# Detecting Derivatives using Specific and Invariant Descriptors

Fabien Poulard, Nicolas Hernandez, and Béatrice Daille

***Abstract*—This paper explores the detection of derivation links between texts (otherwise called plagiarism, near-duplication, revision, etc.) at the document level. We evaluate the use of textual elements implementing the ideas of specificity and invariance as well as their combination to characterize derivatives. We built a French press corpus based on Wikinews revisions to run this evaluation. We obtain performances similar to the state of the art method (n-grams overlap) while reducing the signature size and so, the processing costs. In order to ensure the verifiability and the reproducibility of our results we make our code as well as our corpus available to the community.**

***Index Terms*—Textual derivatives, detection of derivations, near-duplicates, revisions, linguistic descriptors, French corpus.**

## I. Introduction

BEING in the age of information, the information is not only produced but also duplicated, revised and plagiarized at some extent. This redundancy is an hindrance to Information Retrieval (IR) methods in terms of computation, storage and results. Hence, the performance of web search engines could be improved with the filtering of duplicate texts as, meanwhile saving the storage necessary for the index. Moreover, users may not want duplicated (or even near-duplicated) documents in the answer to their search query.

We address the task of detecting text derivatives of a given source document among a collection of suspicious documents, *i.e.* given a collection of suspicious and source documents, one must map the first to the second therefore detecting the derivation links involving a suspicious and a source. This task is usually handled by measuring the n-grams overlap between sources and suspicious. We propose to use textual elements implementing the ideas of specificity and invariance (hapax n-grams, named entities and nominal compounds) instead of n-grams. We report the performance of the classic approach on a corpus we made out of revisions of French news articles. We compare the performances of our propositions to this baseline.

First we introduce the classic signature approach to the problem (Section II). Then we describe the way we built a French corpus (Section III) and present our methods (Section IV) and the evaluation protocol for our experiments (Section V). Lastly, we report the results of our experiments (Section VI) and conclude the paper (Section VII).

Manuscript received November 9, 2010. Manuscript accepted for publication January 15, 2011.

The authors are with the University of Nantes / LINA (CNRS - UMR 6241), 2 rue de la Houssinière, B.P. 92208, 44322 Nantes Cedex 3, France (e-mail: first.last@univ-nantes.fr).

## II. Related Works

The methods to handle the task we address depend on the granularity of the derivation and the transformations involved [1]. Texts that wholly derive from another one are better identified with suffix trees and string alignment methods [2], or using chunks frequency models when rewritting is involved [3]. Texts partially derived are better identified using matching chunks [4].

The n-grams overlap approach usually gives the best results for moderatly rewritten partially derived texts [4], [5]. It has been generalized and formalized by [6] as *w-shingling*. It consists of counting the contiguous subsequences of tokens (w-shingles) two texts have in common using a set theory based similarity metric. The assumption is that the more w-shingles the texts have in common the more probably they derive from each other. The set of the w-shingles of a text is its *signature*. The tokens composing the w-shingles can be any textual elements corresponding to a particular description. A *descriptor* describes the nature of these tokens as well as how they are combined into w-shingles. Several descriptors have been experimented in the litterature: fixed-length characters chunks [7], hashed breakpoints [8], words n-grams [6], [4], [5], sentences [9].

The major limit of this signature approach is its cost. The generated signatures are as large as the text which is inapropriate to handle large amount of data. For example, as word n-grams are not linguistically anchored, the signatures using this descriptor must contain all the overlapping n-grams of a text to match modified texts. This results in a signature even larger than the text itself, impacting the storage and the computational costs. One solution is to hash the tokens and only consider some meaningful bits of the hash therefore reducing the size of the fingerprint [10], [11]. However, the link to the elements in the texts are lost which is acceptable for near-duplicates as the whole document is derived but may not be for other kind of duplicates. We propose to focus on the choice of descriptors that are less numerous in the texts but are more effective at identifying derivations.

## III. Building a Corpus to Evaluate Derivation Detection

A crucial question with NLP studies is the availability of a corpus resource with the wanted language phenomenon annotated in order to be able to infer and to test some hypothesis to retrieve it.

In the domain of the derivation detection, a few corpora with such annotations are available in English (METER [12], NTF and NTF2 [13], PAN [14]), no such resource is currently available in French. We note two major trends in building derivation corpora: (i) artificially generate derivations from a collection of texts by mixing them together [14], (ii) manually retrieve existing derivatives (from the Web for example) [12] or ask human to create some [14]. Both methods offer advantages and drawbacks. On the one hand, the artificial approach allows to quickly get a resource by performing automatically morphological, lexical, syntactic and semantic text edits (deletion, insertion, inversion, substitution) at various degrees and text granularities. The main drawback of this approach is that there is no mean to evaluate how much these transformations stand for natural language and consequently potential derivations. On the other hand, the major advantage of manually writting or searching for existing derivatives is that it may lead to get actual instances of a derivation process. Its drawbacks are that it needs time and fund to build a substantial corpus by searching derivatives and futhermore, it is often impossible to systematically control the search space as well as to be sure about the existence of the derivation links.

We argue that another way of building quickly a substantial corpus with actual derivation relations between documents is to use available corpora which include the annotation of some actual transformations between the documents, such as summarization synthesis, translation, revisions... As the manual simulation of the derivation process, this approach may not cover all the potential types of derivatives but the process to acquire them will be faster and probably cheaper. In this paper, we worked with a corpus made of revision texts.

Working with revisions is interesting for several reasons: the revision is a well-controled derivation type (sources and derivatives are easily identifiable, the derivation degree can be measured by the number of revision), it includes various forms of transformations such as spelling and grammar errors correction, insertion and deletion of contents, rephrasing,... We chose to work with *Wikinews* which is a project of the Wikimedia Foundation. Based on the idea of a collaborative journalism, Wikinews is a multilingual free-content[1] news source wiki. In addition to a head version of a news article, revisions and potential translations of the news are also available. We built our corpus from the data export of the French version of Wikinews[2] in date of November the $13^{rd}$ 2009. All the news articles having more than 10 revisions were selected; this constraint was set in order to reinforce the probability of getting suspicious texts with high degrees of edit operations from an initial source text.

The corpus is structured like the PAN corpus. It distinguishes the source texts and their derivatives. We choose to consider the first version of a news article as the source text and all the following revisions as the derivatives. As a matter of fact, the roles of not-derivative texts of a given source are played by the derivative texts of all other source texts. Since the PAN corpus is currently the reference to hold the evaluation of a derivation detection task, we adopted its file format conventions in order to ensure compatibility with it. The corpus is made of 221 source texts and 2,670 derivatives. On average a news article contains 604 words.

[1]Released under *Creative Commons Attribution 2.5*

[2]The Wikinews dumps can be downloaded from http://download.wikimedia.org. We used the *UIMA mediawiki engine* (http://code.google.com/p/uima-mediawiki-engine) to select and extract the raw texts from the news files.

## IV. Approach

We address the task of detecting the derivatives of a given source. We particularly focus on document level derivatives, *i.e.* texts whose content is mainly derived from the source text as opposed to texts where only some minor passages are derived from the source text. Our goal is to develop a low operational costs method of detection.

As discussed in Section II, for a signature method to be operational, we must reduce the number of its elements. In order to do so, we must find more effective descriptors than word n-grams. In our opinion, this effectivity is a consequence of the specificity and the invariance of the descriptor. The idea that underlies the *specificity* is that a match on a signature element is more worth it if this particular element is only found in the source text that if it is a common element found in almost any text[3]. In other words, the less common a descriptor is the better it will discriminate the document. The *invariance* represents the ability of the descriptor instances to be preserved by the derivative process. In other words, the concept or the reference introduced by the instance should be found in the source and its derivatives.

In this paper we explore the use of descriptors chosen for their specificity or invariance: hapax n-grams, named entities and nominal compounds. We also explore their combination as pros of each may overcome cons of the others.

### A. Hapax n-grams

The hapax n-grams both extend the idea of using word n-grams while implementing the principle of specificity and reducing the number of elements in the signature. Moreover, they can be easily extracted using a reference distribution.

Hapax n-grams are a great example of specificity. They extend the concept of w-shingling by reusing word n-grams as basic units composing the signature, so their implementation is not much different that the w-shingling method. However, a filtering step is necessary as we only keep extremly specific elements : these appearing only once, the hapax. More precisely, we select from the word n-grams of a text the ones with a $\mathrm{df}$ (document frequency) of one or less given a reference distribution. The method hopefully reduces the

[3]It is a direct interpretation of the fact that the more an element derive from the Poisson distribution the more it is useful to discriminate the hidden relationships behind text [15]

number of elements in the signature as it is a filtered version of the original w-shingling.

The only difficulty in building such a signature is to obtain a reference distribution. The reference distribution must be computed over a corpus of the same genre and same language as ours. Using the same corpus is not an option as it would result in an identical distribution while we are interested in variations, and as it is mainly made of derivatives it is not representative of the language (redundancy of reused expressions). This would lead to erroneous results. Instead we use the pages of Wikinews that are not part of the corpus. We only keep one revision per article to avoid derivatives, we especially select the last revision as it is generally the longest and the most correct. The resulting corpus is composed of 1,027 French press articles, representing 289,288 words. The word n-grams are extracted from this reference corpus and stored in an index with their $\mathrm{df}$. Therefore, we considered as hapax the n-grams of our corpus with a $\mathrm{df} <= 1$ in this index.

### B. Nominal Compounds and Named Entities

So far, researchers payed relatively little attention to linguistic-based descriptors. According to us, signatures based on some linguistically motivated descriptors can enhance the detection performance compared to $n$-grams w-shingling signatures.

First, since a linguistic descriptor is defined by some grammatical and semantic constraints, its instances are a subset of the text which is a solution to reduce the size of the signature Second, some linguistic descriptors may be considered to be more relevant than others to describe the content of a document. Among them, we include the nominal compounds and the named entities. Third, since instances of these descriptors result from a linguistic choice of the author, they provide a greater probability to integrate specificities from the author of the source text.

We decide to consider two distinct categories of linguistic descriptors: the named entities (names of persons, organisations, locations) and the nominal compounds. We assume that if the instances of these descriptors from a source are found in a suspicious text they enhance the probability for the suspicious text to be a derivative.

We choose to observe the named entities because they usually designate the referents of the actors or of the context elements of the events reported in news. For named entities extraction, we used the French system Nemesis [16]. Nemesis follows a lexical and grammar-based approach with some automatic learning techniques to enrich the lexicon. It achieves a performance of 95 % in precision and 90 % in recall for recognizing anthroponyms and toponyms in press texts.

Whereas the named entities constitute expressions which stand for referents, the nominal compounds are generally used as the most syntactically plausible class of terminological candidates to model the concepts of the knowledge domains. They constitute more than 80 % of the domain specific terms for the specialized languages [17], but they are also used in the informal language. We use the grammar-based patterns[4] proposed by [18] to extract the nominal compounds: N A (*emballage biodégradable*, *protéine végétale*), N (P (D)) N (*ions calcium*, *protéine de poissons*, *chimioprophylaxie au rifampine*), N à Vinf (*impôts à acquitter*, *fonds à venir*). These patterns are recursive and may admit some variations such as N N A (*forces armées britaniques*), N A (P (D)) N (*lait cru de brebis*) or N (P (D)) N A (*protéine d'origine végétale*, *réunion de la Commission Parlementaire*). In our implementation, we only considered the precited patterns and variants without further recursive variations. Overlapping nominal compounds retrieved by different patterns were allowed in order to enhance the capability of detecting partial rewriting. We used the Apache UIMA Tagger[5] which was trained on the French treebank [19] to compute parts-of-speech on the texts.

## V. Evaluation Protocol

The systems to detect derivatives are usually evaluated as classifiers using the computed similarity scores, whether they categorize pairs of documents [4] or pairs of passages [14]. Usually, the classifier is based on a simple similarity threshold which differs derivatives from not-derivatives. Thus, the underlying comparison method is not evaluated as the focus is on the correct distribution of pairs in their respective classes. This kind of evaluation is appropriate for a decision making system which we believe is not a relevant choice for our problem.

We think derivatives detection systems should be seen as decision support systems and evaluated as such. Therefore, the evaluation must measure how the system sorts out relevant candidates and help a human to take a decision regarding the derivative status of a text by providing relevant insights. In the continuity of the works from [20] and [21], we think an IR-like evaluation is the best choice for such a system. We sort pairs of documents (one source and one suspicious) according to their similarity score computed by the system. The highest scores obtain the highest ranks.

We are interested in three evaluation axes: the quality of the ranking (pairs with derivatives should obtain the highest ranks and not-derivatives the lowest), the discrimination capability of the system (derivatives scores should be very different from not-derivatives ones) and the computation costs (storage cost and execution time of the system). We also present the results of the w-shingling approach that we use as a baseline.

### A. Quality of the Pairs Ranking

The quality of the pairs ranking is the most obvious property to evaluate the quality of our system. It is comparable to the precision and recall measures for the evaluations as

[4]The Part-Of-Speech tag A stands for Adjective, N for Noun, D for Determiner, P for Preposition and Vinf for Infinive Verb. *à* is a specific preposition.

[5]http://uima.apache.org/downloads/sandbox/hmmTaggerUsersGuide/

classification tasks in the sense that it evaluates how well are derivatives identified.

The mean average precision (MAP) metric is well suited to measure the quality of the ranking as it combined precision and recall like notions without the need for a binary categorization between derived and not-derived. It is the average of the regular precision metric computed over a growing window of N ranks starting from rank 1 (Equation 1). We use for N the rank of the last derivation link in the ranking. The recall is expressed here through the denominator N: the highest N is the more there are not-derivatives before the last derivative and the more important is the impact on the MAP.

$$\begin{aligned} &\text{MAP} = \frac{\sum_{r=1}^{N} P(r)}{N} \\ &r \text{ a rank} \\ &N \text{ the highest rank considered for the computation} \\ &P(r) \text{ the precision computed over rank 1 to } r \end{aligned} \tag{1}$$

### B. Discrimination Capability

The discrimination capability of the method reflects how well the method makes a difference between derivation links and not-derivative ones.

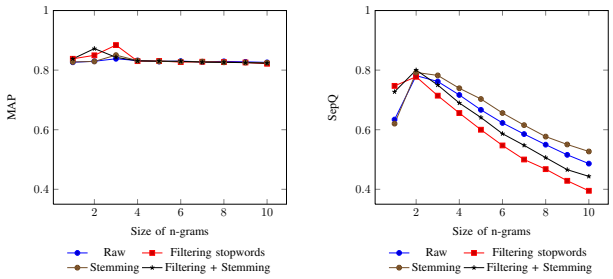
This property of the system is measured as the size of the buffer between derivation links similarity scores and not-derivative ones with the asumption that the larger this buffer is, the more each link is considered differently from the other by our system. The separation is the difference between the similarity score of the highest not-derivative in the ranking and the lowest derivative one. We introduce the SepQ that, instead of using the extrems of each, considers the similarity scores of the third quartile of the derivation links and the first quartile of the not-derivative ones (Equation 2). Indeed, the consideration of a unique individual, in addition an extrem, may not reflect the group. Therefore we prefer to measure the distance between the most significative $\frac{3}{4}$ of each group: the highest similarities for the derivation links and the lowest for the no derivation ones.

$$\begin{aligned} &\text{SepQ} = s_{deriv} - s_{\neg deriv} \\ &s_{deriv} \text{ similarity score of the } 3^{rd} \text{ quartile} \\ &\text{of the derivatives} \\ &s_{\neg deriv} \text{ similarity score of the } 1^{st} \text{ quartile} \\ &\text{of the not-derivatives} \end{aligned} \tag{2}$$

### C. Computational Costs

As our goal is to develop a low operational costs method of detection, we want to measure the computational cost of the system.

The system processes in two steps: the extraction of the signature and the comparison of the signatures pairs. The former is done only once so its impact on the computational costs can be neglected compare to the latter. For our approach, the complexity of the comparisons between two signatures is dominated by the computation of the intersection of the signatures which itself is linear with the size of the signatures ($O(2|s|)$). Therefore, we measure the computational costs by measuring the size of each type of signature.




(a) MAP (b) SepQ

Fig. 1. Results obtained for various size of n-grams.

### D. Baseline Approach

We use the w-shingling approach with word n-grams and the $c_{max}$ similarity metric (*Equation 4*) as a baseline. Experimentations not reported here show that this symetric measure (*Equation 3*) gives better results than the classic containment metric (*Equation 3*) for our corpus.

$$\begin{aligned} &c(a,b) = \frac{|\Pi(a) \cap \Pi(b)|}{|\Pi(a)|} \\ &\Pi(d) \text{ the w-shingling of document } d \end{aligned} \tag{3}$$

$$c_{max}(a,b) = \max(c(a,b), c(b,a)) \tag{4}$$

We explore several parameters regarding the n-grams to obtain the best possible results for the baseline. Thus, we experiment several sizes of n-grams as well as some morphological (stemming) and lexical (stopwords removal) normalizations. The results of these variations are presented in Figure 1. The MAP (Figure 1(a)) is globally constant independantly of the size of n-grams with just two peaks: stopword filtered 3-grams and stopwords filtered and stemmed 2-grams. The SepQ curve (Figure 1(b)) has a totally different shape as results fall with the increasing size of n-grams. The maximum is reached for 2-grams, whatever the type of n-grams. With regard to these results, the stopwords filtered and stemmed 2-grams is the configuration we choose as our baseline. The measured MAP for this configuration is of 0.872 and the SepQ of 0.800. We will consider the size of the corresponding signature as a tare for the next methods.

The Wikinews corpus represents a particular kind of derivations: revisions of press articles in French. The independance of the MAP relative to the size of n-grams is because of the sparse modifications. The revisions globally cover each others, but the raw n-grams are not adapted to capture the variations. The best results are obtained with some normalization and small n-grams. We think that the normalization removes unstable parts especially the endings associated with gender and number, while the small n-grams (2-grams and 3-grams) capture some stable

syntaxic constructions. This particular role of small n-grams is somehow supported by the SepQ results.

## VI. Results

Table I presents the results of the different descriptors described in Section IV and their combination. Results are compared to these of the baseline. In the last section, we discuss the results we obtain for the linguistic descriptors by manually looking at some selected pairs of compared texts.

### A. Results for Each Descriptor

We measure the performance of hapax $n$-grams with $n$ varying from 1 to 10. The MAP of the 2-grams baseline outperforms all the MAP scores of the individual descriptors, *a fortiori* the hapax MAP score. We note that for $n = 1$ and $n = 3$ the hapax MAP gives a better result and for $n >= 4$ they are quite similar. The SepQ is roughly the same whatever $n$ is. The value is decreasing while $n$ is increasing. In Table I, we present only the best results which are obtained with 2-grams and 1-grams respectively for the MAP and the SepQ. As a general trend, the MAP scores of the different descriptors never outperform the baseline but the SepQ ones are better and the corresponding signatures are smaller.

More precisely, Table I indicates that the MAP score of the named entity descriptor decreases of 0.22 points while the discrimination capability increases of 0.03 points. The most interesting observation we note concerns the signature size which corresponds to a significant decrease of the baseline signature size (5 % of this latter). Indeed, this decrease impacts positively the signatures storage cost and so the cost of the signatures comparison.

Turning now to the nominal compound descriptor, we can see in Table I that it gives lower scores than the baseline. Indeed, the MAP of the nominal compound descriptor results in a slight decrease of 0.04 points and its discrimination capability in also a slight decrease of 0.06 points. However, while these results are slightly lower, in comparison there is again a significant decrease of the signature size (15 % of the baseline).

### B. Combination of the Descriptors

The combination of descriptors can be considered at different stages: at the signature building stage by combining all signatures as one or at the similarity measure stage by a simple linear combination. In this paper, we choose to perform the latter for at least two reasons: first, it makes the signature building process easier allowing to compute separately each descriptor signature. Second, it easily allows to control the weight of each descriptor in the combination.

We define the linear function, $\mathrm{sim}_{comb}^{a,b,c}$, to combine the similarity scores we obtained by the differents approaches such that:

$$\begin{aligned}\mathrm{sim}_{comb}^{a,b,c}(t_1,t_2) &= a.\,\mathrm{sim}_H(t_1,t_2) + b.\,\mathrm{sim}_{NE}(t_1,t_2) \\ &\quad + c.\,\mathrm{sim}_{NC}(t_1,t_2)\end{aligned}$$

$t_1, t_2$ two texts on focus

$\mathrm{sim}_H, \mathrm{sim}_{NE}, \mathrm{sim}_{NC}$ scores of the $H$apax, $N$amed $E$ntity and $N$ominal $C$ompound methods

$a, b, c$ coefficients

(5)

We experimented a range of values from 0 to 3 for each coefficient. As shown in Table I, the combination $\mathrm{sim}_{comb}^{2,1,1}$ outperforms the baseline performances. Moreover, the various combinations reported all outperform the results of their individual constituent. This shows that being able to set correctly the coefficient of each descriptor can improve the combination results. Eventually we note that any combination has a significantly lower signature size than the baseline.

### C. Discussion

In order to discuss qualitatively the results we obtain with linguistically motivated descriptors, we manually observed some compared texts: the pairs of texts with no actual derivation link but with the highest similarity rankings and the pairs of texts with an actual derivation link but with the lowest similarity rankings[6].

We found three main reasons why some pairs of texts with no actual derivation link have a high similarity ranking. One reason is attributed to the comparison of signatures with very different size. As a consequence, the more elements a signature has the higher the probability is that this signature includes some elements of the compared signature. Another reason of potential high similarity ranking is due to a low quantity of descriptor instances in the compared texts. This was specifically observed for the named entity descriptor. In general the texts we processed use at most half a dozen of distinct named entities. As a result, one single shared element has strong impact in the score similarity measure. In addition to these remarks, we observed that some of the shared elements between the signatures belongs to a common lexicon which artificially increases the score of similarity. This was the case for the named entities descriptor with common toponyms such as *France*, *United States*, *North*... and also the case for the nominal compounds descriptor with for example some terms related to the model of the document such as "*source*" or "*exclusive right*".

For the named entities descriptor, we found one main explanation about why some pairs of texts with an actual derivation link got a low similarity ranking. Mainly, this was due to the fact that the shared elements were insignificant regarding the signature size. This observation is reinforced by the text variation of the named entities. Indeed *the President of the French Republic* and *the President* count for distinct elements in the signature and do not match if they are

[6]Pairs of texts with a null similarity score were not considered.

TABLE I
COMPARISON OF THE DESCRIPTORS SCORES IN TERMS OF MAP SCORE, SepQ SCORE AND SIZE RELATIVELY TO THE BASELINE SIGNATURE SIZE. FOR EACH SCORE, WE SKETCH ITS EVOLUTION COMPARED WITH THE BASELINE: ↗ INDICATES A SCORE INCREASE, ↘ A SECREASE AND = EQUIVALENT SCORE

| Descriptor(s) | | MAP | | SepQ | | Signature size | |
|---|---|---|---|---|---|---|---|
| Baseline | | 0.872 | | 0.800 | | 100 % | |
| Hapax | max(MAP): 2-grams (H2) | 0.856 | ↘ | 0.807 | ↗ | 78 % | ↗ |
| | max(SepQ): 1-grams (H1) | 0.849 | ↘ | 0.866 | ↗ | 9 % | ↗ |
| Named entities (NE) | | 0.646 | ↘ | 0.833 | ↗ | 5 % | ↗ |
| Nominal compounds (NC) | | 0.831 | ↘ | 0.746 | ↘ | 15 % | ↗ |
| $1 \cdot \mathrm{sim}_{NE} + 1 \cdot \mathrm{sim}_{NC}$ | | 0.846 | ↘ | 1.242 | ↗ | 20 % | ↗ |
| $2 \cdot \mathrm{sim}_{H1} + 1 \cdot \mathrm{sim}_{NE} + 1 \cdot \mathrm{sim}_{NC}$ | | 0.875 | ↗ | 2.906 | ↗ | 28 % | ↗ |
| $1 \cdot \mathrm{sim}_{H2} + 2 \cdot \mathrm{sim}_{NE} + 0 \cdot \mathrm{sim}_{NC}$ | | 0.872 | = | 1.987 | ↗ | 93 % | ↗ |

compared. For the nominal compounds descriptor, this result was due to an intrinsic property of the corpus. Indeed, it seems that some revisions of a piece of news were a translated version. This probably comes when an article was translated from a foreign language. As a consequence, despite the fact they point the same concepts, nominal compounds couldn't match.

## VII. CONCLUSION

This paper has given an account of our work to build a derivation corpus, to set an appropriate evaluation protocole and eventually to evaluate some original descriptors. We believe that the methods we used can open up new paths for the studies of derivation detection. We provide a derivation corpus with revision relation for press texts which constitutes a concrete contribution to the scientific community since no resource were available for studying derivation in French. In addition thanks to an inherent property of the corpus source, it can be extended to include translation derivations. It is freely available and can be easily integrate to the PAN corpus because of its file format compatibility. Concerning our results, we show that descriptors such as 1-gram hapax and nominal compounds can provide a substantial gain in terms of signatures storage and comparison costs with only a slight loss of general performances. Our manual analysis shows that in regard to the text material of a news, these linguistic descriptors can play an important role to discriminate or characterize a text but their impacts remain quite sensitive to the size of the compared texts. Further research should investigate the temporal expressions (dates, times) and the numerical expressions (quantities, monetary values, percentages) as well as the named entities and the nominal compounds variations to enhance the capibility of these descriptors. In addition, more research needs to be undertaken to see whether it is possible to filter the common lexicons, by $\mathrm{tf} \cdot \mathrm{idf}$ for example.

## REFERENCES


[1] S. M. Z. Eissen and B. Stein, "Intrinsic plagiarism detection," in *Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, 2006, pp. 565–569. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.110.5366

[2] A. Aizawa, "Analysis of source identified text corpora: exploring the statistics of the reused text and authorship," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, vol. 1, 2003, pp. 383–390. [Online]. Available: http://portal.acm.org/citation.cfm?id=1075145

[3] N. Shivakumar and H. Garcia-molina, "Building a scalable and accurate copy detection mechanism," in *Proceedings of the 1st ACM International Conference on Digital Libraries (DL 1996)*, 1996, pp. 160–168. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.6064

[4] P. Clough, "Measuring text reuse," Ph.D. dissertation, University of Sheffield, mar 2003.

[5] C. Lyon, R. Barrett, and J. Malcolm, "Plagiarism is easy, but also easy to detect," *Plagiary*, vol. 1, pp. 1–10, 2006.

[6] A. Z. Broder, "On the resemblance and containment of documents," in *Compression and Complexity of SEQUENCES 1997*, 1997, pp. 21–29. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.779

[7] N. Heintze, "Scalable document fingerprinting (Extended abstract)," http://www.cs.cmu.edu/afs/cs/user/nch/www/koala/main.html, 1996. [Online]. Available: http://www.cs.cmu.edu/afs/cs/user/nch/www/koala/main.html

[8] U. Manber, "Finding similar files in a large file system," in *Proceedings of the USENIX Winter 1994 Technical Conference*, October 1994, p. 1–10. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.3222&rep=rep1&type=pdf

[9] S. Brin, J. Davis, and H. Garcia-molina, "Copy detection mechanisms for digital documents," in *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD 1995)*, 1995, pp. 398–409. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.8485

[10] M. Henzinger, "Finding near-duplicate web pages," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, E. N. Efthimiadis, S. T. Dumais, D. Hawking, and J. e. Kalervo, Eds. ACM, 2006, p. 284. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1148170.1148222

[11] Y. Bernstein, M. Shokouhi, and J. Zobel, "Compact features for detection of near-duplicates in distributed retrieval," in *Proceedings of the Symposium on String Processing and Information Retrieval*, 2006, pp. 110–121. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.88.3243

[12] R. Gaizauskas, J. Foster, Y. Wilks, J. Arundel, P. Clough, and S. S. L. Piao, "The meter corpus: a corpus for analysing journalistic text reuse," in *Proceedings of the 2001 Corpus Linguistics Conference*, 2001, pp. 214–223. [Online]. Available: http://nlp.shef.ac.uk/meter/

[13] H. Yang, "Next steps in near-duplicate detection for erulemaking," in *Proceedings of the 7th Annual International Conference on Digital Government Research (DG.O 2006)*, 2006, pp. 239–248. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.111.3732

[14] M. Potthast, B. Stein, and P. Rosso, "An evaluation framework for plagiarism detection," in *Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010*, 2010.

[15] K. W. Church and W. A. Gale, "Inverse document frequency (IDF): A measure of deviations from poisson," in *Proceedings of the Third Workshop on Very Large Corpora*, 1995, p. 121–130.

[16] N. Fourour, E. Morin, and B. Daille, "Incremental recognition and referential categorization of french proper names," in *Proceedings of the Third International Conference on Language Ressources and Evaluation (LREC 2002)*, vol. 3, 2002, pp. 1068–1074.

[17] F. Cerbah, "Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms," in *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, 2000, pp. 145–151.

[18] B. Daille, "Conceptual structuring through term variations," in *Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 2003, pp. 9–16.

[19] A. Abeillé, L. Clément, and F. Toussenel, *Building a treebank for French*. Kluwer Academic Publishers, 2003, pp. 165–187.

[20] T. C. Hoad and J. Zobel, "Methods for identifying versioned and plagiarised documents," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 203—215, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.2680

[21] D. Metzler, Y. Bernstein, B. W. Croft, A. Moffat, and J. Zobel, "Similarity measures for tracking information flow," in *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2005, pp. 517–524.