Les entités nommées événement et les verbes de cause-conséquence

Béatrice Arnulphy[1,2] Xavier Tannier[1,2] Anne Vilnat[1,2]
(1) Univ. Paris-Sud, Orsay, France
(2) LIMSI-CNRS, B.P. 133, 91403 Orsay Cedex, France
{Beatrice.Arnulphy, Xavier.Tannier, Anne.Vilnat}@limsi.fr

Résumé. L'extraction des événements désignés par des noms est peu étudiée dans des corpus généralistes. Si des lexiques de noms déclencheurs d'événements existent, les problèmes de polysémie sont nombreux et beaucoup d'événements ne sont pas introduits par des déclencheurs. Nous nous intéressons dans cet article à une hypothèse selon laquelle les verbes induisant la cause ou la conséquence sont de bons indices quant à la présence d'événements nominaux dans leur cotexte.

Abstract. Few researches focus on nominal event extraction in open-domain corpora. Lists of cue words for events exist, but raise many problems of polysemy. In this article, we focus on the following hypothesis : verbs introducing cause or consequence links have good chances to have an event noun around them.

Mots-clés : Entité nommée, événement, rapports de cause et conséquence.

Keywords: Named entity, event, cause and consequence links.

1 Introduction

L'analyse des entités nommées (EN) se focalise généralement sur les notions classiques de lieu, organisation, personne ou date. Les événements sont rarement considérés, alors même qu'ils ont une grande importance pour les applications habituelles comme la recherche d'information, l'extraction d'information ou la veille technologique. Les événements désignés par des verbes sont traités dans de nombreux travaux comme (Vendler, 1967) ou dans le cadre de TimeML (Pustejovsky *et al.*, 2003). En complément de ces travaux, nous nous intéressons au résultat de la nomination d'un événement, aux noms donnés aux événements, que nous appellerons "entités nommées événement" (EN-E). Les événements nominaux peuvent être de plusieurs types : des noms déverbaux, dérivés de verbes qui font événement tels que *fête* (issu de *fêter*) ; des éléments qui évoquent des événements de façon non ambiguë comme *festival* dans *Festival du film de Berlin* ; ou encore des mots qui prennent un caractère événementiel en contexte, par exemple *salon* dans *La cinquième édition du Salon de l'éducation* ou un nom de lieu comme *Tchernobyl* ou *Copenhague*, désignant, par métonymie, l'incident qui s'y est produit ou la conférence qui s'y est tenue (*Personne ne veut d'un nouveau Tchernobyl* ; *Copenhague se solde par un échec*).

Une étape du travail d'extraction des EN-E est donc de déterminer les expressions désignant potentiellement un événement, puis de déterminer en contexte si c'est bien le cas. Nous formulons ici l'hypothèse que certains verbes introduisant la notion de cause ou de conséquence sont des déclencheurs de noms

d'événements dans leur cotexte.

Nous présentons un survol des définitions données à l'événement et un aperçu de quelques travaux entrepris pour le traitement des EN-E (Section 2), puis nous développons le problème lié à l'extraction des EN-E, ainsi que notre hypothèse de travail (Section 3). Pour finir, nous présentons notre étude lexicale et l'expérimentation mise en œuvre pour valider cette hypothèse (Section 4)[1].

2 État de l'art

Quelques définitions de l'événement ont été avancées en philosophie, histoire ou linguistique. Notons celles développées en journalisme et en linguistique.

Une réflexion importante a été développée depuis les années 70 sur la notion d'événement médiatique[2]. Ces travaux se sont intéressés à "ce qui fait événement" et comment les médias le créent. Neveu & Quéré (1996) présente la notion d'événement, comme une occurrence singulière, imprévue, non répétable, produite "dans un passé plus ou moins proche". Son actualité ou sa réalité passée est tenue pour absolue, singulière, non répétable et contingente. En linguistique, quelques travaux se sont attachés à aborder des problèmes relatifs aux événements et aux EN-E. Velde (2000) introduit la notion de "nom propre de temps", en faisant le parallèle entre les noms propres et "la triade je-ici-maintenant". Il existe bien des noms propres de personnes et noms de lieux, et ceux de temps doivent donc exister également. De plus, des noms de lieux et des dates peuvent, par métonymie, se charger du sens de l'événement qui a eu lieu en cet endroit ou à cette date (Steimberg, 2006). C'est le cas par exemple du toponyme *Tchernobyl* (Lecolle, 2004) qui désigne l'explosion du réacteur nucléaire de la centrale de Tchernobyl en 1986, ou de l'héméronyme *11 septembre* (Steimberg, 2008) qui nomme les attentats de New York. De plus, les travaux d'Ehrmann & Hagège (2009) développent des indices pour l'extraction des expressions temporelles qui ne sont pas des événements. Par opposition, certains de ces indices peuvent permettre de reconnaître des événements. Ces définitions et ces travaux restent pourtant théoriques et sont peu adaptables directement au repérage automatique des EN et à leur extraction par la machine.

Nombre de travaux se sont intéressés à l'extraction d'EN (cf. Ehrmann (2008) pour un historique complet sur la reconnaissance des EN), mais peu d'entre eux se sont focalisés sur la catégorie des événements. Il est à noter que deux campagnes d'évaluation d'extraction d'EN ont abordé le sujet : ACE (Doddington *et al.*, 2004) et Ester (Gravier *et al.*, 2004). Un système de questions-réponses ayant pour objectif la communication homme-machine fondée sur l'oral est développé dans le cadre du projet Ritel (Rosset *et al.*, 2005). L'extraction des EN y a été mise au point, et en particulier celle des événements. Les corpus une fois transcrits sont analysés et enrichis, notamment en EN classiques (lieu, organisation, personne ou date). Une entité nommée y est définie comme une "expression décrivant un modèle spécifique d'un type donné". Dans ce cadre, *le festival de Cannes* est un événement non défini, une entité non-précise, tandis que *le festival de Cannes 2006* est lui une entité nommée (événement précis).

Nous nous intéressons en priorité aux événements sous leurs formes nominales, en tant qu'EN. Les entités uniques comme les noms d'événements historiques (*la Grande Guerre*), celles plus récurrentes (*Festival*

[1]Ce travail a été partiellement financé par OSEO dans le cadre du programme Quaero.

[2]Même si notre corpus de travail est essentiellement constitué d'articles de presse, nous ne nous intéressons pas uniquement aux événements médiatiques ou journalistiques. Notons tout de même que ce sont les médias qui en général nomment les événements.

de Cannes), l'instanciation de ces phénomènes (*les JO de 1996*), les noms de fête (*Noël*). Nous souhaitons aussi nous attarder sur les phénomènes plus anodins comme *la descente de police de demain* ou moins définissables et plus flous comme *le branle-bas de combat mondial*, *le débat* ou *la décision*. Les entités recherchées peuvent être passées, futures ou hypothétiques. Ainsi nous n'écartons pas l'analyse des noms d'éventualités au sens de Vendler.

3 Utilisation des verbes de cause-conséquence

Afin de dégager au mieux des groupes nominaux désignant des événements, nous avons précédemment constitué une liste des déclencheurs de noms d'événement. À cette fin, deux lexiques ont été utilisés : une liste de déclencheurs avérés et une liste formée à partir des substantifs du lexique VerbAction (Hathout *et al.*, 2002). La première liste de déclencheurs (681 termes lemmatisés) a été constituée à partir des mots événements prévus dans Wmatch (Galibert, 2009), un outil conçu dans le cadre du projet Ritel (cf. Section 2) : 39 termes présents dans des grammaires locales descriptives restrictives (permettant par exemple de ne pas récupérer *guerre* en déclencheur de nom d'événement dans les cas où il apparaît dans l'expression *navire de guerre*). Cette liste est enrichie par quelque 588 lemmes du lexique EventNominals de (Bittar, 2009), constitué par des lemmes de substantifs "ayant au moins une interprétation événementielle". De nombreux mots de ce lexique appartiennent à des registres de langue particuliers, comme *anticoagulothérapie*.

Concernant le lexique VerbAction, il est constitué d'une liste de verbes d'action accompagnée des noms déverbaux morphologiquement apparentés à ceux-ci (9393 couples verbe-nom, soit 9200 lemmes nominaux uniques). Les verbes d'action impliquant que quelque chose se produit (*fêter*), les noms déverbaux de ces verbes devraient donc décrire une action (*fête*) et donc potentiellement nommer l'événement qui a lieu lorsque cette action se produit (*la fête de la musique*).

Il est possible d'utiliser des listes de mots déclencheurs événementiels ou des déverbaux pour reconnaître les EN-E. Cependant certaines EN-E ne sont pas détectables par ce moyen, c'est le cas de certaines expressions qui ne renferment à l'origine aucun trait événementiel, comme *le Watergate*, *les frégates de Taiwan*, *le sang contaminé* ou *Clearstream*. Ainsi, *Copenhague*, fin 2009-début 2010, ne désignait plus seulement aux yeux du monde la capitale du Danemark, mais surtout la conférence des Nations Unies sur le changement climatique qui s'y est tenue. La rareté impliquant la qualité, il ne nous est pas possible de concevoir passer à côté de ces termes qui dans l'imaginaire collectif sont étroitement liés (ne serait-ce que sur une courte période) à un événement particulier. Dans le cas de mots polysémiques, la tâche est aussi complexe. En effet, le mot *salon* désigne le mobilier et la pièce qui le reçoit, autant que le lieu d'exposition et l'événement qui s'y déroule. Avec une majuscule, *Salon* est régulièrement analysé en EN de type lieu, car c'est un nom de ville. Dans une phrase comme *Le Salon de l'Agriculture est organisé Porte de Versailles*, on souhaite valider le groupe nominal *Salon de l'Agriculture* en tant que nom d'événement.

L'hypothèse que nous cherchons à vérifier dans le travail présenté ici est que l'utilisation des verbes qui impliquent la cause ou la conséquence pourrait constituer un indice pour la reconnaissance de ces expressions en EN-E et pour l'obtention d'expressions "candidates". Nous appelons "expression candidate" une expression qui ne représente pas un événement en temps normal, mais qui dans un certain contexte en est un (les héméronymes, les toponymes ou les noms polysémiques). Constituer une liste de ces expressions peut bien entendu être précieux pour faciliter l'extraction de ces événements par la suite. Par exemple, *11 septembre*, dans un titre d'article, peut être un événement, mais également une simple date, tandis que

12 septembre, ne peut être *a priori* qu'une date.

Une action ou un événement peut être la cause d'un autre événement : un événement "provoque" ainsi un autre événement en conséquence. Les verbes *entraîner* ou *provoquer* peuvent fonctionner de la sorte. Dans *La crise économique entraînera la famine dans de nombreux pays sous-développés*, le verbe *entraîner* a pour sujet *la crise économique* et pour objet *la famine*. *Famine* est l'événement conséquence de l'autre EN-E de la phrase, *la crise économique*. C'est aussi le mode de fonctionnement du verbe *signer* dans *Le 11 septembre signe la fin de cette hégémonie sur le reste du monde*. Ici, *signer* présente deux événements, l'un (*11 septembre*) cause de l'autre (*fin de cette hégémonie sur le reste du monde*).

Nous souhaitons donc vérifier si les syntagmes nominaux en position sujet ou argument de certains verbes de cause-conséquence sont généralement des événements.

4 Expérimentations

Pour mener à bien notre étude sur l'intérêt des verbes de cause-conséquence dans le cadre de l'extraction d'EN-E, nous avons privilégié une approche lexicale du problème. À partir d'une liste de verbes dégagée au cours d'études de corpus préalables, nous avons prélevé des syntagmes nominaux (SN) issus des contextes gauches et droits au moyen de grammaires locales développées avec Wmatch. Deux annotateurs (expert) ont ensuite filtré manuellement les SN extraits pour ne conserver que les groupes en position sujet et argument, en tenant également compte des sujets inversés. Ceci permet de s'affranchir des éventuelles erreurs du système. Rappelons que le but n'est pas de tester un système (permettant ou non une analyse syntaxique), mais d'évaluer dans quelle mesure certains verbes sont accompagnés de noms d'événements.

En parallèle, les annotateurs ont indiqué si le sujet du verbe (s'il existe) et si l'argument le plus proche de ce verbe (si un argument a été extrait) représentent ou non des noms d'événements. Au total, 4345 verbes ont été annotés en une dizaine d'heures, pour un total de 5016 noms. L'accord inter-annotateur est jugé bon (kappa = 0,79 (Cohen, 1960)). Puis les verbes ont été regroupés en fonction de leur lemme, de leur préposition et de leur pronominalisation (*expliquer* et *s'expliquer par* sont deux entités distinctes étant donné leur fonctionnement syntaxique différent). On obtient ainsi 89 unités verbales.

Les tableaux suivants présentent les verbes qui ont, pour au moins 75% de leurs occurrences dans le corpus, un événement en position sujet (Tableau 1.a) ou en argument (Tableau 1.b). Bien entendu, certains de ces chiffres sont peu significatifs étant donné leur nombre d'occurrences, comme *avoir pour origine* présent dans *que les crises aient pour origine des problèmes de défaillance technique, de santé publique, etc.* ou *tirer les leçons de* dans *le gouvernement se réunira pour tirer les leçons des élections*. Nous avons cependant choisi de les conserver dans cette liste parce qu'ils nous semblent particulièrement pertinents.

Il est intéressant de constater que 305 noms d'événements du corpus ne sont pas présents dans les listes pré-établies. Par exemple, si *conflit* y est présent (*le conflit Danone peut donner naissance à une forme d' alliance entre salariés et consommateurs*), ce n'est pas le cas de *mise en sourdine* dans *cette élection entraînera-t-elle la mise en sourdine des intérêts communaux ?* ni de *tollé* (*provoqué un tollé chez les organisations amérindiennes*) ou de *revers* (*subissent un cuisant revers*). Ces mots peuvent donc être intégrés dans nos lexiques.

Par ailleurs, nous souhaitions également vérifier une autre hypothèse selon laquelle "un événement provoque un événement", c'est-à-dire la configuration dans laquelle sujet et argument d'un verbe sont tous

Verbe infinitif	Occurrences	Pourcentage d'événements à gauche
avoir lieu	89	100%
se produire	45	94%
provoquer	42	76%
s'expliquer par	12	92%
se traduire par	12	80%
affecter	10	83%
aboutir à	7	78%
précipiter	4	80%
se passer	4	80%
avoir pour origine	1	100%
être entraîner	1	100%
rendre à	1	100%
se donner	1	100%

a) Position sujet

Verbe infinitif	Occurrences	Pourcentage d'événements à droite
provoquer	134	87%
organiser	120	94%
permettre	85	79%
subir	84	76%
déclencher	56	100%
conduire à	55	93%
assister à	53	93%
contribuer à	46	81%
aboutir à	38	81%
se traduire par	34	87%
donner lieu à	22	100%
perpétrer	16	80%
inciter à	5	100%
occasionner	1	100%
se précipiter à	1	100%
tirer les conséquences de	1	100%
tirer les leçons de	1	100%

b) Position argument

TAB. 1 – Présence à 75% et plus d'un SN désignant un événement en position sujet ou argument de verbes de cause-conséquence

les deux des événements. Sur 670 verbes présentant une annotation des sujet et argument (31 verbes différents), 181 occurrences seulement présentent la configuration événement-verbe-événement, et aucun verbe ne se détache vraiment pour démontrer notre hypothèse. Le meilleur exemple, le verbe *provoquer*, compte 30 occurrences de ce type pour 45 triplets, comme *son arrestation provoque des manifestations mi-religieuses, mi-politiques*. Le tiers a pour argument des conséquences matérielles comme dans *une autre mini-tornade a provoqué des dégâts à Villeneuve-lès-Maguelone* ou pour sujet des personnes ou assimilés personne, exemple : *le Conseil de prévention et de lutte contre le dopage avait provoqué une petite crise avec l'Union cycliste*. On peut aussi noter *donner lieu à* qui 7 fois sur 10 vérifie cette hypothèse, *le rachat de USA Networks ne donnera lieu ni à création d'actions nouvelles ni à d'importantes sorties d'argent liquide*.

Enfin, les deux dates du corpus représentant des événements ont été repérées au moyen des verbes de cause-conséquence. Il s'agit de *11 septembre* (*le 11 septembre aura précipité une récession*) et *mai 68* (*mai 68 a précipité sa disparition*). Même si les occurrences sont peu nombreuses, ce résultat est intéressant. En effet, une méthode d'extraction basée sur les verbes de cause-conséquence peut conduire à construire une liste de dates ou de lieux qui peuvent potentiellement se comporter comme des événements, et donc d'en améliorer l'extraction.

5 Conclusion et perspectives

Nous envisageons dans un futur proche de mener à terme une autre partie de cette étude qui consiste à vérifier la validité de nos listes de déclencheurs. Une perspective intéressante serait d'utiliser l'analyse syntaxique afin de travailler sur les sujets et compléments d'objet des verbes qui nous intéressent, à la suite de cette étude préalable purement lexicale. Enfin nous prévoyons d'intégrer l'utilisation des verbes de cause-conséquence qui se sont dégagés de notre étude afin de collecter des mots déclencheurs d'événements et d'extraire au mieux nos EN-E.

Références

BITTAR A. (2009). Annotation of events and temporal expressions in french texts. In *ACL-SIGANN*.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

DODDINGTON G., MITCHELL A., PRZYBOCKI M., RAMSHAW L., STRASSEL S. & WEISCHEDEL R. (2004). The Automatic Content Extraction program - tasks, data, and evaluation. In *LREC'04*.

EHRMANN M. (2008). *Les Entités Nommées, de la linguistique au Tal : Statut théorique et méthodes de désambiguïsation.* PhD thesis, Université Paris 7.

EHRMANN M. & HAGÈGE C. (2009). Proposition de caractérisation et de typage des expressions temporelles en contexte. In *Actes de TALN 2009*, Avignon.

GALIBERT O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris-Sud 11, Orsay, France.

GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). Ester, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Proceedings of JEP'04*, Fèz, Maroc.

HATHOUT N., NAMER F. & DAL G. (2002). An Experimental Constructional Database: The MorTAL Project. In P. BOUCHER, Ed., *Many Morphologies*, p. 178–209. Somerville, Mass.: Cascadilla.

LECOLLE M. (2004). Toponymes en jeu : Diversité et mixage des emplois métonymiques de toponymes. In *Studii si cercetari filologice 3 / 2004, Université de Pitesti, Roumanie*.

NEVEU E. & QUÉRÉ L. (1996). Présentation. *Réseaux*, **14**(75), 7–21.

PUSTEJOVSKY J., CASTAÑO J., INGRIA R., SAURÍ R., GAIZAUSKAS R., SETZER A. & KATZ G. (2003). TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5*.

ROSSET S., GALIBERT O., ILLOUZ G. & MAX A. (2005). Interaction et recherche d'information : le projet RITEL. *TAL. Traitement automatique des langues*, **46**(3), 155–179.

STEIMBERG L. C. (2006). La construction de la mémoire historico-médiatique à travers les désignations d'événements. *Studies van de BKL 2006 - Papers of the LSB 2006*.

STEIMBERG L. C. (2008). Les héméronymes. ces évènements qui font date, ces dates qui deviennent évènements. *Mots. Les langages du politique*, **3**, 115–128.

VELDE D. V. D. (2000). Existe-t-il des noms propres de temps ? *Lexique*, **15**, 151.

VENDLER Z. (1967). *Verbs and Times*, In *Linguistics in Philosophy*, p. 97–121. Cornell University Press: Ithaca, NY, USA.