Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers

Sophie Rosset, Cyril Grouin, Karen Fort, Olivier Galibert, Juliette Kahn, Pierre Zweigenbaum

Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers

Sophie Rosset[α], **Cyril Grouin**[α], **Karën Fort**[β,γ]
Olivier Galibert[δ], **Juliette Kahn**[δ], **Pierre Zweigenbaum**[α]
[α]LIMSI–CNRS, France [β]INIST–CNRS, France [γ]LIPN, France [δ]LNE, France
{sophie.rosset,cyril.grouin,pierre.zweigenbaum}@limsi.fr
karen.fort@inist.fr, {olivier.galibert,juliette.kahn}@lne.fr

Abstract

This paper compares the reference annotation of structured named entities in two corpora with different origins and properties. It addresses two questions linked to such a comparison. On the one hand, what specific issues were raised by reusing the same annotation scheme on a corpus that differs from the first in terms of media and that predates it by more than a century? On the other hand, what contrasts were observed in the resulting annotations across the two corpora?

1 Introduction

Named Entity Recognition (NER), and its evaluation methods, constitute an active field of research. NER can be performed on many kinds of documents. On textual data, a few NER applications focus on newspapers, spoken data, as well as digitized data. On specific kinds of data such as historical data, various investigations have been performed to detect named entities (Miller et al., 2000; Crane and Jones, 2006; Byrne, 2007; Grover et al., 2008). From the point of view of both annotation and evaluation campaigns, ACE (Doddington et al., 2004) included NER on OCRed data.

For the French language, an evaluation involving classical named entities was performed a few years ago on old newspapers data (Galibert et al., 2010). More recently, we proposed a definition of structured named entities for broadcast news data (Grouin et al., 2011). We follow this definition in this work.

After a presentation of related work (Section 2), including the definition of structured named entities, this paper presents the construction of a new annotated corpus of old newspapers (Section 3). The main goal of the paper is to report the comparison of structured named entity annotation in two contrasting press corpora: the pre-existing broadcast news corpus and this new corpus of old newspapers. This comparison is performed at two levels: the annotation process itself (Section 4.1) and the annotation results (Section 4.2).

2 Related Work

2.1 Named Entity Definition

Initially, Named Entity recognition (NER) was described as recognizing proper names (Coates-Stephens, 1992). Since MUC-6 (Grishman and Sundheim, 1996), named entities include three major classes: *person, location* and *organization*. Some numerical types are also often described and used in the literature: *date, time* and *amount* (money and percentages in most cases).

Proposals were made to sub-divide existing categories into finer-grained classes: *politician* as part of the *person* class (Fleischman and Hovy, 2002) and *city* in the *location* class (Fleischman, 2001). New classes were added during the CONLL conference. More recently, larger extensions were proposed: *product* by (Bick, 2004) while (Sekine, 2004) defined an extensive hierarchy of named entities containing about 200 types. Numerous investigations concern named entities in historical data (Miller et al., 2000; Crane and Jones, 2006; Byrne, 2007;

Grover et al., 2008). In most cases, the definition of named entity follows the classical definition. Nevertheless, in some cases, new categories were added. For example, in the Virginia Banks project (Crane and Jones, 2006), in order to adapt the definition to the studied period (American Civil War) categories such as *ships*, *regiments*, and *railroads* were added.

2.2 Structured Named Entity Definitions

We proposed a new structure of named entities that relies on two main principles: our extended named entities are both hierarchical and compositional. This structure requires novel methods to evaluate system outputs. Compared to existing named entity definitions, our approach is more general than the extensions proposed for specific domains, and is simpler than the extensive hierarchy defined by Sekine (2004). This structure allows us to cover a large number of named entities with a basic categorization which provides a foundation that facilitates further annotation work. The guidelines are available online (Rosset et al., 2011).

2.2.1 Hierarchy

We defined an extended named entity as being composed of two kinds of element: *types* and *components*. In our definition, *types* refer to a general segmentation of the world into major categories. Furthermore, we consider the content of an entity must be structured as well. From this perspective, we defined a second level of annotation for each category we call *components*.

Types and sub-types refer to the general category of a named entity. We defined this type of element as being the first level of annotation because they give general information about the annotated expression. Our taxonomy is thus composed of 7 types (*person, location, organization, amount, time, production* and *function*) and 32 sub-types (individual person *pers.ind* vs. group of persons *pers.coll*; law, decree, and agreement *prod.rule* vs. political, philosophical and religious belief *prod.doctr*, etc.).

Components can be considered as clues to make it possible the annotations: either to determine the type of an extended named entity (a first name is a clue for the individual person *pers.ind* sub-type), or to set the named entity boundaries (a given token is a clue for the named entity, and is within its scope—e.g., a figure in a date—, while the next token is not a clue and is outside its scope—e.g., a word that is not a month or a part of a date).

Components are second-level elements, and can never be used outside the scope of a type or sub-type element. We specified two kinds of components: transverse components that can be included into all types of entities (*name, kind, qualifier, demonym, val, unit, object* and *range-mark*), and specific components, only used for a reduced set of components (for example, *name.last, name.first, name.middle* and *title* for the *pers.ind* sub-type).

2.2.2 Structure

Three kinds of structures can be found in our annotation schema. First, a sub-type contains a component: the *pers.ind* sub-type (individual person) contains components such as *title* and *name.last*, while the *func.ind* sub-type (individual function) contains other components such as *kind* (the kind of function) and *qualifier* (a qualifier adjective) (see Figure 1).

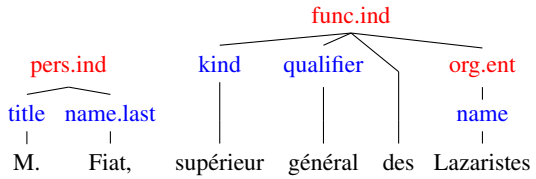

Figure 1: Multi-level annotation of entity sub-types (red tags) and components (blue tags): *Mr Fiat, general Superior of the Lazarists*

Secondly, a sub-type includes another sub-type, used as a component. In Figure 2, the *func.ind* sub-type (individual function), which spans the whole expression, includes the *loc.adm.town* sub-type (administrative location for a town), which spans the single word of the French town *Versailles*.

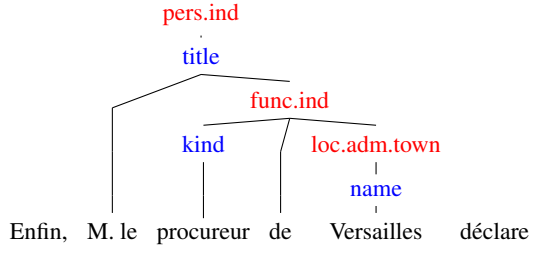

Figure 2: Multi-level annotation of entity sub-types: *At last, Mr the prosecutor of Versailles declares*

At last, in cases of metonymy and antonomasia, a sub-type of entity is used to refer to another sub-type of entity (Figure 3). The sub-type to which the entity intrinsically belongs is annotated (the *loc.oro* sub-type, an oronym location). Then, this entity sub-type is over-annotated with the sub-type to which the expression belongs in the considered context (the *org.adm* sub-type, an administrative organization).

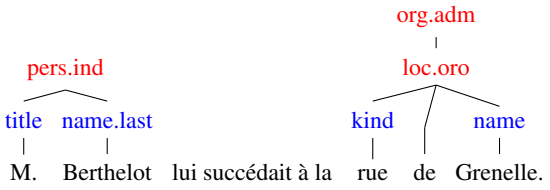

Figure 3: Annotation with sub-types and components including metonymy: *Mr Berthelot was succeeding him at rue de Grenelle*

2.3 Experiments on Broadcast News Data

In (Grouin et al., 2011), we reported a human annotation campaign using the above-mentioned structured entities on spoken data and the resulting corpus. The training part of the corpus is only composed of broadcast news data while the test corpus is composed of both broadcast news and broadcast conversations data. In order to build a mini-reference corpus for this annotation campaign (a "gold" corpus), we randomly extracted a sub-corpus from the training one. This sub-corpus was annotated by 6 different annotators following a 4-step procedure. Table 1 gives statistics about training, test and gold corpora. These corpora ("BN" in the remainder of the paper) has been used in an evaluation campaign (Galibert et al., 2011).

Data / Inf.	Training	Test	Gold
# shows	188	18	-
# lines	43,289	5,637	398
# tokens	1,291,225	108,010	11,532
# entity types	113,885	5,523	1,161
# distinct types	41	32	29
# components	146,405	8,902	1,778
# distinct comp.	29	22	22

Table 1: Statistics on the annotated BN corpora

3 Structured Named Entities in Old Newspapers

We performed the same annotations on a corpus composed of OCRed press archives, henceforth the Old Press (OP) corpus. Human annotation was subcontracted to the same team of annotators as for the BN corpus, thus facilitating the consistency of annotations across corpora.

3.1 Corpus

The Old Press corpus consists of 76 newspaper issues published in December 1890 and provided by the French National Library (Bibliothèque Nationale de France). We used three different French titles: *Le Temps*, 54 documents for a total 209 pages, *La Croix*, 21 documents for a total 84 pages, and *Le Figaro*, 1 document with 2 pages.

A newspaper is composed of various parts (*titles, articles, ads,* etc.), some of which are not useful for named entity annotation. A corpus study allowed us to determine parts in which we considered annotation would be useless: *titles, mastheads, ads, tables of numbers, theater programs, stock exchange results, weather reports, etc*. We designed a procedure to filter out these parts in each document, which is fully described in (Galibert et al., 2012). The result consists in a corpus of about 23,586 blocks extracted from 295 different pages.

3.2 Adaptation of Annotation Guidelines

Given the characteristics of the corpus (OCRed press archives), while the OCR quality rate is good (Character Error Rate at 5.09% and Word Error Rate at 36.59%[1]), we introduced a new XML attribute and a new component into the annotation schema in order to take into account these features and to fulfill annotators' requirements.

Attribute *correction*. Annotators were asked to correct incorrectly recognized entities. To save time and effort, correction was to be performed only on named entities, not on the whole text (see Figure 4 where the entity "*d'Algor*" being of type *loc.adm.town* has been corrected into "*d'Alger*" (from Algiers)).

[1]The CER and the WER was computed in terms of Levenshtein distance (Levenshtein, 1965).

```
00232766/PAG_1_TB000125.png
On nous télégraphie <loc.adm.town correction="d'Alger"> d'Algor </loc.adm.town> Le
<prod.object> Comorin, </prod.object> venant du <loc.adm.reg> Tonkin, </loc.adm.reg>
est arrivé en rade <loc.adm.town correction="d'Agha"> d'Ag'ha </loc.adm.town> à
<time.hour.abs correction="trois heures de l'après-midi;"> <val> trois </val> <unit>
heures </unit> do <time-modifier> l'après-midi; </time-modifier> </time.hour.abs> il
n'a pu être admis à la libre pratique qu'à <time.hour.abs> <val> cinq </val> <unit>
heures </unit> du <time-modifier> soir, </time-modifier> </time.hour.abs> par suite
d'un décès survenu devant <prod.object> Bougie. </prod.object> A <time.hour.abs>
<val> six </val> <unit> heures, </unit> </time.hour.abs> il mouillait dans le port.
Il débarquera ses troupes <time.date.rel> aujourd'hui </time.date.rel> dans la
matinée et appareillera ensuite nour <loc.adm.town> Toulon. </loc.adm.town>
```

Figure 4: Sample of annotations

Component *noisy-entities*. When a character recognition error involves an entity boundary, a segmentation error occurs, either between an entity and other tokens, or between several entities and possibly other tokens. To allow the annotators to annotate the entity in that character span, we defined a new component *noisy-entities* which indicates that an entity is present in the noisy span of characters. A complete description of these adaptations can be found in (Galibert et al., 2012).

3.3 Inter-Annotator Agreement

To evaluate the manual annotations of the annotation team ("Global annotated corpus" in Figure 5), we built a mini reference corpus by selecting 255 blocks from the training corpus. We followed the same procedure than the one used for the BN corpus, as illustrated in Figure 5:

1. The corpus is annotated independently by 2 teams of 2 annotators ("Scientist" boxes).

2. Each team produces an adjudicated annotation corpus from the two teams' annotations ("Institute 1" and "Institute 2" boxes).

3. One team produces an adjudicated annotation corpus from the two previously obtained versions of the corpus ("Institutes" box).

4. Then, one team produces an adjudicated annotation corpus ("Mini-reference" box) from the previous corpus and the corresponding corpus extracted ("Annotated sub-corpus" box) from the global annotated corpus.

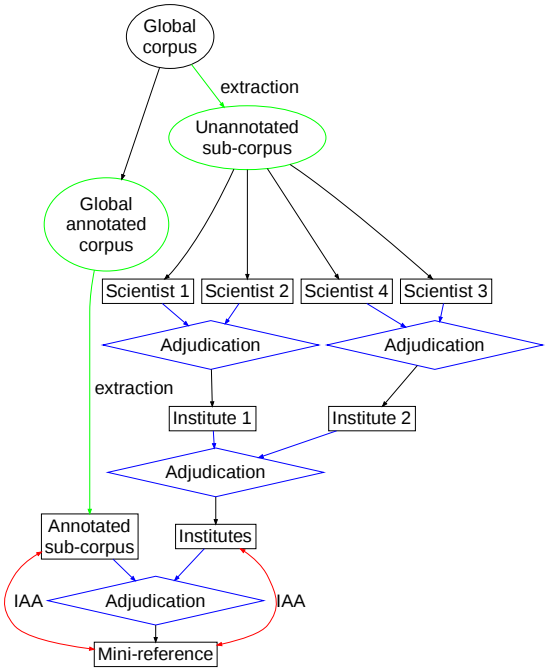

Figure 5: Mini reference corpus constitution procedure. Parts of the figure in green refer to the extraction stage, parts in blue to the adjudication stage and parts in red to the inter-annotator agreement stage

The complete annotated corpus was divided for evaluation purposes into training and test corpora, as described in (Galibert et al., 2012). Table 2 gives statistics about these corpora and the gold corpus.

During the whole annotation process, inter-annotator agreements and disagreements were computed. Here, we present the results in term of inter-annotator agreement between the annotated sub-corpus and the mini reference corpus.

Information \ Data	Training	Test	Gold
# pages	231	64	-
# lines	192,543	61,088	1618
# tokens	1,297,742	363,455	12,263
# distinct tokens	152,655	64,749	5,215
# entity types	114,599	33,083	1,373
# entity types w/ corr.	4,258	1,364	65
# distinct entity types	40	39	29
# components	136,113	40,432	2,053
# components w/ corr.	71	22	51
# distinct components	26	25	23

Table 2: Old Press corpora annotated with extended named entities. *Gold* stands for mini reference corpus; *corr.* for correction attribute

To compute an inter-annotator agreement, we need a 'random baseline' which is dependent on the number of *markables*. We showed that considering as markables all entities annotated at least in one of the two corpora should lead to the lowest possible bound for κ estimation (in our experiment, $\kappa = 0.64705$) (Grouin et al., 2011). In contrast, the F-measure can indicate the highest possible bound ($F = 0.79930$).

4 Comparisons

We have annotated two different corpora using the same definition of extended and structured Named Entity. This gives us an opportunity to analyze differences in (*i*) the annotation campaigns for these two corpora, highlighting specific difficulties linked to corpus properties (Section 4.1), and (*ii*) the obtained annotated corpora (Section 4.2).

4.1 Annotation Campaign

4.1.1 From the Source Material Point of View

As mentioned in Section 3.2, the Old Press annotation included an additional task for the annotators: correcting the incorrectly recognized characters in the annotated named entities. Performing this task properly implies to read not only the OCRed text, but also the corresponding source image, as some errors do not appear as such in the text. This is the case, for example, in "*M. Buis*" (Mr Buis) instead of "*M. Buls*" (Mr Buls) or, more importantly, "*touché*" (touched) instead of "*Fouché*" (last name of a person). In addition to this, segmentation issues had to be dealt with. For example, "*M. Montmerqué,ingénieur des ponts etchauasées*" (Mr Montmerqué, highway engineer) has two tokens and a punctuation glued together (*Montmerqué,ingénieur*).

4.1.2 From the Language Point of View

Specific languages. A set of difficulties was due to the specific language types encountered in the corpus, in particular the religious language from the newspaper *La Croix* (17 issues, 68 pages). Some expressions, like "*mandement de Carême*" (Lent pastoral prayer) were found difficult to annotate and required some checking in external sources. The language used in classified ads from *Le Temps* was also quite difficult to annotate due to their format (see Figure 6) and the abbreviations they contain, which are not always easy to understand. For instance, in the same figure, *Cont.* might stand for contiguous.

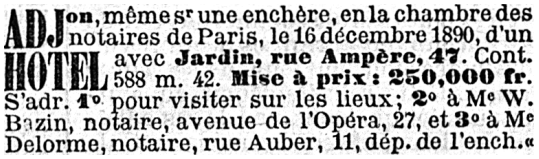
ADJon, même sr une enchère, en la chambre des notaires de Paris, le 16 décembre 1890, d'un HOTEL avec Jardin, rue Ampère, 47. Cont. 588 m. 42. Mise à prix : 250,000 fr. S'adr. 1° pour visiter sur les lieux; 2° à Me W. Bazin, notaire, avenue de l'Opéra, 27, et 3° à Me Delorme, notaire, rue Auber, 11, dép. de l'ench.«

Figure 6: Example of classified ads from *Le Temps*

Cultural context. Another set of difficulties was due to the general cultural context of the time, which is long forgotten now and which the annotators had to rediscover, at least partly. Thus, they had to consult various external sources, like Wikipedia, to check geographical divisions (was "*Tonkin*" a country or a mere region in 1890?), events (in "*le krach Macé*" (Macé crash), does "*Macé*" correspond to a family name?), and even names (is "*Lorys*" a first or last name?).

More generally, the language of the time (end of the 19th century), though apparently close to present French, presents some specificities that required a re-interpretation of the annotation guide. For example, persons were almost systematically designated by their title (e.g., "Mr X", where "Mr" is a *title* component and "X" a *name.last* component).

Annotation difficulties. During the Broadcast News campaign, we noticed that the distinction made in the annotation guide between a function (which does not include a person) and a title (which is included in a person named entity) was in fact not stable and difficult to use. In the Old Press corpus, with the high frequency of the usage of a title with a name of a person, this distinction generated even more questions, mistakes and inconsistencies in the annotations. These differences, though minor and more or less expected, made the annotation more complex, as it depended on realities that are much less frequent nowadays.

Finally, difficulties regarding boundary delimitation were more frequent, most probably due to the written form of the OP corpus (as opposed to the spoken form of the BN corpus). Figure 7 shows a long entity which should probably include *France*.

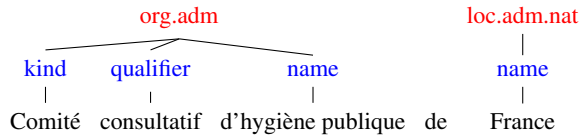

Figure 7: Boundary delimitation difficulties: *consultative committee for public hygiene of France*

4.2 Study of Annotated Corpora

The Broadcast News (BN) corpus and the Old Press (OP) corpus have different temporal and genre properties. Intuitively, we expect these differences to be reflected in the types of named entities they contain.

Having these two corpora consistently annotated with the same types of named entities makes it easier to test this hypothesis. Furthermore, the structure of these entities opens avenues for more detailed contrastive analysis than plain, one-level entities. We sketch some of them in this section.

We used two methods to test the above hypothesis. First, we used a statistical test to compare the distributions of entity types and components across the two corpora. Second, we checked whether documents of these two corpora could be discriminated based on their entity types and components.

4.2.1 Statistical Analysis

We study in this section whether there is a significant difference between the two corpora in terms of entity types and components. Let us stress that we examine only the labels of these entities (e.g., org.adm or name), not their contents (e.g., *Comité consultatif...*).

We first examined the density of entities in documents of the two corpora. For each document, the entity-token ratio is computed as follows: the total number of occurrences of entity types and entity components (*tags*), divided by the number of tokens in the document (*tokens*): $\frac{\text{tags}}{\text{tokens}}$. A Welch Two Sample t-test (computed with the R t.test function) was performed to check whether the mean entity-token ratio was significantly different across the Old Press and Broadcast News documents. It shows that the two means (0.233 and 0.251) have a slightly significant difference (95% confidence interval, $p < 0.01$).

We then applied the same test to each entity type and each entity component. To remove the difference in entity-token ratios, the variable we examine for each entity type or component is the proportion of occurrences of this type or component (tag_i) among all occurrences of entity types and components (*tags*) in a document: $\frac{\text{tag}_i}{\text{tags}}$. Entity types and entity components are all the more over-represented in one of the two corpora as the significance level (p) is high.

Figures 8 and 9 respectively rank entity types and components in decreasing order of p. Bar height reflects the difference in the two means: an ascending bar means that the entity is more present in the Broadcast News corpus, a descending bar in the Old Press corpus. In total, 36 entity types and components out of 73 have a $p < 0.001$, and 6 more have a $p < 0.01$. Therefore, more than half of them have a significant difference across the two corpora.

Entity type density. We can see on Figure 8 that BN has a greater proportion of countries and continents (loc.adm.nat, loc.adm.sup: maybe due to more international news in contemporary press), relative dates and times (time.date.rel, time.hour.rel: possibly linked to the media, audio and television, with more immediate references), companies and administrations (org.ent, org.adm), media names (prod.media). OP has a greater proportion of towns (loc.adm.town), absolute dates (time.date.abs), individual persons and functions (pers.ind, func.ind),

physical addresses, including streets, roads, facilities (loc.add.phys, loc.oro, loc.fac: reference is more often given to where something can be found), hydronyms (loc.phys.hydro), and works of art (prod.art: articles about plays in theaters).

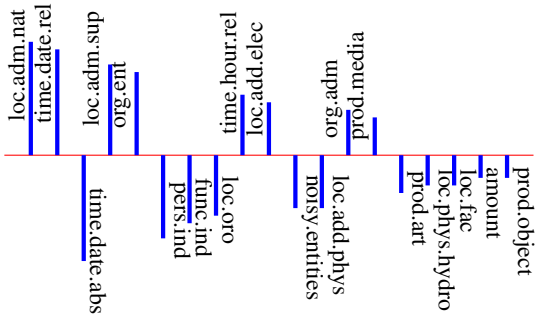

Figure 8: 19 entity types with $p < 0.001$, ranked by decreasing order of significance

Some entity types are only present in one of the corpora. This is indeed the case of the noisy-entities element introduced for OP, but also of electronic addresses and software (loc.add.elec, proc.soft) which did not exist in the nineteenth century.

Entity component density. Figure 9 shows that BN has a greater proportion of first names, middle names, and demonyms (name.first, name.middle, demonym), whereas OP has a greater proportion of titles and last names (title, name.last): this reflects differences in time periods (more titles in the nineteenth century, use of first name in addition to last name in contemporary news) and topics (use of demonyms for sports teams in contemporary news).

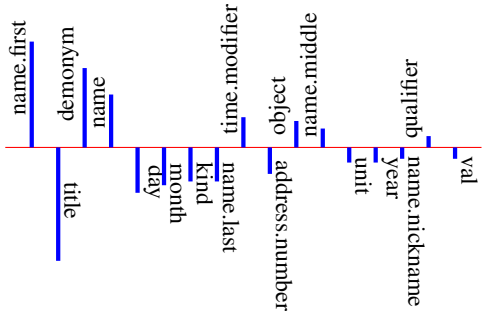

Figure 9: 17 components with $p < 0.001$, ranked by decreasing order of significance

Days, months and years are in greater proportion in OP since they are the components of absolute dates, also in greater proportion in OP (see above).

More precise assessments can be performed based on the rich structure of the entities, with their nested types and components. Among person entities (pers.ind and pers.coll), BN has a much larger proportion (52% vs. 6%) of persons composed of a first and a last name (pers_first_last: entities of the form <pers.*> <name.first/> <name.last/> </pers.*>) and of persons with a first name (pers_with_first: entities where <pers.*> includes a <name.first/>: 69% vs. 19%), whereas OP has a much larger proportion (44% vs. 8%) of persons with a title (pers_with_title: entities where <pers.*> includes a <title/>) and of persons composed of a title and a last name (pers_title_last, 34% vs. 2%: *M. Berthelot*). In contrast, there is no significant difference in the proportion of persons with a last name. This refines the individual observations on name components and types, and shows that although OP has a greater proportion of last names, it has in the same way a greater proportion of persons, so that their ratio is constant across both corpora. On the contrary, the greater proportion of titles in OP is confirmed by its greater proportion of persons who include a title.

In another area, OP has a quite large proportion (29% vs. 6%) of administrations (org.adm) that are composed of exactly one <kind> component (orgadm_kind): the most frequent ones are *la Chambre*, *le gouvernement*, *la République*, *l'Etat*, etc., instead of a kind and some precision (e.g., *la <org.adm> <kind> Chambre </kind> des <func.coll> <kind> députés </kind> </func.coll> </org.adm>* [the Chamber of Representatives]). This reflects a particular administrative (here, governmental) system and a conventional reduction of the full name of some of its instances.

4.2.2 Automatic Classification and Feature Selection

If the distributions of entity types and components are sufficiently different across the Broadcast News and Old Press corpora, it should be possible to use them as features in a classifier which detects whether a document of these corpora belongs to BN or OP.

To test this hypothesis, we used as features for a document the same variables as in the statistical analysis: the $\frac{\text{tag}_i}{\text{tags}}$ ratio for each entity type and component.

We tested several classifiers (using the Weka toolbox (Hall et al., 2009)), with stratified ten-fold cross-validation over the whole training corpus (188 BN documents and 231 OP documents). Table 3 shows the results for One Rule (OneR), decision trees (J48), Naïve Bayes, and SVM (SMO). False Negative (FN) and False Positive (FP) computation assumes that the target class is Old Press (hence FN is the number of OP documents classified as BN).

	FP	FN	FP+FN	Accuracy
One Rule	22	12	34	0.919
Decision Tree	2	5	7	0.983
Naïve Bayes	2	1	3	0.993
SVM	0	0	0	1.000

Table 3: Classification based on tag ratio

Even a baseline classifier (One Rule) obtained a high accuracy (0.919). It chose the title feature and produced the rule "*if title $<$ 0.0255 then BN, else OP*". This is consistent with the above observation that it has the second most significant difference in means across the two corpora.

The Decision Tree classifier obtained a much better accuracy, based on a tree based on features title, then on name.first and loc.adm.sup (also among the most significant differences), and on func.ind, demonym (very significant differences too). The Naïve Bayes classifier did better (0.993), and the SVM obtained a perfect classification: taken together, the 73 tag ratios are indeed discriminant enough to determine the corpus to which a document belongs.

Performing feature selection is yet another way to test which entity types and components are the most discriminant. Using the default feature selection method in Weka (CfsSubsetEval with BestFirst search) selected 21 features, 19 of which had a $p < 0.001$. With only the three features title, demonym, and name.first (the three tag ratios with the most statistically significant differences), the SVM still correctly classified all documents but one. This confirms that some of the entity types and components are highly discriminant. Interestingly enough, the three most discriminant ones are components: this underlines the contribution of this aspect of our structured named entities.

5 Conclusion and Perspectives

We have presented the human annotation of a second reference corpus (Old Press) with Structured Named Entities, using the same annotation scheme as in the previous corpus (Broadcast News). These two corpora have similar overall sizes in tokens and numbers of entity types and components, but are different in terms of time periods and media. This entailed a need to adapt slightly the annotation guidelines.

Having two corpora annotated according to the same annotation scheme makes it possible to perform contrastive studies. We reported a series of observations on the human annotation of these two corpora. We illustrated the impact of OCRed text and of a time-induced cultural distance on the human annotation process. Based on the annotation results, we evidenced significant differences between the entity types and components of the two corpora, as well as discriminant entity types and components.

The structured named entities made it possible to study finer-grained distinctions, such as different naming structures for people (title + last name in Old Press vs. first + last name in Broadcast News), or single-component (in Old Press) vs. multiple-component administrative organizations.

Indeed, the studies reported in this paper are but a small sample of what can be achieved thanks to these structured entities. At the time of writing, we are in the final stages of the paperwork necessary to release the two corpora for free usage by the scientific community. We hope that many colleagues will thus obtain these corpora and use them both to train named entity recognizers and to perform more precise contrastive studies.

Acknowledgments

This work has been partially financed by OSEO under the Quaero program and by the French ANR ETAPE project.

We thank all the annotators *(Jérémy, Matthieu, Orane, and Raoum)* from Elda society; they worked seriously and with professionalism.

References

Eckhard Bick. 2004. A named entity recognizer for Danish. In *Proc. of LREC*. ELRA.

Kate Byrne. 2007. Nested named entity recognition in historical archive text. In *Proceedings of the first IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, California.

Sam Coates-Stephens. 1992. The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 26:441–456.

Gregory Crane and Alison Jones. 2006. The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL'06, pages 31–40, New York, NY, USA. ACM.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program—tasks, data, and evaluation. In *Proc. of LREC*, pages 837–840. ELRA.

Michael Fleischman and Eduard Hovy. 2002. Fine grained classification of named entities. In *Proc. of COLING*, volume 1, pages 1–7. ACL.

Michael Fleischman. 2001. Automated subcategorization of named entities. In *Proc. of the ACL 2001 Student Research Workshop*, pages 25–30.

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean-Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deléger, and Dominique Laurent. 2010. Named and specific entity detection in varied data: The Quaero named entity baseline evaluation. In *Proc. of LREC*, Valletta, Malta. ELRA.

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. 2011. Structured and extended named entity evaluation in automatic speech transcriptions. In *Proc. of IJCNLP*, Chiang Mai, Thailand.

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. 2012. Extended named entities annotation in ocred documents: From corpus constitution to evaluation campaign. In *Proc. of LREC*, Istanbul, Turkey. ELRA.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference - 6: A brief history. In *Proc. of COLING*, pages 466–471.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc. of the Fifth Linguistic Annotation Workshop (LAW-V)*, Portland, OR. Association for Computational Linguistics.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/.

Mark Hall, Eibe Franck, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Vladimir Levenshtein. 1965. Binary codes capable of correction deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 2000. Named entity extraction from noisy input: speech and OCR. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC'00, pages 316–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum, 2011. *Entités Nommées Structurées : guide d'annotation Quaero*. LIMSI–CNRS, Orsay, France. `http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf`.

Satoshi Sekine. 2004. Definition, dictionaries and tagger of extended named entity hierarchy. In *Proc. of LREC*. ELRA.