Evolution of Event Designation in Media: Preliminary Study

Xavier Tannier, Véronique Moriceau, Béatrice Arnulphy and Ruixin He

LIMSI-CNRS
Univ. Paris-Sud, 91403 Orsay, France
firstname.surname@limsi.fr

Abstract

Within the general purpose of information extraction, detection of event descriptions is often an important clue. An important characteristic of event designation in texts, and especially in media, is that it changes over time. Understanding how these designations evolve is important in information retrieval and information extraction. Our first hypothesis is that, when an event first occurs, media relate it in a very descriptive way (using verbal designations) whereas after some time, they use shorter nominal designations instead. Our second hypothesis is that the number of different nominal designations for an event tends to stabilize itself over time. In this article, we present our methodology concerning the study of the evolution of event designations in French documents from the news agency AFP. For this preliminary study, we focused on 7 topics which have been relatively important in France. Verbal and nominal designations of events have been manually annotated in manually selected topic-related passages. This French corpus contains a total of 2064 annotations. We then provide preliminary interesting statistical results and observations concerning these evolutions.

Keywords: events, evolution, media

1. Introduction

Information extraction consists in a surface analysis of text dedicated to a specific application. Within this general purpose, detection of event descriptions is often an important clue. However, events are, in open-domain information extraction, less studied than general named entities like location and person names.

An important characteristic of event designation in texts, and especially in media, is that it changes over time. For example, *"September-11"* is now a common denomination for 2001 attacks, but these attacks were described in a different manner on the first days after the attacks (Fragnon, 2007). Also, the event now known as *"Arab spring"* is the aggregation of many events occurred in Tunisia, Egypt, Libya, Bahrein, etc. that were not talked about as part of this Arab spring when they happened.

Understanding how these designations evolve is important in information retrieval and information extraction. As a single example, imagine a query about "Arab spring"; this query will lead to background documents, analysing ins and outs of this period, but rarely to articles relating actual events.

To the best of our knowledge, no study focused on the evolution of the designations of events. Our first hypothesis is that, when an event first occurs, media relate it in a very descriptive way (using verbal designations) whereas after some time, they use shorter nominal designations instead. Our second hypothesis is that the number of different nominal designations for an event tends to stabilize itself over time.

In this article, we present our methodology concerning the study of the evolution of event designations in French documents from the news agency AFP (Section 2.). We then provide preliminary interesting statistical results and observations concerning these evolutions (Section 3.).

2. Corpus

Our initial corpus is composed of newswire texts in French produced by Agence France Presse (AFP). This corpus covers all areas of the news. It is indexed by the search engine Lucene[1] (Hatcher and Gospodnetić, 2004), with stemming. In the following sections, all examples are translated into English.
For this preliminary study, we focused on 7 topics which have been relatively important in France:

- "la crise grecque" – the Greek economic crisis
- "la révolution arabe: Tunisie et Égypte" – the Tunisian and Egyptian revolutions
- "Wikileaks" – the Wikileaks affair
- "l'affaire Laetitia" – the Laetitia's case (the murder of a young woman, which evolved into a strong judicial polemic in France)
- "la grippe H1N1" – the H1N1 influenza
- "le nuage islandais" – the Icelandic ash cloud

For each topic, we have submitted several queries to Lucene in order to collect the first 100 documents. These queries are composed of topic-related keywords, for example[2]:

- Queries for *"Greek crisis"*: *debt greek*, *deficit greek*, *crisis greek*, *crisis debt greek*,
- Queries for *"Tunisian revolution"*: *arab spring*, *Tunisia disorder*, *jasmine*,
- Queries for *"Icelandic ash cloud"*: *ash volcano iceland*, *iceland eruption*, *icelandic cloud*, *Eyjafjallajokull*.

[1] http://lucene.apache.org/java/docs/

[2] As a stemming algorithm is applied, keyword "greek" will return document containing "Greece", "Tunisia" also leads to "Tunisian", etc.

Topic	Nb of annotated passages	Nb of annotations	
		Verbal	Nominal
Greek crisis	223	131	142
Egyptian revolution	106	23	106
Tunisian revolution	246	28	294
H1N1 influenza	342	296	149
Laetitia's case	409	42	489
Wikileaks affair	152	95	103
Icelandic ash cloud	134	96	70
TOTAL	1412	711	1353
Total number of designations		2064	

Table 1: French corpus description

Among the collected documents, we have manually selected topic-related passages. Each passage is associated with the publication date of the document from which it is extracted.
Verbal and nominal designations of events have been manually annotated with the Callisto annotation tool (Day et al., 2004) by two of the authors. For example:

- Topic "*Greek crisis*":
  - *European leaders are preparing to meet this Thursday for an emergency summit to redefine the rules in the euro zone after the* <EVT_NOUN>*Greek crisis*</EVT_NOUN>.
  - <EVT_VERB>*Drowning in debt, Greece needed help from its partners of the euro area*</EVT_VERB>.
- Topic "*Tunisian revolution*":
  - <EVT_VERB>*Ben Ali left Tunis after a month-long popular protest which Tunisians called the* <EVT_NOUN>*Jasmine Revolution*</EVT_NOUN> </EVT_VERB>

This French corpus contains a total of 2064 annotations. Table 1 describes some of its characteristics. These annotations can be distributed to interested people.

3. Evolution of Event Designations

Measurements on this annotated corpus leads us to several kinds of observations concerning the evolution of event designations in media. The two most important observations concern the choice between a verbal or a nominal designation and the number of different designations.

3.1. From Verbal to Nominal Designations

It sounds quite intuitive that when an event occur, the first designations in relation to what happened take the form of a sentence headed by a verb. The event must be described and contextualized; all the information of location, time, aspect must be given.
When the event is of a relative importance, it is no longer necessary to use the verbal forms. We (media and locutors) use a name for the event: the verbal designations are nominalized. This process can happen in one or few days. Then the event is no more expressed as the main information, but rather as a contextual information, making more difficult the automatic extraction (Arnulphy et al., 2012b; Arnulphy et al., 2012a). It is already incorporated into the context of other related events: responses to this first event, declarations, consequences, etc.
This is a general behavior, that is exemplified by observations in Figure 1. The Figure shows the evolution in time of the number of verbal (continuous, thick red curve) and nominal (dotted, thin blue curve) designations of the Tunisian revolution main events.
The tag ① in the Figure corresponds to president Ben Ali overthrow (Jan. 14th, 2011). This date is preceded by a progression of designations illustrating the worsening of the situation. We can see that many verbal designations on January 14 were relayed by mostly nominal designations the following days.
The second tag ② corresponds to the revolution in Egypt, where a lot of articles refered to the origins of the "Arab Spring".

However, two of our thematic sub-corpora show a different behavior:

- Wikileaks cases (Figure 2), where verbal and nominal designations are simultaneous. This can be explained by the fact that these events were announced and expected before they occurred (Wikileaks announced the leaks a few days in advance). This could explain why designations were already nominalized when the facts happened.
- The financial crisis in Greece (Figure 3). We see several peaks of verbal designations. For example, the case came back at the front page (tag ②) a long time after the first news about financial difficulties in Greece (tag ①). The context must then be reminded.

3.2. The Avarice Principle

The avarice principle is another intuitive notion concerning the evolution of event designation. When an event is repeatedly evoked in media, its identification by readers gets easier, and journalists can take the liberty of using shorter denominations. Finally, following a denomination consensus, all locutors tend to use the same short designations.

3.2.1. Shorter Designations

Nominal description of an event becomes shorter and shorter in time. Journalists use less syntactic relations, and especially spatiotemporal context of the event is often skipped. The reader is actually expected to know already about the event.
A striking example is the case of the murder of a young woman, which evolved into a strong judicial polemic in France. First designated as the "disappearance of a young woman between Tuesday and Wednesday in Pornic (Loire-Atlantique)", it soon became "the disappearance of Laetitia in Pornic", then "the disappearance of Laetitia" and "the Laetitia's case", mentioned at the first time 11 days after the fact.

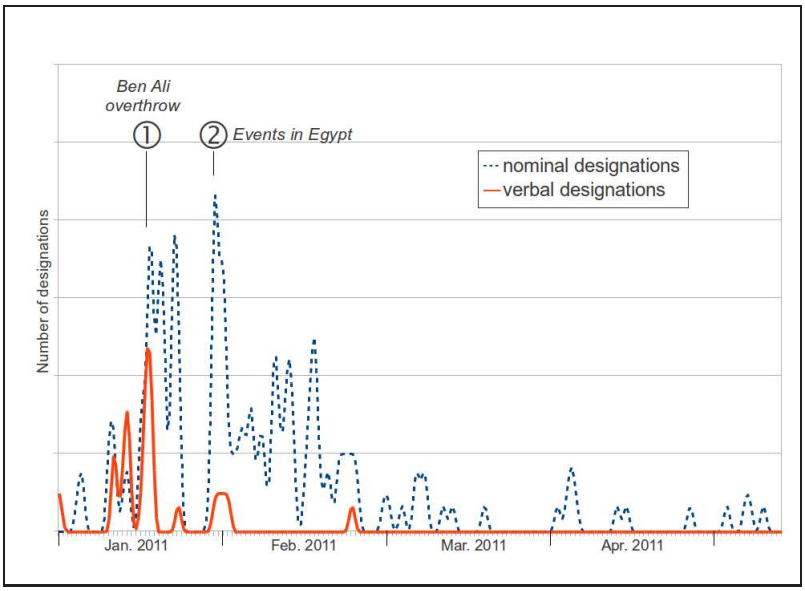

Figure 1: Evolution of event designations, the example of Tunisian events in 2011.

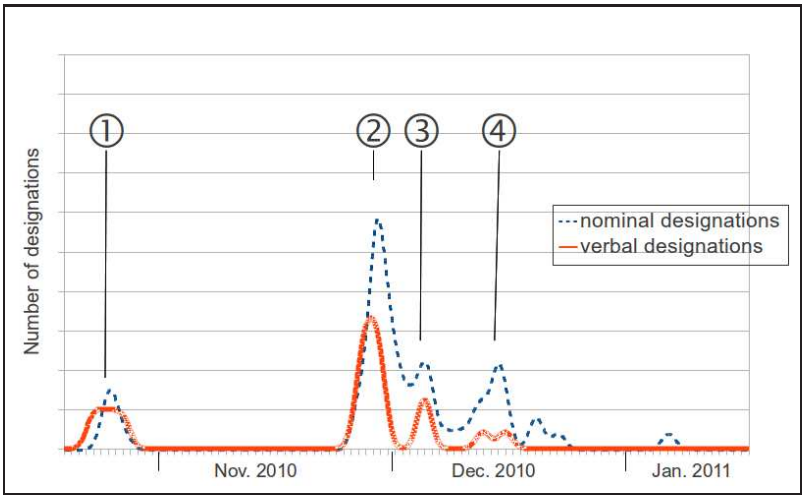

Figure 2: Evolution of event designations, the example of Wikileaks cases. Numbered tags correspond to a divulgation of leaks.

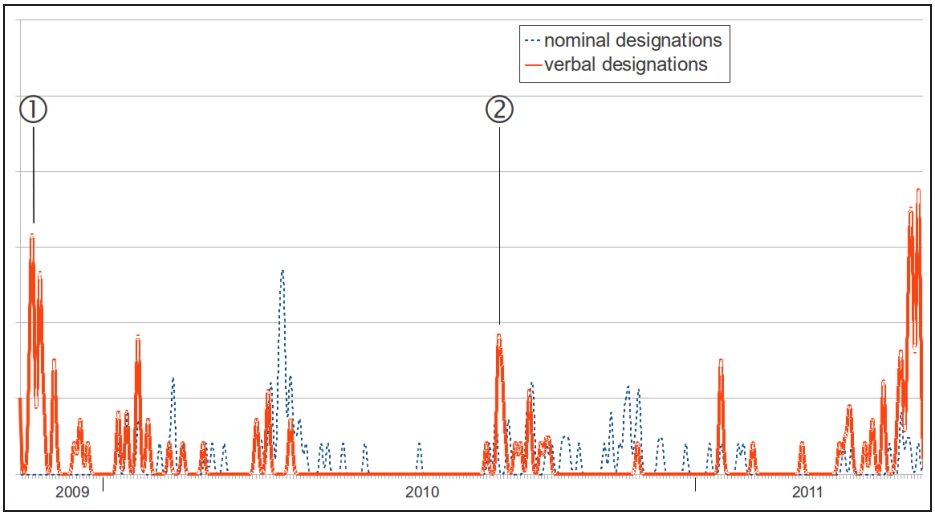

Figure 3: Evolution of event designations, the example of financial crisis in Greece.

But we also showed that contextual elements, as well as verbal designations, come back after a certain period of time if the event is still refered to after several months.

3.2.2. Less Variations

The same process of moving to shorter denominations leads to a (at least partial) freezing of these denominations. This is what happens in "Laetitia's case", but also in almost all events from our study: "Greek crisis", "Jasmine revolution", "Arab Spring", etc. The number of different denominations always decreases in time.

Figure 4 shows an example for "Laetitia's case" example, where we can see that the ratio between the number of different designations of an event and its total number of designations is low at the beginning (lot of different designations), while this rate gets higher at the end (few different designations).

However, the designations of event are not inevitably fixed. According to (Calabrese Steimberg, 2011) (referring to (Bourdon, 2009)), the names of events can change during denominative conflicts, *i.e.* an indecision period when the name does not (or no longer) suit correctly, and when a new denomination is in search. Calabrese takes the example of the "H1N1 flu" which on May 2009 was incorrectly named "grippe porcine" (*swine flu*), and then renamed "grippe A" (*A flu*), "grippe porcine ou nord américaine" (*North-American or swine flu*), and finally *A (H1N1)*.

3.3. The Domino Effects

Other secondary observations concern the evolution of event designations when the consequences of these events turn to be more important than expected:

- The denomination of series of events, when some facts are the consequences of an initial event. In this case, even the name of an object can be used as the designation of an event. For example, "Icelandic ash cloud" is often used for designating the consequences of Eyjafjallajokull's eruption in March 2010 (closure of European airports, etc.). This should constitute a major issue for automatic event extraction systems.

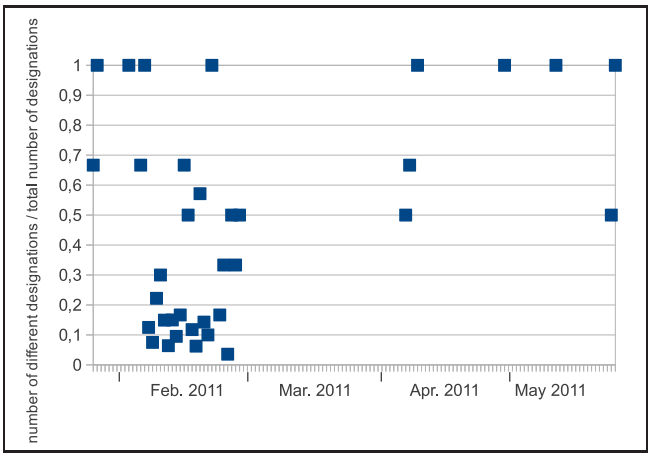

Figure 4: Evolution of the ratio between the number of different designations of an event and its total number of designations.

- Choosing denomination is sometimes a gradual process, when an event takes importance. Events in Tunisia at the beginning of 2011 have then been designated as *agitation*, *demonstrations*, *riots*, *revolt*, etc. and finally *revolution*. Also, events in several countries (Morocco, Tunisia, Egypt, Yemen, etc.), with their own designations, have been named "Arab Spring", a new designation for a new phenomenon including the others.

4. Conclusion

We built an annotated corpus with verbal and nominal designations of seven different events in French newswire articles, in order to study the evolution of these designations in time. This corpus is made freely available to interested researchers.

We also presented some statistical experiments aiming at understanding how designations of events evolve in time. The methodology and the results do not only constitute an interesting linguistic study. It is also a first step to:

- Easier and more precise studies of particular events, in linguistics or in specific applications of information extraction.
- More relevant clues for discovering automatically variations concerning a same event, for news aggregation and news retrieval.

5. Acknowledgements

This work has been partially funded by OSEO under the Quaero program, as well as French National Research Agency (ANR) under project Chronolines (ANR-10-CORD-010). We would like to thank the French News Agency (AFP) for providing us with the corpus.

6. References

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. 2012a. Automatically Generated Noun Lexicons for Event Extraction.

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. 2012b. Event Nominals: Annotation Guidelines and a Manually Annotated Corpus in French. In *Proceedings of the Eighth International Language Resources and Evaluation (LREC'2012)*, Istanbul, Turkey, May.

J. Bourdon. 2009. *Le récit impossible : le conflit israélo-palestinien et les médias*. de Boeck.

L. Calabrese Steimberg. 2011. La nomination d'événements dans le discours d'information : entre activité collective et déférence épistémiologique. In *Colloque Langage, discours, événements*, Firenze, Italy.

D. Day, R. Kozierok, C. McHenry, and L. D. Riek. 2004. Callisto: A Configurable Annotation Workbench. In European Language Resources Association (ELRA), editor, *Proceedings of the Fourth International Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

J. Fragnon. 2007. Quand le 11-Septembre s'approprie le onze septembre. Entre dérive métonymique et antonomase. *Mots. Les langages du politique.*, 85:82–95.

Erik Hatcher and Otis Gospodnetić. 2004. *Lucene in Action*. Manning.