Désignations nominales des événements : étude et extraction automatique dans les textes

Béatrice Arnulphy

École Doctorale d'Informatique de Paris-Sud (edips)
Laboratoire d'Informatique, de Mécanique et de Sciences de l'Ingénieur (limsi)

THÈSE DE DOCTORAT EN INFORMATIQUE

soutenue le *02 octobre 2012*

par

Béatrice Arnulphy

Désignations nominales des événements :
Étude et extraction automatique dans les textes

Version de manuscrit du *08 novembre 2012*

Jury constitué de

Directeur de thèse	Anne Vilnat Université Paris-Sud – Limsi	(Professeur)
Encadrant	Xavier Tannier Université Paris-Sud – Limsi	(Maître de Conférence)
Rapporteurs	Laurence Danlos Université de Paris7 – Alpage (INRIA)	(Professeur)
	Patrice Bellot Polytech – LSIS, Université d'Aix-Marseille	(Professeur)
Examinateurs	Sophie Rosset CNRS, Limsi	(Directeur de Recherche)
	Laura Calabrese Université Libre de Bruxelles – ReSIC	(Maître de Conférence)
	Philippe Muller Université Paul Sabatier, Toulouse – Irit	(Maître de Conférence)

Résumé

Ma thèse a pour but l'étude des désignations nominales des événements pour l'extraction automatique. Mes travaux s'inscrivent en traitement automatique des langues, soit dans une démarche pluridisciplinaire qui fait intervenir linguistique et informatique.

L'extraction d'information a pour but d'analyser des documents en langage naturel et d'en extraire les informations utiles à une application particulière. Dans ce but général, de nombreuses campagnes d'extraction d'information ont été menées : pour chaque événement considéré, il s'agit d'extraire certaines informations relatives (participants, dates, nombres, etc.). Dès le départ, ces challenges touchent de près aux entités nommées (éléments « notables » des textes, comme les noms de personnes ou de lieu). Toutes ces informations forment un ensemble autour de l'événement. Pourtant, ces travaux ne s'intéressent que peu aux mots utilisés pour décrire l'événement (particulièrement lorsqu'il s'agit d'un nom). L'événement est vu comme un tout englobant, comme la quantité et la qualité des informations qui le composent.

Contrairement aux travaux en extraction d'informations générale, notre intérêt principal est porté uniquement sur la manière dont sont nommés les événements qui se produisent et particulièrement à la désignation nominale utilisée. Pour nous, l'événement est ce qui arrive, ce qui vaut la peine qu'on en parle. Les événements plus importants font l'objet d'articles de presse ou apparaissent dans les manuels d'Histoire. Un événement peut être évoqué par une description verbale ou nominale.

Dans cette thèse, nous avons réfléchi à la notion d'événement. Nous avons observé et comparé les différents aspects présentés dans l'état de l'art jusqu'à construire une définition de l'événement et une typologie des événements en général, et qui conviennent dans le cadre de nos travaux et pour les désignations nominales des événements. Nous avons aussi dégagé de nos études sur corpus différents types de formation de ces noms

d'événements, dont nous montrons que chacun peut être ambigu à des titres divers. Pour toutes ces études, la composition d'un corpus annoté est une étape indispensable, nous en avons donc profité pour élaborer un guide d'annotation dédié aux désignations nominales d'événements. Nous avons étudié l'importance et la qualité des lexiques existants pour une application dans notre tâche d'extraction automatique. Nous avons aussi, par des règles d'extraction, porté intérêt au cotexte d'apparition des noms pour en déterminer l'événementialité. À la suite de ces études, nous avons extrait un lexique pondéré en événementialité (dont la particularité est d'être dédié à l'extraction des événements nominaux), qui rend compte du fait que certains noms sont plus susceptibles que d'autres de représenter des événements. Utilisée comme indice pour l'extraction des noms d'événements, cette pondération permet d'extraire des noms qui ne sont pas présents dans les lexiques standards existants. Enfin, au moyen de l'apprentissage automatique, nous avons travaillé sur des traits d'apprentissage contextuels en partie fondés sur la syntaxe pour extraire de noms d'événements.

Mots-clefs – événement – dénomination – désignations nominales – entités nommées – extraction automatique – traitement automatique des langues (TAL)

Table des matières

Introduction générale

Un événement est ce qui arrive, ce qui vaut la peine qu'on en parle. Les événements plus importants font l'objet d'articles de presse ou apparaissent dans les manuels d'Histoire. Un événement peut être évoqué par une description verbale ou nominale. Ainsi, on pourra décrire des élections dans une phrase dont le verbe porte l'événement (description verbale de l'événement) « `le président n'a pas renouvelé son mandat` » ou « `l'UMP a perdu la bataille pour l'Élysée` » ou dans un groupe nominal (description nominale de l'événement) « `La défaite du président sortant` » ou « `l'échec de l'UMP aux élections présidentielles` ». Nous nous intéressons à toutes les désignations nominales référant aux événements, aux dénominations d'événement en général.

Nous utilisons le terme de « désignation nominale d'événement » pour parler des mots qui désignent en langue des événements, soit en terme grammatical ou de lexique, les syntagmes nominaux servant à désigner les événements. « Nom d'événement » est plus général, nous l'utilisons plus souvent pour évoquer les travaux des autres auteurs dont on ne sait pas toujours à quoi ce terme réfère exactement. La « nomination » correspond à l'acte de nommer et la « dénomination » qui en est le résultat existe au travers du discours, en temps que porteur du sens discursif.

Nous travaillons sur des corpus de presse écrite, parce que ce type de corpus est potentiellement plus riche en événement et facilement disponible.

Dans ce manuscrit de thèse, nous proposons une étude des désignations nominales d'événements dans des corpus de presse écrite en français, dont le but est l'extraction automatique de noms d'événements. Une partie de nos travaux a été étendue à l'anglais.

Les travaux de thèse que nous présentons ici se placent en traitement automatique des langues (TAL) et nous évoluons dans le domaine de l'extraction d'information.

Contexte de nos travaux

L'extraction d'information a pour but d'analyser des documents en langage naturel et d'en extraire les informations utiles à une application particulière. Cette tâche du TAL est définie par Poibeau (2003) :

> « L'extraction d'information désigne l'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle »
> (Poibeau, 2003, p. 13)

De nombreuses campagnes d'extraction d'information ont été menées dès la fin des années 1980. Elles avaient pour cadre les conférences *MUC* (Message Understanding Conference) (Grishman and Sundheim, 1996), puis les évaluations *ACE* (Automatic Content Extraction). Dans le cadre des conférences *MUC* et *ACE*, les textes considérés étaient des messages d'opérations navales, des comptes rendus militaires sur des attaques terroristes, des accords commerciaux dans le domaine de la micro-électronique, des rapports lors de lancements de satellites, etc. Ces textes sont des mines d'informations et existent parce qu'ils décrivent quelque chose qui se produit. « Quelque chose qui se produit » est d'ailleurs la définition vulgarisée de l'événement. L'événement joue donc dans le cadre de l'extraction d'information le rôle central.

Ces campagnes d'évaluation avaient pour but de remplir des bases de données : pour chaque événement considéré, une liste d'informations à la structure prédéfinie était à extraire, comme le type d'événement, l'identité des participants, la date de l'événement, les nombres, pourcentages, valeurs monétaires. Ces informations sont souvent décrites au moyen de noms et concernent les entités nommées. Dès *MUC-6* en 1995, les challenges ont porté sur les entités nommées. Voici quelques exemples d'entités nommées (EN) les plus répandues, ces exemples sont issus des guides d'annotation de *MUC-6*. Les EN « personne » sont des noms de personne, comme « `John Doe, Jr.` » ; les EN « locatives » sont des noms de ville, comme « `Kaohsiung` », ou de pays « `Taiwan` » ; les EN « organisation » sont des noms de groupes organisationnels, comme « `European Community` » ou par exemple des noms d'entreprise, comme « `Bridgestone Sports Co.` ». Toutes ces informations forment un ensemble autour de l'événement et ces travaux ne s'intéressent pas aux mots utilisés pour décrire l'événement lorsqu'il s'agit d'un nom. L'événement est vu comme un tout englobant, comme la quantité et la qualité des informations qui le composent.

Contrairement aux travaux en extraction d'informations générale, notre intérêt principal n'est pas porté à cet ensemble d'informations autour de l'événement. Nos travaux

concernent uniquement la manière dont sont nommés les événements qui se produisent et particulièrement à la désignation nominale utilisée. En effet, de nombreux travaux en traitement automatique de langues et extraction d'information se sont focalisés sur la description verbale des événements, mais peu d'entre eux se sont intéressés aux désignations nominales des événements. Il faut noter que lorsque l'événement n'est pas désigné sous une forme verbale, il a de grandes chances de l'être sous une forme nominale. Ne traiter que les verbes, c'est donc passer à côté des noms d'événements et prendre le risque de manquer les informations recherchées.

Par ailleurs, comme nous le décrirons dans ce manuscrit, tout comme les noms de personnes actrices ou actantes des événements ou les noms de lieux où ces événements se produisent sont des entités nommées, les noms d'événements pourraient être considérés comme des entités nommées. D'autant que la notion d'entités nommées n'est pas figée et que dans le cadre du projet *Quaero* [1], cette notion est étendue à de nouvelles catégories d'entités comme les événements (Grouin *et al.*, 2011).

Le programme *Quaero* est décrit comme « un programme fédérateur de recherche et d'innovation industrielle sur les technologies d'analyse automatique, de classification et d'utilisation de documents multimédias et multilingues » [2]. En ce qui nous concerne, une partie du programme dédié aux analyses textuelles est consacrée à un projet général d'extraction d'information dont l'objectif est la constitution d'une base de connaissances et dont les entités nommées sont le pivot. Le projet *Quaero* nous fournit un cadre pour nos travaux sur l'extraction des noms d'événements.

Contributions

Avant tout, nous avons réfléchi à la notion d'événement. Nous avons observé et comparé les différents aspects présentés dans l'état de l'art jusqu'à construire une **définition de l'événement et une typologie des événements** en général qui conviennent dans le cadre de nos travaux et pour les désignations nominales des événements. Nous avons aussi dégagé de nos études sur corpus différents **types de formation de ces noms d'événements**, dont nous montrons que chacun peut être ambigu à des titres divers.

Pour toutes ces études, la composition d'un **corpus annoté** est une étape indispensable, nous en avons donc profité pour élaborer **un guide d'annotation** dédié aux désignations nominales d'événements.

Nous avons étudié l'importance et la qualité des **lexiques existants** pour une ap-

1. Cette thèse a été partiellement financée par *OSEO* dans le cadre du programme *Quaero*.
2. http://www.quaero.org/modules/movie/scenes/home/index.php?fuseAction=article&rubric=presentation

plication dans notre tâche d'extraction automatique. Nous avons aussi, par des règles d'extraction, porté intérêt au **contexte d'apparition des noms** pour en déterminer l'événementialité.

À la suite de ces études, nous avons extrait **un lexique pondéré en événementialité**, qui rend compte du fait que certains noms sont plus susceptibles que d'autres de représenter des événements. Utilisée comme indice pour l'extraction des noms d'événements, cette pondération permet d'extraire des noms qui ne sont pas présents dans les lexiques standards existants. Enfin, au moyen de l'apprentissage automatique, nous avons travaillé sur des **traits d'apprentissage contextuels en partie fondés sur la syntaxe** pour extraire de noms d'événements.

Organisation du mémoire

Notre mémoire se présente en cinq chapitres. Le premier est un chapitre d'état de l'art, le deuxième un chapitre définitoire et les trois suivants des chapitres montrant les réalisations faites en terme de corpus annoté, d'extraction de lexique pour les noms d'événements et d'extraction automatique de noms d'événements au moyen de classifieurs automatique.

Le chapitre 1 présente un tour d'horizon des travaux qui nous ont interpellés dans le cadre de notre recherche pour cette thèse. Tout d'abord, le problème de définition de l'événement (section 1.1) : la notion n'étant pas toujours facile à cerner et souvent même pas définie dans certains travaux, comme si elle était évidente, alors qu'en réalité l'événement est un objet complexe qui nécessite qu'on le définisse. Nous avons aussi porté intérêt aux événements nominaux, pour en venir à une étude des entités nommées et spécifiquement faire un point sur les travaux considérant les événements comme des entités nommées.

Cet état de l'art ne serait pas complet sans un recensement des ressources existantes en lien avec les événements (section 1.2), que ce soient les travaux sur les représentations temporelles et celles des événements (*TimeML*, *ACE* pour ne citer qu'eux), ainsi que les corpus annotés qui sont utilisés dans ces représentations.

La dernière partie de notre état de l'art est consacrée aux travaux en extraction automatique qui touchent à notre cadre de travail (section 1.3) : les approches qui touchent aux événements en général et celles qui se focalisent sur les événements nominaux en particulier.

Le chapitre 2 est dédié à la notion d'événement (section 2.1) et présente notre cheminement entre linguistique et traitement automatique des langues pour dégager une définition générale de l'événement.

Nous avons de plus développé une typologie des événements (section 2.2), qui fait ressortir les informations importantes concernant les événements et qui sont importantes en extraction d'information (modalité, fréquence et moment de réalisation de l'événement).

Pour finir, nous orientons notre travail sur l'événement nominal en particulier (section 2.3), l'objet même de nos travaux, en décrivant les différents types de construction du nom d'événement.

Le chapitre 3 s'attache au recueil, à l'observation et à l'annotation de corpus. Le but de nos travaux étant l'extraction d'information, soit extraire de manière automatique les noms d'événements directement dans les textes, il est important de repérer en contexte l'objet qui nous intéresse, en l'occurrence les noms d'événement. Dans le cadre d'une extraction automatique, il est nécessaire de développer une référence afin de pouvoir évaluer les performances des ressources, règles et autres indices que nous souhaiterions utiliser. Notre référence est un corpus annoté manuellement, qui suit les règles et indices d'annotation présents dans un guide d'annotation que nous avons rédigé (section 3.1).

Notre guide d'annotation pour les noms d'événements est développé dans le cadre du projet Quaero, dans sa partie dédiée aux entités nommées. Notre corpus annoté est composé d'articles journalistiques (section 3.2) issus du quotidien national français *Le Monde* et du quotidien régional *L'Est Républicain*.

Nous avons observé le comportement des noms en lecture événementielle dans notre corpus annoté (section 3.3).

Le chapitre 4 évalue l'importance des lexiques dans la reconnaissance des désignations nominales d'événements. Nous présentons d'abord les lexiques existants (section 4.1), avant de présenter les ressources que nous avons développées (règles d'extraction temporelles et verbales) pour créer une ressource lexicale plus approfondie et très orientée vers l'extraction des événements nominaux (section 4.2).

En appliquant ces règles (très précises et d'un rappel faible) sur un grand corpus, nous créons des lexiques pondérés par une valeur événementielle relative pour l'extraction des noms d'événements (section 4.3).

Le chapitre 5 est un ensemble d'expérimentations de classification automatique des noms suivant qu'ils appartiennent ou non à la classe « événement ». Nous proposons une

approche par apprentissage automatique sur des corpus manuellement annotés et sur des corpus créés de manière automatique sur la base d'amorces.

Dans un premier temps, nous présentons notre choix concernant le type de classifieur utilisé et les traits d'apprentissage (section 5.1). Puis, nous décrivons les corpus d'apprentissage et de tests constitués pour cette tâche (section 5.2), où nous testons l'utilisation des indices contextuels en plus des indices lexicaux pour améliorer l'extraction automatique. Enfin, nous présentons les tests et résultats obtenus (section 5.3), avant de conclure, en nous comparant aux autres travaux existants.

Chapitre 1

État de l'art

Différents traitements de la notion d'événement existent selon les disciplines qui l'étudient. Elle est par exemple abordée et discutée en sciences humaines et en sciences du langage, où l'on crée des typologies de l'événement et des travaux sont menés sur les types d'événements et sur la nomination des événements. Le plus souvent en TAL, la notion d'événement n'est pas définie, on parle d'événement, comme si cette notion était claire et établie pour tous. Elle reste construite de manière empirique, le plus souvent en fonction de l'application qui en a besoin, comme c'est le cas dans le cadre des entités nommées, par exemple.

Dans cet état de l'art, notre première partie est consacrée à l'événement, au problème de définition de la notion en général et aux aspects plus précis liés à la nomination de l'événement, ce qui nous mène à évoquer le traitement des événements dans le cadre des entités nommées en TAL.

Notre deuxième partie est liée à l'existant en terme de représentations temporelles, corpus et ressources utilisables pour des traitements automatiques d'extraction des événements.

La troisième partie est dévolue aux travaux existants en extraction automatique des événements : D'abord les travaux portant sur l'ensemble des événémements indépendamment de la catégorie morpho-syntaxique (travaux généraux sur les verbes, adjectifs, noms, etc.), puis plus précisément ceux portant sur la catégorie grammaticale des noms. Nous décrirons particulièrement les approches linguistiques desquelles nous nous sentons proches.

1.1 L'événement

Si la notion d'événement ne conduit à aucune définition stricte, précise et consensuelle, on retrouve néanmoins toujours les notions importantes suivantes : mouvement, insertion dans le temps, modification d'une situation, idée de fin et d'importance pour une communauté. Rappelons la définition du mot « événement » dans nos dictionnaires de langue, ici le TLFI (via le portail CNRTL [1]) :

> « ÉVÉNEMENT, subst. masc.
> A.– Fait auquel aboutit une situation. [En parlant d'une action dram.] Dénouement. [Dans certaines expr.] Ce qui se produit par la suite.
> B.– Tout ce qui se produit, tout fait qui s'insère dans la durée. Fait d'une importance notable pour un individu ou une communauté humaine. *au plur.* Ensemble des faits plus ou moins importants de l'actualité. »
> http://www.cnrtl.fr/definition/%C3%A9v%C3%A8nement

Dans l'événement est présentée une idée de mouvement, de modification d'une situation. L'événement se produit et dès lors une modification d'un état de chose est opérée, on parle de changement d'état. L'événement est ancré dans le temps, il dure plus ou moins longtemps (un événement instantané, comme le passage à l'an 2000 ou une durée courte comme une expression faciale qui disparaît l'instant d'après sa réalisation ou qui s'insère dans la durée, comme une période historique ou un mandat politique). Un événement peut-être borné dans le temps, parce que dès le départ on connaît sa durée du moins théorique (mandat politique) ou pas (une carrière de chanteur qui débute). L'événement peut être composé d'autres événements (au cours d'un événement « `guerre` » inscrit sur la durée, on observe des événements de « `fusillade` », « `destruction` », « `discours politique` » ou même « `émission de télévision` », etc.). L'événement a une réalité plus ou moins tangible. On peut parler d'événements naturels, qui sont plus ou moins prévisibles et plus ou moins soudains, comme les tsunamis ou les ouragans, l'important étant qu'ils aient une importance sur notre monde. En effet, l'impact de ce qui se produit est prépondérant pour qualifier d'événement ce qui arrive et pour évaluer son importance. De plus dans l'événement est présentée l'idée de fin, ce qui n'implique pas toujours la réalisation de cette fin.

Ainsi, Vendler (1959) propose une étude des classes de verbes en fonction de leur valeur aspectuelle, c'est-à-dire la façon dont le procès ou l'état exprimé par le prédicat est envisagée du point de vue de son développement. La classification de Vendler fait apparaître 4 classes de verbes :

1. http://www.cnrtl.fr/

- les verbes d'état (STATE : les états étant statiques et sans point final) :
  `admirer`, `posséder`, `connaître`, `aimer`.
- les verbes d'activité (PROCESS : pour les actions dynamiques et qui n'ont pas de point final) :
  `manifester`, `discuter`, `se promener`, `courir`.
- les verbes d'accomplissement (ACCOMPLISHMENT : pour les verbes d'action à durée déterminée) :
  `construire` (une maison), `grandir`, `se préparer` (à faire quelque chose), `peindre un tableau`.
- les verbes d'achèvement (ACHIEVEMENT : pour les actions instantanées et qui ont une fin, aussi connue sous le nom de culmination) :
  `reconnaître`, `découvrir` (un objet), `résumer`, `se réveiller`.

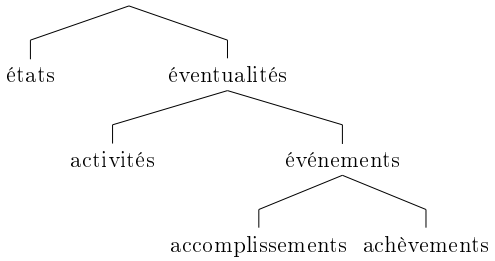


FIGURE 1.1 – Classification des verbes selon (Vendler, 1959) avec ajouts de (Verkuyl, 1989)

Dans (Vendler, 1959), les « éventualités » se distinguent des états par la dynamicité. De nombreux ajouts et modifications ont suivi ces travaux sur la classification des verbes. Chez Verkuyl (1989), l'achèvement et l'accomplissement de par leur télicité remplissent une classe « événement ». Dans Moens and Steedman (1988), on compte cinq classes (quatre types d'événements et les états). Les événements sont distingués suivant deux dimensions contrastives, la première concerne ponctualité et durée, la seconde concerne l'association à un état de conséquence (cf. figure 1.2). Dans cette classification sont distingués « CULMINATION » (correspondant à l'achèvement), « CULMINATED PROCESS » (l'accomplissement), « POINT » (l'événement ponctuel sans durée, comme « `faire un clin d'oeil` ») et « PROCESS » (le procès).

Les travaux sur les verbes sont nombreux depuis Vendler qui se sont focalisés sur les différents aspects de l'événement, mais il se pose toujours le problème de la définition de l'événement que nous allons essayer de cerner ici dans les grandes lignes, en nous orientant vers les études plus précises sur les désignations nominales des événements et, de fil en aiguille, sur les entités nommées.

	EVENTS		STATES
	atomic	extended	
+conseq	**CULMINATION** recognize, spot, win the race	**CULMINATED PROCESS** build a house, eat a sandwich	understand, love, know, resemble
-conseq	**POINT** hiccup, tap, wink	**PROCESS** run, swim, walk, play the piano	

FIGURE 1.2 – Classification aspectuelle des verbes selon (Moens and Steedman, 1988)

1.1.1 Un problème de définition

En histoire, l'événement est considéré comme un fait notable « méritant d'être relaté par les historiens ». Des philosophes, comme Davidson (1993), ont focalisé leur attention sur la notion d'événément. Des sociologues du langages et linguistes se sont attachés à décrire l'événement médiatique, son mode de développement et son importance sur le monde. Des linguistes se sont intéressés à cette notion, aux mécanismes utilisés pour la mise en discours et à comment les mots se chargent d'une valeur événementielle qui n'est pas intrinsèque à ces désignants événementiels.

En philosophie, Davidson (1993) développe « une théorie causale de l'action humaine, de l'action et de l'esprit », dont la catégorie ontologique fondamentale est l'événement. Il définit l'événement de la manière suivante :

> « On peut dire, grossièrement, que les événements sont des choses qui arrivent à un certain moment. »
> (Davidson, 1993, in Essai n°11).

Cette définition est, comme le dit son auteur, grossière, mais finalement peu de définitions dépassent ce lieu commun.

Pour Davidson, l'action est un type d'événement. Si l'événement est « une chose qui arrive » sans implication de quelque agent que ce soit, l'action par contre a un trait agentif activé. L'action y est définie comme « tout ce qu'un agent fait intentionnellement, y compris les omissions intentionnelles ».

Par ailleurs, Davidson considère que l'événement existe en plus d'avoir lieu, il le voit comme un objet qui aurait une existence dans l'espace-temps (van de Velde (2006) s'en

inspire d'ailleurs dans sa grammaire des événements). L'illustration proposée est la suivante : si on parle d'explosion, on présuppose que l'événement « `explosion` » existe, ce qui permettrait à un événement d'être décrit et nommé en langue. On peut donc y référer au moyen des mots « *explosion* » ou « *catastrophe* » indépendamment. Pourtant, comme le fait remarquer (Veniard, 2007), le poids sémantique entre ces deux noms pour le même événement n'est pas le même, la représentation qu'on en fait non plus, « si *explosion* est une description de l'événement, le décrire par le mot *catastrophe* change son rayon d'action ».

Par ailleurs, une réflexion importante a été développée depuis les années 70 sur la notion d'**événement médiatique**, comme un événement qui est porté par les médias, jugé suffisamment important pour qu'on en parle dans les journaux et qu'on en reparle pour le mettre à jour (cf. section 2.1.1). Ces aspects nous intéressent particulièrement. Certes, notre intrêt ne se focalise dans l'absolu pas seulement sur l'événement médiatique ou journalistique, mais d'une part nos corpus d'étude sont tous constitués d'articles de presse, et d'autre part, ce sont souvent les médias qui nomment les événements (d'après Krieg-Planque (2009), les événements permettent la justification de la prise de parole, parce qu'ils sont un apport d'information) ; c'est pourquoi notre traiterons pour l'essentiel de la dénomination médiatique des événements.

Des travaux se sont donc intéressés à « ce qui fait événement » et comment les médias le créent. Neveu and Quéré (1996) présentent l'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 » (il aurait pu ne pas avoir lieu ou l'être autrement).

Dans ses travaux sur l'étude des discours dans la presse quotidienne, Moirand (2007) introduit le critère de la médiatisation dans la définition de l'événement et propose l'introduction de la notion de **moment discursif** dans la définition de l'événement. Un moment discursif est une période de temps au cours de laquelle suite à la survenue d'un fait, se produit une production massive de discours dans les médias. Nous pouvons noter que ce sont les médias qui font de quelque chose qui se produit, un événement. Même si au départ, l'événement est nommé en fonction de son importance, ce sont les journalistes, qui en l'évoquant et en le rendant redondant, accroissent en partie l'importance de l'événement. Dans (Calabrese, 2008), l'événement est défini comme « un ensemble de faits d'intérêt public et identifiés comme un tout unique et non répétable ». Par opposition au fait divers, dit l'auteur, l'événement porte un nom qui assure sa traçabilité au fil du temps. Comme dit précédemment, notre objet d'étude sont les noms donnés aux événements, les dénominations d'événements. Notre base de travail textuelle est constituée

d'articles de presse, c'est pourquoi nous nous intéressons particulièrement aux événements médiatiques, qui sont qualifiés d'importants. Par ailleurs, nous considérons aussi tous les événements désignés par des groupes nominaux, dénominations d'événements plus banals.

La littérature en sciences du langage sur les événements en général est fournie. Loin de proposer ici un état de l'art complet des définitions de l'événement en linguistique ou des différentes conceptions de l'événement en sciences humaines, nous proposons une présentation de certains travaux dont la problématique se rapproche de la nôtre et/ou dont les idées seraient adaptables en TAL (étant donné que notre objectif est l'extraction automatique des désignations nominales d'événements).

Dans (Lecolle, 2009), par exemple, nous est proposé une définition de l'événement, l'auteure s'intéressant principalement à l'usage événementiel des toponymes et donc à certaines dénominations particulières des événements. L'événement doit produire « un changement d'état du monde », il est saillant relativement à un individu ou une collectivité et « se distingue » ; ancré dans une temporalité, il est « en relation avec d'autres événements » ; sa désignation doit de plus faire sens. Cette définition reste suffisamment générale pour convenir aux événements dans leur ensemble. La définition proposée n'est spécifique ni aux toponymes en emploi événementiel ni aux désignations nominales des événement uniquement. Nous nous y retrouvons bien dans notre approche d'extraction d'information.

Dans les travaux sur l'héritage aspectuel des noms déverbaux en français et en espagnol, Huyghe and Marín (2007) prennent en exemple les noms dynamiques (c'est-à-dire non-statiques, dérivés de verbes d'achèvement ou d'accomplissement), qui dénotent des événements pour montrer l'expression de la délimitation temporelle de certains noms et prouver que certains noms héritent des traits aspectuels des verbes dont ils sont la nomination. Dans cette optique, les auteurs proposent une définition proche de notre conception de l'événement.

> « Les événements, au sens où nous l'entendons, sont des entités spatio-temporelles finies, dotées d'une certaine autonomie existentielle. Ce sont des choses qui arrivent, se produisent ou, plus généralement, qui « ont lieu ». Il est possible de leur assigner directement un site de localisation spatiale, en vertu de leur autonomie existentielle. »
> (Huyghe, 2006)

Nous pensons que la dénomination d'événement est intrinsèquement porteur des informations spatio-temporelles de l'événement qu'il décrit. De plus, nous adhérons à l'idée du changement d'état pour qu'il y ait événement. Dans la définition de (Huyghe and Marín, 2007), il n'est pas clairement dit que les états ne sont pas des événements, mais les événe-

ments sont vus comme des « choses qui arrivent, se produisent ou [...] qui ont lieu », ce qui implique un mouvement, un changement d'état. Les auteurs évoquent la possible durée ou par opposition le fait d'être ponctuel des événements portés par les noms : certains événements se déroulent et d'autres, « constatés après-coup » restent ponctuels.

Dans ces définitions, la notion de référence est importante. Nous nous proposons de l'aborder maintenant en lien avec le traitemant des événements.

La référence. En effet, lorsqu'on évoque un événement, que ce soit par sa dénomintaion propre ou par une description nominale définie, on évoque intentionnellement (ou non) à notre interlocuteur des références qui prennent la forme de lieux, de personnes, d'images, de noms, d'activités qui ont participé à cet événement, ou qui y sont rattachées. Nous ne pouvons pas ne pas tenir compte de la valeur référentielle des dénominations d'événements. On la considère dans l'absolu en tant que locuteur, lorsqu'on utilise des dénominations pour référer à des événements. Dans nos travaux présentés ici, nous y accordons forcément beaucoup d'importance en tant qu'annotateur pour déterminer si un syntagme nominale désigne un événement ou non.

En effet, Charolles (2002) définit la référence comme :

> « Tout un arrière de connaissances ou d'expériences plus ou moins partagées qui renvoient à la façon dont les interlocuteurs appréhendent intellectuellement ces situations. »

Il exprime également la nécessité d'analyser l'emploi des expressions référentielles en tenant compte du contexte dans lequel elles sont produites. Pourtant, dans les textes, le contexte de la situation d'énonciation n'est pas toujours présent. Dans le cadre d'une extraction des événements, c'est donc surtout le contexte linguistique ou le cotexte que nous pourrons utiliser, comme par exemple le vocabulaire employé, la modalité ou d'autres indices. Le contexte nous permet ainsi souvent (même si pas toujours) de lever l'ambiguïté sur la référence à un événement, lorsqu'elle est présente.

Lorsqu'un événement est nommé, le contexte permet de désambiguïser dans le cas où le mot n'est pas fondamentalement événementiel, ou pourrait ne pas référer à cet événement en particulier. Mais le mécanisme de référence à un événement est très complexe, tout autant que celui de la nomination de l'événement. Nous nous reportons ici aux travaux sur la nomination en lien avec les événements.

1.1.2 Les événements et les noms

Ce qui nous intéresse particulièrement dans les événements, ce sont leurs noms, les désignations nominales utilisées pour référer aux événements, nous souhaitons pouvoir en définitive extraire les mots qui désignent l'événement.Dans cette section nous nous focalisons sur le lien entre le nom donné à un événement (sa dénomination) et l'événement qui s'est produit. De par ce lien très fort, nous évoquons l'unicité parfois de la dénomination qui en contexte ne peut qu'être mis en rapport avec un événement particulier, ce qui nous mène à évoquer la possibilité de considérer les noms d'événements comme des noms propres. Pourtant les noms d'événements peuvent avant même de désigner des événements, avoir été déjà des noms propres mais d'autres types, nous évoquerons donc les glissements sémantiques de certains désignants d'événement.

1.1.2.1 La dénomination de l'événement

Reprenant les travaux de Kleiber (1997), Krieg-Planque (2009) rappelle la distinction entre dénomination et désignation. La **dénomination** est présentée comme « l'institution d'une association référentielle durable entre un objet et un signe ». La **désignation**, par opposition, « repose sur une association occasionnelle entre une séquence linguistique et un élément de la réalité » En ce qui concerne l'événement, nous pouvons donc établir qu'on passe d'un nom qui est une désignation à une dénomination, lorsque le nom de l'événement se fige dans le temps. Nous présupposons que la désignation nominale de départ est plus proche d'une description définie que d'un nom et qu'au fil du temps, du besoin de stéréotypisation et avec la constance d'y faire référence imposée par l'importance de l'événement, la désignation définie est réduite à une désignation plus courte, plus figée et qui devient dénomination en discours.

D'après Krieg-Planque (2009), il y a sens et utilité des dénominations d'événement pour les productions discursives des journalistes. Les événements permettent la justification de la prise de parole, parce qu'ils sont un apport d'information. En outre, le recours à des dénominations d'événement illustrent le besoin médiatique de *catégorisation* et celui d'*analogie* et de *comparaison*, tout autant que de *prototypicité* : nommer l'événement permet de le faire appartenir à une catégorie et ainsi de le comparer aux autres événements, le rendre représentatif d'une classe d'objets à laquelle il appartient.

Selon Calabrese (2010, p. 117), la dénomination est une **allusion à l'événement**.

> « Ces désignants ont la capacité de stocker les coordonnées de l'événement, et en conséquence d'éveiller la mémoire des faits par la seule mention du nom. »

(Calabrese, 2009)

L'auteure présente et étaie les travaux de Moirand (2004) sur ce rapport entre l'événement et la réalité comme une allusion : la capacité d'un mot à évoquer l'idée d'un événement repose sur l'allusion, parce que ce sont des dénominations partagées qui déclenchent les souvenirs. En effet, une fois le nom déchargé de toute désignation textuelle événementielle (amorce, terme classifieur (Veniard, 2009)) continue par métonymie de rappeler l'événement (« *catastrophe nucléaire de Tchernobyl* » une fois nominativement restreinte à « *Tchernobyl* » continue d'éveiller dans l'esprit des locuteurs l'événement qui s'y est déroulé, si le contexte le permet, bien évidemment). Cette hypothèse est considérée dans ses travaux, comme un constat valide pour fonder son étude sur le processus de nomination des événements et des critères remplis qui font des désignants d'événements des noms propres.

Dans les notions de désignation et d'allusion, nous sommes dans le cadre de la référence, qui touche notamment au contexte. Les dénominations d'événements ont une capacité déictique. En effet, ils font référence au contexte situationnel. Le contexte fait sens avec le désignation d'événement et nous permet d'accéder à la valeur sémantique du mot. En linguistique, on fait une **distinction entre le sens et le contenu**.

Dans ses travaux pour caractériser nom propre et nom commun, Gary-Prieur (1994) définit les notions de sens et de contenu. Le

> « *sens* [qui] caractérise le nom propre en tant qu'unité de langue et [qui] est représenté par le prédicat de dénomination. »

Il s'agit d'une propriété partagée avec le nom commun et qui relève du métalangage. Par contre, le

> « *contenu* des propriétés qui caractérisent le nom propre en tant qu'il est lié à son référent initial »

est une particularité du nom propre par rapport au nom commun. Il s'agit d'une information sur le monde, qui nous rapproche de la notion de référence linguistique* (vue précédemment en section 1.1.1). Même si nous sommes d'accord avec la vision de (Gary-Prieur, 1994, p. 38–42) sur les notions de sens et de contenu, nous ne ferons pas cette distinction très fine dans le mémoire. Nous parlerons plutôt de sens des noms désignant les événements et de la charge sémantique des dénominations d'événements.

Cette distinction sémantique entre les noms communs et les noms propres rapprochent les noms d'événement (comme nous l'entendons) des noms propres. Mais peut-on considérer l'existence de noms propres d'événement ?

1.1.2.2 Le nom propre d'événement ?

Nous introduisons ici un aperçu des travaux sur les noms propres de temps et la place qui pourrait être accordée aux dénominations d'événement dans la classe des noms propres. Rappelons la définition de (Gary-Prieur, 1994, p. 46–51) sur le contenu du nom propre :

> « le contenu d'un nom propre [est] un ensemble de propriétés du référent initial associé au nom propre qui interviennent dans l'interprétation de certains énoncés contenant ce nom [...] dans un univers de croyance »

Dans le cadre de l'événement, il est indéniable que la dénomination en discours est porteuse d'une charge sémantique, qui établit la nature de l'événement qui s'est produit. La nomination de l'événement permet son existence en discours, ce qui nous permet l'allusion à l'événement. Calabrese (2009) traite du processus de dénomination et des critères remplis qui font des désignants d'événements des noms propres.

Veniard (2009, p. 369–370) propose une typologie des noms d'événements sur la base d'étude de corpus en lien avec les événements « *guerre en Afghanistan* » et « *conflit des intermittents* ». Les critères pris en considération sont le figement et l'autonomie référentielle. Les trois types de noms d'événements sont :

- le désignant d'événement qui est transitoire et référentiellement non autonome « `la crise afghane` » ;
- la dénomination discursive d'événement qui est référentiellement autonome bien que non figée, par exemple « `la guerre en Afghanistan` » ou « `la crise (inextricable) des intermittents` » ;
- le nom propre d'événement qui est figé et autonome référentiellement. C'est le cas de « `la guerre d'Afghanistan` ». L'auteur se fonde notamment sur la présence dans des dictionnaires de langue de noms propres pour suggérer qu'il s'agit de syntagmes référentiellement autonomes.

van de Velde (2000) va plus loin et 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 des noms propres de lieux, et selon son raisonnement, tout porte à penser que les noms propres de temps devraient exister également. Ainsi, pour Velde, les années, les noms de mois ou de jours sont des noms propres parce qu'ils appuient la référence par un autre biais que la déixis (information contextuelle nécessaire à la compréhension d'un énoncé). Nous adhérons à l'idée des noms propres de temps qui désigneraient des événements et même à des noms propres de temps qui seraient l'équivalent des entités nommées « date ». Pourtant, nous ne pensons pas que les noms de mois et de jours, lorsqu'ils ne sont pas inclus dans une date complète (qui désigne un

jour particulier) puissent être considérés comme des noms propres de temps, comme c'est évoqué par Velde. Par ailleurs, les noms de mois ou de jour seuls ne peuvent être des dénominations d'événements, mais certaines dates le sont (sous la forme quantième mois « *21 avril* » ou mois année « *mai 68* », par exemple).

La perte de l'amorce de la désignation définie. Contrairement à ce que présente Veniard (2009) au sujet des noms de guerre, la construction d'un nom d'événement sur le modèle « terme classifieur + nom propre individualisant (la guerre de + nom de pays) » n'est pas obligatoire. Le terme classifieur pouvant être défini comme un terme amorce d'événement.

(Calabrese, 2009) introduit l'idée de contamination sémantique dans les désignants d'événement. Le terme classifieur événementiel apporte une valeur sémantique à la désignation nominale définie et ce contenu sémantique est conservé dans le substrat après simplification de la désignation en une dénomination plus courte : les « *élections du 21 avril* » devenues le « *21 avril* » ou « le *mouvement de Mai 68* » pour « *Mai 68* », mais aussi de « *catastrophe nucléaire de Tchernobyl* » à « *Tchernobyl* » seul, en passant par « *catastrophe de Tchernobyl* », la réduction en terme de détails et de longueur de syntagme de la dénomination de l'événement ne change rien à sa compréhension. On réfère toujours à l'événement, le nom étant connoté par l'image de l'événement qui s'est produit.

En effet, au départ, pour nommer un événement, nous formulons en discours une description définie, et au fur et à mesure s'opère une réduction ; nous désignons l'événement par une expression nominale réduite à quelques mots sans pour autant contenir encore de repère temporel et/ou local explicite. Ceci débouche finalement sur un nom évocateur de l'événement, de taille plus réduite et pourtant tout à fait explicite en contexte et dans une société donnée. D'ailleurs, nous avons travaillé sur une étude préliminaire de l'évolution des dénominations d'événements dans une approche en TAL dans (Tannier *et al.*, 2012), avec pour trame de travail un corpus de presse fondé sur les occurrences des désignations des événements « *crise grecque* », « *révolution arabe* », « *Wikileaks* », « *affaire Laetitia* », « *grippe H1N1* » et « *nuage islandais* ». L'idée est d'observer le passage de la désignation verbale à la désignation nominale. Nous souhaiterions par ailleurs pousser cette étude plus en avant dans l'observation du fonctionnement dans les textes du passage de la désignation définie nominale à un nom figé proche du nom propre.

Nous sommes devant le constat que par le biais de la métonymie, la richesse de la langue nous permet toujours de référer à l'événement, même lorsque le terme classifieur a disparu.

1.1.2.3 Des noms désignant d'autres entités devenus dénomination d'événements

Certaines dénominations d'événements sont chargés d'un autre sens avant de désigner un événement. Il s'agit notamment des noms de lieux et des dates qui en contexte peuvent désigner l'événement qui s'y est produit, comme on l'a vu précédemment avec « *Tchernobyl* » et « *Mai 68* ». On doit appréhender ces types de désignants-ci sous un angle différent des autres noms d'événements. Ces mots acquièrent leur sens évenementiel par le biais de la métonymie.

À côté de ces entités qui prennent un sens nouveau en contexte, il y a aussi ces mots étrangers qui entrent dans la langue pour désigner un événement particulier.

Voyons de plus près le fonctionnement des désignants de lieux et des désignants temporels en utilisation événementielle, ainsi que la voie par laquelle ces emprunts linguistiques entrent dans la nouvelle langue avec une charge sémantique événementielle extrêment forte.

a. Les désignants de lieux. (Lecolle, 2006) introduit la notion de « polyréférentialité du toponyme ». En effet, un toponyme peut généralement avoir plusieurs emplois, comme par exemple :

- celui de l'endroit,
- celui de l'endroit à titre d'institution ou de gouvernement dans le cas d'une capitale,
- celui de l'événement qui s'y tient,
- celui d'un ensemble d'individus, le nom de lieu employé comme un nom collectif pour désigner par exemple les habitants d'une ville ou d'un pays par l'utilisation du nom de lieu dans « `La France se mobilise contre la réforme` ».

L'auteur insiste sur l'importance du contexte dans cette nomination, le terme « polysémie » ne prenant pas en compte l'importance de la réalisation en contexte, alors que le toponyme est un « complexe sémantico-référentiel ». C'est le cas, par exemple, du toponyme « *Tchernobyl* » (Lecolle, 2004) qui désigne l'explosion du réacteur nucléaire de la centrale implantée près de la ville de Tchernobyl en 1986 dans l'exemple « `nuage de` *Tchernobyl* ». Le toponyme continue bien évidemment à désigner le lieu dans d'autres contextes, comme dans « `La population proche de` *Tchernobyl* `est évacuée à partir du lendemain de l'incident` », en tant que le nom donné à la centrale nucléaire par extension du nom de lieu où elle se trouve.

(Lecolle, 2004) s'intéresse aux emplois métonymiques des toponymes en particulier, et nous observons que ses conclusions concernant la polyréférentialité du toponyme sont

aussi valides pour les dates en emploi événementiel. Lecolle distingue la polyréférentialité externe et interne :

- La *polyréférentialité externe* désigne l'emploi d'un terme à une seule catégorie métonymique, mais peut avoir des interprétations différentes concurrentes.

  (1.1) `200 000 manifestants anti-mondialisation veulent paralyser *Washington.* (Le Monde 28/09/2002)`

  Dans cet example, l'auteur explique que hormis l'interprétation locative de « `Washington` », ce mot peut en plus désigner le lieu institutionnel, mais dans ce contexte-ci, l'interprétation pertinente est celle de l'événement « réunion des ministres des finances du G7 du 27 septembre 2002 qui a eu lieu à Washington ».
- La *polyréférentialité interne* est un mélange des sens événementiel et locatif ou temporel avec un autre sens métonymique du terme. Il s'agit d'un « épaississement du sens » et non plus d'une concurrence de sens comme dans la polyréférentialité externe.

  (1.2) `Comment *Cannes* constitue son affiche ?`

  Dans 1.2, le nom de ville « `Cannes` » peut référer aux acteurs de l'événement et à l'événement lui même.

Le processus de polyréférentialité interne peut aussi s'appliquer aux désignants temporels. Dans 1.3, `21 avril` réfère aussi bien à la date qu'à l'événement marquant qui s'est produit à cette date, soit l'accession au deuxième tour des élections présidentielles françaises de 2002 par le candidat d'extrême droite, événement vécu comme un réel séisme politique.

(1.3) `Comment le *21 avril* a-t-il changé la face politique française ?`

Cette observation permet de rapprocher le processus métonymique d'acception au sens événementiel des noms de lieux et des dates.

b. Les désignants temporels. Une distinction est faite parmi les désignants temporels. D'une part on évoque les noms de périodes historiques, appelés chrononymes et d'autre part les dates. Des travaux ont été menés dans ces deux voies, nous tâcherons d'éclaircir les points de vue et de donner notre avis.

b.1) Les chrononymes : des noms de périodes historiques

La définition des chrononymes dans (Bacot *et al.*, 2008) oppose les expressions calendaires/dates aux chrononymes. De plus, les auteurs insistent sur l'importance des événements et la valeur sociale dans la notion de chrononyme.

> « une expression, simple ou complexe, servant à désigner en propre une portion de temps que la communauté sociale appréhende, singularise, associe à des actes censés lui donner une cohérence, ce qui s'accompagne du besoin de la nommer. À côté des étiquettes strictement calendaires existe en effet tout un appareil de dénominations seul à même de permettre à une société de penser son histoire [. . .] »
> (Bacot *et al.*, 2008)

Du point de vue morphologique, les chrononymes sont des noms ou des syntagmes nominaux figés ou libres (« `la Belle époque` », « `les Années Mitterand` »). Certaines structures sont formées sur la base d'un « formant » exprimant une durée (« `les Années folles` », « `la décade infâme` »), d'autres indiquent une période référentiellement à un événement (« `l'Avant-guerre` ») ou encore des noms communs qui, au départ, ne possédaient pas la « notion de périodisation » (« `la Renaissance` »).

(Bacot *et al.*, 2008) font une différence entre les chrononymes et les noms d'événements, tout en admettant que :

> « cette affinité rend parfois difficile, par exemple, la distinction pourtant essentielle entre nom de période (*chrononyme*) et nom d'événement (*sorte de praxonyme*). »

Nous souhaitons répondre à ce raisonnement en prenant l'exemple de « `Mai 68` » qui illustre qu'une date peut correspondre à un nom de période et même être une dénomination d'événement. Puis, nous donnerons notre point de vue sur les autres noms de périodes historiques que nous ne jugeons pas être des dénominations événementielles.

La date « *Mai 68* » est un chrononyme dans le sens où ce syntagme de type date, désigne une période temporelle, qui plus est, associée à une image mémorielle importante pour la société française notamment. La simple évocation du groupe de mot « *Mai 68* » renvoie à de nombreux événements, comme des manifestations, grèves, mouvements sociaux, tensions politiques en France et dans le monde. Ce syntagme date évoque en réalité une série d'événements plus qu'une période historique, et fait appel à des images mémorielles ; à ce titre, nous le considérons comme une dénomination d'événement.

Par ailleurs, tous les chrononymes ne peuvent être considérés d'office comme des dénominations d'événements, comme par exemple (à notre avis) : les noms donnés aux périodes de temps font référence à des événements, comme « `post-moderne` » ou « `l'Après-Guerre` » ; les noms de périodes d'Histoire comme « `Renaissance` » ou « `Moyen-Âge` » lorsque ces périodes réfèrent essentiellement à une portion temporelle. Pourtant, en contexte, il est probable que ces noms puissent être assimilés à des

dénominations d'événements.

b.2) Les héméronymes : des dates qui font événement

Certaines dates désignent des événements en contexte. Ce phénomène a été étudié dans (Calabrese, 2008). Il s'agit des **héméronymes**. Toujours, ces expressions semblent être le diminutif d'une description définie plus longue. En effet, un terme classifieur événement a été effacé, en laissant sa charge événementielle sur la date restante.

La date « *11 septembre* » sert à nommer les attentats de New York. Cet exemple est particulier, parce que dans plusieurs langues, la date a servi à désigner le même événement, montrant ainsi l'impact mondial de l'événement par le biais de la nomination et plus particulièrement par l'utilisation du même procédé de nomination utilisé dans les différentes langues : « *11 septembre* » en français, « *September-11* », « *9/11* » dans le monde anglo-saxon, « *11-S* » en Espagne et en Amérique Latine ou encore « *11. September* » en allemand. Dans la dénomination d'événement « *11 septembre* », les mots « `attentat` » ou « `attaque` » ont été supprimés.

Toute date qui désigne un événement est héméronyme. Nous faisons la distinction ici entre les dates qui désignent un jour (« `21 avril` », « `11 septembre` ») et celles qui sont plus englobantes, moins précises comme « `mai 68` », chrononyme parce qu'il s'agit de l'indication temporelle sur un mois entier (une période temporelles) et non un jour en particulier.

c. Les xénismes. Des noms de lieux et des dates peuvent donc, par métonymie, se charger du sens de l'événement qui a eu lieu en cet endroit ou à cette date. C'est aussi le cas des xénismes, qui sont des noms communs dans la langue d'origine mais qui se comportent comme des noms propres dans la langue cible. Cette particularité est opérée de par l'opacité sémantique du mot étranger et de son association étroite avec son référent événement. Cette nomination des événements est souvent liée à un événement extrêmement marquant, comme « *Shoah* » ou « *apartheid* ». Ce mécanisme est aussi utilisé dans l'urgence de la situation, ou parce que l'événement est extrêmement soudain, c'était le cas de « *tsunami* » en décembre 2004, lors du raz de marée meurtrier qui a touché l'Asie du Sud.

Nous nous sommes interrogés sur la dénomination des événements telle que vue dans l'état de l'art sur le sujet. Nous avons observé que certaines dénominations d'événement ont des caractéristiques proches des noms propres, d'autres doivent être cantonnés

plutôt à la catégorie des désignations nominales définies. Pour notre part et dans nos travaux en cours, nous nous intéressons aux dénominations d'événements en général et pas uniquement à ceux qui seraient proches des noms propres standards. Mais si certaines dénominations d'événements sont à considérer comme des noms propres, nous ne pouvons nous empêcher de nous demander comment placer les noms d'événements dans le cadre des entités nommées.

1.1.3 Les entités nommées

Certains noms d'événements sont très figés, et ont un fonctionnement proche de celui des noms propres. Par ailleurs, les entités nommées ont servi, au départ, à relever des groupes de mots qui désignent une entité particulière. Le rapprochement entre nom propre et entités nommées est facile pour les noms de personnes, il est donc simple de s'intéresser aux entités nommées lorsqu'on évoque les noms d'événement et qu'en plus on arrive à les rapprocher des noms propres.

Par ailleurs, les entités nommées ne sont pas constituées que de noms propres, mais plutôt de noms indiquant des entités propres ou catégorisables. On y trouve le plus classiquement des noms de personnes, d'organisation et des dates. Dans cette aperçu rapide sur les entités nommées, nous nous rendons vite compte il n'y aurait aucune raison de ne pas considérer les dénominations d'événements comme une classe particulière d'entités nommées.

1.1.3.1 Définition

Nombre de travaux se sont intéressés à l'extraction des entités nommées (cf. (Ehrmann, 2008) pour un historique complet sur la reconnaissance des entités nommées), mais les définitions ne sont pas nombreuses ou très claires. Souvent, la définition des entités nommées est restreinte aux types d'entités qui sont recherchées.

Ehrmann (2008) rappelle que l'entité nommée n'est pas un objet linguistique, mais bel et bien un objet d'étude inventé de toutes pièces parce qu'il répond à un besoin en traitement automatique des langues. Ce besoin est donc directement lié au domaine de recherche qu'est l'extraction d'information et aux applications correspondantes. Ce constat conduit l'auteure à la définition suivante :

> « Étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. »
> (Ehrmann, 2008)

En ce qui concerne nos travaux, ils sont ancrés dans le projet *Quaero* [2] et plus précisément dans le groupe de travail sur les entités nommées. Cette partie du projet est dédiée à l'extraction d'information, aux entités nommées et aux relations existant entre ces entités, dans le but de construire une base de connaissances (Galibert *et al.*, 2010). L'une des volontés du projet est de référencer des noms propres aussi bien que des désignations autres qui réfèrent aussi à la même entité. Par exemple, il conviendrait dans la base de connaissance de pouvoir répertorier « Jacques Chirac » et « président de la République » comme étant des référents à la même personne. Ainsi, pour le besoin applicatif et pour l'exemple précédent, il conviendrait de pouvoir extraire comme entité nommée, les noms de personnes, les titres et fonctions, aussi bien que les métiers.

Dans ce cadre, la définition des entités nommées classiques a été redéfinie et étendue :

> « par l'extension des entités nommées à de nouveaux types (tels que les civilisations, les fonctions, etc.) ;
>
> et par l'extension de la définition des entités nommées à des expressions construites autour de noms communs ; nous autorisons donc l'inclusion de certaines expressions ne contenant aucun nom propre. »
>
> (Grouin *et al.*, 2011)

Dans cette définition étendue des entités nommées, l'un des nouveaux types d'entités nommées proposé est l'événement. Ainsi, les entités nommées événement sont tous les groupes nominaux désignant des événements. On peut ainsi envisager d'étiqueter en tant qu'événement les expressions « *11 septembre* » aussi bien que « *attaque du World Trade Center* » ou encore le simple nom commun « *attaque* » [3].

La figure 1.3 présente une vue générale de tous les nouveaux types proposés dans le cadre de l'extension de la définition des entités nommées et vise en particulier l'extension appliquée aux noms de personne.

L'objectif du projet *Quaero* étant la constitution d'une base de connaissances, les buts de ces travaux sont l'extraction d'information dans des données de type « nouvelles » (presse écrite ou audio transcrite) et la détection de relations entre les entités auxquelles on s'intéresse. Les entités nommées étendues définies ainsi constituent le pivot de cette base de connaissance.

1.1.3.2 Les entités nommées et les événements

Les campagnes de reconnaissance en entités nommées ont d'abord abordé la nécessité de reconnaître en contexte des données de même type. Au départ, l'intérêt est porté aux

2. http://www.quaero.org
3. pour plus de détails, cf. section 3.1 dédiée à notre guide d'annotation des événements

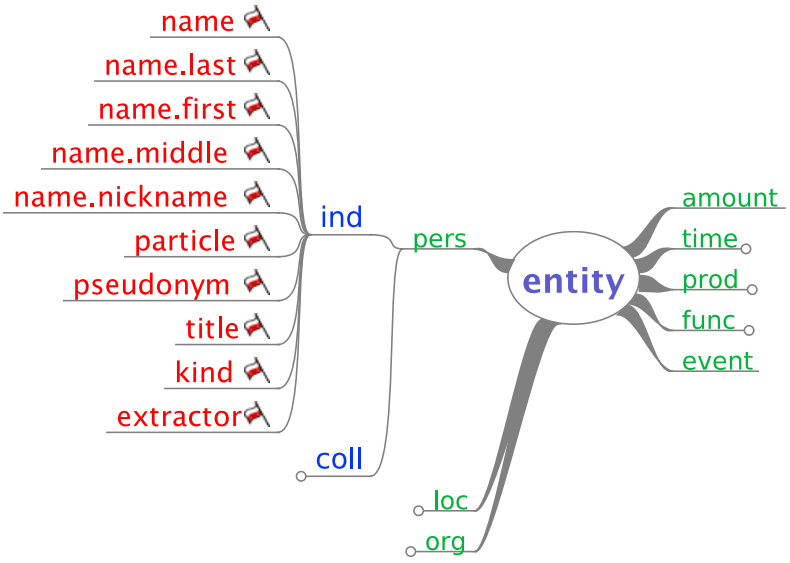


FIGURE 1.3 – Entités nommées étendues définies dans le projet *Quaero* : le cas des noms de personnes

noms de personnes, de lieu et d'organisation (cf. *MUC* (Grishman and Sundheim, 1996)) avant de s'étendre à d'autres types d'entités nommées, comme les montants ou les dates. Quelques campagnes de reconnaissance en entités nommées ont inséré les événements dans leur cadre d'exploration, c'est le cas de *ACE* ou d'*Ester*.

Le programme ***ACE* - Automatic Content Extraction** (Doddington *et al.*, 2004) est financé par l'armée des États Unis, et les corpus sont des comptes rendus militaires de conflits et d'exercices. *ACE* a pour but l'extraction de données précises sur des événements précis dans l'optique de remplir une base de données. Une présentation de l'événement de manière détaillée est proposé en section 1.2.1.4.

Comme pour toute entité nommée, les événements traités dans *ACE* sont les événements qu'on considèrere en fonction de l'utilité et du modèle applicatif qu'on a défini au préalable. C'est pourquoi les thèmes abordés sont limités en nombre, mais extrêmement précis en ce qui concerne les renseignements associés attendus. Il est totalement illusoire (travail de définition titanesque) d'attendre autant de détails pour de plus nombreux thèmes. Ce qui nous amène à nous demander si un tel degré de détail est important dans notre propre démarche.

En effet, *ACE* avait pour démarche une extraction d'information en profondeur, tandis que la nôtre serait plutôt dans la diversité des événements sans y rattacher au premier

abord des arguments comme les participants ou le véhicule utilisé lors d'une attaque terroriste. La seule information que nous envisageons de récupérer dans des perspectives de continuation de nos travuax serait liée au moment auquel se produit cet événement ou au lieu où il prend place. Il serait intéressant aussi d'établir les coréférences entre plusieurs mentions et désignations différentes d'un même événement.

***Ester* - Évaluation des systèmes de transcription enrichie d'émissions radiophoniques** en 2003-2005 constitue la première campagne sur les entités nommées en français. Dans cette campagne, il a été envisagé d'inclure les événements aux entités nommées (Meur *et al.*, 2004), mais par un manque de cohérence inter-annotateurs attesté, sans doute lié à une difficulté de définition et de balisage des noms d'événement, cette démarche a été abandonnée. D'ailleurs, les campagnes suivantes (*ESTER-2* et *ÉTAPE* - Évaluations en traitement automatique de la parole[4]) ne font pas référence aux événements.

Notons quand même que dans le cadre d'un travail préparatoire, deux types d'événements y étaient définis :

- ceux à caractère historique et unique

  (1.4) `les` *attentats du 11 septembre 2001*

  (1.5) `pendant la` *première guerre mondiale*

- ceux à caractère répétitif

  (1.6) `la` *Biennale d'art contemporain de Venise*

  (1.7) `le coup d'envoi des célébrations du cinquantième` *anniversaire de la déclaration des droits de l'homme*

Nous constatons que c'est une vision assez limitée des événements et cependant tout à fait intéressante, d'ailleurs nous retiendrons la distinction sur le caractère répétitif et unique des événements. Nous nous permettons tout de même de critiquer le dernier exemple où de notre point de vue l'un des événements auquel il est fait référence concerne la « *célébration* » de cet anniversaire. Dans l'annotation proposée par *Ester*, il n'est question que de l'« *anniversaire* ». La « *célébration* » est unique car elle a lieu à une date précise en un lieu précis, même si il est vrai que l'« *anniversaire de la déclaration des droits de l'homme* » est lui récurrent. Nous notons également qu'il était prévu que lors de l'annotation de l'événement « *attentat du 11 septembre 2001* », il soit permis l'imbrication des types d'entités, ainsi « *11 septembre 2001* » devait aussi être reconnu comme une date.

4. http://www.univ-paris3.fr/26065257/0/fiche___laboratoire/

Dans cette section d'état de l'art sur l'événement, nous avons constaté les problèmes de définition liés à la notion d'événement et avons observé les points de vue sur la nomination de l'événement.

Dans le chapitre 2, nous reviendrons sur de nombreux points abordés ici pour donner notre vision de la notion d'événement de manière détaillée. Nous avons aussi présenté la passerelle entre les dénominations d'événements et les noms propres et les entités nommées. Dans la section suivante, nous continuons sur ce tracé en présentant les ressources existantes en linguistique et traitement automatique des langues.

1.2 Représentations et ressources : Tour d'horizon de l'existant

Dans cette section, nous proposons un tour d'horizon de l'existant en terme de ressources. Nous présentons des représentations typologiques des événements et les principales représentations temporelles incluant les événements. Suite à ces travaux de représentations, des corpus ont été developpés pour les besoins d'évaluation de systèmes d'extraction notamment. Nous proposons une présentation de certains d'entre eux, avant de présenter un inventaire non exhaustif des autres ressources pouvant être utilisées dans une approche d'extraction automatique des événements nominaux.

1.2.1 Les représentations temporelles et des événements

Nous distinguons ici deux sortes de représentations : les typologies d'événements d'une part et les représentations temporelles qui prennent en compte les descriptions d'événements d'autre part.

1.2.1.1 Une typologie de l'événement en sciences sociales

En psychologie, Moles (1972) dans ses notes pour une typologie des événements appelle « événements » tous les types de variations perceptibles dans un environnement qui n'ont pas été prévues par l'occupant du centre de cet environnement. La notion est centrée autour du ressenti d'une personne et il maintient que par définition, les événements sont imprévisibles et imprévus. L'événement peut être considéré selon une série de dimension dont la plus importance est sa grandeur. Moles distingue plusieurs types d'événement suivant le critère de la grandeur (Moles, 1972, p. 91) :

- les **micro-événements**, qui parviennent à la conscience mais s'effacent dans la mémoire immédiate en suivant les lois de celle-ci ;
- les mini-événements, qui sont retenus pendant un délai variable mais toujours limité dans la durée de vie de l'être : un jour, un mois, un an ;
- les **événements proprement dits**, mémorisés par ceux mêmes qui y ont participé ou en ont été témoins ;
- les **grands événements**, historiques, qui sont inscrits dans des archives sociales de quelque espèce qu'elle soit, agence photographique, agence de nouvelles, journaux, livres d'histoire, et généralement datés dans une quelconque chronologie universelle

Dans la même lignée et concernant la grandeur, (Calabrese, 2010) cite aussi (Farge, 1997) en Histoire (événements tranquilles et événements « bouleversants » en plus des macro et micro-événements) et Oulif dans (Tudesq, 1973) (des événements « faibles » et des événements « forts »). Moles (1972) considère de plus comme caractéristiques de l'événement : son degré d'imprévisibilité, son taux d'implication, son caractère privé ou public, son intelligibilité ou encore son taux d'implication pour un certain nombre d'individus.

Ces hiérarchies des événements se veulent relatives à la société et ces critères sont fortement subjectifs. À cause de notre approche en traitement automatique des langues, nous ne les prendrons pas en compte dans notre analyse des événements, d'autant que nous n'adhérons pas au caractère imprévisible des événements, point à notre sens mal étayé par l'auteur.

1.2.1.2 Une typologie des toponymes événementiels

Dans ses travaux sur l'usage événementiel des toponymes, Lecolle (2005) propose une définition de l'événement (section 1.1.1) et une typologie. On peut étendre cette proposition de typologie à toutes les désignations nominales des événements. L'auteure distingue trois jeux d'oppositions pour faire valoir ses « catégories » d'événements :

- La première distinction concerne la prévision des événements : le « *Festival de Cannes* » a lieu de manière récurrente, il est **prévu**, tandis que « *Tchernobyl* » est un **imprévu**, qui est survenu et qu'on ne peut évoquer qu'au passé. Il est décrit comme soudain.
- La seconde touche au caractère **répétable ou non** d'un événement.
- La troisième se rapporte à la durée de celui-ci : l'utilisation de *pendant* comme introducteur des « *Jeux Olympiques* » permet d'après l'auteure de présenter le caractère **duratif** de cet événement, d'autant qu'on peut aussi dire que « *Cannes* » se déroule, par opposition « *Tchernobyl* » qui est un événement **ponctuel** (une fois qu'il a eu lieu il est déjà passé). L'auteure fait ici une distinction entre l'événement

que l'on pourrait rapprocher de la culmination au sens de Vendler (cf. (Vendler, 1959)) et les événements de type procès.

Cette typologie est proche de ce que nous aimerions faire. Nous adhérons aux idées de prévisibilité ou non, à la répétabilité et la durée possible des événements. Nous revenons d'ailleurs sur ces travaux lors de la présentation de notre typologie (chapitre 2.2).

1.2.1.3 Une typologie sémantique

Dans (Qian *et al.*, 2009), les auteurs proposent une méthode de désambiguïsation sémantique, dont un exemple d'application possible est la désambiguïsation des noms d'événements. Une typologie sémantique est présentée. On y distingue des événements selon trois jeux d'opposition : les événements **humains** et les événements avec **objets inanimés** ; des activités **individuelles** ou **sociales** ; des événements **qui durent** et d'autres **brefs** (cf. tableau 1.1 pour voir des exemples). Ce dernier type d'événement est le seul commun avec la typologie de Lecolle (2005), ce qui montre bien que l'ancrage de l'événement dans le temps est un point important dans sa définition.

human events :	adoption, arrival, birth, betrayal, death, development, disappearance, emancipation, funeral ...
events of inanimate objects :	collapse, construction, definition, destruction, identification, inception, movement, recreation, removal ...
individual activities :	birth, death, execution, funeral, promotion ...
social activities :	abolition, evolution, federation, fragmentation, invasion ...
lasting events :	campaign, development, growth, trial ...
brief events :	awakening, collapse, death, mention, onset, resignation, thunderstorm ...

TABLE 1.1 – Typologie sémantique proposée par (Qian *et al.*, 2009) : six sous-catégories de noms d'événement

Cette typologie est adaptée à la tâche des auteurs, mais pas à la nôtre. Cette typologie sémantique est comme son nom l'indique orientée sur une extraction d'information précise qui cherche à distinguer les événements en fonction de leurs acteurs. La typologie n'est pas dévolue à une extraction d'information plus temporelle, comme nous l'entendons.

1.2.1.4 *ACE* : une typologie détaillée et précise

Le programme *ACE* (Automatic Content Extraction), comme présenté précédemment a pour but l'extraction de données précises sur des événements précis dans l'optique de

remplir une base de données. Ce programme s'inscrit dans une approche d'extraction d'entités nommées. Ces travaux sont fondés sur une typologie thématique détaillée, précise et limitée en nombre de types pris en considération.

Dans *ACE* (Doddington *et al.*, 2004), on compte huit catégories : vie, transaction, affaires, conflit, contact, personnel, justice. Ces catégories sont divisées en sous-catégories, 33 au total. Ainsi, dans la catégorie « vie » se trouve la naissance, le mariage, le divorce, le fait d'être blessé et celui de mourir, ce qui permet de montrer les événements jugés importants dans la vie d'un individu. Seuls les événements qui sont des instances d'une des sous-catégories pré-déterminées sont annotés. À chaque événement est également associé un nombre défini d'arguments dépendant de son type. Pour tous les types d'événements sont extraits des déclencheurs (« event trigger »), correspondant au mot qui symbolise le mieux l'événement (verbe, adjectif, nom ou nominalisation), et un événement dit étendu (« event extent »), correspondant à la phrase dans laquelle l'événement est décrit. Aux événements sont associés des arguments, qui sont les participants, les attributs généraux lieux, date applicables à la plupart des évévénements ou spécifiques à la catégorie d'événement rencontrée (pour le type « transport », on spécifiera le point d'origine et la destination, ainsi que le véhicule par exemple).

En parallèle de *ACE*, Aone and Ramos-Santacruz (2000) proposent une ontologie des événements, reprise dans le tableau 1.2 (comprenant neuf catégories, soit 61 types d'événements, correspondant au nombre de sous-catégories). Les auteurs ont développé un outil d'extraction d'événements, dits à large échelle, pour l'extraction des événements et des relations qui les lient entre eux. Nous considérons que les travaux développés dans (Aone and Ramos-Santacruz, 2000) sont à rapprocher de la typologie *ACE*, même s'ils ne se fondent pas sur la même typologie, ils ont pourtant les mêmes objectifs. En effet, le type de fonctionnement demandé aux outils d'extraction est identique : extraire les événements en même temps que ses participants « who did what to whom when and where ? » (Qui a fait quoi à qui quand et où ?). Dans l'article, l'exemple illustratif de ces travaux est celui d'une transaction financière d'achat : l'outil développé doit extraire les éléments présents en contexte qui désignent l'acheteur, le produit, l'identité du vendeur, le moment et le lieu de l'événement décrit, « buying event ».

SemEval. Dans la continuation des campagnes *MUC* et *ACE* pour l'extraction des entités nommées, en ce qui concerne les événements, *SemEval-2010*[5] s'est intéressé dans le cadre de la tâche 11 aux événements dans le cadre d'une approche pour l'étiquetage de rôles sémantiques et la détection de verbes événementiels dans les *news* en chinois. De

5. http://semeval2.fbk.eu/semeval2.php

1. Vehicle
Vehicle departs
Vehicle arrives
Spacecraft launch
Vehicle crash

2. Personnal Change
Hire
Terminate contract
Promote
Succeed
Start office

3. Crime
Sexual assault
Steal money
Seize drug
Indict
Arrest
Try
Convict
Sentence
Jail

4. Family
Die
Marry

5. Political
Nominate
Appoint
Elect
Expel person
Reach agreement
Hold meeting
Impose embargo
Topple

6. Transaction
Buy artifact
Sell artifact
Import artifact
Export artifact
Give money

7. Business
Start business
Close business
Make artifact
Acquire company
Sell company
Sue organization
Merge company

8. Financial
Currency moves up
Currency moves down
Stock moves up
Stock moves down
Stock market moves up
Stock market moves down
Stock index moves up
Stock index moves down

9. Conflict
Kill
Injure
Hijack vehicle
Hold hostages
Attack target
Fire weapon
Weapon hit
Invade land
Move forces
Retreat
Surrender
Evacuate

TABLE 1.2 – Ontologie *ACE* des événements (Aone and Ramos-Santacruz, 2000)

plus, *TempEval-2* est la tâche 13 du défi SemEval-2010, dans la suite de *TempEval-1*[6] de 2007. Cette tâche s'organise en trois sous-tâches : une sur l'identification des évenements, une autre pour les expressions temporelles et la dernière sur l'identification des relations temporelles. Les événements sont principalement des événements verbaux et suivent la conception des événements *TimeML* (présenté ci-après).

6. http://www.timeml.org/tempeval/

1.2.1.5 TimeML : un langage de spécification pour une représentation temporelle des textes

ISO-TimeML (Pustejovsky *et al.*, 2010) est un langage de spécification pour l'annotation et la normalisation des événevements et des expressions temporelles sur des textes en langage naturel. Originellement, *TimeML* a été développé pour améliorer les performances de systèmes de questions-réponses. Le mot « Event » est dans *TimeML* :

> « a cover term for situations that happen or occur »
> (un terme générique pour désigner les situations qui se produisent)

L'événement est considéré pour sa soudaineté ou sa durée et peut être un état. Dans nos travaux, nous tenons compte des événements qu'ils soient passés, présents ou futurs et nous n'acceptons pas les états dans notre notion d'événement (cf. section 2.1.3). *TimeML* porte un intérêt évident aux verbes et formes prédicatives, alors que nous nous focalisons sur les nominalisations.

Le schéma d'annotation *TimeML* permet l'annotation des

- expressions d'événements (<EVENT>), divisées en 7 classes d'événements différents : ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE, PERCEPTION, REPORTING et STATE (Pour plus de détails, se référer à (Saurí *et al.*, 2005)),
- expressions temporelles et la normalisation de leur valeur (<TIMEX3>),
- relations temporelles qui existent les événements et les expressions temporelles susmentionnées (<TLINK>),
- relations aspectuelles (<ALINK>) et modales (<SLINK>) entre les événements,
- marqueurs linguistiques qui permettent ces relations (<SIGNAL>).

La représentation *TimeML* a d'abord été élaborée pour l'anglais et les schémas d'annotation de départ ont été modifiés en fonction des particularités des autres langues. En ce qui nous concerne, nous nous sommes intéressé à l'adaptation de TimeML au français (Bittar, 2010a), pour lequel un guide d'annotation a été produit [7]. En français, de nombreuses modifications ont été apportées. Par exemple, les modaux en français sont une classe à part et ne sauraient être annotés comme des verbes d'événements, une classe MODAL a donc été rajoutée, qui explicite les traits de temps, mode et polarité. Mais encore, un intérêt conséquent a été apporté à l'analyse des verbes supports, des verbes de cause ou encore l'utilisation du conditionnel pour le choix de la modalité (valeur ajoutée CONJECTURAL), entre autres.

7. (Bittar, 2010b), http://www.linguist.univ-paris-diderot.fr/~abittar/docs/FR-ISO-TimeML-Guidelines.pdf

Dans cette section, nous avons survolé un ensemble non-exhaustif de typologies qui couvrent des choses différentes et dans des buts différents, mais toujours pour une représentation des événements (comme ce qui se produit). Certaines prennent en compte les acteurs, d'autres s'intéressent à toutes les caractéristiques de l'événement et encore d'autres uniquement à l'ordre temporel de l'événement. Nous avons vu aussi que certaines ne s'intéressent pas qu'aux dénominations d'événement mais à toutes les occurrences dans les textes ou juste à l'événement dans l'absolu.

Pour notre part, nous adoptons une vision plutôt générique et moins sémantique que *ACE* ou Lecolle. Nous nous rapprochons plus de la vision *TimeML* des événements, mais par opposition nous n'acceptons pas les états, nous focalisons notre attention essentiellement sur les nominalisations et considérons des aspects liés à l'ancrage temporel des événements (chapitre 2).

1.2.2 Les corpus existants

Les définition et typologie linguistiques des événements n'amènent pas généralement à la création de corpus annotés, mais les représentations temporelles dévelopées dans le cadre du TAL nécessitent la création de telles ressources. En effet, ces représentations ont dès le départ pour but de permettre l'extraction d'informations réutilisables et catégorisées. Pour entraîner, développer et évaluer des systèmes d'extraction automatique, des corpus sont annotés en suivant les directives imposées par ces schémas d'annotation.

Dans cette section, nous mettons en avant quelques corpus en fonction de leurs langues et du schéma d'annotation suivi pour les élaborer. Un récapitulatif de ces ressources est proposé dans le tableau 1.3.

1.2.2.1 Les TimeBank

Jusqu'à *TimeBank1.2*. Le premier corpus annoté suivant le formalisme *TimeML* est *TimeBank*, qui a été régulièrement mis à jour. *TimeBank1.2*[8] (Pustejovsky *et al.*, 2006) est un corpus (généralement utilisé comme *gold standard*) d'environ 200 documents journalistiques : les textes annotés sont extraits de médias écrits comme le *Wall Street Journal*, ou issus de transcriptions de journaux télévisés provenant de *ABC*, *CNN*, *Voice Of America*. Ces documents sont datés entre 1989 et 1998. Le *TimeBank1.2* contient 7 935 annotations `<event>`, verbes et noms confondus, et 1 722 événements nominaux non étiquetés comme un état (les *non-stative nominal events*).

8. http://timeml.org/site/timebank/timebank.html

Corpus	Langue	nbre total d'événements	nbre noms d'événements	
Corpus annotés selon TimeML				
TimeBank 1.2	anglais	7 935	1 792	
FR-TimeBank	français	2 100	663	(Bittar, 2010a)
IT-TimeBank	italien	8 138	3 695	(Russo *et al.*, 2011)
	espagnol	1 677	?	(Wonserver *et al.*, 2012)
TimeBankPT	portugais	7 887	?	(Costa and Branco, 2012)
Ro-TimeBank	roumain	7 926	2 350	(Forascu and Tufis, 2012)
Corpus d'événements nominaux autre				
Creswell_2006	anglais	–	1 579	(Creswell *et al.*, 2006)

TABLE 1.3 – Comparaison des tailles de corpus disponibles en nombre d'événements en fonction du type de représentation temporelle dans les différentes langues. Le nombre de désignations nominales d'événement pour le *TimeBank1.2* correspond aux événements nominaux hors états (*non-stative nominal events*). Le corpus *TimeBank1.2* est disponible sur http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08. Le corpus *ACE 2005* est disponible sur http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T06.

Les corpus TempEval. Le corpus *TempEval-2* (Verhagen *et al.*, 2010)[9] a été créé pour la tâche *Tempeval-2* dans le cadre du défi *Semeval-2010*. Il s'agit d'un corpus manuellement annoté multilingue (chinois, anglais, français, italien, coréen et espagnol) et non parallèle. Le corpus est annoté en suivant les directives *TimeML*. De nombreux corpus *TimeBank* en langues différentes ont d'abord été créés pour *TempEval2010*, comme le corpus pour l'espagnol (auquel nous n'avons pas eu accès, mais dont le guide d'annotation est disponible (Saurí *et al.*, 2009)) ou l'italien (Russo *et al.*, 2011).

Les autres TimeBank. Le **FR-TimeBank**[10] est le corpus français annoté en TimeML, dont le guide d'annotation a été élaboré en suivant les grandes recommandations TimeML avec une adaptation à la langue pour le français. Ce corpus est un ensemble de 109 articles du journal *L'Est Républicain* sur la période 1999-2003. Comme noté précédemment, le formalisme d'annotation TimeML ne s'arrête pas uniquement aux événements nominaux, mais ce corpus en compte tout de même 663. L'**IT-TimeBank**, construit pour l'italien (Russo *et al.*, 2011) et totalement manuellement annoté, compte 3 695 événements nominaux annotés. Sur les autres TimeBank, nous n'avons pas les chiffres du nombre de noms d'événements, mais nous notons l'existence du **TimeBank espagnol** présenté dans (Wonserver *et al.*, 2012) et les TimeBank roumain (**Ro-**

9. http://www.timeml.org/tempeval2/
10. https://gforge.inria.fr/projects/fr-timebank/

TimeBank (Forascu and Tufis, 2012)) et portugais (**TimeBankPT** (Costa and Branco, 2012)), qui sont issus d'une traduction des textes du *TimeBank* anglais dans ces langues respectives.

1.2.2.2 Le corpus Creswell

Creswell *et al.* (2006) présente un corpus annoté en événements nominaux uniquement, par les auteurs et à des fins d'expérimentations que nous présenterons en section 1.3.2. Ce corpus a été développé dans le cadre d'un projet de recherche gouvernemental aux États-Unis et ne peut être distribué. Ce corpus est, d'après les auteurs, constitué de 77 documents de types *news* et issus d'archives en ligne. Il contient 9 381 groupes nominaux annotées, dont 1 579 décrivant un événement et 7 802 qui ne sont pas des événements. 2 319 instances sont différentes et 167 d'entres elles (17 %) sont ambiguës dans l'absolu (tantôt étiquetées événement tantôt pas). La question de l'ambiguïté de certains mots est clairement posée ici. Par exemple, c'est le cas de « `sermon` », « `behavior` » (comportement), « `deal` » (accord/vente), « `violation` » (transgression/infraction/viol).

Les principaux corpus existants sont des corpus *TimeML*, qui plus est seul un corpus est disponible pour le français et prend en compte les événements nomiaux (nos travaux concernent le français et dans une moindre mesure l'anglais). Même si *TimeML* n'est pas une spécification dédiée aux événements nominaux, ceux-ci sont présents et le corpus *FR-TimeBank* peut-être un bon point de comparaison avec nos travaux sur l'extraction des désignation nominales d'événements en français. Pour ce qui est de l'anglais, le corpus de Creswell n'étant pas disponible, seul le *TimeBank1.2* peut être pris en référence dans des évaluations de performances de travaux d'extraction automatique.

1.2.3 Les autres ressources

En plus des corpus, indispensables dans le cadre d'une approche dont le but est l'extraction automatique, nous avons besoin d'autres ressources linguistiques pour élaborer un tel projet. Nous proposons ici un parcours non exhaustif des ressources qui pourraient être utilisées pour l'extraction d'événements nominaux : thésaurus, dictionnaires sémantiques, ainsi qu'observations et/ou critiques sur le développement ou l'utilisation potentielle de ces ressources.

1.2.3.1 Les réseaux sémantiques de type *WordNet*

WordNet[11] est une ressource lexicale développée en anglais (Fellbaum, 1998) par l'Université de Princeton. *WordNet* est en réalité constitué de quatres réseaux sémantiques distincts, un pour chaque catégorie grammaticale de mots pleins : verbes, noms, adjectifs et adverbes. Chaque catégorie grammaticale a un système hiérarchique de classe sémantique à part entière et est structuré par des relations sémantiques. Les unités lexicales de catégories grammaticales différentes ne sont pas reliés. La synonymie et les relations d'hyperonymie et d'hyponymie sont prépondérantes dans le fonctionnement de *WordNet* pour les noms et les verbes, en particulier. Deux expressions font partie du même « synset » quand elles sont synonymes dans un contexte linguistique. En anglais, y sont représentés 117 000 synsets, qui constituent donc autant de relations conceptuelles.

EuroWordNet[12] est le projet ambitieux de créer des ressources lexicales sur le même schéma que *WordNet*, pour d'autres langues européennes (néerlandais, italien, espagnol, français, tchèque et estonien). Ces ressources sont structurées sur le même schéma que la ressource américaine *WordNet* pour l'anglais, avec des synsets et des relations entre eux. *EuroWordNet* se veut une ressource multilingue. Les *WordNet* des différentes langues sont interconnectés via un index inter-langue (basé sur le *Princeton WordNet*), qui permet pour un concept donné de passer d'une langue à l'autre.

L'élaboration de l'*EuroWordNet* du français a consisté en la traduction des ensembles synonymiques de *WordNet* avec une correspondance sur un dictionnaire francophone (cf. (Catherin, 1999)). Y sont représentés 22 745 synsets pour 18 777 entrées lexicales nominales ou verbales. Les critiques formulées dans (Jacquemin, 2003) mettent en avant le manque de données d'ordre syntaxique qui permettrait une désambiguïsation sémantique judicieuse, et « l'étroitesse du lexique couvert ». De plus, le réseau sémantique en français dans *EuroWordNet* est soumis à une licence propriétaire restrictive.

Wolf. Le *Wordnet Libre du Français*[13] (Sagot and Fišer, 2008) a été développé sur la base de *WordNet*. Contrairement à l'*EuroWordNet français*, *WOLF* contient tous les synsets du *Princetown WordNet*, même ceux qui n'ont pas de lexèmes répertoriés en français. Son avantage réside dans le fait qu'il est distribué sous licence libre. Mais cette ressource, bien que disponible, est encore en cours de développement, ce qui la rend difficilement utilisable.

11. http://wordnet.princeton.edu/
12. http://www.illc.uva.nl/EuroWordNet/
13. http://alpage.inria.fr/~sagot/wolf.html

1.2.3.2 Les déverbaux

De nombreux travaux ont traité de l'événementialité potentielle portée par les noms morphologiquement apparentés à des verbes d'action, communément appelés déverbaux (même si le terme n'est pas approprié eu égard au sens de conversion qui n'est pas toujours attesté : *nom* > *verbe* ou *verbe* > *nom*, pour plus d'informations se référer à (Tribout, 2010)). En effet, nous ne distinguons pas ici, les déverbaux (noms morphologiquement dérivés de verbes) et les noms « cousins » selon Meyers (Meyers, 2007) (le nom cousin donne naissance à un verbe, comme « revolución » > « revolucionar ») ou sémantiquement lié à un verbe (« genocidio » (génocide) est lié à « exterminar »(exterminer)). Ces noms seraient porteurs pour bon nombre d'entre eux de la dénotation d'événements. Par exemple, « *fête* » porte la notion d'activité de son verbe morphologiquement apparenté « *fêter* ». Par contre, apparenté à « *rapporter* » qui correspond à une action, le mot « *rapport* » peut désigner l'événement/moment où un rapport est formulé (l'action de rapporter), mais peut aussi désigner l'objet rapport correspondant au document contenant les informations dudit rapport. C'est aussi le cas de « *construction* », dérivé de « *construire* »), qui peut évoquer l'action de construire ou encore le résultat de l'action constitué par l'objet construit.

Ce constat est partagé et étayé dans les travaux de Huyghe et Marín ((Huyghe, 2006) et (Huyghe and Marín, 2007)). Ils se sont intéressés à l'héritage aspectuel des noms déverbaux en français et en espagnol ou à comment expliquer la dénotation d'événements par des noms. Parmi les noms dynamiques (dérivés de verbes d'achèvement « *explosion* » ou d'accomplissement « *accouchement* »), certains expriment une « délimitation temporelle » qui leur permet de référer à un événement. La délimitation temporelle y est présupposée de par la nature du procès qui décrit un point culminant et un changement d'état. Ces noms peuvent apparaître en sujet de « `avoir lieu` » et en complément dans « `le lieu du` ». Les noms dérivés de verbes d'activité pour leur part peuvent être de deux types :

- les noms massifs comme « `natation` » sont non-événement ,
- les noms comptables comme « *manifestation* » sont événement.

On parle ici de noms qui sont porteurs d'une lecture événementielle par essence, mais ceux qui désignent un événement en contexte sont exclus de cette étude.

Le projet *Nomage* (Balvet *et al.*, 2011) avait pour but l'étude sémantique en corpus des nominalisations déverbales, ainsi que l'évaluation de l'héritage aspectuel entre verbes et noms. À partir du corpus *French TreeBank*[14] (Abeillé *et al.*, 2003), les nominalisations dérivées suffixalement de verbes ont été extraites du corpus et ont fait l'objet d'une double

14. http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

annotation :

1. annotation aspectuelle des noms et des verbes de base sur des exemples construits en appliquant des batteries de tests. Les étiquettes utilisées sont fondées sur les classes de Vendler en grande partie (cf. figure 1.4). Selon les auteurs, les noms ayant reçu les étiquettes ACH (achèvement), ACC (accomplissement), ACT (activité), ACH (achèvements suivis d'un état), ACC (accomplissements suivi d'un état) sont bien des événements.
2. annotation des noms en corpus (application d'un batterie de dix tests par des annotateurs « naïfs »). À partir des réponses obtenues à ces tests, des patrons permettant l'attribution automatique d'étiquettes sémantico-aspectuelles ont été appliqués. Cette annotation est beaucoup moins fine que la précédente, les classes sont : ETAT, OBJET, EVENEMENT.

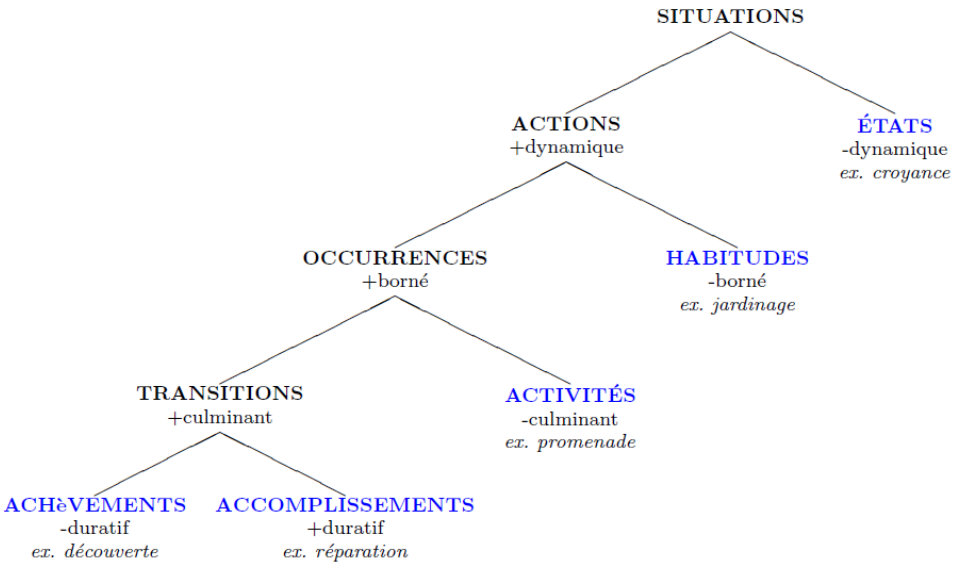


FIG. 2 – Hiérarchie des classes aspectuelles simples pour les noms

FIGURE 1.4 – Schéma de la hiérarchie des classes aspectuelles des noms dans le projet Nomage

Le projet *Nomage*[15] a donné lieu au développement de ressources (qui ne sont pas encore disponibles) : un corpus de phrases où chaque nom déverbal est annoté sémantiquement et syntaxiquement et un lexique de noms déverbaux avec les lexèmes instanciés et le verbe source de chaque nom décrit. Le lexique proposé au format XML contient une première partie destinée aux noms et la deuxième aux verbes. Pour chaque nom et verbe, des indications syntaxiques de leur utilisation et sémantiques sont répertoriées de manière

15. http://sites.google.com/site/nomagesite/

très complète, mais les noms et les verbes dont ils sont issus ne sont pas même liés par leurs sens communs. Ce lexique contient 656 noms déverbaux. Nous n'entrerons pas dans les détails sur ce lexique, car nous avons décidé de ne pas l'utiliser. En effet, nous ne travaillons pas en particulier sur les déverbaux et nous ne nous focalisons pas sur une définition stricte des déverbaux. Ce qui nous intéresse ce sont les noms qui sont apparentés à des verbes d'action (peu importe le sens de conversion *nom* > *verbe* ou *verbe* > *nom*). Nous avons préféré travailler avec le *VerbAction*.

***VerbAction* : le lexique des déverbaux du français.** En français, le lexique *VerbAction* (Tanguy and Hathout, 2002) est constitué d'une liste de verbes d'action accompagnée des noms déverbaux morphologiquement apparentés à ceux-ci (9 393 couples verbe-nom, soit 9 200 lemmes nominaux uniques). Les verbes d'action impliquant que quelque chose se produise (« `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* »). Par ailleurs, ces mots peuvent donc être ambigus, principalement parce qu'ils servent aussi à référer au résultat des actions désignées par ces verbes (« *aération* », « *étalage* »).

1.2.3.3 Des lexiques d'amorces pour le repérage des noms d'événements

Certains mots ont une valeur événementielle indéniable. Ils sont intrinsèquement événementiels et peuvent permettre un repérage à coup sûr d'un nom d'événement (comme « *conflit* », « *fusillade* », « *détonation* », « *cyclone* »). Ces mots amorces de noms d'événements peuvent aussi servir de base à l'apprentissage du repérage de noms d'événements construits à partir de mots qui peuvent désigner un événement en contexte particulier parce qu'ils ne sont pas porteurs uniquement d'une valeur événementielle (comme « `construction` » qui désigne l'action de construire ou le résultat de cette action) ou parce qu'ils ne sont pas voués au départ à indiquer un événement (comme « `intermittents` » en lien avec le conflit des intermittents du spectacle en 2003, « `Outreau` », « `Mai 68` »).

En anglais, deux des auteurs dont les travaux nous intéressent ont utilisé des listes d'amorces ((Creswell *et al.*, 2006) et (Qian *et al.*, 2009)). Ces deux listes ne sont pas fournies de manière complète dans les articles, mais des exemples sont proposés dans le tableau 1.4. Creswell utilise une liste de 95 termes amorces et 295 noms non-événement. Les amorces de Qian consistent en une liste de 15 noms événementiels et 215 noms non-événement. Ces listes servent d'amorce pour un apprentissage en *bootstrapping* de lexique.

Dans les travaux sur l'espagnol et l'anglais de Bel et Resnik, des termes amorces sont aussi utilisés. Les auteurs fournissent dans leurs articles les listes complètes d'amorces :

Amorces	événements	non événements
(Creswell *et al.*, 2006)	`election` (élection) `war` (guerre) `assassination` (assassinat) `dismissal` (démission)	`corpse` (cadavre) `electronics` (électronique) `bureaucracy` (bureaucratie) `airport` (aéroport) `cattle` (bétail)
(Qian *et al.*, 2009)	`appeasement` (apaisement) `happening` (apparition) `crash` (effondrement, crash) `loss` (perte)	`anything` (rien) `ambition` (ambition) `diagonal` (diagonale)

TABLE 1.4 – Exemples de mots amorces employés dans les méthodes d'extraction de noms d'événements de (Creswell *et al.*, 2006) et (Qian *et al.*, 2009) dans leurs travaux sur l'anglais.

dans (Resnik and Bel, 2009) pour l'espagnol (en annexe C.1.2), et dans (Bel *et al.*, 2010) pour l'anglais (en annexe C.1.3).

En français, le lexique *EventNominals*, soit le lexique alternatif des noms événementiels[16] se présente comme complémentaire au *VerbAction* dans la recherche de noms d'événements. Il contient 804 noms (en annexe C.1.1) qui ne sont pas des déverbaux, comme « *anniversaire* » ou « *grève* ». Ces mots ont au moins une lecture événementielle dans le corpus étudié par leur auteur. Comme les déverbaux, certains mots sont ambigus : ils peuvent désigner l'événement ou le procès aussi bien que le résultat ou l'objet de celui-ci, c'est le cas de« *apéro* » et « *feu* ». De plus, certains de ces noms présentent un état (« `absence` »), or les états ne font pas partie de notre définition de l'événement en tant que tel, nous envisageons l'événement sous le biais du changement d'état. Enfin, de nombreux noms appartiennent à des registres de langue spécifiques (« *anticoagulothérapie* »). Ce lexique a été utilisé comme indice pour l'annotation en *TimeML* du corpus *FR-TimeBank* (Bittar, 2010a).

Nous avons fait le tour non-exhaustif des représentations et typologies considérant les événements et des ressources lexicales disponibles. Les représentations et typologies sont toutes intéressantes et relèvent chacune de points de vues particuliers en fonction d'une tâche définie pas toujours compatible avec la nôtre, mais sont des points de départ ou de débat pour la caractérisation de l'objet d'étude de ce manuscrit. Parmi les ressources

16. Nous remercions André Bittar d'avoir mis à notre disposition son lexique complémentaire de noms d'événements.

disponibles, nous avons pointé celles qui nous semblent utilisables en pratique dans nos travaux et dont nous reparlerons plus tard dans les chapitres de mise en œuvre (chapitre 4 et chapitre 5).

1.3 Extraction Automatique

L'extraction automatique des noms d'événement en français est encore peu étudiée. Ce domaine de recherche est nouveau. Pour d'autres langues, quelques recherches ont été consacrées à cette thématique. Nous évoquerons entre autres des travaux qui se sont focalisés sur les moyens de repérer la lecture événementielle des noms déverbaux en langues romanes.

Nous commençons par un aperçu des approches fondées sur les représentations temporelles et des événements (*ACE* et *TimeML*), avant de montrer des travaux d'extraction des événements nominaux en particulier et qui ne s'inscrivent pas dans des représentations temporelles connues, puis nous concluons sur des approches plus lexicales et linguistiques.

1.3.1 Extraction d'événements

Dans cette section nous abordons uniquement quelques travaux, que nous jugeons intéressants du point de vue de notre approche et qui ont pour objectif l'extraction automatique d'informations concernant les événements, suivant les représentations *ACE* et *TimeML*.

1.3.1.1 Les événements de type *ACE*

Comme nous l'avons présenté en section 1.2.1.4, l'objectif des travaux sur les événements *ACE* est de récupérer les informations sur des événements précis. Par opposition, notre approche consiste d'une part à extraire les noms seulement, et nous nous intéressons au procédé de la nomination bien plus qu'aux circonstances et aux participants des événements. D'autre part, nous ne souhaitons pas nous restreindre à un domaine particulier. Néanmoins, certaines approches peuvent présenter des similitudes avec la nôtre.

Aone and Ramos-Santacruz (2000) s'intéressent aux relations et aux événements, tels que décrits dans *MUC-7* (Chinchor, 1998). Un peu comme dans *ACE*, l'idée est de remplir une grille d'informations concernant un événement. Pour chaque événement sont recher-

chées toutes les informations le concernant : pour l'événement « buying », l'outil *REES* doit extraire les données disponibles dans le texte sur l'acheteur, ce qui fait l'objet de l'achat, le vendeur, le moment de la vente et lieu de celle-ci. L'approche est fondée sur un étiquetage à base de patrons en contexte, générés de manière non automatique.

Dans (Ahn, 2006), l'approche présentée se veut modulaire pour l'extraction d'événements de type *ACE*. En effet, un module a été développé pour chacune des quatre sous-tâches définies : identifier des déclencheurs, ancrage des événements (« event anchors ») ; identifier ses arguments ; lui affecter ses attributs ; déterminer les coréférences des événements. Cette approche est à base d'apprentissage, et chaque module correspond à un classifieur automatique, généré avec les outils *TiMBL*[17], une technique fondée sur les plus proches voisins (Daelemans *et al.*, 2004), et *MegaM*[18], un type de modèle d'apprentissage à base d'entropie maximale (Daumé III, 2004).

Dans notre approche de l'extraction des événements nous nous intéressons au premier module concernant l'identification des ancres événement. Les événements recherchés sont des groupes nominaux, verbaux, adjectivaux, adverbiaux, pronominaux, sur la base de déterminants ou de prépositions. L'identification des ancres événement est vue comme une tâche de classification de chaque mot des documents. Les indices considérés sont notamment :

- des traits lexicaux (forme, lemme, profondeur dans l'arbre de dépendance),
- les traits issus de *WordNet* portés par le mot cible,
- les catégories morpho-syntaxiques du mot et des mots à gauche et à droite du mot événement potentiel,
- la dépendance du mot dans un groupe ou s'il est la tête de syntagme.

Même si en français nous n'avons pas accès à une ressource comme *WordNet* (cf. section 1.2.3.1), il nous est possible de considérer la plupart des traits d'apprentissage utilisés dans (Ahn, 2006) dans leur module d'identification des ancres événement pour notre extraction des désignations nominales des événements.

1.3.1.2 Les événements de type *TimeML*

Le deuxième type de représentation temporelle et des événements présenté précédemment est celui selon *TimeML* (section 1.2.1.5). Les travaux qui ont principalement retenu notre attention sont ceux de Saurí *et al.* (2005), Bethard and Martin (2006) sur l'anglais et Parent *et al.* (2008) sur le français.

17. http://ilk.uvt.nl/timbl/
18. http://www.isi.edu/~hdaume/megam/

Le système *Evita* (Saurí *et al.*, 2005) a pour objectif l'extraction des événements de type *TimeML* (toute situation dynamique ponctuelle ou durative, qui a lieu, mais aussi les états quand ils sont temporellement ancrés dans les textes analysés) en anglais. L'annotation *TimeML* tenant compte de tous les événements, les événements nominaux sont aussi l'objet de ce système. Ce système d'extraction combine des techniques linguistiques et statistiques dans le but de remplir au mieux chacune des sous-tâches de la reconnaissance des événements.

Les textes font l'objet d'analyses morphosyntaxiques. L'intérêt de la démarche revendiquée par les auteurs est qu'elle n'est fondée sur aucune liste pré-établie de patron d'extraction. Dans (Saurí *et al.*, 2005), un intérêt particulier est porté à la désambiguisation des noms pouvant avoir ou non une interprétation événementielle. Dans cette tâche, ces travaux sont fondés sur un module statistique avec consultation des informations contenues dans *WordNet*[19]. En effet, sur la base des événements présents dans *SemCor* et *TimeBank1.2*, les auteurs ont identifié 25 sous-arbres du réseau sémantique qui permettrait de retrouver les noms à caractère potentiellement événementiel. Ils ont ausssi identifié les cas de synsets qui annonceraient qu'un mot n'est pas un événement, comme par exemple les mots issus du synset « PHENOMENON » (phénomène) ou du synset « CLOUD » (nuage). Dans le cas où la recherche dans *WordNet* ne donne rien, alors un classifieur automatique de type bayésien fondé sur des règles, appris sur le corpus *SemCor*, est utilisé. Les résultats sont évalués par comparaison avec les étiquettes présentes dans *TimeBank1.2*.

La question de la désambiguïsation entre interprétation événementielle et non événementielle des noms est clairement posée. De plus, les désignants nominaux d'événement considérés sont des groupes nominaux, et pas seulement le nom seul. Cette approche nous intéresse, même si elle est peu reproductible en français, étant donné qu'une ressource telle que *WordNet* n'est pas vraiment disponible dans notre langue (cf. section 1.2.3.1).

(Bethard and Martin, 2006) reprend notamment les travaux menés avec *Evita* pour servir de référence à leurs expérimentations de classification des événements *TimeML*. Cette approche est linguistiquement motivée et utilise l'apprentissage automatique. Les traits d'apprentissage considérés sont le mot, son lemme, ses affixes (3 premières lettres et 3 dernières lettres du mot), le verbe racine du nom, la catégorie morpho-syntaxique, la valeur du label *B-I-O*[20] du mot (Ramshaw and Marcus, 1995), le type de déterminant, si le mot fait partie d'une construction verbale légère (« *light verbs* »), les qualités tem-

19. http://wordnet.princeton.edu/

20. Dans le système d'étiquetage B-I-O : B indique le début d'une étiquette, I que le mot appartient à une étiquette et O que le mot est en dehors de toute étiquette. B_I_ACTION indique débute l'étiquette I_ACTION

porelles du mot (par exemple pour les adverbes modifieurs). Des expérimentations ont été menés par les auteurs pour sélectionner les traits les plus appropriés à la tâche, mais les traits retenus ne sont pas mentionnés. Le classifieur automatique du système *STEP* est développé sur 90 % du corpus *TimeBank* et testé sur les 10 % restants. Dans le sous-corpus de test, seul 28% des `<event>` annotés sont des noms, ce qui d'après les auteurs pourrait expliquer les performances faibles de leur système sur les noms par rapport aux performances sur les verbes.

Ces travaux se sont particulièrement intéressés aux événements nominaux en plus de toutes les autres étiquettes *TimeML*. Et les auteurs sont arrivés à la conclusion que le problème de la détexction des événements nominaux n'est pas trivial, même si leur conclusion est qu'il s'agit d'un manque sur la quantité de données, de noms annotés `<event>`.

Les seuls travaux connus en français sur une extraction automatique des événements de type *TimeML* sont ceux de Parent *et al.* (2008). Ils ont manuellement annoté des textes en suivant les indications du guide d'annotation *TimeML* pour l'anglais [21] Leur modèle est principalement fondé sur l'analyse syntaxique et des *patrons-action* (des segments d'arbre syntaxiques qui permettent la classification des mots du texte en *TimeML*). Ces travaux concernent en priorité les expressions adverbiales de localsation temporelle, ainsi que les verbes, les adjectifs et les noms. En ce qui concerne les événements nominaux, les auteurs s'appuient sur les mots du *VerbAction* (section 1.2.3.2) dont ils ont « éliminé les items indésirables », sur des règles syntaxiques pour annoter les noms qui dépendent d'une préposition temporelle annotée en `<signal>` dans *TimeML*, et un filtrage des noms potentiellement événements qui sont complément du nom.

En plus de ces travaux qui ont plus ou moins traité les noms parmi tous les autres types d'événements *ACE* ou *TimeML*, d'autres se sont focalisés sur l'extraction automatique des dénominations d'événement.

1.3.2 Extraction d'événements nominaux

En parallèle des travaux en lien avec les représentations *ACE* et *TimeML*, des travaux en TAL se sont orientés spécifiquement vers l'extraction automatique des événements dans leur forme nominale.

21. Au moment de ces travaux, le guide d'annotation *TimeML*(Bittar, 2010b) pour le français n'était en effet pas encore disponible.

Creswell *et al.* (2006) s'intéressent uniquement à l'extraction de syntagmes nominaux décrivant des événements, dans un corpus de type journélistique en anglais. Contrairement aux autres approches, le but de départ est bien l'extraction automatique des événements sous leur forme nominale. La technique combine une acquisition automatique de lexique et une désambiguïsation sémantique au moyen d'un algorithme d'apprentissage supervisé.

Un corpus journalistique a été spécialement annoté et un lexique dit de termes amorces (*seed terms*) représentant des noms d'événements et de non-événements a été constitué. Les 95 termes événements ont été trouvés sur la base d'intuition des auteurs avec une vérification dans *WordNet*. Les 295 termes non-événements ont été extraits sur la base de mots figurant dans *WordNet* sous les hyperonymes choisis suivants « GROUP », « PSYCHOLOGICAL », « ENTITY », « POSSESSION » et qui sont présents plus de 800 fois dans le *British National Corpus*. Des exemples de ces mots amorces ont été présentés précédemment dans le tableau 1.4.

L'analyseur syntaxique *Semantex*[22] permet d'extraire des triplets de dépendances utilisés dans une approche, dite de *bootstrapping* (apprentissage en boucle, où la phase d'apprentissage suivante utilise les résultats de la précédente passe). Le corpus d'apprentissage de départ est constitué de 156 000 documents journalistiques ($\simeq$ 100 millions de mots) issues du *Foreign Broadcast Information Service* et d'archives de textes en ligne. Sur la base des triplets de dépendance (w_i, R, w_j), sont extraits du corpus les 48 353 patrons de la forme $(w_i, R, *)$ et $(*, R, w_j)$ présents plus de 300 fois dans le corpus. Les listes servent de base pour trouver d'autres mots dans la même configuration que ceux-ci et leur attribuer la catégorie correspondante, dans le corpus d'apprentissage.

Dans cet article les auteurs évoquent les noms d'événements comme des syntagmes nominaux. La question des frontières des désignations nominales a donc sans doute été posée, pourtant elle n'est pas abordée. De plus, aucune définition de l'événement n'est avancée. Néanmoins, comme dans nos travaux (cf. section 2.1.3), les états ne sont pas retenus dans la catégorie des noms événements.

Dans (Srihari and Novischi, 2008), les auteurs s'inspirent des travaux menés par Creswell *et al.* (2006). Dans un registre différent, Srihari et Novischi s'intéressent au repérage de l'événement majeur de clips vidéos. Les noms d'événements à extraire sont issus d'un corpus de transcriptions écrites de la piste audio des vidéos du Foreign Broadcast Information Service. Le but est de trouver l'événement principal évoqué. Ces travaux sont menés sur l'anglais et utilisent des déclencheurs linguistiques dans un modèle probabiliste « multinominal model ». Leurs travaux sont fondés sur le système d'extraction d'infor-

22. Janya, Inc's information extraction application

mation Semantex[23] qui extrait des entités, événements, relations et attributs, et fait de l'analyse de sentiments sur des types d'événements précis, prédéterminés : « `marriage` », « `divorce` » et « `stealing` ». Les principaux déclencheurs linguistiques utilisés sont :

- la modalité : un événement est réel ou irréel/incertain, les indices sont les verbes modaux (« `may` », « `might` » (pourrait, serait)), les expressions modales (« `it is possible` » (il est possible que)) et les adjectifs d'intention (faux, possible)
- les relations temporelles entre événements portées par les prépositions « `before` » (avant) et « `after` » (après).

Les mots sont classés dans les catégories `<event>` (est un événement) et `<non-event>` (n'est pas un événement), un même mot pouvant être dans l'une ou l'autre des catégories dépendant de leur contexte d'apparition. L'intérêt est porté sur les traits du mot qui ont été donnés par Semantex, et donne les informations sur comment le mot est utilisé. Des vecteurs sont calculés pour les mots de ces listes et par un calcul, dont le résultat est positif ou négatif, on obtient la classification de la nouvelle instance.

1.3.3 Approches plus linguistiques de l'extraction automatique d'événements nominaux

Sur la piste de travaux sur l'extraction automatique de désignations nominales d'événements, on trouve des travaux plus orientés vers la linguistique et la tentative d'expliquer des phénomènes sémantiques. Notons les travaux sur la valeur événementielle des déverbaux, comme ceux de Russo *et al.* (2011) sur l'italien ou de Peris *et al.* (2010a) et Bel *et al.* (2010) sur l'espagnol et l'anglais ou les travaux de Eberle *et al.* (2009) sur la désambiguïsation des noms en « `ung` » en allemand. Par ailleurs, les travaux de Qian *et al.* (2009) en désambiguïsation sémantique ont abordé la désambiguïsation événementielle.

1.3.3.1 Les nominalisations déverbales et les événements

Le constat de départ des travaux suivants porte en partie sur la valeur événementielle des noms déverbaux. Les noms déverbaux issus de verbes d'action peuvent indiquer entre autre chose une action, un événement, le résultat de l'action, l'instrument de l'action, sachant qu'un même mot peut tantôt indiquer un ou un autre sens en fonction du contexte.

Peris *et al.* (2010a) se sont intéressés à l'identification de la dénotation sémantique des déverbaux espagnols par le biais de l'outil ADN-Classifier. Il s'agit d'une tâche de classification des déverbaux en trois classes : « résultat », « événement » et « nom sous-spécifié », avec une information sur l'appartenance du nom à une structure lexicalisée.

23. Janya, Inc's information extraction application

Plusieurs lexiques, ainsi que des traits extraits de manière automatique ou manuelle sont évalués dans un modèle d'apprentissage automatique.

Dans (Peris *et al.*, 2010b) sont présentés les critères utilisés pour annoter manuellement le corpus espagnol *AnCora-Es* les valeurs dénotatives des noms (résultat, événement et noms sous-spécifié) et la struture argumentative des nominalisations déverbales. C'est en contre-pied de ce qui est d'habitude avancé dans la litterature ((Grimshaw, 1990), (Picallo, 1999), (Alexiadou, 2001)), que (Peris *et al.*, 2010b) signifie qu'on ne peut uniquement borner les nominalisations à l'action et au résultat. Pour autant, les critères pris en compte sont ceux communément acceptés et sont d'ordre morphologiques, syntaxiques et sémantiques :

- lorsque le mot cible est dans un syntagme et qu'il lui est rattaché un argument (ARG), alors il s'agit du résultat (result) :

  (1.8) L'*invention*$_{(result)}$ de Juan$_{(ARG\ de\ invention)}$

- la forme plurielle est réservée aux résultats, par exemple : « *perte* » et « pertes ».
- les articles définis et possessifs pour les événements (exemple 1.9) et les indéfinis, démonstratifs et les nombres pour les désignants non événement (exemple 1.10).

  (1.9) la *culmination*$_{(event)}$ *d'un procès de destruction*

  (1.10) une augmentation$_{(result)}$ de 20,47%

- la complémentation par des adjectifs relationnels pour les noms de résultat et les compléments temporels doivent être introduits par « de » (exemple 1.11), alors que pour les événements la préposition n'est pas nécessaire (exemple 1.12).

  (1.11) une négociation$_{(result)}$ de trois heures

  (1.12) depuis sa *construction*$_{(event)}$ *en 1989*

- la classe sémantique du verbe dont est issue la nominalisation.
- d'autres indices, appelés sélecteurs, comme la préposition « durante » (pendant).

Dans la suite de Peris *et al.* (2010a), Russo *et al.* (2011) ont fondé leur étude sur les indices syntagmatiques avancés dans la littérature linguistique comme étant des indices forts de la valeur événementielle des déverbaux, indices sur lesquels les différents spécialistes ne sont pas toujours d'accord. Ces travaux compilent les traits spécifiques de l'identitification des deux lectures possibles de ces noms déverbaux présents dans (Grimshaw, 1990), (Alexiadou, 2001), (Pustejovsky, 1995) afin de les analyser de manière objective et scientifique et d'observer le comportement de ces mots, en italien particulièrement. C'est un point de vue critique sur les critères communément acceptés. Il s'agit d'une vérification de la réelle importance de ces critères fondée sur une étude de corpus. Les paramètres observés sont :

- L'obligation de réalisation de la structure argumentative avec un groupe prépositionnel pour une événement. Contrairement à l'intuition des linguistes, Russo montre que les modifieurs possessifs ont plutôt tendance à déterminer la non-événementialité d'un nom.
- L'idée selon laquelle les noms au pluriel représentent moins souvent des événements est fausse.
- De même, l'étude invalide l'hypothèse énonçant que si l'article qui détermine le déverbal est défini, alors c'est un nom en lecture évènementielle.
- Les modifieurs aspectuels : (Russo *et al.*, 2011) releve 53 adjectifs très fréquents en co-occurrence avec des noms en lecture événementielle et 41 verbes réputés bons indices contextuels pour l'identification de l'événementialité d'un nom. Ces listes ne sont pas divulguées.

Dans une approche d'extraction automatique des noms d'événements et en utilisant un classifieur automatique, (Russo *et al.*, 2011) propose d'utiliser les catégories grammaticales des mots dans l'entourage du mot cible à désambiguïser sur une fenêtre comprenant jusqu'à 5-grams, sans que cette expérience soit plus concluante. De plus, le corpus utilisé pour les différentes observations est en partie l'*IT-TimeBank*, mais il n'est pas précisé que c'est la définition des événements *TimeML* qui est considérée dans ces travaux.

Bel *et al.* (2010) utilisent aussi les caractéristiques reconnues comme pertinentes dans (Grimshaw, 1990), avec quelques modifications issues de Resnik and Bel (2009) (sur l'espagnol). Cette fois, les auteurs travaillent sur l'anglais en plus de l'espagnol. Le but de cette étude est la désambiguisation sémantique automatique entre événement et résultat. L'objet d'étude est un ensemble de noms qui ne sont pas des déverbaux, comme « *party* » (fête), « *conflict* » (conflit). La finalité de ce projet est d'explorer des métodes pour la génération rapide d'un lexique de noms d'événements dans deux langues (anglais et espagnol) en utilisant un corpus de taille réduite. Les traits linguistiques considérés ont été utilisés dans un classifieur automatique à base d'arbre de décisions et adaptés selon la langue :

- la tête de syntagme d'un syntagme prépositionnel actionné par « `durante` »/« `during` » (durant) » ou « `desde el principo de` » (au début de) indique un événement ;
- les noms événements sont arguments de verbes, comme « `produicir` »/« `happen` » (se produire) et « `celebrar` » (célébrer) ;
- des quantifieurs de durée qui introduisent des événements, sous la forme d'expressions temporelles comme « `dos semanas de` » (deux semaines de) ou des verbes aspectuels comme « `begin` » (commencer).

Une fois de plus, aucune définition de l'événement n'est avancée. Cette fois, il n'est pas fait non plus mention de ce qu'est un déverbal pour les auteurs et par opposition un nom non-déverbal.

Eberle *et al.* (2009) présentent un outil utilisant des indices pour la désambiguïsation en allemand des noms en « `ung` » dans leur contexte phrastique. Leur constat de départ est que les nominalisations peuvent avoir une lecture événementielle ou résultative dépendant du groupe sémantique auquel elles appartiennent. Les efforts dans cet article sont concentrés sur les nominalisations de « *verba dicendi* » (mots exprimant le discours ou introduisant une citation, comme « `dire` », « émettre », « `demander` », « `gronder` »). Le but de ces travaux est la désambiguïsation sémantique pour l'extraction d'information. Ils y présentent l'utilisation polysémique de la préposition « nach » (qui signifie « `après` » en tant que préposition temporelle ou « `selon` »/« `d'après` » pour attribuer la source d'une information). L'auteur s'interroge sur les critères de désambiguïsation mutuelle entre la préposition « `nach` » et les nominalisations en « `ung` » considérées. Ces travaux ne sont pas aboutis en outil, mais sont principalement des pistes de réflexion pour la désambiguïsation de ce cas précis de nominalisations en allemand.

1.3.3.2 Désambiguïsation sémantique, hors déverbaux

En marge de ces travaux sur les nominalisations déverbales, notons une approche de désambiguïsation sémantique en général et avec en exemple d'application les événements. (Qian *et al.*, 2009) présentent un algorithme de *bootstrapping* pour découvrir des catégories sémantiques de manière incrémentale en utilisant un lexique source et un corpus d'apprentissage de petite taille. Dans cette approche sur l'anglais, des patrons en contexte phrastique servent d'amorce. Comme un exemple d'application de leur méthode, les auteurs proposent de travailler sur des tâches de classification de noms dont les propriétés sémantiques sont contrastées : les noms événements (event-nouns) ou encore objets physiques *vs.* non objets physiques.

Les traits pris en compte sont de nature morphologique, par exemple : présence de suffixe en « -ion » pour les événements ou « -er » pour les non événements, ou encore la taille des mots qui permettrait d'indiquer la présence d'affixe dans la morphologie d'un mot. Les autres critères sont de nature contextuelle et se fondant sur les propositions verbalisées de (Van Durme *et al.*, 2008), qui consistent en des descriptions abstraites où apparaît le mot cible et qui sont extraites du *British National Corpus* par le système KNEXT[24] (Schubert, 2002). Les propositions verbalisées sont utilisées pour repérer les

24. http://www.cs.rochester.edu/research/knext/

termes de la même classe. Dans la même proposition, si en permutant les mots cibles sur l'axe paradigmatique, la proposition est toujours valide, alors les deux mots cibles sont de la même classe sémantique (dans l'exemple suivant : « destruction » et « protection »).

property may undergo a *destruction*

property may undergo a _ _

property may undergo a *protection*

La liste de noms utilisés pour l'expérience est constituée des 21 512 noms de *WordNet*. Le lexique de termes amorces utilisés pour le bootstrapping est un ensemble de noms manuellement annotés : 15 noms événementiels et 215 noms qui ne décrivent pas des événements. Des exemples de ces mots amorces ont été présentés précédemment dans le tableau 1.4. Le résultat de ces travaux est un lexique sémantique obtenu à moindre coût et qui permettrait la désambiguïsation des noms communs en « désignant d'événements » et « pas désignant d'événement ».

Les auteurs ont classé manuellement les mots cibles des cent premiers patrons selon une typologie sémantique de six classes (human events, events of inanimate objects, individual activities, social activities, lasting events, brief events). Il est dit que certains mots sont difficiles à classer dans cette typologie particulière de par leur ambiguïté, comme « `death` » (la mort) qui peut par exemple être dans les classes présentées, un *individual events* et un *brief* à la fois (dans le cas où l'on évoque le point de culmination mort, passage de vivant à trépas). Mais une autre part de l'ambiguïté n'est pas réellement traitée, celle qui concerne l'aspect des noms événements. En effet, le nom « `construction` » est répertorié dans le lexique d'amorces comme un mot désignant un événement, alors qu'il peut aussi bien désigner le procès et l'édifice construit (résultat du procès).

Dans cette section, nous avons présenté les différents travaux en extraction d'information qui se sont intéressés de près ou de très près aux dénominations d'événements. Nous avons commencé par ceux qui se focalisent sur les événements dans le sens de *ACE* et *TimeML* et qui s'intéressent en partie aux événements nominaux. Puis nous avons évoqué ceux qui ne concernent que les noms en faisant une distinction entre ceux qui sont plus orientés vers l'extraction d'informations et les traitements automatiques de ceux qui sont plus intéressés par les aspects linguistiques des désignations nominales d'événements en s'attaquant par exemple à la valeur événementielle des déverbaux. Ces derniers travaux ont pour objet des noms, mais ils s'attachent à une catégorie de noms particulière.

Conclusion

Nous avons essayé de faire le tour de la question concernant l'événement, les dénominations d'événement et comment on les voit dans les textes, afin de pouvoir extraire les informations liées aux événements.

Dans une première section, nous nous sommes interrogés sur les définitions des autres sur la notion d'événement et les critères considérés pour la cerner. Nous nous sommes aperçus que les critères de définition considérés ne sont pas toujours les mêmes dépendant du domaine de travail de chacun et que de nombreux travaux évoquent l'événement sans le définir, comme si c'était naturellement clair. Mais en réalité, la notion est trouble et souvent définie de manière empirique. Dans le chapitre 2, nous reviendrons sur la notion d'événement, telle que nous la voyons et en relation avec les travaux existants.

Tout au long de cette section, nous sommes revenus plusieurs fois sur la notion de référence, dont nous attestons l'importance. Un événement prend position dans les esprits à l'intérieur d'une réalité. La réalité nouvelle qui s'en suit fait référence à ce qui s'est produit. Une vision nouvelle de la réalité s'impose alors, que l'événement soit important (pour toute une communauté) ou plus faible (dans le cas d'un événement qui ne touche qu'un à plusieurs individus, sachant que nous ne jugeons pas de l'importance des événements). Aucun événement n'est anodin, car tous modifient une réalité, mais bien évidemment voire même heureusement, le plus souvent dans une moindre mesure. De la même façon que notre vision de la réalité est modifiée, les mots utilisés pour nommer l'événement acquièrent des caractéristiques sémantiques nouvelles en référence à un événement et ce faisant une désignation nominale est associée à un événement. Ainsi naît une dénomination d'événement.

Nous avons aussi abordé la question de la nomination et plus précisément du statut des dénominations d'événement. Nous avons fait un tour des travaux qui évoquent la formation des désignations nominales d'événements et nous nous sommes attardés sur la notion de nom propre d'événement et d'entités nommées événement.

Dans la deuxième section, nous nous sommes intéressés aux ressources existantes. Certaines servent à représenter les événements par une typologie ou dans une représentation temporelle plus englobante (comme *ACE* et *TimeML*). D'autres sont constituées de corpus liés à ces représentations temporelles. D'autres encore sont constituées de réseaux sémantiques ou de lexiques, qui apportent des informations pour l'extraction des désignations nominales d'événements. Nous y décrivons les ressources qui nous seront utiles dans le reste du mémoire (les corpus *TimeBank*, le réseau sémantique *WordNet*, les lexiques

VerbAction et *EventNominals* entre autres).

Dans la troisième section, nous avons abordé les travaux dont le but est proche de celui de l'extraction d'informations en lien avec les événements. Certains ne s'intéressant pas particulièrement aux dénominations, mais seulement aux informations connexes, d'autres uniquement aux dénominations ou encore d'autres à la valeur événementielle de certaines catégories de mots.

Nous considérons que notre objet de recherche se rapproche un peu de chacun de ces types de travaux : nous ne considérons que les noms (par opposition à *TimeML*), mais pas que les noms dont les caractéristiques sont proches des noms propres. Nous souhaitons extraire tous les noms qui décrivent des événements, de manière générale. Par ailleurs, dans ces différents travaux, des traits différents sont utilisés et des approches différentes sont proposées, nous nous en inspirons tout au long des études plus pratiques de ce mémoire (chapitre 3, chapitre 4 et chapitre 5).

Chapitre 2

La notion d'événement, proposition de typologie

Comme nous l'avons précisé précédemment, nos travaux sont ancrés en extraction automatique et dans un but d'extraction d'information. C'est pourquoi nous ne nous intéressons pas exactement aux mêmes aspects de l'événement que les travaux en sciences du langage et nous avons choisi une ligne de conduite parfois différente de celle proposée dans *TimeML*.

Nous n'avons pas l'ambition d'apporter une nouvelle définition, mais plutôt de poser les bases d'un cadre de travail qui nous convient et qui est en rapport avec notre domaine de recherche. C'est pourquoi nous avons compilé une définition de l'événement, à partir des définitions apportées par les autres travaux et tenant compte de nos attentes. Notre définition est étayée par une typologie proposée pour les événements en général, et s'adaptant très bien aussi à nos travaux sur la dénomination des événements.

2.1 La notion d'événement

Dans la définition des événements, deux points méritent d'être mis en exergue, en particulier en comparaison aux autres travaux existants : la notion d'importance de l'événement (notamment parce que la notion d'importance médiatique est redondante dans les travaux sur les événements en sciences du langage), et la prise en compte des états dans la notion d'événement (considérés comme des événements dans *TimeML*). Par ailleurs, nous souhaitons éclaircir la notion de non-événement.

2.1.1 De l'importance de l'événement

La question de l'importance médiatique dans la décision d'appeler quelque chose qui se produit un événement peut se poser. En effet, nous travaillons sur un corpus de presse écrite, et nous montrerons d'ailleurs dans le chapitre 3 que cette question nous intéresse. Cependant, dans le cas présent, notre but est d'extraire de l'information d'un texte donné, cette information étant les dénominations d'événement. De ce fait, contrairement à Moirand (2007), nous ne nous focalisons pas sur l'importance médiatique pour définir un événement.

Il est important, de plus, de noter que nous ne jugeons pas de l'appartenance de quelque chose qui se passe à la classe des événements en fonction de son impact sur le monde, nous n'évaluons pas le degré d'importance de ce qui se produit pour en faire un événement. C'est l'application à laquelle notre extraction d'information est dédiée qui définira le contenant (le corpus) et notre tâche est d'extraire les désignations utilisées dans ce corpus pour désigner des événements et non pas d'évaluer en terme de qualité le contenu informationnel du corpus.

Nous considérons qu'il y a événement dès lors que quelque chose se produit et les désignants d'événements qui nous intéressent sont les désignations nominales de ce qui se produit dans notre corpus. Ainsi, nous ne nous intéressons pas à la saillance médiatique d'un événement en particulier. Nous sommes liés au corpus et dépendons de lui.

En effet, si notre corpus contient un éditorial portant sur les grandes guerres du XX$^{\text{ème}}$ siècle, nous souhaiterions extraire de ce texte des dénominations d'événements comme « `seconde guerre mondiale` », « `Guerre du Golfe` » ou « `Guerre du Vietnam` ». Par contre, si notre corpus contient la partie faits divers ou nécrologie d'un journal local, les événements « `mariage de Mr X et Mme Y` » ou « `funérailles de Mr Z` » sont à considérer au même titre que les noms de guerre du précédent corpus.

2.1.2 Les états : des événements ?

2.1.2.1 Éventualités, actions et événements

Si nous étendons au cas de la catégorie grammaticale des noms les travaux de Vendler (1959), Verkuyl (1989) ou encore Moens and Steedman (1988) sur les catégories de verbes et l'événementialité, nous arrivons à la conclusion que nous considérons les éventualités (accomplissements et achèvements) comme faisant partie des événements, comme un tout cohérent, par opposition aux états (cf. section 1.1, pour un rappel de la catégorisation des événements verbaux dans Vendler avec des exemples). Ainsi, nous ne faisons pas de distinction entre accomplissement et achèvement, d'abord parce que cette distinction

semble très difficile à opérer de façon automatique, mais également parce que nous privilégions une typologie plus orientée vers la réutilisation des informations véhiculées par la dénomination d'événement dans le cadre d'un outil d'extraction d'information. Nous avons donc préféré annoter des informations sur la factualité et la temporalité (fréquence et moment de réalisation) de l'événement. Ces informations sont contenues dans notre typologie (section 2.2).

Nous ne distinguons pas non plus les actions et les événements comme dans la théorie de l'action de Davidson (théorie reprise en sciences du langage dans (van de Velde, 2006)). Selon Davidson (1993), la différence entre action et événement est que l'action implique toujours un agent et est intentionnelle, par opposition à l'événement, dont on ne pourrait que dire qu'il a lieu. Nous ne faisons aucune différence de traitement entre les deux : « une *course automobile* », « l' *arrivée d'un cyclone sur une île* et « sa *descente des marches* », événements intentionnels ou pas, causés par un agent ou non sont pour nous des événements à extraire de nos textes de travail. Nous ne nous focalisons pas sur les rapports à l'agent que pourraient avoir l'événement et qui le distinguerait de l'action.

2.1.2.2 D'après les autres travaux

Dans la définition de Huyghe and Marín (2007), il n'est pas clairement dit que les états ne sont pas des événements, mais les événements sont vus comme des « choses qui arrivent, se produisent ou [...] qui ont lieu », ce qui implique un mouvement, un changement d'état. Les auteurs évoquent l'éventuelle durée ou instantanéité des événements portés par les noms : certains événements se déroulent et d'autres, « constatés après-coup », restent ponctuels. Cette notion est proche de notre vision des événements, mais se borne au cas particulier des noms déverbaux, les indices proposés ne sont donc pas les seuls que nous pourrions suivre pour détecter les désignations nominales d'événements ou encore pouvoir repérer de manière automatique les types d'événements qui nous intéressent.

Pour nous, donc, dans l'événement, c'est la notion de mouvement qui est prépondérante. C'est pourquoi nous ne considérons pas les états comme faisant partie des événements. En effet, en extraction d'information, c'est la nouveauté qui est privilégiée, le fait que la situation soit modifiée. Nous n'avons donc pas la même démarche que *TimeML*, où les états aussi sont annotés, lorsqu'ils sont temporellement situés dans le texte, et parce qu'ils sont considérés comme des moments où quelque chose est tenu pour vrai :

> « This includes all dynamic situations (punctual or durative) that happen or occur in the text, but also states in which something obtains or holds true, if they are temporally located in the text (see (Saurí *et al.*, 2004) for a more exhaustive definition of the criteria for event candidacy in *TimeML*). »

Saurí *et al.* (2006)

2.1.2.3 Et le non-événement ?

Le non-événement est évoqué, lorsqu'il ne se passe rien, comme précisé dans l'exemple général suivant, relevé dans le TLFI (Trésor de la langue Informatisé) [1] :

> « **non-*événement***. En pleine guerre d'Algérie, ils diffusaient à « *Cinq Colonnes à la Une* », dix-huit minutes de reportage sur un patrouilleur d'escorte pendant lesquelles il ne se passait rien. Ce « non-événement » était pourtant plein de guerre (L'Humanité, 30 août 1984, p.2, col.1-2). »

Dans cette citation, est illustré le fait que pendant une période de guerre avérée (`la guerre d'Algérie`, qui est d'ailleurs un événement hyperonyme de tous les événements qui le constituent), on a accès à un moment où il n'y a pas d'attaque, de démolition ou de manifestation de lutte explicite, et ce non-événement fait événement étant donné le contexte de guerre en cours.

Le non-événement concerne aussi les événements qui ne se produisent pas et qui sont bien souvent attendus. Dans ce cas, il y a bien événement au sens du changement d'état. Un exemple emblématique du non-événement est le nom d'événement précédé de « `non-` » (« `la` *`non-éclosion de certains oeufs`* `arrive à chaque couvée` ») ou de « `prétendu` » (« `La` *`prétendue agression`* ») ou complété par « annulé » (« `Le` *`rendez-vous annulé de mardi dernier`* `n'a jamais été remplacé` »). Ce type d'événement est défini dans la typologie (section 2.2.1.3) : L'événement non factuel. Nous considérons que le non-événement est aussi un événement, dans le sens où l'absence de mouvement consiste lui-même en un changement d'état par rapport au mouvement attendu. Nous voyons le non-événement comme faisant partie des événements parce que le fait que quelque chose n'arrive pas est en soi un événement.

2.1.3 Synthèse sur la définition de la notion d'événement

Dans le détail, la définition de l'événement selon Lecolle (2009) (section 1.1.1) convient à notre approche en extraction d'information, d'autant qu'elle se veut faire ressortir le « noyau commun de la notion d'événement ». Pour nous, l'événement est donc :

> **ce qui survient, un changement d'état opéré et pas un état. Il peut être récurrent ou unique, prévu ou non. Il peut durer ou être instantané. Il peut aussi se produire dans le passé, le présent ou le futur**

1. Recherche de « non-événement » effectuée sur http://atilf.atilf.fr/

Nous adhérons totalement à l'observation de Mairet (1974) qui définit notre objet d'étude de manière synthétique : « l'événement, c'est ce qui arrive ». Cette remarque simple convient aussi totalement à l'extraction d'information. Ce qui arrive est forcément une modification d'état, cette remarque est donc une raison de plus de ne pas considérer les états de la notion d'événement. Le fait que l'événement arrive ou non est une information essentielle en extraction d'information, en effet elle pose la question de la factualité de l'événement et par extension de son ancrage dans le temps. Ce sont ces points que nous avons voulu faire ressortir dans notre typologie des événements.

2.2 Typologie

Plusieurs sortes de classifications des événements ont été proposées au fil des ans. Celles-ci nous sont présentées dans (Calabrese, 2010) : une classification selon l'appréhension des événements par Molotch and Lester (1996) (événement habituel dit de *routine*, *accident*, *scandale*) ; une autre selon le type de nouvelles pour Tuchman (1973) (*hard news*, *soft news*, *spot news*, *developing news*, *continuing news*) ; et les distinctions selon des critères intrinsèques aux événements avec par exemple (Veyne, 1971) (entre événement humain/naturel, répétitif/individuel) ou (Lecolle, 2009) (prévu/imprévu, répétable/non répétable, duratif/ponctuel).

Comme nous venons de le voir dans la définition (section 2.1.3), on ne s'intéresse pas à l'appréhension de l'événement ni à l'instigateur de l'événement ou encore à la source de déploiement de l'événement en tant qu'événement médiatique. Pour notre typologie, nous rapprochant plutôt de cette dernière catégorie, nous avons décidé de prendre en compte trois critères importants : la modalité, la fréquence de l'événement et le moment de sa réalisation.

2.2.1 Modalité

Par modalité, nous entendons en pratique la réalité ou non de l'événement évoqué. C'est une caractéristique nécessaire lorsqu'on essaie de représenter les événements en discours. Nous partageons ce point de vue avec *TimeML* (Pustejovsky *et al.*, 2010), mais selon des aspects différents. Nous ne distinguons pas les modalités de niveau lexical (les Situation Selecting Predicates, dépendant de la nature dynamique ou statique de l'événement) ou syntaxique (valeur du slink factive, counterfactive, evidential, negative_evidential, modal et conditional).

2.2.1.1 L'événement réel, réalisé

Si l'événement a réellement eu lieu, on y fait référence comme un fait passé ou en train de se dérouler. Il est factuel.

(2.1) Un allemand triche et gagne aux *Jeux Olympiques*$_{factuel}$.

(2.2) Le *festival de Cannes*$_{factuel}$ a sacré Laurent Cantet et son film "Entre les murs".

2.2.1.2 L'événement potentiel, hypothétique, probable

Il y a plusieurs raisons pour lesquels un événement est incertain : soit il n'est pas encore arrivé (il n'est qu'annoncé), soit il est indiqué au passé. Pour autant, celui-ci est marqué comme incertain et on peut noter la distinction entre les événements potentiels passés et futurs.

1. L'événement est dans le **futur** : On envisage qu'il se produise, on le prévoit, mais ce n'est pas encore le cas et tant que ça ne l'est pas, il n'est pas réel.

   (2.3) Les *prochaines élections législatives* se tiendront les dimanches 10 et 17 juin 2012.

   (2.4) [...] une *nouvelle secousse* peut se produire à tout moment.

   (2.5) Le *sommet Chine-Union Européenne* annulé en décembre se tiendra à Prague, en République Tchèque, au mois de mai, selon le journal China Daily.

   Nous ne touchons pas ici à la même distinction prévu/imprévu introduite par Lecolle (2009). Nous privilégions plutôt la réalisation de l'événement. S'il est imprévu, aucun discours ne l'abordera avant son exécution. Nous souhaitons nous intéresser à tous les événements dont on pourrait parler. Mais il est vrai que les notions de prévisions et de réalisations sont tout à fait compatibles.

2. L'événement est **passé** : Il est rapporté comme potentiel. Cette façon de relater un événement est courante dans le langage journalistique. Les intervenants indiquent un fait passé, jugé suffisament important pour être évoqué, alors que l'information au moment de la prise de parole n'a pas pu être correctement vérifiée. Cela peut être indiqué par l'emploi de verbes modaux, du conditionnel, du discours rapporté ou de toute autre construction.

   (2.6) Et pour accroître le désarroi, on apprend qu'un *autre attentat* se serait produit à moins de 100 km d'Alger ...

En revanche, dans le cas de l'exemple 2.7, les deux événements de la phrase ont bien eu lieu, l'incertitude concerne la relation entre ces deux événements factuels.

(2.7) L' *incident*$_{(factuel)}$ ce serait produit lors d'un *forage*$_{(factuel)}$ pour placer des explosifs dans une couche de charbon

Cette catégorie regroupe aussi les événements qui ne doivent pas se produire.

(2.8) La *guerre de Troie* n'aura pas lieu.

2.2.1.3 L'événement non factuel ou qui n'a pas eu lieu

Il nous semble intéressant de noter que certains événements sont abordés du fait qu'ils n'ont pas été réalisés.

(2.9) Cette *prétendue agression* avait débouché sur une grève surprise paralysant l'ensemble de la ligne B du RER.

(2.10) Nous n'avons perçu aucun *changement dans le discours de nos clients*, constate Sophie Romet

(2.11) Cette *tentative*$_{(factuel)}$ *d'assassinat*$_{(nonfactuel)}$

Dans l'exemple 2.11, la tentative est bien réelle, mais l'assassinat ne l'est pas. Il a été imaginé préparé, l'opération ayant échouée, il n'a pas eu lieu. Il prend de l'importance par le non changement d'état/de situation qui aurait pu ou dû se produire par sa survenue.

2.2.1.4 L'événement abstrait

Il s'agit d'événements généraux. Une classe « événement abstrait » dénote les événements sans instanciation. Dans l'exemple 2.12, on parle d'une généralité.

(2.12) La *crise* suit une période de confiance excessive.

Dans l'exemple 2.13, le locuteur évoque la réforme de manière générale et abstraite. Il présente la nécessité de réformer l'État sans pour autant évoquer une réforme en particulier. La permutation des termes « une réforme » avec l'expression « cette réforme-ci » pour le particulier est sémantiquement impossible.

(2.13) je suis persuadé qu'une *réforme en profondeur de l'Etat* est absolument nécessaire .

2.2.2 Fréquence de l'événement

La nécessité de déterminer la fréquence d'un événement et de noter que certains événements sont définis comme appartenant à une suite programmée nous est apparu flagrante,

lors d'études de corpus. Ainsi, par opposition à Neveu and Quéré (1996), nous adhérons à la mention de Calabrese (2010) pour les événements imprévisibles, prévisibles ou répétitifs, exclus par les auteurs au profit d'événement d'occurrence singulière, imprévue et non répétable. Il convient donc de déterminer la fréquence d'un événement (unique, récurrent ou instanciation d'un phénomène récurrent).

2.2.2.1 L'événement unique

Comme son nom l'indique, l'événement unique ne se produit qu'une fois. Il peut s'agir d'un groupe d'événements (les deux *attentats*), ou d'un non-événement (aucun *changement*).

(2.14) Il a publié une quinzaine d'ouvrages sur l'Inde dont le plus récent, « l'Inde, continent rebelle » (Le Seuil, 2000), relate l' *assassinat de Gandhi* par un complot de brahmanes fanatiques.

Quand on parle d'« un autre attentat » qui aurait eu lieu, dans l'exemple précédent 2.6, il ne s'agit pas d'un événement récurrent, mais bien d'un événement unique. Les attentats sont récurrents dans l'absolu dans cette partie de monde sur cette période donnée, mais celui-ci (cet autre attentat) est particulier.

2.2.2.2 L'événement récurrent

Par opposition aux événements uniques, certains événements sont récurrents. Ils se produisent à intervalles réguliers et ont la particularité de pouvoir s'instancier en événements précis. Ainsi, on évoque par « *Jeux olympiques* » l'événement hyperonyme qui se produit tous les 4 ans, mais aussi une instanciation particulière.

(2.15) Les *Jeux paralympiques* se tiennent toujours en marge des *Jeux olympiques*.

À notre sens, qui dit récurrence, dit suite d'instances qui sont par définition factuelles. Les événements récurrents sont le plus souvent de type abstrait. De plus, les événements récurrents peuvent être passés, présents ou futurs.

(2.16) Ils organisent la *campagne d'évaluation*$_{(futur)}$ qui se tiendra cette année, puis tous les ans pendant 10 ans.

(2.17) Les *JO*$_{(recurrent,present)}$ *de 2012*$_{(instance,futur)}$ sont organisés à Londres.

2.2.2.3 L'instanciation d'un phénomène récurrent

Comme présenté ci-dessus, certains événéments récurrents peuvent être individuellement instanciés, c'est le cas de « *JO de 1996* » ou « *Atlanta 96* » qui sont des instanciations de l'événement récurrent « *Jeux Olympiques* ».

2.2.3 Moment de la réalisation de l'événement

Les événements sont ancrés dans la temporalité, il n'est plus besoin de le démontrer.

Nous considérons de plus que l'événement peut être évoqué pendant qu'il se déroule, et il peut durer plus longtemps que le seul moment de son début. Par exemple lorsqu'on parle d'un vent de panique qui a lieu après une explosion : l'explosion est passée, et la panique s'installe. Cet événement caractérisé par une période de panique peut être évoqué alors que l'événement « panique » n'est pas achevé et il s'agit d'un événement présent. L'événement donc en plus d'être passé, peut être présent et même se dérouler dans le futur.

Nous estimons qu'il est opportun de les caractériser par le moment de leur réalisation. Nous évaluons la temporalité des événements en fonction du moment de l'énonciation.

2.2.3.1 L'événement passé

(2.18) `Ils aboutissent à un` *constat d'échec*

2.2.3.2 L'événement présent

(2.19) `Cette` *rentrée*`-ci se place sous le signe de la contestation sociale.`

2.2.3.3 L'événement futur

(2.20) `Le` *sommet Chine - Union Européenne* `se tiendra à Prague.`

La typologie que nous proposons prend en compte les principales données portées par les dénominations d'événement et le cotexte d'apparition du nom. Ce sont les informations principales qui à notre avis seraient nécessaires dans une tâche d'extraction d'information de désignants d'événements.

Notre typologie permet de plus de clarifier la définition, de préciser l'annotation (chapitre 3) et ainsi définir les entités à extraire (chapitre 5), et ce même si nous n'avons pas

annoté tous les documents avec tous les types de la typologie et que nous ne cherchons pas à extraire les types de la typologie pour les dénominations d'événements qui sont l'objet de notre travail d'extraction automatique. D'autant que mener à bien un tel projet nous demanderait de savoir gérer des connaissances du monde que nous ne maîtrisons pas. C'est pourquoi dans le cadre de l'extraction automatique des désignants d'événements nous nous sommes uniquement attelés à la tâche de repérer les désignations nominales d'événement.

2.3 Événement nominal : les trois modes de construction possibles

Il existe une différence de traitement entre les événements verbaux et les événements nominaux. Dans cette section, nous apportons des informations sur ces deux façons de désigner des événements et nous présentons nos observations sur la formation des désignations nominales des événements. Ainsi, nous avons observé trois modes de construction de ces désignations d'événements.

2.3.1 Événements verbaux et événements nominaux

Tout d'abord, il faut noter que les événements sont relatés en langue selon deux méthodes principales : l'usage de syntagmes verbaux et la nomination au moyen de formes nominales, même s'il est possible (comme le fait *TimeML*) de considérer d'autres éléments, comme des adjectifs ou même des nombres, qui peuvent posséder des capacités événementielles, mais de manière marginale.

2.3.1.1 Les verbes

Les verbes (hors verbes d'état) ont pour rôle principal de représenter des événements. En effet, la description d'un événement se fait d'abord au moyen de phrases complètes décrivant des acteurs, des actions, des circonstances, un contexte. Le verbe est voué à cet usage, et renferme notamment des propriétés d'aspect, de mode, de temps (Vendler, 1959). De par ses propriétés intrinsèques, il caractérise les événements, et il est plutôt simple à lier aux expressions temporelles. Les verbes désignent souvent des événements plus communs et moins pertinents. C'est lorsque que les événements prennent de l'importance ils sont nominalement nommés. Toujours et nous l'avons vérifié dans nos corpus, un événement est d'abord désigné par ce qui se produit avant d'être nommé comme tel.

Par exemple, au moment des attaques du 11 septembre à New York, les médias ont

reporté l'incident de manière verbale tout d'abord (exemples 2.21) ; avant de parler d'explosion, d'attaque ou d'attentat plus tard (exemples 2.22).

(2.21) `des bombes` *`ont éclaté`*
`des avions` *`se sont écrasés`* `dans le WTC`

(2.22) `Une` *`explosion`* `a retenti`
`Les attaques du 11 septembre`
`Le` *`11 septembre`*
`L'` *`attentat terroriste`*

En TAL, cette particularité des événements a été largement traitée notamment à travers le formalisme d'annotation des événements *TimeML* (Pustejovsky *et al.*, 2003). Le but de ces travaux est d'extraire les événements, les expressions temporelles, ainsi que la relation existant entre les deux (Verhagen *et al.*, 2007). Les tentatives d'extractions automatiques n'ont pas donné de résultats très satisfaisants, mais grâce à ces travaux, de nombreux corpus annotés ont été créés. Il en existe à des tailles variables sur plusieurs langues, dont le TimeBank pour l'anglais (première langue à avoir bénéficié de ces travaux), l'IT-TimeBank en italien et le FR-TimeBank en français.

2.3.1.2 Les noms

La catégorie grammaticale des noms n'est pas connue pour porter pas toutes les informations véhiculées par le verbe (modalité, temps, factualité, etc.). Indiquer un événement n'est pas une qualité que possèdent tous les noms. De plus, même lorsqu'un nom servir à désigner un événement, même si la désignation qui en résulte est porteuse du sne sévénementiel, le nom lui même ne porte pas toujours un sens événementiel de manière définitoire (`la mairie` par rapport à `les municipales` ou `les élections municipales`). Dans certains cas, en contexte, la désambiguïsation est possible, dans d'autres elle ne l'est pas, il faut souvent faire référence à une connaissance extra-contextuelle. Il existe aussi des syntagmes nominaux qui par métonymie, par exemple, se chargent d'une valeur événementielle non prévue, qui dans l'usage en discours est plus ou moins clair. Si on peut dire que la plupart des verbes désignent des événements, ce n'est pas le cas des noms, ce qui rend les événements nominaux plus difficile à repérer que les événements verbaux, d'autant qu'ils sont moins fréquents.

En TAL, peu de travaux se sont focalisés sur les désignations nominales d'événements. Même si dans *TimeML* des noms sont annotés et décrits, ils ne sont pas nombreux et ne couvrent pas la totalité de ce qui nous intéresse.

2.3.2 Comment sont construits les désignations nominales événementielles ?

Nous nous intéressons aux groupes nominaux qui désignent les entités que nous avons définies comme des événements. Nous considérons l'événement comme ce qui survient, le changement d'état opéré. Il peut être récurrent ou unique, prévu ou non, durer ou être instantané, se produire dans le passé, le présent ou le futur. Parmi les événements, nous distinguons trois classes de constructions d'événements nominaux différentes. Chacune doit, selon nous, être considérée de façon spécifique dès lors que l'on cherche à trouver les dénominations d'événements dans les textes.

2.3.2.1 Nominalisation apparentée à un verbe d'action

Beaucoup d'événements sont construits à partir de noms morphologiquement apparentés à des verbes d'action. Sans parler de déverbaux, ni nous aventurer sur les sens de dérivation possible des conversions (cf. (Tribout, 2010)), nous nous focalisons ici sur les paires nom-verbe issues d'une conversion ou de paires nom-verbe apparentées car formées sur des bases identiques, même si la dérivation morphologique n'est pas avérée (« `fête/fêter` » : « `la` *`fête de la musique`* » ; « `adoption/adopter` » : « `l'` *`adoption d'une réglementation`* » ; « `démission/démissionner` » : « `il a annocé sa` *`démission du gouvernement`* »). Tous ces noms n'ont bien sûr pas une lecture événementielle ; de plus, ceux ayant une lecture événementielle possèdent en général au moins une autre lecture possible, souvent celle du résultat de l'action (« *`construction`* », « *`exil`* »), mais aussi de son instrument (« *`shampooing`* »), du lieu (« *`étalage`* », « *`arrivée`* »), etc.

Dans les exemples suivants, les mots écrits entre parenthèse et en indice corresponde à la valeur sémantique des mots en contexte : « *event* » pour l'usage événementiel, « *obj* » pour l'objet, « ? » quand l'usage est difficile à déterminé.

(2.23) 1. `La` *`construction`*$_{(event)}$ `du métro toulousain a commencé en 1988.`

2. `La vue de cette construction`$_{(obj)}$ `prodigieuse inspire des réflexions philosophiques bien différentes.`

(2.24) `L'étalage`$_{(?)}$ `de marchandises a été interdit sur le port.`

(2.25) `Soumettre une proposition`$_{(?)}$ `de loi`

Ces exemples illustrent bien que le problème ne se situe pas seulement dans l'optique d'une extraction automatique, mais qu'un lecteur humain peut parfois avoir du mal à décider du caractère événementiel ou non d'un mot en contexte. C'est le cas (selon nous) des deux derniers exemples. Le problème qui se pose est donc également un problème de définition et de choix d'annotation en cas d'ambiguïté.

2.3.2.2 Nominalisation autre que dérivée de verbes

Certaines dénominations, autres que ceux apparentés aux verbes d'action, décrivent intrinsèquement des événements (« le *festival de Cannes* » ou « le *match PSG-OM* »). Les mots provenant d'autres langues et introduits en français pour référer à un événement particulier, les xénismes, font partie de cette catégorie (« l' *apartheid* » ou « *tsunami* »). Ces mots n'échappent pas, comme tous les autres, à l'homonymie ou à la polysémie, qui peuvent conduire à l'ambiguïté de leur lecture événementielle :

(2.26) 1. le *salon*$_{(event)}$ *du livre*

2. le salon$_{(obj)}$ de ma grand-mère

(2.27) 1. Il est le nouveau champion du monde de triathlon$_{(sport)}$.

2. Il a remporté le *triathlon*$_{(event)}$ des lacs.

2.3.2.3 Nominalisation métonymique

Enfin, des syntagmes nominaux sans valeur événementielle peuvent, par le résultat d'une métonymie, en contexte, référer à l'événement lié à :

- **un lieu**, comme « Tchernobyl » qui désigne à la fois le lieu (c'est le nom de la centrale nucléaire ukrainienne) et l'incident qui s'y est produit, soit l'accident nucléaire de 1986. Ce sont les toponymes événementiels définis dans (Lecolle, 2009) ;

  (2.28) Personne ne veut d'un nouveau *Tchernobyl*.

  (2.29) *Copenhague* se solde par un échec.

- **une date**, comme « le 11 septembre » pour l'attaque terroriste en 2001 aux États-Unis. Il s'agit des héméronymes, introduits dans (Calabrese, 2008) ;

  (2.30) Les indemnisations pour le *11 Septembre*

  (2.31) On pourrait assister à un *21 avril* à l'envers.

- **l'objet** d'une affaire, comme « les frégates de Taïwan » (qui évoque un scandale politico-financier lors de la vente par des industrie françaises de six frégates destinée à la marine taïwanaise au début des années 1990) dans :

  (2.32) « Les *frégates de Taïwan* s'invitent à Lorient.

L'emploi événementiel par métonymie (cf. section 1.1.2.3) est rarement le plus courant, et l'ambiguïté est par nature systématique. Il s'agit pourtant généralement d'événements importants, du moins médiatiquement parlant, qu'il est nécessaire de pouvoir détecter.

La distinction entre ces trois classes est importante en vue d'une extraction des désignants nominaux d'événement. Pour la première classe, des indications morphosyntaxiques sur une conversion en lien avec un verbe seront utiles. Dans le deuxième cas, la constitution d'un lexique semble nécessaire. La troisième classe pose un problème plus rare, mais plus difficile.

Conclusion

Dans ce chapitre, nous avons défini notre objet d'étude. Nous avons proposé une typologie fondée sur les notions de modalité, fréquence et sur le moment de la réalisation des événements. Cette typologie prend en considération les critères que retiennent les applications d'extraction d'information temporelles, ce qui en fait une typologie adéquate. De plus, nous avons présenté nos observations concernant les événements nominaux et leurs modes de formation.

Sur la base de ces travaux définitoires, et afin de pouvoir proposer une extraction automatique, nous nous devons de passer à la pratique. La définition de l'objet « événement » et la typologie nous permettent de préparer un guide d'annotation de corpus. L'annotation manuelle de corpus est une étape essentielle pour la continuation de nos recherches. Elle permet de se confronter aux données réelles, de vérifier les hypothèses avancées. De plus, nous avons besoin de ressources fiables et homogènes pour faire de nouvelles observations sur les événements en contextes et bien évidemment pour mener à bien une démarche d'extraction automatique des désignations nominales d'événements.

Chapitre 3

Un corpus annoté

Comme nous l'avons vu dans les précédents chapitres, des travaux ont été menés sur l'extraction d'événements qui ont fait suite à des réflexions sur la notion d'événement.

Dans le chapitre 1, nous avons présenté les corpus existants annotés en événements et particulièrement en événements nominaux (cf. section 1.2.2). Les corpus librement distribués et utilisables sont essentiellement les corpus *TimeBank*. Ils sont disponibles en plusieurs langues, incluant l'anglais et le français, les deux langues auxquelles nous nous intéressons. Ces corpus sont annotés selon les directives *TimeML* (Pustejovsky *et al.*, 2010) et sont annotés des diverses marques temporelles des textes, dont entre autres les événements nominaux.

Pourtant, dans les spécifications *TimeML*, l'intérêt est d'abord porté sur le **verbe**, considéré comme le meilleur porteur de l'événement et de la temporalité. Plusieurs travaux constatent que les guides d'annotation *TimeML* ne fournissent pas beaucoup d'information sur l'annotation des noms. Un autre point de contestation avec les événements nominaux annotés par *TimeML* est que **les états** sont annotés. En effet, nous considérons dans notre définition que les événements doivent traduire un changement d'état. D'ailleurs dans les corpus *TimeBank*, une différence est faite entre les noms actifs et les noms statiques. Par ailleurs, dans les corpus TimeBank, **seuls les noms têtes de syntagmes** sont annotés. Les syntagmes nominaux entiers ne sont pas annotés. Il n'y a pas de problème de frontière du nom d'événement puisque ne sont annotés que les noms porteurs d'une dénotation d'événement et non pas les noms d'événements complets. Ainsi dans le texte « `festival de Cannes` », *TimeML* annotera uniquement « `festival` », alors que dans notre démarche (étroitement liée à la notion d'entités nommées), nous annotons

le groupe nominal complet « *festival de Cannes* », comme nom d'événement.

Dans nos travaux, nous nous intéressons uniquement aux désignations nominales et nous appuyons sur la définition et la typologie présentées au chapitre 2. Bien que les corpus en *TimeML* ne traduisent pas toutes nos attentes en terme d'annotation des noms d'événements, nous nous en servirons à des fins d'évaluation et de comparaison avec notre corpus annoté en français.

Nos travaux ont pour cadre l'extraction d'information et l'objectif de notre travail est de parvenir à extraire des textes les événements nominaux de manière automatique. Dans ce but, que ce soit pour le développement, l'apprentissage ou l'évaluation, l'élaboration d'un corpus annoté est une phase nécessaire. Les textes sur lesquels nous travaillons sont issus de corpus journalistiques, parce que ces corpus sont facilement disponibles, en grande quantité, qu'ils ont vacation à transmettre l'information en évoquant des événements qui se produisent. Nous avons donc appliqué notre définition générale concernant les événements au cas particulier des événements nominaux, en annotant des textes journalistiques issus de la presse écrite (articles de presse classique). À cette fin, nous avons développé notre propre guide d'annotation.

Dans ce chapitre, nous décrivons notre guide d'annotation en tenant compte de notre implication dans les entités nommées du projet Quaero et en mettant particulièrement l'accent sur la problématique de l'annotation des désignations nominales. Puis, nous présentons les caractéristiques du corpus annoté. Et nous concluons par une étude approfondie des comportements des noms d'événements dans notre corpus annoté.

3.1 Le guide d'annotation

Notre guide d'annotation en événements nominaux est lié à notre définition des événements et à la typologie détaillée des événements que nous avons développées et précédemment présentées. Nous proposons ici un tour d'horizon des principaux points essentiels du guide, une version temporaire du guide d'annottaion est proposée en annexe (Annexe C.2).

Tout d'abord, nos annotations sont faites selon le langage de marquage XML. Le balisage XML permet la structuration du document et l'attribution d'étiquettes aux zones de textes qui nous intéressent. Nous avons choisi pour les événements de nommer notre balise de marquage `<event>` à laquelle nous avons associé des attributs selon la typologie (cf. section 3.1.2.2). Dans ce chapitre, nous représentons nos exemples sous une forme

indentée ou sous la forme d'arbres. Ces deux formes tiennent compte des étiquettes XML.

Notre guide a aussi été développé sur les bases existantes du guide sur les entités nommées générales édité dans le cadre du projet *Quaero* (Rosset *et al.*, 2011). Le projet nous a fourni un cadre d'application à nos travaux et quelques prérogatives à respecter. Nous énonçons d'abord les particularités du guide des entités nommées générales de Quaero, en les illustrant d'exemples d'annotation de noms d'événements, avant de développer les particularités de notre guide concernant précisément les événements nominaux.

3.1.1 Une annotation sur la base de l'annotation des entités nommées *Quaero*

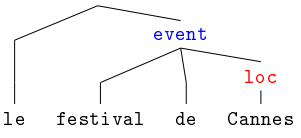
Le guide d'annotation a été élaboré en suivant notre définition des événements, mais en fonction des particularités des événements nominaux et pour les événements nominaux uniquement. Il ne peut être utilisé pour d'autres types d'événements. Il est conçu comme complémentaire au guide d'annotation en entités nommées *Quaero*[1] et doit suivre les recommandations et contraintes présentées dans ce guide, dont voici les plus importantes : l'imbrication des entités nommées et le traitement des déterminants.

3.1.1.1 Les annotations imbriquées avec d'autres entités nommées

Dans les textes, plusieurs entités nommées (possiblement de types différents) peuvent être concomitantes ou présentes dans la même portion de texte. Il est alors inévitable que certaines se retrouvent imbriquées dans d'autres. Les annotations imbriquées avec des entités nommées des autres types sont donc légitimes. Dans « *festival de Cannes* », « Cannes » est un lieu/nom de ville[2], comme on le voit dans l'exemple 3.1.

(3.1)

```
le
    <event> festival de
      <loc> Cannes </loc>
    </event>
```



Les entités événements sont parfois obtenues par métonymie au moyen d'une information liée à l'événement, comme la date ou le lieu où s'est déroulé l'événement. Dans ce cas, deux annotations doivent être portées par le même groupe de mots.

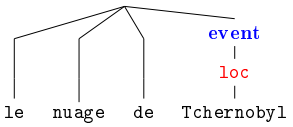
1. Notre guide d'annotation des événements nominaux n'est pas encore intégré dans le guide des entités nommées générales, mais à vocation à l'être.

2. Dans le guide d'annotation *Quaero*, un nom de ville est annoté `<loc.adm.town>`, pour la simplicité de la lecture de ce mémoire et parce que ça ne porte pas à confusion, nous avons réduit l'étiquette à uniquement `<loc>`.

Quand le nom « Tchernobyl » sert à nommer la catastrophe nucléaire de la centrale Tchernobyl, ce nom doit être annoté comme <event>, tout en gardant son étiquette indiquant le lieu (exemple 3.2). Par contre, dans l'exemple 3.3, le mot « Tchernobyl » désigne le lieu et l'événement est nommé par « la *catastrophe de Tchernobyl* ». L'annotation événement est au-dessus de l'annotation de lieu, parce que le but est de sur-annoté les annotations existantes de *Quaero*, mais aussi parce que d'un point de vue historique, Tchernobyl a désigné d'abord le lieu, c'est le nom de la centrale nucléaire avant de désigner l'événement qui s'y est produit.

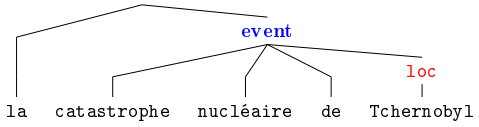
(3.2)

```
Le nuage de
    <event>
      <loc> Tchernobyl </loc>
    </event>
```



(3.3)

```
Cette zone a été polluée par les retombées de la
    <event> catastrophe nucléaire de
      <loc> Tchernobyl </loc>
    </event>
(1986)
```



Prenons l'exemple de « 11 septembre » [3] pour les dates qui servent à nommer des événements. Dans l'exemple 3.4, on montre qu'une date peut aussi désigner un événement, et que celle-ci doit conserver son annotation en tant que date.

(3.4)

```
La découverte d'autres pilotes terroristes avant le
    <event>
      <time> 11 septembre </time>
    </event>
aurait pu limiter les attaques et les pertes en vies humaines
```

Comme pour les noms de lieux, nous sommes tenus de distinguer les occurrences de dates qui nomment l'événement et celles qui sont utilisées en tant que date uniquement. Cette distinction sémantique se fait en contexte et compte tenu de la connaissance du monde du lecteur/annotateur. Elle peut être relativement simple (exemple 3.5) ou plus

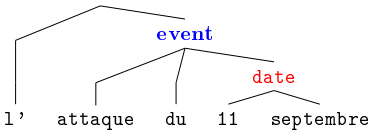
3. Dans *Quaero*, une date comme « 11 septembre » est annotée <time.date.abs>, pour la simplicité de la lecture, nous avons réduit l'étiquette à uniquement <time>.

complexe (exemple 3.6) dans les cas où il est difficile de différencier ces deux usages de la date.

(3.5) Mais le mal était fait, car évoquer un " *21 avril à l'envers*" est inconcevable pour Nicolas Sarkozy. [4]
(où « *21 avril* » réfère à ce jour de 2002 où l'événement marquant a été qu'un représentant de parti d'extrême droite a réussi à se qualifier pour le second tour d'une élection présidentielle en France)

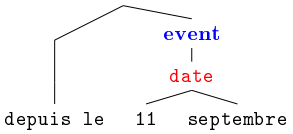
Alors, nous conseillons d'utiliser un test linguistique à cet effet. D'après (Ehrmann and Hagège, 2009), insérer l'expression « l'événement qui s'est produit le » avant la date que l'on présume être un événement est un test efficace. Si la nouvelle phrase est une paraphrase correcte de la précédente, alors c'est bien un événement. Ainsi dans l'exemple 3.6, « 11 septembre » correspond bien à la date, car l'insertion de l'expression « l'événement qui s'est produit le » rend la phrase incorrecte.

(3.6) l'attaque du 11 septembre signe la fin [...]
* l'attaque de *l'événement qui s'est produit le* 11 septembre signe la fin [...]



Par contre dans l'exemple 3.7, les deux phrases sont équivalentes. Donc on en déduit que « 11 septembre » désigne bien l'événement, ce qui ne l'empêche pas aussi de désigner la date.

(3.7) Depuis le 11 septembre, nous n'avons perçu aucun changement
= Depuis *les événements qui se sont produits le 11 septembre*, nous n'avons perçu *aucun changement*.



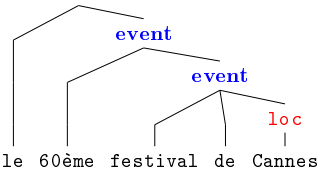
3.1.1.2 Les annotations imbriquées entre événements

Les annotations imbriquées d'événements sont aussi possibles et autorisées par le guide dans le cas de dénominations d'un événement général par rapport à l'événement particulier ou un événement lié à un autre. L'événement hyperonyme est en règle générale

4. Source : Europe1.fr, article daté du 19 mars 2012

étiqueté à l'intérieur de l'événement particulier. Dans l'exemple 3.8, « *60ème festival de Cannes* » est composé de l'événement récurrent « *festival de Cannes* » et de l'instanciation de cet événement récurrent : la 60ème édition de ce festival (cf. section 3.1.2.2 pour plus de détail sur l'annotation suivant la typologie).

(3.8)
```
le
    <event>
    60ème
        <event>
        festival de Cannes
        </event>
    </event>
```



Par ailleurs, la désignation d'un événement peut se faire à l'intérieur du syntagme nominal désignant un autre événement. Ainsi, suivant les contraintes de frontières du nom d'événement que nous avons définies (et qui sont présentées en section 3.1.2.1), un événement est imbriqué dans un autre sans que cette imbrication ne manifeste l'instanciation d'un événement récurrent. Bien évidemment, ceci traduit un lien entre deux événements. Dans l'exemple 3.9, le lien entre « *réunion* » et « *assemblée générale* » est d'ordre succession temporelle.

(3.9)
```
<event> réunion houleuse après une
    <event> assemblée générale à la Bourse du travail </event>
</event>
```

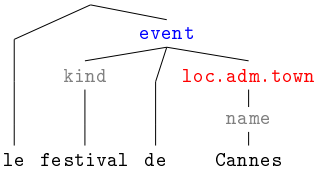
De plus, dans le guide d'annotation *Quaero*, la partie de l'expression qui désigne un **hyperonyme** de l'entité est annoté par la balise <kind>, définie dans le guide des entités nommées de *Quaero*, comme :

> « Partie d'une expression d'entité nommée qui désigne un hyperonyme (genre proche) de l'entité » : « le maire de Paris »
> `le <func.ind> <kind> maire </kind> de <loc.adm.town> Paris </loc.adm.town> </func.ind>`
> Dans cet exemple, maire est une fonction de type <func.ind>

Pour les événements cela signifie que « `festival` » sera annoté <kind> dans l'expression nom d'événement « *festival de Cannes* ». Dans l'exemple suivant, une balise <name> est posée sur le nom de lieu « `Cannes` », pour illustrer l'annotation dans *Quaero*.

(3.10)

le
<event>
<kind> festival </kind> de
<loc>
<name> Cannes </name>
</loc>
</event>



3.1.1.3 Le traitement des déterminants

Dans les directives pour les autres entités nommées définies *Quaero*, les déterminants ne sont pas annotés à l'intérieur des groupes désignants des entités nommées, mis-à-part pour les entités nommées temporelles, de date ou d'heure. Même si les événements sont fortement temporels de par leur ancrage dans le temps, nous avons finalement choisi d'annoter les événements comme la plupart des entités nommées *Quaero* en n'autorisant pas les déterminants dans le bloc annoté. Ainsi, aucun **déterminant** n'est contenu à l'intérieur des balises encadrant le groupe de mot annoté.

Même s'il est vrai que de nombreux déterminants apportent une information supplémentaire à l'entité, nous avons choisi qu'il ne fasse pas partie intégrante du nom d'événement en tant qu'entité nommée événementielle. Pour autant, ces déterminants sont des modifieurs de l'événement, ils héritent de la balise <event-modifier>. Dans les exemples suivants, c'est le cas de « cet » pour le nom d'événement « cet *important remaniement* » (exemple 3.11), mais jamais de l'article défini, comme « la » dans l'exemple 3.12 .

(3.11)

la surprise de <event-modifier> cet </event-modifier>
<event>
important remaniement
</event>

(3.12)

Ça a créé *la*
<event>
surprise
</event>.

De plus, suivant les prérogatives du guide d'annotation des entités nommées étendues de *Quaero*, il conviendra d'utiliser la balise <ordinal> pour annoter les **ordinaux**, comme « 60ème » dans « *60ème festival de Cannes* ». Nous considérons que les ordinaux font partie intégrante de certains noms d'événements, c'est le cas des événements récurrents.

(3.13)
```
le  <event>
    <ordinal> 60ème </ordinal>
     <event> festival de Cannes </event>
 </event>
```

Nous avons présenté dans les grandes lignes, les contraintes imposées par le guide d'annotation des entités nommées *Quaero*, avec lequel nous souhaitons que notre guide soit en correspondance. Mais il ne s'agissait que d'une présentation succinte de certains aspects des règles d'annotation des entités nommées dans le projet *Quaero*.

3.1.2 L'annotation des événements nominaux

Nous présentons maintenant, les particularités du guide concernant les noms d'événement en particulier. Notre guide est accès sur trois points principaux. Le premier concerne la définition des frontières de syntagmes : nous travaillons avec des syntagmes nominaux qui désignant des événements et la question se pose de savoir qu'elles parties du discours doivent être considérées comme étant à l'intérieur du syntagme. Le deuxième point concerne l'annotation des noms d'événements en fonction de la typologie générale des événements que nous avons définie plus tôt dans ce manuscrit (chapitre 2, section 2.2. Et le troisième point est axé sur des indices à suivre pour l'annotation de nos entités événement.

3.1.2.1 La définition des frontières des noms d'événements

Nous avons opté pour une limitation des frontières sur une base syntaxique, bien que les entités nommées soient définies purement sur un point sémantique. Vu que nous avons adopté une définition étendue des entités nommées et qu'ici nous annotons toutes les désignations nominales d'événements, les noms d'événements que nous voulons annoter ne sont pas seulement un nom (tête de syntagme) et ses dépendances nominales simples (adjectif), mais aussi parfois des groupes nominaux plus étendus. Pour autant, nous ne pouvons pas non plus y inclure tous les dépendants nominaux du nom tête de syntagme, comme nous le verrons ci-après.

La question des frontières est essentielle, mais peut facilement enrichir un débat sans espoir de consensus. Mis à part pour les dépendances nominales simples comme les adjectifs, les compléments du nom, il n'est pas facile de s'accorder sur les types de mots à conserver dans l'annotation.Nous avons tranché, parfois non sans mal, devant les observations du corpus que nous avons menées. Nous ne détaillerons pas cette partie ici.

Pour avoir une description de l'étude des frontières des syntagmes nominaux événement, le lecteur pourra se référer à l'annexe C.2, où le guide d'annotation complet est inclu avec une section destinée à ce problème particulier de délimitations de frontières.

Pour synthétiser, nous fondons l'annotation sur une analyse syntaxique des textes pour déterminer les frontières des syntagmes nominaux événements. La tête de syntagme et ses dépendants syntaxiques directs (compléments du nom, adjectifs, attributs, appositions ...) sont annotés. Des termes issus d'autres catégories morpho-syntaxiques sont aussi annotés (participes passés, compléments prépositionnels de temps et de lieu) et d'autres sont exclus (participes présents, subordonnées relative et infinitive). Ce choix est tout à fait discutable.

3.1.2.2 La typologie

Notre typologie a été présentée dans le chapitre 2.2. Nous ne développerons pas dans les détails l'application de la typologie sur le corpus, parce que de nombreux exemples sont présentés dans le guide d'annotation en annexe C.2. Notons simplement que l'annotation s'est faite au moyen d'un balisage XML, où pour la balise `<event>` sont définis trois attributs obligatoires à compléter par une seule valeur parmi les suivantes :

- l'attribut de modalité (`type`) reflète la réalité ou non de l'événement évoqué : `factual`, `hypothetical`, `nonfactual`, `abstract`
- l'attribut de fréquence (`frequence`) permet de différencier les événements programmés des autres et de ceux appartenant à une suite d'événements : `unique`, `recurring`, `instance`
- l'attribut du moment de la réalisation (`temp`) donne une indication concernant l'ancrage sur l'axe temporel de l'événement : `before`, `now`, `after`

L'étiquette `<event>` supporte aussi un attribut optionnel (`source`), dont la valeur désigne la provenance de l'information (`reported` pour un discours rapporté ou `fictive` pour un événement issue d'un récit non factuel). Comme dans le guide d'annotation *Quaero* pour les autres entités nommées, les valeurs `unknown` (signifiant « je ne sais pas quelle valeur choisir parmi celles disponibles ») et `other` (« je sais qu'aucune des valeurs proposées ne convient ») peuvent être utilisées par les annotateurs. Mais dans le cadre de ce corpus, ces valeurs n'ont pas été utilisées.

Voici quelques exemples (texte seul puis exemple balisé) présentant un panel des utilisations des attributs de `<event>` et leurs différentes valeurs :

(3.14) `Sa nomination crée la surprise.`

→ `Sa <event type="factual" frequence="unique" temp="before">` *nomination*

</event> crée la <event type="factual" frequence="unique" temp="before"> *surprise* </event> .

(3.15) Les tâches à dominante féminines aboutissent rarement à la réalisation d'objets durables.
→ Les tâches à dominante féminines aboutissent rarement à la <event type="abstract" frequence="recurring" temp="now"> *réalisation d'objets durables* </event> .

(3.16) À en croire un responsable onusien, quelque vingt-cinq délégués représentant toutes les ethnies afghanes devraient participer à une conférence qui devrait se terminer en "moins d'une semaine" et, "dans le meilleur des cas", pourrait aboutir à un accord de principe sur la mise en place d'un gouvernement transitoire à Kaboul
→ *À en croire un responsable onusien*, quelque vingt-cinq délégués représentant toutes les ethnies afghanes devraient participer à une <event type="factual" frequence="unique" temp="after" source="reported"> *conférence* </event> qui devrait se terminer en "moins d'une semaine" et, "dans le meilleur des cas", pourrait aboutir à un accord de principe sur la <event type="hypothetical" frequence="unique" temp="after"> *mise en place d'un gouvernement transitoire à Kaboul* </event> .

(3.17) A travers une série de fables qui entrent en résonance avec l'actualité, Emilie Valantin met en scène les ficelles de la haine et de la sottise. Jusqu'à l'inéluctable rébellion.
→ A travers une série de fables qui entrent en résonance avec l'actualité, Emilie Valantin met en scène les ficelles de la haine et de la sottise. Jusqu'à l' <event type="abstract" frequence="unique" temp="now" source="fictive"> *inéluctable rébellion* </event> .

La typologie constitue le point d'orgue de notre guide concernant l'annotation de notre corpus. Elle nous aide à annoter, bien plus que l'annotation des types est importante. Tout notre corpus annoté n'est pas riche des informations de la typologie. Seule 232 noms d'événements ont été annotés avec tous les attributs correspondants à la typologie, ce qui correspond à 12,5 % du corpus. Parce que nous avions besoin d'avoir rapidement accès à un corpus annoté en événement, nous avons préféré annoter en quantité suffisante les noms d'événements à défaut de les annoter avec la haute qualité que nous aurions souhaitée (avec les informations issues de la typologie).

3.1.2.3 Les indices à suivre pour l'annotation

La tâche d'annotation des événements étant particulièrement subjective, voici quelques indices à suivre, malheureusement difficile à reproduire automatiquement.

- Il est recommandé aux annotateurs de ne pas hésiter à se référer à des dictionnaires de langue pour vérifier le sens des mots qu'ils cherchent à annoter. Les définitions des dictionnaires (comme *TLFI*) sont souvent très utiles pour lever les ambiguïtés.

- La permutation des mots ambigus avec d'autres mots qui sont, eux, clairement événements ou non événements est un bon stratagème. Dans l'exemple 3.18, « `preuve` » est remplaçable par « `surprise` », et non par « `document` », ce qui en fait un événement :

  (3.18) `Les 37 journalistes français attendus ici pour le suivre en sont une` <event> *preuve supplémentaire* </event>.

  En revanche, c'est le contraire dans « `Les preuves présentes dans le dossier sont accablantes` ».

- Les noms qui sont dérivés de verbes d'action ont plus de chances de décrire des actions et donc des événements que les autres. « `Absence` » est lié à « `être absent` », un état et non à « `s'absenter` », l'action, il s'agirait plutôt de « *abstention* ». Par opposition, « *maintien* » est lié au verbe « `maintenir` » qui présence une action.

- Si une phrase est composée d'une énumération de noms d'événement sûrs et non ambigus, le plus souvent le nom qui semble ambigu dans cette même énumération est à étiqueter comme un événement. De même, dans une énumération de termes non événement et non ambigus, on n'annotera pas les termes ambigus en événement.

- Dans le cas d'une ambiguïté impossible à résoudre, choisir de ne pas annoter. C'est le cas par exemple d'assemblée dans « `sauf décision contraire de l'assemblé générale` ». De nombreux autres exemples sont illustrés dans le guide d'annotation.

Les indices présentés ici, sont d'ordres extrêmement pratiques. Avec la typologie, ils sont d'un grand soutien pour l'annotateur.

Dans cette section, nous avons présenté notre guide d'annotation, les problèmes rencontrés pour l'annotation des événements nominaux et les moyens mis en œuvre pour

les surmonter. Le guide d'annotation que nous avons rédigé contient de nombreuses indications sur la façon d'annoter les noms qui désignent des événements et de nombreux indices pour l'annotation des syntagmes nominaux. Il explique aussi comment annoter en fonction de la typologie que nous avons mise en place. Notre guide été rédigé pour composer un corpus annoté en événements nominaux.

3.2 Notre corpus annoté

Nous avons annoté manuellement 192 articles de presse des journaux *Le Monde* (*LM* – 83 articles) et *L'Est Républicain* (*ER* – les 109 articles utilisés dans le *FR-TimeBank*). Ce corpus représente 1844 noms d'événements. À titre de comparaison, le corpus italien IT-TimeBank (Russo *et al.*, 2011) compte 3695 événements, le *FR-TimeBank* 663 et le corpus (Creswell *et al.*, 2006) 1579. Le tableau 3.1 propose une description chiffrée de notre corpus. Ce travail d'annotation et le corpus ont été présentés dans (Arnulphy *et al.*, 2012b).

	LM	*ER*	*Total*
Textes	83	109	192
Mots	31 449	16 197	47 646
<event>	**1 107**	**737**	**1 844**

TABLE 3.1 – Notre corpus annoté, constitué d'articles des journaux *Le Monde* (*LM*) et *L'Est Républicain* (*ER*). Les articles de l'*ER* sont les textes issus du *FR-TimeBank*, réannoté en suivant notre guide d'annotation

3.2.1 Accord inter-annotateur

Les annotations ont été conduites en fonction du guide d'annotation présenté à la section précédente. Le corpus a été annoté par deux annotateurs, auteurs du guide. La majorité des textes (environ 75 % du corpus) a été annotée par les deux annotateurs sans concertation pendant l'annotation. Ces annotations ont plus tard été révisées conjointement avec les deux annotateurs. Ce corpus est donc extrêmement fiable.

Avant révision du corpus, un bon accord Kappa [5] de 0,808 a été obtenu par comparaison des annotations communes. Cet accord ne peut nous en apprendre beaucoup sur la précision du guide, parce que les annotateurs sont les auteurs et d'autant plus parce que le guide a été modifié au fur et à mesure de la phase d'annotation et des difficultés

5. Nous avons utilisé le cœfficient Kappa décrit dans (Cohen, 1960). Cette mesure compare les accords en rapport avec ce qui serait obtenu par le hasard. Selon (Landis and Koch, 1977), de 0,6 à 0,8 l'accord est considéré comme bon. Au delà de 0,8, il est très bon. (cf. Annexe B.2)

rencontrées. Pour avancer que le guide est bien construit, il faudrait que des textes soient annotés par des annotateurs indépendants (par des personnes qui ne sont pas auteurs du guide d'annotation), ce qui n'a pas encore été le cas.

3.2.2 Taille de notre corpus par rapport aux autres corpus existants

Dans le tableau 3.2, nous présentons les différents chiffres constitutifs des corpus d'événements, auxquels nous avons eu accès. Tous ces corpus ont tous été annotés manuellement et la plupart d'entre eux n'a pas été développé uniquement pour les événements nominaux. Ce tableau est une reprise du tableau 1.3, où la première partie du tableau représente les corpus *TimeML* dans différentes langues et la deuxième est constituée de notre corpus et celui de Creswell qui ont été annotés uniquement en événements nominaux. Des corpus *TimeML*, il est possible de récupérer uniquement les noms d'événements qui ne sont pas des états et qui semblent correspondre à notre définition des événements.

Corpus	Langue	nb. total d'événements	nb. noms d'événements	
Corpus annotés selon *TimeML*				
TimeBank 1.2	anglais	7 935	1 792	
FR-TimeBank	français	2 100	663	(Bittar, 2010a)
IT-TimeBank	italien	8 138	3 695	(Russo *et al.*, 2011)
	espagnol	1 677	?	(Wonserver *et al.*, 2012)
TimeBankPT	portugais	7 887	?	(Costa and Branco, 2012)
Ro-TimeBank	roumain	7 926	2 350	(Forascu and Tufis, 2012)
Corpus d'événements nominaux autre				
Creswell_2006	anglais	–	1 579	(Creswell *et al.*, 2006)
Notre corpus	français	–	1 844	(Arnulphy *et al.*, 2012b)

TABLE 3.2 – Comparaison des tailles de corpus en nombre de noms d'événements en fonction du type de représentation temporelle dans les différentes langues. Le nombre de noms d'événement pour le *TimeBank1.2* correspond aux événements nominaux hors états (*non-stative nominal events*).

Dans les corpus TimeBank, une distinction est faite entre les noms actifs et les noms statiques, il est ainsi assez simple de distinguer les noms d'événements qui ne sont pas des états et les autres. C'est par ce biais que nous pouvons comparer les annotations *TimeML* et les nôtres.

3.2.3 Comparaison de notre annotation avec celle du *FR-TimeBank*

Comme dit précédemment, une partie de notre corpus manuellement annotée est constituée de tous les textes du *FR-TimeBank*. Il s'agit d'articles de presse de l'Est Républicain, soit le même type de texte que tous les autres que nous avons annoté :

des articles de presse écrite de quotidiens français, l'un national (Le Monde) et l'autre régional (L'Est Républicain), mais qui traitent d'informations régionales, nationales et internationales. Nous avons annoté ces textes en même temps que ceux issus du journal Le Monde. Annoter des textes du *FR-TimeBank* avait pour but d'évaluer par comparaison chiffrée notre type d'annotation avec l'annotation des noms en *TimeML* dans sa version française (Bittar, 2010b).

Dans les schémas d'annotation *TimeML*, seules les têtes de syntagmes sont annotées. Ainsi pour la comparaison de notre annotation avec celle les annotations du corpus *FR-TimeBank*, nous avons considéré uniquement les têtes de syntagmes nominaux que nous avions annoté. Dans notre version, 737 noms ont été annotés événement contre 663 dans *FR-TimeBank*. Et 79,8 % des têtes de syntagmes nominaux étiquetés comme des événements dans notre version du corpus FR-TimeBank sont communs avec la version du *FR-TimeBank*. L'accord inter-annotateur calculé entre les deux annotations est considéré comme bon : cœfficient Kappa d'une valeur de 0,704.

Le but de ce corpus est d'annoter les syntagmes nominaux utilisés pour nommer les événements. Notre corpus est d'une taille tout à fait comparable aux autres corpus annotés en noms d'événements des autres langues, et il est d'une plus grande taille que le *FR-TimeBank*. Annoter manuellement un corpus nous a permis de nous rendre compte de manière plus précise du comportement des noms en lecture événementielle.

3.3 Comportement des noms en lecture événementielle

Nous avons observé le comportement des noms d'événements annotés[6]. D'abord en terme de morphologie en nous interrogeant sur la formation des dénominations utilisées, puis en terme sémantique en observant les ambiguïté révélés lors de l'annotation et étudiant le comportement des noms annotés dans le corpus lorsqu'ils sont utilisés comme lexique pour l'extraction et enfin d'un point de vue grammatical en tâchant de vérifier les assertionsgénérales au sujet de l'utilisation par les noms d'événements des pluriels et des déterminants.

6. Cette étude a fait l'objet de (Arnulphy *et al.*, 2011b), (Arnulphy, à paraître)

3.3.1 Processus de composition des noms d'événements

Au chapitre 2 (section 2.3), nous avons défini que la formation des noms d'événements pouvait essentiellement suivre trois processus : la nominalisation apparentée à un verbe d'action, la nominalisation autre que dérivée de verbes et la nominalisation métonymique. Nous avons observé la formation des noms de notre corpus manuellement annoté et comme nous pouvions nous y attendre, les noms d'événements sont principalement formés sur une nominalisation apparentée à un verbe. Ainsi, 67 % des événements nominaux du corpus appartiennent à la première catégorie (nominalisation/verbes d'action), 32 % à la deuxième (nominalisation/non verbes), et seulement 7 occurrences ont un caractère événementiel par métonymie.

3.3.2 Ambiguïté des noms d'événements

Le corpus contient 725 occurrences différentes parmi les 1844 annotations. Parmi ces événements, 273 n'apparaissent qu'une seule fois. Sur l'ensemble des 452 noms d'événements annotés plus d'une fois comme événement, seuls 31 % ont une lecture événementielle à chacune de leur occurrence, les autres sont ambigus. La figure 3.1 donne plus de chiffres sur la proportion des lectures événementielles par rapport au nombre total d'occurrences d'un mot. On y remarque notamment que pour 29 noms du corpus, la proportion entre les occurrences où ils désignent un événement et le nombre total d'occurrences est de moins de 10 %. Pour 129 noms, ce rapport est inférieur à 50 %, et pour 312 noms, ce rapport est de moins de 100 % (soit au plus 99 % de leurs occurrences).

Quelques exemples en fonction du taux d'événementialité des mots extraits sont présentés dans la figure 3.2 et des exemples de mots accompagnés de leur taux d'événementialité dans le tableau 3.3.

Pour plus de clarté, nous avons expliqué l'utilisation de ces mots au moyen d'exemples en contexte :

- « Disparition », « meurtre » et « démission » sont dans notre corpus toujours événement.
- « Campagne » ou « peine » sont événement dans 70 à 99 % des occurrences.

  (3.19) Un juge fédéral casse le verdict condamnant Mumia Abu Jamal à la <event> *peine de mort* </event>. (verdict)

  (3.20) respecter l'esprit de notre législation française qui n'inclut pas la peine$_{(abstrait)}$ de mort " (sanction)

- « Commentaire », « signe », « prescription » et « bombe » sont entre 40 et 69 %.

  (3.21) À l'occasion du <event> *54ème anniversaire des* <event> *bombes*

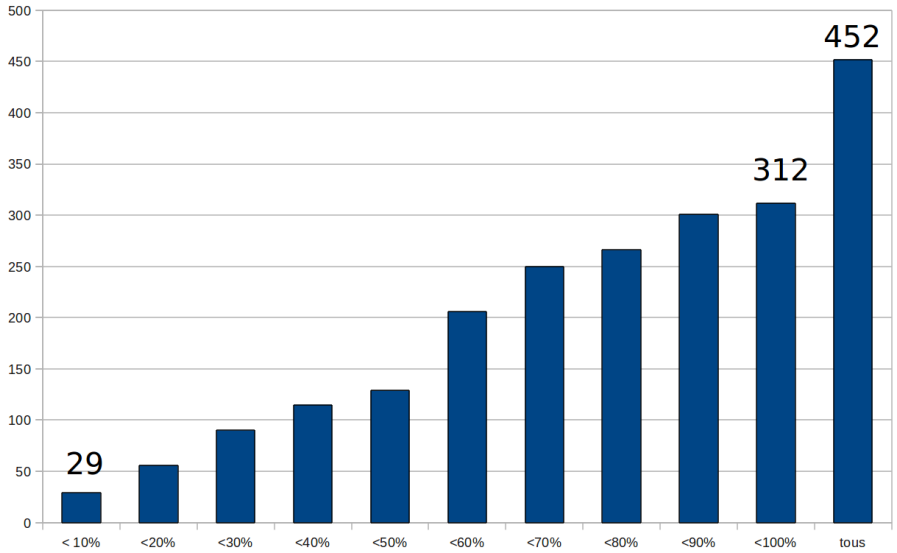


FIGURE 3.1 – Progression du nombre de noms qui ont une lecture événementielle par rapport au nombre d'occurrences totales de chaque nom. Par exemple, 29 noms ont une lecture événementielle dans au moins 10 % de leurs occurrences, alors que 312 noms ont une lecture événementielle dans au plus 99 % de leurs occurrences.

</event> ***de Hiroshima et Nagasaki*** </event> (par métonymie)

(3.22) `Une bombe`$_{(obj)}$ `a explosé hier.`

– « `Prix` », « `mort` », « `conseil` » ou « `triathlon` » sont à moins de 40 %.

(3.23) `Dans ce` <event> ***grand prix*** </event> `au millimètre, seules neuf voitures ont passé la ligne d'` <event> ***arrivée*** </event>. (événement sportif)

(3.24) `ce produit décal [... ] est distribué depuis septembre au prix`$_{(montant)}$ `de 15,90 francs`

(3.25) `Jusqu'à la` <event> ***mort*** </event>. (passage de la vie au trépas)

(3.26) `Un mort`$_{(pers)}$ `dans un dépassement.`

Au moyen de ces exemples, on se rend compte de manière plus précise de l'ambiguïté causée par la polysémie des mots : « `mort` » désigne l'événement, le processus, mais aussi la personne ou le résultat de cet événement. De plus, certains mots se révèlent dans notre corpus plus événementiels que d'autres : « `campagne` » désignant le lieu ou le paysage rural, aussi bien que l'activité politique ou commerciale ; ce mot est plus souvent événementiel (plus de 70 %) que « `conseil` » qui désigne le groupe de personne ou la séance tenu par ce groupe de personnes (moins de 40 %). Il faut ici remarquer que la nature journalistique

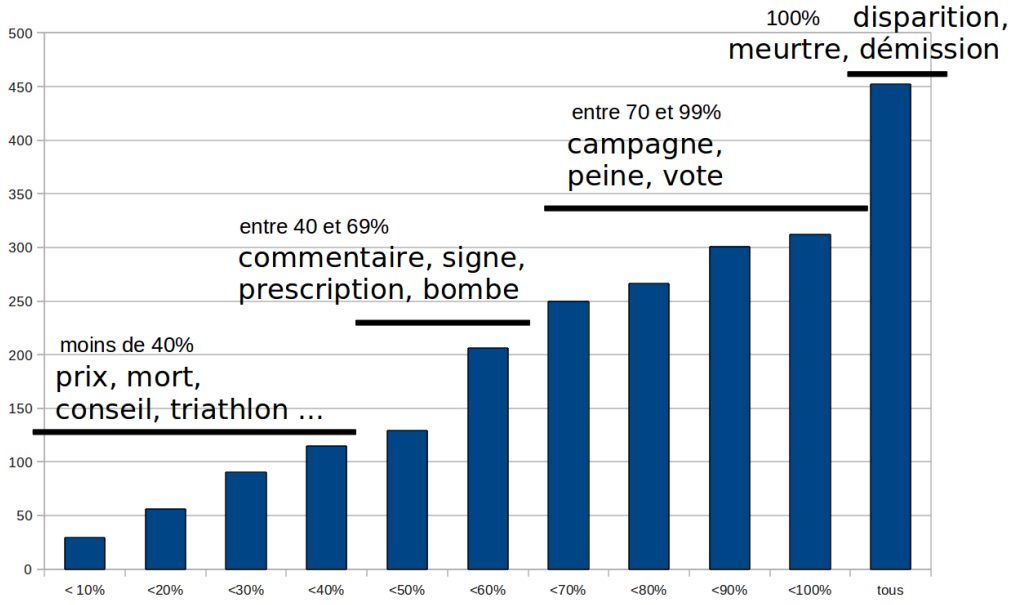


FIGURE 3.2 – Exemples de mots événement en fonction du taux d'événementialité.

	Taux d'événementialité
disparition	100 %
meurtre	100 %
démission	100 %
campagne	88,0 %
peine	88,2 %
vote	80,0 %
commentaire	66,7 %
bombe	50,0 %
signe	44,4 %
mort	37,5 %
prix	22,2 %
conseil	10,7 %

TABLE 3.3 – Exemples de noms ayant quelques fois (moins de 100 %) ou toujours (100 %) une lecture événementielle, accompagnés de leur taux d'événementialité en corpus : rapport entre leur nombre d'occurrence en lecture événementielle et leur nombre d'occurrence lorsqu'ils ne représentent pas des événements.

du corpus a sans aucun doute une forte influence sur les scores de certains mots, et que des proportions différentes seraient probablement obtenues dans d'autres genres textuels.

3.3.3 Utilisation des têtes de syntagmes comme un lexique

Dans le but de montrer la difficulté d'utiliser uniquement des lexiques pour extraire automatiquement des noms d'évenements, nous avons construit un lexique à partir des têtes de syntagmes annotés événement dans le corpus. Deux listes ont été extraites :

1. La liste des noms qui ont toujours une lecture événementielle, soit à 100% de leurs occurrences dans le corpus : LEX_{sure_only}.
2. La liste des noms qui ont une lecture événementielle au moins une fois parmi leurs occurrences dans le corpus : LEX_{all}.

3.3.3.1 Sur le corpus entier

Nous avons appliqué chacun des lexiques (LEX_{sure_only} et LEX_{all}) sur le corpus entier, et indépendamment sur les sources de textes annotés (Le Monde – LM et L'Est Républicain – ER). Vu comment sont construits les lexiques, il est évident que la précision vaut 1 pour LEX_{sure_only} et que le rappel vaut 1 pour LEX_{all}. Tous les résultats sont donnés dans le tableau 3.4.

	P	R	F
LM	1,000	0,434	0,605
ER	1,000	0,403	0,574
Total	**1,000**	**0,421**	**0,593**

a. LEX_{sure_only}

	P	R	F
LM	0,322	1,000	0,487
ER	0,385	1,000	0,556
Total	**0,345**	**1,000**	**0,513**

b. LEX_{all}

TABLE 3.4 – Résultats des lexiques LEX_{sure_only} (a.) et LEX_{all} (b.) construits sur le taux d'événementialité des occurrences de mots annotés manuellement appliqués sur le corpus entier

Loin de prendre ces valeurs pour des résultats réels d'une extraction automatique, étant donné que le même corpus a été utilisé pour extraire les lexiques et pour les tester, nous nous contentons d'établir les valeurs maximales qui pourraient être obtenues en utilisant uniquement les lexiques « les plus complets et parfaits ».

Pour autant, ces statistiques ont le mérite de confirmer les conclusions annoncées par le graphique 3.1 : les lexiques même s'ils sont utiles sont loin de suffire pour une tâche d'extraction automatique des noms d'événements. En effet, le rappel pour LEX_{sure_only} est seulement de 42 %, ce qui signifie que 58 % des occurrences sont ambiguës.

3.3.3.2 Sur un corpus de test

Nous avons mené une expérience comparable, en extrayant les lexiques à partir de 75 % du corpus annoté, ce qui nous donne deux nouvelles listes de mots $LEX^{dev}_{sure\_only}$ et LEX^{dev}_{all}. Ces lexiques ont ensuite été appliqués sur le quart restant de documents, qui sert de test. La même répartition des documents a été appliquée pour chaque lexique (cf. tableau 3.5).

		LM	*ER*	*Total*
Ensemble de développement	Textes	62	81	143
	Mots	22 408	11 680	34 088
	Évenements	741	532	1 273
Ensemble de test	Textes	21	28	49
	Mots	9 041	4 517	13 558
	Événements	366	205	571

TABLE 3.5 – Répartition des corpus de test et développement pour l'extraction de $LEX^{dev}_{sure\_only}$ et LEX^{dev}_{all}

Par cette nouvelle expérience (dont les résultats sont présentés dans le tableau 3.6), nous ne sommes pas étonnés de constater que par rapport à l'expérience précédente sur le corpus entier, toutes les valeurs correspondantes chutent. De plus, nous observons que le déficit est important (entre 5 et plus de 30 points), ce qui nous montre que les lexiques construits sur le corpus de développement peuvent difficilement être considérés comme représentatifs. Seules deux interprétations nous semblent envisageables :

- soit le corpus est de taille insuffisante,
- soit trop nombreaux sont les noms qui prennent une valeur événementielle dans un contexte très particulier.

Il nous semble donc pour le moins illusoire de pouvoir construire un lexique « complet » pour notre tâche d'extraction de formes nominales de noms d'événements.

	P	R	F
LM	0,797	0,268	0,401
ER	0,776	0,254	0,383
Total	**0,789**	**0,263**	**0,395**

a. $LEX^{dev}_{sure\_only}$

	P	R	F
LM	0,287	0,672	0,402
ER	0,323	0,683	0,439
Total	**0,299**	**0,676**	**0,415**

b. LEX^{dev}_{all}

TABLE 3.6 – Résulats des lexiques $LEX^{dev}_{sure\_only}$ (a.) et LEX^{dev}_{all} (b.) construits sur trois quart du corpus manuellement annoté et appliqués sur le quart restant.

3.3.4 Pluriels et déterminants

Comme l'a fait (Russo *et al.*, 2011) sur l'italien, nous avons souhaité vérifier les deux affirmations généralement maintenues dans la littérature :

– Les pluriels ont rarement une lecture événementielle.

– Les articles définis impliquent généralement une lecture événementielle.

Pour le français, nous arrivons aux même conclusions que Russo, à savoir que ces affirmations ne se vérifient pas, comme le montrent les chiffres des tableaux 3.7 et 3.8, résultats. Les proportions de chaque classe sont en réalité équivalentes dans la répartition événement et non événement. La seule différence notable est liée à la fréquence d'utilisation de l'article indéfini pour les événements, plutôt que l'article défini, ce qui est la démonstration du contraire de ce qui est soutenu par l'intuition générale.

	Noms d'événements	*Tous les noms*
Singulier	0,801	0,834
Pluriel	0,199	0,166

TABLE 3.7 – Nombre d'occurrences des noms suivant leur nombre (singulier ou pluriel).

	Noms d'événements	*Tous les noms*
Articles définis	0,279	0,199
Articles indéfinis	0,143	0,062
Déterminants démonstratifs	0,040	0,017
Déterminants possessifs	0,061	0,033

TABLE 3.8 – Nombre d'occurrences des différents types de déterminants introduisant des noms.

3.3.5 Avec les entités nommées *Quaero*

Une partie du corpus entités nommées *Quaero* a été annotée en événement, à des fins de prospection et d'observation sur la compatibilité entre les autres entités nommées et les événements et dans le but d'étudier les relations à considérer entre les événements et les autres entités. Il s'agit de trois textes issus du corpus « broadcast news », retranscription d'émissions radiophoniques. Tout le corpus commun avec les autres entités nommées de *Quaero* a été annoté par les deux même annotateurs que le corpus annoté en événement et selon les mêmes modalités de révision.

Les noms d'événements sont des syntagmes nominaux qui peuvent inclure d'autres entités nommées étiquetées dans le corpus annoté en entités nommées *Quaero*. Avec les

auteurs du guide entités nommées générales (avec qui nous avons mené cette étude), nous avons observé les doubles annotations sur les étiquettes <kind> des événements : 85 % des noms d'événements (soit 169 occurrences) sont isolés des autres entités nommées ; les 15 % restants (28 occurrences) intègrent ou sont intégrées dans un autre type d'entité nommée, comme le montre le tableau 3.9.

	<loc>	<time>	<pers>	<func>	<org>	total <event>
<kind>	6	6	3	3	4	28/169

TABLE 3.9 – Nombre d'entités nommées de chaque type associé à la tête annoté <kind> d'un événement

Conclusion

Élaborer notre propre corpus en français, nous a permis de tester notre typologie et notre définition sur des textes en langue naturelle, mais aussi de vérifier la compatibilité de notre définition avec celle de *TimeML* (c'est la raison pour laquelle une partie des textes que nous avons annotée est constituée de tous les articles du *FR-TimeBank*).

Par ailleurs, annoter notre propre corpus était le seul moyen d'avoir une annotation en syntagmes, qui convienne à notre démarche liée à la notion d'entités nommées. Enfin, la création de notre corpus nous a permis de nous rendre compte directement de l'étendue du problème : entre les difficultés pour définir les frontières de syntagmes, choisir d'annoter ou non un syntagme, ou encore nous aiguiller dans la voie de l'extraction automatique des noms d'événements.

Notre corpus annoté, que nous considérons maintenant comme notre référence nous servira de manière fiable de corpus d'entraînement et de test dans nos projets d'extraction automatique.

Chapitre 4

Vers la création de lexiques pondérés pour l'extraction d'événements nominaux

De nombreuses approches s'appuient totalement ou en partie sur des lexiques pour repérer les noms d'événements. C'est le cas de (Creswell *et al.*, 2006) qui utilise des termes amorces issus de *WordNet* en anglais, (Peris *et al.*, 2010a) et (Resnik and Bel, 2009) en espagnol. De nombreux mots ont une valeur événementielle évidente, cette approche semble intéressante. Pourtant à bien y regarder de près, nous nous rendons vite compte que la plupart des mots de la langue peut avoir une lecture événementielle en contexte. Au départ du travail décrit dans ce chapitre était la volonté de créer un lexique qui pourrait nous fournir les bénéfices des approches lexicales avec l'information supplémentaire consistant en la valeur événementielle du mot en contexte.

Ce chapitre présente les expérimentations menées afin de proposer la création d'un lexique pondéré en événementialité dédié aux noms d'événements. Ainsi nous avons testé les capacités des lexiques existant déjà et que nous avons jugé standards pour notre tâche (en français, *VerbAction*) et *EventNominals* et en anglais, les mots sélectionnés dans *WordNet*). Nous avons décrit des règles contextuelles d'extraction de noms d'événements, nous les avons testées et les meilleures ont été retenues pour nous permettre d'extraire notre nouveau lexique pondéré. Et enfin, nous avons testé les performances de notre nouvelle ressource.

Comme nous l'avions fait dans les études précédemment présentées dans ce manuscrit, nos mesures d'évaluations sont la précision (P), le rappel (R) et la F-mesure (F) (cf. annexe B.1 pour plus de détail).

4.1 Les lexiques existants

Dans le but d'estimer et d'évaluer la possibilité d'utiliser les lexiques existants et compte tenu de nos expérimentations et des observations décrites dans le chapitre 3, nous avons testé une extraction automatique par approche lexicale sur les corpus manuellement annotés. Nous avons utilisé pour seule ressource les lexiques existants (présentés en section 1.2.3.2). Pour le français, notre corpus manuellement annoté a servi de référence, *VerbAction* (section1.2.3.2) et *EventNominals* (section 1.2.3.3) sont les lexiques évalués pour l'extraction de noms d'événements. Pour l'anglais, c'est *TimeBank1.2* qui est le corpus et la liste des noms étiquetés action ou événement dans *WordNet* (section 1.2.3.1) fait office de lexique à évaluer.

Nous proposons ici nos observations sur les lexiques existants dont nous avons fait usage en français et en anglais. Nous appellerons dorénavant lexiques standards, les lexiques qui existaient déjà au moment où nous avons commencé à travailler sur la question des noms d'événements. Nous avons utilisé l'analyseur syntaxique *XIP* pour mener à bien nos expérimentations sur les lexiques (pour plus de détails sur l'outil et l'utilisation que nous en avons faite, se référer à l'annexe C.3).

4.1.1 Les lexiques standards en français : *VerbAction* et *EventNominals*

VerbAction est une liste de couples verbe-nom morphologiquement apparentés. Ce lexique n'a pas été spécialement créé pour l'extraction d'événements nominaux, mais les noms communs présents dans ce lexique étant apparentés à des verbes d'actions, ils sont assimilables à des noms d'actions et donc potentiellement font-ils partie des noms que nous annoterions en tant qu'événements. Par opposition, *EventNominals* a été créé spécifiquement pour l'annotation des événements nominaux selon les spécifications *TimeML*. Ce lexique est décrit comme le lexique alternatif des noms événementiels en français. Il s'agit d'une liste de noms communs amorces de noms d'événements. Il a été utilisé comme indice pour les annotateurs humains. Ce lexique se veut clairement complémentaire au *VerbAction* dans le cadre de l'annotation des noms d'événements.

Comme nous l'avons déjà expliqué, avec ces lexiques, nous sommes confrontés à deux difficultés majeures : celle de l'ambiguïté et celle des mots qui prennent leur caractère événementiel en contexte.

– L'ambiguïté, d'une part, nous est clairement révélée parce que les noms peuvent désigner l'événement, son résultat ou ce qui le sert. Composer un lexique, en particulier de noms non issus d'une conversion verbe/nom, conduit donc à des choix subjectifs dirigés par un nécessaire compromis entre bruit et silence.
– Les mots autres que polysémiques qui prennent leur caractère événementiel en contexte, ne peuvent figurer dans des listes figées de noms d'événements. Par essence, ce sont des groupes nominaux qui n'ont aucun caractère événementiel dans leur formation, ni dans les mots qui les forment. C'est le cas notamment des toponymes, des héméronymes et autres expressions métonymiques en contexte.

Au vu de ces observations, notre constat est que les lexiques seuls ne pourraient suffire pour une extraction de termes événementiels en contexte. Ce que nous parvenons à démontrer par une évaluation des lexiques standards en français.

Évaluation des lexiques français Dans cet exercice, les deux lexiques français (*VerbAction* et *EventNominals*) sont utilisés pour annoter les textes. Nous ne marquons comme événements que les mots du lexique qui ont le même lemme et qui sont de nature nominale. Nous avons appliqué les lexiques sur notre corpus manuellement annoté : une fois avec *VerbAction* uniquement et une fois *VerbAction* et *EventNominals* en même temps, parce que *EventNominals* n'a pas vocation à être utilisé seul, mais il est par définition complémentaire au *VerbAction* pour l'extraction des événements nominaux. Les résultats sur notre corpus annoté sont présentés dans le tableau 4.1.

	P	R	F
LM	0,437	0,651	0,52
ER	0,583	0,691	0,63
LM + ER	**0,487**	**0,668**	**0,56**

a. avec *VerbAction* uniquement

	P	R	F
LM	0,439	0,843	0,58
ER	0,570	0,840	0,68
LM + ER	**0,483**	**0,841**	**0,61**

b. avec *VerbAction* et *EventNominals*

TABLE 4.1 – Résultats d'application des lexiques *VerbAction* et *EventNominals* sur notre corpus manuellement annoté, sur sa globalité et en tenant compte de la source du corpus (*Le Monde* (LM) ou *L'Est Républicain* (ER))

Nous observons tout d'abord que pour les différents corpus (*LM* pour *Le Monde* ou *ER* pour *L'Est Républicain*) les résultats en termes de précision et de rappel sont homogènes. *VerbAction*, appliqué seul, obtient 0,487 de précision sur notre corpus entier ce qui nous montre que les déverbaux ont plus souvent une lecture non-événementielle. Le rappel s'élevant pour ce lexique à 0,668, nous pouvons conclure qu'environ un tiers des noms d'événements de notre corpus ne sont pas construits à partir de déverbaux. Lors de la combinaison du *VerbAction* avec le lexique alternatif *EventNominals*, le rappel est augmenté (de R = 0,668 à R = 0,841) sans affecter la précision (de P = 0,487 à P = 0,48).

Mais 15 % des noms d'événements sont toujours absents et la précision est plutôt basse.

Nous pouvons d'ors et déjà comparer ces résultats obtenus au moyen des lexiques à ceux obtenus par (Parent *et al.*, 2008) (P = 0,547, R = 0,537, F = 0,542) sur la catégorie morphosyntaxique « nom », dans leurs travaux d'extraction automatique de données *TimeML*.

En comparaison avec le test VerbAction *uniquement*, (Parent *et al.*, 2008) présente une meilleure précision (P = 0,487 pour nous contre P = 0,547), un rappel moins élevé (R = 0,668 pour nos travaux contre R = 0,537), par contre les performances en termes de F-mesure restent proches. Nous en concluons que les améliorations apportées dans (Parent *et al.*, 2008) sur le *VerbAction* ne sont en définitive pas si significatives que cela (élimination des items indésirables, règles syntaxiques pour annoter les noms qui dépendent d'une préposition temporelle annotée en <SIGNAL> dans *TimeML*, et suppression des noms potentiellement événements (dans le *VerbAction*) qui sont complément du nom).

En comparaison avec le test *VerbAction et EventNominals*, les performances obtenues par ces tests sont globalement meilleures que celles dans (Parent *et al.*, 2008) : une précision qui reste plus basse (P = 0,483 contre P = 0,547 pour eux), notre rappel qui est encore supérieur (R = 0,841 contre R = 0,537) et une valeur de F-mesure d'environ 7 points supérieure en comparant nos deux systèmes (F = 0,61 contre F = 0,542). Nous n'utilisons dans ce test que l'indice d'appartenance ou non au lexique issu de *VerbAction* et de *EventNominals*, ces autres travaux n'utilisent qu'une version de *VerbAction* filtrée par les auteurs et aussi une règle morphosyntaxique qui écarte toutes les occurrences du *VerbAction* qui sont en position complément du nom. Nous en concluons que le lexique *EventNominals* améliore relativement les résultats obtenus.

4.1.2 Les noms d'actions et d'événements issus de *WordNet* en anglais

Le réseau sémantique des noms de *WordNet* (dans sa version anglophone) contient 5 903 noms étiquetés « act » (les noms d'action) ou « event » (ces noms d'événements sont au nombre de 40 seulement). Cette liste issue de *WordNet* contient des noms décrivant des événements dans la plupart de leurs utilisation (« `war` » (guerre), « `election` » (élection), « `show` » (spectacle), « `carnival` » (carnaval)), des expressions ambigus (« `arts and crafts` » (travaux manuels), « `bet` » (pari), « `coloration` » (teinture/couleur)), des expressions multi-mots (« `a cappella singing` » (le chant a capella)), des noms d'événements (« `Arab-Israeli War` » (la Guerre israelo-arabe de 1948-1949), « `Battle of Britain` » (la Bataille d'Angleterre durant la Seconde Guerre mondiale), « `laser trabecular surgery` » (chirurgie trabéculaire au laser)), mais aussi des mots qui ne

semblent pas correspondre à quelque définition des événements qui soit (« `Attorney General` » (ministre de la justice aux États Unis), « `judo` » (la discipline sportive), « `industry` » (industrie)).

Contrairement aux lexiques français, ce lexique ne contient pas que des noms communs et ces noms sont clairement identifiés comme des noms d'action ou d'événements, de plus il contient des expressions multi-mots et des noms propres. Comme les lexiques français, nous n'avons pas d'information concernant l'ambiguïté événementielle de chaque mot. Pour autant, nous considérons que cette liste de mots est comparable aux lexiques standards que nous avons utilisés dans nos travaux en français.

Évaluation des lexiques anglais Dans cet exercice sur l'anglais, nous avons utilisé les noms issus de *WordNet* qui sont considérés comme des noms d'action et des noms d'événements, en deux listes séparées au départ puis ensemble. Comme pour le français, nous ne marquons comme événements que les mots du lexique qui ont le même lemme et qui sont de nature nominale. Nous avons appliqué les lexiques sur le corpus *TimeBank1.2*. Les résultats sont présentés dans le tableau 4.2.

	P	R	F
Event	0,37	0,163	0,227
Act	0,282	0,567	0,377
Event+Act	**0,28**	0,614	**0,386**

TABLE 4.2 – Résultats d'application des listes de noms extraites de *WordNet* : Évaluation des performances des noms d'action, des noms d'événements et de la combinaison des deux types de noms sur le corpus manuellement annoté *TimeBank1.2*

Nous observons de manière générale que les résultats ne sont pas très engageants quant à l'utilisation de ce type de lexique pour l'extraction en anglais de noms d'événements. En effet, la précision ne dépasse pas 0,37 pour les noms d'événements pour un rappel insignifiant (R = 0,163, ce qui paraît logique étant donné qu'il n'y a que 40 noms étiquetés événements dans *WordNet*). La F-mesure maximum est obtenu par l'utilisation des deux lexiques en même temps, soit F = 0,386.

Même si nous avions l'impression forte que cette liste de noms issue de *WordNet* était comparable aux lexiques utilisés en français, nous observons que les performances en anglais ne sont pas équivalentes à celles sur le français, elles sont bien plus basses (en termes de F-mesure, en français les lexiques atteignent F = 0,61 alors qu'en anglais les lexiques sont à F = 0,386). Pour autant, nous observons que les résultats sont de même type, le rappel est environ deux fois plus élevé que la précision (R = 0,614 et P = 0,28 en anglais et R = 0,841 et P = 0,483 en français).

Nous pouvons comparer les résultats obtenus au moyen du lexique que nous avons extrait de *WordNet* à ceux obtenus par Bethard and Martin (2006) (P = 0,729, R = 0,432, F = 0,543) sur la catégorie morphosyntaxique « nom », dans leurs travaux d'extraction automatique de données *TimeML*. Les performances atteintes sur notre test sont en-dessous des leurs : un rappel supérieur, mais une précision inférieure et une valeur de F-mesure de 15 points plus basse (F = 0,386 contre F = 0,543). Nous n'utilisons dans ce test que l'indice d'appartenance ou non au lexique issu de *WordNet*, ces autres travaux intègrent des indices syntaxiques notamment pour l'élaboration de règles d'extraction.

Nous avons des améliorations à apporter soit par rapport au lexique utilisé, soit par l'utilisation d'autres indices dans le but d'une extraction automatique des noms d'événement plus efficace.

Nous avons donc accès à plusieurs lexiques qui peuvent être utilisés comme piste ou indice pour extraire des noms d'événements. Les noms présents dans ces lexiques sont souvent ambigus et nombreux sont les mots utiles pour l'extraction ou qui sont événements uniquement en contexte qui sont absents de ces listes.

Nous observons dans cette section qu'une approche uniquement lexicale (en utilisant que ces lexiques) ne semble convenir ni en français, ni en anglais. Nous allons essayer dans d'utiliser des critères contextuels pour améliorer ces résultats.

4.2 Nos règles d'extraction

Sur le constat que les lexiques standards posent des problèmes d'ambiguïté, nous pensons qu'ils ne peuvent suffire pour l'extraction automatique des noms d'événements. Alors, nous nous intéressons au contexte d'apparition des noms d'événements. Dans les corpus, nous avons dégagé des contextes particuliers d'apparition de noms d'événements qui nous semblent être des indices forts. Sur cette base, nous avons développé manuellement des règles d'extraction pour relever les occurrences de mots qui apparaissent dans ces contextes événementiels. Nous nous sommes intéressés à deux particularités contextuelles des noms d'événements : ils sont utilisés dans des contextes temporels et des verbes d'événement les introduisent. Nous avons construit des règles contextuelles d'extraction, en utilisant certaines prépositions temporelles et des verbes qui introduisent les événements. Une étude poussée à été conduite pour déterminer quels verbes d'une liste fondée sur l'intuition sont quantitativement en contexte les plus instructifs.

Dans cette section, nous présentons d'abord cette étude sur les verbes d'événements et de cause-conséquence, puis le contenu des règles d'extraction fondées sur les indicateurs temporels et celles sur les verbes avant de les tester sur le corpus manuellement annoté. Pour finir, nous introduisons les règles créées pour l'anglais sur le modèle de celles utilisées en français.

4.2.1 Les verbes d'événement et ceux de cause-conséquence en français

La possibilité d'utiliser des verbes d'événement et de cause/conséquence pour l'extraction des noms d'événements a été étudiée dans (Arnulphy *et al.*, 2010). L'hypothèse que nous avons cherché à vérifier dans ces travaux est que l'utilisation des verbes qui impliquent la cause ou la conséquence pourrait constituer un indice 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. 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.

En effet, 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 l'exemple 4.1, le verbe « `entraîner` » a pour sujet « `la crise économique` » et pour objet « `la famine` ». « `Famine` » est l'événement conséquence de l'autre événement de la phrase, « `crise économique` ». C'est aussi le mode de fonctionnement du verbe « `signer` » dans l'exemple 4.2. Ici, « `signer` » présente deux événements, l'un (« `11 septembre` ») cause de l'autre (« `fin de cette hégémonie sur le reste du monde` »).

(4.1) `La` *`crise économique`* `entraînera la` *`famine dans de nombreux pays sous-développés`*`.`

(4.2) `Le` *`11 septembre`* `signe la` *`fin de cette hégémonie sur le reste du monde`*`.`

Des annotateurs humains ont évalué une liste de verbes définie manuellement et dont nous pensions que les syntagmes en position sujet et/ou argument étaient des événements. 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*[1] (Galibert, 2009). Deux annotateurs (de niveau expert) ont ensuite filtré manuellement les SN extraits pour ne conserver que les groupes en position

1. un outil conçu dans le cadre du projet Ritel (Rosset *et al.*, 2005)

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, 4 345 verbes ont été annotés en une dizaine d'heures, pour un total de 5 016 noms. L'accord inter-annotateur est jugé bon (Kappa = 0,79 (Cohen, 1960), cf. annexe B.2). 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 4.3.a.) ou en argument (tableau 4.3.b.). Bien entendu, certains de ces chiffres sont peu significatifs étant donné leur nombre d'occurrences, comme « `avoir pour origine` » (exemple 4.3) ou « `tirer les leçons de` » (exemple 4.4). Nous avons cependant choisi de les conserver dans cette liste parce qu'ils nous semblent particulièrement pertinents.

(4.3) `[... ] que les` *crises* `aient pour origine des problèmes de défaillance technique, de santé publique, etc.`

(4.4) `Le gouvernement se réunira pour tirer les leçons des` *élections*

Nous avons constaté que de nombreux noms événements du corpus (relevés comme tel par les annotateurs) ne sont pas présents dans les lexiques standards, comme « `conflit` » (exemple 4.5), mais plus difficile à insérer dans une liste « `mise en sourdine` » (exemple 4.6) ni de « `tollé` » (exemple 4.7) ou de « `revers` » (exemple 4.8). Dans l'idéal, ces mots devraient être intégrés dans les lexiques standards.

(4.5) `Le` *conflit Danone* `peut donner naissance à une forme d' alliance entre salariés et consommateurs`

(4.6) `Cette` *élection* `entraînera-t-elle la` *mise en sourdine des intérêts communaux* `?`

(4.7) `[... ] a provoqué un` *tollé chez les organisations amérindiennes*

(4.8) `[... ] subissent un` *cuisant revers*

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

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 entraîné	1	100 %
rendre à	1	100 %
se donner	1	100 %

a. Événement en 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. Événement en position argument

TABLE 4.3 – Présence à 75 % et plus d'un SN désignant un événement en position a. sujet ou b. argument de verbes évalués dans cette étude

d'un verbe sont tous 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 (exemple 4.9). Le tiers a pour argument des conséquences matérielles (exemple 4.10) ou pour sujet des personnes ou assimilés personne (exemple 4.11). On peut aussi noter « donner lieu à » qui 7 fois sur 10 vérifie cette hypothèse (exemple 4.12).

(4.9) Son *arrestation* provoque des *manifestations mi-religieuses, mi-politiques*.

(4.10) *Une autre mini-tornade* a provoqué des dégâts$_{(consq\ \ mat)}$ à Villeneuve-lès-Maguelone.

(4.11) Le Conseil de prévention et de lutte contre le dopage avait provoqué une *petite crise avec l'Union cycliste*.

(4.12) 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 seules dates du corpus représentant des événements ont été repérées lors de cette étude sur les verbes. Il s'agit de « 11 septembre » (exemple 4.13) et « mai 68 » (exemple 4.14). Même si les occurrences sont peu nombreuses, ce résultat est intéressant.

(4.13) le 11 septembre aura précipité une récession

(4.14) Mai 68 a précipité sa disparition

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.

Suivant ces observations sur les verbes, nous les avons utilisés dans des règles de grammaire contextuelles verbales en français d'abord, puis en anglais par transposition des règles et lexiques utilisés.

4.2.2 Contenu des règles d'extraction

Toutes nos règles d'extraction ont été codées selon le formalisme *XIP* et en utilisant l'analyse syntaxique que cet outil nous procure. Des exemples de règles sont présentés en annexe C.3.2. Nous distinguons les règles d'extraction fondées sur des critères temporels et celles utilisant des verbes d'événement ou de cause-conséquence. Ce travail a été fait

pour le français et pour l'anglais. En anglais, sur la base de traduction de texte au départ et d'étude de corpus par l'exemple ensuite, nous avons considéré les indicateurs temporels et les verbes comme c'était le cas en français, mais aussi ce qu'on a appelé des indicateurs de conséquences.

4.2.2.1 Les règles d'extraction temporelles

Les événements sont ancrés dans le temps et les noms donnés à ces événements peuvent être utilisés comme des entités temporelles et être par conséquent introduits par des indicateurs temporels. C'est pourquoi nous nous sommes intéressés à certains introducteurs temporels pour extraire des noms susceptibles d'indiquer des événements.

Les règles d'extraction utilisant les indicateurs temporels sont nommées *règles IT*.

En français, trois types de prépositions ont été utilisées. Elles peuvent indiquer :

- le fait qu'un événement se produise :

  « à l'occasion de », « au moment de »

  (4.15) À Jérusalem, *lors de* la *réunion du gouvernement israélien* [...]

- l'usage référentiel de l'événement :

  « pendant »,« après », « à la suite de », « la veille de »,« le lendemain de »,

  (4.16) La population a été évacuée *avant* l' *arrivée de la lave*.

- un moment interne à l'événement :

  « à l'issue de »,« au commencement de »

  (4.17) Les activistes qu'ils ont libérés *au début de* l' *Intifada* [...]

Certaines de ces prépositions n'indiquent pas toujours le temps : « avant », « après », « au commencement de » sont aussi bien des prépositions temporelles que locatives, alors que « à l'occasion de » ou « la veille de » ont toujours une interprétation temporelle. Les prépositions ambiguës sont à utiliser avec parcimonie, dans des contextes précis.

Nous avons utilisé les prépositions suivantes : « à la suite de », « suite à », « lors de », « à l'occasion de », « au moment de », « au début de », « à la fin de », « à l'issue de », « depuis », « au début de », « à la fin de », « à l'issue de », « au commencement de », « au retour de », « au lendemain de », « au soir de », « au matin de », « à la veille de », « au surlendemain de ». En contexte particulier, nous avons aussi fait usage de « après » et « avant ». Et nous avons utilisé les groupes prépositionnels indiquant la durée suivant le modèle « NOMBRE de NOMS de » et dont les NOMS sont :

« heure », « année », « an », « mois », « période », ou toute « DUREE de » proposé par *XIP*.

En anglais, quelques prépositions et locutions prépositionnelles (avec la possibilité d'y insérer des adjectifs) ont été retenues pour leur non ambiguïté.

- « during » (pendant), « since » (depuis) (lorsque la préposition n'est pas suivie d'une date)

  (4.18) But one local strike *during* the *Normandy invasion year of 1944*, costing 1000 tons of production, took place simply because the miners wanted to get rid of the canteen lady. (Mais une grève locale *lors de* l'*invasion de la Normandie en 1944*, qui a coûté 1000 tonnes de production, a eu lieu tout simplement parce que les mineurs voulaient se débarrasser de la dame de cantine.)

  (4.19) It will be the first time US and Russian troops have soldiered together on a potential combat mission *since* the *Second World War*.(Ce sera la première fois que les troupes américaines et russes combattent ensemble *depuis* la *Seconde Guerre mondiale*.)

- « at the moment of » (au moment de), « at the time of » (lors de), « on the occasion of » (à l'occasion de)

  (4.20) Yet *on the occasion of* the *last fiscal crisis*, the only way found to reduce Government borrowings was to increase the burden of VAT. (Pourtant, *à l'occasion de* la *dernière crise financière*, le seul moyen trouvé pour réduire les emprunts du gouvernement était d'augmenter la TVA.)

  (4.21) I was surprised at the resilience *at the time of* the *flooding*. (J'ai été surpris par la détermination des gens *au moment des inondations*.)

Les indicateurs de conséquence, en marge des indicateurs temporels ont aussi été utilisés. En anglais, certaines expressions verbales ne sont pas traduisibles dans le même type de parties du discours.

- « as the result of », « as a result of » (en conséquence de, en raison de)

  (4.22) Mr Ross of Sea Road, Methil, died at 2.30pm on August 9 at Victoria Hospital, Kirkcaldy, *as a result of* a *blood clot on the lung* and *a massive stroke*. (M. Ross, de Sea Road, Methil, est décédé à 14h30 le 9 août à l'hôpital Victoria, Kirkcaldy, *des suites* d'un *caillot de sang dans le poumon* et d'une *hémorragie cérébrale massive*.)

  (4.23) [...] *as a result of* the *fight by the women's movement*, *as a result of* *major decisions by progressive parties* and *as the result*

of action by enlightened men and women. (*grâce aux luttes du mouvement des femmes*, aux *grandes décisions de partis progressistes* et à *l'action de femmes et d'hommes éclairés*)

– « in the aftermath of » (après, à la suite de, au lendemain de)

(4.24) They also agreed on how to work more collaboratively and to coordinate assistance *in the aftermath of* the *earthquake*. (Les ministres se sont également mis d'accord sur les modalités d'une collaboration accrue dans l'action menée et d'une coordination de l'assistance *au lendemain du séisme*.)

– « aftermath/continuation » (suite/conséquence) dans leur formation complément du nom. Le complémenté est un événement. « the typhoon's aftermath » correspond aux conséquences de l'événement typhon.

(4.25) In the capital Manila the *typhoon*'s *aftermath* raised fears that the country's economy, already staggering from an unexpected surge in inflation, could find the fresh blow unbearable. (À Manille, la capitale, à la suite du typhon, la crainte que l'économie du pays, déjà chancelante d'un gonflement inattendu de l'inflation, pourrait trouver le nouveau coup insupportable, a augmenté.)

4.2.2.2 Les règles d'extraction verbales

Les règles d'extraction utilisant ces verbes de cause-conséquence sont appelées *règles VB90*.

Nous avons finalement considéré trois sortes de verbes :

– les verbes qui introduisent des événements :
« se produire », « avoir lieu »

(4.26) Le *Sommet du G8* est organisé à Deauville.

– les verbes qui présentent des événements comme conséquence ou cause d'autres événements ou d'autre chose :
« occasionner »

(4.27) La *crise économique* **entraînera** la *famine dans les pays sous-développés*.

(4.28) Le *feu* **provoqué** par l' *attaque-suicide*, n'était pas encore éteint que [...]

– les verbes qui présentent un moment de l'événement :
« durer », « commencer »

(4.29) Que le *spectacle* commence !

En français, nous reprenons les 15 verbes ayant une utilisation événementielle à plus de 90% d'après l'étude précédemment présentées (section 4.2.1), en utilisant cette fois une approche syntaxique.

Ces verbes sont :

- En position sujet : « avoir lieu », « se produire », « s'expliquer par », « avoir pour origine », « être entraîné ».
- En position argument : « assister à », « donner lieu à », « inciter à », « occasionner », « se précipiter à », « tirer les conséquences de », « tirer les leçons de ».

En anglais, la distinction des différents types de verbes est moins fine. Ces verbes sont :

- En position sujet : « to begin » (commencer) mais pas « to begin to » (commencer à + infinitif), « to end » (terminer), « to last » (durer), « to happen » (arriver/se prduire), « to occur » (se produire), « to befall » (arriver à/échoir), « to take place » (avoir lieu), « to come about » (arriver comme la conclusion de).

  (4.30) he learned how to prevent himself from registering emotion no matter what *disaster had befallen him* (il a appris à se défendre d'éprouver des émotions, peu importe la catastrophe qui lui était arrivé)

  (4.31) He believed the *revolution would come about* in one of two ways. (Il estimait que la révolution viendrait d'une façon ou d'une autre.)

- En position complément : « to be the result of » (être le résultat de), « to ensure from » (garantir dès), « to perform » (réaliser).

  (4.32) Their lawyers claimed this *was the result of police beatings*. (Leurs avocats affirment que ceci est le résultat de violences policières.)

Notons que les règles sont peu nombreuses et les éléments de lexiques utilisés restreints. Nous ne cherchons pas dans cette démarche à extraire tous les noms d'événements, mais bien à trouver les noms les moins ambigus en utilisant des indices qui apporteraient le moins de bruit possible.

Nous avons étudié certains contextes d'apparition des noms d'événements. Nous avons focalisé notre attention sur les verbes de cause-conséquence et ceux qui introduisent des événements, mais aussi sur des indicateurs temporels les moins ambigus possibles. Nous les utilisons pour extraire des noms en contexte événement. Nous devons tester ces règles sur les corpus annotés afin de vérifier si elles sont suffisamment précises.

4.2.3 Expérimentations sur les règles d'extraction

Dans cette section, nous évaluons les ressources en termes de lexiques standards et de règles d'extraction que nous avons écrites. Nous avons utilisé les règles d'extraction seules, puis en les combinant avec les lexiques. Ces expérimentations ont été menées sur le français uniquement pour le moment et au moyen de *XIP*.

4.2.3.1 Utilisation des règles d'extraction (sans lexique)

Nous avons utilisé uniquement les règles de grammaire contextuelles précédemment présentées pour extraire les noms « candidats » proposés par nos règles d'extraction dans notre corpus manuellement annoté. Nous avons détaillé les résultats suivent les types de règles : temporel et verbal (en s'intéressant aux verbes dont la précision pour les événements est de 75 % et ceux à 90 % d'après l'étude manuelle précédente).

Le tableau 4.4 donne les résultats de l'expérimentation sur les règles (cf. section 4.2), implémentées au moyen de *XIP*.

	P	R	F
IT	0,812	**0,061**	0,11
$VB90$	**0,840**	0,011	0,02
$VB90 + IT$	**0,816**	**0,072**	0,13
$VB75$	0,682	0,032	0,06
$VB75 + IT$	0,762	0,092	0,16

TABLE 4.4 – Performances de l'extraction des noms d'événements annotés de notre corpus manuellement annoté par les règles d'extraction temporelles (IT), verbales utilisant les verbes qui introduisent des événements à une précision supérieure ou égale à 75 % ($VB75$) et 90 % ($VB90$).

Étant donné que les règles sont extrêmement restrictives et les verbes utilisés des déclencheurs sûrs à 90 %, il n'est pas étonnant que contrairement à l'approche motivée par les lexiques, les règles d'extraction ont une bonne précision (toujours supérieure à P = 0,80) et un mauvais rappel (entre R = 0,011 pour les règles à base de $VB90$ suivant leur configuration événementielle préférée et R = 0,072 pour la combinaison IT et $VB90$). À titre de comparaison, nous avons mené la même expérience en utilisant les verbes qui ont présenté des groupes nominaux événementiels à plus de 75 %. Nous pouvons nous rendre compte que les résultats ne sont pas meilleurs : un rappel de 2 points plus élevé pour une précision de 16 points de moins pour les verbes seuls et 5 de moins pour la combinaison avec les règles IT.

4.2.3.2 Combinaison des règles et des lexiques

Comme nous le voyons, l'utilisation des lexiques et des règles est *a priori* destinée à deux buts distincts. Les règles, appliquées sur le corpus de test, n'ont pas vocation à extraire de nouveaux mots qui seraient absents des lexiques, mais plutôt à confirmer qu'un mot appartenant à un lexique est bien un événément. Ceci est confirmé par le fait qu'ajouter les règles aux lexiques dans la phase de test n'améliore que de très peu le rappel (cf. les résultats du tableau 4.1). Les performances en termes de précision passent de P = 0,483 à P = 0,48, sans et avec les règles d'extraction et de R = 0,841 à R = 0,859 pour le rappel. Cependant, comme nous allons le voir à la section suivante, ces règles peuvent permettre la constitution d'un lexique enrichi et pondéré, lorsqu'elles sont utilisées sur un nombre élevé de documents.

Entre le test avec le *VerbAction* utilisé seul (1570 occurrences correctes sur 3206 totales) et celui des deux lexiques (1582 correctes et 3295 totales), il n'y a pas de grande différence dans les chiffres de précision et de rappel. On reste à moins de 50 % de précision et 85,5 % de rappel et une F-mesure de 0,6. Le lexique *EventNominals* n'apporte rien dans le cadre de cette expérience. Les règles sur les verbes et indicateurs temporels semblent récupérer les événements amorcés par les mots de ce lexique.

4.3 Extraction automatique du lexique

Au vu des résultats (P $>$ 80 % et R $<$ 10 %) obtenus par l'utilisation des règles en français, nous avons entrepris de constituer un lexique de façon automatique en utilisant nos règles d'extraction et en les appliquant sur un corpus de grande taille (nous avons évalué la taille du corpus nécessaire en section 4.3.2.3).

Au moyen de nos règles d'extraction, nous avons donc constitué de nouveaux lexiques fondés sur l'apparition en contexte des noms d'événements. Ces lexiques sont de nature différente que les lexiques standards parce que contrairement aux autres lexiques, nos lexiques ne contiennent pas que des noms communs (comme les noms du *VerbAction*), qu'ils ne sont pas de registre particulier (comme *EventNominals*). De plus du fait du mode de construction du lexique, nous pouvons l'enrichir au moyen de la pondération ou poids relatif d'événementialité. Nous nommons cette valeur *Eventiveness Relative Weight* (*ERW*). Cette pondération est une valeur relative d'apparition des noms en position événementielle (selon nos règles d'extraction précédemment décrites). Il est calculé comme un poids pour chaque nom extrait au moins deux fois par les règles d'extraction. $ERW(w)$ est le nombre d'occurrences $e(w)$ du mot w extrait par les règles, divisé par le total

d'occurrences de ce mot dans le corpus $t(w)$:

$$ERW(w) = \frac{e(w)}{t(w)} \tag{4.33}$$

Nous obtenons une valeur d'ERW qui, si elle ne représente bien sûr pas une proportion ou une probabilité de l'usage événementiel du mot (les règles conduisant à un faible rappel), permet, en comparaison des valeurs d'ERW de l'ensemble des mots, d'estimer son degré d'ambiguïté relatif. Ceci nous permet en quelque sorte de prédire à quel point un nom est événement par rapport à d'autres.

Nous avons extraits plusieurs lexiques en français et en anglais. En effet devant le potentiel des règles en français, nous avons entrepris leur traduction en anglais. Dans cette section, nous présentons les lexiques pondérés extraits de manière automatique à partir de nos règles d'extraction et évaluons les performances d'extraction des lexiques pondérés par rapport aux lexiques standards en application directe du lexique pondéré par seuil (arbitraire de valeur d'ERW) sur le corpus manuellement annoté, dans une évaluation au moyen de méthode par apprentissage et nous évaluons aussi la taille de corpus nécessaire à l'élaboration d'un lexique pondéré de qualité.

4.3.1 Plusieurs corpus, plusieurs lexiques

Nous avons créé trois lexiques pondérés, un par corpus. Les caractéristiques des corpus sont présentées dans le tableau 4.5.

Corpus utilisé pour la création de lexique		Nb. de tokens total	noms	différents	Nb. de lemmes dans le lexique pondéré	
(FR)	LM (2001-2002)	61 920 573	19 767	4 843	1 559	(32,1 %)
(FR)	AFP (2005-2011)	390 654 810	166 077	8 053	3 538	(43,9 %)
(EN)	AFP (2004-2011)	426 710 726	543 394	14 619	3 452	(23 %)

TABLE 4.5 – Des corpus aux lexiques : taille en nombre de tokens

Nous avons utilisé des corpus de presse contemporaine. En français, nous avons d'abord utilisé un corpus de 120 246 articles issus du journal *Le Monde*, correspondant à deux années de parution du quotidien (2001 et 2002) (Arnulphy, 2011). En français et en anglais, nous avons travaillé sur les corpus de l'*Agence France Presse - AFP* qui nous étaient disponibles (Arnulphy *et al.*, 2012a). L'intérêt de ce corpus réside dans le fait que sur les deux langues qui nous intéressent nous avons accès à des corpus extrêmement similaires et de la même provenance : en anglais, nous disposions de 1,3 millions de textes sur la période 2004-2011 et en français, un million de textes sur 2005-2011. Les deux

corpus en français sont similaires non pas par leur taille, mais parce qu'ils sont tous les deux des corpus de presse, même si différents car l'*AFP* fournit des dépêches, soit des textes courts et *Le Monde* des articles de presse.

Les tableaux 4.6 et 4.7 proposent une vue d'ensemble des types de mots présents dans les lexiques pondérés. Les tableaux sont doubles, la partie a. englobant les lemmes qui sont présents dans les lexiques standards (*VerbAction*, *EventNominals* et celui des noms d'action et d'événement issu de *WordNet*) et la partie b. pour les lemmes étrangers à ces lexiques.

	Nb. détecté par les règles	Nb. total d'occurrences	**ERW**
chute	434	2620	**0,166**
clôture	63	470	**0,134**
élection	1243	9713	**0,128**
guerre	1126	11542	**0,098**
crise	286	6185	**0,046**
expérience	63	2878	**0,022**
tension	16	1595	**0,001**
coopération	5	1631	**0,003**
subvention	2	867	**0,002**

a. Mots appartenant aux lexiques standards

	Nb. détecté par les règles	Nb. total d'occurrences	*ERW*
Anschluss	3	4	**0,750**
méchoui	3	5	**0,600**
krach	20	169	**0,118**
RTT	14	166	**0,084**
demi-finale	35	553	**0,063**
cessez-le-feu	15	440	**0,034**
difficulté	16	3894	**0,004**
accès	9	2828	**0,003**
11 septembre	12	4354	**0,003**

b. Mots absents des lexiques standards

TABLE 4.6 – Exemples de mots collectés par les règles d'extraction sur le corpus LM (2001-2002) en français. Ces mots sont des noms amorces qui pourraient être des tête de syntagmes de noms d'événements.

Les même observations peuvent être effectuées sur les corpus français et anglais. Beaucoup de ces mots sont présents dans les lexiques standards. On voit que les noms peu ou pas du tout ambigus (ceux qui sont toujours des amorces de noms d'événements) ont un *ERW* relativement élevé (supérieur au rappel moyen décrit dans la section précédente).

	traduction française	Nb. détecté par les règles	Nb. total d'occurrences	**ERW**
overthrow	renversement	383	448	**0,855**
intifada	Intifada	7	11	**0,636**
bombardement	bombardement	6	12	**0,500**
testimony	testament	426	13109	**0,032**
sleepover	soirée pyjama	3	27	**0,111**
publication	publication	154	9337	**0,016**
marathon	marathon	52	8070	**0,006**

a. Mots appartenant au lexique existant

	traduction française	Nb. détecté par les règles	Nb. total d'occurrences	**ERW**
play-off	barrages	73	75	**0,973**
breastfeeding	allaitement	3	4	**0,750**
overheat	surchauffe	3	7	**0,428**
stopover	arrêt	372	1345	**0,276**
cross-examination	examen croisé	53	416	**0,127**
distillery	distillerie	4	126	**0,032**
welcome	bienvenue	66	3884	**0,017**
influenza	grippe	37	6019	**0,006**

b. Mots absents des lexiques standards

TABLE 4.7 – Exemples de mots collectés par les règles d'extraction sur le corpus AFP en anglais. Ces mots sont des noms amorces qui pourraient être des tête de syntagmes de noms d'événements.

C'est le cas de « `chute` », « `élection` », « `krach` », « `overthrow` », « `breastfeeding` ». On note que « `clôture` », fortement ambigu dans le cas général, l'est semble-t-il beaucoup moins dans ce type de corpus (la presse), où l'objet est bien moins évoqué que le fait de clore. En revanche, des mots comme « `tension` », « `subvention` » ou « `accès` » sont très ambigus et obtiennent une valeur d'ERW peu élevée. C'est également le cas de la date « `11 septembre` », mais celle-ci est une des seules dates du lexique obtenu, et a de loin le meilleur ERW, parce que même si cette date est souvent utilisé (déjà pour référer à un jour de chaque année), le contexte événementiel est activé plus souvent que les autres dates.

4.3.2 Évaluation

Nos évaluations constistent en la comparaison des performances des modèles obtenus par apprentissage sur le corpus en n'utilisant que les lexiques standards, que les lexiques

pondérés et l'association des deux sortes de lexiques.

4.3.2.1 Application directe du lexique sur la base de seuils

Nous avons appliqué sur le corpus manuellement annoté ce nouveau lexique obtenu automatiquement, pour le comparer aux lexiques *VerbAction* et *EventNominals*. Dans (Arnulphy *et al.*, 2011a), nous avons utilisé différentes « tranches » de valeur d'*ERW* de ce lexique, pour observer l'évolution des performances : tous les mots ayant une valeur de pondération supérieur à 10 %, puis tous ceux ayant un valeur de pondération supérieur à 8 %, 6 %, etc. Les résultats sont présentés dans le tableau 4.8.

Mots dont l'*ERW* est supérieur à	P	R	F
10 %	0,841	0,166	0,28
8 %	0,836	0,243	0,38
6 %	0,798	0,315	0,45
1 %	0,563	0,710	0,63
0,5 %	0,434	0,801	0,56

TABLE 4.8 – Application du lexique automatique par « tranches » de valeur d'*ERW*.

La précision et le rappel évoluent bien entendu de manière opposée (lorsque le lexique est moins sélectif, le rappel augmente et la précision diminue), et la meilleure F-mesure (pour 1 %) est de 0,63, soit une valeur similaire à la F-mesure des deux lexiques *VerbAction* et *EventNominals* combinés (F = 0,61). Nous obtenons donc, automatiquement, un lexique de qualité comparable aux deux lexiques standards composés de façon semi-automatique, en ajoutant en plus l'information du degré d'ambiguïté sur la lecture événementielle des noms.

Il faut noter que l'autre valeur ajoutée de notre lexique pondéré est qu'il peut être extrait du même type de corpus que celui sur lequel on veut le tester. Il est facilement adaptable à d'autres corpus, ce qui n'est pas le cas des lexiques standards ?

4.3.2.2 Évaluation par apprentissage

Méthode d'évaluation. Nous avons appliqué les lexiques pondérés automatiquement constitués sur les corpus manuellement annotés en conduisant une approche par apprentissage. Nous avons ainsi mené une évaluation de nos lexiques. Nous avons utilisé les valeurs de *ERW* comme trait d'apprentissage dans le classifieur à base de règles *J48*, une implémentation de l'algorithme C4.5 (Quinlan, 1993), tel qu'il est implémenté dans le logiciel Weka (Hall *et al.*, 2009).

Le corpus manuellement annoté a été séparé en deux parties : 75 % destiné à servir d'ensemble d'apprentissage et les 25 % restants sont le corpus de test. L'ensemble de test contient le même nombre de noms d'événements que de noms qui ne le sont pas. La catégorie « YES » correspond aux noms annotés <event> et la catégorie « NO » à celle des noms qui ne sont pas des événements. Les chiffres de répartition des corpus d'apprentissage et de test sont présentés dans le tableau 4.9.

	Ensemble d'apprentissage			*Ensemble de test*		
	total	YES	NO	total	YES	NO
Français	5 226	1 263	1 263	2 700	566	2 134
Anglais	2 182	1 092	1 092	3 246	453	2 793

TABLE 4.9 – Nombre de tokens dans les corpus d'apprentissage et de test

Pour chacune des langues, nous avons implémenté trois modèles simples qui nous permettent de montrer les modifications de performances provoquées par l'introduction des valeurs d'*ERW* :

- M_l n'a accès qu'à l'information sur l'appartenance d'un mot à un lexique standard existant :
  (FR) *VerbAction* et *EventNominals*
  (EN) la liste des noms d'action et d'événements de *WordNet*
- M_r utilise seulement l'*ERW*, comme une valeur réelle. Comme nous avons deux lexiques pondérés en français, un modèle est créé sur chacun d'eux. Ils sont nommés :
  - M_r^{LM}, fondé sur le lexique pondéré issu du corpus constitué de deux années de parution du journal *Le Monde*.
  - M_r^{AFP}, sur le lexique pondéré fondé sur le corpus *AFP*.
- M_{rl} a pour traits d'apprentissage à la fois les lexiques standards et nos lexiques pondérés.

Résultats. Les tableaux 4.10 et 4.11 présentent les résultats chiffrés des évaluations sur les lexiques pondérés créés respectivement sur le français et l'anglais et testés sur les corpus manuellement annotés.

Observations. Tout d'abord, en anglais comme en français, nous remarquons que les lexiques pondérés utilisés seuls (M_r) atteignent des résultats assez similaires à ceux des lexiques standards (M_l) :
En français, le lexique pondéré *AFP* (P = 0,55) est plus précis que celui du *Monde* (P = 0,49) et les lexiques standards (P = 0,53) ; Le lexique du *Monde* (R = 0,89) a un rappel équivalent à celui des lexiques standards (R = 0,88) et ils sont meilleurs de celui de l'*AFP* (R = 0,77) ; La F-mesure pour autant est équivalente pour les trois lexiques

	Nos lexiques pondérés en EWR		*Standard*	*Combinaison*	
	M_r^{LM}	M_r^{AFP}	M_l	M_{lr}^{LM}	M_{lr}^{AFP}
P	0,49	0,55	0,53	0,54	**0,60**
R	0,89	0,77	0,88	0,89	0,84
F	**0,63**	**0,64**	**0,66**	**0,67**	**0,70**

TABLE 4.10 – Évaluation des lexiques pondérés extraits sur les corpus **français** *LM* et *AFP* par comparaison aux lexiques standards existants (*VerbAction* et *EventNominals*) en application sur notre corpus manuellement annoté

	Nos lexiques pondérés en ERW	*Standard*	*Combinaison*
	M_r^{AFP}	M_l	M_{lr}^{AFP}
P	0,36	0,30	**0,36**
R	0,71	0,64	0,77
F	0,476	0,414	**0,493**

TABLE 4.11 – Évaluation des lexiques pondérés extraits sur le corpus **anglais** *AFP* par comparaison avec le lexique de noms d'action et d'événements issus de *WordNet* en application sur le corpus *TimeBank1.2*, annoté selon les spécifications *TimeML*

(F = 0,63 pour le lexique *Le Monde*, F = 0,64 pour l'*AFP* et F = 0,66 pour les standards ensemble).
En anglais, les chiffres du lexique pondéré sont légèrement supérieur à ceux du lexique de *WordNet* (P = 0,36→0,30, R = 0,71→0,64 et F = 0,476→0,414).

Les combinaisons des deux sortes de lexiques (M_{rl}) entraînent une faible mais appréciable augmentation de la précision et du rappel. En français, la combinaison des lexiques standards avec le lexique *AFP* est plus performante que celle avec le lexique *Le Monde* (0,70 pour M_{lr}^{AFP} de F-mesure contre 0,67 pour M_{lr}^{LM}). En anglais, la précision est stable entre l'utilisation du lexique pondéré seul et la combinaison avec le lexique issu de *WordNet* (P = 0,36). Par contre, le rappel varie, ce qui entraîne une augmentation des performances au regard de la F-mesure, on passe de F = 0,41 pour les lexiques standards à F = 0,49 pour la combinaison.

À partir de ces observations, nous confirmons que les lexiques pondérés créés de manière automatique sont aussi précis que les lexiques standards et manuellement annotés du français et de l'anglais.

De plus, afin de faire des comparaisons, nous avons appliqué le modèle M_r^{AFP} séparément sur *FR-TimeBank* et sur notre corpus manuellement annoté. Les performances du lexique pondéré *AFP* sont similaires sur les deux corpus annotés, même si ces deux

corpus ne sont pas annotés suivant les mêmes buts ou guide d'annotation. La précision atteint P = 0,56 sur le corpus *FR-TimeBank* et P = 0,55 sur nos annotations. Le rappel est de R = 0,77 sur les deux corpus annotés, et la F-mesure de F = 0,648 et F = 0,642 respectivement. D'ailleurs, nous observons que les résultats sur l'anglais sont plus faibles que ceux sur le français. Toutefois, cette différence ne peut être causée par une qualité du lexique. En effet, le rapport quantitatif entre les lexiques standards (*VerbAction* et *EventNominals* de (**?**) en français et les noms d'action et d'événements de *WordNet* en anglais) et les lexiques pondérés est similaire. Ce qui signifie que la qualité de chacun des lexiques est bien similaire. Notre présomption de départ qu'une traduction étudiée des règles d'extraction du français vers l'anglais devrait suffire est bel et bien vérifiée. Le fait que le lexique pondéré en anglais fait de bien moins bonne performances que ceux en français semble plutôt prouver que le problème de l'extraction des noms événementiels en anglais doit être considéré différemment qu'en français. Sans forcément être une tâche plus difficile dans une langue que dans l'autre, l'approche lexicale ne semble pas convenir à l'anglais. Il serait intéressant de consacrer du temps à analyser ces différences entre anglais et français.

Par ailleurs, nous avons fait nos premiers tests en partant de l'impression que plus le corpus utilisé pour l'extraction est grand, plus représentatif serait le lexique pondéré. Dans la prochaine section, nous creusons cette hypothèse en évaluant la taille du corpus nécessaire pour un lexique extrait représentatif des événements dans les textes ciblés sur une même période.

4.3.2.3 Impact de la taille du corpus sur la qualité du lexique

Étant donné que la précision des grammaires contextuelles d'extraction est bonne, et vu que le rappel est faible, nous avons supposé qu'un corpus de grande taille est nécessaire pour extraire un lexique pondéré représentatif de qualité. Mais la question essentielle était : à quel point ce corps doit-il être grand ? Afin d'y répondre, nous avons créé plusieurs lexiques pondérés à partir de portions de corpus : d'un mois à une année de parutions de l'*AFP*. Nous avons étudié les performances de ces lexiques dans des modèles de type M_r^{AFP} en fonction de la taille du corpus qui a permis de les extraires (cf. tableau 4.12).

La figure 4.1 montre que, en anglais comme en français, le gain en termes de F-mesure pour les modèles entraînés sur les lexiques pondérés issus d'un an de corpus ont des résultats aussi bons que pour les modèles entraînés sur les lexiques pondérés issus des corpus entiers, eu égard au temps d'exploitation et en fonction de la taille du corpus tellement plus grande (entre une année et huit années). La forme de la courbe semble

Lexique créé sur		1 mois 07 2005	6 mois 07-12 2005	1 an 2005	tout le corpus *AFP* 2004-2011
Français	P	0,665	0,539	0,512	0,56
	R	0,303	0,628	0,692	0,77
	F	**0,416**	**0,58**	**0,588**	**0,648**
Anglais	P	0,36	0,31	0,35	0,36
	R	0,35	0,7	0,76	0,71
	F	**0,36**	**0,43**	**0,48**	**0,48**

TABLE 4.12 – Évaluation des modèles de type M_r^{AFP} appris sur les lexiques pondérés et en fonction de la taille du corpus qui a servi à créer ces lexiques

prouver qu'un corpus de taille supérieurs ne devrait pas augmenter significativement les performances des lexiques pondérés.

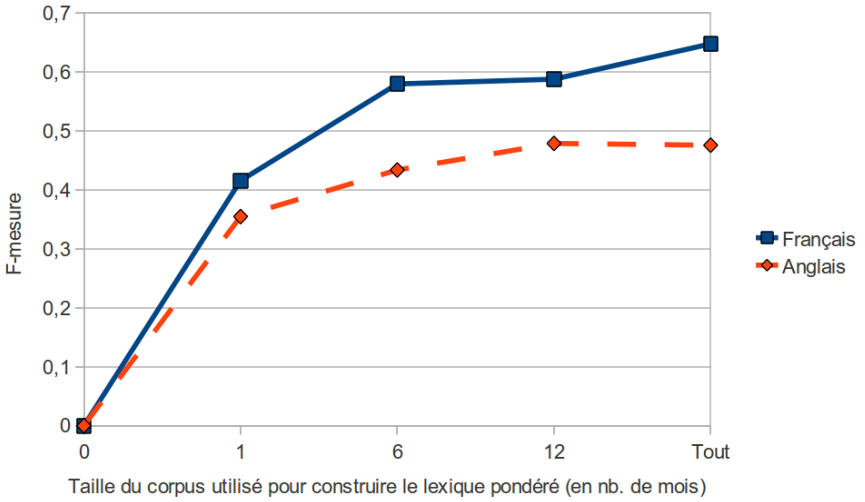


FIGURE 4.1 – Progression de la F-mesure dépendant de la taille du corpus utilisé pour extraire le lexique pondéré

Cependant, même si les performances globales ne sont pas améliorées par un corpus de plus en plus grand, il est toujours intéressant d'extraire les noms d'événements sur de plus longues périodes qu'une seule année. Bien évidemment, il y a un intérêt particulier à mener cette extraction de lexiques d'événements nominaux (et d'amorces de noms d'événements) sur de plus longues périodes et sur des périodes spécifiques, sachant que les événements et leurs dénominations sont ancrés dans le temps et que leurs noms sont parfois utilisés sur de courtes périodes ou à des moments tout à fait particuliers, comme « *tsunami* », « *Printemps arabe* »).

Dans ce chapitre, nous avons rappelé les lexiques standards, qui sont adaptés pour l'extraction des événements. En français, les noms qui sont apparentés à des verbes d'action indiquent en partie des actions et des événements (en partie seulement, parce que bon nombre d'entres eux indiquent aussi le résultat de l'action ou encore l'objet de l'action manifestée par le verbe). Nous avons utilisé le *VerbAction* pour les noms apparentés aux verbes d'action et le lexique *EventNominals*, comme complément. En anglais, nous avons considéré comme un bon indice, les noms étiquetés action et événement dans *WordNet*.

Nous avons proposé une première approche fondée uniquement sur ces lexiques. Plus tard, nous avons proposé une approche plus contextuelle et fondée sur une analyse syntaxique et nous avons comparé les performances de ces deux approches.

Nous avons utilisé des règles de grammaire contextuelles pour l'extraction des noms d'événements en nous fondant sur deux observations. La première concerne la temporalité des événements. Les événements sont ancrés dans le temps, leur manifestation dans les textes sous la forme de dénominations nominales doivent aussi être la plupart du temps temporellement marquées en contexte (d'où les règles temporelles et les règles verbales sur des bases temporelles, qui montrent la durée d'un événement). La seconde observation concerne les verbes dont les arguments sujet ou objet sont des noms d'événements. Nous avons dégagé et évalué une série de verbes soit qui introduisent des événements pour présenter qu'ils ont lieu, ou des verbes de cause-conséquence.

Les résultats étant encourageants, nous avons expérimenté l'extraction d'un nouveau lexique à partir des règles contextuelles que nous avions écrites. Le point fort de nos nouveaux lexiques est la pondération de ces lexiques. En effet, à chaque mot du lexique est associé un score d'événementialité contextuelle des mots qui sont relevés au moins deux fois comme étant des événements potentiels par les règles. Ce score, appelé ERW, est un indice relatif entre les mots.

Plusieurs lexiques pondérés ont été créés, deux en français (sur *Le Monde* et l'*AFP*) et un en anglais (extrait du corpus *AFP*). Nous avons opté pour une évaluation de nos lexiques par une méthode par apprentissage et un test sur notre corpus manuellement annoté en français et sur *TimeBank1.2* en anglais. Notre principal constat est que nos lexiques pondérés construits automatiquement sont aussi performants que les lexiques manuellement obtenus, ce qui implique une diminution du temps passé à valider manuellement des listes de mots (pour nos lexiques, seules les règles sont écrites manuellement). De plus, nos lexiques apportent une information supplémentaire, la pondération. Nous n'avons plus besoin d'évaluer (comme par un seuil binaire) si le mot est événement ou non, la désambiguïsation se fait par le système.

De plus, les lexiques pondérés peuvent être extraits de corpus de types différents, facilement et à moindre coût, ce qui les rend adaptables à tout type de corpus.

Chapitre 5

Extraction automatique

Nous cherchons à reconnaître de manière automatique les noms d'événements suivant la définition que nous avons introduite en section 2.1.3. Comme nous l'avons abordé dans notre état de l'art (chapitre 1), plusieurs systèmes destinés à l'extraction d'informations temporelles et des événements (principalement portés par les verbes) ont vu le jour, mais peu de travaux se sont intéressés particulièrement à l'extraction automatique des dénominations d'événements. Pourtant, comme nous l'avons montré dans (Tannier *et al.*, 2012), les événements importants au bout d'un laps de temps plus ou moins court ne sont souvent plus décrits sous une forme verbale, mais nommés par désignation nominale. Si l'on cherche à remplir par exemple une base de connaissances, il est prépondérant de parvenir à repérer les noms en plus des verbes afin de leur attribuer de manière automatique leurs caractéristiques de temps et de lieu notamment.

Dans ce chapitre, nous proposons une approche à base d'apprentissage automatique pour la tâche de classification des évenements. Dans une première partie nous présentons notre démarche et le choix du classifieur finalement utilisé, ainsi que les traits définis pour l'apprentissage. La deuxième partie de ce chapitre est dédiée aux corpus d'apprentissage et de tests constitués ou développés pour l'apprentissage sur notre tâche. Dans une troisième partie, nous présentons les expérimentations et les résultats obtenus. Pour finir, nous concluons par un bilan et une comparaison aux autres travaux.

5.1 Organisation de la démarche

Nous souhaitons déterminer de manière automatique et pour chaque nom d'un corpus si ce nom désigne en contexte un événement. Cette tâche dont l'extraction automatique

est le but, est une tâche de classification entre deux classes. Elle peut aussi être vue comme une tâche de désambiguïsation sémantique entre événement et non événement.

La démarche que nous avons ici choisi de suivre est une approche par apprentissage automatique en créant des modèles d'apprentissage sur plusieurs corpus et en les testant sur nos corpus manuellement annotés. Nous présentons ici notre choix de classifieur et les traits d'apprentissage retenus pour l'extraction de noms d'événements par apprentissage automatique.

5.1.1 Choix du classifieur

Afin de choisir le meilleur classifieur pour notre tâche de classification des événements, nous avons entrepris de tester plusieurs classifieurs disponibles dans Weka (Hall *et al.*, 2009). Pour cet exercice, nous avons testé plusieurs classifieurs sur les même corpus de développement et de test. En n'utilisant que les valeurs par défaut des classifieurs, nous avons observé les performances des classifieurs développés selon les algorithmes suivants : le modèle naïf bayésien (« NaiveBayes »), les arbres de décision (« J48 », « RandomTree » et « logistic model trees - LMT ») et les classifieurs à base de fonctions (« Logistic », « SimpleLogistic » et « Sequential Minimal Optimization - SMO » pour les SVM). Nous avons choisi ces classifieurs parce qu'ils correspondent au problème d'après la littérature sur les statistiques appliquées au TAL.

Les résultats obtenus par les classifieurs entraînés sont présentés dans le tableau 5.1.

Type	*Classifieur*	*Options*	*P*	*R*	*F*
bayes	NaiveBayes		0,703	0,903	0,79
trees	**J48**	**-C 0.25 -M 2**	**0,815**	**0,827**	**0,821**
...	RandomTree	-K 0 -M 1.0 -S 1	0,734	0,753	0,743
...	LMT	-I -1 -M 15 -W 0.0	-	-	-
functions	Logistic	-R 1.0E-8 -M -1	0,815	0,803	0,809
...	SimpleLogistic	-I 0 -M 500 -H 50 -W 0.0	0,812	0,825	0,819
...	SMO		-	-	-

TABLE 5.1 – Performances de différents types de classifieurs présents dans Weka, sur le même jeu de données.

Deux types de classifieurs n'étaient pas compatibles avec le type de traits de nos données (dont les valaurs sont binaires ou instanciées dans une liste) et ne permettaient donc pas la création de modèles. Il s'agit de « Sequential Minimal Optimization - SMO » et du « logistic model trees - LMT ». En F-mesure, le modèle appris au moyen de l'arbre de décision *J48* obtient les meilleurs performances. Nous avons donc décidé de continuer nos expérimentations en utilisant cette méthode de création de modèles pour l'apprentissage, d'autant que de précédents travaux proches des nôtres ont utilisé le même type de

classifieurs ((Peris *et al.*, 2010a) et (Russo *et al.*, 2011)).

De plus nous avons choisi de développer et de tester nos modèles sans utiliser de validation croisée, contrairement à Bel *et al.* (2010), mais en préférant fournir un corpus de développement et un corpus de test à nos modèles. Même si la validation croisée donne de meilleurs résultats sur les tests que nous avons effectués, nous avons choisi de ne pas utilisé la validation croisée. De plus, certains de nos tests sont incompatibles avec la validation croisée (c'est le cas des expérimentations sur les équilibrages de corpus (explications en section 5.2)). Sans la validation croisée, afin d'obtenir des résultats objectivement comparables les uns avec les autres.

5.1.2 Présentation des traits d'apprentissage considérés

Le but d'extraire des traits pour l'apprentissage est de parvenir à considérer le plus de traits significatifs possibles et qui permettent la description du nom, pour lui même et dans le contexte de la phrase dans laquelle l'occurrence est présente, de par ses relations avec les autres mots de la phrase. Nous avons fait usage de *XIP* (Aït-Mokhtar *et al.*, 2002) (présentation de l'outil en annexe C.3) pour récupérer les traits pour l'apprentissage. *XIP* nous propose une analyse syntaxique, mais aussi un étiquetage en terme d'entités nommées.

Nous avons regroupé les traits d'apprentissage suivant quatre types : syntaxique, sémantique, morphologique et surfacique. Nous proposons dans la section suivante une description des traits retenus pour chaque type. Par défaut, si un trait n'est pas actif dans la description du nom en cours, il est prévu que sa valeur soit nulle.

5.1.2.1 Les traits syntaxiques

Nous avons regroupé les traits syntaxiques suivant la dépendance du trait porté sur notre nom à un verbe ou à un autre nom. Pour ce qui est des relations syntaxiques strictement verbales, **les traits « *sujet* » et « *objet* »** indiquent si le nom en question est le sujet ou l'objet d'un verbe dans la phrase.

Le trait « *préposition* » : donne une indication sur la présence d'une relation entre le nom et le verbe par l'intermédiaire d'une préposition, dans le cadre d'une relation circonstancielle (Cette relation entre un verbe et une préposition dans *XIP* est nommée PREP_OBJ).

Le trait « *est un attribut de* » indique que le nom est dépendant d'un autre nom, **le trait « *a un attribut de* »** indique qu'un autre nom est dépendant du nom que nous sommes en train d'analyser. Ces traits sont déduits en fonction des relations syntaxiques

de *XIP*.

Le trait « *est complément du nom* » désigne un nom qui est rattaché syntaxiquement à un autre nom par la préposition « `de` ».

5.1.2.2 Les traits morphologiques

Le trait « *déterminé* » donne une indication sur le type de déterminant qui détermine le nom en question. Les valeurs possibles pour ce trait (concernant les articles ou les adjectifs) sont définis ou indéfinis, possessifs, numéraux, démonstratifs et quantifieur (ces valeurs sont définies par *XIP*).

Le trait « *suffixe* » prend pour valeur le suffixe du lemme du nom qui nous intéresse, dans une liste des suffixes les plus courants en français. Cette valeur est calculée indépendamment de *XIP* sur la base d'une liste que nous avons constituée.

Le trait « *genre* » a pour valeur féminin ou masculin et **le trait « *nombre* »** indique si le nom est singulier ou pluriel.

5.1.2.3 Les traits sémantiques

Plusieurs sortes de traits sont considérés comme sémantiques. Il y a ceux que nous rattachons à la sémantique générale et ceux que nous rattachons en particulier à des catégorie grammaticales, comme la préposition ou le verbe .

Les traits sémantiques généraux sont constitués des réponses aux questions suivantes : Ce nom est-il un nom propre ? Ce nom appartient-il à une entité nommée ?

Le trait « nom propre » est donné directement par *XIP*, les règles qui indiquent qu'un nom est un nom propre reposent sur des listes de noms de personnes et sur la présence de la capitalisation d'un mot.

Le trait « entité nommée » est aussi fourni par les règles internes de *XIP*. Nous ne considérons que les entités nommées générales (personne, lieu, organisation).

Propriétés sémantiques des verbes : Les traits « *sujet* » et « *objet* », nous indiquent que le nom est par rapport à un un verbe en position sujet ou objet. Nous nous intéressons aussi à certaines propriétés sémantiques du verbe en question. Nous voulons savoir si ce verbe est :

- ***« lié à un verbe du VerbAction »***,
- lié à un verbe de la liste *VB90* des verbes qui introduisent un événement en sujet

ou objet à plus de 90 % de ses occurrences (ces verbes sont présentés au chapitre 4, section 4.2.2.2) (traits « ***sujet d'un verbe de la liste VB90*** », « ***objet d'un verbe de la liste VB90*** »),
- lié à un verbe de la liste *VB* qui est la liste étendue de tous les verbes qui sont moins précis que les *VB90* pour la détection des événements (« ***sujet d'un verbe de la liste VB*** », « ***objet d'un verbe de la liste VB*** »).

Propriétés sémantiques des prépositions : Nous nous intéressons au type de préposition et au type temporel de la préposition.

Le trait « prep_cat » indique le type de préposition : si la préposition est locative, temporelle, si elle correspond à l'une des prépositions les plus courantes et non locatives ni temporelles (que nous avons défini comme étant « `pour` », « `avec` » ou « `de` ») ou autre.

Le trait « prep_cat_temp » correspond au type de préposition temporelle. Nous avons prédéfini des types qui sont liés aux *règles IT* qui nous permettent d'extraire les noms d'événements (pour plus de détail se référer au chapitre 4, section 4.2.2.1). Nous indiquons si la préposition liée syntaxiquement au nom qui nous intéresse indique le fait qu'un événement se produise, l'usage référentiel de l'événement, un moment interne à l'événement.

5.1.2.4 Les traits lexicaux

Les traits lexicaux sont uniquement constitués d'informations lexicales issues de lexiques standards ou des lexiques d'*ERW* que nous avons constitués.

Le trait « *VerbAction* » : Un mot possède le trait *VerbAction* positif lorsque ce mot apartient lexique *VerbAction*.

Le trait « *EventNominals* » : Un mot possède le trait *EventNominals* positif lorsque ce mot appartient au lexique *EventNominals*.

Le trait « *lexique* » : Un mot possède le trait *lexique* positif lorsque ce mot est dans l'un des deux lexiques (*VerbAction* ou *EventNominals*). Il s'agit d'un trait qui combine les valeurs des deux traits précédents.

Le trait « ERW_{AFP} » : Tous les mots qui sont présents dans le lexique pondéré d'ERW [1] constitué sur le corpus AFP et à partir de nos règles d'extraction donnent au trait « ERW_{AFP} » leur valeur d'ERW. Si les mots sont absents de ce lexique pondéré alors, le trait « ERW_{AFP} » est initialisé à 0.

5.1.2.5 Les traits surfaciques

En ce qui concerne les traits surfaciques, nous avons considéré les nombres de lettres des mots dans leur forme de surface et de leur lemme (traits « ***nombre de lettres du lemme*** », « ***nombre de lettres de la forme de surface*** »), un autre indice considéré est la présence ou l'absence de la majuscule dans la forme de surface (**trait « *Capitalisé* »**).

5.1.3 Les regroupements de traits pour les tests

Dans la section précédente, nous avons détaillé le contenu de chaque type de trait et nous les avons présentés. Dans les tests élaborés pour nos expérimentations, nous avons testé chacun des types de traits indépendamment, puis nous avons évalué ces résultats obtenus en enlevant un type de traits différents à l'ensemble des types de traits. Lors de l'étude sur les types de traits utilisés (section 5.3.3), nous avons donc testé le type « syntaxique », le type « sémantique », le type « morphologique », le type « surfacique », et le type « lexical »

Tester chaque type de trait sur nos corpus nous permettra de nous rendre réellement compte de l'impact des traits les plus importants et ceux qui ne le sont pas pour l'extraction des noms d'événements.

5.2 Présentation des corpus utilisés pour l'apprentissage

Nous avons entraîné plusieurs classifieurs. Les premiers modèles ont été développés sur une partie du corpus manuellement annoté (cf. chapitre 3) et testés sur la partie restante de ce même corpus. Plusieurs modèles ont été entraînés en ayant pour corpus de développement des données non équilibrées ou équilibrées en nombre de noms événements ou non-événements et suivant des proportions différentes. Les autres modèles ont été développés sur un corpus de très grande taille créé de manière automatique.

1. L'*Eventiveness Relative Weight* (ERW) est une valeur présentée dans le chapitre 4, à la section 4.3

Nous présentons dans cette section les corpus utilisés pour générer nos classifieurs, compte tenu des répartitions événement-non événement considérées.

5.2.1 Utilisation du corpus manuellement annoté

Pour apprendre nos modèles de classification des noms en événement et non événement, nous avons entrepris d'utiliser les corpus que nous avions manuellement annotés. Les corpus d'apprentissage sont constitués des articles du journal *Le Monde* (*LM*) annotés dans le corpus manuel et les articles de l'*Est Républicain* (*ER*) constituent le corpus de test.

Dans le tableau 5.2 sont présentées les caractéristiques du corpus manuellement annoté, comme le voit *XIP*. Notons que *XIP* ne reconnaît que 1 832 noms d'événements manuellement annotés alors que 1 844 noms ont été en réalité annotés manuellement (cf. tableau 3.1). Les 12 mots non reconnus constituent la petite part d'erreur de reconnaissance dans *XIP* (soit environ 0,65 % d'erreur).

Corpus	noms	`<event>`	non-événements
LM	10 607	1 099	9 508
ER	5 703	732	4 971
total	16 310	1 831	14 479

TABLE 5.2 – Représentation du corpus manuellement annoté pour l'apprentissage automatique. Caractéristiques du corpus manuellement annoté vues par *XIP*.

À la suite des premières expérimentations que nous avons menées, l'importance des choix concernant la proportion d'instances positives dans le corpus d'apprentissage est apparue, et nous avons décidé de constituer plusieurs corpus d'apprentissage qui ont pour distinction l'équilibrage des données qui désignent des événements de celles qui désignent des non-événements :

- équilibrage « *1 sur 2* » (50 % d'événements),
- équilibrage « *1 sur 3* » (33 % d'événements),
- équilibrage « *1 sur 4* » (25 % d'événements),
- équilibrage « *1 sur 5* » (20 % d'événements),
- « sans équilibrage » (soit 10,4 % d'événements parmi tous les noms du corpus). Ici, il s'agit de tous les noms tels qu'ils sont présents dans notre corpus d'apprentissage de départ.

Nous verrons par la suite l'importance de ces différentes configurations.

Le corpus de test est d'aspect normal non équilibré (dans le corpus *L'Est Républicain*

la proportion d'événements par rapport aux non événements est de 12,8 %[2]). La répartition des corpus d'apprentissage et de test est présentée dans le tableau 5.3. La catégorie « YES » correspond aux noms annotés <EVENT> et la catégorie « NO » à celle des noms qui ne sont pas des événements.

	Répartition	total	YES	NO
Ensemble de développement	*1 sur 2*	2 198	1 099	1 099
	1 sur 3	3 297	1 099	2 198
	1 sur 4	4 396	1 099	3 297
	1 sur 5	5 495	1 099	4 396
	non équilibré	10 607	1 099	9 508
Ensemble de test	*Corpus Test ER*	5 703	732	4 971

TABLE 5.3 – Répartition des corpus de test et développement pour l'apprentissage automatique de modèles sur le corpus manuellement annoté, avec des répartitions du corpus d'apprentissage équilibré à nombre d'événements par rapport à non événements : *1 sur 2* (50 %), *1 sur 3* (33 %), *1 sur 4* (25 %), *1 sur 5* (20 %) et *sans équilibrage* (11,5 %).

Tous les tests ont été lancés sur des classifieurs appris sur corpus manuels équilibrés à la « 1 sur 2 ». Sur certains tests, plusieurs types de sous-corpus ont été utilisés comme corpus d'apprentissage. Ainsi, nous pouvons étudier les différences de performance suivant l'équilibrage des données.

5.2.2 Utilisation d'un corpus automatique

Bel *et al.* (2010), dans leurs travaux sur la détection de non-déverbaux événementiels, utilisent des listes de mots événements et non-événements en espagnol (cf. Annexe C.1.2) et en anglais (cf. Annexe C.1.3) pour créer leurs corpus de travail. Ils effectuent une validation croisée sur un classifieur automatique de type arbre de décision. Le corpus d'apprentissage/test est un ensemble de documents constitués uniquement des mots amorces et qui ont été manuellement annotés, c'est-à-dire que les informations contenues dans les traits d'apprentissage ont été remplies à la main (les traits d'apprentissage ne sont pas contextuels, ils ne prennent en compte que les qualités intrinsèques des mots). L'idée étant que comme les mots utilisés sont très certainement des événements ou des non-événements, alors les contextes d'apparition de ces mots sont représentatifs de la classe (événement ou non événement) à laquelle ils appartiennent.

Nous travaillons sur le contexte des mots plus que sur ses qualités intrinsèques, mais cette démarche nous interpelle pour la création automatique de corpus d'apprentissage.

2. La proportion moyenne globale sur tout le corpus (*LM* et *ER* confondus) est de 11,5 % d'événements parmi tous les noms du corpus

Notre postulat est qu'à partir d'une liste d'amorces non-ambigues événements et non-événements, nous pouvons extraire des contextes d'apparition de mots événements et non-événements. Si les contextes et nos traits d'apprentissage sont judicieux, alors il serait possible de créer des corpus de manière automatique pour l'extraction de noms d'événements.

5.2.2.1 Constitution automatique d'un corpus pour l'apprentissage

Nous proposons la création d'un corpus automatiquement annoté de grande taille en partant de deux listes de mots (amorces) constituées sur la base de l'événementialité de ces mots. Nous nous appuyons sur :

- une liste de 67 noms jugés toujours événementiels ou ayant un caractère très fortement événementiel, c'est notre liste d'amorces de noms « sûrs » ;
- une liste de 188 noms non événementiels, qui nous semblent ne jamais désigner d'événements, c'est notre liste d'amorces de noms « sûrs_pas ». Ce sont des noms propres ou des noms communs. Il s'agit par exemple de noms communs des plus banals (comme « `pays` », « `juge` » ou « `texte` »), de nom de lieux (comme « `Washington` », « `Allemagne` », même si dans le cas de « `Tchernobyl` », qui n'est pas dans cette liste, nous avons vu que les noms de lieux peuvent être des événements), de noms de personnes (prénoms : « `Michel` » ou patronyme : « `Bush` », « `Jospin` »), de noms désignant l'appartenance à un groupe, une nationalité, un nom de parti politique (« `taliban` », « `RPR` », « `allemand` »), de noms de chiffres ou de mois (« `quatre` », « `milliards` », « `mai` »).

Ces listes d'amorces (présentées en annexe C.1.4) ont été utilisées pour créer de manière automatique un corpus de développement de grande taille, afin de l'utiliser pour un modèle d'apprentissage. Le corpus *AFP* nous est disponible et sur de nombreuses années de brèves,c'est donc ce corpus que nous avons utilisé pour cette tâche de création d'un grand corpus de développement pour la classification des noms en événements et non-événement. Les particularités du corpus *AFP* ayant servi de base pour extraire les données d'apprentissage, ainsi que les données extraites au moyen de *XIP* et de nos listes d'amorces sont présentées par année de parution et par leurs chiffres totaux dans le tableau 5.4.

5.2.2.2 Caractéristiques des corpus d'apprentissage automatiques utilisés

Le corpus automatique « sûrs et sûrs_pas » contient toutes les infos contextuelles demandées par les traits et données par les phrases contenant les mots des lexiques « sûrs » et « sûrs_pas » et en lien avec ces mots. Ce corpus étant d'une taille particulièrement

Corpus	*Brut*		*Extrait*		
	Nb. de mots	Nb. de noms	total	YES	NO
AFP_2004	47 273 583	1 744 337	1 744 337	168 481	1 575 856
AFP_2005	42 919 059	1 632 669	1 632 669	169 939	1 462 730
AFP_2006	47 987 661	1 514 755	1 514 755	145 064	1 369 691
AFP_2009	95 301 696	3 273 243	3 273 243	292 849	2 980 394
AFP_2010	82 331 603	3 160 620	3 160 620	278 402	2 882 218
AFP_2011	123 692 773	2 948 625	2 948 625	291 496	2 657 129
Total	439 506 375	14 274 249	14 274 249	1 346 238	12 928 011

TABLE 5.4 – Composition des corpus d'apprentissage constitués uniquement de noms « sûrs » et « sûrs_pas », qui ont été construits de manière automatique sur la base d'un lexique d'amorces. Description par année de parution.

grande, la gestion du corpus total entier et non équilibré est bien trop complexe. C'est la raison pour laquelle, pour nos expérimentations, nous avons extrait trois sous-ensembles de corpus issus des données fournies :

- Le premier est un corpus global sur les six années de l'*AFP* correspondant à un équilibrage à « *1 sur 2* » de tous les noms « sûrs » par rapport aux noms « sûrs_pas ». C'est le corpus « *automatique - 1 sur 2* ».
- Le second est un corpus extrait uniquement de l'année de parution 2004 de l'*AFP*, que nous avons intitulé « *AFP_2004* ».
- Le troisième est constitué par l'équilibrage à « *1 sur 2* » du précédent « *AFP_2004* ». Nous l'avons nommé « *AFP_2004* - 1 sur 2 ».

Les particularités des corpus de développement ainsi constitués pour l'apprentissage et la plupart des tests de ce chapitre sont présentées dans le tableau 5.5. Le corpus de test est constitué de tous les noms du corpus manuellement annoté de *L'Est Républicain*, qu'ils soient événements ou non, afin que nos classifieurs entraînés sur nos corpus automatiques soient testés sur la même portion de texte annoté que les classifieurs entraînés sur une partie du corpus manuel, et ainsi que les performances puissent être comparées.

	Corpus	total	YES	NO
Ensemble de développement	*automatique - 1 sur 2*	2 692 476	1 346 238	1 346 238
	AFP_2004	1 744 337	168 481	1 575 856
	AFP_2004 - 1 sur 2	336 962	168 481	168 481
Ensemble de test	*Corpus Test ER*	5 703	732	4 971

TABLE 5.5 – Répartition des corpus de développement et de test pour l'apprentissage automatique de modèles sur le corpus « sûrs et sûrs_pas ».

Nous avons donc deux types de corpus de développement, l'un issu du corpus manuellement annoté et l'autre créé de manière automatique. Nous avons pris soin de créer des corpus à équilibrage différents pour ces corpus d'apprentissage des noms « événements » et « non-événements », afin d'évaluer pour notre tâche l'équilibrage intéressant et/ou la masse de donnée nécessaire pour obtenir les meilleurs performances de classification des événements. Ces performances seront évaluées en fonction de tests considérant une partie ou tous les traits précédemment présentés.

5.3 Résultats

Pour nos expérimentations et nos évaluations de corpus de développement (sur la base d'un corpus manuellement annoté ou sur la base d'amorces), d'équilibrage des données (équilibrage à « *1 sur 2* », « *1 sur 3* », « *1 sur 4* », « *1 sur 5* » ou *non équilibré*), et pour évaluer les meilleurs traits d'apprentissage à utiliser, nous avons mis au point plusieurs critères d'évaluation. Plusieurs batteries de tests ont été lancées sur ces différents corpus d'entraînement, les modèles d'apprentissage obtenus sont testés sur un même corpus de test manuellement annoté (« *Corpus Test ER* »).

- La première série de tests est *Tous les traits* (cf. section 5.3.1), y sont utilisés comme trait d'apprentissage tous les traits que nous avons imaginés et présentés en section 5.1.2.
- La deuxième s'intéresse uniquement aux traits lexicaux : *que lexical* (cf. section 5.3.2).
- La troisième est une étude sur les types de traits que nous avons définis (cf. section 5.3.3). Elle est constitué du test : *Chaque type un par un* et du test : *Tout moins chacun des types un par un*.

Dans cette section, nous comparons donc les performances des classifieurs suivant les corpus de développement utilisés. Comme nous l'avions fait dans les études précédemment présentées dans ce manuscrit, nos mesures d'évaluations sont la précision (P), le rappel (R) et la F-mesure (F) (cf. annexe B.1 pour plus de détail). De plus, les valeurs d'évaluation sur les classifieurs automatiques que nous affichons sont toujours celles obtenues sur la classe « événement » uniquement. Parce que nous ne cherchons pas à faire une classification, mais bien à extraire les noms d'événements, nous n'avons pas présenté de valeurs globales de classification des classes événements et non-événement (qui seraient exagérément élevées du fait de la sur-représentation des non-événements), ni de la classe non-événement,

contrairement à d'autres travaux.

5.3.1 Test : *Tous les traits*

Dans cette section, nous avons développés des classifieurs sur chacun des sous-corpus manuel et automatique précédemment décrits. L'expérimentation consiste à donner tous les traits syntaxiques, morphologiques, sémantiques et surfaciques présentés en section 5.1.2 et d'observer les comportements.

Significativité des résultats obtenus. Afin de nous rendre compte de la significativité des différences de performances des classifieurs ayant appris sur des corpus de développement équilibrés différemment, nous avons appliqué le test t de Student (cf. annexe B.4) sur le point de la précision des modèles dans le cadre de notre test *Tous les traits*. Nous avons comparé les réponses proposées par les différents modèles pour chaque ligne du corpus de test. Ainsi, les performances de chacun des classifieurs ont été évaluées par rapport aux autres classifieurs. Le but étant de vérifier s'il est improbable que ces différences de classification (erreur ou non sur l'attribution de la classe événement à des lignes du test qui sont événement) puisse être obtenues par un simple hasard. Un valeur t inférieure à $0,05$ (ce qui correspond à 5 % de chances d'être obtenue par hasard) marque une différence significative et en dessous de $0,01$, elle est très significative.

5.3.1.1 Utilisation du corpus manuellement annoté

Les résultats sur les sous-corpus manuels sont présentés dans le tableau 5.6.

Répartition	P	R	F
1 sur 2	**0,584**	**0,886**	0,704
1 sur 3	0,662	0,811	0,729
1 sur 4	0,706	0,791	**0,746**
1 sur 5	0,702	0,728	0,715
non équilibrée	**0,824**	**0,596**	0,692

TABLE 5.6 – Performances des différents modèles sur le test : *Tous les traits*. La différence entre les modèles consistant en la différence de répartition en événements et non événements du corpus de développement qui a été utilisée : *1 sur 2* (50 %), *1 sur 3* (33 %), *1 sur 4* (25 %), *1 sur 5* (20 %) et sans équilibrage (11,5 %).

Nous remarquons rapidement la précision basse (P = 0,584) et le rappel haut (R = 0,886) du modèle « *1 sur 2* ». Ainsi, le modèle « *1 sur 2* » s'oppose clairement au modèle « *non équilibré* », qui a une précision haute (P = 0,824) et un rappel moyen

(R = 0,586). Nous notons, de plus, que c'est le modèle « *1 sur 4* » qui donne une performance plus équilibrée (P = 0,706 et R = 0,791) et la meilleure F-mesure sur les corpus manuels (F = 0,746). Nous observons de plus que les performances fluctuent en fonction du sous-corpus utilisé et donc de l'équilibrage en événement et non-événement utilisé. La précision augmente au fur et à mesure que la part en non-événement dans le corpus de développement augmente (de P = 0,584 pour « *1 sur 2* » (50 % de noms qui ne sont pas événements) à P = 0,702 pour le sous-corpus « 1 sur 5 » (80 %) et P = 0,824 « non équilibré » (88,5 %)). Par opposition le rappel décroit, plus le nombre de non-événement des corpus augmente. Ce qui montre que le modèle le plus performant sur ce test en fonction de la f-mesure est le modèle appris sur le sous-corpus « *1 sur 4* ». Le modèle le plus précis est celui créé à partir du corpus non équilibré et le modèle qui a le rappel le plus élevé est celui du corpus « *1 sur 2* ».

Significativité des résultats obtenus. Selon le test t de Student (que nous avons appliqué sur le point de la précision des modèles), nous observons que la plupart des modèles propose des résultats sur les entités manuellement annotées « événement » qui sont significativement différents sur le test : *Tous les traits*.

La question de la significativité se pose seulement sur les modèles « *1 sur 2* » par rapport à « *1 sur 5* » qui avec une valeur de $0,0284$ est considérée comme une probabilité de différence significative, même si la valeur de probabilité donnée par le test est plus haute que les autres comparaisons de modèles. Pour les modèles « *1 sur 3* » par rapport à « *1 sur 4* » avec une valeur de $0,0070$, nous admettons que la différences de performances sur la précision est très significative ($t = 4,3886^{-55}$).

Les résultats de ce test statistique sont présentés dans le tableau 5.7.

	1 sur 2	*1 sur 3*	*1 sur 4*	*1 sur 5*	*non équilibré*
1 sur 2					
1 sur 3	++ $(9,105^{-12})$				
1 sur 4	++ $(5,618^{-16})$	++ $(0,0070)$			
1 sur 5	+ $(0,0284)$	++ $(4,272^{-9})$	++ $(2,067^{-11})$		
non équilibré	++ $(4,388^{-55})$	++ $(5,170^{-36})$	++ $(1,756^{-32})$	++ $(1,072^{-44})$	

TABLE 5.7 – Matrice de confusion des valeurs de la variable t, obtenue par l'application du test de Student sur – Expérimentations des calssifieurs sur le test : *Tous les traits*.

5.3.1.2 Utilisation du corpus automatique

Les résultats sur les sous-corpus automatiques sont présentés dans le tableau 5.8.

Corpus de développement	P	R	F
automatique - 1 sur 2	0,707	**0,657**	**0,681**
AFP_2004	0,664	0,635	0,649
AFP_2004 - 1 sur 2	**0,722**	0,645	**0,681**

TABLE 5.8 – Performances des différents modèles appris sur le corpus automatique pour le test *Tous les traits*.

Nous voyons que le modèle le plus précis est celui développé sur le sous-corpus d'un an et équilibré à « *1 sur 2* » (*AFP_2004 - 1 sur 2*). Sa performance sur la précision est meilleure que l'autre corpus automatique équilibré, mais elle est aussi meilleure que le sous-corpus d'un an complet, non équilibré (*AFP_2004*). Globalement, les performances de ces trois modèles ayant appris sur corpus automatique sont proches.

On observe pourtant que les corpus équilibrés (F = 0,681 pour les deux) ont des performances supérieures à celles du corpus non équilibré (F = 0,649). Meilleure précision et meilleur rappel sont obtenus par les corpus équilibrés : tantôt par le corpus global (rappel élevé), tantôt par l'*AFP_2004 - 1 sur 2* (précision élevée).

Significativité des résultats obtenus. Selon le test de Student sur la précision des modèles, les différences entre les résultats sur les valeurs « événement » données par le corpus de référence (test de Student sur la précision) sont peu significatives entre *automatique - 1 sur 2* et *AFP_2004* ($t = 0{,}1445$), sont significatives entre *AFP_2004* et *AFP_2004 - 1 sur 2* ($t = 0{,}0201$) et sont très significatives uniquement entre *automatique - 1 sur 2* et *AFP_2004 - 1 sur 2*.

5.3.1.3 Comparaison des performances des deux sortes de corpus

Dans la représentation graphique des performances de tous les corpus de développement proposée par la figure 5.1, nous constatons que les performances des modèles issus de corpus automatiques sont relativement moins bonnes que celles sur corpus manuel. Pour autant, en comparant les modèles issus des sous-corpus à « *1 sur 2* », nous observons que les corpus automatiques permettent à leurs modèles d'accéder à une bien meilleure précision que le corpus manuel équilibré à « *1 sur 2* ». Les modèles sur corpus non équilibrés ont des performances qui ne sont pas équivalentes (P = 0,824, R = 0,596, F = 0,692 pour le corpus manuel et P = 0,664, R = 0,635, F = 0,649 pour le corpus automatique). Du point de vue de la F-mesure, le modèle le plus performant est celui qui

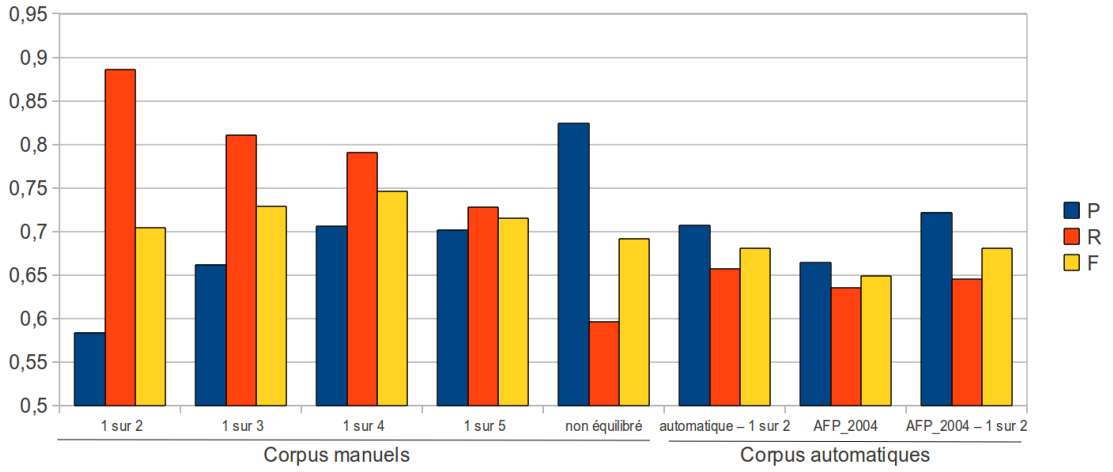


FIGURE 5.1 – Performances des classifieurs en terme de précision (P), rappel (R) et F-mesure (F) en fonction de la répartition des événements dans tous les corpus d'apprentissage, manuels et automatiques – Expériences sur tous les critères, test : *Tous les traits*.

a fait son apprentissage sur le sous-corpus manuel « *1 sur 4* ».

À ce point de notre étude, nous pouvons avancer que les corpus automatiques sont moins bons que les corpus manuels pour l'apprentissage de la classification des événements. De plus les performances des différents équilibrages sont relativement différentes si l'on compare les corpus automatiques ou manuels. Il n'y a pas de constance.

5.3.2 Test : *type lexical*

Nous avons testé les traits portant sur les informations lexicales concernant l'appartenance au lexique *VerbAction* (test $L3$), au lexique *EventNominals* (test $L2$), à l'un des deux lexiques (trait nommé « *lexique* », test $L1$), les valeurs d'ERW du lexique pondéré issu du corpus AFP (les ERW_{AFP}, test $L5$) et la combinaison de ces valeurs de lexique pondéré avec le trait « *lexique* » (test $L7$ avec les valeurs d'ERW_{AFP}).

5.3.2.1 Utilisation du corpus manuellement annoté

Nous avons créé un modèle pour chaque sous-ensemble de corpus issu du corpus annoté manuellement et les résultats obtenus sont présentés dans les tableaux 5.9.

TEST		*1 sur 2*			*non équilibré*		
		P	R	F	P	R	F
$L1$	QUE lexique	**0,578**	**0,878**	**0,697**	0	0	0
$L2$	QUE EventNominals	0,528	**0,157**	0,242	0	0	0
$L3$	QUE Verbaction	0,591	0,721	0,649	0	0	0
$L4$	QUE verbaction et eventnominals	0,578	**0,878**	0,697	0	0	0
$L5$	QUE ERW_{AFP}	0,447	0,802	0,574	0,671	0,574	0,619
$L7$	QUE lexique et ERW_{AFP}	**0,646**	0,858	**0,737**	0,79	0,579	0,668
$L9$	QUE lexique, ERW_{AFP} et règles	0,6	0,9	0,72	0,678	0,72	**0,698**

	1 sur 3			*1 sur 4*			*1 sur 5*		
	P	R	F	P	R	F	P	R	F
$L1$	0,578	0,878	0,697	0,578	0,879	0,698	0,533	0,877	0,663
$L2$	0,525	0,156	0,241	0,528	0,157	0,242	0,496	0,157	0,239
$L3$	0,59	0,722	0,649	0,591	0,722	0,65	0,542	0,728	0,618
$L4$	0,578	0,878	0,697	0,578	0,879	0,698	0,533	0,877	0,663
$L5$	0,528	0,748	0,619	0,63	0,624	0,627	0,623	0,581	0,601
$L7$	0,649	0,85	0,736	0,68	0,823	0,744	0,668	0,732	0,699
$L9$	0,652	0,853	**0,739**	0,652	0,854	0,74	0,652	0,853	**0,739**

TABLE 5.9 – Différents tests sur les données lexicales uniquement au moyen des classifieurs ayant appris sur des corpus de développement équilibrés différemment et provenant du corpus manuel – Expériences sur les tests lexicaux (tests *type lexical*).

Pour tous les classifieurs entraînés sur chaque sous-corpus, nous observons que les performances de $L1$ et $L4$ sont identiques, soit que l'utilisation du seul trait « *lexique* » équivaut à l'utilisation des traits « *VerbAction* » et « *EventNominals* » dans deux traits distincts. En effet, un mot possède le trait « lexique » positif lorsque ce mot est dans l'un des deux lexiques (*VerbAction* ou *EventNominals*).

Dans la représentation graphique de la figure 5.2, nous observons de plus la constance des résultats pour tous les classifieurs sur le test $L1$. Les performances sont les même pour les classifieurs ayant appris sur *1 sur 2*, *1 sur 3* et *1 sur 4* et un peu inférieures sur *1 sur 5*.

Sur le corpus « non équilibré » pour ce test, les performances sont à zéro parce que le modèle classifie toutes les propositions en non événement par défaut. Il semble que sur ce test ajouter plus d'entrées non-événements dans le corpus d'apprentissage ne permet pas de correctement classer les événements. Nous remarquons que le classifieur arrive encore à tenter une classification réfléchie sur *1 sur 5* mais plus sur un corpus de développement équilibré à environ *1 sur 8* (le corpus *non équilibré* contient 11,5 % d'événements, soit un

équilibrage à environ *1 sur 8*).

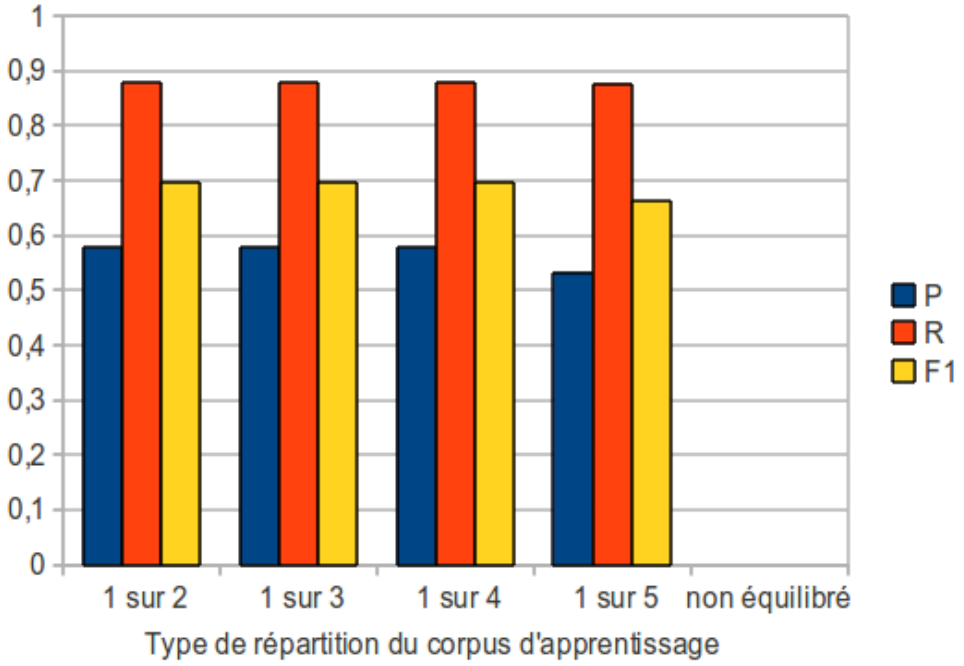


FIGURE 5.2 – Performances des classifieurs en terme de précision (P), rappel (R) et F-mesure (F) en fonction de la répartition des noms d'événement manuellement annotés dans le corpus d'apprentissage – Expériences sur « QUE lexique » (Test $L1$).

Nous constatons de plus que l'utilisation du lexique *EventNominals* en plus du *VerbAction* (test $L4$) provoque une augmentation du rappel et une petite baisse de la précision par rapport au *VerbAction* utilisé seul (test $L3$), comme nous en avons déjà fait état dans un chapitre précédent à la section 4.1.1.

En ce qui concerne les performances du trait « ERW_{AFP} » (test $L5$), les performances fluctuent en fonction du sous-corpus utilisé et donc de l'équilibrage en événement et non-événement utilisé. La représentation graphique 5.3 est particulièrement représentative. La précision augmente au fur et à mesure que la part en non-événement dans le corpus de développement augmente (de « *1 sur 2* » (50 % de noms qui ne sont pas événements) à « *1 sur 5* » (80 %) et « *non équilibré* » (88,5 %)). Le modèle le plus constant pour ce test est le modèle construit sur le sous-corpus « *1 sur 4* » : sa F-mesure est proche de celle du modèle « *1 sur 3* » (F = 0,619 pour « *1 sur 3* » contre F = 0,744 pour « *1 sur 4* »), mais ce dernier présente une rappel plus fort (R = 0,748 contre R = 0,624) et une précision plus basse (P = 0,528 et P = 0,63).

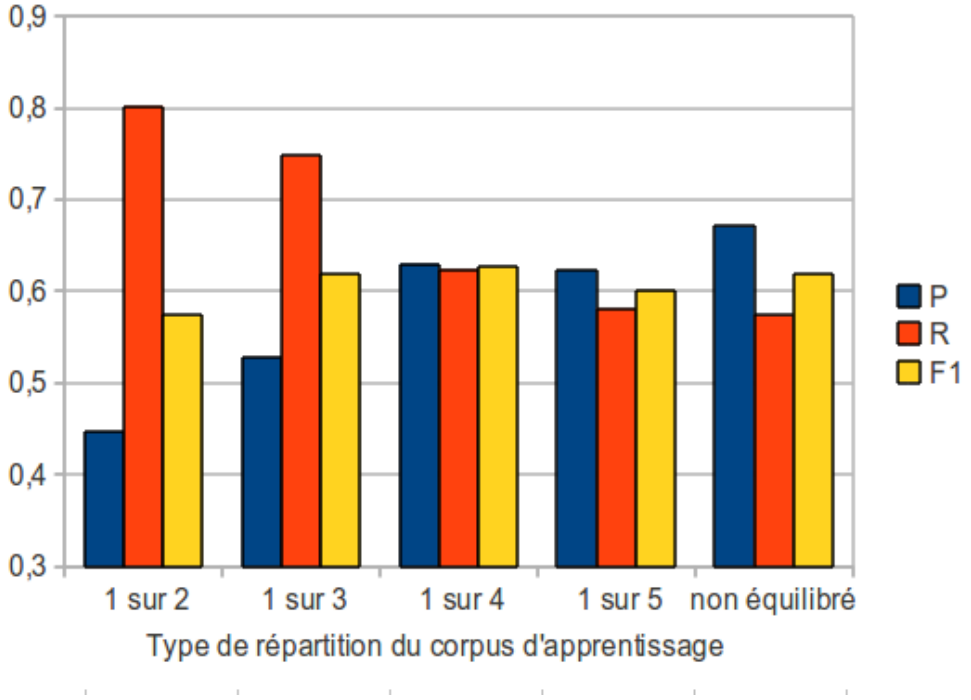


FIGURE 5.3 – Performances des classifieurs en terme de précision (P), rappel (R) et F-mesure (F) en fonction de la répartition des noms d'événement manuellement annotés dans le corpus d'apprentissage – Expériences sur « QUE ERW_{AFP} » (Test L5).

Dans le cas du test *L7* de combinaison du trait « *lexique* » et du trait « ERW_{AFP} », soit la présence de toutes les informations lexicales, les résultats sont plus complexes. Nous pouvons les observer dans la figure 5.4. En effet, les modèles sont de plus en plus précis avec l'augmentation du nombre de données classées non-événement dans le corpus de développement, avec une précision homogène de P = 0,646 et P = 0,649 pour « *1 sur 2* » et « *1 sur 3* », un légère augmentation sur « *1 sur 4* » (P = 0,68, valeur très proche du P = 0,668 de « *1 sur 5* ») et un pic avec le corpus non équilibré avec son P = 0,79.

Pour ce qui est du rappel, nous observons la progression inverse, le rappel augmente plus la proportion d'événement baisse. La meilleure F-mesure est obtenue pour le classifieur entraîné sur le corpus équilibré à « *1 sur 4* », sachant que ses performances sont très proches des équilibrages à proportion supérieures de noms d'événements.

Notons que le test *L7* est strictement exactement le même test que le test *Que_Lex* de la section 5.3.3.

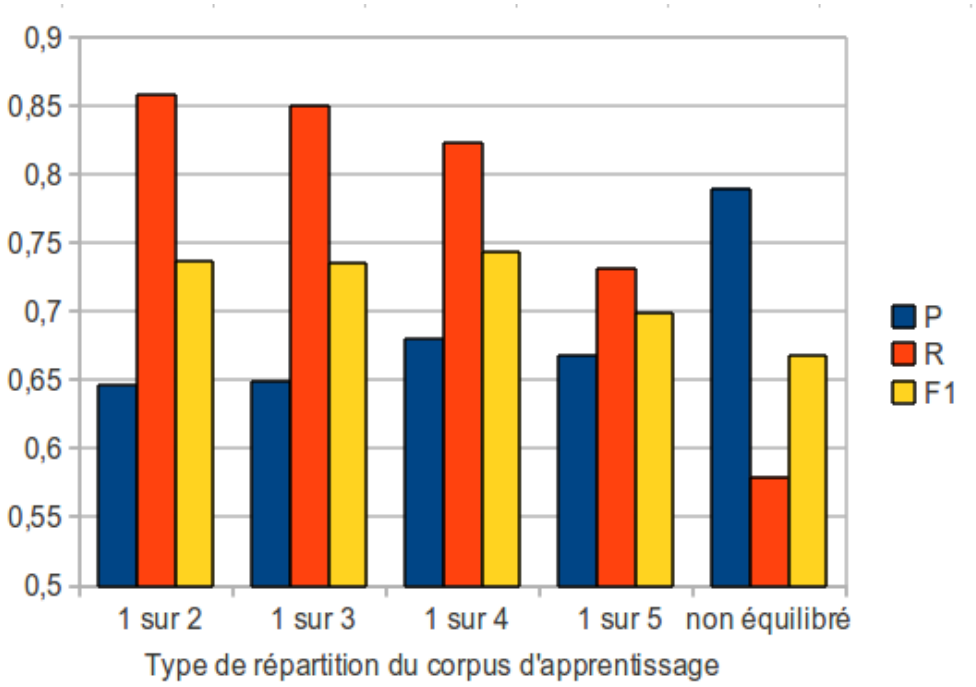


FIGURE 5.4 – Performances des classifieurs en terme de précision (P), rappel (R) et F-mesure (F) en fonction de la répartition des noms d'événement manuellement annotés dans le corpus d'apprentissage – Expériences sur « QUE lexique et ERW_{AFP} » (Test L7).

Dans le test $L9$, les classifieurs ont accès au résultat des règles d'extraction en plus des traits lexicaux « lexique » et « ERW_{AFP} ». Les résultats des tests $L7$ et $L9$ sont comparables et permettent d'évaluer l'apport des règles sur l'apprentissage par rapport aux informationq lexicales utilisées seules.

Nous observons que le rapport entre les résultats sur $L7$ et $L9$ n'est pas constant. Les performances des classifieurs ayant accès à l'information supplémentaire des résultats des règles est, en terme de F-mesure, sont tantôt légèrement inférieures (pour « *1 sur 2* » et « *1 sur 4* »), tantôt légèrement supérieures sur « *1 sur 2* » ou bien clairement supérieures pour « *1 sur 5* » (de F = 0,699 à F = 0,739) et « non équilibré » (de F = 0,668 à F = 0,698).

Pourtant, l'utilisation des règles uniquement mène à une précision de P = 0,8 pour un rappel insignifiant de R = 0,066.

5.3.2.2 Utilisation du corpus automatique

Les mêmes tests ont été menés sur les sous-corpus automatiques, ils sont présentés dans le tableau 5.10.

TEST	*automatique - 1 sur 2*			*AFP_2004*			*AFP_2004 - 1 sur 2*		
	P	R	F	P	R	F	P	R	F
$L1$	0,578	0,877	**0,697**	0,578	0,877	**0,697**	0,578	0,877	**0,697**
$L2$	0,528	0,157	0,242	0,528	0,157	0,242	0,528	0,157	0,242
$L3$	0,591	0,727	0,649	0,591	0,72	0,649	0,591	0,72	0,649
$L4$	0,578	0,877	0,697	0,578	0,877	0,697	0,578	0,877	0,697
$L5$	0,587	0,653	0,618	0,639	0,62	0,63	0,639	0,62	0,63
$L7$	0,752	0,66	**0,702**	0,723	0,65	0,685	0,748	0,63	0,684
$L9$	0,601	0,791	0,683	0,601	0,791	0,683	0,601	0,791	0,683

TABLE 5.10 – Différents tests sur les données lexicales uniquement au moyen des classifieurs ayant utilisé des données d'apprentissage équilibrées différemment.

Pour les tests concernant les traits « *VerbAction* » et « *EventNominals* » et « *lexique* » sur des classifieurs développés sur les corpus automatiques, peu importe la masse de données ou l'équilibrage du corpus, les performances sont exactement les même.

De plus, les même observations que celles sur le corpus manuel peuvent être faites concernant les tests $L1$ et $L4$: « *EventNominals* » est une valeur ajoutée par rapport à « *VerbAction* » seul. Le test utilisant les traits correspondant à ces deux lexiques donne des résultats exactement semblables au test avec uniquement le trait « lexique ».

Nous remarquons aussi que les classifieurs *AFP_2004* et *AFP_2004 - 1 sur 2* ont exactement les mêmes performances mis-à-part sur le test $L7$ de combinaison des traits « ERW_{AFP} » et « lexique », dont la f-mesure est quasi identique (F = 0,685 et F = 0,684).

Sur le test $L7$ encore, nous voyons que les performances du classifieur *automatique - 1 sur 2* sont supérieures à celles de *AFP_2004* et *AFP_2004 - 1 sur 2*, plus précis et meilleur F-mesure (F = 0,702 contre F = 0,685 pour les autres).

Sur le test $L9$, nous voyons que toutes les performmances sont égales. Une fois de plus, ni la taille du corpus ni l'équilibrage ne fait varier les résultats sur le corpus artificiel. Qui plus est la F-mesure sur $L9$ est toujours inférieure ou très proche de celle de $L7$.

Finalement, l'équilibrage ou non du corpus *AFP_2004* ne modifie pas les performances des modèles pour les tests lexicaux. Sur la plupart des tests, les performances des sous-corpus *AFP_2004* et du sous-corpus *automatique - 1 sur 2* sont proches. La quantité de donnée n'influe pas sur les résultats.

5.3.2.3 Comparaison des performances des deux sortes de corpus

Sur le test $L1$, les performances sont fréquemment les même pour les modèles sur corpus manuels et ceux sur corpus automatique (P = 0,578,R = 0,878 et F = 0,697 pour tous les modèles sur corpus automatiques et les meilleurs modèles sur corpus manuel). Sur le test $L5$, utilisant seulement les valeurs du ERW_{AFP}, les performances des modèles fondés sur les sous-corpus automatiques *AFP_2004* et *AFP_2004 - 1 sur 2* ont des résultats équivalent aux résultats du meilleur modèle issu des corpus manuels (F = 0,62 contre F = 0,63). En ce qui concerne le test $L7$, le corpus équilibré à « *1 sur 4* » a les meilleures performances de tous les classifieurs.

Globalement, nous pouvons à ce point avancer que nos corpus automatiques de grande taille utilisés pour entraîner des classifieurs automatiques ont des perfomances équivalentes au corpus manuellement annoté. Cette observation est intéressante, parce qu'elle sous-entend que sur ce test un corpus automatique comme nous avons constitué le nôtre est suffisant pour l'apprentissage de classifieur pour notre tache de classification des événements. Ce qui veut dire qu'il est possible d'éviter l'annotation manuelle de corpus de grande taille.

Nous avons aussi observé que l'équilibrage et la quantité de données pour le corpus automatique ne donne pas des performances très différentes. Néanmoins, sur le corpus manuel, l'équilibrage donne une modification des performances substancielle. En effet, les expérimentations menées sur les corpus manuels montrent qu'en jouant sur l'équilibrage des données d'apprentissage, nous obtenons un système de haute précision ou de haut rappel, particularité intéressante qu'on ne retrouve pas avec le corpus automatique. Par ailleurs, au vu de la F-mesure, l'équilibrage en événement et non-événement qui convient le mieux pour le corpus manuel est celui à *1 sur 4*.

Nous observons aussi que l'utilisation des informations lexicales seules conduit à de meilleurs résultats que « *Tous les traits* », ce qui montre que notre recherche de contextualisation n'est pas couronnée de succès, ce que nous allons confirmer dans l'étude de la section suivante.

5.3.3 Étude sur les types de traits utilisés

Dans cette section, nous décrivons et testons des classifieurs appris sur les types de traits présentées en section 5.1.2, en utilisant les traits selon le type auquel ils appartiennent : syntaxiques, sémantiques, morphologiques et surfaciques. Le premier test

TEST		*1 sur 2*			*non équilibré*		
		P	R	F	P	R	F
Que_synt	QUE tout syntaxique	0,247	0,836	0,381	0	0	0
Que_sem	QUE tout sémantique	0,237	0,765	0,362	0,724	0,057	0,106
Que_morph	QUE tout morphologique	0,263	0,791	0,395	0	0	0
Que_surf	QUE tout surfacique	0,215	0,883	0,341	0,824	0,057	0,107
Que_lex	QUE **tout lexical**	**0,646**	**0,858**	**0,737**	**0,79**	**0,579**	**0,668**

TEST	*automatique - 1 sur 2*			*AFP_2004*			*AFP_2004 - 1 sur 2*		
	P	R	F	P	R	F	P	R	F
Que_synt	0,278	0,66	0,391	0	0	0	0,266	0,654	0,378
Que_sem	0,240	0,538	0,332	0,689	0,057	0,106	0,227	0,817	0,355
Que_morph	0,358	0,501	**0,418**	0,442	0,301	0,358	0,305	0,510	0,381
Que_surf	0,153	0,602	0,244	0,109	0,12	0,115	0,155	0,598	0,247
Que_lex	**0,752**	**0,66**	**0,702**	**0,723**	**0,65**	**0,685**	**0,748**	**0,63**	**0,684**

TABLE 5.11 – Performances en terme de précision (P), rappel (R) et F-mesure (F) de modèles ayant appris sur le set de développement issu du corpus manuellement annoté, avec une répartition du corpus d'apprentissage à 50 % des éléments qui sont annotés événement et 50 % de non événement – Expériences sur les regroupements de traits : test : *Chaque type un par un*.

consiste à donner à l'apprentissage automatique uniquement un type de trait à la fois (test : *Chaque type un par un*). Le deuxième test consiste à lui donner tous les types de traits sauf celui à tester (test : *Tout moins chacun des types un par un*).

Les tableaux 5.11 et 5.12 présentent les performances des modèles entraînés suivant ces tests et selon le corpus d'apprentissage utilisé.

Les performances sur les classifieurs qui n'utilisent qu'un regroupement de trait (test : *Chaque type un par un*, dont les résultats sont présentés dans le tableau 5.11) sont médiocres, voire nulles : la F-mesure ne décolle jamais à plus de 0,418 pour le test *Que_morph* sur le sous-corpus « *automatique - 1 sur 2* ». C'est effectivement le cas pour tous les tests sauf les tests qui ne tiennent compte que des tests lexicaux (test *Que_Lex*[3]), où les performances grimpent en flèche (F = 0,668 pour la plus petite F-mesure des tests *Chaque type un par un* obtenue par le corpus manuel « *non équilibré* »et F = 0,737 pour la plus grande F-mesure atteinte par le corpus manuel « *1 sur 2* »).

Sur les tests : *Tout moins chacun des types un par un* (résultats présentés dans 5.12), c'est très logiquement l'inverse, toutes les performances sont bonnes, mais les résultats

3. Le test *Que_Lex* est strictement exactement le même test que le test lexical *L7*.

TEST		*1 sur 2*			*non équilibré*		
		P	R	F	P	R	F
Sans_synt	TOUT sauf syntaxique	0,646	0,841	0,731	0,854	0,464	0,602
Sans_sem	TOUT sauf sémantique	0,595	0,866	0,705	0,812	0,538	0,647
Sans_morph	TOUT sauf morphologique	0,592	0,851	0,698	0,814	0,602	0,692
Sans_surf	TOUT sauf surfacique	0,639	0,851	0,730	0,804	0,522	0,633
Sans_lex	TOUT **sauf lexical**	0,306	0,793	**0,442**	0,84	0,057	**0,107**

TEST	*automatique - 1 sur 2*			*AFP_2004*			*AFP_2004 - 1 sur 2*		
	P	R	F	P	R	F	P	R	F
Sans_synt	0,671	0,678	0,674	0,657	0,667	0,662	0,668	0,673	0,671
Sans_sem	0,671	0,678	0,674	0,657	0,667	0,662	0,668	0,673	0,671
Sans_morph	0,668	0,678	0,673	0,669	0,654	0,662	0,681	0,661	0,671
Sans_surf	0,669	0,727	0,697	0,662	0,665	0,663	0,696	0,664	0,680
Sans_lex	0,191	0,585	0,288	0,231	0,407	0,294	0,213	0,600	0,314

TABLE 5.12 – Performances en terme de précision (P), rappel (R) et F-mesure (F) de modèles ayant appris sur l'ensemble de développement issu du corpus manuellement annoté, avec une répartition du corpus d'apprentissage à 50 % des éléments qui sont annotés événement et 50 % de non événement – Expériences sur les regroupements de traits : test : *Tout moins chacun des types un par un*.

des classifieurs qui n'ont pas accés aux lexiques sont fort bas (F = 0,107 pour le classifieur appris sur corpus manuel non équilibré et la meilleure F-mesure atteint F = 0,442 pour le classifieur appris sur le corpus manuel « *1 sur 2* »).

Ces tests effectués en parallèle prouvent que les traits lexicaux sont les plus déterminants dans tous les classifieurs que nous avons développés.

5.3.4 Combinaison de classifieurs

Sur le même test, nous avons vu que certains classifieurs ont une précision élevée et un rappel faible et par opposition d'autres ont une précision faible et un rappel élevé. Ces deux types de résultats nous intéressent, alors nous souhaitons combiner les performances des classifieurs pour obtenir de meilleurs résultats.

Les modèles appris sur le corpus de développement « 1 sur 2 » et « non équilibré » sur le test : *Tous les traits* obtiennent pour le premier la plus basse précision (P = 0,584) et le plus élevé des rappels (R = 0,886) de cette batterie de test et pour le deuxième modèle la précision la plus élevée (P = 0,824) et le rappel le plus bas (R = 0,596) (cf. tableau 5.6).

De plus, eu égard au test t de Student (que nous avons appliqué sur le point de la précision des modèles), nous observons que ces deux modèles sont significativement

différents sur le test : *Tous les traits*. Nous avons donc utilisé ces deux modèles pour la combinaison de classifieurs dans le but d'obtenir une précision élevée et un rappel moins faible.

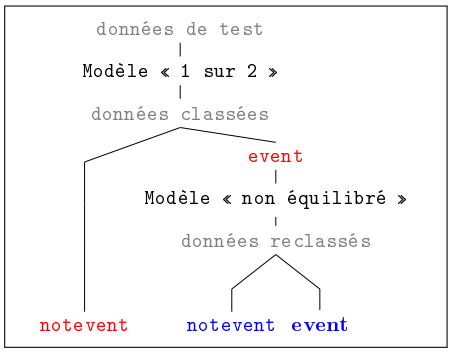
Un modèle qui a un rappel élevé et une précision basse va collecter tous les noms d'événement de façon trop large, en revanche on peut lui faire confiance lorsqu'il prédit un non-événement. D'un modèle qui a une précision élevée mais un rappel bas, on peut dire qu'il se trompe peu lorsqu'il prédit un événement, mais qu'il en laisse beaucoup de côté. Ainsi le modèle « *1 sur 2* » ne se trompe que rarement lorsqu'il dit qu'une description par trait qui lui est proposée est un événement, et le modèle « *non équilibré* » ne se trompe que rarement sur les non événements.

Combinaison en cascade des modèles « 1 sur 2 » et « non équilibré ». La technique suivie consiste en une simple combinaison en cascade de deux modèles. L'un est utilisé sur toutes les données de test et l'autre uniquement sur une partie des données, celles pour lesquelles la réponse à la classification est de la classe pour laquelle le premier modèle appliqué est moins doué. Si le premier modèle a une forte précision sur la classe événement, alors les items classés comme non événement sont donné au deuxième modèle, qui lui est doué pour le rappel.

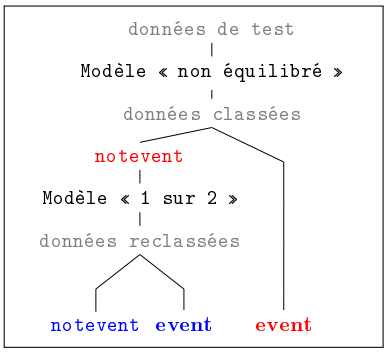
Les deux modèles utilisés sont de conception très proches : le même type de classifieur (*J48* de Weka), mais avec des performances significativement différentes et un corpus de développement différent car l'un manuellement annoté en contexte phrastique et l'autre automatiquement généré sur la base de noms amorces (événementiels certains et à caractère événementiel non certain). Mais les corpus d'apprentissage utilisés sont issus d'un même type de corpus journalistique. Même si nos deux modèles n'ont finalement pour différences que le fait que leur ensemble d'apprentissage est différent, nous combinons ces deux modèles dans le but d'obtenir de meilleur performances par l'uilisation combinée de nos modèles.

Nous avons choisi de combiner nos modèles de deux façons et selon les schémas du tableau 5.13 :

- La *combinaison 1* correspond à l'application du modèle ayant appris la classification des événements sur un corpus de développement de type « *1 sur 2* » (qui présente un rappel élevé), puis l'application du modèle issu du corpus de développement « *non équilibré* » sur les lignes du test qui ont été étiquetées comme événement.
- La *combinaison 2* correspond quant à elle à l'application du modèle basé sur le corpus de développement « *non équilibré* » d'abord, puis l'application du modèle « *1 sur 2* » sur les lignes du test qui n'ont pas été étiquetées comme événement.



Combinaison 1



Combinaison 2

TABLE 5.13 – Schémas de l'exécution des combinaisons de performances des modèles « 1 sur 2 » et « non équilibré ».

Résultats. Les résultats obtenus sont présentés dans le tableau 5.14.

Modèle	P	R	F
1 sur 2	0,584	**0,886**	**0,704**
non équilibré	**0,824**	0,596	**0,692**
Combinaison 1	**0,841**	0,594	**0,696**
Combinaison 2	0,579	**0,889**	**0,701**

TABLE 5.14 – Performances des combinaisons – Expérimentations sur le test $B1$.

Nous observons que les résultats de la *combinaison 1* présentent une précision supérieure à celle du classifieur le plus précis (P = 0,841 contre P = 0,824 pour le modèle appris sur corpus d'apprentissage *non équilibré*) et le rappel est équivalent à celui de ce même modèle (R = 0,594 pour la *combinaison 1* contre R = 0,594 pour le modèle « *non équilibré* »). Nous remarquons le cas inverse pour la *combinaison 2* : une précision inférieure à celle du modèle « *1 sur 2* », le modèle doué pour le rappel, (P = 0,579 au lieu de R = 0,584) et un rappel à peine plus élevé (R = 0,889 au lieu de R = 0,886).

Cependant, le test de Student montre que ces différences restent peu significatives (valeur $t = 0{,}1574$ pour le modèle « *non équilibré* » par rapport à la *combinaison 1* et pour le modèle « *1 sur 2* » par rapport à la *combinaison 2*).

De plus nous constatons qu'en terme de F-mesure, toutes les combinaissons se valent (F est compris entre 0,692 et 0,704). Nous ne sommes pas parvenus à l'objectif voulu de créer une méthode plus perfomante par la combinaison de nos deux meilleurs classifieurs (l'un pour la meilleure précision et l'autre pour le rappel le plus élevé). Toutes les com-

binaison sont finalement équivalentes à l'un des classifieurs utilisés pour la combinaison.

Nous constatons qu'une combinaison naïve en cascade ne fonctionne pas pour améliorer les performances de nos classifieurs, mais un autre type de combinaison plus élaboré pourrait peut-être faire meilleure office. De plus, pour ces tentatives de combinaison, nous avons utilisé des classifieurs ayant appris sur le même test (les même traits d'apprentissage définissant le test), il est envisageable d'aller au-delà et de tenter la tâche plus ardue de combiner de classifieurs portant sur des tests différents.

5.4 Discussion

5.4.1 Nos résultats

Dans le tableau 5.15 sont présentées les meilleures performances de nos classifieurs automatiques.

Test	*Corpus*	P	R	F
Tous les traits	*1 sur 4*	0,706	0,791	**0,746**
...	*1 sur 2*	0,584	**0,886**	0,704
...	*non équilibré*	**0,824**	0,596	0,692
Que_lex = L7	*1 sur 4*	0,68	0,823	**0,744**
QUE lexique - *L1*	*AFP_2004*	0,578	0,877	0,697
QUE lexique - *L1*	*AFP_2004 - 1 sur 2*	0,578	0,877	0,697
Que_lex = L7	*AFP_2004 - 1 sur 2*	0,748	0,63	0,684
Que_lex = L7	*automatique - 1 sur 2*	0,752	0,66	0,702
Combinaison 1		0,841	0,594	0,696
Combinaison 2		0,579	0,889	0,701

TABLE 5.15 – Les meilleures performances de nos classifieurs automatiques en terme de précision (P), rappel (R) et F-mesure (F).

La première de nos observations concerne le fait que les meilleurs performances sont obtenues avec l'aide des traits lexicaux. Aucun des traits contextuels seul ne parvient à des performances intéressantes. Nous pouvons en conclure que le contexte semble inutile au regard des tests entrepris ici.

Nous observons que le classifieur développé sur un corpus équilibré à « *1 sur 4* » (soit un événement pour trois noms qui ne le sont pas dans le corpus d'entraînement) est en règle générale le type de classifieur qui a les meilleures performances, sachant que cet équilibrage n'est pas l'équilibrage le plus proche de la réalité (11,5 % d'événements par rapport au nombre total de noms). Nous voyons aussi que le classifieur d'équilibrage « *1 sur 4* » testé uniquement sur les traits « *lexique* » et « ERW_{AFP} » (test lexical *L7*) est

aussi performant que le meilleur classifieur qui considère tous les traits et qui est encore une fois appris sur le corpus équilibré à « 1 sur 4 » (P = 0,68 et P = 0,706 – la valeur au test t de Student est très significative de 0,00055 – , F = 0,744 et F = 0,746). Nous en déduisons que les traits lexicaux « *lexique* » et « ERW_{AFP} » sont les traits les plus performants de tous les traits proposés.

Par ailleurs le clasifieur le plus précis de tous est celui appris sur corpus non équilibré (P = 0,824) et celui qui a le rappel le plus élevé (R = 0,886) celui appris sur le corpus « *1 sur 2* », ces deux classifieurs étant développés avec tous les traits. Nous en concluons que le jeu d'équilibrage nous permet de gagner en précision ou en rappel.

De plus, nous avons démontré que même si les performances des classifieurs appris sur corpus automatique sont moins bonnes en terme de précision que celles apprises sur corpus manuel, il est possible de créer des corpus automatiques sur la tâche de classification des noms en tant qu'ils sont événement ou non événement, car leurs performances sont assez satisfaisantes en terme de F-mesure (la F-mesure atteint quand même F = 0,702 sur le test $L7$, qui correspond au test *Que_lex*).

5.4.2 Comparaison à l'état de l'art

Nous sommes confrontés à une difficulté dans le cadre d'une analyse comparative des résultats de nos travaux avec les autres travaux existants sur l'extraction des événements nominaux ou se rapprochant thématiquement des nôtres. Se posent le problème de la définition de l'événement, celui des ressources considérées, celui des mesures considérées et enfin celui de la performance sur les noms d'événements en particulier dans les travaux plus généraux sur les événements en général paortés par des verbes ou des noms.

Premièrement, vu que la définition d'événements n'est souvent pas clairement posée dans ces différents travaux ou alors ne recouvre pas les même aspects de l'événement, il est objectivement très difficile de faire de comparaison entre les résultats obtenus.

Secondement, même quand il s'agit de la même langue, les corpus utilisés dans les différents travaux ne sont pas les mêmes. Les résultats obtenus en terme de précision, rappel et F-mesure ne peuvent donc être comparés, d'autant que les corpus annotés le sont selon des critères différents et une représentation temporelle non équivalente. Il aurait été envisageable de comparer des approches sur des langues différentes, uniquement dans le cas où elles auraient été évaluées par les auteurs sur des corpus annotés avec la même représentation des événéments.

Troisièmement, dans les publications antérieures, les chiffres de performances ne sont pas toujours donnés de manière complète ou précise. Par exemple, certains travaux éva-

luant la tâche de classification évenement et non événement dans sa globalité, ne donnent pour toute performance qu'une valeur d'« accuracy » [4] globale calculée par les classifieurs sur les performances de classification en événement et non événement à la fois comme c'est le cas dans (Bel *et al.*, 2010). Cette méthode surévalue les résultats, en raison de la sur-représentativité des non-événements, qui souvent ont de très bons résultats car l'annotation des non-événements ets souvent évidente. C'est aussi le cas dans (Wonserver *et al.*, 2012), où nous sont données les performances des classifieurs sur les « nominal events » sans préciser à quoi correspondent exactement les chiffres : uniquement à la reconnaissance des noms d'événement ? ou à la reconnaissance globale des noms événement et ceux qui ne le sont pas ?

Pour finir, les travaux de classification des événements existent, mais les performances en particulier sur les noms ne sont pas toujours données. Ainsi, nous savons qu'Evita (Saurí *et al.*, 2005) en anglais fait un bon score sur la classification des événements (P = 0,74, R = 0,87 et F = 0,80), mais le corpus d'évaluation n'est pas indépendant du corpus d'entraînement et les performances pour chaque catégorie lexicales ne sont pas données.

Comparaison possible avec certains ? Il semble cependant que nous pouvons comparer nos résultats à ceux de : Russo *et al.* (2011) qui a travallé sur l'italien, Peris *et al.* (2010a) et Resnik and Bel (2009) sur l'espagnol. Les deux premiers travaux ont qui plus est utilisé le même classifieur que nous (*J48* dans Weka). De plus, ces travaux ont été élaborés sur des langues latines, comme le français. De ce fait, nos travaux sont un peu plus comparables aux leurs. En anglais, Creswell *et al.* (2006) ont orienté ses travaux sur l'extraction des syntagmes nominaux en anglais. Ces travaux sont les travaux les plus proches des nôtres thématiquement. (Bethard and Martin, 2006), en anglais, et (Parent *et al.*, 2008), en français, donnent leurs résultats sur les événements nominaux dans leurs travaux concernant l'extraction automatique de données *TimeML*.

Le tableau 5.16 fait la synthèse des meilleures performances affichées dans les travaux antérieurs. Avec nos 0,746 de meilleure F-mesure, nous n'égalons pas les 0,80 du *modèle 1* de (Russo *et al.*, 2011), ni les 0,836 de (Resnik and Bel, 2009). Pourtant à y regarder de plus près, nous voyons que nos travaux ne sont pas vraiment comparables.

Russo *et al.* (2011) s'intéresse à la reconnaissance de la valeur événementielle des noms déverbaux. Les performances de deux modèles ont été observées, celles d'un modèle développé sur un corpus dont les noms déverbaux uniquement ont été étiquetés en événement

4. Dans certains travaux en anglais, il semble y avoir confusion entre « Accuracy » et « précision ». « Accuracy » est parfois considéré à défaut comme l'équivalent de précision (P) (cf. annexe B.1).

et non événement (*modèle 1*) et celles dont le corpus de développement est constitué de l'IT-TimeBank (*modèle 2*). Seul le modèle 2 est comparable à nos travaux. Leur corpus de test est un corpus du même type que le corpus de développement du *modèle 1*, soit issu de *La Republica* et dont seuls les noms déverbaux onté été annotés en temps qu'ils sont événements ou pas événements. La tâche étant différente, les performances de nos classifieurs et des leurs sont finalement difficilement comparables.

Dans (Peris *et al.*, 2010a), la tâche de classification des noms déverbaux est plus fine parce que les auteurs cherchent à classer les noms déverbaux suivant quatre classes (événement, résultat, non-spécifié et lexicalisé). En ce qui concerne la reconnaissance des événements uniquement, leur modèle obtient des résultats très moyens sur la classe (145 noms d'événements trouvés sur les 255 attendus, soit une précision de P = 0,569).

Resnik and Bel (2009) présente un outil de détection automatique de noms d'événement qui ne sont pas des déverbaux en espagnol. Leurs corpus de développement et de tests sont constitués uniquement de noms non-déverbaux, pour moitié événement et pour moitié pas événement. Aucun mot ambigu n'y est présent.

En ce qui nous concerne et par opposition à tous ces travaux, nous ne nous concentrons pas ni sur les noms déverbaux, ni sur les noms qui ne sont pas des déverbaux. Nous nous intéressons à tous les noms, c'est pourquoi nos corpus de développement et de tests sont plus généralistes. Nous remarquons aussi que nos corpus contiennent des mots ambigus qui ont été supprimés des corpus de ces autres travaux. Nous avons une tâche plus globale, moins « aidée », mais néanmoins très proche. Si on considère toutes ces différences, nous considérons que nos résultats sont bons par rapport aux autres, sans parvenir donner une explication chiffrée statistiquement justifiée.

Parmi les travaux sur l'anglais, on dénombre notamment ceux de Creswell *et al.* (2006) et Bethard and Martin (2006). Même si l'anglais est une langue germanique et donc moins proche du français, il est intéressant d'observer les performances de leurs méthodes. Ces travaux-ci utilisent des corpus manuellement annotés, qui concernent tous les mots (déverbaux ou non) qui désignent des événements selon la définition TimeML et les mots ambigus ne sont pas écartés du corpus.

Dans (Creswell *et al.*, 2006) est proposée une approche de la détection des noms d'événements (expressément), mais les chiffres avancés ne sont pas très clairs et seule la précision est donnée. Sur la classe événement, la précision oscille entre P = 0,798 et P = 0,893. Pour les performances sur chaque classe (événement et non-événement), il semble que les classifieurs soient uniquement testés sur les occurrences du corpus de test qui sont événement. Par ailleurs, Bethard and Martin (2006) travaillent sur les corpus

TimeBank et cherchent à classer tous les éléments d'intérêt de *TimeML* et pas que les noms, mais ces auteurs font part de leurs résultats sur cette catégorie syntaxique qui nous intéresse. Cette fois les résultats sont exprimés en précision, rappel et F-mesure, comme les nôtres.

Si tant est que nous puissions de manière objective comparer nos résultats sur le français avec ceux de travaux sur l'anglais et que les chiffres de précision de (Creswell *et al.*, 2006) sont comparables aux nôtres, nous admettons que même nos modèles les plus précis ne sont pas aussi précis que le modèle *ALL word+context* qui assigne à tous les items du test une valeur événement ou non événement (et où les valeurs pour lesquelles le classifieur est resté indécis sont considérées incorrectes). En effet, notre modèle le plus précis obtient P = 0,824 et le leur atteint P = 0,888. En comparaison aux performances sur les événements nominaux dans Bethard and Martin (2006), nos performances sont meilleures sur la plupart de nos meilleurs modèles. Mais cette remarque ne peut être conclusive vu que nous ne comparons pas les performances de travaux sur les même langues et que nous avons déjà remarqué (chapitre 4 sur les différences de résultats des lexiques pondérés *ERW* en français et en anglais) que nos performances en utilisant les informations lexicales uniquement sont moins bonnes en anglais que celles sur le français.

Dans (Parent *et al.*, 2008), les travaux portent sur l'extraction automatique des structures *TimeML* dans des textes en français. Leurs corpus d'apprentissage et de tests sont constitués de biographies et de nouvelles annotées par les auteurs en suivant le guide d'annotation *TimeML* de l'anglais. En effet, au moment de ces travaux, le corpus *FR-TimeBank* et le guide d'annotation *TimeML* pour le français n'était pas encore disponible. Pour tout point de comparaison, à d'autres travaux, Parent *et al.* (2008) se comparent aux travaux sur l'anglais de (Bethard and Martin, 2006) et (Saurí *et al.*, 2005). Cette fois, nous pouvons nous comparer à eux avec moins de difficultés, nous travaillons sur la même langue, même si nos corpus sont de types différents et annotés selon des représentations différentes (pour eux, selon *TimeML*).

Les performances de reconnaissance d'événements nominaux dans (Parent *et al.*, 2008) sont moyennes tant en précision, rappel et F-mesure (P = 0,547,R = 0,537,F = 0,542). D'ailleurs les auteurs marquent bien dans leur article qu'ils ne font pas aussi bien sur les événements verbaux que nominaux. Touts nos meilleurs modèles présentés dans le tableau récapitualtif 5.16 ont de meilleures performances que celles de (Parent *et al.*, 2008).

Les performances de (Parent *et al.*, 2008) reposent sur l'emploi des noms du *VerbAction* dont les auteurs ont « éliminé les items indésirables », sur des règles syntaxiques pour annoter les noms qui dépendent d'une préposition temporelle annotée en <SIGNAL> dans *TimeML*, et sur une suppression des noms potentiellement événements

Les meilleures performances des autres					
Référence	Modèle	P	R	F	Acc
(Russo *et al.*, 2011)	*modèle 1*	0,72	0,88	0,80	0,715
...	*modèle 2*	0,40	0,63	0,49	0,635
(Peris *et al.*, 2010a)	*Lexicon+Corpus*	0,569	0,525	0,546	0,800
(Resnik and Bel, 2009)		0,824	0,848	0,836 *	
(Creswell *et al.*, 2006)	*ALL word+context*	0,888			
(Bethard and Martin, 2006)	*Event Identification*	0,729	0,432	0,543	
(Parent *et al.*, 2008)		0,547	0,537	0,542	

Nos meilleurs modèles				
Test	Corpus	P	R	F
Tous les traits	*1 sur 4*	0,706	0,791	0,746
...	*1 sur 2*	0,584	0,886	0,704
...	*non équilibré*	0,824	0,596	0,692
QUE lexical - *L7*	*1 sur 4*	0,68	0,823	0,744

TABLE 5.16 – Synthèse des performances des travaux antérieurs comparables, en comparaison des meilleures performances de nos classifieurs. Les performances sont données en terme de précision (P), rappel (R) et F-mesure (F), Accuracy (Acc) pour les travaux qui le mentionne (cette valeur correspond à une moyenne global des performances des classes événement et pas événement). La mention * indique que le chiffre présenté a été calculé par nos soins, alors qu'il n'était pas donné par les auteurs.

(dans le *VerbAction*) qui sont complément du nom. Par opposition, pour notre meilleure performance (test QUE *lexical - L7*), nous avons considérés tous les mots du *VerbAction* et les mots du lexique *EventNominals* et les valeurs de notre lexique pondéré ERW_{AFP}.

Bilan

Nous avons mené de nombreux tests sur les différents types de traits d'apprentissage qui nous paraissaient intéressants pour l'extraction des noms d'événements. Nous voulions intégrer les contextes phrastiques des noms dans les traits d'apprentissage, mais nous ne sommes pas parvenus à prendre en compte le contexte pour obtenir de meilleures performances. Finalement, nous arrivons à la conclusion que l'utilisation des lexiques reste la solution la plus efficace en français.

Le lexique pondéré en *ERW* est un bon complément aux informations des autres lexiques. Les performances des modèles sont améliorées lors de l'ajout de notre lexique pondéré comme trait d'apprentissage en plus des lexiques standards. De plus, le développement de sous-corpus suivant l'équilibrage en nombre de noms qui sont événements et

ceux qui ne le sont pas nous permet de développer des classifieurs en privillégiant soit la précision, soit le rappel.

Nous avons aussi développé un corpus de manière automatique sur la base d'amorces. Ces amorces sont des noms, triés sur le volet par nos soins, qui désignent toujours un événement ou alors qui ne désignent jamais d'événements. Nous avons créé ce corpus automatique dans l'objectif d'avoir un grand corpus de développement sans devoir annoter d'encore plus grandes masses de données. Un plus grand corpus correspond à un plus grand nombre de contextes d'apparition de noms événements ou non-événements et ainsi une quantité de données plus importante pour l'apprentissage. Les performances des sous-corpus automatiques sont quasiment toujours légèrement en-dessous des performances des sous-corpus manuels, et ce même lors des tests sur les lexiques. Nous pensons que parce que ce sont toujours les même mots-amorce qui sont utilisés pour créer ce corpus d'apprentissage, l'importance du lexique est limitée. Ce qui nous interpelle sur la nécessité d'améliorer les traits contextuels et leur prise en charge. Même sur un gros corpus, les traits contextuels ne suffisent pas à compenser la perte d'emprise du lexique sur ce type de corpus.

Conclusion générale

Synthèse

Dans ce mémoire, nous avons présenté les travaux de thèse qui avaient pour but l'étude des désignations nominales d'événements et leur l'extraction automatique.

Événement et typologie. Nous avons défini l'événement en reprenant les principaux points explorés par les travaux disponibles en sciences humaines et en traitement automatique des langues. Loin de vouloir proposer une nouvelle définition, nous avons plutôt compilé ce qui a déjà été avancé par les précédents travaux en retenant dans notre **définition** des caractéristiques qui conviennent dans le cadre de nos travaux et pour les désignations nominales des événements. Nous y intégrons le changement d'état, l'idée de récurrence ou d'instanciation, la durée, le moment de sa réalisation.

En lien avec la définition des événements et pour l'intégrer dans l'application que nous visons, nous avons développé **une typologie**, qui semble se dégager naturellement de la définition. Cette typologie fait ressortir les points importants en extraction d'information, les indications que nous souhaiterions pouvoir dégager de l'occurrence d'événement et qui sont toujours intrinsèquement portées par l'événement (modalité, fréquence et moment de réalisation).

Nous avons aussi dégagé de nos études sur corpus différents **types de formation de ces noms d'événements**, dont nous montrons que chacun peut être ambigu à des titres divers.

Guide d'annotation et corpus. Nous avons rédigé un **guide d'annotation** dédié aux événements et en particulier dans leur forme nominale. Ce guide est d'une aide précieuse pour l'annotation, il cible les difficultés d'annotation et permet de les outrepasser. Sur la base de ce guide d'annotation, nous proposons un nouveau **corpus dédié uniquement aux événements nominaux**. Dans notre corpus, seules les désignations nominales des

événements sont annotées. Notre corpus est plus grand que le corpus *FR-TimeBank* (en termes de nombre de mots), qui comporte tous les types d'annotations *TimeML*.

Nous avons étudié la capacité des **lexiques standards existants** à détecter les noms d'événements dans le but de les inclure dans nos travaux d'extraction des noms d'événements. Cette étude a été conduite en français sur notre corpus manuellement annoté et en anglais sur les noms du corpus *TimeBank1.2*. Nous avons créé des **règles d'extraction contextuelles** fondées en partie sur la syntaxe pour déterminer l'événementialité des mots. Ces règles nous ont permis de développer des lexiques pondérés en valeur événementielle.

Les lexiques pondérés que nous avons créés sont une valeur ajoutée pour l'extraction des noms d'événements, quand ils sont utilisés avec les lexiques standards. Les règles qui permettent d'extraire les lexiques pondérés sont adaptées à tous les textes en langage naturel en français (ou anglais), ainsi il est facile d'en extraire en passant d'un corpus à l'autre et en changeant de style (autre que journalistique).

Dans cette optique, un lexique pondéré pour les événements nominaux appris sur un autre corpus viendrait d'autant plus renforcer les performances des lexiques standards étant donné qu'il pourrait être issu de corpus de même type ou de type proche contrairement aux lexiques standards qui sont figés.

Pour extraire de manière automatique les noms d'événements, nous avons fait appel aux techniques de classifications automatiques. Nous avons utilisé des traits d'apprentissages contextuels, mais ceux-ci ne font jamais aussi bien que les lexiques standards.

Par ailleurs, nos expérimentations sur des corpus construits de manière automatique à partir de listes d'amorces choisies nous montrent que les lexiques ne sont pas à toute épreuve. Il faut donc persévérer en cherchant à obtenir de meilleures performances en utilisant des indices contextuels plus efficaces.

Même si les indices contextuels que nous avons utilisés ne sont pas très performants en termes de F-mesure, nous restons persuadés que par l'utilisation d'indices contextuels, nous pourrons améliorer les performances des classifieurs qui tiennent compte des lexiques standards et de notre lexique pondéré. De plus, au fil de nos expérimentations, nous avons développé des modèles qui ont une haute précision et ce type de modèle est de toute façon, en extraction d'information, une valeur ajoutée.

Améliorations et perspectives

Dans cette section, nous présentons succinctement les améliorations à apporter à nos travaux et nous donnons un aperçu des perspectives possibles en lien avec nos travaux.

Pour une meilleure extraction du lexique pondéré, les règles pourraient être améliorées et plus de vocabulaire ajouté afin d'atteindre une couverture plus large en français et surtout en anglais.

Par ailleurs, le contexte peut prendre une importance accrue si on se limite à un domaine particulier (ce que fait l'extraction d'information en général), et donc relancer l'espoir de pouvoir s'en servir.

De plus, de nombreuses améliorations sont à apporter en effet pour perfectionner les résultats sur l'anglais.

Passage à l'anglais. Nos travaux d'extraction automatique sont aboutis uniquement sur le français et nous envisageons de les porter sur l'anglais. D'autant que nous avons vu dans nos études sur les lexiques pondérés en anglais que l'approche lexicale est moins efficiente en anglais qu'en français. L'utilisation de traits contextuels pourrait amélorer significativement les résultats obtenus en anglais par le lexique. D'ailleurs dans (Bethard and Martin, 2006), les performances de leurs modèles utilisant des ressources syntaxiques ont de bien meilleures performances que celles que nous avons obtenues avec nos lexiques pondérés (section 4.1.2).

Intégration dans le projet *Quaero*. Même si ces travaux s'inscrivent dans le projet *Quaero*, nous aurions pu aller plus loin dans l'intégration des événements. En effet, notre guide d'annotation devrait être amélioré afin que des annotateurs autres que les auteurs puissent annoter des textes de manière efficace. De plus trop peu de documents du corpus en entités nommées générales annoté au cours du projet ont été annotés avec les événements. Il serait tout à fait envisageable d'enrichir tout le corpus des entités nommées au moyen des annotations en noms d'événements.

Évaluation par l'intégration dans une application. L'évaluation menée est basée sur la comparaison des performances des classifieurs automatiques par rapport à notre corpus manuellement annoté, considéré comme la référence. Nous envisageons d'intégrer les résultats de nos classifieurs les plus performants dans des applications existantes afin d'évaluer l'apport de notre approche. Les systèmes de question-réponse, par exemple, pourraient bénéficier sur des questions portant sur des événements de la reconnaissance

des noms d'événements. À la question « Quand a été célébrée l' *ouverture des Jeux Olympiques de Londres* ? », on attend une réponse sous forme de date (« le vendredi 27 juillet 2012 »). Mais il faut savoir reconnaître l'événement « *Jeux Olympiques* », savoir quand se sont déroulés les « *JO de Londres* », soit en 2012. Il faut aussi reconnaître que la « *célébration de l'ouverture des Jeux Olympiques* », correspond à la « *cérémonie d'ouverture* ». La reconnaissance des noms d'événements pourrait permettre une réponse plus fine aux questions portant sur les événements.

Appariement verbes-noms. Comme nous l'avons vu précédemment dans ce manuscrit, de nombreux travaux se sont focalisés sur les verbes désignant des événements. Ce n'est pas parce que nos travaux sont axés sur les noms d'événements, que nous ne considérons pas l'importance des verbes pour décrire les événements. Les verbes servent à décrire les événements et les noms servent à nommer les événements. Pour ne pas passer à côté d'informations sur les événements, il faut savoir extraire les noms comme les verbes.

De la même manière, une autre tâche intéressante est de parvenir à relier les différentes mentions des événements : mentions verbales et nominales, mais aussi les différentes désignations nominales des événements entre elles. Par exemple, les désignations suivantes réfèrent toutes au même événement : « la fuite de Ben Ali », « Ben Ali a fui », « Ben Ali a quitté la Tunisie ». « fuite » est à relier à « fuire », mais aussi à « quitter ». Les problèmes classiques de variation perdurent (« quitter » *vs* « fuir ») mais la normalisation a besoin d'un travail approfondi sur les désignations nominales.

Évolution des noms d'événements. Nous étudions l'évolution des noms des événements. Comme nous l'avons vu précédemment dans ce manuscrit, nous partons du principe que la désignation de l'événement est d'abord verbale avant de migrer vers le nominal, lorsqu'elle acquiert plus d'importance. Nous pensons qu'étudier la nomination des événements peut nous fournir des indices afin de repérer les noms d'événements et bien évidemment pouvoir relier les différentes mentions d'un même événement de manière automatique. Par exemple, nous souhaiterions repérer de manière automatique lors des événements qui ont été nommés « les *révolutions arabes* » et qui ont commencé au début de l'année 2011, les mots « agitation », « démonstrations », « émeutes », « révoltes », etc. avant le mot « révolution ».

Dans le même ordre d'idées, (Calabrese, 2011) illustre le phénomène de renomination d'événements au fil du temps. Son exemple est celui de la « *grippe H1N1* » qui avait été incorrectement nommée « *grippe porcine* » en mai 2009, et ensuite renommée au fil du temps « *grippe A* », « *grippe porcine ou nord américaine* » et finalement en « *A (H1N1)* ».

Situer les événements dans le temps est un projet intéressant et porteur. En effet, pouvoir intégrer la reconnaissance des événements nominaux et verbaux avec la reconnaissance temporelle nous permettrait de savoir quand un événement a eu lieu. Ce type d'information nous permettrait de retrouver des chronologies des événements dans un texte, ce qui rapproche nos intentions d'applications de type *TempEval*.

De plus, il serait alors possible d'élaborer des chronologies sur des thèmes particuliers (projet *Chronolines* sur lequel nous avons également travaillé).

Il serait aussi envisageable de répondre à des questions plus précises dans le cadre d'applications de question réponse et notamment de pouvoir répondre aux questions en considérant des intervalles de temps. Il serait aussi imaginable que nos travaux puissent améliorer les outils de résumé automatique multi-documents, car connaître l'enchaînement des événements permet de placer les phrases en provenance de plusieurs documents dans le bon ordre. En effet, les outils de résumé multi-documents automatiques pâtissent entre autres de la difficulté à repérer les événements relatés et leur position dans le temps. Même si les systèmes retiennent des phrases pertinentes pour constituer le résumé, les événements étant mal repérés, les systèmes se trompent souvent dans l'ordre dans lequel ils les positionnent. Le résultat est donc peu cohérent du point de vue chronologique. Ce défaut de performance est en partie causé par l'utilisation d'expressions temporelles relatives, mais aussi par une détection non approfondie des événements en général.

Annexes

Annexe A

Concepts évoqués / Définitions

- Allusion (cf. section 1.1.2.1, p. 22) : Selon http://www.cnrtl.fr/definition/allusion. Figure rhétorique par laquelle certains mots ou tournures éveillent dans l'esprit l'idée d'une personne ou d'un fait dont on ne parle pas expressément.
  Faire allusion à (qqn ou qqc.). Évoquer indirectement quelqu'un ou quelque chose ; faire (une rapide) mention de.
- Chrononyme (cf. section 1.1.2.3, p. 27) : Dans (Bacot *et al.*, 2008), nom de périodes historiques
- Contexte (cf. section 1.1.1, p. 21) : lié à la situation d'énonciation
- Cotexte (cf. section 1.1.1, p. 21) : contexte linguistique à l'intérieur du texte, ce qu'il y a autour du texte qui fait l'objet d'intérêt.
- Dénomination (cf. section 1.1.2.1, p. 22) : le nom donné, résultat de la nomination.
- Entités nommées (cf. section 1.1.3.1, p. 30) : Dans (Ehrmann, 2008).
- Entités nommées étendues dans *Quaero* (cf. section 1.1.3.1, p. 31) : Dans (Grouin *et al.*, 2011)
- Héméronyme (cf. section 1.1.2.3, p. 29) : Dans (Calabrese, 2008), les dates qui font événement, ou les noms d'événement formés sur la base essentielle et unique de la date à laquelle l'événement s'est produit.
- Moment discursif (cf. section 1.1.1, p. 19) : Dans (Moirand, 2007)
- Nominalisation (cf. section 1.2.3.2, p. 44) : le passage à un groupe nominal.
- Nomination (cf. section 1.1.2.1, p. 22) : le processus de nommer, l'acte de nommer, dont le résultat est une dénomination.
- Polyréférentialité du toponyme (cf. section 1.1.2.3, p. 26) : Dans (Lecolle, 2006).
- Praxonyme (cf. section 1.1.2.3, p. 28) : Dans (Bacot *et al.*, 2008), nom propre d'événement
- Référence (cf. section 1.1.1, p. 21) : définition de (Charolles, 2002)

Annexe B

Mesures utilisées

Certaines des informations reprises dans cette partie des annexes ont été récupéré de l'encyclopédie libre en ligne *Wikipedia*[1].

B.1 Précision, rappel et F-mesure

Toutes nos expérimentations sont évaluées en fonction des valeurs de précision (P), le rappel (R) et la F-mesure (F).

La précision est définie comme la probabilité observée pour un élément d'être correct.

$$P_{(i)} = \frac{nombre\ documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ de\ documents\ attribués\ à\ la\ classe\ i} \tag{B.1}$$

Le rappel est la probabilité observée pour un élément référencé d'avoir été trouvé dans la classe correspondant à la référence.

$$R_{(i)} = \frac{nombre\ documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ de\ documents\ appartenant\ à\ la\ classe\ i} \tag{B.2}$$

La F-mesure est la pondération combinée de précision et rappel. La F-mesure utilisé est le *F1-score*, où précision et rappel sont combiné de façon égale.

$$F_1 = \frac{2\ *\ (P * R)}{P + R} \tag{B.3}$$

1. http://fr.wikipedia.org

Les valeurs d'évaluation sur les classifieurs automatiques que nous donnons dans ce mémoire sont toujours celles obtenues sur la classe EVENT. Parce que ce que l'on cherche à évaluer c'est notre capacité à extraire les noms d'événements, nous n'avons pas présenté de valeurs globale de classification des classes événements et non-événement, ni de la classe non-événement.

Accuracy et précision Dans une classification binaire, les mesures statistiques d'Accuracy et de précision ne sont pas équivalentes (cf. http://en.wikipedia.org/wiki/Accuracy_%26_precision sur l'encylopédie libre *Wikipedia*).

		Condition as determined by *Gold* standard		
		True	**False**	
Test outcome	**Positive**	**True positive**	**False positive**	→ Positive predictive value or Precision
	Negative	**False negative**	**True negative**	→ Negative predictive value
		↓ Sensitivity or recall	↓ Specificity (or its complement, Fall-Out)	Accuracy

FIGURE B.1 – Schéma en anglais de définition de précision et accuracy en fonction des valeurs positives et négatives. Source : *Wikipedia* sur http://en.wikipedia.org/wiki/Accuracy_%26_precision.

Les *vrais positifs* étant les résultats correctement attribués à la classe et les *vrais négatifs* étant les résultats attribués à l'autre classe et qui sont dans la référence attribués à l'autre classe. Par opposition, les *faux négatifs* sont les résultats attribués à l'autre classes qui sont d'après la référence à attribuer à la classe qui nous intéresse et les *faux positifs* sont les résultats attribués par notre modèle à la classe qui nous intéresse, mais qui d'après la référence sont à attribuer à l'autre classe.

D'une part, l'accuracy (Acc) correspond à la proportion de résultats positifs (vrais positifs et vrais négatifs) dans les résultats :

$$Acc = \frac{nombre\ de\ vrais\ positifs + nombre\ de\ vrais\ négatifs}{nombre\ de\ vrais\ positifs + faux\ positifs + faux\ négatifs + vrais\ négatifs} \tag{B.4}$$

D'autre part, la précision (P) est définie comme la prportion de vrais positifs par rapport à tous les résultats positifs (vrais positifs et vrais négatifs) :

$$P = \frac{nombre\ de\ vrais\ positifs}{nombre\ de\ vrais\ positifs + faux\ positifs} \tag{B.5}$$

B.2 Accord inter-annotateur

Les annotations manuelles de corpus sont évaluées par un accord inter-annotateur. Nous avons utilisé le coefficient *Kappa* décrit dans (Cohen, 1960) et présenté sur *Wikipedia* sur http://fr.wikipedia.org/wiki/Kappa_de_Cohen. Cette mesure compare les accords en rapport avec ce qui serait obtenu par le hasard.

Le calcul du Kappa se fait de la manière suivante :

$$\kappa = \frac{\Pr_{(a)} - \Pr_{(e)}}{1 - \Pr(e)} \tag{B.6}$$

Où $Pr_{(a)}$ est l'accord relatif entre codeurs et $Pr_{(e)}$ la probabilité d'un accord aléatoire.

Selon (Landis and Koch, 1977), de 0,6 à 0,8 l'accord est considéré comme bon. Au delà de 0,8, il est très bon.

B.3 Pondération *ERW*

La pondération Eventiveness Relative Weight (ERW) est utilisée pour calculer l'événementialité des noms des lexiques ERW (section 4.3).

$ERW(w)$ est le nombre d'occurrences $e(w)$ du mot w extrait par les règles, divisé par le total d'occurrences de ce mot dans le corpus $t(w)$:

$$ERW(w) = \frac{e(w)}{t(w)} \tag{B.7}$$

B.4 Test t de Student

Nous avons évalué la significativité des performances de nos classifieurs automatiques au moyen des valeurs obtenues par la test de Student, lors de la comparaison de deux classifieurs sur le point de la précision.

Nous nous sommes fondé sur les valeurs de ce test, notamment lors des essais de combinaison en cascades de deux modèles appris sur les même jeux de traits d'apprentissage (section 5.3.4).

Description du principe du test dans *Wikipedia* sur http://fr.wikipedia.org/wiki/Test_t **:**

« Le principe du test t est le suivant : on veut déterminer si la valeur d'espérance μ d'une population de distribution normale et d'écart type σ non connu est égale à une valeur déterminée μ_0. Pour ce faire, on tire de cette population un échantillon de taille n dont on calcule la moyenne $\overline{x}$ et l'écart-type empirique s. Selon l'hypothèse nulle, la distribution d'échantillonnage de cette moyenne se distribue elle aussi normalement avec un écart type $\sigma = s/\sqrt{n}$. La variable :

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \tag{B.8}$$

suit alors une loi de Student avec n-1 degrés de liberté. »

Interprétation de la valeur du test t : En statistiques, il est admis qu'à partir d'une probabilité de $0,05$ (ce qui correspond à 5 % de chances d'être obtenue par hasard), une valeur est significative et qu'elle est très significative en-dessous de $0,01$. Ceci vaut pour la valeur du test t de Student.

Annexe C

Ressources

C.1 Listes d'amorces

C.1.1 La liste *EventNominals* de (Bittar, 2010a)

aberration abnégation absence absentéisme abus accalmie accident accord acrobatie activisme acupression adénoïdectomie adultère affaire allègement allopathie amniocentèse amygdalectomie anaphorèse anaphylaxie anastylose androgenèse ânerie angiogenèse angiographie angioplastie anniversaire anomalie anthropogenèse anthropométrie anticoagulothérapie apéritif apéro aphérèse apocalypse apologie apoplexie apoptose apostasie apothéose appendicectomie aquaplaning artériographie arthrographie arthroplastie arthroscopie arythmie assemblée assertion assises astreinte atelier atrocité attachement attentisme attirance audience audioconférence aumônerie autoallumage autocensure autocontrôle autocorrection autocritique autodafé autodiagnostic autoévaluation autolyse automassage autoplastie autoradiographie autorégulation autoritarisme auto-sondage autothérapie avalanche avarie aventure baby-sitting bacchanale backup bâillon bal ballade ban banqueroute banquet barbaque barbare barbarie bavure bêtise beuverie bévue biennale bioérosion biogenèse biopiraterie biopsie biosynthèse bioterrorisme bizarrerie blacklistage blague blessure blocus boom bordel boucherie bouchon bouffonnerie boum bourde brainstorming bronchectasie bronchospasme brûlure brunch brutalité budget calamité campagne candidature canicule cannibalisme canular capacitation caprice caravaning carême carnage carnaval carrière cassure catabolisme cataclysme cataphorèse catastrophe catéchise catéchisme cauchemar célèbre centenaire cérémonie césarienne césure challenge championnat chaos charte châtiment chimiosynthèse chimiothérapie chimisorption chirurgie cholangiographie cholécystectomie chromatographie chromoblastomycose clitoridectomie coéducation cofinancement coincement collapsus collimation colonoscopie coloscopie

colostomie colposcopie coma come-back commanderie commemoration commotion comparatisme compétition complication composant compromis concile concordat conférence configuration conflit congé congestion congrès conjoint connerie conscience conscription conseil consensus constipation contentieux contracture contre-attaque contrecoup contremesure contre-passation contre-performance contre-révolution contre-transfert contrevisite convalescence conviction corrida corvée cotutelle coup couper-coller craniotomie crapuleux crash crémaillère crémation crépi crime crise croisade croisière crue cryochirurgie cure cyberguerre cyclone cytolyse débalancement débauche débilitation débogage déboguage debriefing débriefing début décade décathlon déclic décoction décrantage décret décriminalisation défection déflation défragmentation dégobage délabrement délai délaminage délamination délation délettrage délit déluge déméchage demi-réaction demitour démo démultiplexage dépigmentation dépilage-empilage déplaisir déportement dépotentionalisation dépressage dépression dépulsage derby déréglage dérégulation déremboursement désaccord désastre déséquençage désincitation désinscription désinsectisation désinstallation désintérêt détonation deuil dévolution dévotion diarrhée diaspora dichotomie dictature diète différence dimension discrétisation disfonctionnement dispense dissidence distance diversité divorce domaine domicile dommage donation-partage drame drogue droitisation dynastie dysfonctionnement échéance échec échographie électrocardiogramme électrochoc électrodéposition électromyogramme éliminatoire éloge emailing embargo embolie émeute emmerde émolument emplette encéphalogramme énervement engourdissement enrochement entorse entracte entrainement entrevue épidémie éruption eurotransfert événement évènement excentricité exécration exhalaison exode expérience exposé famine fanfare farandole farce fausse-couche félonie festival festivité feu fibrage fibrillation fibrose fieldwork filiation film fin finale fixing flatulence flop fonctionnairicide fonctionnariat footing formalité forum fracas frasque fumisterie funéraille funérailles gabegie gâchis galvanoplastie gastroscopie gêne génocide génotypage génuflexion géocodage gestation geste glandouille goujaterie gouvernance grand-messe grève grossesse grutage gueulante guidonnage hagiographie hallucination hémorragie hernie histogenèse histoire holocauste hoquet hostilité hyper-pigmentation hypersécrétion hypoglycémie hystérie immunocapture implémentation impolitesse imprécation impunité inaction inactivité inaptitude incartade incendie inceste incident incivilité indiscrétion infanticide infarctus inflammation initiative injustice insuccès interblocage interchangement intransigeance intro intrusion inventaire investiture jeux journée kermesse laryngoscopie leçon libation lithogenèse litige loterie loto lunaison lustration magasinage mailing maillage maladie maladresse malaise malentendu malfonctionnement malformation malheur malnutrition mammographie mandat manif manip mappage marathon mascarade mastering match maturité méchoui mécontentement médiation meeting méiose ménopause messe meurtre mi-

cropaiement micropayement microtraumatisme miction miracle modif mondial monstruosité morphogenèse mousson multi-recomposition multitexturing musculation mutagénèse myomectomie nazisme neige néo-colonisation néovascularisation neuropiratage neutropénie nidation non-accentuation non-autorisation non-conclusion non-dénonciation non-exécution non-fécondation non-harmonisation nuptiales obligation obsèques occurrence ollie orage oraison orgasme ostéoplastie ouragan pagaille pandémie panne papilonnage paralysie paramétrage paraplégie parcellarisation parricide partenariat partition partitionnement parturition pataquès pâtée pélerinage pénalité pénurie perfection performance périple péristaltisme perpétration pétition photocopie photosensibilisation photosynthèse pitrerie pixelation planétarisation plangage pluie polycondensation pot pourparlers préalphabétisation précarisation précompréhension préconsultation prédiagnostic prédimensionnement prédisposition préenregistrement préfaçonnage pré-inscription préinscription prélimage prépublication prérecrutement préréglage pré-réservation préréservation présidentielle problème procédé procédure procès processus pseudo-défaite pupaison putsch quarantaine quart question quinquennat raccord raid ramadan rancard rando ratonnade réaffichage réappropriation réattribution rebuffade recadrage recalibrage recapitalisation récession rechercher recolonisation recompte reconnexion redécollage redéveloppement redimensionnement redirection redoux rééchantillonage rééchantillonnage rééchelonnement ré-élection réenclenchement réencodage réestimation réétiquetage refacturation référendum refondation reformatage regain réhospitalisation réimputation réinitialisation relookage relooking remappage remèchage rémission rémunération rendez-vous rendormissage rendu renfoncement réordonnancement repas repéchage repositionnement répulsion resérialisation ressentiment restylage retaxation retéléchargement rétribution retrogradage rétrolavage réunion-concertation revers rhinoplastie rite rituel rixe rotation sabbat safari saisine saison salai salât salon sandboxage sauvagerie scénario scène schisme sécession sécheresse sédition séisme semi séminaire séquestre session sexocide shoah show siège signalement sinistre soirée soirée-débat solution sommet sonde sottise spectacle stage stagnation steppage subvention succès supercherie suraccumulation surcommercialisation surcompactage surconsommation surdosage surendettement surmédiatisasation surmédicalisation surpâturage surpêche surpopulation surréglementation surrénalectomie surséchage sus-occlusion symbiose téldéclaration téléchargement téléconférence télédéclaration télédémarche télépaiement téléprocédure télérèglement téléréunion téléthon télétransmission téléversement télévirement tempête tendinopathie tétrahydropyranylation tohu-bohu tombola tornade tournoi tracas tragédie trahison trajet transition traumatisme travaux trêve triathlon tsunami upgrade upload urgence valuation vandalisme vantardise vautrage veulerie vice vidéoconférence vigile virée voyage-lecture walkathon web-conférence zoom

C.1.2 Les amorces de (Resnik and Bel, 2009) en espagnol

Noms événement : fiesta misa serenata sequía lío feria vacaciones espectáculo catástrofe follón festival receso show cataclismo problema boda excursión programa desastre motín funeral trayecto película tragedia huelga velorio/velatorio travesía ciclo holocausto incidente ceremonia clase discurso drama boicot evento conferencia sermón incendio pánico picnic curso torneo accidente miedo cóctel taller campeonato impacto pasión té workshop carrera siniestro furor banquete congreso rally caos rabia festín simposio tormenta crisis siesta ágape jornadas tempestad guerra frío tertulia tumulto temporal batalla calor campaña coloquio borrasca conflicto hambre cónclave entrevista terremoto paz pereza cumbre audiencia sismo silencio dolor asamblea concierto huracán ruido fiebre sesión ópera maremoto escándalo gripe

Noms non-événement : mapa madera garganta economía herramienta antología cifra maqueta figura factura característica habitación dato mar miembro droga fotocopia volcán pancarta forma plasma vivienda cárcel grupo tema teléfono gas familia arma fuente montaña literatura dinero informe temperatura tubo especie estereotipo diario euro estética paisaje tarifa trama ilusión cliente diferencia compañía zona punto escena carretera justicia misterio batería colectividad seguridad humo facultad silueta canal red balneario cadáver unidad arquitectura contraseña paquete nivel organismo cara rodilla prensa pista norma levedad virus vehículo columna vía estadio cantidad dueño combustible planta batuta provincia prejuicio estructura autobús súbdito detalle banda ruta perspectiva ciudad público consorcio alimento antena

C.1.3 Les amorces de (Bel *et al.*, 2010) en anglais

Noms positifs (événements) : accident assembly audience battle boycott campaign catastrophe ceremony cold collapse conference conflict course crime crisis cycle cyclone change choice decline disease disaster drought earthquake epidemic event excursion fair famine feast festival fever fight fire flight flood growth holiday hurricane impact incident increase injury interview journey lecture loss meal measurement meiosis marriage mitosis monsoon period process program quake response seminar snowstorm speech storm strike struggle summit symposium therapy tour treaty trial trip vacation war

Noms négatifs (non-événements) : agency airport animal architecture bag battery bird bridge bus canal circle city climate community company computer constitution country creature customer chain chair channel characteristic child defence director drug economy ecosystem energy face family firm folder food grade grant group health hope hospital house illusion information intelligence internet island malaria mammal map market mountain nation nature ocean office organism pencil people perspective phone pipe plan plant profile profit reserve river role satellite school sea shape source space star statistics store technology television temperature theme theory tree medicine tube university visa visitor water weather window world

C.1.4 Les listes « surs » et « surs_pas » pour la création de notre corpus artificiel d'apprentissage automatique

Noms « surs » (événements) : abandon changement explosion plaidoirie révolution accident condamnation festival poursuite saccage acquisition confrontation frappe procès scandale agression décès incident profanation sortie amélioration défaite inculpation ramassage suicide annulation démission manifestation récession surprise arrestation déplacement massacre recluson tentative arrivée diffusion meurtre recul tournoi ascension disparition nomination rencontre transaction attentat élection panique renouvellement trêve atterrisage élections panne retour viol bataille enlèvement parachutage retrait bombardement exaction perte réussite cambriolage exploit pique-nique révélation

Noms « surs_pas » (non-événements) : afghanistan compagnie homme niveau rpr alain compte image nom rue allemagne corps indice nombre saint allemand cours information nord secteur allemande directeur internet nuit seine argent directrice jacques objet sens arme dirigeant janvier onu septembre art dix jean ordre service auteur dollar jospin page site avenir droit jour paris situation banque droite juge parole sud ben enfant justice partie sujet bernard entreprise laden paul taliban besoin euro ligne pays taux bourse exemple liste personne technique britannique face livre philippe temps bush faim loi pierre terrain cadre famille lundi place texte campagne femme mai point titre candidat fille maire police tour candidate film maison politique travail capitale finance marie presse trente carte fond mars prix type cas force membre produit udf cause forme mercredi programme un centre franc michel projet valeur chef france milliard ps vie chiffre gens million quatre ville chirac gouvernement ministre question voix chose groupe mois raison washington cinq heure monde rapport zone claude histoire moyen responsable

C.2 Le guide d'annotation en événements nominaux

Entités Nommées Événement : guide d'annotation du projet Quaero

Béatrice ARNULPHY et Xavier TANNIER

Version 3.2 du 08/11/2012

Ce guide traite des annotations pour les entités nommées événement (EN-É). Il est développé dans le cadre du projet de recherche Quaero et fera partie intégrante du guide d'annotation entités nommées étendues Quaero [Rosset *et al.*, 2011]. Ce guide est une version temporaire de travail.

Nous avons annoté les entités nommées (EN) autre qu'événement dans des cas particuliers qui le nécessitaient. En effet, les EN-É sont à annoter en plus des entités nommées Quaero déjà définies dans le guide d'annotation entités nommées étendues Quaero [Grouin *et al.*, 2011].

Dans une première partie, nous présentons les annotations de manière simplifiée. Tous les exemples sont repris dans la deuxième partie. Ils sont alors annotés compte tenu de la typologie.

1 Définition

selon notre définition dans [Arnulphy, 2012], l'événement est défini comme :

> **ce qui survient, un changement d'état opéré et pas un état. Il peut être récurrent ou unique, prévu ou non. Il peut durer ou être instantané. Il peut aussi se produire dans le passé, le présent ou le futur**

L'EN-É est déterminée comme le syntagme nominal qui désigne l'événement. Il s'agit du nom de l'événement, tel qu'on le désigne ou l'une de ses dénominations et uniquement les groupe nominaux.

(1) `le <event> festival de Cannes 2010 </event> s'est déroulé du 12 mai au 23 mai 2010.`

Les expressions verbales qui désignent un événement (selon Vendler [Vendler, 1959] ou dans le cadre de TimeML [Pustejovsky *et al.*, 2003]) ne sont pas à annoter. Dans l'exemple suivant, le verbe "*jouer*" n'est pas considéré, le nom "*repas*" oui.

(2) `le chat joue avec une balle pendant le <event> repas de son maître </event>.`

1.1 Imbrication de balises

1.1.1 Informations supplémentaires sur les composants de <event> : composants transverses

Une partie de l'expression d'une entité nommée qui désigne un hyperonyme (genre proche) de l'entité est annotée avec <kind>[1]. Ol est à noter que lorsqu'un <kind> est annoté dans une annotation événementielle, il s'agit toujours de la tête de syntagme événementielle. L'annotation du mot en <kind> est maintenant uniquement symbolisée par le souligné.

(3)
```
<event>
    <kind> festival </kind> de Cannes
</event>
```

(4)
```
<event>
    <kind> réunion </kind> houleuse à la Bourse du travail
</event>
```

Certains articles sont annotés comme partie intégrante de l'<event> et reçoivent la balise particulière <event-modifier>. Par exemple, les tantièmes sont annotés <ordinal>. Ces phénomènes est présentés dans la section 1.5 dédiée au traitement des déterminants.

1.1.2 Annotations imbriquées d'EN

Les annotations imbriquées **avec des entités nommées des autres types** sont légitimes. Dans "*festival de Cannes*", "Cannes" est un lieu/nom de ville, il devra être annoté <loc.adm.town>. Cette annotation comlète à l'intérieure du syntagme événementiel ne sera plus présentée plus loin, mis à part dans des cas particulier de représentation.

(5)
```
<event> festival de
    <loc.adm.town> Cannes </loc.adm.town>
</event>
```

Les **annotations imbriquées d'événements** sont autorisées, pour des dénominations d'un événement général par rapport à l'événement particulier ou un événement lié à un autre, par exemple "*le 60ème festival de Cannes*". L'événement hyperonyme étiqueté en règle général à l'intérieur de l'événement particulier.

(6)
```
<event>
    <ordinal> 60ème </ordinal>
    <event> festival de Cannes</event>
</event>
```

1. Définition du guide EN étendues Quaero : Partie d'une expression d'entité nommée qui désigne un hyperonyme (genre proche) de l'entité : "`le maire de Paris`" : `le <func.ind> <kind> maire </kind> de <loc.adm.town> Paris </loc.adm.town> </func.ind>` Dans cet exemple, maire est une fonction de type <func.ind>

(7)

<event> réunion houleuse après une
<event> assemblée générale à la Bourse du travail </event>
</event>

Exemple particulier dans la typologie (section 2.2) : Les événements frequence="recurring" sont à annoter à l'intérieur des événements frequence="instance".

(8)

au moment des
<event type="factual" frequence="instance" temp="before"
source="notreport">
<event type="factual" frequence="recurring" temp="now"
source="notreport">
Jeux olympiques
</event>
de <time.date.abs> 1988 </time.date.abs>
</event>

(9)

le
<event type="factual" frequence="instance" temp="after"
source="notreport">
<event type="factual" frequence="recurring" temp="now"
source="notreport">
Mondial
</event>
</event>
s'achève le 30 juin.

1.2 Connaissance du monde

Certains événements ne peuvent être correctement annotés sans faire appel à sa connaissance du monde, différente pour chacun et aussi pour ls systèmes automatiques. Il est délicat de demander aux annotateurs de ne pas prendre en considération leur connaissance du monde et de s'en tenir uniquement aux informations fournies dans le cotexte d'un événement. De plus, on ne peut limiter les ressources qui pourrait être utilisées dans le cadre de l'extraction automatique. C'est pourquoi, l'annotateur est invité à annoter en fonction de ses connaissances de la réalité des événements.

Cette règle s'applique lorsqu'on doit appliquer des annotations en fonction de la typologie définie en section 2. Prenons l'exemple de la récurrence pour illustrer notre point de vue.

Dans l'exemple suivant, mis à part notre connaissance personnelle des événements, aucune indication dans la phrase ne nous permet de noter la récurrence du festival de Cannes, pourtant cet événement notable est récurrent.

(10) Au <event frequence="recurring"> Festival de <loc.adm.town> Cannes </loc.adm.town> </event>, il est presque passé inaperçu.

(11) Le <event frequence="instance" temp="after"><event frequence="recurring" temp="now"> sommet de l'Union pour la Méditerranée (UpM) </event></event> a été reporté au mois de novembre.

1.3 D'autres entités nommées qui sont événement

Certaines entités nommées (temps <time.date. > ou de lieu <loc. >) peuvent être en réalité des événements en situation particulière.

1.3.1 Les entités nommées <loc.>

Prenons l'exemple de ''Tchernobyl" pour décrire ce phénomène pour les noms de lieu.

(12) À quand la vérité sur les retombées radioactives de <event> <loc.adm.town> Tchernobyl </loc.adm.town> </event> !

(13) Le nuage de <event> <loc.adm.town> Tchernobyl </loc.adm.town> </event> ne s'est pas arrêté aux frontières françaises, comme l'avaient affirmé à l'époque les médias et les politiques.

Ces exemples nous montrent qu'un nom de ville peut référer dans certains cas à un événement et non plus directement à un lieu. Il conviendra de l'annoter en <event> tout en gardant les indications sur le nom de lieu.

Par contre, dans "Cette zone a été polluée par les retombées de la <event> catastrophe nucléaire de <loc.adm.town> Tchernobyl </loc.adm.town> </event> (1986)", le mot *Tchernobyl* désigne le lieu et l'événement est nommé dans *la catastrophe de Tchernobyl*.

1.3.2 Les entités nommées <time.date. >

"11 septembre", exemple pour les dates qui sont événement.

(14) La découverte d'autres pilotes terroristes avant le <event><time.date.abs> 11 septembre </time.date.abs></event> aurait pu limiter les attaques et les pertes en vies humaines

(15) On ne parlait que de ça : le <event><time.date.abs> 11 septembre </time.date.abs></event>, l'islam, l'Afghanistan, l'anthrax, l' <event> absence des Américains </event>, les incertitudes économiques.

Par contre, *11 septembre* est uniquement une date dans "l'attaque du 11 septembre signe peut-être la fin de cette hégémonie sur le reste du monde", le syntagme complet "*attaque du 11 septembre*" étant l'événement.

(16) l'<event> attaque du <time.date.abs> 11 septembre </time.date.abs></event> signe peut-être la fin de cette hégémonie sur le reste du monde

Dans l'exemple suivant, les deux cas sont possibles pour "11 septembre" : date et événement. Il faut les annoter tous les deux. Pourtant d'après le contexte global (un article de presse relatant les difficultés rencontrées par les compagnies aériennes depuis les attentats du 11 septembre aux États-Unis), on en conclu qu'il s'agit de l'événement "11 septembre".

(17) "Nous n'étions pas dans une bonne situation avant le <event> <time.date.abs> 11 septembre </time.date.abs> </event>, avec nos coûts dépassant nos recettes" , a expliqué James Goodwin, selon la revue spécialisée Air Transport World, mais depuis, "nos coûts excèdent nos revenus quatre fois plus qu'avant le <event> <time.date.abs> 11 septembre </time.date.abs> </event> " , a -t-il ajouté. '.

(18) "Depuis le <event><time.date.abs> 11 septembre </time.date.abs></event> , nous n'avons percu aucun <event> changement dans le discours de nos clients </event>, constate Sophie Romet, directrice générale associée de l'agence de design Dragon rouge et experte en "packaging"."

Des tests linguistiques peuvent permettre de différencier date et événement certains événements en situation ambigüe. D'après l'un des tests simples présentés dans [Ehrmann and Hagège, 2009], il suffit de tenter d'insérer l'expression "l'événement qui s'est produit le" avant la date que l'on présume être un événement. Si la nouvelle phrase est une paraphrase de la précédente, alors c'est bien un événement. Ainsi dans "l'attaque du 11 septembre signe la fin [...]", "11 septembre" correspond bien à la date, car l'expression suivante est incorrecte :

(19) * l'attaque de *l'événement qui s'est produit le* 11 septembre signe la fin [...]

Par contre dans la phrase "Depuis le 11 septembre, nous n'avons perçu aucun changement", Il n'est pas incorrect de dire :

(20) Depuis *les événements qui se sont produits le* <date> 11 septembre </date>, nous n'avons percu aucun <event> changement </event>

1.3.3 Des expressions de durée

Certaines expressions de durée peuvent aussi désigner des événements :

(21) Jean-Michel Pilc est , après <event> vingt ans de musique hautement personnelle </event> , salué avec insistance .

(22) Après <event> quelques semaines de baisse de tension </event> , de <event> véritables négociations </event> pourraient reprendre. [2]

(23) Le duo d'adieu qui l'unit à Rodelinda à l'acte II et passe dans un souffle , à l'instar de ces <event> trois heures de musique écoutées comme en rêve </event>

(24) Cet <event> opéra en trois actes </event> - <event> trois heures de musique, sublime pour l'essentiel </event> - est une partition qui fleure bon Wagner tout en étant d'une inventivité folle.

Pour suivre le fil d'exemples avec "musique", on note que seule durée de musique est événement. Ainsi, dans l'exemple suivant, il ne désigne pas un événement.

(25) Et que la *musique* commence.

1.4 Les frontières d'ENE

1.4.1 Les dépendances syntaxiques

Dans le cadre de l'annotation, il peut y avoir un problème de frontières. On fonde l'extraction sur la tête de syntagme et ses dépendants directs (compléments du nom, adjectifs, attributs, appositions ...). On ne conserve pas les compléments circonstanciels de temps et de lieu, car ils dépendent du verbe. **Seuls les dépendances nominales** peuvent être annotés (avec leur tête de syntagme) comme événements.

1. **Les dépendances nominales à conserver**

   – Les **participes passés**
   Il est plus difficile de trancher entre l'utilisation en tant que verbe ou nom pour un participe passé que pour un participe présent, c'est pourquoi, nous avons choisi d'annotés comme partie intégrante des <event>, tous les participes passés rattachés à des noms. Ils seront annotés ainsi que leurs accompagnants complément d'agent et sujet inversés, le cas échéant.
   Pour nous, les participes passés suivants sont équivalents :

   (26) <event> interview accordée pour le compte du journal Le Monde </event>

   (27) <event> interview accordée par le ministre fédéral </event>

   Exemples complets :

2. NB. *Baisse de tension* n'est pas annoté en particulier, il correspond au <kind>.

(28) L'UE - l'ANASE: <event> Interview *accordée par le ministre fédéral des Affaires étrangères, Frank-Walter Steinmeier*, au quotidien "Nürnberger Nachrichten" </event>.

(29) Le <event> sommet Chine - Union Européenne *annulé en décembre* </event> se tiendra à Prague, en République Tchèque, au mois de mai, selon le journal China Daily.

– Les **compléments de lieu**

(30) L' <event> attaque du train postal *dans la banlieue de Londres* </event> se déroula aux petites heures du matin.
Le complément de lieu *banlieue de Londres* est rattaché à *attaque*.

(31) La <event> nomination de la ministre de l'Intérieur Michèle Alliot-Marie *au poste de garde des Sceaux* </event>.
Au même titre que *la ministre de l'Intérieur*, *au poste de garde des Sceaux* est complément du nom *nomination*.

(32) Lors de la <event> réunion *à Paris* </event>, aucune décision n'a été prise.

(33) Malgré la <event> poursuite de la <event> " <event> sale guerre </event>" *en Tchétchénie* </event> </event>

En revanche,

(34) *À Paris*, les <event> défilés </event> sont toujours réussis.

(35) Les forces russes menaient une <event> opération de nettoyage particulièrement brutale </event> *dans la ville d'Argoun* .

– Les **compléments de temps** :
Dans "*les JO de 1996*", le <time.date.abs> *1996* est complément de temps de *JO*, il complète le nom particulier de ces Jeux Olympiques. De plus, cette indication de temps permet de placer ces JO sur l'axe temporel et de nommer une instance d'un événement récurrent.

(36) Les <event> <event> JO </event> *de 1996* </event>

En revanche, dans l'exemple suivant, *le 3 janvier dernier* est complément circonstanciel de temps donc à ne pas annoter.

(37) La banque a subi une <event> nouvelle attaque </event> *le 3 janvier dernier*.

2. Les dépendances nominales à ne pas conserver

– Les **subordonnées relatives**

(38) Le <event> sommet de l'Union pour la Méditerranée (UpM) </event> *qui devait avoir lieu à Barcelone les 6 et 7 juin* a été reporté au mois de novembre, a-t-on appris le 20 mai de source égyptienne.

(39) Le ministre de l'Ecologie Jean-Louis Borloo a annoncé samedi soir la <event> mise en place d'une "cellule de crise" </event> *qui réfléchit à la possibilité d'utiliser des aéroports militaires, quand la météo le permettra, pour soulager les aéroports civils qui seraient encombrés*.

Dans l'Ex. 39, si on avait conservé les subordonnées relatives, il y aurait sans doute un problème de frontières. On prendrait tout depuis "mise en place" jusqu'à "encombrés".
Ne pas oublier que les subordonnées relatives peuvent être introduites par "dont" :

(40) Lord Russell-Johnston justifiait la <event> décision prise le 26 janvier par l'Assemblée parlementaire du Conseil de l'Europe </event> , *dont il est le président*.

– Les **participes présents**
À moins bien sûr qu'ils soient utilisés en adjectif qualificatifs (s'accordent en genre et en nombre avec leur nom).
la <event> "vaine querelle" </event> *opposant les pionniers des soins palliatifs à ceux du droit à mourir dans la dignité*.

– Les **subordonnées infinitives**

(41) L' <event> incident </event> se serait produit lors d'un <event> forage </event> *pour placer des explosifs dans une couche de charbon*.

(42) lors de sa <event> première journée à Monterrey </event> , mardi, il a mentionné l' engagement de la France en faveur de la <event> création d'une taxe internationale </event> *pour financer les besoins des pays pauvres*

(43) Lord Russell-Johnston justifiait la <event> décision </event> *de restaurer tous ses droits à la délégation russe*.

(44) le <event> procès du colonel Boudanov </event> , seul militaire à ce jour *à avoir fait l'objet de* <event> *poursuites pour le* <event> *viol* </event> *et le* <event> *meurtre d'une jeune fille tchétchène* </event> </event> , tourne à la <event> farce tragique </event>

1.4.2 Les guillemets

Plusieurs cas de positionnement de guillemets sont possibles. Voici des indications sur la manière de s'y prendre, même si dans le cadre d'une évaluation, il faudrait considérer qu'une erreur de frontière due à la ponctuationdevrait être pas ou moins pénalisante, qu'une autre erreur d'annotation.

1. " <event> EVT </event> "
   Si le nom d'événement et uniquement lui est entre guillemets, on annote pas les guillemets.

   (45) La maison devient le prolongement du corps , proclament quelques capteurs de tendances priés de cogiter sur le thème de la “ <event> revalorisation des tâches ménagères </event> ”.

2. " <event> EVT </event> mots "
   Si le nom d'événement fait partie d'un groupe de mot entre guillemets, on n'annote pas le guillemet même s'ils se jouxtent.

   (46) métaphore musicale de la tension exponentielle développée par un inexorable leitmotiv , “l'important , ce n'est pas la <event> chute </event> , c'est l' <event> atterrissage </event> ” , cette bande originale ne se contente pourtant pas d'illustrer l'intrigue.

3. <event> " EVT- " -EVT </event>
   On annote naturellement les guillemets inclus dans le nom d'événement, sans soucis d'annoter le guillemet ouvrant ou fermant manquant, s'il ne jouxte pas le groupe de nom annoté.

   (47) Malgré la <event> poursuite de la <event> " <event> sale guerre </event>" *en Tchétchénie* </event> </event>

   (48) Présence renforcée du FSB, notamment pour des “ <event> opérations spéciales </event> ” , <event> arrestations </event> et <event> meurtres “ciblés” </event> .

1.5 Les déterminants

Dans les directives pour les autres entités nommées définies *Quaero*, les déterminants ne sont pas annotés à l'intérieur des groupes désignants des entités nommées, mis-à-part pour les entités nommées temporelles, de date ou d'heure. Même si les événements sont fortement temporels de par leur ancrage dans le temps, nous avons finalement choisi d'annoter les événements comme la plupart des entités nommées *Quaero* en n'autorisant pas les déterminants dans le bloc annoté.

Même s'il est vrai que de nombreux déterminants apportent une information supplémentaire à l'entité, nous avons choisi qu'il ne fasse pas partie intégrante du nom d'événement en tant qu'entité nommée événementielle. Pour autant,

ces déterminants sont des modifieurs de l'événement, ils héritent de la balise <event-modifier>. Dans les exemples suivants, c'est le cas de “cet” pour le nom d'événement “cet important remaniement” (Ex. 58), mais pas de l'article défini “le” dans l'Ex. 50 .

(49)
```
la surprise de <event-modifier> cet </event-modifier>
    <event>
        important remaniement
    </event>
```

(50)
```
Ça a créé la
    <event>
        surprise
    </event>.
```

(51) Autres exemples :
```
Ça a créé la <event> surprise </event> .
La <event> réunion </event> se tiendra [...]
C'est une <event> formidable surprise </event> !
Lors du <event> conseil de laboratoire de mardi dernier
</event> [...]
```

Notons que contrairement à certaines idées reçues, la détermination d'un nom d'événement pas un **article indéfini** est tout à fait possible.

(52) Les dirigeants européens pourraient contraindre Moscou à accepter *un* <event> cessez-le-feu </event> et *des* <event> négociations politiques avec les autorités tchétchènes </event> .

Attention ! Dans l'Ex. 53, le mot “manifestation”, déterminé par l'article indéfini “un” désigne l'objet.

(53) Lors d'une manifestation syndicale, il y a toujours [...]

Dans l'Ex. 54, l'article indéfini “un” n'est pas annoté dans le syntagme “un autre attentat”.

(54) Et pour accroître le désarroi, on apprend qu'un autre <event> attentat </event> se serait produit à moins de 100 km d'Alger ...

L'article “un” n'est pas annoté dans le syntagme “un autre attentat”, le nombre “deux” serait à annoter. On ne fera pas de différence entre le nombre et le déterminant dans ce genre de syntagmes ambigus. En effet, “un autre” est considéré comme le déterminant de “attentat” dans sa globalité.

Les **nombres** sont annotés comme le veut le guide d'annotation des entités nommées étendues de Quaero comme des <unit>, à l'intérieur d'un <amount> (cf. Ex. 55). La désignations nominale d'événement dans ce contexte est donc à l'intérieur de la balise <object> (cf. Ex. 56).

(55)
```
<amount>
    | <unit> deux </unit>
    | <object>
    |   <event> attentats </event>
    | <object>
<amount>
```

(56)
```
les <amount>
        | <unit> deux </unit>
        | <object>
        |   <event> attentats de <loc.> NY </loc.>
        |   </event>
        | <object>
    <amount>
```

(57)
```
les <amount>
        | <unit> deux </unit>
        | <object>
        |   <qualifier> précédentes </qualifier>
        |   <event> élections
        |   </event>
        | <object>
    <amount>
```

Nous considérons que les **ordinaux** font partie intégrante de certains noms d'événements, c'est le cas des événements récurrents. C'est pourquoi, suivant les directives du guide d'annotation des entités nommées étendues de Quaero, il conviendra d'utiliser la balise <ordinal> pour annoter *60ème* dans "60ème festival de Cannes". Nous n'illustrerons pas ce point plus en détail dans ce guide.

(58)
```
le <event>
    <ordinal> 60ème </ordinal>
    <event> festival de Cannes</event>
</event>
```

1.6 Les conjonctions de coordination

Les conjonctions de coordinations qui relient deux événements marquent une frontière entre les deux événements différents.

(59) Gestion des <event> troubles </event> et des <event> perturbations </event> pédiatriques du sommeil.

Si deux évenements ou plus sont déterminés par un même groupe de mots et que ces deux événements sont coordonnés par une conjonction de coordination, les deux événements seront annotés indépendamment et l'événement contigu au groupe qui les qualifie sera annoté avec (cf. Ex. 61).

Exception faite lorsqu'il s'agit d'une éventualité, comme dans l'Ex. 62, où on ne peut séparer "retour" et "non", vu qu'il sont fortement dépendant sémantiquement et que "non" perd de tout son sens sans sa coordination avec "retour".

(60) Gestion des <event> troubles </event> et des <event> perturbations </event> pédiatriques du sommeil.

(61) la liste des <event> *méfaits* </event> ***et*** <event> *exactions* commis entre 0h30 et 4h mardi </event> ressemble à une <event> litanie sans <event> fin </event> </event> ...

(62) la fine arithmétique électorale dépend du <event> *retour ou non* au Parlement des néocommunistes du PDS, héritiers du Parti communiste au pouvoir dans l'ex-RDA </event> .

1.7 Les constructions à verbe support

Dans le cas des constructions à verbe support, on peut se poser la question de savoir si l'événement est porté par le verbe ou le nom. Pour les entités nommées dans Quaero, on annote le nom "attaque" dans "*mener une attaque*". Même si il y a un verbe support, l'événement reste porté par le nom.

mener une <event> attaque </event>
lancer une <event> attaque </event>
effectuer une <event> rentrée </event>
créer la <event> surprise </event>

Autre exemple : Il y a ici trois événements à annoter : "***nomination***, "***surprise***" et "***remaniement***". Du point de vue sémantique, la "surprise" et la "nomination" sont le même événement, même si la surprise est une conséquence et perçue par un autre public. La nomination est le résultat-conséquence du remaniement.

(63) La <event> nomination de la ministre de l'Intérieur Michèle Alliot-Marie au poste de garde des Sceaux </event> crée la <event> surprise de cet <event> important remaniement </event> </event>

2 Typologie

L'annotation se fera sous le format .xml. Une balise <event> peut comporter plusieurs attributs. Trois attributs sont obligatoires et à compléter par une seule valeur parmi les suivantes :

- la modalité (type=) : factual, hypothetical, nonfactual, abstract
- la frequence (frequence=) : unique, recurring, instance

– le moment de sa réalisation (temp=) : before, now, after

Par défaut, les valeurs de ces balises sont type="factual" frequence="unique" temp="before". Dans certaines configuration, des valeurs par défaut sont proposées.

Une information supplémentaire peut être apporté en lien avec la provenance de l'information (attribut source=) au moyen de la valeur reported, une autre dans le même attribut concernera la valeur fictive.

Comme dans le guide d'annotation Quaero pour les autres entités nommées, les valeurs unknown (je ne sais pas quelle valeur choisir parmi celles disponibles) et other (je sais qu'aucune des valeurs proposées ne convient) peuvent être utilisées par les annotateurs.

2.1 Modalité

Par modalité, nous entendons réalité ou non de l'événement évoqué.

2.1.1 Événement réel, réalisé → type="factual"

Si l'événement a réellement eu lieu, on y fait référence comme un fait passé ou en train de se dérouler.

(64) <event type="factual" frequence="unique" temp="before"> Tchernobyl </event> ne doit pas se reproduire !

(65) Le <event type="factual" frequence="recurring" temp="now"> plus grand festival du cinéma français </event> se déroule chaque année à Cannes.

(66) L'UE - l'ANASE: <event type="factual" frequence="unique" temp="before"> Interview accordée par le ministre fédéral des Affaires étrangères, Frank-Walter Steinmeier, au quotidien "Nürnberger Nachrichten" </event>.

(67) Le second à banaliser l'enjeu que constitue, pour ceux qui souhaitent une <event type="hypothetical" frequence="unique" temp="after"> victoire de la gauche </event> , l'orientation d'une nouvelle expérience après les <event type="factual" frequence="instance" temp="future"> <event type="abstract" frequence="recurring" temp="now"> élections </event> </event> .

(68) Il pense que la <event type="factual" frequence="unique" temp="after" source="reported"> réunion </event> devrait se tenir le jour même.

2.1.2 Événement potentiel, hypothétique, probable, possible → type="hypothetical"

On peut noter la distinction entre les événements potentiels passés et futurs.

1. L'événement est dans le futur. On envisage qu'il se produise, on le prévoit, mais ce n'est pas encore le cas et tant que ça ne l'est pas, il n'est pas réel.

   (69) On se trouve donc en face d'une véritable jungle de programmes et de matériels de toutes générations, auxquels il est pratiquement impossible de toucher sans risque d' <event type="hypothetical" frequence="unique" temp="after"> effondrement </event>.

   (70) Le ministère de la fonction publique semble pourtant avoir de grandes ambitions quant à l' <event type="hypothetical" frequence="unique" temp="after"> utilisation des technologies </event> pour améliorer le fonctionnement de l'Etat ?

   (71) Le second à banaliser l'enjeu que constitue , pour ceux qui souhaitent une <event type="hypothetical" frequence="unique" temp="after"> victoire de la gauche </event> , l'orientation d'une nouvelle expérience après les <event type="factual" frequence="instance" temp="future"> <event type="abstract" frequence="recurring" temp="now"> élections </event> </event>.

2. L'événement est passé. Il est rapporté comme potentiel. Cette façon de relater un événement est courante dans le langage journalistique. Les intervenants indiquent un fait passé, jugé suffisament important pour être évoqué, alors que l'information au moment de la prise de parole n'a pas pu être correctement vérifiée.

   (72) Et pour accroître le désarroi, on apprend qu'un autre <event type="hypothetical" frequence="unique" temp="before"> attentat </event> se serait produit à moins de 100 km d'Alger.

   NB. Dans le cas de l'exemple suivant, les deux événements de la phrase ont bien eu lieu, l'incertitude se pose dans la relation entre ces deux événements factuels. Le contexte de l'article de presse dont est issu l'exemple présente une enquête sur les causes d'un incident qui a eu lieu.

   (73) L' <event type="factual" frequence="unique" temp="before"> incident </event> se serait produit lors d'un <event type="factual" frequence="unique" temp="before"> forage </event> pour placer des explosifs dans une couche de charbon.

De plus, un événement qui ne doit pas se produire est de type hypothetical.

(74) La <event type="hypothetical" frequence="unique" temp="after"> guerre de Troie </event> n'aura pas lieu.

2.1.3 Événement qui n'a pas eu lieu → type="nonfactual"

Il nous semble intéressant de noter que ces événéments sont relatés parce qu'ils n'ont pas été réalisés. Ils prennent de l'importance par le non changement d'état/de situation qui aurait pu ou dû se produire.

(75) Cette <event type="nonfactual" frequence="unique" temp="before"> prétendue agression </event> avait débouché sur une grêve surprise paralysant l'ensemble de la ligne B du RER.

(76) " Nous n'avons perçu aucun <event type="nonfactual" frequence="unique" temp="before"> changement dans le discours de nos clients </event> , constate Sophie Romet "

Dans "*cette tentative d'assassinat*", la tentative est bien réelle, mais l'assassinat ne l'est pas. Il a été imaginé préparé, l'opération ayant échoué, il n'a pas eu lieu. Il est évoqué.

(77) Cette <event type="factual" frequence="unique" temp="before"> tentative d' <event type="nonfactual" frequence="unique" temp="before"> assassinat </event> </event>

2.1.4 Événement abstrait → type="abstract"

Il s'agit d'événements généraux. Une classe "événement abstrait" dénote les événements dont les instanciations importent peu, comme dans "*La crise suit une période de confiance excessive*" (de façon générale).

(78) La <event type="abstract" frequence="unique" temp="now"> crise </event> suit une période de confiance excessive.

Dans l'exemple suivant, le locuteur évoque la réforme de manière générale et abstraite. Il présente la nécessité de réformer l'État sans pour autant évoquer une réforme en particulier. La permutation des termes "une réforme" avec l'expression "cette réforme-ci" pour le particulier est sémantiquement impossible.

(79) je suis persuadé qu'une <event type="abstract" frequence="unique" temp="after"> réforme en profondeur de l'Etat </event> est absolument nécessaire .

2.2 Fréquence de l'événement

Certains événements sont définis comme appartenant à une suite programmée, de nature cyclique, comme d'autres sont uniques. Il convient de déterminer la fréquence d'un événement.

2.2.1 Événement unique → frequence="unique"

Comme son nom l'indique, l'événement unique ne se produit qu'une fois. Il peut s'agir d'un groupe d'événements (Les deux attentats), ou d'un non-événement (aucun changement).

(80) Durant leur <event type="factual" frequence="unique" temp="before"> chute accélérée par gravité </event>, elles passent devant la fenêtre de mesure.

C'est la catégorie qui sera utilisée par défaut, si aucune information ne nous permet de préciser la fréquence d'un événement.

Quand on parle d'*un autre attentat* qui aurait eu lieu, il ne s'agit pas d'un événement récurrent, mais bien d'un événement unique. Les attentats sont récurrents dans l'absolu dans cette partie de monde sur cette période donnée, mais celui-ci (cet autre attentat) est particulier.

(81) Et pour accroître le désarroi, on apprend qu' un <event type="hypothetical" frequence="unique" temp="before"> autre attentat </event> se serait produit à moins de 100 km d'Alger.

2.2.2 Événement récurrent → frequence="recurring"

Par opposition aux événements uniques, certains événements sont récurrents du genre cyclique ou périodique. Ils ont la particularité de pouvoir s'instancier en événements précis. Ainsi, on évoque l'événement hyperhonyme *Jeux Olympiques* sans pour autant être chamboulés dans nos références.

(82) Les <event type="abstract" frequence="recurring" temp="now"> Jeux paralympiques </event> se tiennent toujours en marge des <event type="abstract" frequence="recurring" temp="now"> Jeux olympiques </event>.

Remarque : Qui dit récurrence, dit suite d'instances qui sont par définition factuelles. Les événements récurrents sont le plus souvent de type abstrait.

(83) C'est comme si tu disais que tu as vu <event type="factual" frequence="unique" temp="before"> France-Allemagne 82 </event> et que tu as la preuve que les français ont été volés à l'arbitrage.

Dans l'Ex.83, on pourrait considérer "France-Allemagne 82" comme une suite d'événements, dont le match en 1982 serait une instance. Mais ces matchs de foot ne sont pas organisés selon un schéma cyclique prévisible et parfait. En effet, ces match entre des équipes nationales sont décidés par le hasard dans des tirages au sort.

Remarque : Le événements récurrents peuvent être passés, présents ou futurs, même si on s'accorde sur le fait que la valeur par défaut de celui-ci est le présent.

(84) Ils organisent, en ce moment, la <event type="factual" frequence="recurring" temp="after"> campagne d'évaluation </event> qui se tiendra cette année, puis tous les ans pendant 10 ans.

(85) Pendant son rêgne, une <event type="factual" frequence="recurring" temp="before"> vente d'esclaves </event> était organisé tous les mois.

Les événements récurrents qui sont toujours d'actualité, qui ont eu lieu précédemment, qui sont prévus encore, sont par défaut de type présent.

(86) Les <event type="factual" frequence="instance" temp="after"> <event type="abstract" frequence="recurring" temp="now"> JO </event> de 2012 </event> sont organiés à Londres.

Un événement qui a lieu tous les jours sera un événement récurrent, comme le "journal télévisé".

(87) et on reviendra bien évidemment sur [...] dans le <event type="factual" frequence="instance" temp="after"> <event type="abstract" frequence="recurring" temp="now"> journal de vingt heures </event> </event> .

Contre-exemple : "vacances". Les vacances de Pâques sont un événement récurrent, car ont lieu chaque année. Mais dans l'exemple suivant, "vacances" évoque l'idée générale et pas l'événement récurrent, cyclique.

(88) il s'agissait en fait de pompiers espagnols en <event.factual.unique.before.notreport> vacances </event> .

2.2.3 Événement instanciation d'un phénomène récurrent → frequence="instance"

Comme présenté ci-dessus, certains événéments récurrents peuvent être individuellement instanciés, c'est le cas de "*JO de 1996*" ou "*Atlanta 96*" qui sont des instanciations de l'événement récurrent et périodique *jeux Olympiques*.

(89) Les <event type="factual" freq="instance" temp="before"> <event type="abstract" type="recurring" temp="before"> Jeux Olympiques</event> de 1996 </event> ont été un succés.

Dans l'exemple suivant, on parle de "jeux Olympiques", mais il s'agit en réalité d'une instance des jeux, en particulier. Il faut donc annoter l'instance et l'événement récurrent sur le même syntagme.

(90) Un allemand triche et gagne aux <event type="factual" frequence="instance" temp="before"><event type="abstract" frequence="recurring" temp="now"> Jeux Olympiques </event></event>.

(91) Le <event type="factual" frequence="instance" temp="before"><event type="abstract" frequence="recurring" temp="now"> festival de Cannes </event></event> a sacré Laurent Cantet et son film "Entre les murs".

(92) Le <event type="factual" frequence="instance" temp="after"><event type="abstract" frequence="recurring" temp="now"> sommet de l'Union pour la Méditerranée (UpM) </event></event>qui devait avoir lieu à Barcelone les 6 et 7 juin a été reporté au mois de novembre, a-t-on appris le 20 mai de source égyptienne.

(93) lors de l' <event type="factual" frequence="instance" temp="before"><event type="abstract" frequence="recurring" temp="now"> université d'été du PS </event></event>, le 31 août, François Hollande enregistre une émission de radio .

(94) <event> 14 victoires </event> en <event type="instance"> 17 <event frequence="recurring"> Grands Prix </event> </event> [3]

(95) <event type="instance"> 17 années de <event type="recurring"> festival de Cannes </event> </event>

2.3 Moment de la réalisation

Les événements sont ancrés dans la temporalité, il est judicieux de les caractériser par le moment de leur réalisation. Il seront annotés en fonction du moment de l'énonciation, même lorsqu'il s'agit de discours rapporté ou d'un événement fictif (cf. section 2.4).

2.3.1 Événement passé → temp="before"

(96) <event type="factual" frequence="unique" temp="before"> Sa nomination </event> crée la <event temp="before"> surprise </event>

(97) Ils aboutissent à un <event type="factual" frequence="unique" temp="before"> constat d'échec </event>

(98) Mais en avril 2000 deux gestionnaires de fonds anglo-saxons, Guy Wyser-Pratte et Nathaniel Rothschild, sont parvenus lors d'une <event type="factual" frequence="unique" temp="before"> assemblée générale mémorable </event> à prendre le contrôle d'André, malgré la <event type="factual" frequence="unique" temp="before"> bataille menée par MR. Descours et ses actionnaires amis </event>.

(99) vingt ans plus tard , son successeur , Daniel Vaillant , devait ouvrir , vendredi 1er février à Marseille , le <event type="hypothetical" frequence="unique" temp="before"> cycle anniversaire de <event-modifier> cet </event-modifier> <event type="factual" frequence="unique" temp="before" source="reported"> " acte fondateur " </event> </event> .

2.3.2 Événement présent → temp="now"

(100) Les <event type="abstract" frequence="recurring" temp="now"> Jeux paralympiques </event> se tiennent toujours en marge

3. NB. Ce sont 17 instances de "*Grand Prix*".

des <event type="abstract" frequence="recurring" temp="now">
Jeux olympiques </event>.

(101) Les tâches à dominante féminines aboutissent rarement à la <event type="abstract" frequence="recurring" temp="now">
réalisation d'objets durables </event>

(102) Cette <event type="factual" frequence="unique" temp="now">
rentrée</event>-ci se place sous le signe de la contestation sociale.

(103) le programme d'action gouvernemental pour la société de l'information (PAGSI) a toujours soutenu que l' <event type="factual" frequence="unique" temp="now"> usage des nouvelles technologies </event> était essentiel.

2.3.3 Événement futur → temp="after"

(104) Le <event type="factual" frequence="unique" temp="after">
sommet Chine - Union Européenne annulé en décembre </event>
se tiendra à Prague.

2.4 Infos supplémentaires

Dans certains cas et en contexte, des informations supplémentaires sont intéressantes à annoter et du point de vue de l'énonciation. Il s'agit des indications sur le discours rapporté et le discours fictif.

2.4.1 Discours rapporté → source="reported"

On appelle discours rapporté tous les propos d'un texte qui sont issus d'une situation de communication différente de celle du texte où ils se trouvent. Ces propos sont donc repris et cités dans le texte étudié. Il est possible de reprendre le discours de manière intégrale ou en les reformulant.

Les marques du discours rapporté sont le plus souvent le temps verbal conditionnel présent (le spectacle *se tiendrait* tous les soirs) ou passé (le drame *se serait produit* ...) ou encore des marques lexicales de reprise de discours (selon X, Dans l'édition du journal daté du ..., raconté par Y, « ... » propos recueillis par Z, etc.)

L'événement "*assassinat de Gandhi par un complot de brahmanes fanatiques*" est de type "factual" car présenté comme une certitude, mais comme il est relaté par l'auteur du livre, cet événement est aussi "reported". De plus cet événement a bien eu lieu, le doute subsiste sur les auteurs de l'assassinat. L'information du discours rapporté ne force pas toujours que le type soit "hypothetical" :

(105) Il a publié une quinzaine d'*ouvrages* sur l'Inde dont le plus récent, "l'Inde, continent rebelle" (Le Seuil, 2000), *relate* l' <event type="factual" temp="before" source="reported">

assassinat de Gandhi par un complot de brahmanes fanatiques </event>.

(106) " Depuis le <event type="factual" frequence="unique" temp="before"> 11 septembre </event> , nous n'avons perçu aucun <event type="nonfactual" frequence="unique" temp="before" source="reported"> changement dans le discours de nos clients </event> , *constate Sophie Romet* . "

(107) *À en croire un responsable onusien*, quelque vingt-cinq délégués représentant toutes les ethnies afghanes devraient participer à une <event type="factual" frequence="unique" temp="after" source="reported"> conférence </event> qui devrait se terminer en "moins d'une semaine" et, "dans le meilleur des cas", pourrait aboutir à un accord de principe sur la <event type="hypothetical" frequence="unique" temp="after"> mise en place d'un gouvernement transitoire à Kaboul </event>.

Ici, ce n'est pas que le sommet ait lieu qui est rapporté, mais bien les conditions de sa mise en place : le lieu, la date.

(108) Le <event type="factual" frequence="unique" temp="after">
sommet Chine - Union Européenne annulé en décembre </event>
se tiendra à Prague, en République Tchèque, au mois de mai, *selon le journal China Daily*.

(109) Cette forme de carpe diem restera -t-elle dans l'air du temps, malgré le <event type="hypothetical" frequence="unique" temp="after" source="reported"> tassement *annoncé* de la croissance </event> et les lourdes incertitudes nées des <event> attentats de New York </event> ?

La spécification "source="reported"" est ajoutée dans le cas où l'apport d'information est rapporté dans la globalité. En effet, dans les exemples suivant, le discours rapporté est évoqué par la présence de guillemets, ce qui implique un détachement de l'auteur par rapport aux propos tenus. Dans l'Ex. 110, ce sont les participants à l'événement qui sont la source du discours rapporté, pas l'événement lui-même. Dans l'Ex. 111, nous avons choisi d'annoter l'événement comme atant rapporté, parce que le terme "fondateur" qui qualifie le nom "acte" confère à l'événement une autre dimension, et rend l'événement plus important qu'un simple acte.

(110) Le 25 mars, à l'initiative de la Présidence allemande du Conseil de l'Union européenne, les vingt-sept États membres ont fait une <event *source="notreport"*> déclaration commune *"des responsables de l'Union"* </event>

(111) Vingt ans plus tard, son successeur devait ouvrir le <event type="hypothetical" frequence="unique" temp="before"> cycle anniversaire de cet <event type="factual" frequence="unique" temp="before" *source="reported"*> acte "fondateur" </event> </event> .

2.4.2 L'événement fictif → source="fictive"

Certains événements semblent impossible à situer ailleurs que dans la fiction, une source à part entière d'information. C'est le cas de "rébellion", dans cet article évoquant une pièce de théâtre et la mise en scène d'un spectacle. La "rébellion" n'a lieu que dans l'histoire contée.

(112) A travers une série de fables qui entrent en résonance avec l'actualité, Emilie Valantin met en scène les ficelles de la haine et de la sottise. Jusqu'à l' <event type="abstract" frequence="unique" temp="now" source="fictive"> inéluctable rébellion </event> .

3 Annexes

3.1 Exemples supplémentaires et contre-exemples

– (113) Ce n'est pas <event type="factual" frequence="unique" temp="before"> Pearl Harbor </event>

– Exemple portant sur l'imbrication factual - nonfactual (hors contexte) :

(114) Cette <event type="factual" frequence="unique" temp="before"> tentative d' <event type="nonfactual" frequence="unique" temp="before"> assassinat </event> </event>

– (115) pour contribuer au <event type="abstract" frequence="unique" temp="now"> maintien de la paix dans le monde </event>, elle a besoin d'une défense commune.

(116) la délégation française s'est ainsi illustrée par l'<event type="factual" frequence="unique" temp="past"> absence de 12 de ses membres sur 16</event>.

– (117) La voie militaire aboutit fatalement à une *impasse* ce qui n'empêchera pas certains de vouloir l'emprunter.
Dans cet exemple, l'*impasse* est le constat d'échec, pas événement.

– Un report de date pour un événement équivaut à un événement attendu dans le futur : c'est le même événement, qui a simplement été reporté.

(118) Le <event type="factual" frequence="instance" temp="after"><event type="abstract" frequence="recurring" temp="now"> sommet de l'Union pour la Méditerranée (UpM) </event></event> qui devait avoir lieu à Barcelone les 6 et 7 juin a été reporté au mois de novembre, a-t-on appris le 20 mai de source égyptienne.

– *Cellule de crise*, quelques exemples annotés :

(119) La mission de la *cellule de* *<event>crise</event>* et les moyens déployés.

(120) Organisation des *cellules de* *<event>crise</event>* centrales et locales.

(121) Le ministre de l'Ecologie Jean-Louis Borloo a annoncé samedi soir la <event type="hypothetical" frequence="unique" temp="after"> mise en place d'une "*cellule de* *<event>crise</event>*" </event> qui réfléchit à la possibilité d'utiliser des aéroports militaires, quand la météo le permettra, pour soulager les aéroports civils qui seraient encombrés.

– (122) Rejetant la <event type="factual" frequence="unique" temp="before" source="reported"> "diabolisation" de l' <event type="abstract" frequence="unique" temp="now"> euthanasie </event> </event>, mais aussi la <event temp="before" source="reported"> "vaine querelle" opposant les pionniers des soins palliatifs à ceux du droit à mourir dans la dignité </event> , il salue l' <event type="factual" frequence="unique" temp="now"> arrivée d'une nouvelle génération, en particulier parmi les réanimateurs </event> , qui s'engagent dans "la logique du double effet" qui "donne au malade l'assurance de n'être jamais abandonné" à ses souffrances, d'être calmé par une sédation aussi poussée que nécessaire.
Nous avons considéré ici que "en particulier" fait partie de l'expression événementielle "arrivée d'une nouvelle génération, en particulier parmi les réanimateurs", que "en particulier" est rattaché à "génération" au même titre que l'est "réanimateurs"

– (123) Gaston Defferre présente ainsi le *grand dessein* de la loi de <event type="factual" frequence="unique" temp="before"> décentralisation </event> , qui sera promulguée le 2 mars 1982.
Cet exemple est issu d'un article rétrospective. Du point de vue de l'énonciation, c'est un événement passé, même si la phrase est au futur (de narration).

– (124) Franck Williams et Ron Dennis estiment aussi que <event> certaines propositions de la FIA </event> auront des <event> incidences sur la sécurité des pilotes </event> .
Incidences = répercussions, conséquences. On l'étiquette <event> même si c'est un type d'événement et non un événement en particulier.

– Les activités ...

(125) Dès dimanche, petits et grands pourront venir s'exercer au crève-ballons, à la loterie, au tir ou encore se régaler à la confiserie ou essayer le manège.

Dans l'exemple précédent, loterie, crève-ballon, etc. sont des catégories d'activités proposées lors de la kermesse, alors qu'à l'école on propose des activités qui ont lieu et qui sont faites, il y a une réelle actualisation.

(126) [...] en proposant aux enfants et aux adolescents des <event> activités en dehors des <event> heures scolaires </event> </event>

(127) Il présente les schémas de <event> développement de multiples activités , diverses et variées </event> .

(128) Ces contrats concernent l' <event> aménagement du temps de l'activité de l'enfant </event>.

– (129) Nos <event> sincères salutations </event>.

(130) <event> Nos sincères condoléances </event>.
C'est une expression performative, dire ou écrire cette formule permet de faire ses condoléances en même temps. Les condoléances sont un témoignage de sympathie à l'occasion d'un décès, le témoignage peut être un événement, d'autant que l'expression décrivant le moment où les condoléances ont lieu est désigné par *faire ses condoléances*. De plus, dans faire ses condoléances / son mea culpa / ses excuses, l'objet du verbe faire correspond à des événements.

– (131) les <event frequence="instance"> 19es <event frequence="recurring"> Jeux Olympiques d'hiver </event> </event>
Jeux Olympiques, seul, fait référence implicitement soit aux JO d'été, soit aux JO d'hiver.

3.2 Les indices à suivre pour annoter <event>

La tâche d'annotation des événements étant particulièrement subjective, voici quelques indices à suivre, malheureusement difficile à reproduire automatiquement.

3.2.1 Permutations avec d'autres mots qui eux sont clairements événement ou non événement

Dans l'exemple suivant, *preuve* ne peut être remplacé par "formalités/document" sans une modification sémantique de la phrase, ce mot est par contre permutable avec "surprise", qui est événement dans ce contexte.

(132) Les trente-sept journalistes français attendus ici pour le suivre dans ce <event> déplacement éclair </event> en sont , s'il en était besoin , une <event> preuve supplémentaire </event> .

En revanche, c'est le contraire dans l'exemple :

(133) Les preuves présentes dans le dossier sont accablantes

3.2.2 Les énumérations

Si une phrase est composée d'une énumération de noms d'événements sûrs et non ambigus, le plus souvent le syntagme nominal qui semble ambigu est à étiqueter en événement.

Si dans une énumération de termes, l'un semble irrémédiablement événement *violences policières*, les autres ne sont pas forcément à étiqueter en événement : *racisme* ne peut l'être et *engrenage malsain* peut paraître ambigu, mais c'est "enchaînement de circonstances" est événement.

(134) Pour ces enfants des quartiers Est de Londres , qui , depuis 1995 , n'ont cessé de dénoncer racisme , <event> violences policières </event> et <event> engrenage malsain des ensembles suburbains </event> , cette histoire filmée au coeur d'une cité sous tension de la banlieue parisienne ("la cité des muguets") a tout de suite trouvé une <event> résonance </event> .

Dans le même genre, dans une énumération de termes non événement et non ambigus, on n'annotera pas les termes ambigus en événement.

3.2.3 Les termes fortement ambigus

Certains termes en contexte sont fortement ambigus, il est alors difficile de trancher pour l'annotation ou non de son gn.

– Période de temps ou état ?
Dans cet expression, le purgatoire peut être considéré comme l'état dans lequel on peut être, ou la durée de sa présence dans le lieu purgatoire

(135) Ohana connaît aujourd'hui une *période de purgatoire* contre laquelle tentent de s'insurger une poignée d'interprètes , tels Roland Hayrabédian et son ensemble vocal Musicatreize , basé à Marseille .

– Objet ou action ?
Création peut faire référence à "l'objet, la chose créé" et "l'acte, le fait de créer". Dans le premier exemple suivant, on considère le projet de créer quelquechose, la création est donc un événement. dans le deuxième, c'est l'objet créé.

(136) une <event> manifestation </event> leur demandait de réfléchir à un projet de <event> *création* </event>

(137) le metteur en scène pressait le groupe et sa maison de disques (Labels) de donner une <event> représentation parisienne de *cette création* </event> .

Fusion désigne ici, le résultat d'une fusion, qu'est cette musique.

(138) *fusion* de rythmes traditionnels , de distorsion rock , de nonchalance dub , de surtension hip hop et électronique , leur musique décrit le mélange possible des cultures et la violence d'une société .

(139) Interroger les victimes et les auteurs eux-mêmes pour mieux évaluer la réalité de la délinquance figurent au rang des <event> *propositions* </event> que les députés ont remises hier au Premier ministre. Ces <event> *propositions* </event> interviennent avant que le bilan 2001 de la délinquance et de la criminalité soit rendu public.

Dans certains cas, l'ambiguïté est forte entre l'objet désigné et l'action. À des fins d'homogénéité, il a été décidé que les *propositions de loi* seraient annotées en événement, vu qu'il est difficile de différencier les propositions présentées à l'oral et celles remises par écrit.

– Qualité ou événement ?

(140) Le ministre de l'Ecologie Jean-Louis Borloo a annoncé samedi soir la <event> mise en place d'une "cellule de crise" </event> qui réfléchit à la *possibilité*

d'utiliser des aéroports militaires, quand la météo le permettra, pour soulager les aéroports civils qui seraient encombrés.

Dans cet exemple ; "la possibilité" n'est pas événement. Ce terme est permutable avec capacité, il s'agit d'une qualité.

(141) Quels sont les principaux freins au changement ? Il en existe trois, principalement . La résistance du fonctionnaire au changement, tout d'abord. Ensuite, le manque d'équipements, ou des équipements pas au niveau. Enfin, le manque de volonté politique.

résistance au changement est à considérer en contexte comme une qualité imputée au fonctionnaire et donc pas annoté.

– Groupe de personne ou événement éponyme ?
Les noms d'événement et groupes de personnes, comme "assemblée" ou "conseil" peuvent être ambigus. Voici quelques exemples annotés :

(142) Mais en avril 2000 deux gestionnaires de fonds anglo-saxons, Guy Wyser-Pratte et Nathaniel Rothschild, parviennent lors d'une <event> assemblée générale mémorable </event> à prendre le contrôle d'André, malgré la <event> bataille menée par MR. Descours et ses actionnaires amis </event>.

(143) La loi Solidarité renouvellement urbains (SRU), votée le 13 décembre 2000, a souhaité généraliser le compte bancaire séparé, sauf décision contraire de l'*assemblée générale*.

Autre exemple, somme toute peut-être plus parlant : "secours", "équipe de secours" ou "action de secourrir quelqu'un".

(144) Le plus urgent c'est l' <event> envoi des *secours* pour faire face à une terrible situation humanitaire </event>

(145) L' <event> opération </event> a mobilisé 25 hommes issus des centres de Belfort Nord et Sud, de Giromagny et du service départemental d' <event> incendie </event> et de <event> *secours* </event> .

Références

[Arnulphy, 2012] Béatrice Arnulphy. *Désignations nominales des événements : Étude et extraction automatique dans les textes.* PhD thesis, Université Paris-Sud - École Doctorale d'Informatique de Paris Sud (EDIPS) / Laboratoire LIMSI, OCT 2012.

[Ehrmann and Hagège, 2009] Maud Ehrmann and Caroline Hagège. Proposition de caractérisation et de typage des expressions temporelles en contexte. In Adeline Nazarenko, Thierry Poibeau, and Haïfa Zargayouna, editors, *Actes de TALN2009*, Avignon, June 2009. ATALA.

[Grouin *et al.*, 2011] Cyril Grouin, Olivier Galibert, Sophie Rosset, Ludovic Quintard, and Pierre Zweigenbaum. Mesures d'évaluation pour entités nommées structurées. In *Actes EvalECD/QDC'2011*, 2011.

[Pustejovsky *et al.*, 2003] James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. Timeml : Robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics.*, 2003.

[Rosset *et al.*, 2011] Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. *Entités Nommées Structurées : guide d'annotation Quaero.* LIMSI–CNRS, Orsay, France, 2011. `http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf`.

[Vendler, 1959] Zeno Vendler. Verbs and Times. *The philosophical Review*, 66(2) :143–160, apr 1959. (In Mani, Pustejovsky and Gaizauskas, 2005, chapter 1).

C.3 L'analyseur syntaxique *XIP*

C.3.1 Présentation

XIP (Aït-Mokhtar *et al.*, 2002) est un analyseur syntaxique robuste qui permet une analyse pour le français et l'anglais des relations de dépendances et la reconnaissance d'entités nommées classiques. Les événements n'y sont pas traités. Ce produit développé par Xerox Research Centre Europe est distribué avec des grammaires encryptées, inaccessibles à l'utilisateur. Pourtant, il est possible d'enrichir l'analyse par l'ajout de ressources et la création de ses propres règles de grammaire.

Dans toutes les expérimentations du chapitre 4, nous avons utilisé *XIP*. pour tester les lexiques uniquement, il est vrai que nous aurions pu réaliser une simple projection du lexique sur le corpus (après lemmatisation de celui-ci). Nous avions précédemment tenté une approche en utilisant *Tree Tagger*, qui fournit une annotation de corpus en parties du discours, mais l'ambiguïté sur les noms et verbes n'est pas toujours évidente à faire tomber, cette distinction nom-verbe étant plus difficile à faire en anglais. Alors, nous avons choisi d'utiliser XIP, disponible pour le français et l'anglais.

De plus, par souci de cohésion de nos résultats sur la durée, nous avons souhaité utiliser le même outil et ce même si il est vrai qu'une marge d'erreur a été observée, par exemple, des noms qui sont analysés comme des verbes et inversement, ce qui est préjudiciable lorsqu'on travaille sur des noms. Mais l'analyse syntaxique fournie par XIP nous est par ailleurs d'une grande aide pour la suite de nos travaux, comme l'utilisation des règles de grammaires contextuelles pour l'extraction des arguments de verbes présélectionnés.

C.3.2 Les règles XIP

Ajout de vocabulaire dans Xip Il est possible d'attribuer des traits à du lexique dans *XIP*. Il faut déclarer le trait qui nous intéresse et ensuite l'attribuer aux mots du lexique dans un fichier de type « Vocabulaire », comme dans l'exemple suivant.

(C.1)
```
Vocabulaires :

abaissement : noun += [VerbAction=+].
abalourdissement : noun += [VerbAction=+].
abandon : noun += [VerbAction=+].
abandonnement : noun += [VerbAction=+].
abasourdissement : noun += [VerbAction=+].
abâtardissement : noun += [VerbAction=+].
...
zippage : noun += [VerbAction=+].
zonage : noun += [VerbAction=+].
zonzonnement : noun += [VerbAction=+].
zozotement : noun += [VerbAction=+].
zwanze : noun += [VerbAction=+].
```

Repérage d'une locution prépositive au moyen des règles XIP Prenons l'exemple suivant C.2.

(C.2) `au lendemain de` *`<event> son arrivée </event>`*`, il est reparti.`

Les dépendances données par *XIP* pour le groupe nominal événement qui nous intéresse dans cet exemple sont :

- `ATTRIBUT_DE[POSIT1 ](lendemain(2), arrivée(8))`
  relation attribut par l'intermédiaire de la préposition « `de` » entre « `lendemain` » et « `passation` ».
- `PREPOBJ(arrivée(8), de(4))`
  → La préposition « `de` » est rattachée à « `passation` » : « `passation` » est dépendant de la préposition « `de` ».
- `PREPOBJ(lendemain(2), à(0))`
  → La préposition « `à` » est rattachée à « `lendemain` » : « `lendemain` » ets dépendant de la préposition « `à` ».
- `DETERM[POSS ](arrivée(8), son(6))`
  → Le nom « arrivée » est déterminé par le déterminant possessif « son ».

Nous souhaitons y repérer l'événement « son arrivée ». Nous souhaitons qu'une règle *XIP* repère ce mot. Par ailleurs, « Arrivée » est un nom appartenant au *VerbAction*, comme nous l'indique le trait VERBACTION. Rappelons que pour le moment, nos travaux ont privilégié le repérage des têtes de syntagme. Nous considérons que le syntagme no-

minal entier peut être recomposé par la suite. L'indice d'événementialité de cet exemple est ce que nous avons appelé un indicateur temporel, qui dans ce cas est la locution prépositive « au lendemain de ». Il ne s'agit pas seulement de réperer cette dépendance « ATTRIBUT_DE » pour repérer le nom d'événement, mais aussi de répérer que le mot « lendemain » est bien utilisé dans le contexte événementiel qui nous intéresse, soit « au lendemain de ». Nous le faisons au moyen d'une règle *XIP*.

Pour repérer dans les phrases cette locution prépositive, il faut d'abord donner au nom « lendemain » un trait qui sera par la suite utilisé pour créer dans *XIP* un déclencheur *IT* et repérer le nom qui est en relation avec cette préposition. Le trait utilisé est « moment_a_de ».

(C.3)
```
Vocabulaires :

lendemain : noun += [moment_a_de = +].
```

Dans une « grammaire, dite de déduction », nous définissons l'usage attendu de « `lendemain` » pour repérer le nom événementiel. Cette description vaudra pour tous les mots portant ce trait. Ainsi on demande à ce que le mot portant le trait « `moment_a_de` » ait une dépendance syntaxique de type `PREPOBJ` avec la préposition « `à` ». Le nom d'événement est de plus rattaché à ce mot portant le trait « `moment_a_de` ». Ce mot pour être événement doit être un nom, qui ne possède pas le trait « `time` ». Il ne doit pas non plus être une entité nommée « lieu », ni « date »

(C.4)
```
if ( ATTRIBUT_DE(#1[moment_a_de],#2[noun, time:~])
    & PREPOBJ(#1[moment_a_de],#3[prep,lemme:à])
    & ~LIEU(#2)
    & ~DATE(#2)
    )
EVENT[IT=6](#2).
```

Bibliographie

A. Abeillé, L. Clément, and F. Toussenel. *Treebanks*, chapter Building a French treebank, pages 165–188. Kluwer, Dordrecht, 2003. 44

David Ahn. The stages of event extraction. In *ARTE '06 : Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Morristown, NJ, USA, 2006. 49

Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. Robustness beyond Shallowness : Incremental Deep Parsing. *Natural Language Engineering*, 8 :121–144, 2002. 125, 190

Artemis Alexiadou. *Functional Structure in Nominals : Nominalization and Ergativity.* Linguistik aktuell. J. Benjamins, 2001. 54

Chinatsu Aone and Mila Ramos-Santacruz. REES : A large-Scale Relation and Event Extraction System. In *Proceedings of ANLP 2000*, pages 76–83, Seattle, Washington, USA, 2000. 37, 38, 48

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. Les entités nommées événement et les verbes de cause-conséquence. In *Actes de TALN 2010*, Montreal, Canada, 2010. 103

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. Un lexique pondéré des noms d'événements en français. In *Actes de TALN 2011*, Montpellier, France, 2011. 116

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. Vers une annotation automatique des événements dans les textes. In *Colloque international - Langage, discours, événements*, Firenze, Italy, 2011. Communication orale (article en cours de publication). 88

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. Automatically Generated Noun Lexicons for Event Extraction. In *Proceedings of the 13th International Conference on*

Intelligent Text Processing and Computational Linguistics (CicLing 2012), New Delhi, India, 2012. 113

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. Event Nominals : Annotation Guidelines and a Manually Annotated Corpus in French. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. 86, 87

Béatrice Arnulphy. A weighted lexicon of french event names. In *Proceedings of the 2nd Student Research Workshop associated with RANLP 2011*, pages 9–16, Hissar, Bulgaria, 2011. 113

Béatrice Arnulphy. Vers une annotation automatique des événements dans les textes. *Les carnets du Cédiscor*, à paraître. 88

Paul Bacot, Laurent Douzou, and Jean-Paul Honoré. Chrononymes. la politisation du temps. *Mots. Les langages du politique*, 87 :5–12, 2008. 27, 28, 163

Antonio Balvet, Lucie Barque, Marie-Helene Condette, Pauline Haas, Richard Huyghe, Rafael Marín, and Aurélie Merlo. Nomage : an electronic lexicon of french deverbal nouns based on a semantically annotated corpus. In *Proceedings of the 1st International Workshop on Lexical Resources (WoLeR 2011)*, 2011. 44

Nuria Bel, Maria Coll, and Gabriela Resnik. Automatic Detection of Non-deverbal Event Nouns for Quick Lexicon Production. In *Proceedings of the International Conference on Computational Linguistics (COLING 2010)*, pages 46–52, Beijing, China, 2010. 8, 47, 53, 55, 125, 130, 150, 173

Steven Bethard and James H. Martin. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154, Sydney, Australia, 2006. 49, 50, 102, 150, 151, 152, 153, 157

André Bittar. *Construction d'un TimeBank du français : Un corpus de référence annoté selon la norme ISO-TimeML*. PhD thesis, Université Paris Diderot - École doctorale de Sciences du Langage / Laboratoire ALPAGE, 2010. 8, 39, 41, 47, 87, 169

André Bittar. *ISO-TimeML Annotation Guidelines for French – Version 1.0*. ALPAGE & Université Paris 7 Diderot, 2010. 30 septembre 2010. 39, 51, 88

Laura Calabrese. Les héméronymes. Ces évènements qui font date, ces dates qui deviennent évènements. *Mots. Les langages du politique*, 3 :115–128, 2008. 19, 29, 73, 163

Laura Calabrese. Nom propre et dénomination évènementiele : quelles différences en langue et en discours ? *Corela*, 7(1), 2009. 23, 24, 25

Laura Calabrese. *Le rôle des désignants d'événements historico-médiatiques dans la construction de l'histoire immédiate. Une analyse du discours de la presse écrite.* PhD thesis, Université Libre de Bruxelles - Faculté de Philosophie et Lettres, 2010. 22, 35, 65, 68

Laura Calabrese. La nomination d'événements dans le discours d'information : entre activité collective et déférence épistémiologique. In *Colloque Langage, discours, événements*, Firenze, Italy, 2011. 158

Laurent Catherin. The french wordnet. Rap. tech. deliverable 2d014, EuroWordNet, 1999. 43

Michel Charolles. *La référence et les expressions référentielles en français.* Collection L'essentiel Français. Ophrys, Paris, France, 2002. 21, 163

Nancy Chinchor. Overview of muc-7. In *Proceedings Seventh Message Understanding Conference (MUC-7)*, Fairfax, Virginia, 1998. 48

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1) :37–46, 1960. 86, 104, 167

Francisco Costa and António Branco. Timebankpt : A timeml annotated corpus of portuguese. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. 41, 42, 87

Cassandre Creswell, Matthew J. Beal, John Chen, Thomas L. Cornell, Lars Nilsson, and Rohini K. Srihari. Automatically Extracting Nominal Mentions of Events with a Bootstrapped Probabilistic Classifier. In *Proceedings of the International Conference on Computational Linguistics (COLING 2006)*, pages 168–175, Sydney, Australia, 2006. 41, 42, 46, 47, 52, 86, 87, 97, 150, 151, 152, 153

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. Timbl : Tilburg memory based learner, version 5.1, reference guide. Ilk technical report ilk-0402, University of Tilburg, 2004. 49

Hal Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. 2004. 49

Donald Davidson. *Actions et Événements.* Épithémée. Presses Universitaires de France, Paris, France, 1993. (traduit de l'américain par Pascal Engel). 18, 63

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ace) program - tasks, data, and evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal, 2004. 32, 37

Kurt Eberle, Gertrud Faaß, and Ulrich Heid. Corpus-based identification and disambiguation of reading indicators for German nominalizations. In *Proceedings of Corpus Linguistics 2009*, Liverpool, UK, jul 2009. 53, 56

Maud Ehrmann and Caroline Hagège. Proposition de caractérisation et de typage des expressions temporelles en contexte. In *Actes de TALN'09*, Avignon, 2009. 79

Maud Ehrmann. *Les Entités Nommées, de la linguistique au Tal : Statut théorique et méthodes de désambiguïsation*. PhD thesis, Université Paris 7, JUN 2008. 30, 163

Arlette Farge. Histoire, événement, parole. *Socio-Anthropologie*, Communauté et/ou ensemble populationnel.(2), 1997. 35

Christiane Fellbaum. *WordNet : An Electronic Lexical Database*. Bradford Books, 1998. 43

Corina Forascu and Dan Tufis. Romanian timebank : An annotated parallel corpus for temporal information. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. 41, 42, 87

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean-Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deléger, and Dominique Laurent. Named and specific entity detection in varied data : The quæro named entity baseline evaluation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. 31

Olivier Galibert. *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, 2009. 103

Marie-Noëlle Gary-Prieur. *Grammaire du nom propre*. Presses Universitaires de France, 1994. 23, 24

Jane Grimshaw. *Argument Structure*. The MIT Press, Cambridge, Massachussets, 1990. 54, 55

Ralph Grishman and Beth Sundheim. Message Understanding Conference : A Brief History. In *Proceedings of the International Conference on Computational Linguistics (COLING 1996)*, pages 466–471, Kopenhagen, Danmark, 1996. 10, 32

Cyril Grouin, Olivier Galibert, Sophie Rosset, Ludovic Quintard, and Pierre Zweigenbaum. Mesures d'évaluation pour entités nommées structurées. In *Actes EvalECD/QDC'2011*, 2011. 11, 31, 163

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software : An Update. *SIGKDD Explorations*, 11(1), 2009. 116, 124

Richard Huyghe and Rafael Marín. L'héritage aspectuel des noms déverbaux en français et en espagnol. *Faits de Langues*, 30 :265–274, 2007. 20, 44, 63

Richard Huyghe. Les noms génériques d'espace en français. Master's thesis, Université de Lille 3, 2006. 20, 44

Bernard Jacquemin. *Construction et interrogation de la structure informationnelle d'une base documentaire en français*. PhD thesis, Université Paris III Sorbonne Nouvelle, Paris, 2003. 43

Georges Kleiber. Sens, référence et existence : que faire de l'extralinguistique ? *Langages*, 127 :9–37, 1997. 22

Alice Krieg-Planque. À propos des « noms propres d'événement ». Événementialité et discursivité. *Les Carnets du Cediscor, Le nom propre en discours*, 11 :77–90, 2009. 19, 22

J R Landis and G G Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33 :159–174, 1977. 86, 167

Michelle Lecolle. Toponymes en jeu : Diversité et mixage des emplois métonymiques de toponymes. In Roumanie Université de Pitesti, editor, *Studii si cercetari filologice 3 / 2004*, 2004. 26

Michelle Lecolle. Éléments pour la caractérisation des toponymes en emploi événementiel. In *Actes du Colloque international - Représentations du sens linguistique III*, Bruxelles, nov 2005. (début 2009 : à paraitre) ń Éléments pour la caractérisation des toponymes en emploi événementiel ż. In Ivan Evrard, Michel Pierrard, Laurence Rosier, Dan Van Raemdonck (éds), /Les sens en marge / /Représentations linguistiques et observables discursifs : actes du colloque international de Bruxelles, 3-5 novembre 2005/, Paris, L'Harmattan, 2009, pp. 29-43. 35, 36

Michelle Lecolle. Polyvalence des toponymes et interprétation en contexte. *Pratiques*, 129/130 :107–122, 2006. 26, 163

Michelle Lecolle. Éléments pour la caractérisation des toponymes en emploi événementiel. In Ivan Evrard, Michel Pierrard, Laurence Rosier, and Dan Van Raemdonck, editors, *Les sens en marge - Représentations linguistiques et observables discursifs : actes du colloque international de Bruxelles, 3-5 novembre 2005*, pages 29–43. L'Harmattan, Paris, 2009. 20, 64, 65, 66, 73

Gérard Mairet. *Le Discours et l'historique : essai sur la représentation historienne du temps*. Bibliothèque Repères : Sciences humaines, idéologies. Mame, 1974. 65

Céline Le Meur, Sylvain Galliano, and Edouard Geoffrois. Conventions d'annotations en Entités Nommées - ESTER. Technical report, Centre d'Expertise Parisien de la Délégation Générale de l'Armement (DGA), jul 2004. (15 juillet 2004). 33

A Meyers. Anotation guidelines for nombank-noun argument structure for propbank. Online Publication : http ://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf, 2007. 44

Marc Moens and Mark Steedman. Temporal Ontology and Temporal Reference. *Computational linguistics - Special issue on tense and aspect*, 14(2) :15–28, jun 1988. (In Mani, Pustejovsky and Gaizauskas, 2005, chapter 15). 17, 18, 62

Sophie Moirand. *Le discours rapporté dans tous ses états*, chapter La circulation interdiscursive comme lieu de construction de domaines de mémoire par les médias, pages 373–385. l'Harmattan, 2004. 23

Sophie Moirand. *Les discours de la presse quotidienne. Observer, analyser, comprendre*. PUF, 2007. Isbn : 2-13-055923-9. 19, 62, 163

Abraham A. Moles. Notes pour une typologie des événements. *Communications*, 18 "L'événement" :90–96, 1972. 34, 35

Harvey Molotch and Marilyn Lester. Informer : Une conduite délibérée de l'usage stratégique des événements. *Réseaux*, 14(75) :23–41, 1996. traduction de NEWS AS PURPOSIVE BEHAVIOR : ON THE STRATEGIC USE OF ROUTINE EVENTS, ACCIDENTS AND SCANDALS ©American Sociological Review, vol. 39 (février), 1974 pour la version originale. 65

Erik Neveu and Louis Quéré. Présentation. *Réseaux*, 14(75) :7–21, 1996. 19, 68

Gabriel Parent, Michel Gagnon, and Philippe Muller. Annotation d'expressions temporelles et d'événements en français. In *Traitement Automatique des Langues Naturelles (TALN'08)*, Avignon, 2008. 49, 51, 100, 150, 152, 153

Aina Peris, Mariona Taulé, Gemma Boleda, and Horacio Rodríguez. Adn-classifier :automatically assigning denotation types to nominalizations. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. 53, 54, 97, 125, 150, 151, 153

Aina Peris, Mariona Taulé, and Horacio Rodríguez. Semantic annotation of deverbal nominalizations in the spanish corpus ancora. In Markus Dickinson, Kaili Mrisep, and Marco Passarotti, editors, *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, volume 9, pages 187–198, University of Tartu, Estonia., 2010. Northern European Association for Language Technology (NEALT). 54

M. Carme Picallo. La estructura del sintagma nominal : las nominalizaciones y otros sustantivos com complementos argumentales. In I. Bosque and V. Demonte, editors, *Gramàtica Descriptiva de la Lengua Española*, volume 1, pages 363–393, Madrid, Espasa, 1999. 54

Thierry Poibeau. *Extraction automatique d'information. Du texte brut au web sémantique*. Hermès, 2003. 10

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. Timeml : Robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics.*, 2003. 71

J. Pustejovsky, M. Verhagen, R. Saurí, J. Littman, R. Gaizauskas, G. Katz, I. Mani, R. Knippen, and A. Setzer. *TimeBank 1.2*. Linguistic Data Consortium, 2006. 40

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. ISO-TimeML : An International Standard for Semantic Annotation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. 39, 65, 75

James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, Massachussetts, 1995. 54

Ting Qian, Benjamin Van Durme, and Lenhart Schubert. Building a Semantic Lexicon of English Nouns via Bootstrapping. In Association for Computational Linguistics, editor, *HLT-NAACL Student Research Workshop and Doctoral Consortium*, pages 37–42, Boulder, Colorado, JUN 2009. 36, 46, 47, 53, 56

R. Quinlan. *C4.5 : Programs for Machine Learning*. Morgan Kaufman Publishers, 1993. 116

Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040, 1995. 50

Gabriela Resnik and Núria Bel. Automatic detection of non-deverbal event nouns in spanish. In Istituto di Linguistica Computazionale, editor, *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pisa, 2009. 8, 47, 55, 97, 150, 151, 153, 172

Sophie Rosset, Olivier Galibert, Gabriel Illouz, and Aurélien Max. Interaction et recherche d'information : le projet RITEL. *TAL. Traitement automatique des langues*, 46(3) :155–179, 2005. 103

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. *Entités Nommées Structurées : guide d'annotation Quaero*. LIMSI–CNRS, Orsay, France, 2011. http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf. 77

Irene Russo, Tommaso Caselli, and Francesco Rubino. Recognizing deverbal events in context. In *Proceedings of 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011), poster session*, Tokyo, Japan, feb 2011. Springer. 41, 53, 54, 55, 86, 87, 94, 125, 150, 153

Benoît Sagot and Darja Fišer. Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de TALN 2008*, Avignon, France, 2008. 43

R Saurí, J Littman, R Knippen, R Gaizauskas, A Setzer, and Pustejovsky. Timeml annotation guidelines. Technical report, Linguistic Data Consortium (LDC), 2004. 63

Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. Evita : A Robust Event Recognizer for QA Systems. In *Proceedings of the HLT05*, Vancouver, Canada, OCT 2005. 39, 49, 50, 150, 152

Roser Saurí, Marc Verhagen, and James Pustejovsky. Annotating and recognizing event modality in text. In *The 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida, USA, may 2006. American Association for Artificial Intelligence (www.aaai.org). 64

Roser Saurí, Olga Batiukova, and James Pustejovsky. Annotating events in spanish timeml annotation guidelines. Technical report, Barcelona Media, 2009. 41

Lenhart K. Schubert. Can we derive general world knowledge from texts ? In *Proceedings of the Second International Conference on Human Language Technology Research (HLT 2002)*, pages 94–97, San Diego, CA, March 2002. 56

Rohini K. Srihari and Adrian Novischi. Visual semantics for reducing false positives in video search. In *Proceedings of AAAI Fall Symposium - Multimedia Information Extraction*, pages 31–35, Arlington, VA, nov 2008. 52

Ludovic Tanguy and Nabil Hathout. Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In Jean-Marie Pierrel, editor, *Actes de TALN 2002*, pages 245–254, Nancy, June 2002. ATALA, ATILF. 46

Xavier Tannier, Véronique Moriceau, Béatrice Arnulphy, and Ruixin He. Evolution of event designation in media : Preliminary study. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. 25, 123

Delphine Tribout. *Les conversions de nom à verbe et de verbe à nom.* PhD thesis, Université Paris Diderot - École doctorale de Sciences du Langage, 2010. 44, 72

Gaye Tuchman. Making news by doing work : Routinizing the unexpected. *American Journal of Sociology*, 79(1) :110–131, 1973. 65

A.-J. Tudesq. *La presse et l'événement.* Maison des sciences de l'homme de Bordeaux, Paris/La Haye, 1973. 35

Danièle van de Velde. Existe-t-il des noms propres de temps ? *Lexique*, 15 :151, 2000. 24

Danièle van de Velde. *Grammaire des événements.* Sens et structures. Presses Universitaires du Septentrion, 2006. 18, 63

Benjamin Van Durme, Ting Qian, and Lenhart K. Schubert. Class-driven attribute extraction. In *the International Conference on Computational Linguistics (COLING 2008)*, 2008. 56

Zeno Vendler. Verbs and Times. *The philosophical Review*, 66(2) :143–160, apr 1959. (In Mani, Pustejovsky and Gaizauskas, 2005, chapter 1). 16, 17, 36, 62, 70

Marie Veniard. *La nomination d'un événement dans la presse quotidienne nationale. Une étude sémantique et discursive : la guerre en Afghanistan et le conflit des intermittents dans Le Monde et Le Figaro.* PhD thesis, Paris 3, 2007. 19

Marie Veniard. La dénomination propre la guerre d'afghanistan en discours : une interaction entre sens et référence. *Les Carnets du Cediscor, Le nom propre en discours*, 11 :61–76, 2009. 23, 24, 25

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Graham Katz, and James Pustejovsky. Semeval2007 task 15 : Tempeval temporal relation identification. In *In SemEval-2007 : 4th International Workshop on Semantic Evaluations*, 2007. 71

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13 : Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 57–62, Uppsala, Sweden, JUL 2010. 41

Henk J. Verkuyl. Aspectual Classes and Aspectual Composition. *Linguistics and Philosophy*, 12(1) :39–94, feb 1989. (In Mani, Pustejovsky and Gaizauskas, 2005, chapter 15). 17, 62

Paul Veyne. *Comment on écrit l'histoire. Essai d'épistémologie.* Seuil, Paris, 1971. 65

Dina Wonserver, Aiala Rosá, Marisa Malcuori, Guillermo Moncecchi, and Alan Descoins. Event annotation schemes and event recognition in spanish texts. In Alexander Gelbukh, editor, *Proceedings of Cicling 2012*, volume 2, pages 206–218. Springer LNCS, MAR 2012. 41, 87, 150