Analyzing Variation Patterns In Quotes Over Time

Aurelien Lauf[1],[2], Mathieu Valette[1], and Leila Khouas[2]

[1] ERTIM (INALCO) – 49bis av. de la Belle Gabrielle, 75012 Paris
[2] AMI Software – Immeuble "Le Cristal", 1475 av. A. Einstein, 34000 Montpellier

Abstract. In the past few years, there has been a growing interest in detecting quotation families and in automatically analyzing transformation patterns. However, no work has provided a complete qualitative analysis of these variations. Through a comprehensive linguistic analysis, the goal of this paper is to study and categorize the way quotes from newspapers tend to evolve and deform over time. In order to help in observing patterns and variability, we apply global sequence alignment techniques, commonly used in bioinformatics. Recurrent patterns, such as the common deletion of words expressing modality, paraphrases, or strong synonymic variations, are listed and discussed. In addition to providing a better understanding of cultural dynamics in media, we believe that the categorization of variation regularities in quotes can help further enhance the quality of similar quotations clustering algorithms and other NLP tasks such as paraphrase identification.

Keywords: Discourse analysis, quotes, memes, multiple sequence alignment, linguistics, co-reference, modality

1 Introduction

Quoting is a very common practice, especially in journalism. A given quote can have many different versions: depending on the context, one can decide to focus on the beginning or the end of a quote, or on the contrary to write the whole segment. Quotes may also shrink or grow over time, for various reasons. Furthermore, one would usually assume that quotes are faithful to the original but this may not always be true: they are sometimes quoted from memory or willingly modified in order to convey a stronger meaning.

In the past few years, following [1] framework for "meme-tracking", there has been a growing interest in detecting quotation families and in automatically analyzing transformation patterns [2][3]. However, to our knowledge, no work has provided a complete qualitative analysis of these variations. The goal of this paper is, through a comprehensive linguistic analysis, to study and categorize the way quotes tend to evolve and deform over time. In order to help in observing patterns and variability, we apply global sequence alignment techniques, commonly used in bioinformatics. Recurrent patterns, such as the common deletion of words expressing modality, paraphrases, or strong synonymic variations, are

listed and discussed. In addition to providing a better understanding of cultural dynamics in media, we think that these variation regularities in quotes should be known in the NLP community; we believe that it can help further enhance the quality of similar quotations clustering algorithms and, to a lesser extent, paraphrase identification.

First, we will focus on previous works on this subject. We will then describe the textual data we used, and explain our approach. We will finally present and discuss the results of our analysis, focusing on the most interesting patterns.

2 Previous Works

In [1], authors present a clustering approach capable of identifying quotation families, i.e. all textual variants of each quotation, leading to the first large scale quantitative analysis of memes. In [2], hyperlinks between sources are added in order to adress the fidelity of information according to the type of source. Interestingly, authors show that most changes are introduced by media and that blogs are less likely to do so because they tend to simply copy and paste quotes. A new algorithm for quotation clustering is introduced in [3]. This method is based on [1] but relies on a linguistic approach. None of the above papers provide a comprehensive linguistic analysis to study and categorize the way quotes actually evolve and deform over time.

3 Description Of The Dataset

Our textual data is about the case relating to allegations of sexual assault against the former IMF managing director, during the year 2011. The documents have been collected using a metasearch engine with the following query: *dsk OR strauss-kahn OR strauss-khan*. The corpus comprised 27 439 news articles written in English.

We extract quotes from all documents, i.e. strings between quotation marks. We store for each of them the number of times they occur, the days they appear in (with the number of occurrences for each day) and the corresponding documents IDs in order to return to the text if needed: checking the context may be useful, e.g. to check whether two short quotes are really linked, or to understand why the journalist chose this particular segment of the full quote instead of another one. Using the approach described in [1], we produce quotes clusters which are groups of similar quotes allowing close textual variations. Using [4], quotes were lemmatized and stop words were filtered to help the clustering task[3]. In the remainder of this paper, we will talk about *quotes* instead of *quotes clusters*. Each *quote* has one or more *version(s)*. Table 1 shows an example of a *quote* which has 9 *versions*.

[3] Lemmatization and stop word filtering only occur during the quote clustering task. To avoid neutralizing differences, original quotes are used when analyzing variations.

Table 1. Illustration of a quote with all its 9 different versions, sorted by number of occurrences. *Analyzing Variation Patterns In Quotes Over Time*

Version of the quote	Occurrences	First and last dates
"Offered a compelling and unwavering story"	66	05-20 to 06-06
"Compelling and unwavering story about what occured in the defendant room"	46	05-22 to 08-26
"Compelling and unwavering"	28	07-01 to 09-28
"The complainant in this case has offered a compelling and unwavering story about what occured in the defendant room"	3	05-24 to 07-03
"The victim has given a compelling and unwavering story about what happened in the defendant room"	2	05-20 only
"Offered compelling and unwavering story about what occurred in the defendant room"	2	05-24 only
"A compelling and unwavering story"	1	08-03 only
"Compelling and unwavering story"	1	08-03 only
"She offered a compelling and unwavering story"	1	08-23 only

Out of the 27 439 documents, our system has detected 22 099 quotes, many of them having only one version. We define the weight of a quote as the sum of the number of occurrences of each of its versions. For the present study, we decided to focus on the 100 best quotes according to this computed weight. In total, these 100 quotes represent 1039 different versions and 13 958 occurrences.

16% of quotes are translations, mostly from French. It is an interesting case because even though what was actually said was in a different language, it is reproduced between quotation marks. Quotation, which is strongly linked to intertextuality, is common and convenient in journalism to distantiate from what is being said (objectivity).

4 Description Of Our Approach

In order to observe patterns and variability, we perform global sequence alignment techniques commonly used in bioinformatics to identify similarities between sequences of DNA or protein [5]. These alignments are most of the time represented as rows within a matrix; similar elements are aligned on the same column and gaps may be introduced when an element has no match in the other sequences. Gap-to-gap matches are not allowed (a column cannot have only gaps). We are interested here in a word-to-word correspondence between all versions of a given quote instead of a residue-to-residue correspondence but the way to achieve it is similar [6]: the order of the words has to be preserved; when a word has no match in the other versions, a gap is introduced. Similar words are on the

same column. A simple example of perfectly aligned versions is shown in table 2.

Table 2. Illustration of the global alignment of the 9 versions of the quote shown in table 1. The last words are troncated in this example to stay on the same line. Dashes represent gaps. We can observe 2 replacements: *complainant*/*victim* and *offered*/*given*.

-	-	-	-	-	-	offered	a compelling and (...)
-	-	-	-	-	-	-	- compelling and (...)
-	-	-	-	-	-	-	- compelling and (...)
the	complainant	in	this	case	has	offered	a compelling and (...)
the	victim	-	-	-	has	given	a compelling and (...)
-	-	-	-	-	-	offered	- compelling and (...)
-	-	-	-	-	-	-	a compelling and (...)
-	-	-	-	-	-	-	- compelling and (...)
she	-	-	-	-	-	offered	a compelling and (...)

Using [7] dynamic programming algorithm, it is possible to easily get the best global alignment of two sequences. It is theoretically possible to generalize this method for more than two sequences using a hyper cube instead of a simple matrix. However, this generalization is unfeasible because it is known to be exponential in complexity. The only way to perform multiple sequences alignment is then to use heuristic methods which make locally optimal choices at each step, but are not guaranteed to find the optimal alignment [8][9]. The most used approach is the progressive alignment technique and the most famous implementation is ClustalW [10], which we use due to its efficiency.

The output alignment is noisy, especially when versions share only a few similarities. We thus had to manually correct all the mistakes. Nevertheless, it was a far less time consuming task than aligning all versions by hand. Results of our analysis are discussed in the following section.

5 Results And Discussion

Before discussing the observed variations and deletions between versions, we show here the results of some automatic quantitative analysis.

5.1 Diachronic Analysis

We store for each version of a quote the number of times it occurs, the days it appears in, and the number of its occurrences for each day. Using these data, we automatically calculate dispersion diagrams in order to check whether there is a correlation between the number of words of a version and the number of times it is reused. Figure 1 displays the relation between the number of words of a version and how many times it occurs. It also focuses on its lifespan, i.e. the

number of days it appears in the Press. Shorter versions seem to occur far more and have a better lifespan than long ones.

Furthermore, it is worth noticing that most of the versions appear right from the first day, i.e. there does not seem to be one original quote gradually modified through time.

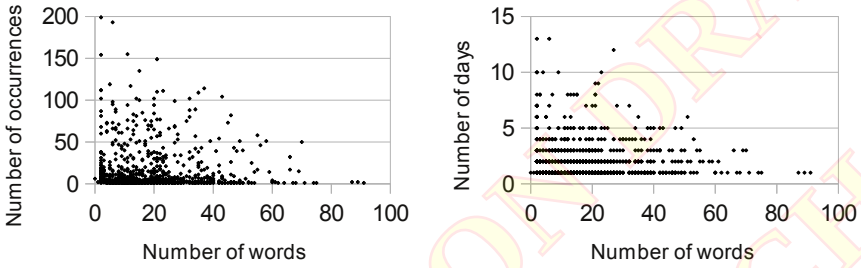

Fig. 1. Correlation between the number of words of a version and (left) the number of times it occurs, (right) the number of days it stays. The shorter a version is, the more and the longer it seems to be used.

5.2 Size And Variations

Next, using the sequence alignmnent technique described earlier, it is possible to count the number of deletions (gaps) and variations (replacements) within the different versions. Figure 2 displays the relation between the size of a quote (the number of words of its longest version) and the way it is altered. There is a tendency for longer quotes to be modified more than shorter ones, which is surprising because one would think that longer quotes are copy-pasted and not quoted from memory like shorter ones [2].

5.3 Convergence

We observed a tendency for quotes to converge to 2-3 words long phrases. About 60% of quotes staying more than one month are concerned (only 3% for shorter quotes). These phrases are what remains of the whole quote on the last day and are meaningful enough to recontextualize. Examples are shown in table 3.

5.4 Analyzing Variations

In order to better understand the way quotes are actually altered over time, we performed a comprehensive manual linguistic analysis of the variations and the deletions (see section 5.5 below) between versions.

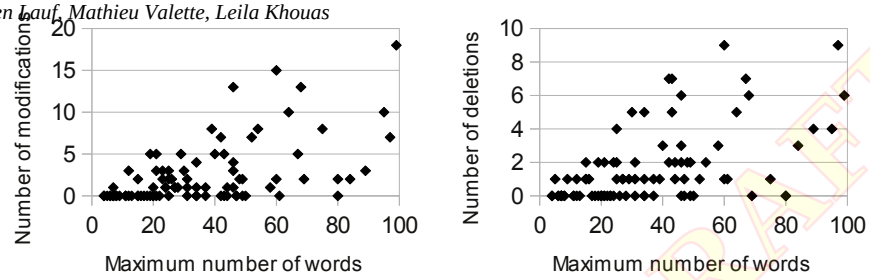

Fig. 2. Correlation between the number of words of the longest version of a quote (i.e. maximum number of words), and the way it is altered. Longer quotes tend to be less stable than short ones.

Table 3. Examples of what remains of the whole quote on the last day (3-4 months after the appearance of the quote).

He said he was leaving his IMF post with "infinite sadness" *so that he could devote full time to proving his innocence.*
That night he is in the custody of the New York Police Department facing the humiliating "perp walk".
How then, did she go from "compelling and unwavering" *to having her case dismissed due to lack of credibility?*

On our data set, 215 variations were observed. Table 4 shows the different types. We will focus on the three most important variations, i.e. synonymic variations, co-reference, and reformulation.

Synonymic variations are the most frequent type of modifications, mostly verbs and nouns. We noticed that translated quotes, which we discussed earlier, follow the same patterns than English ones, except for synonyms: about half (44%) of synonymic variations come from translated quotes. Furthermore, words from translations tend to have more variants, up to six:

- "It was a moral failing/failure/weakness/error/mistake/fault".
- "He tried to open/undo/remove my jeans".
- "It's important for a politician/man in politics/political man to be able to seduce".

Co-reference is a well known linguistic phenomenon: multiple elements (persons, actions or things) may have the same referent. Most of the time, later mentions of a previously introduced element are simpler, often reduced to pronouns. For example, consider the following version: "I'm rather proud of my husband reputation as a seducer". On some other versions, the "reputation as

Table 4. List of all the observed variation types, sorted by rate of occurrence. Most of the phenomenons can be subdivided. An example is given for each subtype.

Variation type	Subtype	Example
Synonyms (28.37%)	Verbs (40.98%)	*happen/occur*
	Nouns (37.70%)	*relationship/liaison*
	Adverbs (18.03%)	*gravely/seriously*
	Adjectives (3.28%)	*unjust/unfair*
Co-reference (16.74%)	Person: pronoun (41.67%)	*dsk/he*
	Person: reformulation (25%)	*woman/victim*
	Person: abbreviation (16.67)%	*district attorney/DA*
	Action (8.33%)	*what happened/it*
	Thing (8.33%)	*this incident/it*
Reformulation (12.09%)	Paraphrase (57.69%)	*has no idea/doesn't know*
	Syntactic variation (42.31%)	*a man with/this man has*
Spelling (11.16%)	(Common) misspelling (37.5%)	*whatsoever/what so ever*
	UK vs. US spelling (37.5%)	*honour/honor*
	Typo (25%)	*candidate/cadidate*
Determiners (8.84%)	Def. art./dem. (57.89%)	*the/this*
	Def. art./poss. (15.79%)	*the/my*
	Indef. art./dem. (10.53%)	*a/this*
	Indef. art./quantifier (5.26%)	*a/one*
	Def. art./indef. art. (5.26%)	*the/a*
	Dem./dem. (5.26%)	*this/that*
Conjugation (8.37%)	Tense (83.33%)	*have been/were*
	Person (11.11%)	*are/is*
	Mood (5.55%)	*put/puts*
Linking words (6.51%)	Prep./prep. (71.42%)	*on/in*
	Subord./subord. (21.43%)	*although/even if*
	Coord./coord. (7.14%)	*or/and*
Contractions (4.19%)	/	*was not/wasn't*
Number (1.86%)	/	*skill/skills*
Inversions (1.86%)	/	*still is/is still*

a seducer" is introduced out of the quote: *Anne Sinclair seemed forgiving of his reputed behavior. "No, I'm rather proud of it!" she told.* Sometimes, co-reference occurs for stylistic reasons to avoid repetitions of a word, e.g. "man" instead of "defendant". It may introduce some nuances though, willingly or not, e.g. "victim" does not carry the same implication than "complainant".

Reformulation can be divided into two subtypes : syntactic variations and paraphrases. The former includes many phenomenons, e.g. transition from direct to indirect speech or constituent order modification:

– "I just want to know if I need a lawyer" vs. "do I need a lawyer".

– "We expect him to be released tomorrow" vs. "we expect he will be released tomorrow".
– "Of which I am not proud" vs. "and I'm not proud of it".

Paraphrastic reformulation goes beyond simple syntactic modifications; whole phrases may be altered without modifying its global meaning:

– "I have no doubt" vs. "I am certain".
– "She has no idea what" vs. "she doesn't know what".
– "There were many reasons to believe" vs. "we continue to believe".
– "It was not just" vs. "it was more than".

5.5 Analyzing Deletions

129 deletions were observed. Table 5 displays the different types. We decided to ignore deletions at the beginning or the end of a version because they can be easily explained by indirect speech. We will focus on the most recurrent phenomenon: deletion of words expressing modality.

Modality [11] [12] is what allows speakers to express subjectivity. There is no consensus among researchers regarding categorisation of modality but most works agree on two main types: necessity and possibility. *Deontic modality* refers to permission and obligation (or moral desirability). *Alethic modality* is about (im)possibility and logical necessity. *Epistemic modality* indicates the speaker's judgment. Alethic and epistemic modality are often mixed because it might not be relevant to oppose what is logically true and what the speaker believes to be true [13]. We observe a strong tendency for words expressing epistemic modality to be omitted:

– "Forensic evidence (we believe) will not be consistent with a forcible account".
– "He is (obviously) not in a position to run the IMF".
– "(I think) it was a moral failing".

6 Conclusion

In this paper, we provided a comprehensive linguistic analysis and categorization of the variations in newspaper quotes over time. Quotes were semi-automatically aligned using a multiple sequence alignment technique, in order to help in detecting similarities, variations (replacements) and deletions (gaps).

We have detected some recurrent patterns, such as the common deletion of words expressing modality, paraphrases, or strong synonymic variations mostly among nouns and verbs. Furthermore, we highlighted an interesting tendency for quotes to converge to 2-3 words long phrases, powerful enough to summarize the whole context. We believe that a complete categorization of variation regularities

Table 5. List of all the observed deletion types, sorted by rate of occurrence. Deleted words are shown between parentheses. *Analyzing Variation Patterns In Quotes Over Time*

Deletion type	Subtype	Example
Modality (20.15%)	Epistemic (80.77%)	*(I think) it is*
	Alethic (11.54%)	*(may) have*
	Deontic (3.85%)	*(have to) face*
	Affective (3.85%)	*I'm (sorry I'm) not*
Modifiers (18.60%)	Adjectives (50%)	*(physical) evidence*
	Adj. phrases (41.67%)	*influence (througout the world)*
	Noun adjuncts (8.33%)	*(selection) process*
Linking words (15.50%)	Coordinators (90%)	*my children (and) my friends*
	Conj. adverbs (10%)	*(indeed), we were intent on*
Determiners (10.85%)	Def. articles (50%)	*to (the) prosecutors*
	Possessives (28.57%)	*our guest and (our) staff*
	Indef. articles (21.43%)	*with (a) complete conviction*
Completives (10.07%)	Verbs (53.85%)	*I felt (that) I*
	Adjectives (30.77%)	*important (that) the*
	Nouns (15.38%)	*the idea (that) she*
Enumerations (7.75%)	/	*my strength (and all my energy)*
Time expressions (6.98%)	/	*I feel compelled (today) to*
Repetitions (5.43%)	Same referents (71.43%)	*this man (Mr. Strauss Kahn)*
	Same words (28.57%)	*a very (very) defensible case*
Intensity (4.65%)	/	*changed (a single) thing*

in quotes can help further enhance the quality of similar quotations clustering algorithms and other NLP tasks such as paraphrase identification.

Further studies have to be conducted on other sets of data though in order to validate our observations. Furthermore, we are aware that our results may be strongly influenced by the genre of our corpus made from news articles only; results should thus also be compared with different kind of news sources, i.e. blogs articles or tweets.

References

1. Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD'09 - Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 497–506, Paris, 2009.
2. Matthew Simmons, Lada Adamic, and Eytan Adar. Memes online: extracted, subtracted, injected, and recollected. In *ICWSM 2011 - Proceedings of the 5th international AAAI conference on weblogs and social media*, Barcelona, 2011.
3. Elisa Omodei, Thierry Poibeau, and Jean-Philippe Cointet. Multi-level modeling of quotation families morphogenesis. In *Proceedings of the 2012 ASE/IEEE international conference on social computing (SocialCom 2012)*, Amsterdam, 2012.

4. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, Manchester, 1994.
5. David Mount. *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor Laboratory Press, 2004.
6. Robert Irving. Plagiarism and collusion detection using the Smith-Waterman algorithm. Technical report, 2004.
7. Saul Needleman and Christian Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.
8. Sing-Hoi Sze, Yue Lu, and Qingwu Yang. A polynomial time solvable formulation of multiple sequence alignment. *Journal of computational biology*, 13(2):309–319, 2006.
9. Cédric Notredame, Desmond Higgins, and Jaap Heringa. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, 2000.
10. Julie Thompson, Desmond Higgins, and Toby Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994.
11. Paul Portner. *Modality*. Oxford University Press, 2009.
12. William Frawley, Erin Eschenroeder, Sarah Mills, and Thao Nguyen. *The expression of modality*. Mouton de Gruyter, 2006.
13. Frank Palmer. *Mood and modality*. Cambridge University Press, 1986.