RECITAL 2008, Avignon, 9–13 juin 2008

Annotation des informations temporelles dans des textes en français.

André Bittar
Université Paris 7 Diderot - ALPAGE
andre.bittar@linguist.jussieu.fr

Résumé. Le traitement des informations temporelles est crucial pour la compréhension de textes en langue naturelle. Le langage de spécification TimeML a été conçu afin de permettre le repérage et la normalisation des expressions temporelles et des événements dans des textes écrits en anglais. L'objectif des divers projets TimeML a été de formuler un schéma d'annotation pouvant s'appliquer à du texte libre, comme ce que l'on trouve sur le Web, par exemple. Des efforts ont été faits pour l'application de TimeML à d'autres langues que l'anglais, notamment le chinois, le coréen, l'italien, l'espagnol et l'allemand. Pour le français, il y a eu des efforts allant dans ce sens, mais ils sont encore un peu éparpillés. Dans cet article, nous détaillons nos travaux actuels qui visent à élaborer des ressources complètes pour l'annotation de textes en français selon TimeML - notamment un guide d'annotation, un corpus de référence (*Gold Standard*) et des modules d'annotation automatique.

Abstract. The processing of temporal information is crucial for the understanding of natural language texts. The specification language TimeML was developed to facilitate the identification and normalization of temporal expressions and events in texts written in English. The aim of the various TimeML projects was to formulate an annotation scheme able to be applied to free text, such as that which is found on the Web, for example. Recently, efforts have been made to apply TimeML to languages other than English, namely Chinese, Korean, Italian, Spanish and German. Some efforts have been made in this direction with respect to French, but they remain somewhat scattered. In this paper, we detail our ongoing work, which aims to establish comprehensive resources for the annotation of French texts according to TimeML - an annotation guide, a *Gold Standard* corpus and modules for automatic annotation.

Mots-clés : Annotation temporelle, repérage des événements, TimeML.

Keywords: Temporal annotation, event recognition, TimeML.

1 Introduction

Depuis quelques années, l'importance des informations temporelles dans la compréhension de la langue naturelle a motivé le développement d'outils pour le repérage et la normalisation de ce type de données avec le but d'améliorer les performances et la couverture de certaines applications, telles que les systèmes de questions-réponses ou l'extraction d'information.

Le langage de spécification TimeML, qui est l'aboutissement de divers projets dans le domaine de l'annotation des informations temporelles, fournit un schéma d'annotation pour les expressions temporelles et les événements ainsi que les relations auxquelles ils participent. Les travaux

ont été menés sur l'annotation de l'anglais. Un corpus de référence (*Gold Standard*), TimeBank, a été élaboré, ainsi que des outils d'annotation manuelle et automatique pour les événements, les expressions temporelles et les relations qui gouvernent les deux.

Étant donné l'absence de convergence des différentes ressources pour le français et leur incomplétude partielle, le développement d'un ensemble cohérent d'outils nous a semblé pertinent. Dans cet article, nous présentons nos travaux pour la création de ressources TimeML pour le français. La section 2 consistera en une description du langage TimeML et des travaux sur l'anglais. La section 3 fera le tour des travaux préexistants sur d'autres langues, y compris le français. La section 4 couvrira l'élaboration d'un corpus de référence pour le français. La section 5 détaillera les modules que nous sommes en train d'élaborer pour le repérage et l'annotation automatique des événements et des expressions temporelles.

2 Le projet TimeML

2.1 Le schéma d'annotation

TimeML (Pustejovsky *et al.*, 2003) est un langage de spécification issu de l'atelier TERQAS[1] dans le cadre du projet AQUAINT. Ce projet vise à améliorer les systèmes de questions-réponses en leur permettant de répondre à des questions de nature temporelle sur les entités et les événements. Dans ce cadre, TimeML a été élaboré comme langage d'annotation pour faciliter le raisonnement et l'inférence sur le temps.

Le schéma d'annotation prévoit les fonctionnalités suivantes : l'annotation des événements, l'étiquetage des expressions temporelles et la normalisation de leur valeur ainsi que la mise en évidence des relations qui existent entre ces deux types d'entités temporelles. Les traits de temps verbaux, la polarité et la modalité ainsi que la classe d'événement peuvent aussi être annotés. Contrairement à des notions traditionnelles, TimeML adopte une conception large des événements qui regroupe les événements et certains états (ce qui correspond plus à la notion d'*éventualité* (Bach, 1986)). En plus de la plupart des verbes, cette définition comprend des noms événementiels comme *destruction* et *guerre*, ainsi que des adjectifs (*malade*) et les groupes prépositionnels (*à bord*) qui désignent typiquement des états. TimeML compte 7 classes d'événements différents : `ASPECTUAL`, `I_ACTION`, `I_STATE`, `OCCURRENCE`, `PERCEPTION`, `REPORTING` et `STATE`. Nous orientons le lecteur vers (Saurí *et al.*, 2005) pour une description de ces classes. Les événements sont annotés avec la balise `EVENT`. Le schéma d'annotation TimeML précise que c'est la tête lexicale du chunk événementiel qui doit être annotée. Ce choix est fait afin de simplifier l'annotation (le processus et le résultat), notamment en vu des difficultés présentées par les propositions enchâssées ou celles contenant plusieurs verbes[2]. Cette simplification constitue une première étape dans le repérage des événements. (Pustejovsky *et al.*, 2006) ont proposé une extension du schéma d'annotation qui consiste en l'ajout de balises pour capturer les arguments des événements, mais cette proposition n'a pas encore été intégrée.
Les informations sur la polarité (attribut `polarity`), l'aspect (`aspect`) et la modalité (`modality`) sont également représentées à l'intérieur de la balise `EVENT`.

Les expressions temporelles sont marquées par la balise `TIMEX3`. Elles se divisent en 4 classes :

[1] http ://www.timeml.org/site/terqas/

[2] Voir http ://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf p135-137 pour une discussion.

les dates (type DATE, *le 15 janvier, 15.01.2008*), les heures (type TIME, *15h20, l'après-midi*), les durées (type DURATION, *5 jours, deux ans*) et les ensembles (type SET, *tous les jours, chaque année*). TimeML permet aussi le raisonnement avec des expressions temporelles sous-spécifiées, comme *lundi prochain* et *l'année précédente*, dont la valeur doit être déterminée par rapport à un point temporel de référence.

Les événements et les expressions temporelles sont mis en relation par trois sortes de liens (*links*) : liens temporels (TLINK), aspectuels (ALINK) et de subordination (SLINK). La première sorte capture des relations temporelles entre deux entités (EVENT–EVENT, TIMEX3–TIMEX3, ou EVENT–TIMEX3), la deuxième capture les phases dans le déroulement d'un événement et la dernière est essentielle pour les raisonnements qui dépendent de la véracité ou la certitude des propositions qui dénotent des événements. Les mots fonctionnels qui signalent explicitement un de ces liens sont annotés avec la balise SIGNAL. Le plus souvent ce sont des prépositions comme *avant, après, pendant* ou *lors de*.

Ci-dessous figure un exemple simplifié d'annotation avec les balises principales y compris un lien (TLINK) pour la relation temporelle entre l'événement et l'expression de date, pour la phrase *Jean est arrivé avant le 11 février 2004* :

```
Jean est <EVENT id="e1" class="OCCURRENCE" pos="VERB" tense="PAST"
polarity="POS">arrivé</EVENT> <SIGNAL>avant</SIGNAL> le
<TIMEX3 id="t1" val="2004-02-11">11 février 2004</TIMEX3> .
<TLINK relType="BEFORE" event="e1" time="t1"/>
```

2.2 Les ressources pour l'anglais

Pour l'anglais, un corpus de référence, TimeBank, a été élaboré manuellement[3]. Des modules d'annotation automatique ont également été développés dans le cadre du projet TARSQI[4]. Evita (Saurí *et al.*, 2005) est un système de reconnaissance d'événements, basé sur des stratégies linguistiquement motivées et des données statistiques. Sa performance a été mesurée à 74.03% de précision et 87.31% de rappel, une F-mesure de 80.12%.
GUTime (Mani & Wilson, 2000), basé sur la reconnaissance de patrons, étiquette les expressions temporelles et normalise leur valeur. Il a été évalué à 85% et 82% de F-mesure respectivement pour les tâches de reconnaissance et de normalisation. Des outils pour l'annotation des liens ont également été développés[5], mais des chiffres précis sur leur évaluation ne sont pas disponibles. L'ensemble de ces modules vient d'être intégré dans le TARSQI Toolkit. Plus récemment, lors de l'atelier TempEval 2007[6], une évaluation d'un ensemble de systèmes a été effectuée pour trois tâches portant sur l'identification des relations temporelles dans des textes en anglais. Les participants ont été évalués sur une section du corpus TimeBank.

[3]Il existe désormais d'autres corpus manuellement annotés - le corpus AQUAINT et celui de TempEval 2007 (Voir http ://www.timeml.org/site/timebank/timebank.html)

[4]http ://www.timeml.org/site/tarsqi/

[5]voir http ://www.timeml.org pour les détails

[6]http ://timeml.org/tempeval/

3 Travaux sur d'autres langues

3.1 Le chinois, le coréen, l'italien, l'espagnol et l'allemand

Le projet TimeML a été adopté par l'ISO et une norme est en cours d'élaboration avec une perspective d'adaptation à d'autres langues que l'anglais. Dans l'établissement de cette norme des efforts ont déjà été faits pour son adaptation au chinois, au coréen et à l'italien avec des guides d'annotation pour ces langues [7]. Il existe également divers travaux sur l'annotation automatique TimeML pour certaines langues.

(Caselli *et al.*, 2007) ont mené une étude sur l'annotation manuelle des événements en italien et les bénéfices d'un lexique sémantique pour guider cette tâche. (Negri & Marseglia, 2004) ont développé un module à base de règles pour la reconnaissance et la normalisation des expressions temporelles. (Saquete *et al.*, 2006) ont développé, dans une optique multilingue, un système semblable pour l'espagnol et l'italien.

Pour le chinois (Cheng *et al.*, 2007) détaille un guide d'annotation pour la création d'un corpus annoté en relations temporelles ainsi que la construction d'un modèle d'apprentissage pour l'annotation automatique de ces relations.

(Jang *et al.*, 2004) décrit un étiqueteur temporel pour des articles de presse en coréen. Ce module se focalise sur le repérage et la normalisation des expressions temporelles. Le système, basé sur des techniques d'apprentissage, atteint une F-mesure de 87% pour cette tâche.

Une nouvelle approche, qui permet de pallier l'absence de corpus annotés en allemand, est décrit dans (Spreyer & Frank, 2008). Les auteurs proposent une méthode statistique pour la création d'un étiqueteur temporel basée sur la projection des annotations dans un corpus aligné anglais-allemand issu d'EuroParl (Koehn, 2005). Pour la tâche de reconnaissance des événements le système atteint une précision de 83.95% et un rappel de 34.44% (F-mesure 48.84%) et le repérage des expressions temporelles une précision de 89.52%, 51.79% de rappel (F-mesure 65.62%). Cette évaluation a été réalisée sur un corpus de 236 phrases alignées.

3.2 Travaux précédents sur le français

Même s'ils n'exploitent pas le schéma TimeML, nous citons d'abord les travaux de (Gagnon & Lapalme, 1996), qui présentent une méthode pour la génération de textes en français qui véhiculent des informations temporelles. Les auteurs expliquent en détail la base formelle sur laquelle repose l'implémentation d'un prototype de générateur de textes et précisent également les étapes suivies dans la génération des différents éléments temporels - les temps verbaux et les adverbiaux temporels.

Plus récemment, (Battistelli *et al.*, 2006) propose une analyse fonctionnelle pour la représentation formelle des expressions calendaires. Les auteurs présentent un projet pour l'implémentation de cette formalisation dans le cadre d'un système de navigation temporelle dans des textes biographiques.

Les travaux menés sur le français par (Muller & Tannier, 2004) se concentrent sur une évaluation de la faisabilité et de la complexité de la tâche d'annotation des relations temporelles ainsi que

[7] Voir http ://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf pour une version préfinale

sur une méthode de calcul de ces relations. Ces travaux décrivent un module qui vise à expliciter les relations temporelles qui existent entre les verbes finis d'un texte et détaillent une méthode d'évaluation sur un corpus d'articles de presse manuellement annotés. Muller a travaillé depuis ces travaux sur le repérage des expressions et des relations temporelles et a collaboré avec Michel Gagnon sur le repérage des événements. Leurs résultats n'ont pas encore été rendu publics.

4 Corpus de référence et guide d'annotation pour le français

L'établissement d'un corpus de référence *Gold Standard* est essentiel pour l'évaluation de modules d'annotation automatique et pourrait, si sa taille et sa qualité le permettent, servir de base à des algorithmes d'apprentissage. Un de nos objectifs est d'annoter un ensemble de documents en français pour arriver à un corpus de référence de bonne qualité. Les types de texte qui nous intéressent particulièrement sont les articles ou les dépêches de presse, les documents encyclopédiques ou historiques et les textes biographiques du fait d'un fort contenu temporel. Philippe Muller a annoté manuellement un ensemble d'articles de presse et de biographies (non redistribuables) en TimeML. Nous avons récemment annoté les mêmes textes et envisageons une comparaison de nos annotations respectives afin de déterminer le taux d'accord, ce qui donnera une idée des performances que l'on peut attendre d'un module d'annotation automatique. Notre corpus de référence est constitué pour l'instant de 30 articles de presse (15 876 mots au total, 529 mots par article en moyenne) annotés à la main avec les outils Callisto et Tango (Verhagen *et al.*, 2005) conçus pour l'annotation de l'anglais. Nous comptons élargir ce corpus avec des textes biographiques et encyclopédiques afin de varier sa couverture.

Afin de garantir une annotation la plus uniforme possible, nous élaborons un guide d'annotation qui explique ce qu'il faut et ne faut pas annoter, et quelles annotations donner au différentes formes linguistiques pertinentes. Les particularités du français doivent être prises en compte. Par exemple, contrairement à l'anglais, les verbes modaux du français tels que *devoir* et *falloir* peuvent être conjugués à tous les temps. Cela a pour conséquence qu'il est nécessaire d'expliciter leurs traits de temps, mais aussi de mode et de polarité dans l'annotation. Nous les annotons comme `EVENT`, même si ce ne sont pas des événements au sens strict du terme. Nous utilisons donc une huitième classe, `MODAL`, pour ces verbes. Une autre particularité du français (qui le distingue notamment de l'anglais) concerne l'utilisation du conditionnel pour marquer une certaine modalité par rapport à un fait prétendu, comme dans la phrase suivante : *Les gardes du corps des parlementaires auraient fait plus de victimes que l'attentat-suicide selon un rapport interne*. Pour repérer cet emploi du conditionnel (fréquent dans les articles de presse) et le distinguer de l'emploi plus habituel, nous attribuons la valeur `CONJECTURAL` à l'attribut `mood` de la balise `EVENT` (cette valeur `CONJECTURAL` a aussi été postulée pour le coréen).

Nous adhérons à la norme ISO-TimeML dans l'élaboration de ce guide. Si cette norme ne diffère pas beaucoup du langage TimeML original, elle est développée dans un esprit multilingue et ne contient pas certaines redondances de son prédécesseur (notamment par la suppression de la balise `MAKEINSTANCE` qui sert à réaliser les différentes instances d'un événement donné).

Ce guide d'annotation est agrémenté au fur et à mesure de l'annotation du corpus de référence avec les éventuelles adaptations nécessaires. Notre objectif est qu'il soit basé à la fois sur la théorie linguistique et sur une étude de corpus.

5 Modules d'annotation automatique

5.1 Approche adoptée

Nous optons pour une approche qui repose sur des règles linguistiques. Ce choix s'impose tant pour des raisons théoriques que techniques. Premièrement, il nous semble intéressant d'un point de vue linguistique d'aborder les problématiques spécifiques au français vis-à-vis des tâches en question. Cela est nécessaire pour l'élaboration de ressources qui puissent être partagées, et surtout pour la création d'un guide d'annotation. Deuxièmement, nous ne disposons pas de la quantité de corpus annoté qui serait nécessaire pour une approche basée sur l'apprentissage automatique. Enfin, une approche par projection (cf section 3.1) donne pour l'instant des résultats moins bons que ceux obtenus par des systèmes symboliques.

5.2 Annotation automatique d'expressions temporelles

Nous concevons actuellement un module pour l'annotation des expressions temporelles qui se base sur les graphes du package `Time_French` (Gross, 2002) pour Unitex[8]. Les graphes de `Time_French` reconnaissent, de façon très exhaustive, des patrons de date, d'heure, de durée et de fréquence. Néanmoins, la sortie d'Unitex ne garde pas la trace du type d'expression temporelle. Or le type d'expression temporelle (`DATE`, `DURATION`, `TIME` etc.) est une information essentielle en TimeML. Les graphes de Maurice Gross ne correspondant pas aux catégories des expressions temporelles de TimeML, il a été nécessaire de réorganiser totalement ces graphes et de leur rajouter les balises `TIMEX3` afin de respecter la typologie TimeML, ce qui correspond techniquement à transformer des automates en transducteurs. Par exemple, dans `Time_French`, *l'an 2000* et *l'an prochain* sont reconnus dans le même graphe alors que ces deux expressions sont de types différents en TimeML. Le premier est de type `DATE` (absolue) et le deuxième de type `DATE-TemporalFunction` (une fonction temporelle dont la valeur doit être résolue par rapport à un point temporel de référence, comme *hier* ou *la semaine prochaine*).

Enfin, de nombreuses expressions temporelles de `Time_French` n'ont pas été utilisées car elles sortent du cadre de TimeML qui s'accroche aux relations surfaciques et calculables de façon opérationnelle. Par exemple, *à diverses reprises, à l'heure convenue, en toute éventualité*. Cela ne veut pas pour autant dire que ces expressions ne devraient pas être annotées.

Si Unitex s'avère très utile pour la création de nos transducteurs, nous n'utilisons néanmoins pas cette application pour la suite de notre traitement pour des raisons expliquées dans (Sagot *et al.*, 2008). L'ensemble de transducteurs sert, après conversion, de grammaire d'entrée à SxPipe (Sagot & Boullier, 2005), un système pour le prétraitement de textes destinés à l'analyse syntaxique, qui effectue la reconnaissance et l'étiquetage des patrons en TimeML sur du texte brut. Après l'étape de reconnaissance une deuxième passe (avec un script Perl) sur le texte effectue la normalisation des valeurs des `TIMEX3` et détermine un éventuel point temporel de référence pour le calcul des expressions temporelles sous-spécifiées.

[8] http ://www-igm.univ-mlv.fr/ unitex/

5.3 Annotation automatique des informations événementielles

Le but de ce module est de repérer et baliser (EVENT) les occurrences d'événements (au sens TimeML) et de les classifier selon l'ontologie définie par le schéma d'annotation. Il doit également détecter la présence d'éléments à polarité négative, déterminer la catégorie aspectuelle des verbes et d'éventuels éléments portant sur la modalité de l'événement. Dans l'annotation d'un événement, le module doit annoter la tête lexicale du constituant événementiel, comme montré dans les exemples qui suivent, où les attributs de la balise <EVENT>sont omis,

Le chat a <EVENT>*mangé*</EVENT> *la souris*
La violente <EVENT>*destruction*</EVENT> *de la ville*
Jean était <EVENT>*malade*</EVENT> *pendant deux jours*

Un certain nombre de prétraitements sont nécessaires avant l'intervention de ce module. Son entrée doit être un texte segmenté en phrases avec une tokenisation, un étiquetage en parties du discours, une analyse morphologique flexionnelle et une analyse syntaxique de surface (chunking). Le système SxPipe, utilisé pour l'annotation des expressions temporelles, n'effectuant pas encore ces pré-traitements, nous utilisons la chaîne de traitement Macaon (Acosta & Nasr, 2007) pour l'annotation des informations événementielles. Nous espérons disposer dans le futur d'une chaîne de pré-traitements complète permettant de réaliser à la fois l'annotation d'expressions temporelles et celle des informations événementielles et donc de permettre le calcul des relations entre les deux (cf section 5.4).

Dans son état actuel, ce module repose fortement sur des données lexicales. Pour la reconnaissance des noms événementiels, il dispose d'un lexique (en cours d'élaboration) des noms ayant au moins *une* interprétation événementielle. Il sera à terme bénéfique de séparer les noms dont la seule interprétation est celle d'événement (*mort, guerre, accident*) de ceux qui ont d'autres sens *présentation, description, repas, etc*. Il sera nécessaire de désambiguïser les noms ayant plusieurs interprétations possibles. Plusieurs sources ont contribué à la constitution de ce lexique. Une première source de noms potentiellement événementiels est le lexique VerbAction (Hathout *et al.*, 2002) qui contient 9 393 couples verbe-nom déverbal pour un total de 9 200 lemmes nominaux uniques. Pour compléter ce lexique, nous avons extrait semi-automatiquement de requêtes sur des moteurs de recherche les éléments apparaissant dans des patrons tels que *X avoir lieu, X se produire, lors de X, le/la X de...par*, ou X est susceptible d'être un nom événementiel. Une première application de cette méthode a donné 769 lemmes différents qui ne figurent pas dans VerbAction. Elle sera réutilisée pour enrichir progressivement le lexique. La taille de ce lexique (9 969 entrées) est du même ordre de grandeur que celle du lexique correspondant pour l'anglais (13 495 entrées).

Un lexique de verbes sert de base à la classification des événements sous forme verbale. Il contient les lemmes de verbes correspondant à 6 des 7 classes TimeML (plus la classe MODAL, voir section 4). La classe OCCURRENCE est la classe par défaut pour les verbes qui n'appartiennent à aucune autre classe. Il existe également une liste des verbes qui n'ont aucune interprétation événementielle au sens TimeML (la classe NON_EVENT contient des verbes tels que *durer, sembler, suffire, etc*). Ce lexique, comme celui des noms, contient des ambiguïtés dans la mesure où certains verbes peuvent appartenir à plusieurs classes ou ne doivent pas être annotés dans certains contextes. Par exemple, *expliquer* appartient à la classe REPORTING quand il introduit un fait rapporté ou du discours (*"La situation risquerait de s'aggraver", a expliqué le porte-parole*). Cependant, lorsqu'il a un sujet humain et un événement comme objet (*l'ex-capitaine a expliqué le renouvellement de l'équipe*), il doit être annoté I_ACTION. Enfin, s'il

a un sujet et un objet événementiels (*le réchauffement climatique explique la fonte des glaces*) il doit être annoté `I_STATE`. Citons également l'exemple du verbe *savoir* qui doit être annoté `I_STATE` selon son interprétation habituelle (*Max sait que la terre est ronde*), ou `OCCURRENCE` s'il est conjugué au passé composé avec une interprétation voisine d'*apprendre* (*Max a su que Léa était partie*).

Nous nous voyons confrontés au problème non trivial de la désambiguïsation lexicale pour identifier les acceptions des emplois des noms et des verbes. Dans un premier temps nous abordons ce problème pour les verbes par l'application d'heuristiques simples portant sur chacune des différentes classes de verbes TimeML. Par exemple, un verbe de la classe `ASPECTUAL` dans le lexique (*commencer, terminer, continuer*), doit être balisé `OCCURRENCE` si le premier nom à sa droite n'est pas balisé `EVENT`. Cela rend compte de la différence entre *Jean a commencé la discussion* et *Jean a commencé son livre* (où un événement sous-jacent est non spécifié). Ces heuristiques consistent en des généralisations grossières sur la structure argumentale des verbes et nous sommes conscient de leurs limites. Étant donné l'impraticabilité d'une description exhaustive des règles de désambiguïsation, nous envisageons, pour les noms et les verbes, des méthodes probabilistes de désambiguïsation lexicale.

L'algorithme actuel fonctionne par l'application de deux couches : lexicale et contextuelle. La première consiste en une recherche lexicale pour les noms - ceux qui figurent dans le lexique sont balisés `EVENT` dans le texte. Tous les verbes, à part les auxiliaires et ceux dont le lemme figure dans la classe `NON_EVENT` du lexique sont annotés `<EVENT>`. Ces verbes se voient attribuer une classe TimeML selon le lexique des verbes. Ensuite s'applique la couche d'analyse contextuelle. Pour l'instant cette couche d'analyse ne traite que les verbes. Cette étape élimine les mauvais candidats parmi les événements annotés par la couche lexicale. Des règles excluent de l'annotation certains emplois non événementiels de verbes impersonnels, comme *il s'agit, il convient de/que, il importe de/que etc*. Elle corrige également des erreurs de classification, comme dans le cas des verbes *expliquer* et *savoir* ci-dessus. Un autre ensemble de règles détermine la présence d'éléments à polarité négative (adverbes de négation, déterminants négatifs etc), valeur représentée dans l'attribut `polarity` des noms et des verbes événementiels. La classe aspectuelle (trait `aspect`) est également déterminée par un système de règles se basant sur les temps verbaux et d'autres éléments contextuels.

5.4 Annotation automatique des relations

À terme, notre objectif est d'expliciter les relations (*links*, voir 2.1) qui existent entre les entités annotés en TimeML - les événements et les expressions temporelles. L'annotation de ces relations mettra en évidence la structure temporelle du texte, notamment l'ordre d'occurrence et la durée des événements. À l'heure actuelle, nous n'avons pas commencé cette tâche, qui dépend de la résolution des tâches décrites dans les sections 5.2 et 5.3. De plus, en vue de nos travaux sur l'anaphore événementielle (Bittar, 2006), il serait intéressant d'envisager l'intégration de relations référentielles, notamment l'anaphore et la coréférence événementielles (Danlos, 2006), dans le schéma d'annotation. Nous ne proposons pas pour autant de résoudre les tâches de résolution concernées, mais plutôt de fournir un moyen d'annoter ces types de relation référentielle, ce qui n'est pas, pour l'instant, prévu par le schéma TimeML.

6 Conclusion

Nous avons présenté un projet, en cours de réalisation, pour le développement d'un ensemble cohérent de ressources pour l'annotation des événements et des expressions temporelles en français selon la norme TimeML. Nous sommes en train d'élaborer un corpus de référence de bonne qualité, annoté selon cette norme, ainsi qu'un guide d'annotation linguistiquement fondé. Nous soulignons l'importance d'avoir un guide d'annotation, qui permettra aux divers actants de la communauté de se focaliser sur la même tâche, ce qui est primordial pour assurer la pertinence d'éventuelles comparaisons de systèmes. Les modules d'annotation automatique que nous élaborons reposent à l'heure actuelle majoritairement sur des données lexicales. Le module de repérage des événements consiste aussi en l'application d'un ensemble d'heuristiques portant sur la structure syntaxique locale. Afin d'améliorer la performance de ces modules, il sera essentiel d'intégrer un plus grand nombre de règles. L'emploi de méthodes probabilistes est également à considérer pour traiter la tâche difficile de désambiguïsation lexicale.
Au fur et à mesure du développement de nos modules nous avons effectué plusieurs évaluations, mais n'avons pas pour le moment de résultats complets à présenter.

Références

ACOSTA A. & NASR A. (2007). Le projet MACAON : une architecture ouverte pour le développement d'outils de TAL. Voir http ://code.google.com/p/macaon/.

BACH E. (1986). The Algebra of Events. *Linguistics and Philosophy*, **9**(1), 5–16.

BATTISTELLI D., MINEL J.-L. & SCHWER S. R. (2006). Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. *TAL*, **47**(3), 11–37.

BITTAR A. (2006). Un algorithme pour la résolution d'anaphores événementielles. Master's thesis, Université Paris 7 Diderot, Paris, France.

CASELLI T., PRODANOF I., RUIMY N. & CALZOLARI N. (2007). Mapping SIMPLE and TimeML : Improving Event Identification and Classification Using a Semantic Lexicon. In *Prooceedings of the 4th International Workshop on Generative Approaches to Lexicon*, Paris, France.

CHENG Y., ASAHARA M. & MATSUMOTO Y. (2007). Constructing a Temporal Relation Tagged Corpus of Chinese Based on Dependency Structure. In *Proceedings of The 21st Annual Conference of the Japanese Society for Artificial Intelligence*, p. 311–315, Miyazaki, Japan.

DANLOS L. (2006). Verbes causatifs, discours causaux et coréférence événementielle. *Lynx*, p. 233–246.

GAGNON M. & LAPALME G. (1996). From Conceptual Time to Linguistic Time. *Computational Linguistics*, **22**(1), 91–127.

GROSS M. (2002). Les déterminants numéraux, un exemple : les dates horaires. *Langages*, (145), 21–38.

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

JANG S. B., BALDWIN J. & MANI I. (2004). Automatic TIMEX2 Tagging of Korean News. *ACM Transactions on Asian Language Information Processing*, **3**, 51–65.

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit 2005*.

MANI I. & WILSON G. (2000). Processing of News. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, p. 69–76.

MULLER P. & TANNIER X. (2004). Annotating and Measuring Temporal Relations in Texts. In *Proceedings of Coling 2004*, volume 1, p. 50–56, Geneva, Switzerland : Association for Computational Linguistics.

NEGRI M. & MARSEGLIA L. (2004). *Recognition and Normalization of Time Expressions : ITC-irst at TERN 2004*. Rapport interne, ITC-irst, Trento.

PUSTEJOVSKY J., CASTAÑO J., INGRIA R., SAURÍ R., GAIZAUSKAS R., SETZER A. & KATZ G. (2003). TimeML : Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics*.

PUSTEJOVSKY J., LITTMAN J. & SAURÍ R. (2006). Argument Structure in TimeML. In G. KATZ, J. PUSTEJOVSKY & F. SCHILDER, Eds., *Annotating, Extracting and Reasoning about Time and Events*, number 05151 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany : IBFI, Schloss Dagstuhl, Germany.

SAGOT B. & BOULLIER P. (2005). From Raw Corpus to Word Lattices : Robust Pre-parsing Processing. In *Proceedings of the 2nd Language & Technology Conference (LT'05)*, p. 348–351, Poznan, Poland.

SAGOT B., DANLOS L. & DÉSIR A. (2008). Traitements de surface à l'aide de graphes de transitions récursifs : de Unitex à SxPipe. In *Actes de TALN 2008*, Avignon, France. Soumis.

SAQUETE E., MUÑOZ R. & MARTÍNEZ-BARCO P. (2006). Event Ordering Using TERSEO System. *Data Knowledge Engineering*, **58**(1), 70–89.

SAURÍ R., KNIPPEN R., VERHAGEN M. & PUSTEJOVSKY J. (2005). Evita : A Robust Event Recognizer for QA Systems. In *Proceedings of HLT/EMNLP 2005*, p. 700–707.

SPREYER K. & FRANK A. (2008). Projection-based Acquisition of a Temporal Labeller. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP-2008)*, Hyderabad, India.

VERHAGEN M., MANI I., SAURÍ R., KNIPPEN R., LITTMAN J. & PUSTEJOVSKY J. (2005). Automating Temporal Annotation with TARSQI. In *Proceedings of the ACL 2005*.