas previous works engaged in homogeneous corpora, e.g., a collection of microbiology papers, LN is multi-disciplinary, and naturalist articles comprise only a few percent of the total. No single domain in LN is more than a few percent of the total.

— The LN corpus is only accessible as low-resolution scanned images. These images must go through an OCR process (*Optical Character Recognition*) before further treatment, and this introduces *noise* in the data. Whereas previous works exploit born-digital sources, the LN corpus is only accessible as a noisy source.

The purpose of this paper is to study how binomial names could be incorporated into the semantic annotation of the diachronic, heterogeneous, and noisy LN corpus. Figure 1 illustrates the task and its context. To cope with the LN complexity, we have developed a model that we call the *Competent Reader*, which represents the ability to recognize positive occurrences despite the fact that they are obsolete, ill-formed, or defaced by noise. This model is used for manually creating a gold standard for evaluating automatic annotations. We show that approaches proposed for similar tasks but different contexts do not work well for this gold standard. Thus, in this paper, we develop our approach, which we call the *Competent Reader Imitator* (CRI), that involves combining a rule-based approach with a frequency argument. We show that this innovative method is robust to numerous variations and consistently achieves an F-measure of about 70% despite diachronicity, heterogeneity, and noise, which are all known to impede named entity recognition. We present evidence that the task, the proposed method to solve it, and the evaluation methodology are all tied together. These findings support the idea that different methods cannot be compared in the absolute; they can only be compared with respect to a task in context, and the characteristics of the corpus are part of the context.

Additionally, we strive for inclusive and frugal computing placed in a context of *digital sufficiency*. This includes choosing less computing-intensive approaches and aiming for just good enough results (Gabrys et al., 2016, Becker, 2023, Santarius et al., 2022).

Section 2 presents the Linnean classification, the LN corpus we are working with, and some difficulties this conjunction induces. Section 3 presents the Competent Reader Hypothesis (CRH), on which the manual annotation strategy for evaluation is based. This section also details the task and the annotation process. Section 4 presents the state of the art and the performance of available existing methods. Sections 5 and 6 present our proposal for coping with the task and the evaluation of this proposal. Finally, Section 7 discusses our results and presents future work.

2. Context information

We present the elements that structure the work presented in this article, namely, the Linnean classification and the LN corpus.

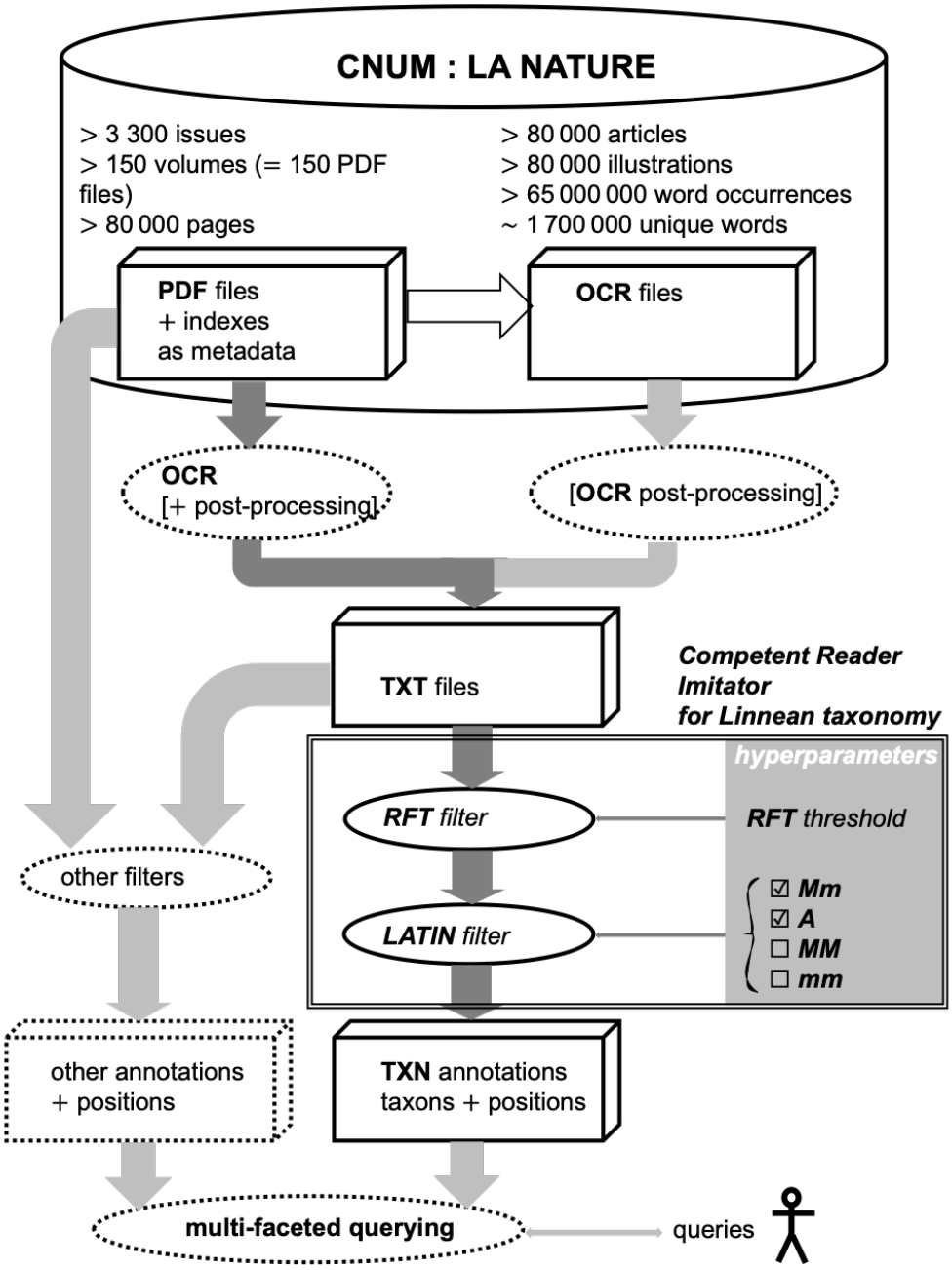

Figure 1 – The CRI process in its context. Details are clarified progressively in the text. Dotted boxes and light grey arrows are for context only.

2.1. The Linnean classification

The Linnean classification of living organisms is organized as a hierarchy in which each level has a name, starting with *domains* under the root (two to six according to taxonomists), then *kingdoms* under domains (five to ten, e.g., animals, plants and fungi), down to *genera* (a few hundreds of thousands) and *species* (about 15% already named species from an estimated ten million total yet to be discovered). The first attempts at such a classification date from antiquity, but Carl von Linné is recognized for giving it its modern form. Consequently, this classification is often called Linnean classification or Linnean taxonomy. Genus-species pairs are called *binomial names*, and the species part is called the *specific epithet* or *specific name*.

At each level of the hierarchy, classification is done after observable criteria, but as observations accumulate and change in nature (surface-level observation, microscope, microbiology, genetics), classification adapts to new discoveries. In fact, this adaptation is frequent, and our experiments (see 4.3) show that less than 40% of all binomial names found in *La Nature* are still in use today.

A well-known example of a binomial name is *Homo sapiens* for the human species. Besides being unique inside a kingdom, binomial names follow syntactic and typographical rules that have been consolidated over time. A simplified summary of these rules is as follows (ordered from very superficial traits to deeper linguistic ones):

— The names of the Linnean classification should always be written in italics: e.g., *Cupressus* vs. "cupressus". This is, in fact, an application of the more general rule that foreign expressions must be written in italics. Since binomial names are reputedly Latin, they must be composed in italics.

— Binomial names must be written in the Roman alphabet, without accents nor ligatures, even if names are derived from original words with accents or ligatures, e.g., *Chamaerops* should not be written *Chamærops*.

— They are *binoms* (i.e., consist of two parts). The first part designates the genus, and it is capitalized (e.g., *Homo*). The second part designates the species in the genus. It is written with a lower-case initial letter (e.g., *sapiens*), even if it is derived from a proper name (e.g., *Trachycarpus fortunei*, in honor of the botanist Robert Fortune). Names of different species can share the same specific name provided the genus names are different.

— Binomial names can be abbreviated by replacing the genus name with its initial followed by a dot. For instance, the abbreviated version of *Homo sapiens* is *H. sapiens*. In principle, abbreviated forms must be preceded by a fully spelled out form,. It is common that a fully spelled binomial name

introduces a genus name, and is followed by abbreviated names for the same genus but different species.

— The name of the discoverer, possibly abbreviated, can be added after the species name. For example, *Bombyx neustria Linn* is composed of the genus name *Bombyx*, the species name *neustria* followed by an abbreviation of the name of the discoverer, in this case Carl von Linnée.

— Binomial names must be written in "Latin" with agreement in gender (masculine, feminine, and neuter in Latin) of the species name with the genus name. However, this is Latin sounding and not actual Latin words, e.g., *Feijoa sellowiana* is an evergreen shrub named after two renowned botanists of the early 19th century: João da Silva Feijó and Friedrich Sellow.

— Grammatically, genus-species binomials can be declined into nominative-adjective (e.g., *Helleborus niger*, black hellebore or Christmas rose), nominative-nominative (e.g., *Panthera leo*, the lion) and nominative-genitive (e.g., *Trachycarpus fortunei*, Robert Fortune's trachycarpus).

The reality, however, is much more complex. For example, many genera and species are divided into sub-genera and sub-species. Different variants of the rules apply in botany, zoology, virology, or for crops, and they are regularly updated. They have been formalized in international codes. For instance, the first code for botanics dates from 1867 with the latest update in 2018 (Turland et al., 2018).

An additional difficulty is that non-specialists tend to relax these rules. In particular, species names derived from proper names (e.g., name of discoverer or name of place of discovery) are often capitalized in practice, even if it is ruled out in modern codes. Similarly, ligatures are often used against current codes.

Localizing binomial names in a corpus should take into consideration the above rules, and also the deviation observed in the corpus studied.

2.2. The corpus of La Nature

This section presents the structure of the LN corpus and its volumetry. This corresponds to the top of Figure 1.

2.2.1. *Structure of the corpus*

La Nature is a popular science magazine founded by Gaston Tissandier and published from 1873 to 1962 (Vautrin, 2018). The topics dealt with in the magazine were mainly in the fields of natural sciences (geology, botany, meteorology, etc.), humanities and social sciences (ethnology, medicine, hygiene, etc.), or technology (scientific apparatus, energy, mechanics, photography, transportation, etc.). Ab-

stracts of communications published in scientific academies or from other journals were also included. During the publication period of *La Nature*, there were major developments in several fields of modern science and technology, as well as in the political and social history of the world such as colonization, industrialization, the two World Wars, decolonization, the phylloxera crisis in the European vineyards, etc.

LN publication can be analyzed in two periods: pre-WW II and post-WW II. LN was a weekly magazine during the first period and then became a monthly. In fact, LN never fully recovered from the war, and during the second period, LN experienced different formats. The general rules that follow apply better to the first period. Each issue of LN was composed of two folios: an outer one and an inner one. The outer one was for "consumable" content and advertisements, which were renumbered for each issue. Recurrent topics of the outer folio are meteorologic tables and astronomic observations. The inner folio was for archival content and its numbering ran on a semester basis. The inner folios were assembled and sold as separate *volumes* every semester. For each volume, several indexes were compiled (authors, topics, and articles) and added after the inner folios. Moreover, parts of the outer folios were often added after the indexes, e.g., astronomical ephemerides. In fact, there are fewer volumes than semesters because of troubled periods that reduced the publishing rate (war and post-war periods and the beginning and end of the life of the magazine). These volumes form the raw material of this work.

Figure 2 shows the inner folio of the June 2nd, 1923 issue. Even though the actual formatting of LN has changed a lot over the years we will use this folio as a model for the organisation of LN. In particular, the figure is not meant to show the contents but rather to show the organization. This inner folio appears on pages 337-352 of volume 104.

Each volume is a sequence of articles in which the boundaries of the original issues are hardly visible. There are two kinds of articles: *long* and *short*. Long articles span several pages, and short articles span only several lines. Short articles are often assembled into longer articles with titles such as "Chroniques de l'Académie des Sciences" (*Chronicles of Science Academy*). However, the editor's index added at the end of every volume shows indistinguishably all articles, long and short. Moreover, articles in this sense (long and short) are the largest chunks of semantically homogeneous text. Consequently, we consider the volumes as collections of long and short articles, and we set our goal to annotate articles.

Figure 2 – The 16 pages of the inner folio of the June 2nd, 1923 issue

For instance, the June 2nd issue of 1923 contains the following articles (our composition of the French titles reflects the original composition, including capitalization and italics):

— "LE SAHARA", 337-341: a long article on the geography and climate of the Sahara desert. It contains five pictures and a map. The map raises several difficulties. First, the OCR process tries to parse the labels, and second, several labels are obsolete, e.g., *Tripolitaine* for Lybia.

— "COMMENT ON FABRIQUE UNE LAMPE DE T. S. F.", 341-347: a long article on the making of vacuum valves. It contains ten pictures.

— "L'UTILITÉ DE LA FOURMI DES BOIS (Formica rufa L.)", 347-348: a plea in favor of the red wood ant (*Formica rufa L.*) as a formidable insect killer. It contains a table.

— "ACADÉMIE DES SCIENCES", 348: a report on the April 1923 session of Académie des Sciences. It contains five short items, each of which has a title, and is referenced in the index of the volume as an article.

  — "*La dénaturation de l'alcool éthylique*", 8 lines: on the denaturation of ethyl alcohol. Note several difficulties here. Ethyl alcohol is now better known as ethanol, and denaturation is often known as methylation after the use of methanol as a denaturant.

  — "*Le tremblement de terre du Kansou*", 12 lines: on the evaluation of the intensity of a severe earthquake in China. Note that Kansou is the former French spelling of the northwestern region of China that is now called Ganzu, both in French and in English. A long article on this earthquake was published at the beginning of the same year, and this short article only reports on the difficulty of measuring its intensity.

  — "*Un nouveau minerai radioactif*", 8 lines: on a newly discovered radioactive mineral. The article refers to a Russian location called Sludjanka which is now spelled Slioudianka in French and Slyudyanka in English. Over the years, the same title has come back several times.

  — "*La variation des parfums sous l'influence du greffage*", 12 lines: on the effect of grafting on plant scents. Note that the article refers to *Artemisis absinthium* (common wormwood), which is now called *Artemisia absinthium*.

  — "*La préparation du vin par fermentation continue*", 18 lines: on a new process for wine-making.

  Two of these small articles are not ascribed to the right page by the editor's index.

— "LA VISION DES OISEAUX [*Suite et fin*]", 348-350: on bird vision. This is the last of a series of six long articles. Despite its natural science title, it is an ophthalmology article, and it contains no taxonomic reference.
— "LA MACHINE À CALCULER ADDIATOR", 351-352: on a slide calculator which was popular from 1889 to 1968. It contains two pictures and a diagram.

2.2.2. *Volumetry of the corpus*

The LN archive consists of 157 volumes and more than 80,000 pages. This amounts to about 600-700 pages per volume. This average is a fair estimation of volume size. Indeed, the maximum is about 800 pages, and only a few volumes are significantly shorter than 500 pages, e.g., during wartime or at the beginning and end of the life of LN. The ocerized archive contains more than 65,000,000 words for 1,760,000 unique words. For instance, the June 2nd, 1923, issue of LN contains about 11,600 words, 3,000 unique words, among which 1,900 have only one occurrence. Some of these are OCR artifacts, but this is not the only explanation. Indeed, articles in LN, especially short ones, often mention named entities only once.

Scanned files of the archive are available (CNUM, ca 2000), but they have a very low resolution (12 pixels per character height, x-height) and are structured as a flat sequence of pages. In other words, they do not explicitly show articles or weekly issues. Each original page has been scanned as a PNG image, and each archival volume is represented by a PDF file whose pages are the images. Every PDF file also contains a detailed index in its metadata, which is a transcript of the indexes added at the end of the printed volumes. A source of difficulty is that there are errors in the scans, e.g., missing pages, and in the index, e.g., faulty page numbers.

The metadata also contains the domain of each article, as indicated in the table of contents for the printed volume. However, these classification attempts are often confusing, with compound classes like "Physiology and zoology" that seem awkward in modern days (note that "Physiology and botanics" does not exist in the index) and perplexing subtleties like "Medicine and hygiene" and "Hygiene and health".

The total number of articles is of the same order as the total number of pages, which amounts to an average length of one page per article, i.e., 80,000 articles. However, the average length of articles is a poor summary of the structure of LN because this length follows a bimodal distribution. It is much more precise to say that about half of the articles are short (several lines long), and the other half are

long (several pages long). For instance, the June 2nd, 1923 issue of LN contains 5 short articles and 5 long ones for 16 pages.

LN is strongly multi-disciplinary, and it is not limited to the so-called hard sciences. As a consequence, any domain is a minority. In volumes that we manually annotated, there is only one binomial name occurrence every 3.5 pages. Again, this is an inadequate statistic because binomial name occurrences arrive in bursts. Many articles contain no occurrence of binomial names, but a few contain many.

It is also worth noting the more than 80,000 illustrations, about one illustration per page, again following a bimodal distribution. For instance, the June 2nd, 1923 issue of LN contains 19 illustrations for 16 pages, and these illustrations are concentrated on 3 articles out of 10. This remarkable number of illustrations was a hallmark of LN at its creation and an indicator of progress in printing technology. The high density of illustrations and their burst occurrences also have an impact on the OCR process because it makes the composition of pages less predictable, and the OCR process tries to read texts in illustrations.

As scanned pages go through an OCR process, this introduces noise. An ocerization is available with the LN archive. It shows a high error rate, and we tried to improve things by performing our own ocerization, using Tesseract. In both cases, we estimate the rate of OCR errors to be 10% at the word level. This error rate is even worse if one considers the long distant reading order, i.e., page parsing. In fact, the long distant reading order is seldom properly reconstituted. This does not originate with the ocerizer but rather is the result of the complexity of the original document: a composition in two columns, interrupted by many illustrations, pages tainted with library marks, pencil annotations, shades, and copyright marks by the CNUM (operator of the scan process) and the low resolution of the scan.

Ocerization leads to two main difficulties. Firstly, OCR ignores the typography of the source text, especially the italics, which are the first clue for recognizing binomial names. Secondly, OCR introduces a lot of spelling errors: e.g., words cut off by a space, 't' turned into 'l', 'm' into 'rn' or vice-versa, nonexistent ligatures being added, etc. For instance, "Rhinocerus" can be ocerized as "Ahinocerus" or "Amanita" as "Amanila". It even happens that the incorrect forms are more frequent in the corpus than the correct ones. For instance, in the June 2nd, 1923 issue, the specific name *rufa* (the red wood ant) appears only once under its proper form and twice under the form *rvfa*. It also appears once under each of the forms *ru fa* and *ru fan*. We have decided to consider OCR and post-OCR correction as a distinct task. To avoid mixing issues in this article, we consider only raw post-OCR text files (see

the OCR stage in Figure 1). Any improvement in the OCR chain will improve the global semantic annotation task.

Moreover, the composition and typography of LN vary a lot with time. It is almost always a two-column based composition (except for tables of contents, which are three-column), but with very different strategies for marking article limits: e.g., titles in thin capitals on one column width, or titles in capitals on two columns width, or titles in bold, etc. The strategy for composing illustrations in text is also very unstable. As a result, and considering the errors in the metadata, the delimitation of articles is a more difficult task than might be expected. In this paper, the delimitation of articles is considered to belong to another phase and will not be covered, but ongoing work suggests that this task has an error rate of several percent. As for errors in the OCR process, any improvement in the delimitation of articles will highly improve the whole semantic annotation process.

The two elements developed in this section, namely the Linnean classification and the LN corpus, show the specifics of the task: rules that change and are relaxed for the Linnean classification, noise, diachronicity, and heterogeneity for the corpus. This must be taken into account to define what is a true positive occurrence.

3. Creation of a gold standard

We present in this section the hypotheses and the methods that have been used to create a *gold standard* for evaluating different classifiers. First, a true positive occurrence is defined by the *Competent Reader Hypothesis* (CRH), which we define shortly. Then, this hypothesis is used to specify the task, that is, to imitate the Competent Reader. Finally, it is used to define the manual annotation process for building a gold standard.

3.1. The Competent Reader Hypothesis - CRH

Definition 1 (Competent Reader) *In the context of this research, we call a Competent Reader a person who knows the linguistic codes of a popular science magazine. They are not experts, but they recognize place names even if they do not know the place, and they recognize chemical names, even at first encounter. They also recognize taxonomic binomials, even if only by their textual aspect.*

In particular, we assume the Competent Reader is knowledgeable enough to differentiate the everyday use of a Latin genus name from its scientific use. For instance, "cupressus" (from genus *Cupressus* for cypress) is often used in French

instead of the proper French vernacular name "cyprès". A Competent Reader can thus distinguish the formal use of *Cupressus* from the vernacular usage of "cupressus". Similarly, the vernacular term "géranium" often designates the species of genus *Pelargonium*, even though *Geranium* also exists as a genus. It is also generally the case in paleontology texts: *Diplodocus* vs "diplodocus'. Recall that the OCR process eliminates italics so that an important clue disappears. That only leaves capitalization as a clue, but OCR often mistakes a 'C' for a 'c'.

Definition 2 (Competent Reader Hypothesis) *The Competent Reader Hypothesis (CRH) is the assumption that the objective of a classifier is to imitate the behavior of the Competent Reader. In the case of recognizing binomial names, it is the ability to recognize the intention of ascribing to a living being a position in the taxonomic classification, even by using obsolete, or ill-formed binomial names.*

The name CRH is inspired by the *Competent Programmer Hypothesis* introduced for error detection in programs (DeMillo et al., 1978). Competent Programmers may incorrectly apply programming rules, but they know the rules. Similarly, Competent Readers may incorrectly interpret binomial names, but they can identify them.

Thus, according to the CRH, our goal is to design an automatic classifier with similar performances to the Competent Reader, and the evaluation of that classifier using a gold standard that reflects the behavior of the Competent Reader. Therefore, OCR errors, deprecated classifications, and incorrect practices with respect to nomenclature codes (modern or not) should be accepted.

3.2. The task

In theory, named entities are referred to by *rigid designators* of some reality (in a very broad sense, including myths and fiction). However, binomial names do not fit this definition because they are not rigid, and even the reality they should refer to is not rigid. Indeed, life classification is always in progress, and the very notion of genus and species is debated. So, we will not try to find to which entity a binomial name refers. Instead, we will consider that binomial names refer to themselves.

Consequently, the expected output of this approach to *Named Entities Recognition* (NER) is the *localization* of entity names rather than the *identification* of named entities. Henceforth, this will be referred to as *Entity Name Recognition*, ENR, to highlight the difference. This is important to notice because it dictates what the gold standard should be and the objectives of the evaluation phase. Localization

found by ENR can then be passed to a downstream application that performs multi-faceted querying (see bottom of Figure 1). This does not forbid trying to identify the referred species; it only means it is another task.

Localizing binomial names amounts to designing an algorithm that acts as a *classifier*, i.e., a program that parses a text and determines for each occurrence (positions) whether or not it is *positive*, i.e., effectively contains a binomial name. Localization in *La Nature* can be made at different levels: the character position, the page, the article, the weekly issue, and the volume. Our main target is the article, but we have evaluated all classifiers at the character level within articles to facilitate comparison with current practices. This is the baseline; any coarser localization can only yield better results. Moreover, localization at the character level can be used to derive all other localization levels.

Note that localizing occurrences of binomial names is only a facet of a broader task: to summarize an article as a set of attributes that could be used by a multi-faceted querying application. To this end, other classes of named entities must be recognized, keywords are extracted from titles and articles, etc. The balance between all these operations is not the topic of this article, but the main idea is that every facet compensates for the weakness of the others. For instance, a weakness in detecting binomial names could be compensated by the computation of keywords using TF-IDF (Jurafsky & Martin, 2009).

As a classifier is never perfect, its performances are evaluated against expected results. A classifier signals occurrences as positive (P) or negative (N) according to whether or not they match the class of interest. The evaluation against a ground truth qualifies positive and negative occurrences as true (TP and TN) or false (FP or FN) according to whether or not the classification is correct. The proportion of true and false positive and negative occurrences is the basis of classical performance indicators (Jurafsky & Martin, 2009) that are called *precision* ($TP/(TP+FP)$) and *recall* ($TP/(TP+FN)$), and their harmonic mean, the *F-measure* ($2\times(precision\times recall)/(precision+recall)$, the inverse of the arithmetic mean of the inverses of precision and recall). Another classical performance indicator is *accuracy* ($(TP+TN)/(TP+TN+FP+FN)$, the ratio of valid decisions on all decisions) but we will not use it because the distribution of binomial names is highly biased (see Section 2.2.2). Indeed, to never recognize binomial names (hence, accuracy = $TN/(TN+FN)$) would be extremely accurate because they are very rare in the corpus (hence, accuracy = $TN/(TN+\varepsilon)$), but given our task, it would be absurd to do so.

Since F-measure is a function of precision and recall, it could be left implicit. However, it is an interesting measure on its own. It even shows behaviors that cannot be read easily in precision and recall alone. For instance, we will show later that certain decisions affect precision and recall, but not the F-measure.

In the following, the ground truth is represented by a gold standard composed of manually annotated volumes.

3.3. The manual annotation

For the sake of evaluating different classifiers, we manually annotated four volumes (volume 12, 1879, first semester, 432 pages; volume 83, 1912, second semester, 671 pages; volume 126, 1934, first semester, 610 pages; and volume 155, 1960, whole year, 558 pages) of LN. We chose the four volumes at about a 30-year distance from each other to cover all the publication periods of LN and to cope with variations in style, taxonomy, topics, etc.

Following the CRH, the annotation instruction was to signal all binomial names, including those that do not strictly follow the rules of binomial nomenclature, but that a Competent Reader would recognize as such. For instance, genus-species pairs where the species is capitalized are recognized as binomial names. Note that since facsimiles are available, the annotator can see the italics that the classifier will never see. They do not suffer from OCR errors (the Competent Reader's eyes perform the OCR). This helps solve ambiguities.

Note that following the CRH implies that some occurrences must be annotated as positive, even though it is evident from the start that they will be nearly impossible to recognize. If a classifier is based on a dictionary, any OCR error in a binomial name will make it unrecognizable. For instance, *Homo sapieus* (with a *u* for an *n*, a frequent OCR error) does not belong to any dictionary. Similarly, if a classifier is based on Latin declensions, most OCR errors in the last letters of a binomial name will make it unrecognizable. For instance, *Phorminm tenax* (with a *n* for an *u*, also a frequent OCR error) does not match any Latin declensions.

The two authors performed the manual annotation over a few weeks, one reading mainly the ocerized text and the other the facsimile. Initially, assuming the CRH was not intuitive because the annotators were always tempted to refuse non-conform occurrences, but the annotators were eventually habituated to it. The CRH provided the basis for a manual annotations consensus. The manual annotation discovered 671 positive occurrences in the four test volumes. This amounts to about one positive occurrence every 3.5 pages and one positive occurrence every 2742 words, or

0.036%. As for other statistics, these average values are poor indicators because occurrences of binomial names appear in bursts.

This is *ex-ante* manual annotations. We also performed *ex-post* manual annotations to evaluate the classifiers at the corpus level. The principle is as follows: in validation mode, a classifier builds a *concordancer* (Barrière, 2016) that contains all positive occurrences in their context (30 characters before the occurrence and 30 after). A sample is randomly selected and then annotated manually in a few hours. By definition, these sampled occurrences are positive, but some are true sampled positives ($\widehat{TP}$), and others are false sampled positives ($\widehat{FP}$), and this is enough to compute sampled precision ($\widehat{TP}/(\widehat{TP}+\widehat{FP})$) over the whole corpus at a very low cost because binomial names are very rare (see Section 2.2.2).

It is to be noted that for this work, the automatic delimitation of articles had not yet been done on the LN archive. To save time and manual operations, we performed an ad-hoc delimitation that precisely delimited only the articles that contained at least one binomial name. The parts of the text between the identified articles are considered single virtual articles even if they concatenate several actual articles. This approximation can only lower precision since the task considers the identification of the positions of binomial names in each article. Moreover, *ex-post* evaluation shows that this approximation has no visible effect (see Section 6.2).

The CRH has helped us define the task and the gold standard for evaluating candidate classifiers. Candidates classifiers are (executable) classifiers described in the literature that are compatible with the LN corpus. We also consider an *adhoc* classifier that we developed in order to measure the taxonomic drift.

4. State of the art and baseline

This section presents the state of the art in named-entities recognition (NER) applied to the class of binomial names. We identify three available classifiers for comparison; two are existing classifiers, and one is a new classifier for measuring the taxonomic drift.

4.1. State of the art

Named entities of the binomial name class have been studied for applications in the field of biodiversity (Koning et al., 2005, Sautter et al., 2006, Little, 2020), in the biomedical domain (Akella et al., 2012, Pafilis et al., 2013) and microbiology (Nédellec et al., 2006). However, it has never been studied within a historical corpus with as much thematic diversity as in LN. Quite the opposite, nu-

merous works in the domain have an extremely specific finality, such as the study of biodiversity in the Philippines (Nguyen et al., 2019) or the medicinal herbs in Maghreb (Seideh et al., 2016). These studies define an intricate network of very specific tasks, such as the identification and geographic or temporal localization of the discovery of a specimen or the recognition of a binomial name and its medicinal role.

Nowadays, most state-of-the-art Natural Language Processing methods, in general, and NER methods, in particular, use a deep learning approach (Nasar et al., 2021, Ehrmann et al., 2021). This commonly includes fine-tuning a pre-trained *(Large) Language Model* (LLM), typically a Transformer, for the specific task to be solved. Commonly used models include *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al., 2019) or its specialized variants such as BioBERT (Lee et al., 2019) in the biomedical domain or CamemBERT (Martin et al., 2020) or FlauBERT (Le et al., 2020) for the French language. Examples of works based on such a method for NER in historical data include (Yu & Wang, 2020, Labusch et al., 2019).

However, (Cunha et al., 2021) shows that symbolic approaches can outperform neural-based approaches in cases of scarce data. In our case, there are only 671 positive examples in the more than 2,200 pages of our gold standard corpus. The observations on the LN corpus (few positive examples, a wide variety of binomial name forms, and the noise in the data) place our task in a context of scarce data. Thus, following (Cunha et al., 2021), we have chosen to investigate symbolic approaches only. Furthermore, even if the data to look for were not scarce, numerous environmental and ethical concerns about *Deep Learning* (DL) and LLMs make us prefer symbolic approaches.

(Strubell et al., 2019) first raised concerns about the carbon footprint of training DL models, and (Sevilla et al., 2022, Thompson et al., 2023, Schwartz et al., 2020) show that there have been structural tendencies towards more computationally expensive models. Additionally, there are significant life cycle impacts resulting from the hardware needed to run such computations (Gupta et al., 2020, Luccioni et al., 2023). Apart from the environmental concerns raised by the increasing use of DL, there are also ethical questions related to LLMs and DL in general (Bender et al., 2021), including academic ethics (Schwartz et al., 2020, Birhane et al., 2022). Academic ethics concerns also stem from the concentration of power in the hands of *Big Tech* companies that are the only ones capable of creating and maintaining the latest and biggest models (Abdalla & Abdalla, 2021, Birhane et al., 2022, Abdalla et al., 2023).

We place our work in a context where there are growing expectations by civil society on the social benefits of science for everyone. Taking this context into consideration requires thinking about the accessibility of our work and its potential harmful and positive impacts (COMETS, Ethics Committee of the CNRS, 2022). Beyond our work on LN, we hope this paper provides a set of tools and methods that are easily understandable, frugal, and usable for the general public. And, acknowledging (Schwartz et al., 2020) call for a more inclusive AI, and in a similar approach to (Castellan et al., 2023), we will therefore investigate more symbolic, or *rule-based*, approaches such as the use of dictionary, trigger words or pattern matching.

Striving for digital sufficiency first asks to check if there does not already exist a good enough tool. Thus, we present the performances of three tools that represent various approaches that could be followed. Two are tools from the literature, LINNAEUS (Gerner et al., 2010) and QUAESITOR (Little, 2020), that can readily be used to process our data. A third one, the TAXREF classifier, is an ad hoc classifier that allows us to explore various hypotheses. These tools will be presented in the order LINNAEUS, TAXREF classifier, and QUAESITOR.

4.2. LINNAEUS

LINNAEUS (Gerner et al., 2010) is a NER system that aims at finding all mentions of binomial names and vernacular names in scientific publications (e.g., *Quercux ilex* and evergreen oak, holly oak or holm oak in English, and *chêne vert* or *yeuse* in French). It uses a dictionary based on the NCBI taxinomy (NCBI, 2008). It contains binomial and vernacular names of more than 380,000 species from public nucleotide and protein sequence databases. LINNAEUS performance has been tested on a large micro-biology corpus (10 millions abstracts and 100,000 articles) that contains a very high density of binomial name occurrences (3 binomial names per abstract and 40 per article). This corpus is highly biased, with the human species accounting for 47 % of all occurrences, and the top ten species accounting for more than 71 % of the total number of occurrences. This means that simply searching for these 10 species names in this corpus achieves a 100 % precision and 71 % recall. According to the authors, LINNAEUS achieves more than 90 % precision and recall.

We tested LINNAEUS on our manually annotated volumes (see 3.3). It obtains really poor results with a precision of 16.05%, a recall of 19.83%, and an F-measure of 17.74%. This shows that the task of LINNAEUS is very different from our task.

The challenges induced by OCR noise and the diachronicity of our corpus made us hypothesize that no dictionary-based approach could be satisfying. The experiment with LINNAEUS is the first confirmation of this intuition. We found a second confirmation by using a more natural-history-oriented taxonomy: TAXREF, the taxonomy referential edited by the French "*Museum Histoire naturelle*" (Gargominy et al., 2021).

4.3. TAXREF classifier

The TAXREF thesaurus describes a large part of the natural classification hierarchy, from kingdoms to species plus vernacular names when it makes sense, of about 535,000 different species. In fact, TAXREF contains 733,000 entries, but about 200,000 describe the upper levels of the classification hierarchy, and a few entries only differ by vernacular names, which our task ignores.

A first experiment is to use the TAXREF thesaurus as a classifier: to be or not to be in the thesaurus. Unsurprisingly, precision is 100%, and recall has also improved over LINNAEUS, 33.8%, which yields an F-measure of 50.6%. The improved recall probably comes from using a more natural-history-oriented thesaurus.

This experiment shows that more than 60% of the binomial names found in our gold standard are not in a modern thesaurus. Several factors can explain this:

1. real changes in the Linnean classification,
2. OCR noise,
3. relaxed usages that do not obey taxonomy codes (capitalization of species names, accents, ligatures, spelling, etc.),
4. TAXREF is simply incomplete.

We examine points 4 and 2 in this order. Point 3 will be examined later in Section 5.3.

The missing element in a thesaurus like TAXREF is that abbreviations are not in the thesaurus. However, they are completely legal and of rather frequent use. We refined the TAXREF classifier so that it also recognizes abbreviated forms of its binomial names. This improves recall, 37.1%, and the F-measure, 54.1%.

Note that it could have deteriorated precision if an abbreviated form had coincided with a "M. Lastname" form ("M." is the abbreviation for Mister in French). For instance, *Malva Ludwigii* abbreviates in *M. Ludwigii*, which makes a plausible proper name. The 100% precision shows that this has not happened in the annotated volumes, but it does not exclude that it may occur elsewhere, and we will see in the following that similar kinds of coincidence really happen.

Estimating the effect of OCR noise is not easy, especially if one wants to avoid an enumeration of all OCR errors. We tried approaching this question by refining the TAXREF classifier with an edition distance. In this version, the classifier does not look for a perfect match with binomial names in the thesaurus, but it allows for a match modulo edition distance. In this experiment, we used the Python `regex` implementation of matching regular expressions modulo a *Levenshtein* distance. It appears that the cost of this variant is enormous and prevents anyone from basing an effective large-scale solution on this principle. In short, computation time doubles with each increment of the distance.

So, instead of running a full-fledged TAXREF classifier modulo distance, we took the binomial names in the gold standard and checked whether they matched an entry in the thesaurus, and which one. We observed that each increment of the distance recognized more positive occurrences, but also that for distances greater than 2, it is more extrapolating than simply correcting OCR errors. For instance, *Cygnus ferus* (a swan) is recognized as *Cionus netus* (a bug). Recall that our ultimate task is to find keys for indexing articles in LN, so such confusion must not be introduced.

Altogether, at distance 4, which is unreliable and highly computationally costly, there is still around a third of the true positive binoms in the sample that are not recognized. At distance 2, which seems to be the maximum safe distance, 40% of the true binoms (i.e., binomial names) of the sample are not recognized in TAXREF modulo distance. It is even worse when combining the abbreviation refinement with the distance refinement. For instance, the true positive *D. punjabicus* (an extinct great ape) is recognized as *Dasyscyphus pudicus* (a mold). However, even at this distance, about 30% of the true abbreviated binomial names are not recognized in TAXREF.

These results confirm that the important drift in classification combined with noise renders taxonomic thesaurus ill-suited for the task of detecting binomial names in a historical and noisy corpus, as is LN. Nevertheless, we have shown in (Morand & Ridoux, 2023) that a thesaurus such as TAXREF could be used to learn valid Latin declensions.

Finally, QUAESITOR proposes a solution that is not dictionary-based.

4.4. QUAESITOR

Similarly to LINNAEUS, QUAESITOR (Little, 2020) aims at finding binomial names in scientific publications but with a more complex approach. It uses a com-

bination of pattern matching (regular expressions), a Bloom filter, and a trio of complementary ensembled neural networks. To evaluate whether the factors that prevent LINNAEUS from working in our context are only a matter of technological development, we also tested QUAESITOR. It is to be noted that since it aims at identifying the species mentioned, QUAESITOR outputs a list of *normalized* names. This is a normal part of a NER system, but not of our ENR task. Therefore, we had to undo the normalization to find the position of the binomial names it found.

QUAESITOR obtains the following results: a precision of 12.38%, a recall of 53.42%, and an F-measure of 20.10%.

We can see that QUAESITOR obtains better results than LINNAEUS. However, its F-measure is less than half that of the TAXREF classifier, and both have very unbalanced precision and recall. It seems that QUAESITOR has traded precision for recall. However, such a low precision cannot be accepted for our task of annotating articles because almost 90% of the tags would be irrelevant. (Little, 2020) compares QUAESITOR with other approaches, including LINNAEUS, on three benchmarks. The precision of LINNAEUS is evaluated at 70-90 %, and the recall is evaluated at 5-60 %, depending on the benchmark. This evaluation confirms that the observed performances of a tool are highly dependent on the task used for evaluation.

Altogether, neither LINNAEUS and QUAESITOR nor our TAXREF classifier can cope with the characteristics of our task, which is to annotate binomial names in a large, noisy, heterogeneous, and diachronic corpus. This baseline, combined with observations from the state of the art, justifies the need and choice of a new approach for our task.

5. A Competent Reader Imitator

We propose a method that yields significantly better results than the methods tested in the previous section. We call it *Competent Reader Imitator* (CRI). It combines two heuristics:

1. To use the Latin structure of binomial names as a search pattern. We call this heuristics *LATIN*.
2. To observe that binomial names are rare words and thus restrict the search to rare words. We call this heuristics *Rare-Frequent Threshold* (RFT).

Taken separately, these two heuristics do not yield better results than other methods. However, their combination yields much better results. In short, the CRI heuristic is to combine LATIN and RFT.

5.1. LATIN

The LATIN heuristic is to combine formal patterns that binomial names follow.

As in (Koning et al., 2005, Sautter et al., 2006), we use a first pattern, being that binomial names must look Latin. They must end with a Latin declension, and the species name declension must agree with the genus name according to Latin grammar rules. This pattern does not work well alone because many Latin declensions are very common in French, such as ending with an 's' or an 'e'. Note that these French declensions can be only superficially similar to Latin declension: e.g., "Whymper recommença ..." ("Whymper tried again ...", intentionally left non-italicized). However, this conjunction of a subject proper name and a third-person past tense verb is very frequent in narration, and 'a' is a plausible declension for a Latin adjective.

A second pattern is that the first word of a pair should start with a capital letter, and the second should not.

Definition 3 (Mm) *In the sequel, Mm (Majuscule-minuscule) refers to name pairs that respect the Latin declensions and the capitalization of the first world only.*

This pattern alone is insufficient because it represents too many situations that have nothing to do with binomial names. Most sentences begin with a sequence that satisfies this pattern, and most proper names initiate a similar sequence.

A third pattern is that the genus can be abbreviated in its capitalized initial letter.

Definition 4 (A) *In the sequel, A (Abbreviation) refers to pairs whose first part is a capital letter with a dot and whose second part respects the Latin declensions of specific names (i.e., adjective, nominative noun, or genitive noun).*

This pattern is also too broad since it is frequent to encounter "M. Lastname" (French for "Mr Lastname") or even "F. Lastname" where "F." is the initial of the first name ("Fabiola Lastname"). In fact, a side-rule of this pattern is that the genus name must occur in full closely before its abbreviation.

Taken individually, these patterns are way too permissive. Taken collectively, they are more precise, but in practice, not precise enough. Binoms respecting the *A*

or *Mm* pattern are the conventional binoms, e.g., the ones that would be considered valid with respect to modern codes.

To recognize patterns in text, we use *regular expressions* (Jurafsky & Martin, 2009). There are the three Latin genera to be considered, the five Latin declensions, and the three authorized forms: nominative-nominative, nominative-genitive, and nominative-adjective. The regular expression that describes all possible combinations is too large to be easily written by hand. However, it is very systematically constructed, and a rather simple program can generate it automatically from the authorized forms and the Latin declension rules. This amounts to thousands of binomial name patterns combined in a huge regular expression, which is then passed to methods of the `regex` Python library to compile it into an automaton.

The CRI classifier works in two passes. The first one aims at finding all the *Mm* patterns. This will return positive entries, some true, others false. After the first pass, the classifier performs a second pass intending to find binomial names with an abbreviated genus name. To do so, it generates a new regular expression that matches the first letter of all genus names that have been found in the *Mm* path, followed by a dot and then any lower-case word that respects a specific name declension. To cope with the frequent "M. Lastname", this second pass ignores binomial name candidates that start with a capitalized "M". This is a slight loss in recall but a significant gain in precision.

Used naively, all these purely syntactic patterns are not very discriminating on their own and give terrible results. Indeed, they generate a lot of false positives, hence a low precision (about 20 %). However, as we observed in section 3.3, with only around one binomial name every 3.5 pages, binomial names are rare in the statistical distribution of all the words in the corpus. To account for this, we only seek binoms that match the LATIN heuristic and that do not contain any frequent words. It remains to determine what it is to be frequent.

5.2. A Rare-Frequent Threshold

The frequency heuristic is to ignore all words that are more frequent than a Rare-Frequent Threshold (RFT in the sequel).

Definition 5 (Rare-Frequent Threshold, RFT) *Considering an ordering of all words of the collection (about 1,000,000 unique words) from the most frequent to the least frequent, the RFT is the rank r that separates words that are excluded, i.e., all words w with rank $r_w < r$, from words which are considered as possible genus and species names, i.e., all words w with rank $r_w \geq r$.*

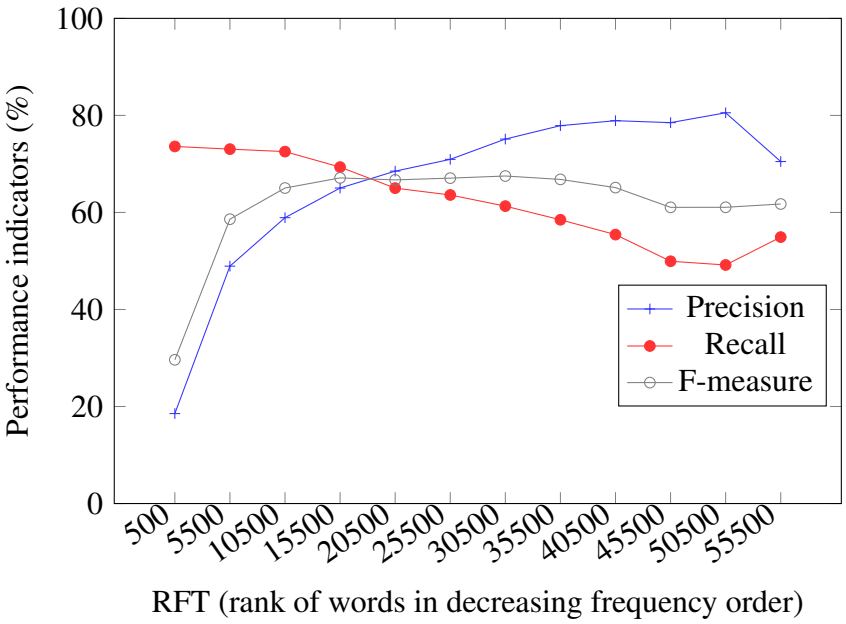

Figure 3 – Evolution with respect to increasing RFT of the different performance indicators for the modified LATIN classifier that implements the RFT

The RFT only applies to words of length strictly greater than one letter to cope with abbreviations.

To identify a useful RFT, we repeatedly apply a modified LATIN classifier that takes frequency into account to a sequence of increasing frequency thresholds. Applied to manually annotated volumes, this application will return a range of performance indicators for the whole sequence, which will then help us to choose the best RFT value.

The estimation of RFT also implies that the gold standard is now split into two subsets. The first subset comprises volumes 12 and 126 and serves as a "*training set*" to calibrate the RFT. The second comprises volumes 83 and 155 and serves as the *evaluation* or test set.

Figure 3 shows the evolution of the different performance indicators of the modified classifier as the RFT increases. Best choices can be read at the maxima of the F-measure curve. The LATIN patterns are very inefficient for a low RFT (left part of the graph). Indeed, although LATIN classifies most of the binomial names of the calibration corpus correctly (indicated by a high recall), precision is

very low. This low precision suggests that LATIN accepts many pairs that are not true binomial nouns (even in the weak sense of the CRH).

The precision improves greatly after removing the most frequent words (e.g., the 15,000 most frequent words or the first 1.5% of the vocabulary). At the same time, recall slightly decreases as expected because a stronger condition is applied. However, the decrease in recall is much less than the increase in precision, and the F-measure increases.

For an RFT between 15,000 and 40,000, the F-measure is stable, and choosing an RFT is mainly a matter of prioritizing precision or recall. These observations suggest that our future experiments can be set to always use an RFT of 15,000, which is the point of precision-recall equilibrium. Thus, this experiment shows that syntactic patterns that were initially not very discriminating become relevant when applied to sufficiently rare words.

The modified LATIN classifier with an RFT set at 15,000 is the core of the CRI classifier (see the lower half of Figure 1). At this point, CRI implements a classifier that strictly follows taxonomic codes. The next step involves coping with frequently observed deviations from the code.

5.3. Handling deviations from the code

As mentioned in Section 2.1, the LN corpus presents many variations from modern taxonomic rules. Such variations depend on the authors and change over time. Consequently, many entirely legitimate binoms (according to the Competent Reader Hypothesis) do not respect established conventions. For instance, it is common that when the specific name refers to a place or a person, such as in *Trachycarpus fortunei* (a palm tree named after Mr Robert Fortune, hence the genitive), it is capitalized as is the genus name. To account for this usage, an option is added for accepting Latin-like binoms where both words are capitalized. symmetrically, another option is added for accepting Latin-like binoms where both words are not capitalized.

Definition 6 (MM and mm) *In the sequel, MM refers to word pairs that respect the Latin declension rules where both words are capitalized, and mm refers to word pairs that respect the Latin declension rules where words are not capitalized.*

Unfortunately, the *MM* and *mm* patterns invite too many false positives. Restricted Latin patterns are therefore used to compensate for these inconsistent patterns, i.e., restricted variants of the (*ending*, *gender*) tables are created that exclude endings that are too French-like. For instance, the ending `"s"` is excluded while the ending

`"us"` is retained. These restricted variants are then used to generate the regular expression used to recognize relaxed patterns in the text, such as *MM* and *mm*, in order to avoid producing an altogether over-generalizing pattern.

To account for these new patterns, the CRI classifier is modified as follows. The first pass tries to identify *Mm* but also *MM* and *mm* binoms. The second pass now tries to find abbreviations corresponding to only the candidate binoms that start with a capitalized letter (i.e., *Mm* or *MM* patterns), meaning that if the classifier's first pass found "Homo sapiens", "Canis Lupus" and "felix gatus", the classifier will generate the expression for any "H." or "C.", followed by a space and any word. It will not try to find "F." followed by a space and any word since `felix` is not capitalized in the match. To accept non-capitalized genus names (here, `felix` with a small `'f'`) is already a relaxation of the rule that generates many false positives. For instance, "*qui*" ("who/that/which") is a frequent occurrence in French. It matches a Latin declension, but has nothing to do with Linnean taxonomy. It is rare behind a capitalized word, and the LATIN classifier will filter out much of them. However, it is much more frequent behind a non-capitalized word, and the LATIN classifier will let some of them go through (e.g., in "vapeur d'eau qui accompagne", French for "Steam that accompanies", *eau qui* is composed of plausible latin declension, and might be accepted by the *mm* pattern). As a rule, we do not want a relaxation apply to another relaxation, generating still more false positives.

Another relaxation is to use accentuated letters in binomial names. Again, this goes against established conventions, but it happens. However, the situation differs a lot from the *MM* and *mm* cases. While the *MM* and *mm* cases are frequent, the use of accentuated letters in binomial names remains rare (about 1% in the sample studied in Section 6.2), so the gain in recall is small. Accepting this relaxation would also lead to far too many ordinary French words being accepted, so the loss in precision would be enormous. We therefore decided to ignore this relaxed usage.

Another idea that comes to mind is to use trigger words to detect binomial names that the preceding methods might have missed. However, we observe that the words *genre* (French for genus) and *espèce* (French for species), which should serve as triggers for the trigger-based approach, have too many other unrelated occurrences, as shown in Figure 4 (the average number of mentions of the trigger words in a page does not grow with the number of binomial names in the same page). This can be easily observed for the word *genre*, whose distribution is almost flat, i.e., independent from the density of binomial names. For this reason, we have decided to disregard this possibility.

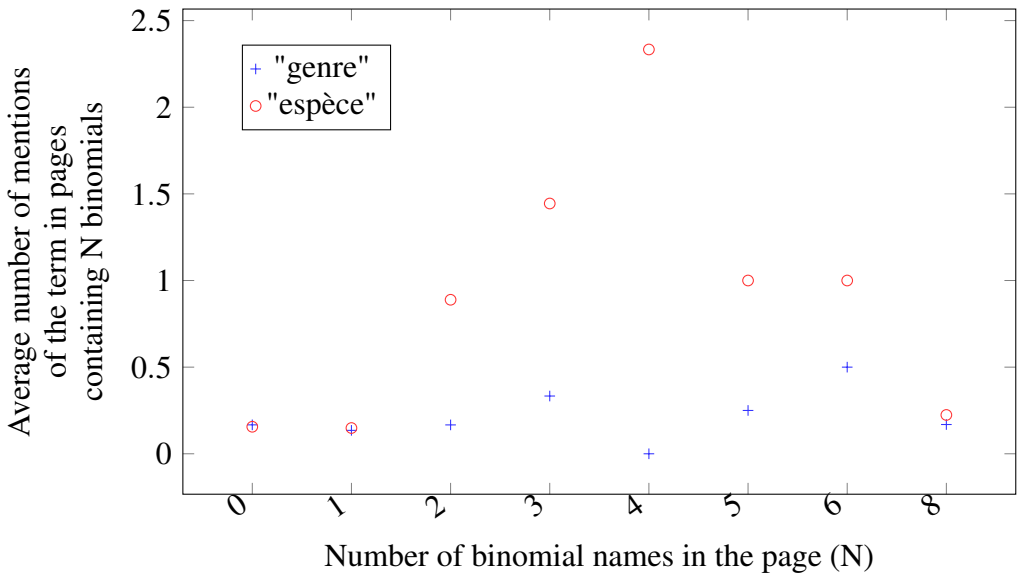

Figure 4 – Average number of mentions of the terms "genre" and "espèce" in volume 103 of LN in function of the number of binomial names (tagged by CRI) in the pages

We have presented a new classifier with many possible variants. In the next section, we evaluate whether it helps solve the task of finding the position of binomial names in a diachronic, heterogeneous, and noisy corpus.

6. Evaluation

The evaluation of the CRI classifier is presented as follows. First, in Section 6.1, the performances of the CRI classifier (precision, recall, F-measure) are compared with methods from the state-of-the-art section (LINNEAUS and QUAESITOR, see 4.1) with respect to the gold standard, i.e., the *ex-ante* manual annotations of four volumes (see section 3.3). Then, in section 6.2, we perform an *ex-post* validation to estimate precision over the whole LN archive. Then, as we have tried and discussed numerous variations (*Mm*, *A*, *MM*, *mm*, for majuscule-minuscule, abbreviation, majuscule-majuscule, and minuscule-minuscule, all defined in sections 5.1 and 5.3), Section 6.3 evaluates the sensitivity of the CRI classifier to the types of patterns that are recognized. Finally, Section 6.4 explores the sensitivity of our classifier to heterogeneity in the LN archive.

Classifier	LINNAEUS	QUAESITOR	CRI (*Mm A*)	CRI (*Mm A MM mm*)
precision (%)	18.92	13.61	**68.95**	60.10
recall (%)	23.71	57.65	69.98	**79.93**
F-measure (%)	21.04	22.02	**69.47**	68.61
real time (minutes)	**3'23**	12'54	4'43	~ 20
user time (minutes)	7'44	14'30	**4'43**	~ 20

Table 1 – Comparison of CRI with LINNAEUS and QUAESITOR on the evaluation corpus

6.1. Comparison with the state of the art

Table 1 compares the performances of LINNAEUS, QUAESITOR, and two variants of CRI: the basic variant with *A* and *Mm*, and a relaxed variant that adds *MM* and *mm*. In this table, row "real time" shows the perceived computation time from the outside of the computer (aka "wall-clock time" or "real-world time"). Row "user time" shows the total work required to compute binomial names for the evaluation corpus on the several cores of a personal laptop (ASUS 2019, Core i7-7700HQ, 2.80 GHz, eight cores, 8 GB memory). Row "user time" shows a greater value than row "real time" when several cores are used simultaneously. We deliberately did not seek an optimized computation time for this work. Even if our lab is equipped with a high-performance computing system, our goal is that less equipped institutions, or even individuals, could repeat the same type of experiments (Gundersen et al., 2018).

We observe that the two variants of CRI yield much better performances than both LINNAEUS and QUAESITOR. The two variants of CRI yield equivalent F-measures, but the basic one has better precision, and the relaxed one better recall, at the price of a longer run-time.

Note that LINNAEUS uses several processor cores, though CRI only uses one. This shows that the run-time of CRI with *A* and *Mm* can still be improved, though it is already the solution that demands the least amount of work.

6.2. Ex-post evaluation

The previous evaluation is based on the gold standard, i.e., the *ex-ante* manual annotations of four volumes. The goal of this section is to evaluate whether these results

generalize to all volumes. The principle of this evaluation is to compute a sample of all positive answers and check manually whether the answers in the sample are true positives or not. This can easily be done for precision. In this section, we test the basic variant of CRI (*A* plus *Mm*).

A random sample of 5% of all binoms extracted from the whole collection by the CRI classifier was computed. In order to make the manual checking easier to perform, the sampled binoms were extracted with their context, in a concordancer style (Barrière, 2016): 30 characters on the left of each positive binom and 30 characters on the right.

Then, we played the role of the Competent Reader. In other words, we checked whether each binom could be read as an intended genus-species name, regardless of whether it is an actual binom of a modern taxonomy or adequately written. Consider, for instance, the following outputs and their evaluation by the Competent Reader (the matched strings are underlined for ease of reading, though they are not in the original output):

- `tares envahie par les mulots (Wus sylvaticus) et par les campagnols (Arvic`
- `s secteurs déterminés : ainsi Suæda frulicosa, espèce méditerranéenne, ne s`
- `es Canards sauvages colverts (Anas boschas) et même les Oies sauvages de`
- `Crescens, en 1474, publia un Opus ruralium commodorum plein de renseigne`
- `. ..,..,..... posssnsrssse 46 Arkllerie prussienne. ........4......sssses es " H`
- `lations.  - On connait, dit la Gazelle hebdo madaire de médecine, les lois`

Wus for *Mus* (the genus of mice) is accepted as a true positive despite an OCR error (*W* for *M*, a frequent error). Similarly, *Suæda frulicosa* is accepted despite a widespread OCR error (*l* for *t* in *fruticosa*) and a deprecated spelling (*æ* for *ae* in *Suaeda*). And *Anas boschas* is accepted despite being an obsolete name for *Anas platyrhynchos*, the mallard or wild duck. Note also the spurious capitalization of "Canards" (ducks) and "Oies" (geese) in the context. In the next sample, *Opus ruralium* is part of a Latin book title, which the concordancer shows well. The last items exemplify what word monstrosities or unintentional puns OCR can create. Notice in particular how the same widespread OCR error as above (*l* for *t*, twice in a row) in conjunction with another OCR error (inserting a blank character) transforms

types of binoms searched	*Mm*	*Mm*, *MM*	*Mm*, *A*	*Mm*, *MM*, *A*	*Mm*, *mm* *A*	*Mm*, *MM*, *mm*, *A*
precision (%)	**75.70**	69.39	68.95	62.53	65.44	60.10
recall (%)	67.66	73.30	69.98	76.12	73.80	**79.93**
F-measure (%)	**71.45**	71.29	69.47	68.66	69.37	68.61
time (minutes)	$\simeq 5$	$\simeq 5$	$\simeq 5$	$\simeq 5$	15	20

Table 2 – Variation of the performance indicators with respect to binom patterns (Vol. 83 and 155)

Gazette hebdomadaire (weekly newsletter) into *Gazelle hebdo* which has a definite Latin look if one only considers word endings, as the classifier does. Note also that *Gazelle* is only one letter apart from the valid genus *Gazella* and that taking an *a* for an *e* is a frequent OCR error.

This *ex-post* analysis showed a sampled precision of 68%, which is in the same order as for the test volumes 83 and 155. This shows that the precision estimated for the whole LN archive is similar to the precision measured with respect to the gold standard.

6.3. Sensitivity to accepted patterns

The choice of a variant among combinations of *Mm*, *A*, *MM*, and *mm* is a hyperparameter of the CRI classifier (see on the right side of the lower half of Figure 1). This section presents the sensitivity of the classifier with respect to this hyperparameter.

Table 2 presents the variations in the different quality indicators when looking for different patterns on volumes 83 and 155 (the evaluation corpus) with an RFT of 15,000 words.

The table can be analyzed in terms of the basic patterns as follows:

— Pattern *Mm*: when used alone, it corresponds to binomial names that could appear in a modern taxonomic thesaurus. This yields the best precision and the worst recall. It is the basis of all pattern combinations.

— Pattern *A*: it accepts abbreviated forms. Recall that they are not relaxed forms but plain legal forms. This pattern increases recall by about 3%. It confirms that abbreviated forms are frequent.

— Pattern *MM*: it accepts more relaxed forms of binomial names. It causes a drop in precision and an increase in recall. Computation time increases by about 20%. For the first four columns, it remains within the minute round-

ing. Note that *MM* leads to a significant increase in the recall at the price of many false positives, which come from confusion with proper names. The RFT barely filters proper names because they are also rare.

— Pattern *mm*: like *MM* it is a relaxed form but behaves differently. Pattern *mm* impacts little the results, by comparison with the pattern *MM*. It brings a smaller increase in recall and a little loss in precision. Its cost is the increase in computation time because many more word pairs must be tested.

Overall, the F-measure is robust to changes in the patterns searched. Therefore, depending on one's objectives, one can confidently choose a combination of basic patterns. If precision is very important (i.e., false positives are costly), then one must restrict to the sole *Mm* pattern. On the contrary, if false negatives are costly combining all patterns will maximize recall.

6.4. Sensitivity to thematic variability

Several authors mention that thematic heterogeneity impedes recognizing named entities (Ehrmann et al., 2021, Nadeau & Sekine, 2007). This section confirms this observation for the case of identifying binomial names in the LN corpus.

Tables 3a and 3b present the results obtained by the CRI classifier on the four manually annotated volumes when trying to maximize recall (*Mm*, *A*, *MM* and *mm*) where volumes 12 and 126 only served for calibration, and volumes 83 and 155 for the evaluation. Table 3a shows the results when considering the whole corpus, while Table 3b shows the results when only articles with actual occurrences of binomial names are kept.

This last experiment measures the negative effect of heterogeneity in our corpus. Recall that *La Nature* is a generalist magazine, and that Life Science articles correspond to a small subset of it. The experiment shows that a considerable part of the false positives are due to articles that do not mention binomial names at all. This is indicated by the almost 20-point increase in precision when only considering relevant articles and the resulting 10-point increase in the F-measure. This is why we argue that the corpus statistics are part of the definition of the task. In the present case, diachronicity, noisiness, and heterogeneity of the corpus are as much part of the definition of the task as are the characteristics of binomial names. Indeed, by only changing the distribution of articles that contain binomial names the same evaluation protocol shows a 20-point difference.

	vol. 12	vol. 83	vol. 126	vol. 155	vol. 83 & 155	vol. 12, 83, 126 & 155
precision (%)	67.43	61.25	43.38	58.76	60.10	58.16
recall (%)	82.81	81.99	78.33	77.58	79.93	80.43
F-measure (%)	74.33	70.12	55.84	66.87	68.61	67.50

(a) Scores on the corpus considering all articles

	vol. 12	vol. 83	vol. 126	vol. 155	vol. 83 & 155	vol. 12, 83, 126 & 155
precision (%)	86.45	86.27	69.12	72.19	79.28	79.17
recall (%)	82.81	81.99	78.33	77.58	73.93	80.43
F-measure (%)	84.59	84.08	73.44	74.79	79.60	79.80

(b) Scores on the corpus only considering life science articles

Table 3 – Scores on the corpus with the RFT set to 15,000 words and looking for patterns *Mm*, *MM*, *mm* and *A* (column 83 & 155 corresponds to the evaluation corpus while column 12, 83, 126 & 155 corresponds to both the training and evaluation corpora)

7. Conclusion & future work

When describing the state-of-the-art in section 4.1, we discussed the impacts, positive and negative, of our research. We start this conclusion by estimating the carbon footprint of the work presented in this article. We then discuss the academic results of this work and, finally, some directions for future work.

7.1. Carbon impact

Regarding recent AI work, several authors object to the untold environmental damages of these technologies (Strubell et al., 2019, Schwartz et al., 2020, Bannour et al., 2021). They call for a systematic reporting of AI research impacts. We apply this recommendation to our work, but it is only a tentative effort because we lack the tools and norms for a more fine-grained evaluation.

We have evaluated the carbon footprint of our work using the Green Algorithms calculator (`calculator.green-algorithms.org` v2.2) (Lannelongue et al., 2021). We estimate the design and the (multiple) runs of all of the exper-

iments in this work to have necessitated 100 h on 1 CPU Intel core i7-7700HQ. This amounts to an energy consumption of 1.43 kWh. Based in France (with an average carbon intensity of the electricity mix of 51.28 g CO_2e/kWh), this leads to a carbon footprint of 73.23 g CO_2e. Since the i7-7700HQ is not in the database of the calculator, we declared an A8-7680 CPU because it has a similar energy consumption profile.

Applying the CRI classifier to the whole LN archive necessitated 10 hours on the same computer, hence an estimated consumption of 0.15 kWh and a supplementary carbon footprint of 7.3 g CO_2e.

7.2. Conclusion / discussion

In this work, we have introduced the Competent Reader Hypothesis (CRH) to define the task of recognizing binomial names in a diachronic, heterogeneous, and noisy corpus. The idea is that a Competent Reader knows the linguistic codes of a popular science magazine and, therefore, recognizes binomial names, even on the first encounter, even when depreciated or ill-formed, and can differentiate the colloquial use of a Latin genus from its scientific use. Nevertheless, the Competent Reader is not an expert reader. This is coherent with the task of analyzing the articles of a popular science magazine.

CRH serves as a guideline for the evaluation of candidate classifiers for our task. In particular, it was used for manually annotating part of the LN archive, forming a gold standard corpus.

We reviewed several available classifiers against this gold standard and found their performance insufficient for our task. This can be easily explained by the constant evolution of the Linnean classification, by OCR noise, and by frequent deviations from the taxonomic codes in the LN corpus.

We have introduced CRI, a rule-based approach that imitates the CR. Since binomial names are very rare in the corpus, the CRI approach that we have developed combines a filter that recognizes the expected form of binomial names (LATIN) with a frequency filter (RFT) that eliminates frequent words. Taken separately, these two filters are very weak, but their combination provides acceptable performances.

The performance of CRI is about 70% regarding precision, recall, and F-measure. We have observed that the F-measure is stable with respect to changes in hyperparameters, though precision and recall may vary. Ex-post annotation in section 6.2 has shown that precision evaluated on test volumes is confirmed on random samples of all answers for the whole collection. Our experiments also confirm

the negative impact of thematic heterogeneity on performance. Indeed, when non-life-science articles are suppressed from the corpus, precision increases by about 20 points.

Our experiments show that the performances heavily depend on the task. This requires specifying the task with great precision, including the properties of the corpus. Our ultimate task is not only to recognize binomial names but, more generally, to facilitate digital humanities research. Consequently, an actual evaluation should be done with reference to this ultimate task.

The program codes and data of the experiments described in this article are available at `https://github.com/oridoux/TAXONER/tree/main/LI2024/CRI`.

7.3. Future Work

Our method does not yet take into account the textual context of candidate binoms. Possible contextual information that could be used to improve the method is the presence/absence of the vernacular names of living organisms close to the candidate binoms. To do so, thesauruses like TAXREF or the Catalog of Life (Bánki et al., 2023) could greatly help.

The Competent Reader Hypothesis has many potential applications. For instance, it could be applied to other classes of named entities with similar characteristics to those of binomial names. In the field of chemistry, chemical names are highly recognizable as such, even if one does not understand what they mean. Like biology, chemistry has adopted international nomenclature codes since the end of the 19th century (Eltyeb & Salim, 2014). Similarly to biology, the nomenclature has changed a lot over time, and the adoption of these codes in popular science magazines has been inconsistent. Similarly, names of scientific and technical artifacts can be easily recognized by a Competent Reader even without knowing what those objects do (e.g., "...scope" or "...tron"). We believe that these classes, chemical names, and technical innovations could benefit from the CRI approach.

A part of the difficulty of our task came from surface confusion between Latin and French. It would be interesting to test the CRI classifier on archives with a similar structure to LN but written in a language other than French. For instance, *The Scientific American* and *Nature* started publication in the 19th century with similar objectives to LN.

The validation mode presented here is called *intrinsic* (Clark et al., 2012). It is the only possible choice since there are no digital humanities applications that ex-

ploit the semantic annotations we compute. As soon as such an application is available, it will be possible to proceed to an *extrinsic* validation, i.e., a measure of how a downstream application is affected by the imperfections of semantic annotation. Such an application is currently under development. It consists of a "conceptual" navigation interface, i.e., one that allows a user to elaborate a query in a dialogue with the application (see bottom of Figure 1).

We are satisfied with the results of our approach for detecting binomial names in LN. Consequently, we think future work should strive for the detection of a greater variety of attributes and combine them for an effective digital humanities application without forgetting our global objectives of digital sufficiency.

References

Abdalla M. & Abdalla M. (2021). The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, p. 287–297, New York, NY, USA: Association for Computing Machinery.

Abdalla M., Wahle J. P., Ruas T. L., Névéol A., Ducel F., Mohammad S. M. & Fort K. (2023). The elephant in the room: Analyzing the presence of big tech in natural language processing research. In A. Rogers, J. L. Boyd-Graber & N. Okazaki, Eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, p. 13141–13160: Association for Computational Linguistics.

Akella L. M., Norton C. & Miller H. (2012). Netineti: Discovery of scientific names from text using machine learning methods. BMC Bioinformatics, **13**, 211.

Bánki O., Roskov Y., Döring M., Ower G., Hernández Robles D., Plata Corredor C., Stjernegaard Jeppesen T., Örn A., Vandepitte L., Hobern D., Schalk P., DeWalt R., Ma K., Miller J., Orrell T., Aalbu R., Abbott J., Adlard R. & Adriaenssens E. e. a. (2023). Catalogue of Life Checklist.

Bannour N., Ghannay S., Névéol A. & Ligozat A. (2021). Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In N. S. Moosavi, I. Gurevych, A. Fan, T. Wolf, Y. Hou, A. Marasovic & S. Ravi, Eds., Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, SustaiNLP@EMNLP 2021, Virtual, November 10, 2021, p. 11–21: Association for Computational Linguistics.

Barrière C. (2016). Natural Language Understanding in a Semantic Web Context. Springer.

Becker C. (2023). Insolvent: How to Reorient Computing for Just Sustainability. Cambridge: The MIT Press.

Bender E. M., Gebru T., McMillan-Major A. & Shmitchell S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, p. 610–623, New York, NY, USA: Association for Computing Machinery.

Birhane A., Kalluri P., Card D., Agnew W., Dotan R. & Bao M. (2022). The values encoded in machine learning research. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, p. 173–184, New York, NY, USA: Association for Computing Machinery.

Burdick A., Drucker J., Lunenfeld P., Presner T. & Schnapp J. (2012). Digital Humanities. The MIT Press.

Castellan S., Käfer J. & Tannier E. (2023). Back to the trees: Identifying plants with Human Intelligence. In Ninth Computing within Limits 2023: LIMITS. https://limits.pubpub.org/pub/sapyi15v.

Clark A., Fox C. & Lappin S. (2012). The handbook of computational linguistics and natural language processing, volume 118. John Wiley & Sons.

CNUM (ca. 2000). Conservatoire numérique des Arts et Métiers. HTTP links to scanned fac-simile of LA NATURE: http://cnum.cnam.fr/CGI/ redira.cgi?4KY28.

COMETS, Ethics Committee of the CNRS (2022). AVIS n 2022-43, Intégrer les enjeux environnementaux à la conduite de la recherche - Une responsabilité éthique.

Cunha W., Mangaravite V., Gomes C., Canuto S., Resende E., Nascimento C., Viegas F., França C., Martins W. S., Almeida J. M., Rosa T., Rocha L. & Gonçalves M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. Information Processing & Management, **58**(3), 102481.

DeMillo R., Lipton R. & Sayward F. (1978). Hints on test data selection: Help for the practicing programmer. Computer, **11**(4), 34–41.

Devlin J., Chang M., Lee K. & Toutanova K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding.

Ehrmann M., Hamdi A., Pontes E. L., Romanello M. & Doucet A. (2021). Named entity recognition and classification on historical documents: A survey. CoRR, **abs/2109.11406**.

Eltyeb S. & Salim N. (2014). Chemical named entities recognition: a review on approaches and applications. Cheminform., **6:**(17).

Gabrys J., Pritchard H. & Barratt B. (2016). Just good enough data: Figuring data citizenships through air pollution sensing and data stories. Big Data & Society, **3**(2), 2053951716679677.

Gargominy O., Tercerie S., Régnier C., Ramage T., Dupont, P., Daszkiewicz P. & Poncet L. (2021). TAXREF v15, référentiel taxonomique pour la France : méthodologie, mise en œuvre et diffusion.

Gerner M., Nenadic G. & Bergman C. M. (2010). Linnaeus: a species name identification system for biomedical literature. BMC Bioinformatics, **11**(1), 1–17.

Gundersen O. E., Gil Y. & Aha D. W. (2018). On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. AI Magazine, **39**.

Gupta U., Kim Y. G., Lee S., Tse J., Lee H.-H. S., Wei G.-Y., Brooks D. & Wu C.-J. (2020). Chasing carbon: The elusive environmental footprint of computing.

Jurafsky D. & Martin J. H. (2009). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Pearson Prentice Hall.

Koning D., Sarkar I. N. & Moritz T. (2005). Taxongrab: Extracting taxonomic names from text. Biodiversity Informatics, **2**, 79–82.

Kuhn T. S. (1962). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Labusch K., Neudecker C. & Zellhofer D. (2019). Bert for named entity recognition in contemporary and historic german. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, p. 1–9, Erlangen, Germany: German Society for Computational Linguistics & Language Technology.

Lannelongue L., Grealey J. & Inouye M. (2021). Green algorithms: Quantifying the carbon footprint of computation. Advanced Science, **8**(12), 2100707.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L. & Schwab D. (2020). Flaubert: Unsupervised language model pre-training for french. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, Eds., Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, p. 2479–2490: European Language Resources Association.

Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H. & Kang J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, **36**(4), 1234–1240.

Little D. (2020). Recognition of Latin scientific names using artificial neural networks. Applications in Plant Sciences, **8**.

Luccioni A. S., Viguier S. & Ligozat A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176b parameter language model. Journal of Machine Learning Research, **24**(253), 1–15.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D. & Sagot B. (2020). CamemBERT: a tasty French language model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 7203–7219, Online: Association for Computational Linguistics.

Morand C. & Ridoux O. (2023). Extraction dans des textes anciens d'entités nommées de type binômes de la classification linnéenne du vivant : une étude de cas. Revue des Nouvelles Technologies de l'Information, **Extraction et Gestion des Connaissances, RNTI-E-39**, 417–424.

Mozzherin D., Myltsev A. & Patterson D. (2017). "gnparser": A powerful parser for scientific names based on parsing expression grammar. BMC Bioinformatics, **18**.

Nadeau D. & Sekine S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, **30**, 3–26.

Nasar Z., Jaffry S. W. & Malik M. (2021). Named entity recognition and relation extraction: State of the art. ACM Computing Surveys, **54**.

NCBI (2008). The national center for biotechnology information taxonomy.

Nédellec C., Bessières P., Bossy R. R., Kotoujansky A. & Manine A.-P. (2006). Annotation guidelines for machine learning-based named entity recognition in microbiology. In Proceeding of Data and Text Mining for Integrative Biology Workshop 17.European Conference on Machine Learning 10. European Conference on Principles and Practice of Knowledge Discovery in Databases, Workshop on data and text mining for integrative biology: Springer - Verlag.

Nguyen N. T. H., Gabud R. & Ananiadou S. (2019). Copious: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodiversity Data Journal.

Pafilis E., Frankild S. P., Fanini L., Faulwetter S., Pavloudi C., Vasileiadou A., Arvanitidis C. & Jensen L. J. (2013). The species and organisms resources for fast and accurate identification of taxonomic names in text. PloS one, **8**(6), e65390.

G. M. Sacco & Y. Tzitzikas, Eds. (2009). Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience, volume 25 of The Information Retrieval Series. Springer.

Santarius T., Bieser J. C. T., Frick V., HÃ¶jer M., Gossen M., Hilty L. M., Kern E., Pohl J., Rohde F. & Lange S. (2022). Digital sufficiency: conceptual considerations for icts on a finite planet. Annals of Telecommunications, **78**(5-6), 277–295.

Sautter G., Böhm K. & Agosti D. (2006). A combining approach to find all taxon names (FAT). Biodiversity Informatics, **3**.

Schwartz R., Dodge J., Smith N. A. & Etzioni O. (2020). Green AI. Commun. ACM, **63**(12), 54–63.

Seideh M. A. F., Fehri H. & Haddar K. (2016). Named entity recognition from arabic-french herbalism parallel corpora. In T. Okrut, Y. Hetsevich, M. Silberztein & H. Stanislavenka, Eds., Automatic Processing of Natural-Language Electronic Texts with NooJ, p. 191–201, Cham: Springer International Publishing.

Sevilla J., Heim L., Ho A., Besiroglu T., Hobbhahn M. & Villalobos P. (2022). Compute trends across three eras of machine learning. In 2022 International Joint Conference on Neural Networks (IJCNN), p. 1–8.

Smil V. (2021). Grand Transitions: How the Modern World Was Made. Oxford University Press.

Strubell E., Ganesh A. & McCallum A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, p. 3645–3650, Florence, Italy: Association for Computational Linguistics.

Thompson N., Greenewald K., Lee K. & Manso G. F. (2023). The Computational Limits of Deep Learning. In Ninth Computing within Limits 2023: LIMITS. https://limits.pubpub.org/pub/wm1lwjce.

Tissandier G. (1873-1962). LA NATURE : Revue des sciences et de leurs applications aux arts et à l'industrie.

N. J. Turland, J. H. Wiersema, F. R. Barrie, W. Greuter, D. L. Hawksworth, P. S. Herendeen, S. Knapp, W.-H. Kusber, D.-Z. Li, K. Marhold, T. W. May, J. McNeill, A. M. Monro, J. Prado, M. J. Price & G. F. Smith, Eds. (2018). International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code). Glashütten: Koeltz Botanical Books.

Vautrin G. (2018). Histoire de la vulgarisation scientifique avant 1900 (History of Science Popularization before 1900 - in France). EDP sciences.

Yu P. & Wang X. (2020). Bert-based named entity recognition in chinese twenty-four histories. In G. Wang, X. Lin, J. Hendler, W. Song, Z. Xu & G. Liu, Eds., Web Information Systems and Applications, p. 289–301, Cham: Springer International Publishing.

Summary

La Nature (1873 - 1962) is a French popular science magazine that spanned a large time period and a large range of topics. It is available via ocerized archives so that it forms a corpus that is simultaneously diachronous, heterogeneous, and noisy. Although these characteristics make it complex to analyze, La Nature is of great interest for *digital humanities* studies on the evolution of thoughts in science, technology, and even politics. The work presented in this article is part of research on the semantic annotation of these archives, which is discovering clues for exploring them. One type of clue that has not been explored in a complex corpus such as La Nature is *binomial names*, or more specifically, the *named entities* that refer to the Linnean classification of life, e.g., *Escherichia coli*. To overcome this complexity, the concept of a *Competent Reader*, who can detect binomial names even when obsolete, non-standard or defaced by OCR, is introduced. By imitating a Competent Reader, our approach, which we call the *Competent Reader Imitator* (CRI), involves combining a rule-based approach with a frequency argument. We show that this innovative method is robust to numerous variations and consistently achieves an F-measure of about 70% despite diachronicity, heterogeneity, and noise, which are all known to impede named entity recognition. Our method has many potential applications, such as in the study of chemical names and names of scientific and technical artifacts, which could benefit from the Competent Reader imitation approach. Beyond our work on La Nature, we hope this paper provides a set of tools and methods that are easily understandable, frugal, and usable for a general public interested in exploring similar historical corpus.

Keywords

named-entity recognition, binomial names, historical corpus, digital sufficiency

Mots-clés

reconnaissance d'entités nommées, noms binominaux, corpus ancien, sobriété numérique

Author addresses:
Clément Morand
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique clement.morand@lisn.upsaclay.fr

Olivier Ridoux
Université de Rennes - IRISA
olivier.ridoux@irisa.fr