Textometry and Information Discovery: A New Approach to Mining Textual Data on the Web

E. MacMurray[1], **M. Leenhardt**[1,2],
[1]SYLED/CLA²T EA2290 UFR ILPGA Université Sorbonne Nouvelle Paris 3, France
[2]Le Semiopôle, Montreuil, France
[1] erin.macmurray@gmail.com, [2] marguerite.leenhardt@gmail.com

Abstract - *Most Text Mining tasks focus on local linguistic rules for detecting such elements as named entities, events and opinions: the goal here is to go beyond these local context boundaries by taking global dimensions into account. A robust method to mine textual data known as Textometry is not constrained by external resources and avoids problems such as the coverage limitations of standard dictionaries and at a higher level, domain-dependant resources. Textometry provides a new approach of exploring and comparing textual data. This paper studies the Textometric method and how it can be applied to the industrial context of mining named entities and their trends (opinions or events) in both French and American online news media: Le Monde and the New York Times. This paper focuses on bypassing certain costly steps in tasks related to mining information on Named Entities.*

Keywords: Textometry, quantitative linguistics, textual statistics, named entity mining, opinion mining

1 Introduction

It's no scoop that data- or the "quiet revolution" as Bollier [2] puts it- has grown tremendously since the availability of computing and databases, even more so since the dawn of the Internet. Data is not just conveniently stored in structured databases, it comes in the form of natural language: articles, blogs, forums are among some of the many formats in the mobile network for sharing information. This growing collection of content demands computer processing in order to dig or render visible information of interest. The detection and extraction of named entities in large compilations of text helps pin point potential zones of information corresponding to intense activity of the Named Entity in the media. In this paper we compare two intelligence application use cases where known statistical algorithms are applied as a method for mining information on named entities in online news articles from both Le Monde and the New York Times. Named Entities were used as an entry point for the analysis of the corpora. Then, statistical tools provided results for creating new Linguistic Resources (LR) in French and English for both the context of opinion analysis and event detection. This research puts forth a new approach to textual data analysis through methods for more industrial contexts, such as business and communication intelligence needs.

1.1 Mining techniques and natural language processing

Today, there are many natural language mining techniques: machine learning and information extraction through automatic semantic and morpho-syntactic patterns to name just a couple as discussed during the Message Understanding Conferences (MUC) [8]. Text Mining, generally seen as a subfield of Data Mining, is roughly defined as the processes used to extract and structure unstructured data [6]. Early work in text mining tried simply applying the algorithms developed for data mining without considering their specific unstructured nature [5],[9]. These applications showed how it was possible to use the methods of extraction sequences to identify new trends in a database [11]. However, textual data presents very different challenges from pre-structured data. Text Mining techniques often use a structuring phase of the information expressed in natural language in order to apply standard data mining strategies[6],[11]. The units of analysis used by these techniques rarely go beyond the sentence level and sometimes fail to consider their object of analysis, the text, as a component in and of itself. Here, our goal is to shift the focus from the sentence level to the text level by applying existing statistical strategies to discover patterns at this higher level in a corpus of unstructured textual data.

1.2 Data heterogeneity and web mining

Beyond the notion that textual data is unstructured, there is another major difficulty when putting in place a mining strategy dedicated to the qualitative analysis of textual content on the web- the heterogeneous nature of the data. As defined above textual data is unstructured information that simply keeps on expanding. Several factors must therefore be taken into account when developing a mining strategy. The first factor is the variety of physical media used to convey content: websites, social networks, forums, blogs, information portals. The structure of the content differs greatly depending on the medium. Although these media use meta-data in order to structure their online display and search possibilities, html/xml are only weak representations of the actual textual content. A second important factor in the process of producing content is the different writing strategies web-users exercise when exchanging on the web. They can, for example, generate a full text segment in writing an article or blog or simply intervene by leaving a comment on text already produced. Thirdly, in the search process, different types of textual data must be considered: headline, column name, lead, article, date and time, legends and paragraphs, among others. These types of data will yield different results when being mined for information. These three factors (physical media, writing process, and text segments) make the task of pertinent information extraction complex, in other words, the development of robust systems plays a detrimental role in the management of these different variables. Moreover, in order to perform analyses, some structure will have to be given to the bulk of information gathered from these online media.

The goals for mining natural language are therefore twofold: (i) structuring free text for use by other computer applications, and (ii) providing strategies for following the trends and/or patterns expressed in the text.

We focus here on the latter, presenting the Textometric approach and showing the advantages of this method for text and information analysis issues. To this end, two subtasks of Information Discovery are considered: (i) mining Named Entities and (ii) gathering information for opinion detection and trend analysis.

1.3 Named Entity Mining for Opinion Detection and Information Discovery

Information Extraction systems have long attempted to group textual elements into Named Entities and relationships or template scenarios between these entities [8], [15]. Named Entity Recognition (NER) and Relation Templates continue to be hot topics today as they were during the MUCs, which can be noted by the number of open source technologies that have begun to undertake this task. The definitions attributed to what are called entities and relationships remain unsatisfactory. Entities are roughly defined as names of people, organizations, and geographic locations in a text [8]. They are perceived as rigid designators that reference 'real world' objects organized in an ontology [16]. However, these definitions fail to take into account the semantic complexity of named entities in terms of their surface polysemy and their underlying referentiality which aims at combining both the linguistic designation of an entity and the extra-linguistic level or the 'real world' object an entity refers to [16].

The situation is similar for Opinion Mining (OM). There is a terminologic instability resulting from the coexistence of *sentiment analysis* versus *opinion mining*, *evaluative stances* versus *opinion expressions*. The objects of OM and Sentiment Analysis are thus not based on consensus. Simply put, the technologies supporting sentiment analysis are related to classification tasks, whereas OM is derived from mining tasks. Named Entity Mining is deeply related to OM and Sentiment Analysis tasks. In following definitions given in [22] annotation objects, such as *agent annotation* or *target annotation,* rely on NER for their information discovery tasks in commercial applications. Although our method has yet to provide a satisfactory definition of named entities, combining both linguistic and computer science considerations, these objects remain vital access points for uncovering zones of information in the corpus.

2 A new approach to mining the web

As previously discussed here, information extraction techniques have often used semantic or content annotations to structure information of interest [6],[8],[11]. However, using qualitative coding- usually in the form of such morpho-syntactic or semantic annotations- to drive quantitative conclusions almost defeats the purpose of discovering unknown information in the text. Content annotations are not an abstraction of what is actually expressed in the text, but rather the vision of annotator creating them. This calls into question the accurate interpretation of results acquired using such basic information extraction techniques. Following MUC guidelines, precision and recall remain the gold standards for measuring such systems. However, "one man's noise is another man's data" [2], which clearly points out the difficulty in creating a generic system that can objectively process large quantities.

2.1 Textometric approach

Textometry is already well rooted in social science studies and quantitative linguistic research [10][11], mostly developed in France with numerous pioneers, Pierre Guiraud, Charles Muller, Jean-Paul Benzécri, Ludovic Lebart and André Salem. According to this approach, a text posseses its own internal structure that would be difficult to analyze by manual means alone. By applying statistical and probabilistic calculations directly to the textual units of comparable texts in a corpus [10][20] it becomes possible to analyze patterns and trends that would otherwise be obscured by the quantity of the textual units.

Information extraction techniques using qualitative coding can, therefore, be bypassed when studying textual data. Indeed, even basic preprocessing steps, such as lemmatization, can potentially hide distinctive features of textual units. Although, Textometry is not generally considered a text mining technique by the industrial community, because it is not fully automated, in following broader Text Mining definitions [6],[20], it seems an appropriate strategy for discovering related elements in a corpus when no predetermined information model is available. The Textometric analysis process relies on the interaction between an expert user and the system. The validity of the result interpretation is provided by and depends entirely on the expert.

2.2 Textometric objects

Textometry consists of seeing the document through a prism of numbers and figures, producing information on the frequency counts of words, otherwise known as **occurrences**, whereas **forms** are a single graphical unit corresponding to several instances in the text [10]. This corresponds to the type/token distinction in Corpus Linguistics. It is also possible to calculate **Repeated Segments** (RS) which returns sequences of at least two consecutive units, or more, that occur several times in the corpus. These objects, forms, occurrences, RS, can be grouped together to create ad-hoc LR generated by the analyst. The resulting resources make up the analyst's "equipment" providing access to textual tendencies that could otherwise remain hidden by the quantity of data.

2.3 Textometric methods

The Textometric method segments a corpus into comparable zones of text. The news corpora used in the following examples are broken down into smaller groups of articles according to date and in one case according to the writing process (whole article attributed to one author versus user's comments on the article attributed to several authors). Using statistic and probabilistic calculations on the units within each zone, quantifiable information is derived providing the analyst with new knowledge of the textual data. Trends or patterns in the quantifiable information can therefore be observed across the predefined zones.

In this paper, the hypergeometric model[1] is applied to both text zones as well as their forms. In the first case, the calculation shows the statistical probability of a form to appear in a specific zone of the corpus in order to represent the form as having a degree of *specificness* or statistical significance for the zone it appears in. The result is a graphical representation of the *specificness* distribution for the selected forms as will be seen in the figure 1. In the second case, the same calculation is applied directly to a single form, otherwise known as pivot-form, in order to obtain graph or network of interrelated forms from the corpus as a whole or a single zone, table 3. The resulting relations are known as co-occurrences, or the statistical attraction of two or more words in a given span of text (sentence, paragraph, entire article) [14].

Both calculations (*specificness* and co-occurrences) will be used to observe named entities and their various designations over a period of time. Chronological analyses using these methods have already been carried on numerous news media sources ranging from portals such as Lexis-Nexis and Factiva [3] to the French national newspaper Le Monde [15].

In comparison with approaches that use qualitative coding, textual statistics would have a relatively low maintenance cost, due to the minimum amount of actual processing or human development of annotation models. Using relatively simple tokenizers these tools can be applied to a wide range of languages [4].

3 Intelligence Applications

The two use-cases presented here show how the results of Textometric calculations can be used for interpreting vital information from textual data. Whether an analyst is attempting to compile LR for future use or trying to discover relationships that certain points of interest entertain in the data, Textometric methods help shed new light on the intrinsic properties of text elements.

3.1 Named Entity reference detection for opinion mining

The raw material under study comes from a corpus archive of various French online newspapers originally made for commercial purposes of seeking insight on the image of the Socialist Party[2] from November 2008 to August 2009. The data from Le Monde were stripped from the archive for use in the following experiments. The textual data has been automatically extracted from the XML feed for each article and its available user's comments. The main entry of analysis for this kind of task consists in looking for information on NE, particularly political personalities. Typically, here, paraphrases of a NE are valuable because they allow the discovery of semantic variations related to how the NE is perceived. Indeed, whether the focus is on journalistic paraphrases or on nicknames given by Internet users, paraphrases are a major entry point to opinion mining. In the current use-case, this detection step provides candidates for the analysis of how the French President is depicted in web news and is seen through user's conversations.

A first entry to the Textometric analysis consists of using full-text search in the generated dictionary of lexical frequencies, coupled with search based on *regexp*. As a result, one can obtain a list from which derived forms can be selected and grouped in a set. These derived forms are highly informative concerning how a NE is represented in the textual material, with no regards to the linguistic performance of the comment and article writers. A second and complementary entry consists of calculating the *repeted segments (RS).* The **RS calculations** take text material as input and return an ordered set of objects that can be analyzed contextually. For example, Sarkozy has a number of lexical derivations (*Sarkozyste*, *Sarkoland*) as well as paraphrases (*M. Sarkozy*, *Président de la République*) that help determine the various directions an image analysis must follow. In this case, the lexical derivations for *Sarkozy* portray chiefly a negative image of the President; whereas, the RS show a relatively neutral image that requires further investigation. RS calculations can therefore be used to fruitfully detect NE and how they materialize in the text.

Table 1 – Example of forms and RS extracted sets

Form	***Freq.***	***Repeted Segment***	***Freq.***
Sarkozy	278	Nicolas Sarkozy	85
Sarko	98	de Sarkozy	29
Sarko	19	président de la République	27
Sarkosy	12	de Nicolas Sarkozy	23
Sarkozy	8	M. Sarkozy	19
Sarkozyste	3	de Sarko	16
Sarkozystes	3	Mr Sarkozy	14
Sarkosysme	2	Président de la République	10
Sarkoland	1	le président de la République	9
Sarkoland	1	Le Président de la République	2

[1] The hypergeometric distribution as described in P. Lafon (1980), "Analyse Lexicométrique et recherche des cooccurrences", Cahiers de Lexicologie n°36

[2] The *Parti Socialiste (Socialist Party)* is traditionally on the left-hand political side of the French landscape. The focus is set from November 2008, when a new head of the party is elected, to august 2009, following a deep intern crisis after the defeat at the European Elections.

The above constructed paradigms (lexical derivations, RS) can be seen as a subset creating a new LR. These linguistic phenomena allow the identification of discursive figures that can be contextually analyzed, as shown in [3][15].

Textometric tools allow the analyst to quickly build lexical paradigms. This step is an advantage in itself from a very pragmatic point of view, especially when one is in the position of having to acquire knowledge and accurate linguistic information for building LR from scratch. Transposed to the industrial context, such a process bypasses industrial impediments such as cost-cutting and production time.

The defined paradigms are then set as *Textometric objects*, on which *specificness* calculation can be applied. As seen in 2.3, this **statistical method** is aimed at extracting, for a given **subset of a corpus**, the objects that are over or under represented compared to all the other subsets of the corpus.

Table 2 - Example of extracted paraphrases

Mr Sarkozy	15
Président de la République	10
président de la République	27
Nicolas Sarkozy	85

This kind of results provides the analyst with accurate information on when and **how the *discursive figures* attached to a personality evolve in media news through time**. The shifts of the opinion can be sensed through this evolution, indicating where the attention should be focused, thus acting as a "metal detector" indicating where to dig.

For example, in the articles[3] that make up the corpus, the results (Fig.1) clearly show that the civil paraphrase *Nicolas Sarkozy* and the status paraphrase *président de la République* are highly specific of the Le Monde discourse during intense times of the political agenda (M_904_A to M_907_A), here the European Elections in June[4]. The newspaper discourse itself evolves from April to August, focusing on the person *Nicolas Sarkozy* in April (M_904_A) to emphasizing his status by the segment *president de la République* in June (M_906_A).

On the other hand, in the user's comments of these same articles, the *Président de la République* paraphrase is distinctively over represented in June (M_906_C). The capital letter in the word *Président* is highly informative as indicating a particular attachment to the normative form for writing status words. This paraphrase also evolves into *Mr Sarkozy* in August (M_908_C), while both disappear between June and August. In such cases, it is necessary to go back to the textual material to deliver a more accurate analysis, as specific knowledge of the media situation must be known to interpret the meaning of this trend. Textometric frameworks allow the user to navigate back and forth between statistical results, graphic representations and raw analysis material- the text. It is thus interesting to see the online audience of Le Monde modifying the linguistic material used to refer to *Nicolas Sarkozy*, preferring the latter civil paraphrase *Mr Sarkozy* to the status *Président de la République*. In fact, given that Le Monde is traditionally on the political left side, this shift can be explained by two factors: (i) the rise of provocative or off-topic messages,[5] supporting UMP, Sarkozy's party, resulting

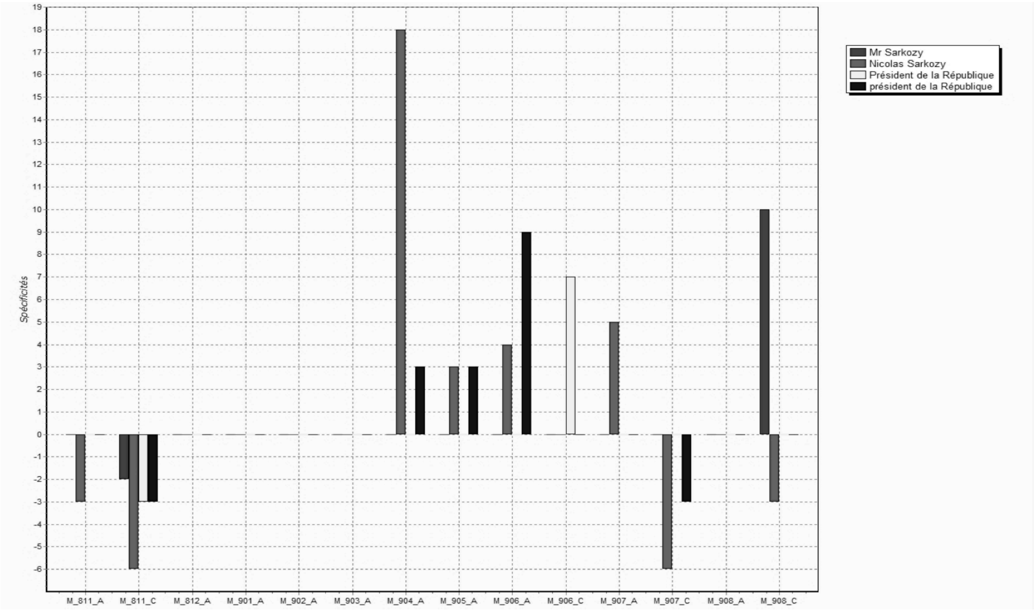

Figure 1 - Monthly variation of *specificness* on the paraphrases for the NE *"Nicolas Sarkozy"*.

[3] It must be specified that the user's comments for some months could not be retrieved due to the fact that Le Monde became a paying newspaper the year we collected the corpus, and though did not provide access to the comments associated to the collected articles.

[4] This European election was punctuated by the defeat of the Socialist Party resulting in violent media confrontations within the party. Nicolas Sarkozy is already President of France at this time.

[5] In Internet slang, this kind of attitude among users is know as a "troll", defining a user who posts provocative or off-topic messages, specifically in discussion forums.

in a *specificness* peak for *Président de la République*; (ii) the rise of unsatisfied Socialist Party supporters stemming from media confrontations in the party after their defeat. Through these confrontations, Sarkozy's position is reinforced in the political arena, resulting in a *specificness* peak for *Mr Sarkozy*. This segment is far from completely neutral for image interpretations as it seems to remove Sarkozy from his status as President without going so far as to use insulting paraphrases found in other French newspapers at the time.

3.2 Named Entities and current events for business intelligence applications

As demonstrated in 3.1, Textometry derives new information from the trends in the various forms of an NE, but how can we access information that the analyst is not specifically looking for? This is akin to the problem discussed above, standard information extraction techniques use qualitative coding to derive interpretations of the data [6]. This leads to potentially missing unknown information, in other words, semantic annotations provide only as much enriched content as the resource has been designed for.

This second use-case follows the Firthean inspiration [7]: "You shall know a word by the company it keeps", where a great deal of research has been done on lexical affinities (collocations or co-occurrences) between words. Here co-occurrences is understood as the statistical attraction of two or more words as discussed in 2.3. This calculation allows for the precise description of the lexical environment of a pivot-form through several variables left up to the analyst: (i) **co-frequency**, indicating the lowest number of times two words must appear together in the corpus to be considered as a co-occurrence; (ii) **threshold**, designating the probability level that a co-occurrence relationship must have to be considered; (iii) **segmentation context**, giving the punctuation boundary for the pivot-form, sentence, paragraph, other. What results is a list or network of co-occurring forms that can be interpreted depending on their statistical attraction (table 3).

The hypothesis here is that as a current event is discussed in the media, the lexical network produced by the co-occurrence calculation will be greater during an event than during periods of calm or low activity of the NE. This is similar to a sort of "buzz" effect where it has been shown that the more an NE is discussed by the media, the more likely it is that an event involving the NE is taking place [12]. However, the frequency of an NE alone may not be enough information to determine if an event may be taking place at a given time in the data. The high frequency of an NE could simply denote a popular topic. Two factors are thus important to discovering events, (i) the lexical network and (ii) chronological trend in the data.

The corpus used for this study is a sub-section of articles from the NYT Annotated corpus [19]. The articles correspond to Business/Financial Desk from 2001-2002 and were stripped of the xml to be put into txt format for more efficient analysis by Textometric tools. In a method similar to [12] and [4] the co-occurrence network for an NE can be calculated month by month, showing emerging information through the resulting network. In comparing both the variations in the number of different co-occurrences produced using the same input criteria, as well as the lexical units themselves, it is possible to determine what events an NE is involved in at a given time. The following example shows the monthly trend in articles mentioning the NE *Xerox* from January 2001 to December 2002, which corresponds to 160 articles in the NYT.

As can be observed in the distributions in figure 2, the form *Xerox* fluctuates greatly over the course of two years. These peaks in the number of occurrences show potential zones of interest for this NE for the periods of Febuary-March 2001 and April-July 2002. This "buzz" corresponds to the accounting scandal *Xerox* was involved in with the firm KPMG. When studying the distribution of number of different co-occurrences (each interrelated form counts as a single co-occurrence), a sharp peak can be seen for the month of April (33 co-occurrences), meaning the lexical network is much more abundant for this period and that, in following the hypothesis, an event may be taking place.

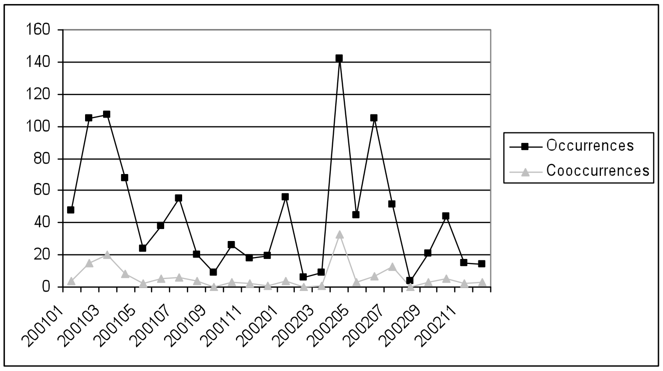

Figure 2 – Monthly variation of the number of occurrences for the NE and the number of co-occurrences for the pivot-form *Xerox*

To verify this idea, the lexical network for the month of April was compared to other months. April shows a higher number of *unexpected* vocabulary relating to the NE *Xerox* (table 3), in other words few co-occurrences actually describe the NE (*leases*, for example).

The majority of co-occurrences for this month are in keeping with the complaint filed against *Xerox* by the SEC in April 2002 (*complaint*, *kpmg*, *revenues*, *1997*, *accounting*).

Table 3 – Co-occurrences for Xerox, April 2002, co-freq 5, threshold 5

Form	Frequency	Co-Freq	Specif	Context
kpmg	100	25	23.43	19
complaint	18	9	12.68	9
pay	184	12	5.47	11
leases	23	5	5.63	5
numbers	55	7	5.66	7
that	2606	71	6.29	49
corporation	116	11	6.68	11
future	56	7	5.61	4
fine	42	10	10.15	9
its	882	49	14.44	38
restate	24	10	12.92	9
securities	134	12	6.88	12

revenue	98	10	6.51	10
cents	87	8	5.20	8
1997	69	9	6.92	9
revenues	27	5	5.28	5
accounting	329	32	16.52	30
earnings	138	12	6.74	10
share	116	13	8.46	11
agreed	39	9	9.17	9
method	20	5	5.95	4
investigation	91	12	8.74	10
commission	126	16	10.85	16
had	635	27	6.47	24
it	1406	51	8.43	37
settlement	29	7	7.62	7
auditor	25	5	5.45	5
financial	276	17	6.72	17
filed	61	9	7.38	9
settle	16	5	6.48	5
exchange	93	13	9.65	13
results	77	15	13.03	14

When analyzing the depleted lexical networks for the other months (on average between April 2001 and March 2002, only 3 co-occurrences are found using the same criteria), there is much more *expected* vocabulary: *computing, sales, services, representatives*, for example. In a manner similar to the use-case discussed in 3.1, co-occurrences are used here as the 'metal detector' for finding potential events that involve the NE. However, it remains necessary for the analyst to determine what vocabulary can be *expected* for a given NE.

4 Discussion

In this paper we compared two intelligence application use-cases applying Textometry as a method for mining information on named entities present in online news articles from Le Monde and the New York Times. Named Entities were used as an entry point for the analyses of the corpora and Textometric tools provided results for creating new Linguistic Resources as well as identifying relationships with other entities. These methods use quantitative information to formulate qualitative interpretations and thus can be included among other text mining strategies.

Both use-cases illustrate how Textometry can help media analysis tasks through two different, but complementary approaches (*specificness* and co-occurrence analysis). In a more industrial context, the analyses presented here yield promising results for business and communication intelligence applications. Three main contributions are established here:

- corpus-driven Linguistic Resource building and adaptation or update by using the Repeated Segments and lexical derivation exploratory functions of Textometric tools ;
- identification of trends with *specificness* calculation to detect over or under represented segments in a subset of the corpus in order to guide qualitative analyses of current events ;
- chronologically emerging information through the co-occurrence network of a specific NE to target zones of activity or events.

These points can help analysts in the desicion making process by shedding light on evolving trends in the corpus and potential critical information.

However, this method should be distinguished from other robust NLP approaches due to the important emphasis on the role of the user. Contrary to other mining techniques, this approach is not fully automated which raises interoperability issues with other computer processing tasks. Textometry demands the "return of the expert in the system". This explains why it is often not inculded among commercialized applications.

For future research, several venues must be explored:

- evaluating the interoperability of Textometric tools with other robust NLP applications. A combined approach using both Textometry and precoded information requires further experimentation ;
- confronting results acquired with Textometric methods against results obtained through NLP methods such as building ontologies for opinions or NE-relationship extraction ;
- analyzing the results obtained with different NLP applications such as tokenizer or syntactic taggers through Textometric methods.

In sum, deriving knowledge from corpora without predefined information models, often provided through qualitative coding, is easier said than done. This paper demonstrated how such annotations can be skirted with statistical calculations and Textometric methods, cutting production time. These methods provide adequate functions enabling interaction between the expertise of the user and the processing tools. The analyst, therefore, can achieve more in depth research.

References

[1] Bloom K., Stein S. & Argamon S., Appraisal extraction for news opinion analysis at NTCIR-6, Proceedings of NTCIR-6, 2007, p 279-289.

[2] Bollier, D. The Promise and Peril of Big Data. Washington, DC : The Aspen Institute, 2010.

[3] Delanoë, A. 2010. Statistique textuelle et series chronologiques sur un corpus de presse écrite. Le cas de la mise en application du principe de précaution. Proceedings, JADT'2010.

[4] Delaplace R., Leenhardt M. & Wu L-C., Méthode de conception d'une application de veille et d'Analyse Linguistique Assistée par Ordinateur, VSST Conference, Toulouse, France, 2010.

[5] Fayyard, U.M, Piatesky, G., Smyth, P. & Uthurusamy, R. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[6] Feldman R. & Sanger J., The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data, Cambrigde University Press, 2006, 422 p.

[7] Firth, J.R. A Synopsis of Linguistic Theory 1930-1955, Linguistic Analysis Philological Society, Oxford, 1957.

[8] Grishman, R. & Sundheim, B. Message Understanding Conference- 6 : A Brief History. Proceedings of the 16th International Conference on Computational Linguistics (COLING), I. Kopenhagen, 1996 p.466–471,.

[9] Kodratoff, Y. Knowledge discovery in texts: A definition and applications, Proceedings of the International Symposium on Methodologies for Intelligent Systems, 1999, volume LNAI 1609, p. 16–29.

[10] Lebart, L. & Salem, A. Statistique textuelle. Paris, Dunod, 1994.

[11] Lent, B., Agrawal, R., & Srikant, R. Discovering trends in text databases, Proceedings KDD'1997, AAAI Press, 14–17 p. 227–230.

[12] MacMurray E. & Shen L., Textual Statistics and Information Discovery: Using Co-occurrences to Detect Events, VSST Conference, Toulouse, France, 2010.

[13] Martin J.R. & White P.R.R., The language of evaluation: appraisal in English, Palgrave, London, 2005.

[14] Martinez, W. Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels, Thèse pour le doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3, 2003.

[15] Née, E. Insécurité et élections presidentielles dans le journal Le Monde, Lexicometrica numéro thématique « Explorations Textuelles », S. Fleury, A. Salem. 2008

[16] Poibeau T. Extraction automatique d'information. Du texte brut au web sémantique. Paris : Hermès Sciences, 2003.

[17] Poibeau, T. Sur le statut référentiel des entités nommées, Proceedings TALN'05. Dourdan, France, 2005.

[18] Salem A., Introduction à la résonance textuelle, In Actes des JADT 2004 (7 èmes Journées internationales d'Analyse Statistique des Données Textuelles), 2004, p 986-992.

[19] Sandhaus, E., The New York Times Annotated Corpus. Philadelphia: Linguistic Data Consortium, 2008.

[20] Tufféry, S., Data mining et statistique décisionnelle: l'intelligence des données. Paris : Editions Technip, 2007.

[21] Stoyanov, V., Cardie, C., Litman, D. and Wiebe, J. Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus. Working Notes of the 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004

[22] Wilson, T., Ruppenhofer, J., Wiebe, J., Documentation for MPQA Corpus version 2.0, [online] http://www.cs.pitt.edu/mpqa/databaserelease/Database.2.0.README Date consulted: May, 12th, 2011

[23] Wright, K., Using Open Source Common Sense Reasoning Tools in Text Mining Research, the International Journal of Applied Management and Technology, 2006 vol 4 n°2 p.349-387.