




Article

Assisting Forensic Identification through Unsupervised Information Extraction of Free Text Autopsy Reports: The Disappearances Cases during the Brazilian Military Dictatorship

Patricia Martin-Rodilla ^{1,*} , Marcia L. Hattori ²  and Cesar Gonzalez-Perez ² 

¹ Research Center on Information Technologies (CiTIUS), University of Santiago de Compostela, Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, Spain

² Institute of Heritage Sciences (Incipit), Spanish National Research Council (CSIC), Avda. Vigo, s/n, 15705 Santiago de Compostela, Spain

* Correspondence: patricia.martin.rodilla@usc.es

Received: 31 May 2019; Accepted: 3 July 2019; Published: 5 July 2019



Abstract: Anthropological, archaeological, and forensic studies situate enforced disappearance as a strategy associated with the Brazilian military dictatorship (1964–1985), leaving hundreds of persons without identity or cause of death identified. Their forensic reports are the only existing clue for people identification and detection of possible crimes associated with them. The exchange of information among institutions about the identities of disappeared people was not a common practice. Thus, their analysis requires unsupervised techniques, mainly due to the fact that their contextual annotation is extremely time-consuming, difficult to obtain, and with high dependence on the annotator. The use of these techniques allows researchers to assist in the identification and analysis in four areas: Common causes of death, relevant body locations, personal belongings terminology, and correlations between actors such as doctors and police officers involved in the disappearances. This paper analyzes almost 3000 textual reports of missing persons in São Paulo city during the Brazilian dictatorship through unsupervised algorithms of information extraction in Portuguese, identifying named entities and relevant terminology associated with these four criteria. The analysis allowed us to observe terminological patterns relevant for people identification (e.g., presence of rings or similar personal belongings) and automate the study of correlations between actors. The proposed system acts as a first classificatory and indexing middleware of the reports and represents a feasible system that can assist researchers working in pattern search among autopsy reports.

Keywords: information extraction; named entity recognition; terminology extraction; autopsy reports

1. Introduction

The development and improvement over the last few decades of natural language processing (hereafter NLP) algorithms, both in performance and precision in some relevant tasks (named entity recognition, open information extraction, part-of-speech tagging, among others), has allowed a more systematic coverage and application of these approaches to domains where large textual sources are common and highly specific terminology and discursive structures exist. Thus, it is common to find applications of NLP algorithms to legal, administrative, or medical reports with varying levels of supervision. Specifically, there are some available ad hoc corpora [1,2], reference indicators [3,4], and applications for these tasks in the medical domain [5–7], although with few applications of NLP to the treatment of forensic reports (autopsies and related reports). However, large investigation efforts into natural catastrophes, war massacres, or political conflicts generally involve an open universe of victims

and a high volume of forensic reports. In these contexts, professionals often require software-assisted treatment that allows them to jointly analyze a large volume of reports, looking for certain patterns in them. In addition to this need (given by the high volume of documents), most of these reports contain information about unsolved cases, which makes accurate and reliable processing of high humanitarian importance. For these reasons, the textual analysis and processing of reports like these is crucial.

In some Latin America countries like Brazil, there has been a studied [8–11] systematic strategy by the state, using the omission of information in forensic analysis or related forensic practices as a bureaucracy system to force people disappearances. This strategy possibly makes it difficult to identify the forensic cases tagged as NN (a category used for tagging person cases who we consider as potential missing persons without identification) cases, since the information presented in the forensic reports are vague or even filled with all information that could help to individualize that body. These situations have been categorized as negligent forensic practices [12].

In this paper, we present an analysis based on unsupervised information extraction algorithms in Portuguese of a collection of about 3000 forensic reports of persons buried as NN during 1971–1975, the most repressive period of the dictatorship. Our goal was to create a system that can act as a middleware to analyze the information contained in the reports based on the criteria established by anthropologists and forensic experts, who subsequently can perform a preliminary validation of the results.

The paper is organized as follows: The remainder of this Introduction section presents the historical context, analysis criteria, and motivation for this work, as well as a review of existing initiatives of natural language application in similar forensic contexts. Section 2 describes the materials and methods employed, including the particularities of the natural language suite *Linguakit* [13] for Portuguese, which was used as the basis for information extraction, as well as the forensic corpus analyzed. Section 3 presents the results obtained according to the expert criteria adopted: (1) Common causes of death, (2) relevant body locations, (3) personal belongings terminology, and (4) correlations between actors. Due to the limited output visualization options of the *Linguakit* suite, these results are subsequently treated and visualized with Google Data Studio [14] dashboards. Section 4 discusses the overall application and their forensic and technological implications. Finally, Section 5 summarizes the contributions and details future directions of work.

1.1. Historical Context and Motivation

On March 31 1964, making use of a coup and the force of the tanks in the streets, the military of Brazil institutionalized execution, torture, and murder. The civilian president, João Goulart, was exiled, people were arrested, and had their political and civil rights suspended. According to different studies made by the Amnesty Commission (the Amnesty Commission was installed in the Ministry of Justice on August 28 2001 and aims to examine and assess the amnesty applications, issuing an opinion intended to subsidize the Minister of Justice in the decision on the granting of Amnesty Policy) (2019), more than 75,000 persons demand compensation from the state because they were directly affected by the repression. This number shows the impact of the dictatorship and the consequences today in Brazilian society [15].

In 1990, the mass grave of Perus came to public attention [9]. A long investigation was carried out with the documentation of the legal medicine institutions of São Paulo [16], specifically about the names of the politically disappeared, who were buried with false names in the cemetery of Perus. However, it is recognized by the final report of the Brazilian national Truth Commission [17] that from the year 1973, the strategy of disappearance became systematic [18] and even documents related to death were produced (albeit with false information). The suspicion is that some of these cases could have been buried as NN, a category commonly used by institutions in cases where identity of a body is not known [9,19]. The forensic anthropology analysis carried out in the 1990s and 2000s were conducted by different research groups that ultimately did not complete the work and further

contributed to making more identification even more difficult [12]. Since 2014, an agreement between institutions has made significant progress [20].

Since the beginning of this last investigation in the Perus mass grave, only two people have been identified. The huge volume of cases only in Perus, not to mention the whole of São Paulo city, highlights the difficulty of the document analysis task, the possible impact on Brazilian society, and the need for software assistance. The present article aims to advance in this direction by applying named entity recognition and information extraction algorithms. Our final goal is to provide the experts with a simple software assistance system that allows them to analyze a corpus of forensic reports according to their own criteria and take, as a starting point, the textual information automatically extracted with unsupervised techniques. In the next section, common algorithms used and their applications to similar domains are described.

1.2. Automatic Information Extraction of Medical, Anthropological, and Forensic Information

The medical domain constituted one of the first domains for massive application of NLP studies for a number of reasons, including their narrative tradition in reporting or the controlled output formats, vocabularies, and linguistic structures and constructions. As we described in the introduction, it is common to find specific NLP corpora, metrics, and developments for medical purposes. However, it is difficult to find similar works for forensic information, especially in interdisciplinary contexts in which the authors of the reports are not only doctors, but also police officers, anthropologists, or archaeologists, which to a greater or lesser extent contributes to the collaborative reporting activities. There are some recent information extraction applications in analyzing separately police reports [21] and medical autopsies [22], with different levels of supervision and mainly in English. There are also some attempts for Portuguese [22], in which we can find a good review of other cross-domain approaches for NLP tasks, especially information extraction, but without references to previous works on forensic applications. The challenge of extracting information from different contributors from various backgrounds and based on expert criteria is presented in all revised works. These current efforts are clearly aligned with our proposal here and constitute a good basis to continue working on the same direction.

2. Materials and Methods

In this section, we describe the characteristics of the analyzed corpus, the conceptualization of the underlying information, and the NLP suite used for the unsupervised information extraction applications, which constitutes the pipeline of the proposed system.

2.1. The “Disappearances in São Paulo” Corpus

The corpus we adopted is composed of forensic reports compiled over many years of research about disappearances during the Brazilian military dictatorship in São Paulo. It includes a total of 2980 forensic reports of cases categorized as NN. It is necessary to emphasize that the reports contained in the corpus come from a great diversity of authors (doctors in their majority), who in turn worked from various sources of information coming in from police stations or testimonies from local people, etc. This implies that, although a certain degree of homogeneous treatment is necessary in order to build a computational solution, the reports do not share a common philosophy or make up a “designed” corpus; quite the contrary, they respond to different administrative practices of this past political regime. This heterogeneity presents an additional challenge, both for the human expert and for the computer system that must assist and analyze the texts.

As shown in Figure 1, each forensic report corresponds to one missing person case and contains information about first police impressions during the moment of finding the body, including an initial description of the body context and situation; and medical information from the autopsy and related analysis. Reports were written on paper by the doctors in charge and signed by two doctors. Reports usually follow a narrative schema that includes body location, person profile (gender, age, skin color,

etc.), primary and secondary hypothesis about cause of death, identification information (personal belongings, body marks, etc.), and administrative information (responsible institutions, doctors and police officers involved in each case).

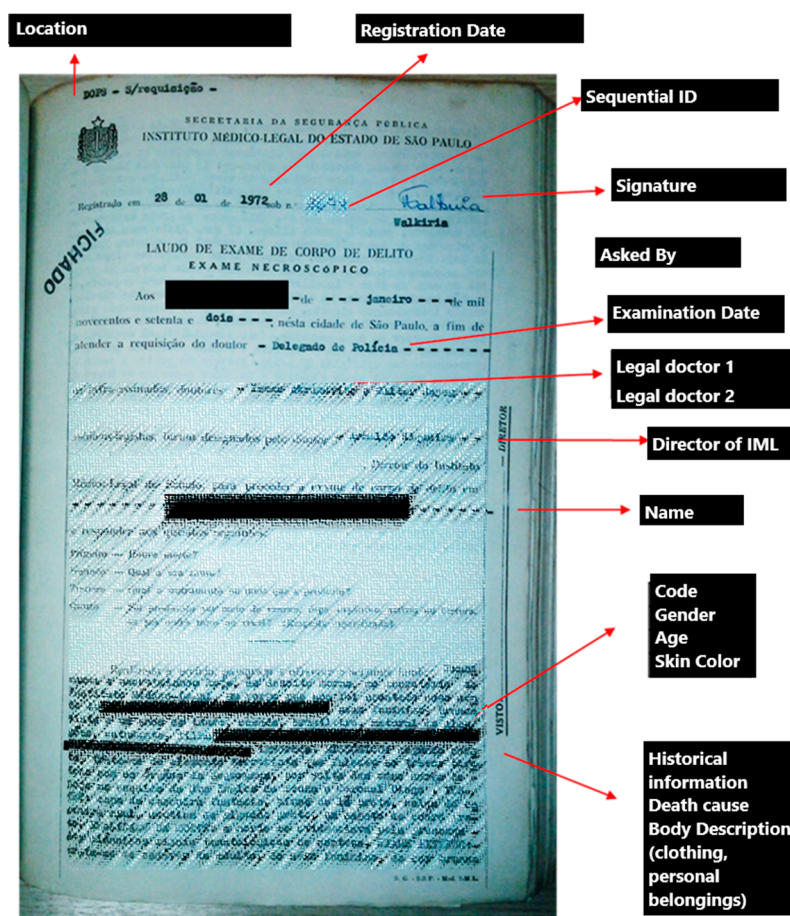


Figure 1. Sample of a forensic report (blurred to protect personal details of a missing person) in the “Disappearances in São Paulo” corpus.

One of biggest challenges from these reports is the non-systematic way in which some aspects of the reports are described. For example, the place of death is very ambiguous in some cases, sometimes appearing with just street, road, river, but not with the exact place. The same happens for clothes descriptions or death causes.

Due to the high volume of reports and the mentioned heterogeneity, the first step in our work was to construct a conceptual model of the major concepts and relationships underlying the corpus. Figure 2 shows the conceptual model (using the object-oriented paradigm [23–25]) that was created to represent the major concepts, properties, and relationships that are mentioned in the corpus. The model is expressed in ConML [26,27], a conceptual modeling language that allows for expressing models in humanities and social sciences domains [28–30]. We chose ConML for two major reasons. Firstly, it is one of the very few modelling languages that focuses on conceptual modelling rather than the specification of software systems (such as UML [31]), while maintaining a significant degree of formalization, which makes it a good candidate for software system implementations. Secondly, ConML incorporates the capacity to represent “soft” issues such as temporality, subjectivity, and vagueness, which are especially relevant to the humanities, whereas other modelling languages lack this ability. The model contains two main categories: Person and institution (modelled as classes in ConML terminology), as well as the actors involved in the reports and their possible roles, including direction roles of institutions and labor relations over time. Also, the necessary attributes have been

added to cater for the disappeared persons profile, including the location where the body was found, the cause of death, a legal medicine description about contextual information, police department in charge, initial evidences, and physical description (clothes, personal objects, etc.). Once the conceptual model was ready, a data model was obtained to prepare the necessary software storage for the textual information extracted from the reports.

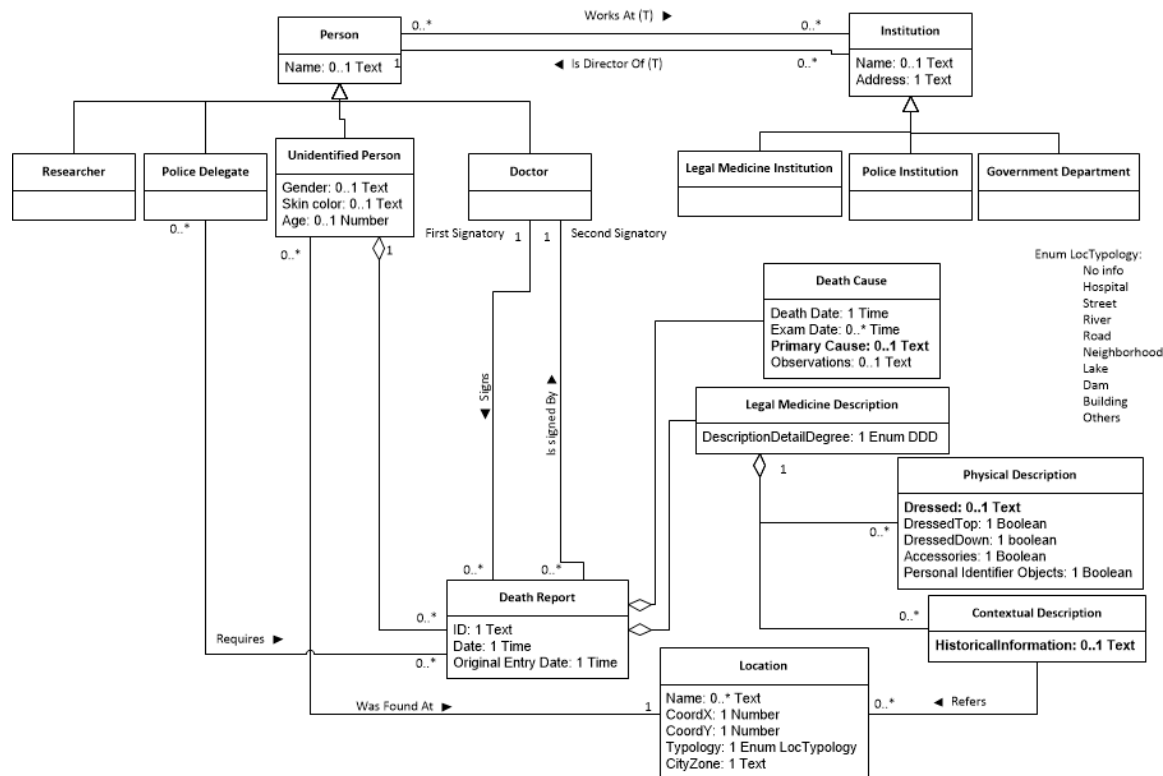


Figure 2. Conceptual model representing the major concepts, properties, and relationships that are mentioned in the corpus.

Then, we designed a pipeline architecture to apply information extraction algorithms to the selected corpus, aiming to extract the information stipulated by the data model and without losing sight of the four criteria defined by the experts that guided the analysis. In particular, we found that having an explicit conceptual model greatly helped us to focus the extraction and analysis efforts around the necessary concepts, properties, and relationships, acting as a set of guidelines that established what semantic fields we were interested in, what kinds of characterization was expected for each one, and what connections were expected to appear. In the next sections, we detail this NLP pipeline and the subsequent analysis of results.

2.2. Unsupervised Information Extraction in Portuguese: Linguakit Suite

As well as the rest of NLP tasks and algorithms, the development of methods and resources for Portuguese are increasing day by day. Some important examples are HAREM and Second HAREM [32], Linguakit [13], or SIEMÉS [33] algorithms and resources for unsupervised named entity recognition, joint with well-known suites such as FreeLing [34] or Stanford CoreNLP [35] for Portuguese and related supervised initiatives based on conditional random fields [36]. It is important to mention here similar works only focused on semantic relation extraction [37].

Most of these NLP resources and algorithms developed for Portuguese and based on machine learning methods require a huge effort to annotate and establish a training set, plus the subsequent training phase. This dependence on annotating information constituted our first handicap in using neural-based algorithms for the NLP analysis, since the contextual annotation of forensic corpus

is extremely time-consuming, difficult to obtain due to the needed for experts in the field and in contact with the bodies and the locations, and highly dependent on the annotator for some of the parts. While the medical information (such as cause of death) usually presents a greater degree of agreement and objectivity, descriptions such as previous situation of the body, clothes and personal belongings, or socio-cultural descriptions of the person present a greater degree of variation among annotators. Precisely, it is these more variable details that are especially relevant in terms of a possible identification of the person or detection of negligent forensic practices. Thus, unsupervised methods offer us the best alternative in our case, as they remove the need for human annotators. Specifically, the *Linguakit* suite was evaluated and compared to supervised approaches [38,39] for the four analyzed languages (English, Portuguese, Spanish, and Galician), using different corpora, with results close to those of FreeLing and Stanford CoreNLP, and clearly surpassing OpenNLP. Thus, we determined that *Linguakit* constituted a good candidate for the necessary linguistic analysis and information extraction. The *Linguakit* suite presents a modular structure that is depicted in Figure 3, highlighting the sub-modules that were used in this work.

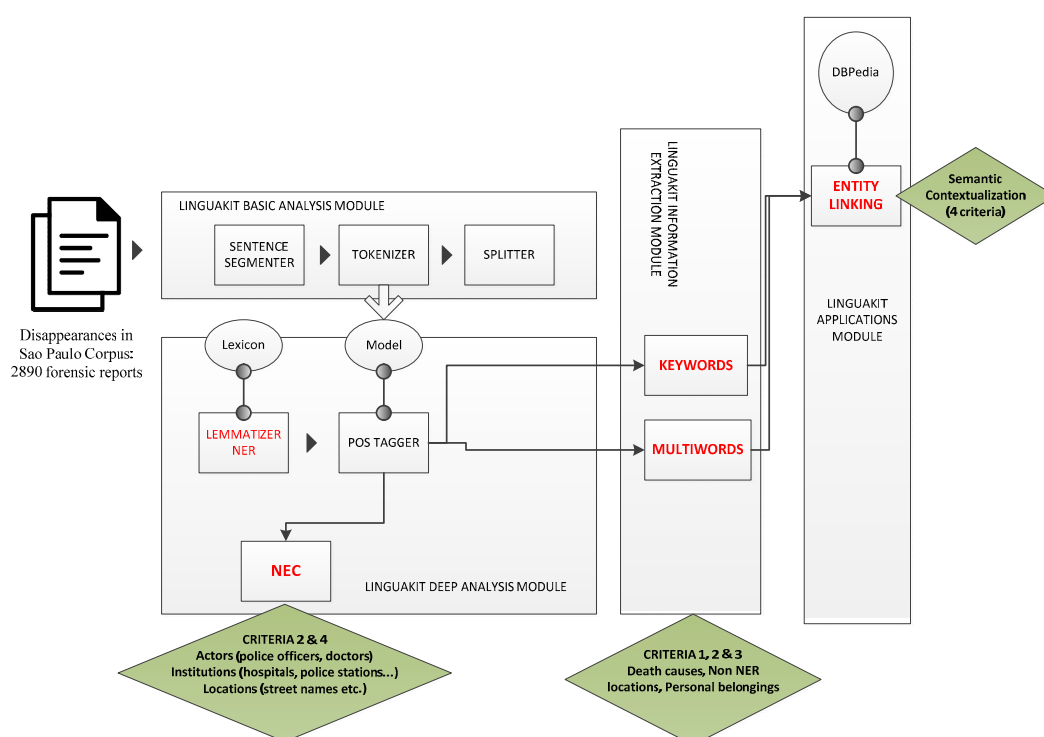


Figure 3. Linguakit modules used for building the natural language processing (NLP) system, and the sequential pipeline and interdependencies between modules output and criteria of analysis taken.

Firstly, we treated the input reports from the corpus with some tools of the basic analysis module. This module is highly dependent on the language of the source. Due to the specialization of *Linguakit* in Portuguese (also Galician and Spanish), this module presents a very high performance on text segmentation into sentences, subsequent tokenization process, and splitting (with basic rules of splitting, such as separation in case of contractions) for Portuguese. Based on these previous outputs, the deep analysis module was used for its lemmatizer. Then, data were disambiguated by the PoS tagger. This disambiguation process allowed the identification of proper names through named entity recognition (NER) and their subsequent named entity classification (NEC).

The joint output between the NER entities recognized and their classification in NEC offered the first valid output, making it possible to achieve some of our goals related to extraction information from the forensics reports and according to expert criteria: We had a list of the named entities present in each report, especially actors (doctors and police officers), institutions (mainly hospitals names and

police offices), and some locations such as street names. This information helped us address criteria 2 (relevant body locations) and 4 (correlations between actors) defined by the experts.

However, this first analysis was not enough to obtain non-named entities that are however relevant to the experts, and which often appear in narrative sections of the forensic reports. To address this, we carried out a keyword and multiword combined analysis, with a filter on semantic information for three topic areas: Death causes, personal belongings found on the bodies, and non-NER locations (i.e., words that refer to places or locations without a proper name but that are potentially relevant, such as hill, river, etc.).

This analysis responds mainly to criteria 1 (causes of death), 2 (relevant body locations), and 3 (personal belongings terminology) of the experts. We used extractors of relevant terms (keywords and multiword expressions) from the information extraction module and an application based on them, the semantic annotator or entity linking (hereafter EL). The final application allowed us to link the terms mentioned in the text with the concepts of the ontological and encyclopedic database DBpedia. *Linguakit* Entity Linking module was evaluated for Portuguese language on this task [40], giving rise to similar results to state-of-the-art EL systems, such as DBpedia Spotlight [41]. The final use of the EL module responded to the need of having a link between the information extracted from the reports to a reference ontology for future applications in the same domain or for a re-analysis of the current corpus or their future versions.

Note that, due to the modular architecture of *Linguakit*, we obtained an output from each submodule used. In the case of NER and NEC outputs, the list of named entities identified and their classification is directly usable for our middleware system to classify, conceptualize (following the ConML model), and visualize the information of the corpus. In the case of the final part of the pipeline, based on focalized keywords, multiword, and their semantic links, the middleware layer of the systems acts as an aggregator of the information extracted from the three modules for each forensic report. We describe the final outputs and their interpretation in the Results section.

3. Results

This section describes the results of the forensic corpus analysis, organized in two parts. First, we show the results corresponding to the algorithmic outputs of the pipeline, organized by expert criteria, also including additional information about grammatical properties and the study of frequencies by output. Subsequently, these outputs are visualized through data analytics dashboards, in order to better assist forensic experts in the global analysis of the corpus.

3.1. Results Analysis

We organized the analysis of the results following the criteria defined by experts. Regarding criterion 1, the experts need assistance on extracting common causes of death, the terminology employed for their description, and links to the specific set of reports related to each family of terms. Thus, we performed a combined keywords and multiword analysis for the entire corpus (2890 reports), looking in the *Histórico* and *Local da morte* part of the reports.

This process produced two sets of words about cause of death terminology: 100 keywords and 665 multiword combinations. Figure 4 shows the first 60 entries for each list. The system combined the entries in function of the relevance ranking obtained from *Linguakit* and offered the ranked terms to the expert. The resultant ranked list provides the experts with an anchoring terminology link to the semantic components of the death causes, which, combined with the DBpedia entity linking analysis, constitutes a first ontological reference for the underlying death causes associated with the corpus.

Regarding criterion 2, software assistance focused on identifying relevant body locations. The information extraction process for this criterion was divided into two different parts: On the one hand, locations were found as named entities, such as names of hospitals (i.e., *Hospital de as Clínicas* or *Instituto Paulista*) offered by *Linguakit* NEC. On the other hand, we performed an analysis of the *Historico* part of the report (that inherits information from the initial location for each body) by

looking up semantic links of the associated entities in DBpedia corresponding to locations terminology (for example *rio Tietê*). Figure 5 shows part of both outputs obtained for criterion 2 analysis.

Regarding criterion 3, the extraction and characterization of personal belongings is performed from the *Vestes* part of the reports. We performed a combined keywords and multiword analysis for the entire corpus (2890 reports). This process resulted in two sets of words about clothes and personal belongings (hats, bags, rings, etc.) terminology: 100 keywords and 1762 multiword combinations were extracted.

```

OutputKEY_LK_CausaCAA.txt
1 instrumento 60186.3069407659 N
2 contundente 58607.856492402 A
3 traumático 34132.4894475608 A
4 encefálico 22646.0392521293 A
5 agente 21869.9678359951 N
6 crânio 21208.0766491386 N
7 traumatismo 17709.6700137893 N
8 broncopneumonia 14375.0816061176 N
9 choque 12179.1293324875 N
10 hemorragia 10423.9111725604 N
11 corpo 9902.00552868766 N
12 lesão 8876.06533898282 N
13 mecânico 6755.49760770354 A
14 asfixia 6581.71764832343 N
15 putrefação 5903.63861726946 N
16 mortis 5659.31880663923 A
17 cirrose 5290.13864084445 N
18 projétil 4193.4556854629 A
19 afogamento 4177.23135690063 N
20 crânio-encefálico 4152.73866926092 N
21 agudo 4152.73866926092 N
22 hepático 3993.44284314732 A
23 toxemia 3908.43015921692 N
24 impossível 3508.66909909279 A
25 agudo 3490.62263798679 A
26 perfuro 3134.83782709087 N
27 interno 3082.5559290465 A
28 hemorrágico 2931.26332685058 A
29 edema 2849.83405577017 N
30 esmagamento 2726.10827711274 N
31 pulmonar 2282.53816733003 A
32 anemia 2279.83762229358 N
33 cranio 2239.12415811351 N
34 avançado 2017.51975726198 N
35 perfuro-contundente 1994.84495999334 N

OutputMWE_LK_CausaCAA.txt
1 corpo contundente 42940.7372343738 N-A
2 choque traumático 13770.1710287002 N-A
3 agente contundente 9751.20927890915 N-A
4 arma de fogo 6674.45969884161 N-P-N
5 instrumento contundente 4891.94368696638 N-A
6 avançado estado 4008.5570830901 N-N
7 crânio encefálico 3979.12517422805 N-A
8 instrumento agente 3430.38807600018 N-N
9 causa mortis 3316.74125288473 N-A
10 cirrose hepática 2847.6284960553 N-A
11 traumatismo crânio-encefálico 2441.5199404084 N-N
12 crânio encefálico 1334.83487664949 N-A
13 hemorragia interna 1236.96753839661 N-A
14 morte natural 1054.03759302752 N-A
15 infarto de miocárdio 996.603521819731 N-P-N
16 traumatismo crânio 601.316760029102 N-N
17 tuberculose pulmonar 491.610360385334 N-A
18 traumáticas crânio-encefálicas 374.934182585922 A-N
19 lesões traumáticas 370.639989823236 N-A
20 lesões crânio 335.60435581757 N-N
21 decurso de tratamento 334.729018240853 N-P-N
22 asfixia mecânica 304.21364684532 N-A
23 Asfixia mecânica 290.639590335853 N-A
24 edema agudo 283.08588735629 N-A
25 choque hemorrágico 280.864224345938 N-A
26 perfuro contundente 280.169008693956 N-A
27 toxemia por broncopneumonia 238.607756869468 N-P-N
28 traumatismos múltiplos 214.873389328687 N-A
29 anemia aguda 166.98030900808 N-A
30 estado de putrefação 144.062846220215 N-P-N
31 broncopneumonia bilateral 136.922334517029 N-A
32 insuficiência cardíaca 131.883115563997 N-A
33 fraturas múltiplas 131.013850371117 N-A
34 origem traumática 129.70735497581 N-A
35 crânio encefálico 119.375724347554 N-A

```

Figure 4. Linguakit output for keywords and multiword analysis of causes of death in the ‘Disappearances in São Paulo’ corpus.

```

OutputKEY_LK_HistoricoCAA.txt
1 atropelamento 15434.2617146319 N
2 falecer 14795.4246462759 V
3 setenta 11572.3132917912 N
4 vítima 11115.0452417069 N
5 vinte 10973.8819257615 N
6 trintar 4088.14235659327 V
7 Hospital de as Clínicas 3196.85479766048 ENTITY
8 encontrado 2871.61926410839 A
9 quinze 2552.81890750407 N
10 encontrar 2177.14246798081 V
11 trem 2020.85010195352 N
12 dezenove 2000.81430300758 N
13 trinta 1885.8163828278 N
14 dezessete 1862.81692454387 N
15 atropelada 1769.13957282749 N
16 falecido 1685.77817524101 N
17 atropelado 1655.82368624305 N
18 vir 1640.51819567456 V
19 quarenta 1609.82564992228 N
20 internado 1586.82669463567 N
21 dezesseis 1540.82890809043 N
22 boiar 1519.64301227716 V
23 doze 1494.83129264507 N
24 Instituto Paulista 1471.83254693581 ENTITY
25 putrefação 1425.83518126255 N
26 dezoito 1356.8394471141 N
27 minuto 1355.979271145 N
28 S 1333.84095280938 ENTITY
29 noventa 1310.84250058009 N
30 auto 1284.39925717137 N
31 P. 1264.84572171937 ENTITY
32 morte 1189.66799265759 N
33 onze 1172.85266695016 N
34 suspeito 1147.72953568067 A
35 treze 1126.85639104782 N

OutputMWE_LK_HistoricoCAA.txt
1 P. S 3687.540773844 N-N
2 morte natural 3363.91829915585 N-A
3 via pública 2453.54545605325 N-A
4 vítima de atropelamento 1636.62775529301 N-P-N
5 terreno baldio 1562.96454215445 N-A
6 adiantado estado 799.072140224355 N-N
7 arma de fogo 687.210605143125 N-P-N
8 queda acidental 663.266245631773 N-A
9 estado de putrefação 504.268968430195 N-P-N
10 morte suspeita 335.227435234062 N-A
11 encontro de cadáver 309.612032421087 N-P-N
12 rio Tietê 281.232617039862 N-N
13 requisição de exame 215.583501186411 N-P-N
14 corrente ano 170.625668462442 A-N
15 dia vinte 154.73763254214 N-N
16 dezenove horas 148.995809407819 N-N
17 agosto de setenta 140.243620788375 N-P-N
18 cinquenta minutos 116.824419829329 A-N
19 maio de setenta 115.588691181061 N-P-N
20 ano em curso 112.246410275212 N-P-N
21 dezessete horas 100.122840054488 N-N
22 onze horas 99.9488521846006 N-N
23 dezoito horas 97.4530320079642 N-N
24 distrito policial 96.6329011134617 N-A
25 atropelamento por auto 93.4564531640244 N-P-N
26 julho de setenta 92.1939750188382 N-P-N
27 treze horas 90.24983669112293 N-N
28 novembro de setenta 87.6050105962535 N-P-N
29 duas horas 84.7819324823095 A-N
30 quinze horas 80.9920711361392 N-N
31 vítima de agressão 80.6123075470239 N-P-N
32 Instituto paulista 80.4575530010187 N-N
33 vítima encontrado 79.922243872747 N-A
34 setembro de setenta 77.1869207129654 N-P-N
35 Santo Amaro 71.9674575698542 N-N

```

Figure 5. Linguakit output for keywords and multiword analysis of historical information (*Histórico*) in the ‘Disappearances in São Paulo’ corpus.

Figure 6 shows the first 40 entries of each list. The system combined the entries in function of the relevance ranking obtained from *Linguakit* and offered the ranked terms to the expert. In a first approach, the experts envisioned that the personal belongings analysis would offer them some clues towards identification of bodies, such as special materials, wedding rings, etc. They defined a subset of them that could provide these clues. However, in most cases, the experts realized that the terms extracted by this analysis are also useful for an additional purpose: The description of the belongings are very poor in most of the reports (with generalist terms such as “algodão”, “calça”, etc.) with an absence of description about gender-specific clothes and belongings, sizes, specific textures, materials, or print patterns on clothes. This is very unusual and goes clearly against good practices in forensic descriptions, especially in cases that deal with someone who is going to be buried unidentified (NN). In these cases, one of the most important aspects for recognition by relatives and later the identification is the description of personal belongings. Thanks to our work, the experts decided to perform a cross-analysis confronting which reports presented these unusual characteristics and who signed each one, in order to look for patterns on authorship by doctors that signed negligently described reports. This cross-analysis offers some responses regarding possible implications of doctors with unresolved disappearances and negligent forensic practices, with the discovering of pairs of co-working doctors that is necessary to deeply analyze in the near future.

```

OutputKEY_LK_VesteesCAA.txt
1 calça 59893.4422883658 N
2 algodão 50881.5756034645 N
3 camisa 36010.4406995547 N
4 azul 33202.1038157351 A
5 preto 23950.9491618104 A
6 branco 21983.5381425983 A
7 blusa 20484.8430061231 N
8 fantasia 19957.2980465045 N
9 despir 19404.2897344712 V
10 lá 13959.1453259701 N
11 cinza 13901.2687619382 A
12 marrom 13126.8935351192 A
13 sapato 12929.8416660056 N
14 casemira 9868.44022955578 N
15 paletó 9287.78952572543 N
16 couro 9247.22802933027 N
17 meia 8628.83511619914 N
18 brim 7578.20775709259 N
19 cueca 6900.87253791643 N
20 verde 6734.61614716499 A
21 vermelho 5953.85865914857 A
22 tergal 5707.53623346304 N
23 lençol 5173.12422335139 N
24 cinza 4811.73811802635 N
25 amarelo 3572.84260254704 A
26 saia 2941.66772623862 N
27 marrom 2740.67466341801 A
28 short 2643.93765399902 N
29 vestido 2449.62670578997 N
30 rosa 2448.90902990153 A
31 preta 2289.23986599861 N
32 nylon 2289.23986599861 N
33 cor 2165.12379483732 N
34 malha 2031.02053238839 N
35 cinto 1907.74754695579 N

OutputMWE_LK_VesteesCAA.txt
1 calça calça 4195.6493717294 N-N
2 azul de algodão 1705.69812287236 N-P-N
3 blusa algodão 990.305912291943 N-N
4 algodão cinza 924.060052893613 N-N
5 blusa de lá 715.074294805872 N-P-N
6 camisa de fantasia 577.871387235087 N-P-N
7 camisa de algodão 561.25006419873 N-P-N
8 algodão marrom 554.38450002474 N-N
9 algodão branco 502.429004722446 N-A
10 sapatos pretos 470.579002454919 N-A
11 sapatos de couro 459.073682297965 N-P-N
12 algodão branco 442.219052375251 N-N
13 couro preto 411.263958122978 N-A
14 calça de casemira 396.503904396624 N-P-N
15 calça de brim 319.17123103471 N-P-N
16 branca algodão 293.112908675495 A-N
17 azul fantasia 280.94236561137 N-N
18 calção algodão 266.258105820573 N-N
19 algodão fantasia 258.72983237682 N-N
20 algodão preto 255.877332491048 N-N
21 meias pretas 231.378204676027 N-A
22 brim azul 230.731885692985 N-A
23 calça comprida 229.004983343883 N-A
24 azul branco 219.173843145337 A-N
25 lençol branco 218.573818779603 N-A
26 calça de lá 175.50707024697 N-P-N
27 blusa de fantasia 164.119850794976 N-P-N
28 caixa de papelão 155.86675868252 N-P-N
29 cinza fantasia 152.65630679081 N-N
30 pé de sapato 152.383114490999 N-P-N
31 calça de paletó 151.861302188782 N-P-N
32 paletó de calça 151.861302188782 N-P-N
33 algodão azul 148.421433158456 N-A
34 calça fantasia 143.2137490462 N-N
35 calção de algodão 140.445261664753 N-P-N
  
```

Figure 6. *Linguakit* output for keywords and multiword analysis of body description: Clothes and personal belongings (*Vestees*) in the ‘Disappearances in São Paulo’ corpus.

Finally, criterion 4 responds to the correlation of doctors’ analysis. Thus, we took the NEC output of *Linguakit* that automatically classify PERSONS and automatically checked the names of the doctors from a list that the experts had. Thus, we established a matrix of signatory doctors for each report, thus having access to information about how many reports each pair of doctors had signed together, in what year, and working for which organizations.

The next section shows the visualization obtained from the outputs, which we consider to be prototypes for a future system that will allow experts to analyze and interpret the obtained information more deeply.

3.2. Visualizing Results in an Interactive Dashboard

It is common that the results provided by NLP algorithms presents an output format in free text or, in some cases, in a textual-based structure output stored in formats such as XML, JSON, or similar. Although most of the current software systems can process these formats in an efficient way, it is necessary to remember here that the final users of the information contained on the corpus under use are forensic experts from different background with knowledge generation necessities, but with no NLP training whatsoever. In other words, although the volume and relevance of the information extracted through NLP are highly relevant, future software-assisted knowledge generation systems will require certain visualization features to facilitate reasoning and interpretation by the experts, and without which the previous efforts of analysis would be most likely wasted.

As a prototype phase to explore the possibilities of visualization and software assistance of the outputs obtained through the analysis with *Linguakit*, we have visualized the information in different dashboards of Google Data Studio, one per forensic criterion. Dashboards were created to visualize:

- The information extracted by NEC and categorized as PERSON and their correlation matrix (see Figure 7, top). Here, each row and/or column represents a doctor that signed forensic reports. Each cell specifies the number of reports signed together. Colored cells in reddish tones show emergent islands of intensive cooperation between the same couple of doctors in negligently described reports. This has allowed the experts to focus their study on certain doctors and their cases as possible collaborators of the regime's practices.
- The information extracted in the studies of keywords and multiword combinations has been visualized as word clouds of separate terms for each of the criteria: Common causes of death, terminology related to clothing, and terminology related to locations that do not respond to proper names (see Figure 7, bottom left).
- Finally, the NEC information about proper names of locations has been displayed by the number of related reports (such as police stations responsible for requesting each report) (see Figure 7, bottom right) and the terms extracted from the reports. The information extracted about locations that do not correspond to proper names have also been visualized in a similar way (see Figure 8).

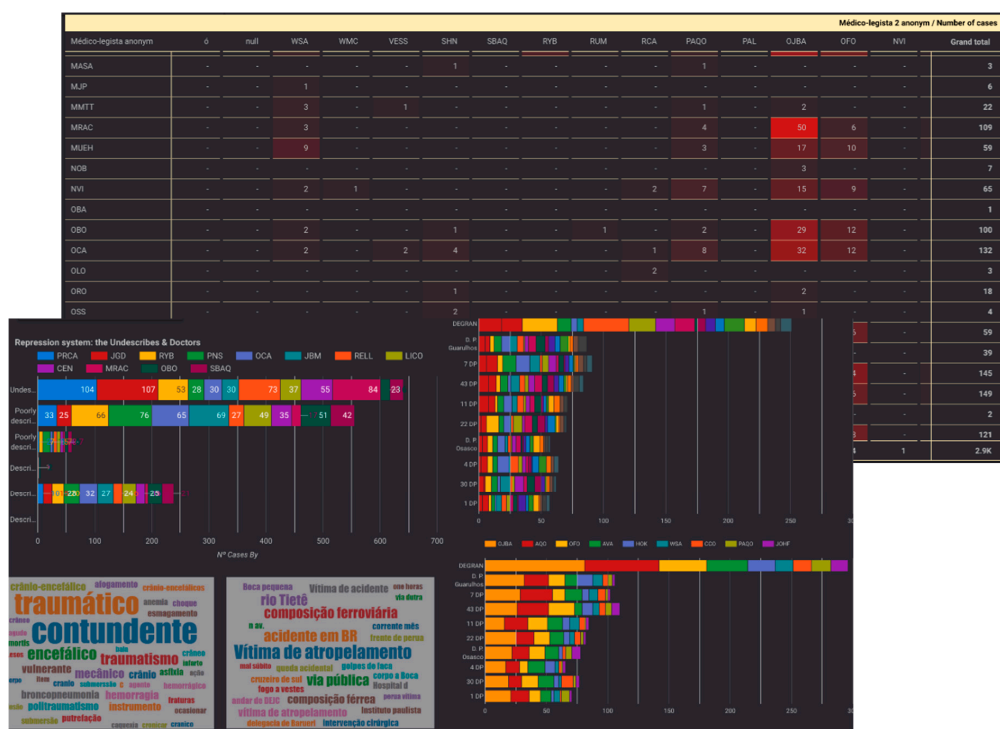


Figure 7. Google Data Studio dashboards created as prototypes to visualize NLP results.

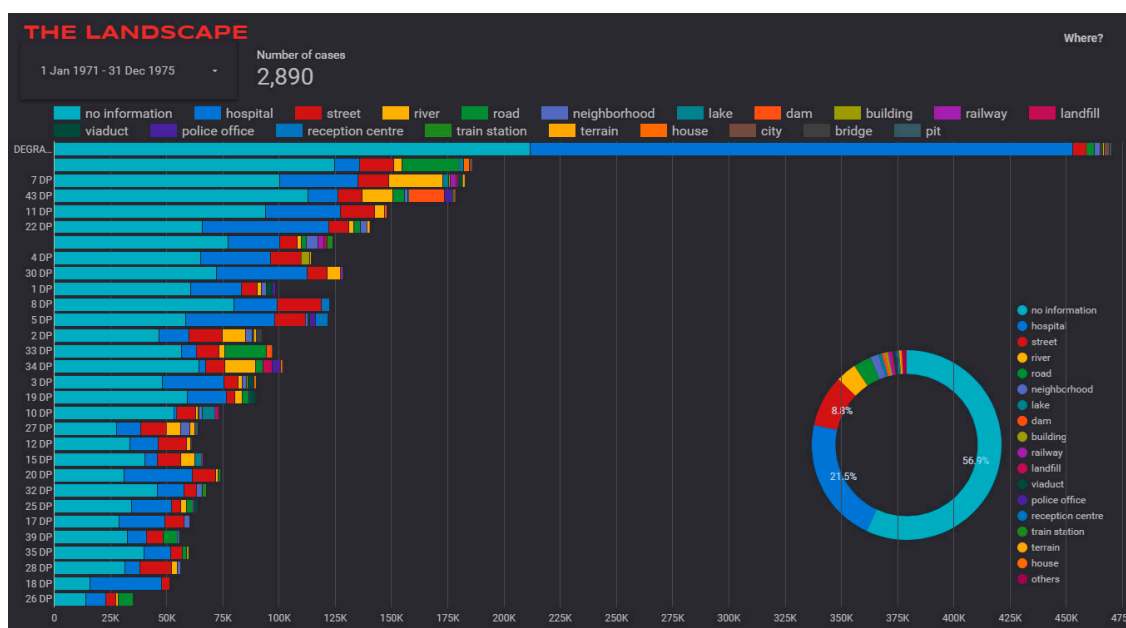


Figure 8. Google Data Studio dashboard about locations, created as a prototype to visualize NLP results.

The design and choice of the components of the dashboards was always carried out through the collaboration of forensics experts and software engineers, although it is limited by the technological possibilities of the Google Data Studio tool. We want to remark that any other visual analytics tool would most likely present limitations too. However, these dashboards have allowed us to observe the advantages of the NLP analysis of information extraction and its future possibilities, as well as identify some visualization requirements for this kind of forensic data.

4. Discussion

From the forensic, anthropological, and archaeological point of view, the NLP study helped us to see how some clues about the repression practices performed in São Paulo emerged from the textual reports, further reflecting on the action of repression on bodies “that do not matter” [42] and their identity. It was possible to observe mechanisms of disappearance in operation formed by the steps of an institutional and bureaucratic chain that makes it difficult to trace bodies, perpetuating their concealment (reflected, for instance, in the negligent description practices detected). This, in turn, produces the loss of identity of the citizens and camouflages the circumstances in which the deaths took place (with lack of several proper names of locations and generalized use of common terminology, i.e., river, hill, etc. for referring specific locations). NER and NEC mechanisms also constitute an efficient way to automate some analyses about the actors involved in the conflict, allowing us to easily obtain the resultant matrix of correlations in the case of doctors, which constitutes a valuable output for the current investigation of the unsolved disappearances cases.

Regarding the technological implication of the study, we are conscious that the approach described in this paper is dependent on the structure of the reports in the corpus (as reflected by the case-specific conceptual model), and heavily guided by the criteria previously defined by experts. We are also conscious that, from the natural language point of view, it is possible to go beyond the presented work and, for example, continue working on the sematic links or keep exploring the possibilities of relationship extraction. In addition, it is necessary to critically discuss the difficulty to properly evaluate this kind of NLP application. The reason for this is twofold: On the one hand, we can evaluate the quality of the information extraction algorithms in this specific domain and for the Portuguese language; *Linguakit* has been evaluated with results similar to those of the state of the art for all the used submodules, but there are no gold standards for the forensic domain. On the other hand,

the proposed approach is a mix of NLP and visualization for assisted knowledge generation, which constitutes a complex scenario to help researchers and domain experts to perform cognitive inferences from (especially unstructured) data. Systems like this are usually called software-assisted knowledge generation systems [43,44], and have, as one of their main characteristics, the difficulty to validate their level of achievement and quality, due to the cognitive nature of their functionality. Thus, it is common to perform empirical validations [45–48] involving final users (i.e., domain experts) in order to extract some evaluation information. We plan this kind of evaluation with the domain experts as a next step.

5. Conclusions and Future Work

In summary, it is difficult to find in the literature, as far as we know, software applications to forensic corpus analysis using natural language processing techniques, especially of non-English sources. This paper presents an innovative application of NLP to forensic sources in Portuguese. In addition, it represents one of the most complete approaches for real use of NLP in the forensics domain, constituting not only an application of NLP algorithms, but also including a conceptualization of the information that is necessary to extract and a proposal for visualizing the outputs. Thus, the main contributions are the exemplification of a real-world application in the forensics domain, with valuable results for the forensics experts, as well as an automatic method (through the pipeline) to extract and study correlations between actors in the forensics domain. The pipeline provided could be applied to similar forensic corpus regardless of the underlying source language.

In the future, we will continue our work on NLP in the forensics domain. This will imply a detailed analysis of the generalization of the pipeline to other similar and/or related corpora and source languages.

Also, a comparison is needed between NLP systems similar to *Linguakit* in general tests (such as the Stanford suite or Freeling), in order to check whether the presented pipeline offers comparable results when the conceptualized domain and the textual sources are focused only on forensic discipline and corpora. Finally, and as we have indicated above, the empirical evaluation of the system by experts (once several corpora have been analyzed) will also constitute an area of great value and interest.

Author Contributions: Conceptualization, M.L.H. and C.G.-P.; Data curation, M.L.H.; Formal analysis, P.M.-R. and M.L.H.; Funding acquisition, P.M.-R. and M.L.H.; Investigation, P.M.-R. and C.G.-P.; Methodology, P.M.-R. and M.L.H.; Project administration, P.M.-R.; Supervision, P.M.-R. and C.G.-P.; Validation, M.L.H.; Visualization, P.M.-R.; Writing – original draft, P.M.-R. and M.L.H.; Writing – review & editing, P.M.-R. and C.G.-P.

Funding: This research was partially funded by Spanish Ministry of Economy, Industry and 5 Competitiveness under its Competitive Juan de la Cierva Postdoctoral Research Programme, grant FJCI-2016-6 28032 and from the European Union, through the Marie Skłodowska-Curie Innovative Training Network ‘CHEurope: Critical Heritage Studies and the Future of Europe’ H2020 Marie Skłodowska-Curie Actions, grant 722416.

Acknowledgments: The authors want to thank here the data, experiences and research hypothesis sharing of the researchers who work in the forensic study and identification of people during the Brazilian military dictatorship, and to The Special Commission for Death and Disappeared People (*Comissão Especial sobre Mortos e Desaparecidos Políticos*) to allowed this work. Thanks also to *Linguakit* contributors for their algorithmic support.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

- Ogren, P.V.; Savova, G.K.; Chute, C.G. Constructing evaluation corpora for automated clinical named entity recognition. In *Building Sustainable Health Systems, Proceedings of the Medinfo 2007: 12th World Congress on Health (Medical) Informatics, Brisbane, Australia, 20–24 August 2007*; IOS Press: Amsterdam, The Netherlands, 2007.
- Neamatullah, I.; Douglass, M.M.; Lehman, L.H.; Reisner, A.; Villarroel, M.; Long, W.J.; Szolovits, P.; Moody, G.B.; Mark, R.G.; Clifford, G.D. Automated de-identification of free-text medical records. *BMC Med. Inform. Decis. Mak.* **2008**, *8*, 32. [[CrossRef](#)] [[PubMed](#)]

3. Uzuner, O.; Solti, I.; Xia, F.; Cadag, E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 519–523. [[CrossRef](#)] [[PubMed](#)]
4. Deleger, L.; Li, Q.; Lingren, T.; Kaiser, M.; Molnar, K.; Stoutenborough, L. Building gold standard corpora for medical natural language processing tasks. In Proceedings of the AMIA Annual Symposium Proceedings, Chicago, IL, USA, 3–7 November 2012.
5. Torii, M.; Waghlikar, K.; Liu, H. Using machine learning for concept extraction on clinical documents from multiple data sources. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 580–587. [[CrossRef](#)] [[PubMed](#)]
6. Bodnari, A.; Deléger, L.; Lavergne, T.; Névéol, A.; Zweigenbaum, P. A Supervised Named-Entity Extraction System for Medical Text. In Proceedings of the CLEF, Valencia, Spain, 23–26 September 2013.
7. Jiang, M.; Chen, Y.; Liu, M.; Rosenbloom, S.T.; Mani, S.; Denny, J.C.; Xu, H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 601–606. [[CrossRef](#)] [[PubMed](#)]
8. Teles, M.A.D.A.; Lisboa, S.K. A Vala de Perus: Um Marco Histórico na Busca da Verdade e da Justiça. In *Vala Clandestina de Perus: Desaparecidos políticos, um Capítulo Não Encerrado da História Brasileira*; Instituto Macuco: São Paulo, Brazil, 2012; pp. 51–102. (In Portuguese)
9. Teles, J. *Mortos e desaparecidos políticos: Reparação ou impunidade*; Humanitas FFLCH/USP: São Paulo, Brazil, 2001. (In Portuguese)
10. Somigliana, C. Apuntes sobre la importancia de la actuación del Estado burocrático durante el período de la desaparición forzada de personas en la Argentina. *Taller Rev. Soc. C. Y Política* **2000**, *5*, 9–19. (In Spanish)
11. Crenzel, E.A. Otra literatura: Los registros burocráticos y las huellas de las desapariciones en la Argentina. *Estudios Teor. Lit.* **2014**, *3*, 29–42. (In Spanish)
12. Hattori, M.L.; de Abreu, R.; Tauhyl, S.A.P.M.; Alberto, L.A. O caminho burocrático da morte e a máquina de fazer desaparecer: Propostas de análise da documentação do Instituto Médico Legal-SP para antropologia forense1 2. *Rev. Do Arq.* **2014**, *6*, 1–21. (In Portuguese)
13. Gamallo, P.; Garcia, M.; Pineiro, C.; Martinez-Castaño, R.; Pichel, J.C. LinguaKit: A Big Data-based multilingual tool for linguistic analysis and information extraction. In Proceedings of the 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 15–18 October 2018.
14. Google Data Studio 2019. Available online: <https://datastudio.google.com/> (accessed on 5 July 2019).
15. Mezarobba, G. Entre reparações, meias verdades e impunidade: O difícil rompimento com o legado da ditadura no Brasil. *Rev. Int. Direitos Hum.* **2010**, *7*, 7–26. (In Portuguese)
16. Ministério da Justiça e Segurança Pública—Sobre a comissão. 2019. Available online: <https://www.justica.gov.br/seus-direitos/anistia/sobre-a-comissao/sobre-a-comissao> (accessed on 25 May 2019). (In Portuguese)
17. Comissão Nacional da Verdade (CNV). *Relatório Final da Comissão Nacional da Verdade*; DF. 3350 ISBN; Comissão Nacional da Verdade: Brasília, Brazil, 2014. (In Portuguese)
18. Barcellos, C. O Globo Repórter sobre a vala de Perus. In *Mortos e desaparecidos políticos: Reparação ou impunidade*; Humanitas FFLCH/USP: São Paulo, Brazil, 2001; pp. 213–226. (In Portuguese)
19. Godoy, M. *A Casa da Vovó: Uma Biografia do DOI-Codi (1969–1991), O Centro de Sequestro, Tortura E Morte da Ditadura Militar*; Alameda Casa Editorial: São Paulo, Brazil, 2015. (In Portuguese)
20. Asociación Latinoamericana de Antropología Forense. *Guía latinoamericana de buenas prácticas para la aplicación en antropología forense*; ALAF: Antigua, Guatemala, 2016. (In Spanish)
21. Carnaz, G.; Quaresma, P.; Nogueira, V.B.; Antunes, M.; Ferreira, N.N.M.F. A Review on Relations Extraction in Police Reports. In Proceedings of the New Knowledge in Information Systems and Technologies, La Toja, Spain, 6–19 April 2019; Springer: Cham, Switzerland, 2019.
22. Mujtaba, G.; Shuib, L.; Raj, R.G.; Rajandram, R.; Shaikh, K. Automatic Text Classification of ICD-10 Related CoD from Complex and Free Text Forensic Autopsy Reports. In Proceedings of the 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016.
23. Partridge, C. *Business Objects: Re-Engineering for Re-Use*; Butterworth-Heinemann: Oxford, UK, 1996.
24. Armstrong, D.J. The quarks of object-oriented development. *Commun. ACM* **2006**, *49*, 123–128. [[CrossRef](#)]
25. Surya, M.; Padmavathi, S. A Survey of Object-Oriented Programming Languages. Available online: <http://users.soe.ucsc.edu/~vrk/Reports/oopssurvey.pdf> (accessed on 16 December 2014).

26. Gonzalez-Perez, C. A conceptual modelling language for the humanities and social sciences RCIS'12. In Proceedings of the Sixth International Conference on Research Challenges in Information Science, Valencia, Spain, 16–18 May 2012.
27. Gonzalez-Perez, C. *Information Modelling for Archaeology and Anthropology: Software Engineering Principles for Cultural Heritage*; Springer: Berlin, Germany, 2018.
28. Martin-Rodilla, P.; Gonzalez-Perez, C. Assessing the learning curve in archaeological information modelling: Educational experiences with the Mind Maps and Object-Oriented paradigms. In Proceedings of the 45th Computer Applications and Quantitative Methods in Archaeology (CAA 2017), Atlanta, GA, USA, 13–16 March 2017.
29. Gonzalez-Perez, C.; Martin-Rodilla, P. Teaching Conceptual Modelling in Humanities and Social Sciences. *Rev. Humanidades Dig.* **2017**, *1*, 408–416. [CrossRef]
30. Gonzalez-Perez, C.; Martin-Rodilla, P. Using model views to assist with model conformance and extension. In Proceedings of the 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), Grenoble, France, 1–3 June 2016.
31. OMG. UML 2.4.1 Superstructure Specification. August 2012. Available online: <http://www.omg.org/> (accessed on 5 June 2019).
32. Freitas, C.; Mota, C.; Santos, D.; Oliveira, H.G.; Carvalho, P. *Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese*; European Languages Resources Association (ELRA): Valletta, Malta, 2010.
33. Sarmiento, L. SIEMÊS—A Named-Entity Recognizer for Portuguese Relying on Similarity Rules. In *International Workshop on Computational Processing of the Portuguese Language*; Vieira, R., Quesada, P., Nunes, M.D.G.V., Mamede, N.J., Oliveira, C., Dias, M.C., Eds.; Springer: Berlin, Germany, 2006; pp. 90–99.
34. Padró, L.; Stanilovsky, E. Freeling 3.0: Towards wider multilinguality. In Proceedings of the LREC2012, Istanbul, Turkey, 21–27 May 2012.
35. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 22 June 2014.
36. Pirovani, J.; de Oliveira, E. CRF+LG: A Hybrid Approach for the Portuguese Named Entity Recognition. In *Intelligent Systems Design and Applications*; Abraham, A., Muhuri, P.K., Muda, A.K., Gandhi, N., Eds.; Springer: Berlin, Germany, 2018; pp. 102–113.
37. Collovini, S.; Machado, G.; Vieira, R. Extracting and Structuring Open Relations from Portuguese Text. In *Computational Processing of the Portuguese Language*; Silva, J., Ribeiro, R., Quesada, P., Adami, A., Branco, A., Eds.; Springer: Cham, Switzerland, 2016; pp. 153–164.
38. Gamallo, P.; Garcia, M. A resource-based method for named entity extraction and classification. In Proceedings of the Portuguese Conference on Artificial Intelligence, Lisbon, Portugal, 10–13 October 2011; Springer: Berlin, Germany, 2011.
39. Garcia, M.; Gamallo, P. Yet Another Suite of Multilingual NLP Tools. In *Languages, Applications and Technologies, Proceedings of the 4th International Symposium, SLATE 2015, Madrid, Spain, 18–19 June 2015*; Springer: Cham, Switzerland, 2015; pp. 65–75.
40. Gamallo, P.; Garcia, M. Entity Linking with Distributional Semantics. In Proceedings of the 12th International Conference, PROPOR, Tomar, Portugal, 13–15 July 2016; Springer: Berlin, Germany, 2016.
41. Mendes, P.N.; Jakob, M.; Garcia-Silva, A.; Bizer, C. DBpedia spotlight: Shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems, Graz, Austria, 7–9 September 2011.
42. Butler, J. *Frames of War: When is life Grievable?* Verso Books: Brooklyn, NY, USA, 2016.
43. Martin-Rodilla, P. *Digging into Software Knowledge Generation in Cultural Heritage*; Springer: Basel, Switzerland, 2018.
44. Odena, O. Using Specialist Software to Assist Knowledge Generation: An Example from a Study of Practitioners' Perceptions of Music as a Tool for Ethnic Inclusion in Cross-Community Activities in Northern Ireland. In *Advancing Race and Ethnicity in Education*; Race, R., Lander, V., Eds.; Palgrave Macmillan: London, UK, 2014; pp. 178–192.
45. Juristo, N.; Moreno, A.M. *Basics of Software Engineering Experimentation*; Springer: Berlin, Germany, 2013.
46. Martin-Rodilla, P.; Panach, J.I.; Gonzalez-Perez, C.; Pastor, O. Assessing data analysis performance in research contexts: An experiment on accuracy, efficiency, productivity and researchers' satisfaction. *Data Knowl. Eng.* **2018**, *116*, 177–204. [CrossRef]

47. Wohlin, C.; Runeson, P.; Höst, M.; Ohlsson, M.C.; Regnell, B.; Wesslén, A. *Experimentation in Software Engineering*; Springer: Berlin, Germany, 2012.
48. Panach, J.I.; España, S.; Dieste, O.; Pastor, O.; Juristo, N. In search of evidence for model-driven development claims: An experiment on quality, effort, productivity and satisfaction. *Inf. Softw. Technol.* **2015**, *62*, 164–186. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).