

Comment le text mining peut-il aider à gérer le risque militaire et stratégique ?

[Ramona PAUNA](#) (*), [Sylvie Guillemin-Lanne](#) (*),
<mailto:ramona.pauna@temis.com>, <mailto:sylvie.guillemin-lanne@temis.com>

(*) TEMIS, 164 rue de Rivoli 75001 Paris (France)

Mots clefs :

Fouille textuelle, détection d'événements, composant de connaissance, risque stratégique et terroriste

Keywords:

Text mining, event detection, skill cartridge, strategic and terrorist risk

Palabras clave:

Búsqueda textual, detección de eventos, skill cartridge, riesgo estratégico y terrorista

Résumé

Cette communication a pour objet de proposer un modèle de description de la notion d'*événements* dans une perspective de *text mining*. L'étude est centrée sur les événements qui constituent un risque dans le domaine de la sécurité militaire. Nous présentons dans un premier temps les fondements linguistiques qui sont à la base de notre modèle, notamment les paramètres d'analyse des événements proposés par Gross et Kiefer ([9]). Dans cette optique, les événements sont divisés en événements *fortuits*, *créés* et *cycliques*. Après cette introduction théorique, nous rappelons quelques modèles d'analyse qui ont été proposés pour les événements. Enfin, nous présentons un outil d'extraction d'événements mis au point par la société TEMIS, notamment le composant de connaissance **Events**TM. Ainsi, nous décrivons de façon succincte l'architecture sémantique du composant de connaissance, plus précisément les entités et les relations de type événementiel détectées par ce produit. Le but final est de montrer qu'avec des outils de *text mining*, nous pouvons apporter une aide efficace à la détection des événements représentant un risque sécuritaire pour la population d'un Etat. Enfin, nous présentons une application en ligne mettant en jeu les événements de type militaire et stratégique que nous avons modélisés.

1 Introduction

L'étude du lexique a pris dans les vingt dernières années une dimension nouvelle. Avec l'explosion des Technologies de l'Information et de la Communication (TIC), on est passé de l'analyse de mots isolés à des traitements plus complexes : recherche d'informations textuelles à l'aide de moteurs de recherche, indexation de documents, résumé automatique, fouilles et génération automatiques de textes. Le but de ces techniques et méthodes, correspondant grosso modo au traitement automatique de la langue, consiste à apporter une aide efficace à l'analyse de grandes masses de données, afin d'en faciliter la lecture et l'exploitation. L'objet de cette communication est de montrer une application de ces techniques à la détection d'événements du domaine militaire et stratégique. Nous présenterons une approche qui combine une méthodologie linguistique de traitement des événements inspirée du modèle proposé par G. Gross et F. Kiefer *La structure événementielle des substantifs* ([9]) et une solution pour le text mining tel qu'il est pratiqué par la société Temis.

La première partie de la communication présentera les fondements linguistiques de la description des événements (paramètres des événements, classes sémantiques). La deuxième partie sera consacrée à la présentation d'une modélisation de la notion d'événement.

2 Qu'appelle-t-on un événement ?

Il est difficile de donner une définition conceptuelle de la notion d'*événement*, en raison de leur complexité et du fait que des états ou des actions peuvent aussi, dans certaines conditions, avoir une lecture événementielle ([8]). A première vue, on désigne par *événement* un « fait » qui se produit dans un cadre spatio-temporel et qui a un témoin implicite ou explicite. Nous rappelons ici brièvement la définition linguistique des événements proposée par Gross et Kiefer, qui nous a servi pour mettre au point notre modèle descriptif.

Selon ces auteurs, un fait est considéré comme un événement grâce à la présence d'un témoin qui l'atteste. Celui-ci peut être visuel, auditif ou encore désigner un média (télévision, radio ou presse). Un événement peut être décrit, en outre, à l'aide de deux autres paramètres définitoires : le lieu où il se passe et la date à laquelle il se produit. Enfin, le domaine dont relèvent les événements peut jouer un rôle dans leur définition : un événement météorologique (*une tempête*, par exemple) aura une syntaxe différente d'un événement du domaine politique (*des élections présidentielles*), puisque les acteurs impliqués dans ces événements sont constitutifs de la nature de l'événement lui-même.

Sur la base de ces paramètres, Gross et Kiefer distinguent trois grands types d'événements: les événements *fortuits*, les événements *créés* et les événements *cycliques*. Si les premiers sont le fait du hasard et n'impliquent aucun sujet humain agentif, les événements créés reposent essentiellement sur la présence d'un tel agent humain. Enfin, les événements cycliques sont définis en fonction de leur récurrence. Pour chaque type d'événements, ces auteurs procèdent à une classification plus fine, en fonction de leur aspect temporel, d'une part, et de leur nature sémantique, de l'autre. Ainsi certains événements fortuits sont soit ponctuels (*déclat, éclair*) soit duratifs (*incendie, pluie, épidémie*). Au niveau sémantique, ils énumèrent sept sous-ensembles : les catastrophes naturelles, les incidents techniques, les phénomènes météorologiques, les phénomènes acoustiques, les phénomènes visuels, les épidémies, les maladies contagieuses et les changements d'états.

Les événements créés ont par définition un sujet humain qui participe activement à la réalisation de l'événement en tant qu'organisateur implicite ou explicite. De cette catégorie d'événements se dégagent, entre autres, quatre classes : les événements politiques et diplomatiques, les événements sociaux, les

fêtes, les cérémonies et les spectacles. Les événements cycliques, eux, se distinguent par leur récurrence. C'est le cas des fêtes religieuses (*Noël, Pâques*), nationales (le *14 juillet*) ou privées (*anniversaire*) et de certains événements de la vie sociale (*rentrée des classes*).

Nous considérons cette typologie événementielle comme une contribution importante à la définition de la notion d'événement. La division en événements fortuits, cycliques et créés d'une part et l'importance accordée au domaine, de l'autre, ont été des critères importants dans l'établissement de notre typologie événementielle.

3 Vers un modèle de représentation de l'événement en *text mining*

La littérature spécialisée abonde en propositions de traitement des événements. La plupart des approches proposent l'utilisation des modèles d'apprentissage (moteurs de catégorisation ou de clustering) combinés à une description linguistique, basée essentiellement sur des ontologies ou sur d'autres ressources linguistiques. ([1], [5], [7], [16], [17]). De ces publications, nous retenons la typologie proposée par ACE (*Automatic Content Extraction*) pour les événements, présentée par David Ahn dans : "The stages of Event Extraction" ([1]). Cet auteur distingue 8 types d'événements avec 33 sous-types :

"ACE Events, like ACE entities, are restricted to a range of types. [...]. The eight event types (with subtypes in parentheses) are Life (Be-Born, Marry, Divorce, Injure, Die), Movement (Transport), Transaction (Transfer-Ownership, Transfer-Money), Business (Start-Org, Merge-Org, Declare-Bankruptcy, End-Org), Conflict (Attack, Demonstrate), Contact (Meet, Phone-Write), Personnel (Start-Position, End-Position, Nominate, Elect), Justice (Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon)." ([1] : 2)

Comme on peut le remarquer, les types d'événements sont assez hétérogènes et il est difficile de rendre compte des domaines dont ils relèvent. En revanche, D. Ahn mentionne plusieurs éléments qui permettent de comprendre leur définition : un événement est considéré comme tel, si les éléments ci-dessous sont présents :

"event trigger : the main word which most clearly expresses an event occurrence
event arguments : the mentions that are involved in an event (participants)
event mention : a phrase or sentence within which an event is described, including trigger and arguments"

Nous retenons de cette présentation qu'un événement est désigné par un déclencheur (*event trigger*) et des participants qui correspondent à des arguments de l'événement même.

Cette optique nous semble assez proche de notre propre conception de la notion d'événement, à cette différence près que nous adoptons la position théorique de Gaston Gross, qui à la suite de Z. Harris ([12]) et de M. Gross ([10]), postule que l'unité minimale d'analyse est la phrase simple, cadre dans lequel l'ambiguïté est quasiment inexistante. Par phrase simple, on entend la combinaison d'un prédicat (en text mining cette notion s'appelle une *relation*) et de la suite la plus longue de ses arguments (qui sont donc des *entités*). Or, cette relation ne peut pas être déterminée sur la base de catégories abstraites, car un prédicat est identifié par la nature *sémantique* et non simplement *morphologique* de ses arguments. Le fait de décrire la phrase comme constituée d'un prédicat appartenant à une classe sémantique donnée et acceptant un nombre déterminé d'arguments, définis eux aussi par leur classe sémantique, présente beaucoup

d'avantages. Pour ce qui est des prédicats, leur définition est reproductible : ce sont des lexèmes qui génèrent des arguments et qui sont soumis à la temporalité. A partir de ces propriétés, on est en mesure d'établir une première subdivision des prédicats, qui seront répartis en trois classes : les actions, les états et les événements.

Après cette exposition des fondements linguistiques, nous présentons maintenant notre modélisation de la notion d'événement. Pour ce faire, nous proposons l'analyse de la phrase ci-dessus :

Selon l'agence de presse AFP, hier, à 6h30, un attentat à la bombe contre le consulat américain de Peshawar, a tué 10 personnes et a blessé 4 civils, dont deux touristes français et deux habitants de la ville.

Dans cet exemple, nous distinguons un événement principal et deux événements secondaires. Le premier événement est désigné par *l'attentat à la bombe*, avec son cadre temporel, *Hier, à 6h30*, et le lieu où il s'est déroulé, en l'occurrence la ville de *Peshawar* et un témoin qui le rapporte *selon l'agence de presse AFP*. Le même événement est décrit par les dégâts qu'il a causés, notamment les *10 personnes* qui ont été *tuées* (deuxième événement) et les *4 civils* qui ont été *blessés* (troisième événement).

Le calcul de ce type de construction événementielle nécessite plusieurs étapes. La première consiste à définir le type d'événement même. Il s'agit, dans ce cas, d'un *événement organisé*, puisque tout attentat nécessite un auteur, même si dans la situation présente celui-ci est implicite. Deuxièmement, il est nécessaire de déterminer le mode opératoire de cet événement : *l'attentat* a été perpétré à l'aide d'une *bombe*. Ensuite, il faut identifier son cadre spatio-temporel, ici *hier, à 6h30* et *Peshawar*. En même temps, un attentat implique une cible : dans notre exemple, *le consulat américain de Peshawar*. Enfin, les dégâts humains sont désignés par les substantifs : *personnes, civils, touristes, habitants*. Les dégâts sont évoqués par les verbes *tuer* et *blessé*.

Cette segmentation de la phrase complexe permet, d'une part, de voir la structuration interne de l'événement et, d'autre part, de rendre compte des éléments qui le constituent. Chacun de ces éléments représente une classe sémantique d'arguments, donc d'entités organisées autour du prédicat *attentat*. Le schéma prédicatif de cette phrase peut être représenté comme ceci :

Exemple 1

Témoin (*selon l'agence de presse AFP*) / Temps (*hier, à 6h30*) / Événement (*un attentat à la bombe*) / Cible (*contre le consulat américain de Peshawar*) / Effet (*a tué 10 personnes et a blessé 4 civils, dont deux touristes français et deux habitants de la ville*).

En appliquant différentes restructurations à cette phrase complexe (effacements, réductions), nous pourrions avoir plusieurs schémas prédicatifs :

Exemple 2

Temps (*hier, à 6h30*) / Événement (*un attentat à la bombe*) / Cible (*contre le consulat américain de Peshawar*) / Effet (*a tué 10 personnes et a blessé 4 civils, dont deux touristes français et deux habitants de la ville*).

Exemple 3

Temps (*hier, à 6h30*) / Événement (*un attentat à la bombe*) / Cible (*contre le consulat américain de Peshawar*)

Exemple 4

Événement (*un attentat à la bombe*) / Cible (*contre le consulat américain de Peshawar*)

Exemple 5

Événement (*un attentat à la bombe*) / Cible (*contre le consulat américain*)

Exemple 6

Événement (*un attentat à la bombe*)

La dernière construction présuppose l'existence d'un verbe support qui a été effacé :

Il y a eu un attentat à la bombe.

Les exemples que nous avons donnés ci-dessus expliquent notre modèle de représentation de la notion d'événement. Pour qu'un événement soit considéré comme tel, il suffit d'avoir au minimum un élément (argument) qui l'accompagne. Ces éléments sont assimilés soit aux personnes participant à l'événement (*auteur, cible, témoin*), soit à ses moyens (*à la bombe*) ou à son cadre locatif et temporel. Nous sommes partis d'une structure complexe et, en effaçant successivement des arguments, nous sommes parvenus à la séquence *un attentat à la bombe*, qui tout en étant moins riche que les exemples 1 à 5, représente néanmoins l'élément constitutif de notre définition des événements. Cependant les schémas prédicatifs de 1 à 5 constituent les prototypes des événements que nous modélisons.

4 De la modélisation vers l'extraction de l'événement

Précisons que, dans ce qui suit, nous adoptons la typologie des événements présentée en première partie de cette communication. (cf. 2.), notamment la division des événements en *fortuits, créés* et *cycliques*. A la différence de l'approche ACE, nous limitons notre description des événements aux risques terroristes et stratégiques.

Chez Temis, le modèle d'extraction de l'information repose sur l'utilisation des patrons d'extraction. Nous modélisons et organisons l'information à extraire selon une hiérarchie de composants linguistiques modulaires intégrables à différents domaines d'activité ou à d'autres langues. Ces composants sont appelés des *Skill Cartridges*TM. Un composant de connaissance inclut, d'une part, des dictionnaires (afin d'assigner un sens aux mots et aux phrases) et, d'autre part, des règles d'extraction, qui calculent les concepts et les relations existant entre ces concepts. Les composants de connaissance sont couplés au moteur d'extraction, *Insight Discoverer*TM *Extractor* (IDE) et à l'analyseur morpho-syntaxique Xelda®, qui identifie la langue et procède à l'étiquetage grammatical des séquences de mots.

Le moteur d'extraction est intégré dans une plateforme d'enrichissement de contenu et de découverte de la connaissance, appelée Luxid®. L'un des objectifs de cette plateforme technologique est d'annoter de grandes masses de documents de différents formats sur la base des composants de connaissance. Une fois les documents annotés, le moteur d'analyse et de découverte Luxid® permet de cibler les documents les plus pertinents et de faire des analyses en

profondeur des informations extraites (analyses temporelles, analyses croisées et de proximité entre les entités, navigation sémantique pour les relations entre les entités, etc.). Ainsi, l'information pertinente n'est pas seulement identifiée, mais elle représente le point de départ pour des analyses plus détaillées. Ajoutons à cela que notre système est capable de gérer plusieurs langues et que les modèles d'extraction sont adaptés à plusieurs domaines d'activité. Parmi les domaines d'application de nos composants de connaissance, nous énumérons les sciences de la vie, l'intelligence économique, l'analyse des sentiments et le domaine du risque. C'est dans cette dernière catégorie que s'inscrit le composant de connaissances **Events™**.

5 Le composant de connaissances Events™ : Principes méthodologiques et spectre de la cartouche

La cartouche *Events™* a été développée afin de permettre la détection d'événements des domaines militaire, policier et stratégique. Les objectifs de la cartouche sont d'extraire, d'une part, des entités nommées des domaines mentionnés, et, d'autre part, les relations dans lesquelles sont impliquées ces entités. Les entités et les relations de la cartouche *Events™* ont été organisées sous la forme de classes, sur la base de leurs propriétés syntaxiques et sémantiques. Pour la mise au point de ces classes, nous nous sommes servi essentiellement de sources ouvertes, notamment des dictionnaires de spécialité, des ouvrages de référence pour certaines entités nommées ([4]) et des sites de référence des organismes nationaux ou internationaux.

5.1 Organisation des entités

L'architecture sémantique des entités qui sont extraites avec la cartouche **Events™** a été déterminée par la nature même du domaine dont relève ce produit, en l'occurrence le risque militaire et stratégique. Plusieurs axes définissent l'organisation de ces entités : nous nous intéressons d'abord aux acteurs impliqués dans événements terroristes et stratégiques (*forces terroristes, forces armées d'un état, criminels, civils* et *organismes politiques*). Ensuite, nous décrivons le mode opératoire (*armement, moyens d'attaque*) associé aux actions terroristes et aux opérations militaires. Troisièmement, nous nous intéressons à la nature des lieux où ces événements se produisent (*lieux géographiques, lieux fonctionnels, lieux géophysiques, lieux commerciaux*, etc.). Enfin, nous visons la détection des produits qui font l'objet de trafic et d'activités illégales (*drogues, objets d'art, différentes armes*). A chacune des entités mentionnées peuvent correspondre des sous-classes. Les *Armées* et les *Forces de police*, par exemple, sont subdivisées en collectifs (*régiments, patrouille, bataillon, forces de sécurité*) et en individus (*policier, gendarme, soldat*), à quoi il faut ajouter parallèlement les grades qui caractérisent ces individus (*officier de police, lieutenant d'aviation*).

Notons aussi que la classification des entités n'a pas été faite à partir de mots-clés, mais sur la base des propriétés syntactico-sémantiques de chaque type. Pour donner un exemple concret, les armes ont été sous-divisées en *Armes blanches, Projectiles, Armes explosives, Armes biologiques*, etc., selon leurs propriétés linguistiques. Une arme blanche n'a pas les mêmes constructions appropriées qu'une arme à feu : on *tire* avec un fusil, on *poignarde* avec un couteau, etc. Il y a donc un intérêt évident à faire des classes sur la base de la syntaxe et de la sémantique.

Parmi les autres classes d'entités de la cartouche *Events*, nous mentionnons les *Organisations Terroristes*, recensées dans des ouvrages de référence comme l'*Encyclopédie des terrorismes et violences organisées* ([4]) ou l'*Atlas du terrorisme* (Courrier International 2008) ou encore les *Phénomènes de risque* (piraterie, terrorisme, crime organisé). Chacune de ces classes d'entités est extraite de façon autonome ou dans le cadre d'une relation événementielle.

5.2 Organisation des relations

Comme nous l'avons signalé, une relation dans le cadre de la cartouche *Events* est désignée sous le nom d'*événement*. Dans notre perspective, un événement est caractérisé par un cadre spatio-temporel et des acteurs dont le rôle peut varier (auteur, assaillant, belligérant, cible, etc.). Avant de préciser la nature de ces acteurs, il est nécessaire de préciser les types de relations événementielles extraits à l'aide du composant de connaissance *Events*. Grosso modo, il y en a trois grandes classes. La première désigne des événements impliquant la participation des organismes de protection d'un pays. Il peut s'agir de l'une des trois armées (*armée de l'air, armée de terre, marine*), de forces de police ou bien d'organisations politico-militaires nationales ou internationales (*OTAN, GIGN*).

Ces acteurs, qui combattent de façon générale les phénomènes de risque comme le terrorisme ou la piraterie et, de façon spécifique, les groupes ou les organisations terroristes et criminelles, sont les acteurs-clés dans les événements suivants : *Etat conflictuel* (*Conflict*), *Combat* (*Fighting*), *Opération de police* (*Police Operation*), *Opération Militaire* (*Military Operation*). En voici quelques exemples :



Figure 1: Type de relation événementielle dans Luxid® : la classe des conflits (conflit, lutte, guerre)



Figure 2: Type de relation événementielle dans Luxid® : la classe des combats (combat, accrochage, affrontement)



Figure 3 : Type de relation événementielle dans Luxid® : la classe des opérations militaires (embuscade, tir, mouvement de troupes)

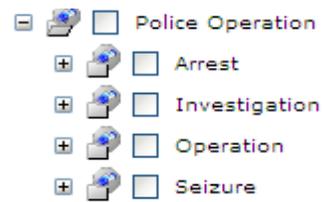


Figure 4: Type de relation événementielle dans Luxid® : la classe des opérations de police (arrestation, enquête)

Les Figures 1, 2, 3, 4 illustrent la structure hiérarchique des classes d'événements, qui peuvent comporter des sous-classes (*Military Operation, Police Operation*). Pour que soit extrait un événement relevant de ces classes, il faut qu'une ou plusieurs entités parmi celles que nous avons décrites plus haut soient calculées. Un exemple de modélisation de ce type d'événements est donné par la Figure 5, où la phrase : *L'armée italienne a été déployée samedi le 3 Mars dans la région de Caserte* est extraite comme un événement militaire.

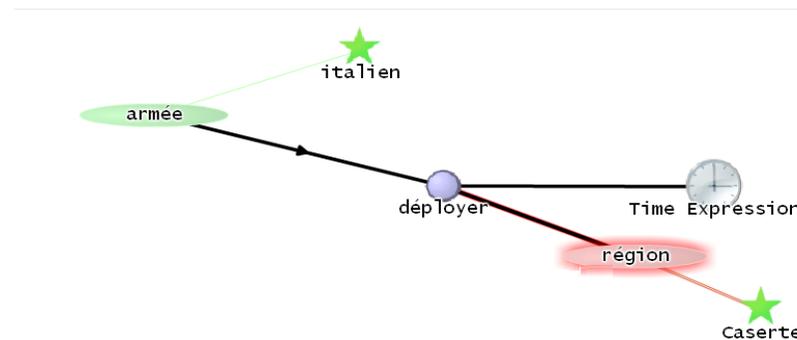


Figure 5 : Exemple d'un événement de type militaire dans Luxid®

La figure 5 désigne un graphe sémantique où un acteur, *l'armée italienne*, participe à un événement de la classe des *Opérations militaires*, désigné par le verbe *déployer*. Cet événement a lieu dans *la région de Caserte* et comporte un cadre temporel relatif, exprimé par *samedi le 3 Mars* (Time Expression). Notons, d'une part, que les occurrences des entités et des relations qui apparaissent dans le graphe sont lemmatisées (la forme *a été déployée* est remplacée par son lemme *déployer*) et que, d'autre part, la classe des *Opérations Militaires* est désignée par l'occurrence de *déployer*, qui fait partie de la sous-classe d'événements *Mouvements de troupes*.

Nous venons de présenter un cas de figure décrivant des événements dont les acteurs désignent des organisations ou des individus chargés de protéger la population d'un Etat. Nous évoquons rapidement les autres types d'événements détectés par la cartouche *Events*. Il en existe dont les acteurs principaux sont des groupes ou des individus terroristes ou criminels, qui projettent de mettre en danger la sécurité d'un Etat, d'une population ou des personnalités politiques et militaires. En voici quelques illustrations sous la forme de captures d'écran, avec des exemples d'extraction :



Figure 6 : Type de relation événementielle dans Luxid® : la classe des crimes (assassinat, massacre, génocide : *Le génocide arménien a eu lieu d'avril 1915 à juillet 1916*).

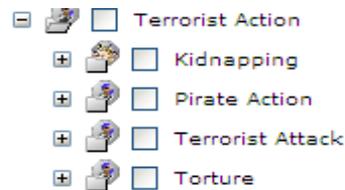


Figure 7 : Type de relation événementielle dans Luxid® : la classe des actions terroristes (attentat, enlèvement, acte de piraterie, acte de torture : *Un attentat à la bombe contre l'ambassade américaine à Nairobi a fait hier trois morts*).

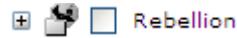


Figure 8 : Type de relation événementielle dans Luxid® : la classe des rébellions (rébellion, coup d'état, révolution : Une rébellion sanglante a éclaté dans le pays il y a moins d'une semaine).

Dans la même lignée d'événements désignant des risques, se situent les activités illicites (*export illégal des armes*), le trafic de drogues, la contrebande ou la contrefaçon. Pour ce type de relations événementielles, nous nous intéressons aux entités qui font l'objet de trafics comme les drogues, les médicaments, les armes, les œuvres d'art, le tabac, etc. Parmi les autres événements que nous avons modélisés, nous mentionnons ceux qui représentent la suspension ou la fin des hostilités (*trêve, cessez-le-feu*), des explosions accidentelles ou criminelles (par exemple : *Un kamikaze s'est fait exploser devant le palais de justice de Peshawar, tuant 35 personnes*).

Les classes d'événements que nous venons de présenter définissent le spectre de la cartouche *Events*TM, orientée essentiellement vers le domaine de la gestion du risque stratégique. Le composant de connaissance *Events*TM a été désigné afin de détecter des événements uniquement en français. Cependant, la généralité du modèle de représentation de la notion d'événement rend possible une adaptation facile à d'autres langues et à d'autres domaines d'activité.

6 Evaluation et corpus

Le corpus de développement pour *Events*TM a été constitué de *news* de différentes agences de presse (*Le Monde, Reuters, AFP*) portant sur l'actualité politique nationale et internationale et de documents de spécialité, notamment le bulletin électronique d'actualité, RENSEIGNOR, qui offre un panorama complet de l'actualité du renseignement, du terrorisme, de la criminalité dans le monde.

L'évaluation qualitative des extractions effectuées avec le composant de connaissance *Events*TM a été faite dans trois étapes. Dans un premier temps, le corpus de développement (une centaine de documents) a été annoté manuellement, à l'aide de l'outil d'annotation textuelle Knowtator, <http://knowtator.sourceforge.net/quickstart.shtml>, qui permet de reproduire une arborescence de classes sémantiques identique aux entités et aux relations de la cartouche *Events*TM. Cette étape a débouché sur la mise en place d'un corpus annoté manuellement, avec un taux de rappel et de précision de 100%.

Dans un deuxième temps, nous avons calculé la précision des extractions de la cartouche *Events*TM via un outil d'évaluation interne à Temis, le *Skill Cartridge Quality Tool* (SCQT). Ce dernier a permis de valider l'exactitude ou l'inexactitude des entités et des relations extraites avec notre produit, grâce à l'utilisation d'une évaluation tridimensionnelle. Pour donner un exemple concret, pour invalider ou valider une extraction, nous avons mis une étiquette « Wrong » sur l'entité ou la relation qui n'a pas été correctement extraite par la cartouche, une autre étiquette « Good » a été mise sur les extractions précises et enfin une troisième étiquette, « Good Type », a été mise sur les extractions correctes mais incomplètes. Cette approche a donc donné le taux de précision de la cartouche qui se situe entre 75% et 80%.

La troisième étape a consisté en une comparaison entre le corpus de référence (le corpus annoté) et le corpus comprenant les extractions validées avec le SCQT. La moyenne des différences entre les deux corpus a conduit au calcul du taux de rappel. Cette évaluation a montré que le taux de rappel du composant de connaissance *Events*TM est de 70%.

7 Conclusion

Dans cette communication, nous nous sommes donné comme objectif de présenter une méthodologie appropriée à l'extraction d'événements d'un domaine spécifique, en l'occurrence le domaine militaire et stratégique, dans une perspective de veille stratégique. Après avoir présenté différentes approches de la détection des événements, nous avons proposé notre propre modèle, qui permet de détecter des événements liés au risque terroriste.

Nous avons montré que notre méthode de représentation des événements est innovante de par sa granularité. Alors que la plupart des méthodes de détection des événements ne visent pas une description fine de la nature des entités impliquées dans un événement, notre approche propose une description en extension des éléments composant un événement.

La prise en compte de l'ensemble des entités qui participent à la définition de la notion d'événement (acteurs, mode opératoire, lieu de déroulement), nous permet de saisir chaque événement sous tous les angles par lesquels il peut être appréhendé : son auteur, le témoin, la victime ou la cible, ce qui peut présenter de nombreux avantages pour la veille stratégique. Sur de grandes masses de données, la factorisation de ces informations peut apporter des pistes intéressantes aux analystes qui visent le démantèlement des actes terroristes et qui ont besoin de descriptions très détaillées du mode opératoire des organisations terroristes ou de la nature de leur cible.

Ainsi, avec le composant de connaissance *Events*TM, nous avons une illustration du fait que le text mining vient ainsi apporter une aide précieuse aux organisations qui combattent le terrorisme et le crime organisé.

8 Bibliographie

- [1] AHN, D. *The stages of Event Extraction* in Proceedings of the Workshop on Annotating and Reasoning about Time and Events, ACL Workshops, 2006 , pp 1-8.
- [2] APPELT D., HOBBS J., BEAR J., ISRAEL D., KAMEYAMA M, TYSON M. *FASTUS: a finite-state processor for information extraction from real-world text* in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'93), Chambéry, 1993, pp. 1172-1178.
- [3] AUBRY C., GRIVEL L., GUILLEMIN-LANNE S., LAUTIER C. , F. *Aide à la construction de composants de connaissance pour l'extraction d'information : méthodologie et environnement*, CIFT 2002 Colloque International sur la Fouille de Textes, Hammamet- Tunisie, 2002.
- [4] BAUD J. *Encyclopédie des terrorismes et Violences politiques*, Charles Lavauzelle, 2003.
- [5] BLACK J.A., RANJAN N. DAVID, B. *Automated Event Extraction from Email*, <http://nlp.stanford.edu/courses/cs224n/2004/jblack-final-report.pdf>, 2004.

- [6] DELECROIX B., GUILLEMIN-LANNE S., SIX A. *Veille concurrentielle et veille stratégique : deux applications d'extraction d'information* VSST 2004 Veille Scientifique et Stratégique, Toulouse, 25-29 oct 2004, pp 117-128.
- [7] GRISHMAN R., JI H. *Refining Event Extraction through Cross-document Inference* in Proceedings of ACL-08: HLT (June 2008), 2008, pp. 254-262.
- [8] GROSS G. *Actions, états et événements constituent-ils des ensembles disjoints*. In LARRIVEE Pierre, (ed), Variation et stabilité du français. *Des notions aux opérations. Mélanges de linguistique français offerts au professeur Jean-Marcel Léard par ses collègues et amis*, Bibliothèque de l'Information grammaticale, Peeters. Louvain, 2007, pp. 107-114.
- [9] GROSS G., KIEFER F. *La structure événementielle des substantifs*, Folia linguistica 29 (1-2), 1995, pp. 43-65.
- [10] GROSS M. *Méthodes en syntaxe*, Hermann, 1975.
- [11] GUILLEMIN-LANNE S, SIX A. *La normalisation : nouveau challenge en extraction d'information* VSST 2006 Veille Scientifique et Stratégique, Lille, 16-17 janvier, 2006.
- [12] HARRIS Z. *Notes du cours de syntaxe*, Le Seuil, 1973.
- [13] JUNGERMANN F., MORRIK K. *Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 335-336, 2008.
- [14] LEE S.H. 2001 *Les classes d'objets d'événements. Pour une typologie sémantique des noms prédicatifs d'événements en français*, thèse de doctorat, Université Paris 13, 2001.
- [15] PAUNA R. *Les causes événementielles*, thèse de doctorat, Université Paris 13, 2007.
- [16] PISKORSKI J., TANEV H., ATKINSON M., VAN DER GOOT E. *Cluster-Centric Approach to News Event Extraction*, In Proceedings of MISSI 2008, Wroclaw, Poland, 2008.
- [17] PISKORSKI J., TANEV H., ATKINSON M. *Real-Time News Event Extraction for Global Crisis Monitoring*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, pp. 207-218.
- [18] POIBEAU T. *Extraction d'information à base de connaissances hybrides*, Thèse, Université Paris-Nord, 2002.
- [19] TRONCY R. *Bringing The IPTC News Architecture into the Semantic Web*. In 7th International Semantic Web Conference (ISWC'08), Karlsruhe, Germany, 2008, pages 483–498.
- [20] TRONCY R. *Explorer des actualités multimédia dans le web de données*, <http://homepages.cwi.nl/~troncy/Publications/Troncy-ic09.pdf>, 2009.
- [21] YANGARBER R., GRISHMAN R. *Customisation of Information Extraction Systems* In Paziienza, M. T., editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Springer Verlag, Heildelberg, 1997, pp. 1-11.