Was it better before? Automated Quotation Detection in Ancient Texts Evaluating today's approaches on yesterday's content

Samuel Gesche, Előd Egyed-Zsigmond, Sylvie Calabretto

Was it better before? Automated Quotation Detection in Ancient Texts

Evaluating today's approaches on yesterday's content

Samuel Gesche, Előd Egyed-Zsigmond, Sylvie Calabretto

Univ Lyon
INSA de Lyon, LIRIS
7 av Capelle
69100 Villeurbanne
Email: {prénom.nom}@liris.cnrs.fr

RESUME. *Cet article s'intéresse à l'application des travaux en recherche de citations dans le contexte de documents anciens en langue grecque. La notion de citation est définie dans le contexte de ces documents, et les approches automatiques –statistiques et sémantiques– qui permettent de les découvrir sont évaluées à la lumière de cette définition et des ressources disponibles. Nous étudions également les effets spécifiques à notre corpus sur les métriques de performance.*

ABSTRACT. *This work evaluates quotation finding approaches in the context of ancient Greek documents. We explore the notion of quotation as relevant to the first centuries of our era, and we discuss the efficiency and usability of unsupervised approaches, both statistical and semantic, used in more modern cases. We also discuss some characteristics of our corpus relatively to performance metrics.*

MOTS-CLES : *détection de citations, corpus antique, rappel, évaluation*

KEYWORDS: *quotation detection, ancient corpus, recall, evaluation*

1. Introduction

Detecting quotations and references has become the focus of numerous projects, and the applications of this field of research are as various as detecting plagiary, studying the influences between authors or artists, or even predicting news spread across the Internet.

This work stems from the encounter between two fields: Patristics, which studies ancient documents and the quotation relationships between them, and Information Retrieval, which uses digital algorithms to match queries and documents. The story may have been that simple: the ancient documents have been digitalized, they are in a relatively homogeneous format, and the issue is a known one: detecting quotation between them. Moreover, the quotations were mostly *already* detected, so the issue was merely to find their exact boundaries.

In this paper, we present the challenges brought by quotations in ancient documents. We discuss the definition of a quotation in the context of Antiquity as opposed to the modern definition, and we try to map the modern approaches towards the ancient material. We also present the challenges brought by the corpus itself, and how it impacts the usual performance metrics such as precision, recall and f-measure. Our work has been mainly directed towards statistical linguistics, but we also discuss statistical semantics based on a few results.

2. Related Work

2.1. Related fields

Quotation detection is a research field that grew increasingly important during the last years, benefiting from the more large-scale interest for text reuse and plagiary detection that is crucial to fields such as the patent business or copyright enforcing.

One great asset of this field is that it can be considered as a specific subfield of information retrieval, using any part of a whole document as a query. Of course, using such large-scale querying means that specific measures must be taken to ensure that the answers can be found in a reasonable amount of time; however, both the approaches and the metrics (notably precision, recall and f-measure) can be reused to a certain extent, as long as the core of the problem remains similarity detection (a process for which (Lukashenko *et Al.*, 2007) or (Bao *et Al.*, 2006) recall most approaches and paradigms).

However, this means that plagiary and quotation detection also suffer from the same difficulty as information retrieval, which is the translation of the query -here some part of a text- into a form that will allow finding relevant documents -those that quoted that part of text- in a large corpus. Therefore, if the detection of literal quotations is as easy as matching words, in most cases more advanced strategies are required. There is a gradation between literal, erroneous literal, literary or

paraphrased quotations and allusions (references that sometimes require large context awareness) that is not unlike the gradation between lexical, syntactic, semantic and pragmatic levels of analysis. Plagiary detection has shown advances, for example in paraphrase detection ((Faisal *et Al.*, 2012), (Jo *et Al.*, 2007)), but even then relies more on knowledge of the actor's mindset -the psychological mechanisms of plagiarism and paraphrasing- than on crude text analysis heuristics. There is also the case where text reformulation merely serves the purpose of merging the quotation within a new discourse, and the work is far easier then. (Ernst-Gerlach *et Al.*, 2008) work on this kind of reformulation, including errors, word addition, deletion or change, and language evolution. Lastly, semantic approaches emerge as semantic resources are made available. For instance, (Nawab *et Al.*, 2012) uses synonymy to enhance n-grams overlap detection.

Automatically detecting allusions is still very difficult unless specific semantic or pragmatic resources are tailored to the task. *Meme* (recurrent ironic representation) tracking on the Internet (Leskovec *et Al.*, 2009) may be the beginning of an answer, even though it focuses on quick mapping of the spread of a widely-used, quickly mutating formula and not on detecting less widely used formulae over a large time frame.

2.2. Quotation detection in antique texts

(Ernst-Gerlach *et Al.*, 2008) proposes an approach for discovering references in a Latin text corpus. This work provides a typology for the text differences inherent to the mechanics of referencing (regular and irregular differences, deletions, insertions and substitutions). The proposed method assumes the availability of simple resources such as proper nouns (characters, authors, ethnic groups, places *etc.*) and number format variations (digits, letters, abbreviations). Their algorithm is based on the concept of sliding overlapping windows and is very tolerant, two identical words within a window being sufficient to assert a quotation. The method is tested on quotations taken from a Latin dictionary (basically using the examples of an entry as the quotations to find), and these quotations are looked for in the Perseus corpus[1]. The results have a good recall but a varying precision. Among the reasons for the lack in precision is the impact of short words or that of not taking word order into account. They suggest taking into account the amount and frequency of terms, their orders and stop-words. The expected results are also far worse on short quotations (especially one word quotations, which are more than half their corpus).

(Lee J., 2007) takes another approach to antique text and provides many elements on the specificities of these texts and their languages. Applying cosine similarity to verses of the Greek New Testament, and consolidating highly similar verse groups, this work is able to find many quotations within the three synoptic

[1] Perseus Digital Library, *http://www.perseus.tufts.edu*, accessed on 2013-02-07, updated in 2013. This is the digital library of Tufts University, containing 69M words worth of resources in Ancient Greek and Latin. The 'Perseus Hopper' provides not only the resources, but also text processing tools to work on them online.

Gospels. One of the main hypotheses is that quotations follow the same text order in both the quoted and the quoting texts. The approach is evaluated against nine quotation tables obtained from nine different experts, and the block approach is thus validated empirically.

(Büchler M. *et Al.*, 2010) studies text reuse for the purpose of discussing the available versions of the texts. This work explains the difference in the practice of quotation between Antiquity and nowadays (and in particular the absence of the explicit reference). The work also adds a visual analysis over the text processing aspect, which is enhanced itself using the Perseus lemmatizer. The described algorithm is based on n-grams and prefix filtering and aims at finding similar areas in the texts. It then performs a more semantic analysis using significant words co-occurrence in sentences. The approach is able to take into account language evolution and dialect change, as well as word omission and substitution. However, it cannot discriminate between quotations and large idiomatic expressions, and is not good at finding the exact quotation boundaries.

Recently there have been methods using machine learning and neural networks to create numerical (vector) representations of words (Mikolov*et al.*, 2013-1; Mikolov*et al.*, 2013-2) and short texts (Le *et al.*, 2014) based on a large corpus. We are studying them.

3. Project Context

The purpose of our research is to find references of a work within another. Our focus is on ancient (*Koiné* Greek) documents that reference both the Tanakh and the New Testament (which will eventually be known together as the Bible). We do not delve deeply into the issue of versioning, although it is a really relevant problem in this research field, and one we had to clear beforehand (we still present it among the challenges we had to overcome).

Quis dives salvetur, by Clement of Alexandria
...
chapter 42
paragraph 1
line 1: Ἵνα δὲ ἐπιθαρρήσῃς οὕτω μετανοήσας ἀληθῶς ὅτι σοὶ μένει
line 2: σωτηρίας ἐλπὶς ἀξιόχρεως ἄκουσον μῦθον οὐ μῦθον ἀλλὰ ὄντα
line 3: λόγον περὶ Ἰωάννου τοῦ ἀποστόλου παραδεδομένον καὶ μνήμῃ πεφυλαγμένον
paragraph 2
line 1: ἐπειδὴ γὰρ τοῦ τυράννου τελευτήσαντος ἀπὸ τῆς Πάτμου
...

Figure 1 - Example of a patristic structured text (beginning of chapter 42 of 'Quis Dives Salvetur' by Clement of Alexandria)

Our research corpus is composed of the Septuagint (the *Koiné* translation of the Tanakh that was available at the time), the Greek New Testament, and around 700 works, mainly of Church Fathers, which amounts to approximately ten million words.

Documents are structured and therefore can be referenced easily. Biblical texts are structured as books/chapters/verses, and patristic texts as work/chapter/paragraph/line (we discarded the page structure because it depends too much on which edition is used). This allows a single word to be located unambiguously. Figure 1 shows an example of patristic text with its logical structure.

4. Challenges

4.1. Koiné Greek

Koiné Greek is a language that gained relevance following the conquests of Alexander the Great, went global during the hellenistic period and the Roman empire, was supplanted by Latin around 300 A.D., and developed further within the Byzantine empire, becoming the much different Medieval Greek in the process.

Koiné Greek was therefore the global language during the first centuries in the Roman Empire, which means that it was constantly evolving to meet new needs, new markets and new concepts, much like English nowadays. This means that *koiné* Greek is really different from both the former Classical Greek and, of course, the Modern Greek. The consequence is that semantic resources tailored to either classical or modern Greek cannot be used easily, if at all.

There are, however, some things that remained constant:

– The alphabet did not evolve much: as a matter of fact, the Unicode standard for polytonic Greek works for both classical and *koiné* Greek (we will point out however that there are still other encoding standards, which still causes conversion issues);

– The language itself remains based on inflection, which means that the word order within a sentence bears no importance, at least concerning the meaning of the sentence (this means in particular that a quotation can reorder words); unfortunately, we have to add here that in most antique manuscripts, there is no visual indication of sentence beginning or end. Depending on the digital source, we thus may or may not be able to split the text into sentences.

4.2. Quotation in the Antiquity

Quotation finding, and more generally finding any kind of text similarities between documents (including, of course, plagiary), is a research field that still mostly applies to quotation, and text similarity, as it is defined today. We quickly realized that, in order to work on ancient documents, we had to define what a quotation was at the time.

Quoting other people -be it written text, vocal speech, gestures, or other means of communication- is intrinsically a part of the communication process. While the

core concept stayed mostly the same (which is, recalling or telling what someone has expressed), the actual process of quotation, and the final result, has evolved because of several tremendous revolutions. Ancient texts come after the development of written language (with oral tradition still being very strong and as likely to be quoted as written text), but since then, the world has seen the development of the printing press, which made copy of written texts available for every scholar and every library, and the digital revolution, which made accessing texts, copying, quoting, transforming and broadcasting them trivial, to the point where the concept of document itself may not be clear anymore (Pédauque R. T., 2007).

In the Antiquity, texts (physically tablets, parchments, papyri or codexes) were not as readily available as nowadays. Copying was a long -and therefore expensive- process. However, during the first centuries of our Era, road were mostly secure and travel was quick along the commercial axes, so travelling to read a specific book -and to copy some parts of it- was possible, as was transporting a specific text (apostles' epistles, for instance, initially were transported from community to community to be read, while the Great Library of Alexandria, as well as numerous others, were busy with scholar visitors).

Taking this context into account, we can divide references into three types: literal quotations, non-literal quotations, and allusions.

– Literal quotations -today a portion of text put between brackets that must be lexically *strictly identical* to the original text- was at the time merely *exactly corresponding* to the original text, as a result of the writer either having the original text before his eyes, or having a faithful memory of the text. In other words, an ancient literal quotation says exactly what the original says, but not necessarily with the same words.

– In contrary, non-literal quotations are quotations that do not strictly correspond to the original text; either because of a partial memory, or by a deliberate change.

– References are not quotations *per se*, or they are implicit quotations -or cross-references. While lexically they do not contain any form of similarity to the original text, they do reference this text. Often, an analysis at semantic level is not sufficient to detect such a reference: the proper level is the pragmatic level. This is for example where the use of that very characteristic formula leads to considering that this author has been influenced by some work of this other author -which can be a material for thesis rather than automated digital process.

If we were to provide a single relevant number, it would be that 20% of the references we worked on do not have even a single lemma in common (yet lemmatization drops this number from an initial 43%).

What remains clear in that context is that searching for lexical similarities, while helpful, will not be nearly as successful as in a modern context. And since semantic resources that may be available for Modern Greek are irrelevant for *Koiné*, to use any approach by this angle we had to build them first.

4.3. Versioning

We will not present this challenge in details, since there is little we can do about it: these issues must be solved by the researchers in Humanities. They do, however, put a hard limit on our ability to find some references, so we will mention them briefly.

There are mainly two aspects of versioning:

– Present versioning, or the issue of the availability today of the documents that were quoted yesterday;

– Past versioning, or the knowledge of which of the then available versions of the document was quoted.

4.4. Available resources

4.4.1. Lemma and morpho-syntactic analysis

The available resources were mainly lemma and morphological analysis for a number of word forms in ancient Greek. We used three main sources: Perseus, initially through the Archimedes Project Morphology Service[2], then through their database dump available for download; BibleWorks[3], a Bible-focused software who provided an export for morphologically analysed texts; finally, the lemma tables of Sources Chrétiennes[4], built over the last years. In total, we got over 380 000 lemmatized ancient Greek forms, discounting homonymy and polysemy.

Without a morphological analyzer for our own texts, and without a morphological analysis accompanying the lemma tables of Sources Chrétiennes, we decided to limit our language processing to matching a term with a lemma from these lists. Of course, we were then vulnerable to the issues of homonymy and polysemy. We thus took a priority rule as follow: first, if the term is already a lemmatized form (that is, if one of the lists has the form identical to one of its possible lemmas), the form itself is kept. If it is not the case, we take the first lemma present in these tables for the form, prioritizing the tables according to their focus to Patristics:

– Sources Chrétiennes first, with 21 505 lemmatized terms in *koiné* Greek;

– then BibleWorks, with 115 498 lemmatized terms in *koiné* Greek;

– finally Perseus and its 318 584 lemmatized terms spread over multiple forms of ancient Greek language, including *koiné*;

All in all, we were able to find a lemma for 40% of the forms of our entire text corpus (however, several texts were entirely lemmatized, and we experimented on them), and 24% of the available lemmatized forms were present in the text corpus

---

[2] http://archimedes.mpiwg-berlin.mpg.de/arch/doc/xml-rpc.html
[3] http://www.bibleworks.com/
[4] http://www.sources-chretiennes.mom.fr/

(78% if we discount Perseus due to its language spread). Figure 2. presents the distribution of the lemmatized forms within the different sources, and the contribution of these sources to the lemmatization of the forms present in the corpus.

As a side effect, we got the list of around 5 000 proper nouns used in the Septuagint.

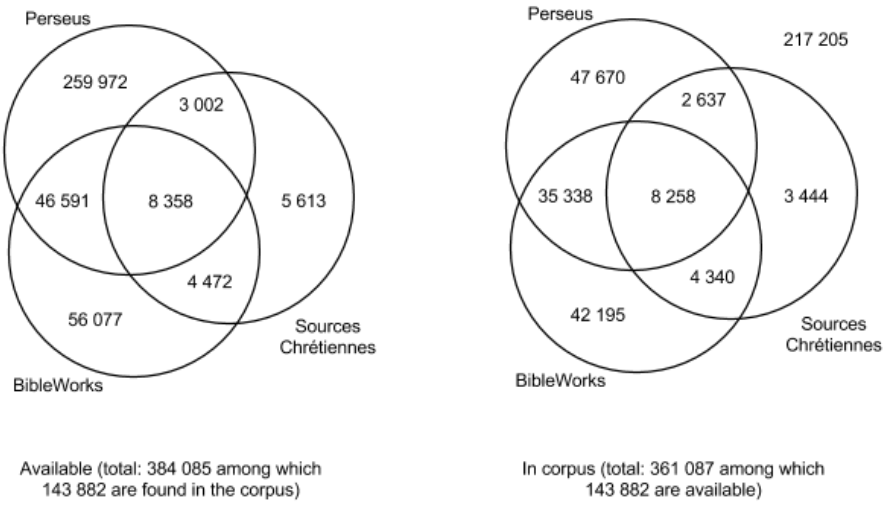

Figure 2. Distribution of the available and necessary lemmatized forms among the three sources Perseus, BibleWorks and Sources Chrétiennes (left: forms available for lemmatization; right: forms both available and actually present in our corpus - the outer number represents the amount of forms for which none of the sources provide a single lemma)

Automated lemmatization (Dimitrios *et Al.*, 2008) may greatly enhance these results but we have not delved into it currently, since it was not the focus of our research. The same is true for stemming, which we could not achieve with simple approaches (Greek is a not an easy language to stem), and for which we could not find lists of already stemmed forms.

4.4.2. Stopwords and other recurring terms

We have a very narrow list of stopwords, the experts having determined that a larger list did not reflect the list of basically disposable terms. We have 16 lemma in this list (οὖν, τε, ὁ, καί, δέ, γάρ, τίς, ἤ, μέν, μήν, δή, οὗς, γέ, ἄρα, αὐτός, πότε). Figures 3 to 6 show this list to be sufficient (there are not many common words left between the references in the control set).

Using a table of the most frequent terms, we built two more lists, a list of recurring terms in the language (111 lemma such as ὅσιος - hallowed), and a list of recurring terms in the corpus (870 lemma such as κύριος - lord). These lists allow us

to better adjust the sensibility of our algorithm, and better, to use each list at the relevant steps (for instance, a stopword can be eliminated from the start, but recurring words will typically be used to discard results that would only contain them).

On top of that, we listed recurring expressions as well (24 in the language and 202 in the corpus).

We called those recurring formulae, terms or expressions nonquotes. While they do contribute to the meaning of the text, and as such cannot be simply discarded, their presence cannot be used alone as an argument to support a quotation.

4.4.3. Semantic resources

Semantic approaches are greatly limited as long as they rely on resources that are highly language-dependent. However, statistico-semantical approaches remain usable as long as their underlying hypotheses on language are valid for koiné Greek. (Büchler *et Al.*, 2010) actually uses the statistico-semantical tool of significant terms co-occurrence, and (Ernst-Gerlach *et Al.*, 2008) uses names as hints for a greater chance of quotation. Other semantic analysis methods such as LSA may yield results too.

4.5. Corpus-specific challenges

4.5.1. Boundaries detection and precision

Precision is defined as the proportion of found answers that are good answers, and recall as the proportion of good answers that are found. In our case, however, asserting whether a found quotation is a good result is difficult, as is asserting whether a quotation was found.

Firstly, we have a list of results to find -results that are taken from already published material. These results are, however, often approximate, to the point where one of our goals is to find the actual boundaries of the quotations within them. We know where these results begin, and often not where they end. So, how do we decide that a quotation is good? Fortunately, in our small test sample (329 references total, including allusions that do not have any word, lemma or otherwise in common), a manual work has been done to ensure that we know the boundaries at the granularity of words. But how is this number sufficient to assert the performance of our algorithm?

We eventually decided that if a quotation that we found overlaps with a quotation that must be found, even with a single word, we will take it as a good result. Obviously, overlapping must occur in both texts.

Secondly, another of our aims is to find new quotations -quotations that have not been detected, or documented, in the previous centuries. Digital processing allows being exhaustive, if biased by the limits of an algorithm, and we already found several new quotations during our tests. While recall does not suffer from this aspect

of the project, we have to first pass the results to a second process -in this case, a manual study- to assert whether a false result (as defined by the control list) is actually a true negative that should enrich the original list, thus irremediably corrupting the independence between the expected results and the algorithm that we want to test: we can only include the true negative if we find them, which leads to an artificial increase in performance that does not reflect an actual improvement. We cannot artificially enhance its performance by biasing the test sample.

We can of course tag these results as 'good but not mandatory', thus keeping the independence, and simply not count them. But it still means that we have to manually control the results before asserting a precise precision score.

4.5.2. Self-quotation and recall

Lastly, we have a specific bias that comes from the very text we study: the Bible is a web of self-quotations, with sometimes more than 20 candidate passages for a single quotation. In this case, which of the texts is the right one? The one in the list of results is, obviously. But can we hold an algorithm -that basically matches strings- accountable for deconvolving the strata of successive quotations and finding the right author that was quoted, as opposed to those he quoted and those who quoted him? In many cases, a pragmatical analysis is necessary. In some extreme cases, manual research is necessary.

We first sought to differentiate real quoted text from mere recurring formulae -either in the language or in the specific corpus- and it brought us to expand the notion of stopwords to nonquotes as defined in Section 4.4.2. Then, we looked with the experts for a threshold for the amount of candidates; either there were less candidates, and they were all considered right, for needing a manual validation (we had actually quotations that were considered enriching by the experts, quotations that they wanted to have as results, even if not formally the right one); or the amount of candidates exceeded the threshold and they were all considered false, for having found the right answer for wrong reasons.

Intertwining both processes, searching for nonquotes in the numerous candidates, was very effective at building our lists of recurring expressions. The drawback, however, was the risk for them to become highly sensitive to the specific kind of text we were testing with, and to bias the results once more. This means that the search for recurring formulae must be integrated as a learning process in the algorithm itself.

We thus defined both a *raw* score -precision and recall using the amount of found answers- and an *efficient* score, factoring that last bias and the according threshold. The efficient score basically takes into account the fact that the results, in our case, will be studied by human experts.

It is worth noting that while in the first case a single number serves as the dividend of both precision and recall, in the refined version the dividend of precision is greater than the dividend of recall (factoring the numerous cases where more than one answer is deemed correct). Besides, the impact of this correction on the recall

depends on the amount of found answers (which normally only impacts precision). This impairs sets of parameters that increase the amount of candidates: in addition to being subject to low precision, they also get low recall. Empirically, we obtained a factor of 2/3 between raw recall and efficient recall with a threshold of 5.

5. Evaluation

5.1. Algorithm

In order to test the approaches that we presented in section 2, we designed a general algorithm based on the usual steps:

– Pre-processing, including documents import and encoding resolution, tokenizing, and optionally stopwords deletion and language processing.

– Main processing: matching word n-grams from both texts and computes the exact quotation boundaries.

– Post-processing: preparing the found quotations to match the quotation granularity of the control set, then computing the relevant metrics (both raw and efficient precision, recall and f-measure) as well as a more verbose report (allowing us to detect for instance which approach is better at finding some set of quotations).

During the main processing, we tested various approaches and heuristics, among which:

– (Ernst-Gerlach *et Al.*, 2008): overlapping windows algorithm, fundamentally equivalent to a n-gram approach where we search for 2 shared terms in a n-gram the size of the window, and then merge the overlapping found results; we also tested their heuristic of expanding the window whenever a proper noun is found; we did not test the impact of having different formats for numbers. We did test the impact of word order as discussed in their perspectives.

– (Lee J., 2007): merging two non-overlapping quotations if there were less than 50 words between each other. We did not however use the full consolidation process due to the difference between the corpuses (the method described in this work is especially effective in the studied case).

– (Büchler M. *et Al.*, 2010): using co-occurrence of significative words as a hint for a more semantic measure. We used the approach from (Mousselly-Sergieh, H. *et Al.*, 2013) to infer the semantic similarity.

5.2. Experimental sets

We defined two experimental sets to test the approaches. They differ both on their scale and on the accuracy of the control set of quotations.

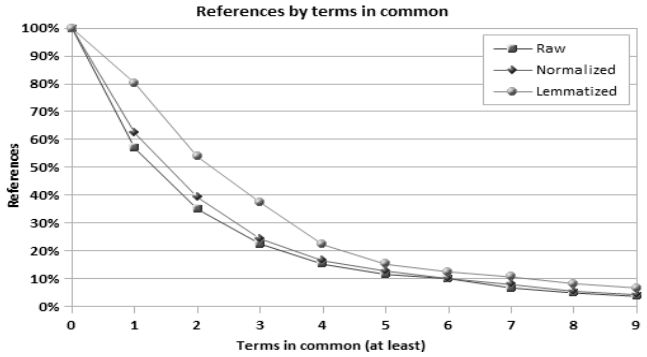

Figure 3. Amount of references to the Bible within Quis Dives Salvetur where both texts have at least a given amount of common terms (either the terms themselves, their normalized version, or their corresponding lemma). Stopwords are discounted. Raw: refers to the original text, Normalized to the normalized text, and Lemmatized to the text where words are reduced to their lemmatized form.

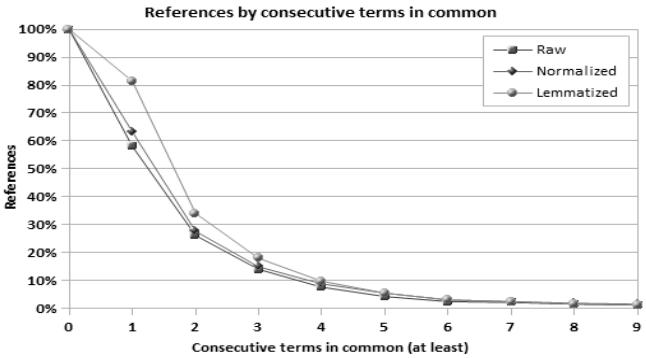

Figure 4. Amount of references to the Bible within Quis Dives Salvetur where both texts have in common at least a given amount of consecutive terms (either the terms themselves, their normalized version, or their corresponding lemma). Stopwords are discounted.

The first experimental set uses a single document, *Quis dives salvetur*, quoting the Bible as composed of the Septuagint and the Greek New Testament. To control our results, we have a narrow set of 329 references, the exact boundaries of which have been produced. These references include 19 allusions. The breakdown of the similarities between the quoting text and the quoted text are presented in Figure 3 (identical terms) and 4 (amount of consecutive identical terms).

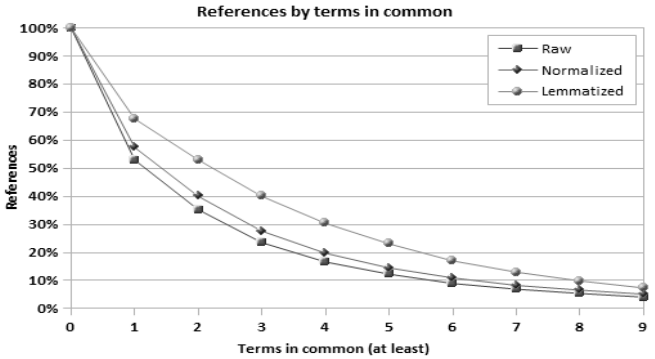

Figure 5. Amount of references to the Torah within the works of Philo of Alexandria where both texts have at least a given amount of common terms (either the terms themselves, their normalized version, or their corresponding lemma). Stopwords are discounted.

The second experimental set uses a whole author's publication (namely Philo of Alexandria), quoting the Septuagint. We only use the references to the Torah (the first five books of the Septuagint) as a control set, and we only know in which paragraph of the Philo texts the reference is. We do not know how many of them are allusions. The breakdown of the similarities between the quoting text and the quoted text are presented in Figure 5 (identical terms) and 6 (amount of consecutive identical terms). It is worth noting that this case shows even less similarities between the texts than the previous case, even though the quotations are more loosely defined.

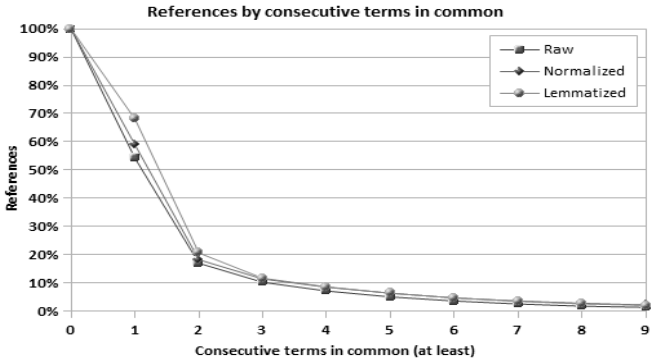

Figure 6. Amount of references to the Torah within the works of Philo of Alexandria where both texts have in common at least a given amount of consecutive terms (either the terms themselves, their normalized version, or their corresponding lemma). Stopwords are discounted.

5.3. Results

We run the algorithm on those sets several hundreds of times with varying parameters. Overall, we were able to find up to 30% of the references (without resorting to extremely loose settings). The results were comparable between both sets of experiments.

Some of the parameters had massive impact:

– Discarding stopwords had the most noticeable impact, always bringing better results (especially in the case of our effective scores). We quickly stopped experimenting with stopwords allowed, so the following results all include stopword deletion.

– Merging found references that were less than 50 words distant (in both texts) greatly increased precision and effective recall (limiting the amount of candidates for a given reference). We also quickly included this action by default in the experiments.

– The size of the n-grams (how many similar words have to be found) and the tolerance (haw many differences are allowed between similar fragments) had obviously a definite influence on both precision and recall.

- The former set a compromise between them, smaller n-grams bringing lower precision and higher recall. However, the increase in effective recall receded quickly when the size of the n-grams went lower than 3, so we kept 3-grams as an optimal size, from the original 5 taken from (Ernst-Gerlach *et Al.*, 2008). The F-measure was actually better for larger n-grams, but the amount of results was too low to be of use (on the first set, a mere 20 results for 5-grams compared to several hundred for 3-grams).

- Tolerance showed a sizeable increase in recall (averaging 60% all other conditions equal) without affecting precision nearly as much (a mere 20% loss). Interestingly, the numbers are the same in the case of both the usual recall and our more constrained effective recall. Our optimal setup was searching for 3 common terms within 5-grams.

Among the parameters that had either low or negligible impact:

– Nonquotes filtering had no significant impact outside of experiments which did not discard stopwords (in which case the filtering was not enough to warrant interesting results).

– Using proper nouns as a hint to widen the frame had no discernible impact, probably due to the fact that few of the references that were not found displayed a name.

– Discounting word order had a larger impact on lowering precision than on increasing recall, especially with lemmatized text.

– Text processing, rather than simply improving the results, moved the compromise between precision and recall. Figures 3 to 6 show the beneficial impact of lemmatization in references, but the same impact is true outside of them. In

average, the results with effective precision and recall were better with unprocessed or normalized text.

– We only tested a single semantic approach, but the results were along the lines of text processing and n-gram size: the better the measure was at merging terms, the farther the results moved along the precision versus recall axis (and in the case of effective scores, towards lower overall scores).

These results tend to show the following features of our corpus:

– There is, as we showed in section 4.3, a proportion of quotations in which there are common terms, but these terms do not follow one another exactly. In other words, there will often be an edit distance between a source text and the quoting text.

– The general heuristic that stopwords limit the efficiency of statistical text matching holds true. However, there is no need to have a large list (since filtering other common words does not have a significant impact).

– The general heuristic that merging terms that have a close meaning (through text processing, semantic measures and other means) increases recall at the cost of precision also holds true. In our case, the benefit is not apparent, especially because we have constrained the recall to take into account a manual processing of the results and the highly self-quoting nature of the source document.

– Most of the references still cannot be found, even with highly permissive parameters. This is to be expected when half of them do not even have two lemma in common.

6. Conclusion and perspectives

We discussed the issue of detecting quotations within ancient documents. We took into account the specificities of language, culture (through the practices of quotation) and of the corpus itself. Using several approaches that were used in similar cases, we evaluated the main tools offered by statistical computation and found that even though they were not a sufficient solution, they worked as usual in the case of quotations that show text similarities.

This work leaves us three options to increase our performance at detecting quotations in ancient Greek texts. The first two revolve around the heuristic of merging similar terms using statistical methods. We can use our partnership with Greek experts to build a stemmer for this language, and we can use new semantic algorithms to improve our similarity measure. Increasing the amount of common terms may allow us to increase the similarity threshold of quotation detection and improve precision. The third option is finding heuristics within the metadata that can be obtained on the texts themselves. For example, if we know that some text is a commentary of some book –and there are many such cases, we can rely on this information to resolve most cases of multiple source candidates.

Another perspective is the study of the efficiency and scalability of statistico-semantic methods. We are currently working on the implementation of methods

such as word2vec (Mikolov,*et al.*,2013) and the improvement of memory efficiency of the algorithm of (Mousselly-Sergieh *et al.*,2013) based on stream processing methods.

7. References

Bao J.P. *et Al.* (2006). A fast document copy detection model. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 10, n° 1, p. 41-46.

Büchler M. *et Al.* (2010). Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, vol. 1, n° 2, p. 1-17.

Coulie B. (1996). La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs. *Byzantion : revue internationale des études byzantines*, vol. 66, p. 35-54.

Dimitrios P. L. *et Al.* (2008). Applying similarity measures for automatic lemmatization: a case study for modern Greek and English?. *Proceedings of Int. J. Artif. Intell. Tools 2008.*

Ernst-Gerlach A. *et Al.* (2008). Identifying quotations in reference works and primary materials. *Research and Advanced Technology for Digital Libraries*, vol. 5173, p. 78-87.

Faisal A. *et Al.* (2012). Analysis and extraction of sentence-level paraphrase sub-corpus in CS education. *Proceedings of ACM SIGITE 2012.*

Jo C. *et Al.* (2007). New Functions for Unsupervised Asymmetrical Paraphrase. *Journal of Software*, vol. 2, n° 4, p. 12-23.

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of The 31st International Conference on Machine Learning* (pp. 1188–1196). Retrieved from http://jmlr.org/proceedings/papers/v32/le14.html

Lee J. (2007). A computational model of text reuse in ancient literary texts. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Leskovec J. *et Al.* (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.*

Lukashenko R. *et Al.* (2007). Computer-Based Plagiarism Detection Methods and Tools : An Overview. *Proceedings of the International Conference on Computer Systems and Technologies* 2007.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Computation and Language*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., 0010, K. C., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *In NIPS* (pp. 3111–3119). Retrieved from http://dblp.uni-trier.de/db/conf/nips/nips2013.html#MikolovSCCD13

Mousselly-Sergieh, H. *et Al.* (2013). Tag Similarity in Folksonomies. *Proceedings of the XXXI INFORSID congress*, p 319-334.

Nawab R. *et Al.* (2012). Detecting Text Reuse with Modified and Weighted N-grams. *Proceedings of the ACM First Joint Conference on Lexical and Computational Semantics* 2012.

Pédauque R. T. (2007). *La redocumentarisation du monde*. Paris : Éditions Cépadues.