Projet TourInFlux. Annotation des expressions temporelles

Lucie Drat

Projet TourInFlux
Annotation des expressions temporelles

Drat Lucie

UFR LLASIC

Mémoire de master 2 professionnel – Sciences du langage – Industrie de la langue

Parcours : Traitement automatique de la langue écrite et parlée

Sous la direction de : Agnès Tutin, Mickaël Coustaty, Alain Couillault

Année universitaire 2013-2014

UNIVERSITE Stendhal GRENOBLE 3

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : DRAT

PRENOM : Lucie

DATE : 29 Août 2014 SIGNATURE : [redacted]

Sommaire

Introduction

Ouverture du camping En principe le camping est ouvert du premier avril jusqu'au mi-novembre, mais ces dates sont surtout fonctions de la présence de campeurs sur le terrain, nous pouvons ouvrir avant et après. S'il fait beau ou sur réservation les dates d'ouverture peuvent être un peu plus amples.

Stéphanie Weiser

Bien qu'étant marginal, cet exemple extrait de la thèse de S. Weiser (Weiser, 2010) est un parfait exemple de l'expression de la temporalité dans les documents touristiques, en particulier lorsqu'ils sont issus du web.

En effet, si les expressions temporelles sont exprimées clairement en une succession de jours et d'heures d'ouvertures sur les dépliants et autres documents que l'on peut trouver en Office de Tourisme, ce n'est pas nécessairement le cas sur le web qui permet une plus grande permissivité.

Cela fait inévitablement partie des difficultés que l'on rencontre lorsque l'on souhaite représenter formellement des informations exprimées en langue naturelle.

C'est pourquoi l'étude linguistique est une étape fondamentale dans l'extraction d'information. Si l'on ne peut pas typer ce que l'on cherche, peu importent les techniques utilisées pour la détection, l'annotation et le stockage des informations, les résultats seront nécessairement incomplets et insatisfaisants.

Ce stage fait partie du projet Tourinflux qui a pour objectif d'apporter des outils aux acteurs du tourisme pour leur permettre de gérer leurs données internes ainsi que les informations disponibles sur le web, notamment pour avoir des éléments sur comment leur territoire est perçu.

En effet, le tourisme constitue un secteur important pour l'économie française et il est essentiel de connaître l'opinion des touristes sur les différents territoires de France afin de pouvoir influer sur la fréquentation touristique et rentabiliser les sommes investies en prenant des décisions efficaces.

La solution proposée est un tableau de bord complet intégrant l'ensemble des connaissances disponibles sur un territoire touristique. L'ergonomie envisagée le rendra facilement utilisable aussi bien par les spécialistes que les non-spécialistes du domaine du tourisme.

Ce projet présente plusieurs verrous dont :
- Collecte d'informations sur le web
- Traitement de grands volumes de données
- Fiabilité et normalisation des informations récoltées
- Analyse, manipulation et échange entre organismes et interfaces homme-machines correspondantes
- Adaptation à de nouveaux objets touristiques (restaurants, hôtels etc.)

Il réunit plusieurs partenaires :
- APROGED : Association de la Maîtrise et de la Valorisation des Contenus.
- L3i : Laboratoire de recherche du domaine des sciences du numérique de l'université de La Rochelle.
- Proxem : Éditeur de solutions stratégiques d'analyse de contenu pour l'entreprise.
- Syllabs : Spécialiste du traitement sémantique pour le web.

L'objectif de ce stage est de développer une grammaire d'extraction de marqueurs temporels dans le domaine du tourisme.

La plupart des informations contenues dans une base de données touristiques (évènements, manifestations, hôtels, restaurants etc.) contiennent des marqueurs temporels (date, durée, horaires d'ouverture, conditions d'ouverture etc.) qu'il s'agira d'identifier au moyen d'une grammaire hors-contexte adaptée en utilisant les outils Unitex et GramLab.

Faisant partie des entités nommées, les expressions temporelles ont fait l'objet de plusieurs travaux d'extraction et d'annotation d'informations.

Deux thèses ont été particulièrement intéressantes au vu de la mission confiée.

La première thèse est celle de S. Weiser qui s'attèle au repérage et à l'annotation d'expressions temporelles contenues dans des pages web touristiques.

Elle propose pour cela un ensemble de transducteurs développés dans Unitex qui se chargent de détecter et annoter les expressions temporelles, ainsi que les objets touristiques (restaurant, hôtel etc.) et leurs adresses, suivant un schéma d'annotation mis au point selon une ontologie du tourisme modélisé par Mondeca pour le projet Eiffel. Les informations concernant une même offre touristique sont regroupées à l'aide de transducteurs de liage.

Le schéma mis au point n'étant pas un standard, il ne correspondait pas à nos attentes et n'a donc pas été utilisé pour le projet TourinFLux.

Cependant, le travail de classification des expressions temporelles selon les différents objets touristiques est particulièrement intéressant, notamment pour la partie sur les inférences à faire [1] sur les données présentes dans le texte pour

1. Ce travail ne fait pas partie de la mission de stage et sera réalisé en aval de la grammaire dans la chaîne de traitement.

en déduire les dates et/ou horaires d'ouverture et fermeture.

En effet, elle explique que les inférences doivent être faites en fonction du type d'objet touristique. Par exemple, si les horaires d'ouverture d'un restaurant sont les suivants : *"ouvert tous les jours, sauf le soir."*, on doit en déduire qu'il est ouvert seulement le midi car un restaurant n'est pas ouvert toute la journée ou la nuit.

Cela a été pris compte lors de la création théorique de fonctions permettant de calculer la valeur exacte de certaines dates [2] (voir section 2.1).

La deuxième thèse est celle d'A. Bittar dont le but était de développer des ressources pour permettre le traitement des expressions temporelles dans des textes en français, ainsi que de construire un corpus de référence annoté selon le standard ISO-TimeML.

Un guide d'annotation ISO-TimeML a été conçu pour le français et a permis d'apporter des amélioration au standard. Un annotateur automatique à base de transducteurs a été mis au point afin de pré-annoter les textes qui ont été ensuite corrigés par des annotateurs humains pour élaborer le corpus de référence French TimeBank.

Ici, le schéma d'annotation utilisé est un standard particulièrement intéressant car il permet de décrire minutieusement les expressions temporelles et les relations qu'elles entretiennent entre elles et d'autres éléments du texte (évènements etc.). Cependant, les textes utilisés pour ce corpus étaient extraits d'articles de journaux et il n'a donc pas pu être utilisé comme référence au sein du projet TourinFlux. De plus, les graphes développés pour la détection et l'annotation étaient trop larges pour notre tâche. Ils incluaient, par exemple, les expressions temporelles relatives aux siècles.

Néanmoins, le travail effectué sur l'annotation en ISO-TimeML pour le français a permis de définir ce standard comme schéma d'annotation suite à l'étude de corpus (voir section 2.2).

Ce mémoire professionnel présente donc le travail effectué de mars à juillet 2014 au sein du laboratoire L3i et du project TourinFlux.

Dans la partie *Schéma et guide d'annotation*, nous aborderons le schéma d'annotation choisi ainsi que l'étude de corpus, réalisé en amont, que l'on décrira à l'aide du schéma.

Dans la partie *Grammaire d'extraction et d'annotation*, nous décrirons le développement de l'annotateur automatique sous forme de grammaire hors-contexte à l'aide de la présentation de la chaîne de traitements. Nous aborderons également les limites d'une grammaire hors-contexte pour notre tâche.

Dans la partie *Évaluation de la grammaire hors-contexte*, nous aborderons le processus de création du corpus de référence pour l'évaluation ainsi que les résultats obtenus par la grammaire hors-contexte.

2. La création de ces fonctions est théorique car elles ont uniquement été nommées et décrites selon les besoins, mais n'ont pas été encore développées et implémentées.

Chapitre 1

Présentation du laboratoire

Ce stage a été effectué au sein du Laboratoire Informatique, Image, Interaction (L3i) de l'université de La Rochelle. Créé en 1993, il compte environ 100 employés (chercheurs, doctorants, maîtres de conférence, professeurs etc.)

Les thématiques de recherche du laboratoire, centrées autour de la gestion interactive et intelligente des contenus numériques, sont les suivantes :

- Ingénierie des connaissances
- Analyse et gestions de contenus
- Interactivité et dynamique des systèmes

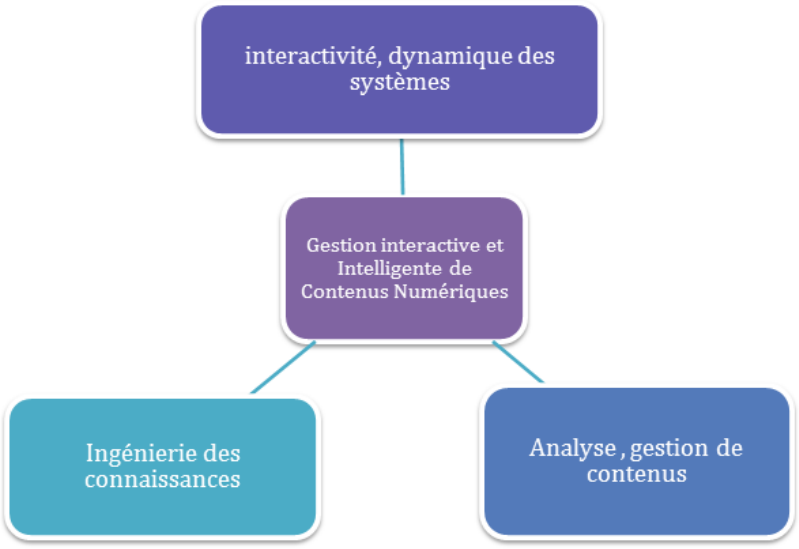


FIGURE 1.1 – Structuration du laboratoire (Université de La Rochelle, 2014)

Chapitre 2

Schéma et guide d'annotation

Sommaire

Avant de développer la grammaire, il était nécessaire de faire une étude de corpus afin de définir les types d'expressions temporelles et d'informations présents dans les documents touristiques, pour ensuite définir un schéma d'annotation adapté.

En effet, le schéma utilisé devait être complet pour représenter les phénomènes présents dans les documents, mais il fallait aussi qu'il soit clair et parfaitement structuré pour faciliter son utilisation dans le reste de la chaîne de traitement, notamment lors de transduction vers d'autres schémas ou normes.

Une part importante du stage a donc été consacrée à l'élaboration du guide d'annotation qui réunit à la fois le schéma d'annotation choisi et l'étude de corpus réalisée, et c'est sur celui-ci que s'est basé le développement de la grammaire

hors-contexte. Ce guide d'annotation a également été utilisé pour l'élaboration de schémas XML et pour une journée d'annotation (voir section 4.2).

La section 2.1 présentera le langage de balisage TimeML utilisé en tant que schéma d'annotation.
Les objets de l'annotation, définis lors de l'étude de corpus, ainsi que leurs annotations en TimeML seront décrits en section 2.2.

2.1 TimeML

TimeML est un langage de spécification permettant d'annoter les évènements et les expressions temporelles d'un document ainsi que les relations existant entre ces entités. Il a été conçu en majeure partie par le laboratoire de linguistique de l'université Brandeis lors du workshop TERQAS (Time and Event Recognition for Question Answering Systems) organisé par James Pustejovsky en 2002. L'objet de ce workshop était l'amélioration des systèmes de question-réponse en langue naturelle sur des questions relatives au temps sur des évènements ou des entités dans des articles de journaux.

Ce langage est une référence pour l'annotation d'informations temporelles et a connu plusieurs améliorations suite à son utilisation par la communauté scientifique. En Mars 2009, l'ISO-TimeML a été approuvé comme standard international.

Étant particulièrement exhaustif, nous l'avons adapté à nos besoins dans le cadre de la détection des objets touristiques et de leurs informations temporelles. C'est pour cela que seuls les éléments qui nous sont pertinents seront décrits. Pour plus de détails sur TimeML veuillez vous référer à (Saurí et al., 2006) et à (Bittar, 2010b). Des ajouts au schéma d'annotation ont aussi été effectués afin :

- de combler certains manques de TimeML pour notre tâche d'annotation de marqueurs temporels,
- d'annoter certains empans ne décrivant pas des informations temporelles mais nécessaires à la visée applicative.

2.1.1 EVENT

Selon le guide d'annotation TimeML v1.2.1 (Saurí et al., 2006), la balise EVENT regroupe les situations ayant lieu et les états ou circonstances dans lesquels un évènement est ou devient vrai.

- **Attribut class** : définition du type d'évènement.

  OCCURENCE : situation ayant lieu

  STATE : circonstance selon laquelle un élément est ou devient vrai

- **Attribut eid** : identifiant unique de l'évènement, formé de la lettre "e" suivie d'un ou plusieurs chiffres.
- **Attribut eiid** : identifiant unique de la réalisation de l'évènement (fusion des étiquettes EVENT et MAKEINSTANCE dans le format ISO-TimeML), formé de la lettre "ei" suivie d'un ou plusieurs chiffres : un événement du type **eid** a une occurrence particulière **eiid** .

L'exemple ci-dessous présente l'annotation d'un horaire d'ouverture [1].

Ouverture est annoté en tant que situation ayant lieu à l'aide de la balise EVENT et de l'attribut class. L'expression temporelle qui se rattache à cette ouverture est *24/24* qui indique une durée d'une journée exprimée dans l'attribut value de la balise TIMEX3 (section 2.1.2). Le rattachement de ces deux éléments est fait à l'aide de leur ID dans la balise TLINK (section 2.1.4) qui permet également expliciter leur type de relation.

```
<EVENT eid="e1" eiid="ei1" class="OCCURENCE">
Ouverture
</EVENT>
<TIMEX3 tid="t1" type="DURATION" value="P1D">
24/24
</TIMEX3>
<TLINK lid="l1" relType="IDENTITY" eventInstanceID="ei1"
relatedToTime="t1"/>
```

2.1.2 TIMEX3

Les balises TIMEX3 servent à annoter les expressions temporelles.

Attributs obligatoires

- **tid** : identifiant unique, formé de la lettre "t" suivie d'un ou plusieurs chiffres.
- **type** : type d'expression temporelle annotée. Il y a quatre types de TIMEX3 :

  DATE : date calendaire

  DURATION : durée

  SET : ensemble de dates ou d'horaires

  TIME : horaire

- **value** : valeur normalisée au standard ISO-8601 [2]. Lorsqu'un élément de la valeur est inconnu, il doit être transcrit à l'aide de "X".

1. Il s'agit ici de l'horaire d'ouverture d'un hôtel. L'information a été oblitérée pour simplifier l'exemple. L'annotation des objets touristiques tels que les hôtels sont présentés ultérieurement.
2. Correspondant au format : YYYY-MM-DD

DATE	Exemple	value
Année	2014	2014
Semestre	le deuxième semestre	XXXX-H2
Trimestre	le troisième trimestre	XXXX-Q3
Quart	le troisième quart	XXXX-Qu3
Saison	automne hiver printemps été	XXXX-FA XXXX-WI XXXX-SP XXXX-SU
Mois et année	avril 2014	2014-04
Mois	avril	XXXX-04
Semaine	la semaine	XXXX-WXX *
Weekend	le week-end	XXXX-WXX-WE
Date complète	24/04/2014 (mardi) 24 Avril 2014	2014-04-24 2014-04-24
Date incomplète	24 avril	XXXX-04-24
Jour de la semaine	mardi	XXXX-WXX-2

TABLE 2.1 – TIMEX3 DATE value

* La valeur de *WXX* correspond au numéro de la semaine de l'année. Il peut varier selon les années, donc, sauf mention claire ("12ème semaine de 2014"), elle doit être calculée par une temporalFunction voir section 2.1.2

DURATION	Exemple	value
t $\geq$ jour	un an	P1Y
	un an et demi	P1.5Y
	deux mois	P2M
	trois semaines	P3W
	quatre jours	P4D
t<jour	deux heures	PT2H
	une heure et demie	PT1.5H
	une heure trente	PT1H30
	1h30	PT1H30
	trois minutes	PT3M
	quatre secondes	PT4S

TABLE 2.2 – TIMEX3 DURATION value

SET	Exemple	value	quant
"Tous"	tous les ans	P1Y	EVERY
	tous les jours	P1D	EVERY
	tous les mardis	XXXX-WXX-2	EVERY
"Chaque"	chaque jour	P1D	EVERY
	chaque vendredi	XXXX-WXX-5	EVERY
"Certain"	certaines semaines	P1W	SOME
	certains dimanches	XXXX-WXX-7	SOME

TABLE 2.3 – TIMEX3 SET value

TIME	Exemple	value
Heures	15h	T15 :00
	15h30	T15 :30
	trois heures et demie de l'après-midi	T15 :30
Périodes	la journée	TDT
	le matin	TMO
	le midi	TMI
	l'après-midi	TAF
	le soir	TEV
	la nuit	TNI

TABLE 2.4 – TIMEX3 TIME value

Exemple :

```
le
<TIMEX3 tid="t1" type="DATE" value="XXXX-04-28"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
28 Avril
</TIMEX3>
```

"t0" correspond à une étiquette TIMEX3 de référence (functionInDocument) pour les calculs effectués par les temporalFunction (voir section 2.1.2).

Attributs facultatifs

- **functionInDocument** : sert à obtenir un point de référence, s'il est absent, pour le calcul d'autres expressions temporelles.

  CREATION_TIME : création du document

  MODIFICATION_TIME : dernière modification

  PUBLICATION_TIME : publication

  RELEASE_TIME : autorisation de publication

  RECEPTION_TIME : réception du document par un lecteur
- **beginPoint et endPoint** : seulement présent(s) dans un TIMEX3 de type DURATION. Ils correspondent respectivement aux *tid*s du début et de la fin de l'intervalle représenté par une durée. Il peut n'y avoir qu'un seul des deux attributs de présent.
- **freq** et **quant** : seulement présent(s) dans un TIMEX3 de type SET. Ils correspondent respectivement à la fréquence et au quantifieur modifiant l'expression temporelle.

– **mod** : sert à capter une modification ne pouvant pas être exprimée dans l'attribut value ou à l'aide d'un lien ou d'une fonction.

	Exemple	mod
Point	il y a plus d'un an	BEFORE
	il y a moins d'un an	AFTER
	il n'y a pas plus d'un an	ON_OR_BEFORE
	il n'y a pas moins d'un an	ON_OR_AFTER
Durée	dans moins de deux heures	LESS_THAN
	dans plus de deux heures	MORE_THAN
	dans pas plus de deux heures	EQUAL_OR_LESS
	dans au moins de deux heures	EQUAL_OR_MORE
Les deux	en début de semaine	START
	mi-novembre	MID
	fin-août	END
	environ une heure	APPROX

TABLE 2.5 – TIMEX3 mod

– **temporalFunction** : sert à spécifier qu'une fonction temporelle doit être utilisée sur cette TIMEX3 afin de calculer sa valeur, pour cela il faut lui donner la valeur "*true*".
– **valueFromFunction** : correspond au nom de la fonction temporelle à utiliser. L'étude linguistique a mis en avant des expressions temporelles pouvant s'avérer particulièrement complexes lors de la représentation formelle leur valeur avec le standard ISO-8601 [3]. Nous avons donc créé les valeurs suivantes pour les *temporalFunction*, qui ne sont pas décrites dans TimeML, afin d'expliciter le travail de complétion des valeurs à effectuer :

  tf1 : calcul d'une date qui ne peut avoir qu'une seule valeur possible ("*à la Pentecôte*", "*1er samedi du mois de juillet 2013*")

  tf2 : calcul d'une date qui a plusieurs valeurs possibles ("*pendant les jours fériés*")

  tf3 : calcul du point manquant et de la durée d'un intervalle avec une seule borne (voir section 2.2.4) et n'ayant qu'une valeur ("*jusqu'à 22h*")

  tf4 : calcul du début/fin et de la durée pour un intervalle sans borne (voir section 2.2.6) et qui ne peut avoir qu'une valeur ("*haute saison*")

  tf5 : calcul des débuts/fins et des durées pour un intervalle sans borne et ayant plusieurs valeurs ("*aux heures des repas*")

  tf6 : calcul d'une date exprimée par l'ancrage d'une durée à une expression temporelle (voir section 2.2.3) ("*deux jours avant Noël*")

– **anchorTimeID** : permet d'indiquer le *tid* à laquelle la fonction temporelle doit se référer pour effectuer le calcul, si ce point de référence existe.

3. Rappel : ce format correspond à YYYY-MM-DD

Exemple :

```
<TIMEX3 tid="t0" type="DATE" value="2013-12">
Décembre 2013
</TIMEX3>
:
<TIMEX3 tid="t1" type="DATE" value="XXXX-12-24"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
24 Décembre
</TIMEX3>
```

2.1.3 SIGNAL

Cette balise sert à marquer les éléments indiquant une relation entre deux EVENT, ou deux TIMEX3, ou un EVENT et une TIMEX3.

- **Attribut sid** : identifiant unique, formé de la lettre "s" suivie d'un ou plusieurs chiffres.

Exemple :

```
<TIMEX3 tid="t1" type="DURATION" value="P2D"
temporalFunction="true" valueFromFunction="tf6">
deux jours
</TIMEX3>
<SIGNAL sid="s1">
avant
</SIGNAL>
<TIMEX3 tid="t2" type="DATE" value="XXXX-12-24"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
Noël
</TIMEX3>
<TLINK lid="l1" relType="IBEFORE" timeID="t1"
relatedToTime="t2"/>
```

2.1.4 SLINK et TLINK

Nous utiliserons deux types de liens dans notre tâche d'annotation : les SLINK et les TLINK.

Le standard TimeML ne permet pas l'imbrication de balises, l'annotation se fait donc "à plat". Il est donc impossible d'encadrer deux éléments à relier au sein d'une même balise qui représenterait leur relation.

En TimeML, les liens sont annotés à l'aide de balises auto-fermantes référençant les balises à relier à l'aide de leur identifiant unique. Ces balises devront être placées à la suite du dernier élément faisant partie du lien.

SLINK

Ce lien permet de créer une subordination entre deux EVENT afin d'exprimer une condition sous la forme : *évènement A "si" évènement B.*

- **Attribut lid** : identifiant unique du lien, formé de la lettre "l" suivie d'un ou plusieurs chiffres.
- **Attribut relType** : il y a plusieurs valeurs possibles mais nous n'utiliserons pour notre annotation que la valeur "*CONDITIONAL*".
- **Attribut eventInstanceID** : contient l'*eid* de l'évènement principal (*évènement A*).
- **Attribut subordinatedEventInstance** : contient l'*eid* de l'évènement subordonné (*évènement B*).
- **Attribut signalID** : correspond à le *sid* de l'éventuel élément marquant le lien.

Exemple :

```
<OI oid="o1" type="ASC" eid="e1" eiid="ei1">
Visite
</OI>
sur
<EVENT eid="e2" eiid="ei2" class="STATE>
rendez-vous
</EVENT>
<SLINK lid="l1" relType="CONDITIONAL" eventInstanceID="ei1"
subordinatedEventInstanceID="ei2"/>
```

Etiquette OI non-incluse dans le standard. Voir section 2.2.8.

TLINK

Les TLINK sont les liens entre deux EVENT, ou deux TIMEX3, ou un EVENT et un TIMEX3.

- **Attribut lid** : identifiant unique, formé de la lettre "l" suivie d'un ou plusieurs chiffres .
- **Attribut relType** : il existe différents type de liens :

TimeML TLINK relType	Relation	Représentation graphique
BEGINS	A commence B	
BEGUN_BY	B commence par A	
ENDS	A finit B	
ENDED_BY	B finit par A	
BEFORE	A est avant B	
IBEFORE	A est juste avant B	
AFTER	A est après B	
IAFTER	A est juste après B	
SIMULTANEOUS	A et B sont simultanés	
IDENTITY	A et B sont simultanés où A et B font référence au même évènement	
INCLUDES	A inclus B	
IS_INCLUDED	B est inclus dans A	

TABLE 2.6 – TLINK relType

Différences par rapport au standard

Les phénomènes que nous avons à annoter nous ont conduit à utiliser des schémas légèrement différents du standard :

- "*DURING*" : Pour cette campagne d'annotation, nous n'utiliserons pas le type "*DURING*" car il semble être redondant avec le type "*IS_INCLUDED*". De plus, du fait de sa redondance avec un autre type de relation, son utilisation est litigieuse et incohérente au sein même de la communauté scientifique travaillant sur TimeML (Bittar, 2010a, p. 78).
- *IS_EXCLUDED* : Des phrases comme "*ouvert tous les jours sauf le samedi*", dans lesquelles une période est décrite comme étant exclue d'une période plus grande sont fréquentes dans les corpus et ne sont pas couvertes par le standard. Nous ajoutons donc le type *IS_EXCLUDED* lorsque la période A, qui, par défaut, est incluse dans B, en est explicitement exclue.

  - *XOR* : Plusieurs exemples comme "*le matin ou l'après-midi*", décrivent qu'un événement a lieu à une date ou une autre de façon mutuellement exclusive : si l'événement a lieu à la date A, il n'aura pas lieu à la date B. Pour décrire ces exemples, il est donc nécessaire d'introduire une disjonction exclusive.
  - *AROUND* : Dans certains cas, il n'est pas possible de savoir si un évènement a lieu avant ou après une certaine date, par exemple : *le dimanche le plus proche de Noël*. Il faut donc pouvoir modéliser cette incertitude pour le calcul ultérieur de la date, c'est pour cela que nous avons rajouté ce type de lien.

- **Attributs eventInstanceID** et **timeID** : contient l'ID de l'évènement (*eid*) ou de l'expression temporelle (*tid*) principal. Il ne peut y avoir que l'un des deux.
- **Attributs relatedToEventInstance** et **relatedToTime** : contient l'ID de l'évènement (*eid*) ou de l'expression temporelle (*tid*) auquel l'évènement ou l'expression temporelle principal est relié. Il ne peut y avoir que l'un des deux.
- **Attribut signalID** : correspond à le *sid* de l'éventuel élément marquant le lien.

Exemple :

```
<TIMEX3 tid="t1" type="DURATION" value="PT2H">
Deux heures
</TIMEX3>
<SIGNAL sid="s1">
avant
</SIGNAL>
<TIMEX3 tid="t2" type="TIME" value="TMI">
midi
</TIMEX3>
<TLINK lid="l1" relType="IBEFORE" timeID="t1" relatedToTime="t2"
sid="s1"/>
```

2.2 Étude de corpus et annotations

2.2.1 Expressions temporelles

Cette section présente les principaux types d'informations temporelles recensés dans les corpus utilisés pour rédiger le présent guide d'annotation.

Ces corpus sont constitués de pages web récoltées par les partenaires du projet ainsi que de données ouvertes mises à disposition par le gouvernement.

Pour chaque expression est indiquée son annotation en TimeML.

Éléments atomiques

– Années

20/03/**2014**

20/03/**14**

```
<TIMEX3 tid="t1" type="DATE" value="2014-03-20">
20/03/14
</TIMEX3>
```

L'abréviation des années peut causer des ambiguïtés si l'expression temporelle est dépourvue de contexte. Dans le cas ci-dessous(1.), il est impossible de déterminer si la date indique le 1er décembre ou janvier 2012. En (2.), l'information sur le siècle est supprimée, la date peut alors désigner l'année 2018 ou l'année 1918 [4].

1. 01/12
2. 20/03/**18**

Dans le cas de l'annotation manuelle, cette abréviation n'est pas problématique car l'annotateur humain dispose nécessairement du contexte et peut donc faire un choix. Ce qui n'est pas le cas pour l'annotateur automatique (c'est-à-dire la grammaire), ce problème a donc été pris compte lors de son développement.

– Saisons, trimestres, semestres etc.

pendant **l'été**

```
pendant l'
<TIMEX3 tid="t1" type="DATE" value="XXXX-SU"
temporalFuntion="true" valueFromFunction="tf1
anchorTimeID="t0">
été
</TIMEX3>
```

le troisième **trimestre** de l'année

```
le
<TIMEX3 tid="t1" type="DATE" value="XXXX-Q3"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
troisième trimestre de l'année
</TIMEX3>
```

– Mois

4. Il peut sembler plus probable que l'abréviation désigne l'année du siècle en cours, mais les références à des évènements historiques ne sont pas rares dans les documents touristiques

20/**03**/2014

20 **mars** 2014

1er **juil.** 2014

15 **avr** 2014

```
<TIMEX3 tid="t1" type="DATE" value="2014-04-15">
15 avr 2014
</TIMEX3>
```

- Semaine et week-end

  en **semaine** sur RDV

```
en
<TIMEX3 tid="t1" type="DATE" value="XXXX-XX-XX"
temporalFuntion="true" valueFromFunction="tf4"
anchorTimeID="t0">
semaine
</TIMEX3>
sur
<EVENT eid="e1" eiid="ei1" class="STATE">
RDV
</EVENT>
<TLINK lid="l1" relType="IDENTITY" eventInstanceID="ei1"
relatedToTime="t1"/>
```

Ici, *"en semaine"* fait référence à tous les jours de la semaine sauf le week-end. Donc, les début et fin ainsi que la durée de l'intervalle doivent être calculés à l'aide d'une *temporalFunction* (section 2.1.2).

le 1er **weekend** de juillet

```
le
<TIMEX3 tid="t1" type="DATE" value="XXXX-WXX-WE"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
1er weekend de juillet
</TIMEX3>
```

- Dates des jours
  - Ordinaux pour les premiers du mois

    Le mardi **premier** octobre

    Le **1er** avril
  - Cardinaux pour les autres jours du mois

    Du **6** juillet au **10** août 2013

– Jours de la semaine
Ils sont annotés uniquement lorsqu'ils sont seuls dans les expressions temporelles. S'ils sont accompagnés d'une date du jour (*"jeudi 21"*), l'annotation portera sur cette date et non sur le jour de la semaine, considérée comme une information complémentaire sur la date.

mardi et **vendredi**

mer. et **jeu.**

lun et **mar**

```
<TIMEX3 tid="t1" type="DATE" value="XXXX-WXX-1"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
lun
</TIMEX3>
et
<TIMEX3 tid="t2" type="DATE" value="XXXX-WXX-2"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
mar
</TIMEX3>
```

Tout comme l'abréviation des années, l'abréviation des jours de la semaine peut entraîner des ambiguïtés pour la grammaire, en particulier lorsque les règles d'abréviations ne sont pas nécessairement respectées. L'abréviation *"jeu"* pour jeudi sera en temps normal considéré par la grammaire comme le nom faisant référence au divertissement.

Date précise

Une date précise est un point qui peut être positionné sur un calendrier car les éléments atomiques apportent suffisamment d'informations, bien qu'ils ne soient pas tous obligatoirement présents.

le **30/03/2014**

Septembre 2013

le **24 avril**

Date floue

Une date floue est un point imprécis faisant référence au calendrier. Les fêtes mobiles, dont la date change selon les années, sont un bon exemple.

de **mi-juin** à **fin-août**

```
de
<TIMEX3 tid="t1" type="DATE" value="XXXX-06" mod="MID"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
mi-juin
</TIMEX3>
à
<TIMEX3 tid="t2" type="DATE" value="XXXX-08" mod="END"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
fin août
</TIMEX3>
<TIMEX3 tid="t3" type="DURATION" value="P2.5M" beginPoint="t1"
endPoint="t2"/>
```

le **1er samedi du mois de juillet 2013**

```
le
<TIMEX3 tid="t1" type="DATE" value="2013-07-XX"
temporalFuntion="true" valueFromFunction="tf1">
1er samedi du mois de juillet 2013
</TIMEX3>
```

deux semaines avant **Pâques**

à la **Pentecôte**

```
à la
<TIMEX3 tid="t1" type="DATE" value="XXXX-XX-XX"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
Pentecôte
</TIMEX3>
```

Date ancrée

C'est une date liée à une autre date par un marqueur tel que : avant, après, etc. (Saurí et al., 2010, pp. 4-5 et 20-21)

dimanche le plus près du 25 octobre

le **dimanche** précédant le premier dimanche de l' Avent

```
le
<TIMEX3 tid="t1" type="DATE" value="XXXX-WXX-7"
temporalFuntion="true" valueFromFunction="tf1"
anchorTimeID="t0">
dimanche
</TIMEX3>
<SIGNAL sid="s1">
précédant
</SIGNAL>
le
<TIMEX3 tid="t2" type="DATE" value="XXXX-XX-XX"
temporalFuntion="true" valueFromFunction="tf1">
premier dimanche de l'Avent
</TIMEX3>
<TLINK lid="l1" relType="BEFORE" timeID="t1" relatedToTime="t2"
sid="s1"/>
```

2.2.2 Horaire

Les horaires sont construits par concaténation d'un nombre pour l'heure, un séparateur (généralement la lettre h), et, éventuellement, un nombre pour les minutes :

à **20h**

14 :00-17 :30

```
à
<TIMEX3 tid="t1" type="DATE" value="T17:30">
17h30
</TIMEX3>
```

Une information sur la période de la journée peut être présente (*"6h du soir"*). Elle peut être complémentaire ou bien essentielle lorsqu'elle modifie la valeur de l'horaire. Elle doit donc être conservée lors de l'annotation pour le calcul et la bonne conversion de l'horaire pour l'attribut *value*. Par exemple, *"6h du soir"* doit être normalisé en *"T18 :00"* et non *"T06 :00"*. Certains horaires sont lexicalisés, par exemple :

De 7h30 à **midi**

minuit

```
à
<TIMEX3 tid="t1" type="DATE" value="T00:00">
minuit
</TIMEX3>
```

Les horaires et les périodes (section 2.2.6) ont été décrites de manière à les différencier. En effet, celles-ci peuvent paraître identiques mais leur valeur sera différente. Par exemple, "à midi" est un horaire et "le midi" est une période. Leurs annotation respectives seront donc les suivantes :

```
à
<TIMEX3 tid="t1" type="TIME" value="T12:00">
midi
</TIMEX3>
```

```
le
<TIMEX3 tid="t1" type="TIME" value="TMI">
midi
</TIMEX3>
```

2.2.3 Durée

Une durée s'exprime par un ordinal ou un cardinal suivi d'une unité de temps (jour, semaine etc.) ou d'une quantité (dizaine, quinzaine etc.) (Saurí et al., 2010, pp. 4-5 et 20-21).
Cette durée peut permettre d'exprimer une date en s'ancrant à une autre expression temporelle, la date implicite doit être calculée à l'aide d'une *temporalFunction*.

deux jours avant Noël

1ère quinzaine de Juillet

```
<TIMEX3 tid="t1" type="DURATION" value="P2W"
temporalFunction="true" valueFromFunction="tf6"
anchorTimeID="t2">
1ère quinzaine
</TIMEX3>
de
<TIMEX3 tid="t2" type="DATE" value="XXXX-07"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
Juillet
</TIMEX3>
```

7/7 jours

24h/24

```
<TIMEX3 tid="t1" type="DURATION" value="P1D">
24h/24
</TIMEX3>
```

2.2.4 Intervalle

Un intervalle s'exprime généralement en associant un marqueur de borne à une date ou un horaire.

Bornée à gauche et à droite

Son annotation se fait à l'aide de deux TIMEX3 *DATE* ou *TIME* représentant la date ou l'heure de début et de fin de l'intervalle et d'une *TIMEX3 DURATION* auto-fermante contenant en attribut *value* la durée entre ces deux points et en attribut *beginPoint* et *endPoint*, les IDs de deux TIMEX3 formant l'intervalle.

du 25/12 **au** 03/01/2014

Du lundi **au** vendredi

Du 6 **au** 17 juillet

de mai **à** octobre

de 9h **à** 18h

```
de <TIMEX3 tid="t1" type="TIME" value="T09:00">
9h
</TIMEX3>
à
<TIMEX3 tid="t2" type="TIME" value="T18:00">
18h
</TIMEX3>
<TIMEX3 tid="t3" type="DURATION" value="P7H" beginPoint="t1"
endPoint="t2"/>
```

Bornée uniquement à gauche ou à droite

L'annotation s'effectue comme pour un intervalle borné à gauche et à droite, sauf qu'il n'y aura qu'une TIMEX3 *DATE* ou *DURATION*, donc l'attribut *value* de la TIMEX3 *DURATION* sera inconnu et il manquera un des attributs *beginPoint* ou *endPoint*.

à partir du 25/03/14

jusqu'au 30 mai 2014

à partir de 19 h le vendredi et **à partir de** 15 h le samed

dès 20h

jusqu'à 22h

```
jusqu'à
<TIMEX3 tid="t1" type="TIME" value="T22:00">
22h
</TIMEX3>
<TIMEX3 tid="t2" type="DURATION" value="PXH" endPoint="t1"
temporalFunction="true" valueFromFunction="tf3"/>
```

2.2.5 Ellipses

Il est fréquent que des informations temporelles soient supprimées par des ellipses, principalement afin d'éviter la redondance.

Dans ce cas, les expressions temporelles avec des informations manquantes devront faire appel à une *temporalFunction* en ayant pour ancre l'ID de l'expression temporelle ayant gardé les informations.

18 et 19 Mai 2013 (= 18 mai 2013 et 19 mai 2013)

Du 9 au 11 décembre et le 21 janvier 2014 (= Du 9 décembre 2013 au 11 décembre 2013 et le 21 janvier 2014)

Lun au ven 8h 12h30 14h 18h (= Lun au ven 8h-12h30 et 14h-18h)

Du 6 au 17 juillet (= Du 6 juillet au 17 juillet)

```
du
<TIMEX3 tid="t1" type="DATE" value="XXXX-XX-06"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t2">
6
</TIMEX3>
au
<TIMEX3 tid="t2" type="DATE" value="XXXX-07-17"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
17 juillet
</TIMEX3>
<TIMEX3 tid="t2" type="DURATION" value="P11D" beginPoint="t1"
endPoint="t2"/>
```

2.2.6 Période

Une période est un intervalle dont le début et la fin ne sont pas précis. Dans les cas où l'on ne peut pas représenter la valeur de la période dans l'attribut *value*, il faut faire appel à une *temporalFunction* qui s'occupera de calculer cette valeur.

pendant les **vacances scolaires**

haute saison : 8h-20h, **basse saison** : 10h-12h et 14h-16h

Fermé le dimanche **midi**

```
<EVENT eid="e1" eiid="ei1" class="OCCURRENCE">
Fermé
</EVENT>
le
<TIMEX3 tid="t1" type="DATE" value="XXXX-WXX-7"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
dimanche
</TIMEX3>
<TLINK lid="l1" relType="IDENTITY" eventInstanceID="ei1"
```

```
relatedToTime="t1"/>
<TIMEX3 tid="t2" type="TIME" value="TMI">
midi
</TIMEX3>
<TLINK lid="l2" relType="IS_INCLUDED" timeID="t2"
relatedToTime="t1"/>
```

2.2.7 Ensemble

Un ensemble est une expression temporelle regroupant à elle seule plusieurs dates ou horaires, précis ou non.

Tous les jours

tous les 1er mai

```
tous les
<TIMEX3 tid="t1" type="SET" value="XXXX-05-01" quant="EVERY"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
1er mai
</TIMEX3>
```

2.2.8 Informations non-temporelles

Objets et évènements touristiques

Les informations temporelles contenues dans les corpus touristiques se réfèrent toutes à des *Objets touristiques* au sens de la norme TourinFrance.

La norme TourinFrance a été initiée par le ministère du Tourisme afin de faciliter l'échange d'un maximum de données touristiques sur les objets touristiques (offres d'hébergements, activités culturelles, festivals etc.) telles que l'adresse, les prestations et tarifs etc.

Ces objets touristiques peuvent être considérés comme des événements ayant lieu, auxquels peuvent se rattacher des dates et horaires d'ouverture/fermeture ainsi que des modalités.

Seuls les objets touristiques avec une ou plusieurs expressions temporelles nous intéressent, donc s'ils n'en ont pas, il ne faut pas les annoter. Cependant, cette distinction ne relevait pas de ma mission de stage qui se concentrait sur l'annotation des expressions temporelles. Cette tâche a donc été déléguée aux partenaires se chargeant de la récupération des pages web touristiques.

Nous avons choisi de créer une étiquette **OI** spécifique pour marquer l'empan correspondant à un objet touristique. Celle-ci possèdera un *eid* et *eiid* comme pour un EVENT et définira le type d'évènement selon la norme TourinFrance.

La BNF [5] pour cette étiquette **OI** est :

```
attributes ::= oid type
oid ::= o<integer>
type ::= 'ASC' | 'DEG' | 'FMA' | 'HLO' | 'HOT' | 'HPA' | 'ITI'
| 'LOI' | 'MUL' | 'ORG' | 'PCU' | 'PNA' | 'RES' | 'VIL'
eid ::= ei<integer>
eiid ::= ei<integer>
stem ::= CDATA
```

Les différents types d'objets touristiques sont les suivants :

ASC : Activités sportives / culturelles / séjour itinérants

DEG : Dégustation (tous produits)

FMA : Fêtes et Manifestations

HLO : Hébergement locatif (Gites et chambres d'hotes)

HOT : Hotellerie

HPA : Hôtellerie de plein air (camping)

ITI : Itinéraires touristiques

LOI : Activités et équipements de loisirs

MUL : Multimedia

ORG : Organismes

PCU : Patrimoine culturel

PNA : Patrimoine naturel

RES : Restauration

VIL : Villages de vacances

```
<TIMEX3 tid="t1" type="DATE" value="2011-07">
Juillet 2011
</TIMEX3>
:
<TIMEX3 tid="t2" type="DATE" value="XXXX-07-02"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t1">
2 juillet
</TIMEX3>
```

5. *"La forme de Backus-Naur (souvent abrégée en BNF, de l'anglais Backus-Naur Form) est une notation permettant de décrire les règles syntaxiques des langages de programmation."* - Wikipédia

```
: ESTISSAC -
<OI oid="o1" type="FMA" eid="e1" eiid="ei1">
Concert
</OI>
<TLINK lid="l1" relType="IDENTITY" eventInstanceID="ei1"
relatedToTime="t2"/>
de l'atelier rock à
<TIMEX3 tid="t3" type="TIME" value="T20:30">
20 h 30
</TIMEX3>
<TLINK lid="l2" relType="IS_INCLUDED" timeID="t3"
relatedToTime="t2">
à la salle des fêtes d'Estissac
```

Lors de l'annotation des évènements touristiques à l'aide de l'étiquette OI, il convient d'être précis sans être spécifique. Les noms propres ne doivent donc pas être annotés, sauf s'ils constituent l'unique information sur l'évènement (exemple 5 et 6)

1. **Réveillon du Nouvel An**
2. **Course pédestre de haut niveau**
3. **Représentation de la Chorale** "Les Fa sans Dièse"
4. **Exposition** Antoni TAPIES
5. **Téléthon** au foyer familial organisé par le CCAS.
6. Le 18 juillet : **Festi'Coccinelle**

Dans le cas des énumérations d'évènements, s'ils ont tous lieu au même endroit et en même temps, il faut alors les annoter comme un seul évènement.

Vente de plants à repiquer (fleurs et légumes), marché du terroir, animations...dans le Centre Ville d'Ervy-le-Châtel tous les 8 mai

Informations complémentaires

Des éléments lexicalisés donnent des informations complémentaires sur les évènements touristiques, permettant ainsi d'identifier si les expressions temporelles indiquent une ouverture, un rendez-vous, un départ etc.

Ces éléments sont considérés comme des évènements ayant lieu et doivent être annoté en tant que tels.

Lundi à samedi : 9h - 12h et 13h30 - 18h sauf **fermeture** lundi après midi et mercredi toute la journée

Fermeture le mercredi. **Ouverture** de 12h à 16h

Fermé en juillet et août

Dimanche 9 février : Randonnée pédestre. **Rendez-vous** au lavoir à 8 h 45 pour un **départ** à 9 h.

```
<TIMEX3 tid="t1" type="DATE" value="XXXX-02-09"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
Dimanche 9 Février
</TIMEX3>
:
<OI oid="o1" type="ASC" eid="e1" eiid="ei1">
Randonnée pédestre
</OI>.
<TLINK lid="l1" relType="IDENTITY" eventInstanceID="ei1"
relatedToTime="t1"/>
<EVENT eid="e2" eiid="ei2" class="OCCURENCE">
Rendez-vous
</EVENT>
au lavoir à
<TIMEX3 tid="t2" type="TIME" value="T08:45">
8 h 45
</TIMEX3>
<TLINK lid="l2" relType="IDENTITY" eventInstanceID="ei2"
relatedToTime="t2"/>
<TLINK lid="l4" relType="IS_INCLUDED" timeID="t2"
relatedToTime="t1">
pour un
<EVENT eid="e3" eiid="ei3" class="OCCURENCE">
départ
</EVENT>
à
<TIMEX3 tid="t3" type="TIME" value="T09:00">
9 h
</TIMEX3>.
<TLINK lid="l3" relType="IDENTITY" eventInstanceID="ei3"
relatedToTime="t3"/>
<TLINK lid="l5" relType="IS_INCLUDED" timeID="t3"
relatedToTime="t1">
```

Modalités

Les modalités précisent les conditions auxquelles une expression temporelle doit répondre pour être valable. A ce titre, elles doivent être considérées comme

des états ou circonstances dans lesquels un évènement est vrai et être annotées à l'aide d'EVENT.

Visites **sur rendez-vous**

Ouvert toute l'année **sur simple appel téléphonique** aux heures des repas

repas pour groupes **sur réservation** le week-end

```
<EVENT eid="e1" eiid="ei1" class="OCCURENCE">
repas pour groupes
</EVENT>
sur
<EVENT eid="e2" eiid="ei2" class="STATE">
réservation
</EVENT>
<SLINK lid="l1" relType="CONDITIONAL" eventInstanceID="ei1"
subordinatedEventInstanceID="ei2"/>
le
<TIMEX3 tid="t1" type="DATE" value="XXXX-WXX-WE"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
week-end
</TIMEX3>
<TLINK lid="l2" relType="IDENTITY" eventInstanceID="ei2"
relatedToTime="t1"/>
```

2.2.9 Combinatoire

La combinatoire décrite ici est celle exposée dans (Teissèdre, 2012). Elle permet d'associer plusieurs expressions temporelles afin d'en créer de nouvelles. En TimeML, la combinatoire se fait à l'aide des liens SLINK (section 2.1.4) et TLINK (section 2.1.4).

Fonction de spécification

La fonction de spécification indique qu'une expression temporelle en précise une autre, avec pour **opérant** l'expression qui spécifie et pour opérande celle à laquelle s'applique la spécification.

le matin **à 9h**

le 26 mars **à 15h45**

Ouvert toute l'année **sur simple appel téléphonique aux heures des repas**

Lien de concaténation

Le lien de concaténation permet de joindre des expressions temporelles.

de 9h à 12h **et** de 14h à 16h

Ouvert en mars **et** en septembre

Ce type de fonction combinatoire n'existe pas en TimeML, il suffit d'annoter séparément les expressions temporelles concaténées.

Lien de disjonction exclusive

Le lien de disjonction exclusive indique une indécision entre plusieurs informations temporelles sur un objet touristique. Il s'effectue grâce au TLINK *relType XOR* que nous avons rajouté au stantard (section 2.1.4).

Dimanche 19 **ou** 26 janvier

En février **ou** en mars

```
En
<TIMEX3 tid="t1" type="DATE" value="XXXX-02"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
février
</TIMEX3>
ou en
<TIMEX3 tid="t2" type="DATE" value="XXXX-03"
temporalFunction="true" valueFromFunction="tf1"
anchorTimeID="t0">
mars
</TIMEX3>
<TLINK lid="l1" relType="XOR" timeID="t1" relatedToTime="t2"/>
```

Rôle d'exception

Le rôle d'exception indique qu'une **expression temporelle** échappe à l'expression temporelle générale. Il n'y a pas obligatoirement un marqueur tel que *sauf*, *hors* etc. pour indiquer une exception, comme dans le troisième exemple.

Lundi à samedi : 9h - 12h et 13h30 - 18h sauf fermeture **lundi après midi et mercredi toute la journée**

Ouvert tous les jours de 7h30 à 23h, **les mardi et vendredi de 9h30 à 23h**

toute l'année, hors **périodes de vacances scolaires**

La notion d'exception étant absente du standard, nous avons rajouté un *relType IS_ EXCLUDED* (section 2.1.4).

```
<TIMEX3 tid="t1" type="DURATION" value="P1Y">
toute l'année
</TIMEX3>
, hors périodes de
<TIMEX3 tid="t2" type="DATE" value="XXXX-XX-XX"
temporalFunction="true" valueFromFunction="tf5"
anchorTimeID="t0">
vacances scolaires
</TIMEX3>
<TLINK lid="l1" relType="IS_EXCLUDED" timeID="t2"
relatedToTime="t1"/>
```

2.2.10 A exclure de l'annotation

Lors de l'étude de corpus, plusieurs expressions temporelles n'étant pas pertinentes pour notre tâche étaient présentes dans les corpus. Une liste d'expressions à exclure de l'annotation, manuelle et automatique, a donc été dressée.

- Évènements historiques :
  le 6 Juin 1944
- Évènements autres (date de création d'entreprise etc.) :
  Créé depuis le 1er Janvier 1987
  l'établissement ouvert au printemps 2008
  Ce vieux moulin à farine de 1524
  rénovée durant l'hiver 2009-2010
- Durées exprimant une distance :
  à 1h30 de Paris
- Fréquences non-temporelles :
  je suis allée une fois à New-York
- Toponymes et expressions utilisées en nom :
  rue du 6 mai 1956
  Le 11 septembre est encore très présent dans les esprits

Conclusion

L'étude de corpus a permis de typer les expressions temporelles utilisées dans les documents touristiques.

Ce travail a permis de choisir le schéma d'annotation le plus adapté à nos besoins ainsi qu'aux expressions temporelles à représenter, bien qu'il ait fallu étendre ce schéma pour parfaire son utilisation dans le projet TourinFlux.

Avec le schéma d'annotation, il est la base sur laquelle s'est reposée le développement de la grammaire hors-contexte utilisée pour la détection et l'annotation des expressions temporelles.

Chapitre 3

Grammaire d'extraction et d'annotation

Sommaire

La détection et l'annotation des expressions temporelles et des informations touristiques s'est fait à l'aide d'une grammaire hors-contexte.

Le travail préliminaire sur le guide d'annotation a permis de relever des phénomènes nécessitant une attention particulière lors du développement de la grammaire. Cependant, les principaux enjeux étaient d'assurer une couverture la plus large avec le minimum de bruit possible et la maintenabilité de la grammaire une fois le stage terminé.

3.1 GramLab Unitex

Unitex est un système de traitement de corpus permettant de construire des grammaires hors-contexte à l'aide de réseaux de transitions, proches des automates à états finis, appelés "graphes".

Il a été créé en 2002 à l'université Paris-Est Marne-la-Vallée, afin de reproduire les fonctionnalités d'Intex qui n'était pas open-source, et a été amélioré à maintes reprises à l'aide du développement collaboratif.

Ces transitions indiquent soient des éléments terminaux (voir figure 3.1) soit d'autres réseaux de transitions (voir figure 3.2) appelés alors "sous-graphes".

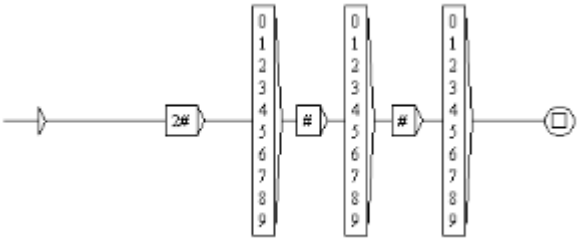


FIGURE 3.1 – Exemple de réseaux de transitions indiquant des éléments terminaux

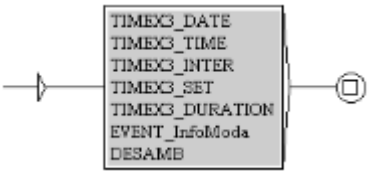


FIGURE 3.2 – Exemple d'appels de sous-graphes

Il permet aussi l'utilisation de dictionnaires électroniques de format DELA contenant des entrées lexicales simples et/ou composées (DELAS) ou fléchies (DELAF) accompagnées ou non d'informations morpho-syntaxiques.

- DELAS : lemme,code grammatical +code
- DELAF : mot, lemme.catégorie+sous classe :flexion

3.2 Processus

La détection et l'annotation par la grammaire hors-contexte s'effectue en deux passes principales.

3.2.1 Pré-traitement

Nous avons ajouté un pré-traitement qui permet :

- la normalisation des apostrophes, guillemets etc. afin de pouvoir les utiliser, si besoin, dans les graphes sans devoir déclarer chaque type possible.
- l'ajout des balises déclarant le début et fin de document XML pour pouvoir les utiliser dans l'outil permettant l'évaluation entre la sortie de la grammaire et la référence (section 4.1).

– la séparation des séquences avec le marqueur STOP à chaque début de nouvelles phrases qui oblige la grammaire à ne pas faire de liens entre les annotations d'une phrase avec celles de la phrase précédente ou suivante.
– l'ajout d'une annotation OI en début de chaque phrase en attendant les annotations des objets et évènements touristiques par les partenaires.

3.2.2 Annotation des expressions temporelles

Cette première passe détecte et annote les expressions temporelles et les informations touristiques telles que décrites en 2.2. Pour cela, nous avons utilisé le logiciel GramLab IDEling qui est un environnement ajouté à Unitex dans le cadre du projet GramLab (http ://apoliade.com/Gramlab) afin de lui apporter des fonctionnalités de gestion de projet utiles dans un contexte de production industrielle.

Afin de faciliter la lisibilité et la maintenance, le système de sous-graphes a été utilisé pour diminuer la taille des graphes qui détectent et annotent. La division a été faite de manière à regrouper les objets similaires, par exemple (voir figure 3.3), le graphe TIMEX3_DATE est chargé d'appeler les sous-graphes annotant les dates alphabétiques (*"22 août"*), numériques (*"22/08/14"*), énumérées (*"21 et 22 août"*) et ancrées (*"deux jours avant dimanche"*).

FIGURE 3.3 – Graphe d'annotation des dates

Toujours pour permettre une meilleure maintenance, nous avons utilisé les dictionnaires électroniques pour définir les éléments terminaux lorsque cela était possible. Ainsi, si des modifications de ces éléments sont nécessaires, il suffira de modifier directement les dictionnaires. L'appel au dictionnaire dans la figure 3.4 permet de définir en tant qu'éléments terminaux tous les déterminants féminins et singuliers (*DET :fs*) et tous les prépositions (*PREP*) des dictionnaires.

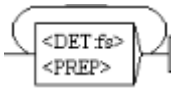


FIGURE 3.4 – Exemple d'appel au dictionnaire

3.2.3 Annotation des liens

Cette deuxième passe se base sur la sortie de la première passe en utilisant les annotations ajoutées au texte et rajoute les liens (TLINK et SLINK) entre

les différentes annotations. Nous avons dû changer de logiciel pour utiliser la dernière version d'Unitex 3.1beta afin de pouvoir utiliser CasSys, un système de cascades de transducteurs qui n'était pas encore intégré dans GramLab IDEling lors du développement de la grammaire.

Ce système de cascade CasSys a été développé par N. Friburger et D. Maurel en 2004 (Friburger and Maurel, 2004) et permet de définir une succession de graphes dans laquelle chaque graphe prend en entrée la sortie du graphe précédant.

Le développement de ces graphes a été particulièrement hasardeux à cause de l'utilisation de <TOKEN>* et de son comportement gourmand. En effet, pour permettre l'ajout des liens dans cet exemple : *"24 Décembre : Chant de Noël sur la place de l'église à 20h"*, nous devions décrire une règle permettant la présence de n'importe quel caractère entre la date, l'évènement et l'heure.

Cependant, le masque <TOKEN>* permettant de reconnaître cette suite possible de caractère incluait également les balises à détecter, donc aucun lien ne pouvait être ajouté.

Ce problème a été résolu en utilisant des contextes négatifs gauche et droit permettant de décrire ce qui ne peut pas suivre ou précéder la séquence à reconnaître. Cependant, cette solution peut considérablement allonger le temps de traitement et ne devrait être utilisée qu'en dernier recours.

Pour un exemple de graphe de liens, voir annexe B.

3.2.4 Post-traitement

Une grammaire hors-contexte ayant ses limites (section 3.3), les sorties de la grammaire devront être enrichies. Cette tâche est prise en charge par un membre de l'équipe en post-traitement. Elle comprend :

- Compléter les attributs dont la valeur est présente dans le texte annoté. Afin de faciliter cette complétion, nous avons rajouté au maximum des balises contenant les informations nécessaires (voir figure 3.5).
- Calcul de l'attribut *value* lorsque celui-ci nécessite l'utilisation d'une *temporalFunction*.
- Correction de l'incrémentation des ID des annotations.

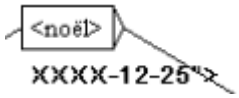


FIGURE 3.5 – Balise contenant l'information pour la complétion de l'attribut value

3.3 Limites des grammaires hors-contexte

La grammaire hors-contexte a permis la détection et l'annotation d'une majorité des phénomènes mais elle montrait ses limites sur certains points.

Étant donné que nous la paramétrons en mode MERGE, les annotations sont ajoutées au texte selon leur position dans les graphes (voir figure 3.6). Il est donc impossible de compléter certains attributs alors que l'on possède leurs valeurs dans le texte entre les balises d'annotation [1].

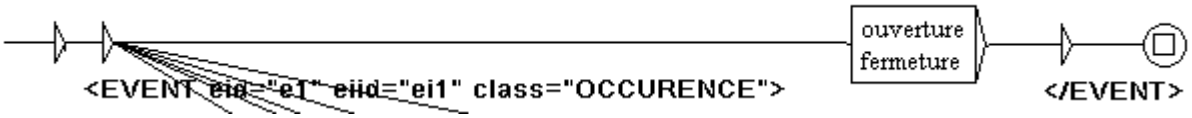


FIGURE 3.6 – Placement des annotations lors de la détection

L'incrémentation automatique des identifiants uniques est également impossible car l'utilisation des variables dans Unitex n'est pas très développée. En effet, il est possible d'enregistrer le texte détecté dans une variable pour la réutiliser plus loin dans le graphe, mais il est impossible d'inclure une variable qui servirait de compteur et qui s'incrémenterait à chaque annotation rajoutée au texte. Or, les ID sont essentiels en TimeML car c'est grâce à eux que les liens entre les différentes annotations peuvent être faits ; la correction en post-traitement doit donc être effectuée avec précaution.

De plus, lors de la création des graphes pour ajouter les liens entre les annotations, nous nous sommes rendus compte que certains cas nécessitaient l'utilisation des contextes.

Par exemple, dans *"Dimanche 8, 15 et 22 Décembre : Marchés de Noël de 14h à 18h"*, il faut pouvoir ajouter autant de liens temporels que de dates, c'est-à-dire trois liens qui reliraient chaque date en début de phrase avec l'intervalle d'horaires en fin de phrase.

Or, cela nécessite l'utilisation d'une grammaire contextuelle.

Pour finir, les transducteurs composant une grammaire peuvent rapidement se multiplier et devenir imposants. La maintenance peut donc devenir difficile, en particulier pour ceux qui n'ont pas développé la grammaire. Il est alors essentiel de structurer la grammaire le plus clairement possible, comme cela a été précisé en 3.2.2. Il est également recommandé de créer un inventaire et une carte des transducteurs qui expliciterons les relations entre les différents transducteurs (père-fils, appels au dictionnaire etc.), même si ces documents peuvent être fastidieux à lire (voir annexes C et D).

3.4 Perspectives

Il reste un travail sur certains types de liens qui n'ont pas eu le temps d'être implémentés : les "inverses" de certains liens existants, par exemple :

1. Ici, il n'est pas question des informations qui doivent être calculées à l'aide de temporalFunction.

- Date-OI (*"23 juillet : Concert"*) : ce lien sera nécessaire lorsque l'on aura les annotations des objets et évènements touristiques annotés par les partenaires, au lieu de l'annotation par défaut du pré-traitement.
- Time-Date (*"à 15h le 23 juillet"*) : ce lien n'était pas présent dans les corpus mais est possible.

Cependant, ces ajouts doivent être faits prudemment car il existe un risque de conflits ou de sur-annotations, par exemple :
Musée ouvert du lundi au vendredi 8h à 18h, le samedi de 10h à 16h
"de 8h à 18h" doit être relié avec *"du lundi au vendredi"* mais pas avec *"le samedi"*. Or, un lien Time-Date pourra tout de même le créer.

Conclusion

En s'appuyant sur le guide d'annotation créé à partir de l'étude linguistique, nous avons pu prendre en compte certains phénomènes présents dans les corpus, par exemple les abréviations des jours de la semaine pouvant causer des ambiguïtés pour la grammaire hors-contexte[2].

Nous avons pu également structurer les graphes constituant la grammaire de manière à la rendre la plus lisible et maintenable par une personne extérieure à son développement.

2. "Jeudi" en "jeu", "mercredi" en "mer" etc.

Chapitre 4

Évaluation de la grammaire hors-contexte

Sommaire

En parallèle au développement de la grammaire-hors contexte, il était nécessaire de créer un corpus annoté de référence permettant d'évaluer les annotations de celle-ci.

Dans un premier temps, nous nous sommes tournés vers l'annotation manuelle par des membres du laboratoire L3i lors d'une journée d'annotation (section 4.2). Cette tâche s'étant avérée plus complexe que prévue, la référence a été créée uniquement à partir de l'annotation manuelle d'un unique annotateur expert (section 4.3) afin d'obtenir une indication sur la qualité de l'annotation produite en sortie de la grammaire développée (section 4.3.1), dans l'attente d'une réalisation plus adaptée du corpus de référence (section 4.4).

4.1 Outil pour l'annotation manuelle

Nous avons décidé d'utiliser GATE (General Architecture for Text Engineering) comme outil d'annotation manuelle pour plusieurs raisons.

Il permet d'importer des schémas XML qui permettent ainsi d'annoter des textes avec le langage de balisage choisi. Nous avons donc pu modifier des schémas existants pour TimeML en supprimant certains éléments que nous n'avons

pas utilisé et en ajoutant nos apports au standard comme décrit en (voir Annexe A pour un exemple de schéma).

Il accepte en entrée plusieurs formats dont le XML, ce qui permet de pouvoir utiliser des fichiers déjà annotés manuellement ou automatiquement pour modifier les annotations présentes ou pour utiliser l'outil Diff Tool calculant la précision, le rappel et la F-mesure de deux documents.

Cependant, il a quelques défauts d'ergonomie notamment pour l'ajout de balises auto-fermantes[1] puisque l'on ne peut pas directement créer une annotation avec un empan de zéro. Il y a également un problème de pop-up pour l'annotation qui s'échappe si on ne l'épingle pas systématiquement et cela peut entraîner une modification du texte à annoter si l'utilisateur est en train d'écrire. De plus, GATE ne permet pas d'annuler une modification, l'annotateur doit donc corriger lui-même le texte, ce qui n'est pas souhaitable.

4.2 Journée d'annotation

En se basant sur le guide d'annotation créé en amont, une journée d'annotation a été mise en place au sein du laboratoire.

4.2.1 Processus

Le travail précédant la journée d'annotation a été basé sur la méthodologie pour l'annotation exposée dans (Fort, 2012), cependant il a été adapté à la taille réduite et aux circonstances de notre campagne d'annotation. En effet, dans sa thèse, K. Fort présente une méthodologie de campagne d'annotation telle qu'un client la demande pour tester ou entraîner ses outils. Or, dans notre projet, la campagne visait à évaluer notre propre outil. Tous les acteurs de cette campagne étaient donc membres du laboratoire L3i.

- **Le corpus à annoter**
  Chaque fichier était numéroté et comportait trois à quatre phrases représentatives et extraites des corpus touristiques à disposition. Les fichiers à annoter étaient divisés en plusieurs niveaux allant crescendo dans la complexité des expressions temporelles et des liens à annoter.

  1. Expressions temporelles complètes et incomplètes
  2. Ajout de suites d'expressions temporelles (ellipses) et du TLINK *reltype XOR*
  3. Ajout des intervalles bornés ou non et du TLINK *reltype IS_ EXCLUDED*
  4. Ajout des modalités (SLINK)

- **Le guide d'annotation**
  Le guide d'annotation avait été finalisé par les retours lors de la création d'une mini-référence, c'est-à-dire un extrait du corpus annoté par des

1. Les balises auto-fermantes sont utilisées pour exprimer les intervalles et les liens en TimeML.

annotateurs experts, dans le sens où ils ont pris connaissance du guide d'annotation.
La journée d'annotation aurait permis d'effectuer de nouvelles modifications au guide grâce à l'accord inter-annotateurs qui aurait pu mettre en évidence des incohérences.
Une version raccourcie du guide avait été faite afin de faciliter sa lecture par les annotateurs. Les deux versions étaient cependant à disposition sur les postes de travail, ainsi que des versions papier.

- **Les annotateurs**
  Les membres du laboratoire étaient invités à venir annoter des textes en fonction de leur temps libre. De par les activités menées au sein du laboratoire, les participants avaient déjà participé à ce type de journée d'annotation, cependant ils étaient non-experts dans le domaine en question.
- **La formation des annotateurs**
  Les annotateurs étaient formés lors de leur arrivée et pour la plupart de manière individuelle ou en binôme. La formation se voulait courte car la majorité des annotateurs ne pouvaient pas rester plus d'une demi-heure.
  Elle comprenait une explication du but de la journée d'annotation, de ce qu'il fallait annoter ainsi que d'un exemple fait ensemble pour la prise en main du logiciel GATE. Les annotateurs avaient la possibilité de poser des questions lors qu'ils annotaient.
  Les premiers fichiers à annoter étaient dédiés à l'entraînement. Ensuite, des numéros de fichiers leur étaient attribués de manière à obtenir des annotations d'au moins deux annotateurs différents.

4.2.2 Analyse du résultat

Le résultat de cette journée d'annotation a été peu concluant pour la création d'un corpus annoté de référence. Cependant, cela a été instructif quant à la mise en place d'une campagne d'annotation et de la création d'une référence pour l'évaluation.

D'un point de vue quantitatif, nous n'avons pas obtenu assez d'annotation, pour les raisons suivantes :

- La prise en main du logiciel GATE était trop longue et les annotateurs disposaient en général d'un temps réduit puisqu'ils prenaient une pause sur leur propre travail pour participer à la journée d'annotation.
- Le guide d'annotation était également trop long à lire, malgré la version raccourcie, et n'a pas été utilisé par tous les annotateurs, de ce fait certaines annotations étaient incomplètes.
- Il y a eu également des problèmes vis-à-vis de la langue. En effet, certains annotateurs n'étaient pas des locuteurs natifs du français et certains mots dans les textes leur posaient problème, notamment pour définir si cela correspondait à l'évènement touristique de la phrase ou non.

De ce fait, parmi les quatre niveaux du corpus à annoter, aucun annotateur n'a pu atteindre un niveau d'aisance suffisant pour annoter des textes au-dessus du premier niveau.

4.3 Annotation par un annotateur expert

Face au besoin imminent d'avoir une indication chiffrée pour évaluer la grammaire hors-contexte, nous avons décidé que le corpus de référence serait annoté par un seul annotateur expert malgré les biais que cela entraîne.

J'ai donc fait passer le corpus de référence dans la version finale de la grammaire et ait ensuite corrigé la sortie afin que l'annotation soit plus rapide.
Nous avons pu ensuite utiliser ce corpus de référence et la sortie de la grammaire sur ce même corpus dans Diff Tool.

Bien que l'annotation de la référence ait été faite uniquement en suivant le guide d'annotation, les résultats ont nécessairement un biais puisque l'annotateur de la référence est également l'auteur du guide et de la grammaire hors-contexte qui ait évaluée. Nous manquons probablement de recul sur le guide d'annotation puisque la campagne d'annotation n'a pas pu amener à une remise en question ou validation de celui-ci par des annotateurs n'ayant pas participé directement ou indirectement à sa conception.

4.3.1 Résultats de l'accord inter-annotations

Dans cette section, nous présentons les résultats de l'accord inter-annotations effectué à l'aide de Diff Tool.

Mesures utilisées

Le rappel permet d'évaluer le nombre d'annotations supplémentaires (donc erronées) par rapport au corpus de référence, ce phénomène est aussi appelé du bruit. La précision permet d'évaluer le nombre d'annotations absentes par rapport au corpus de référence, on parle ici de silence. La pondération de ces deux mesures s'appelle la F-mesure. Dans ces trois mesures, plus le résultat tend vers 1, meilleur il est.

L'outil Diff Tool présente également le nombre d'annotations correctes, partiellement correctes, manquantes (silence), et supplémentaires (bruit).

Corpus évalués

Nous avons évalués la grammaire-hors contexte sur plusieurs corpus.

- corpus_exemple : est un des fichiers utilisés pour la campagne d'annotation manuelle contenant trois à quatre phrases chacun. Il a été formé à partir des corpus touristiques mis à disposition par les partenaires.
- corpus GALevt1, GALevt2 et GALevt2013 : contenaient les agendas des évènements touristique en Othe Armance.
- référence : est le corpus de référence créé par l'annotateur expert. C'est à celui-ci que les autres corpus ont été comparés.

Tableaux des résultats

Ces résultats fournis par Diff Tool sont faussés à cause de différences dans les index des caractères. En effet, Diff Tool compare les annotations selon leur position dans le texte. Si les index sont différents, les annotations seront considérées comme partiellement correctes, mais au-delà d'un certain nombre de caractères, il considère que les annotations sont différentes. Une des annotations est alors considérée comme manquante et l'autre comme faux-positif (voir exemple en annexe E).

Les résultats peuvent donc être particulièrement bas pour certains corpus alors que ce n'est pas le cas. Cependant, les tableaux des résultats seront quand même inclus dans ce mémoire car ils permettent d'évaluer approximativement la grammaire hors-contexte.

EVENT

	corpus_exemple	corpusGALevt1	corpusGALevt2	corpusGALevt2013
Rappel	1	0.0526	1	0.7143
Précision	1	0.0179	0.5	0.4167
F-mesure	1	0.0267	0.6667	0.5263
Corrects	1	1	4	5
Partiellement corrects	5	18	0	2
Manquants	0	0	0	0
Faux positif	0	37	4	5

TIMEX3

	corpus_exemple	corpusGALevt1	corpusGALevt2	corpusGALevt2013
Rappel	0.8043	0.2043	0.9348	0.6826
Précision	0.8043	0.1908	0.9149	0.6462
F-mesure	0.8043	0.1973	0.9247	0.664
Corrects	37	66	129	168
Partiellement corrects	4	173	0	65
Manquants	5	79	9	13
Faux positif	5	102	12	27

SLINK

	corpus_exemple	corpusGALevt1	corpusGALevt2	corpusGALevt2013
Rappel	1	0.0667	manquant	1
Précision	1	0.0667	manquant	1
F-mesure	1	0.0667	manquant	1
Corrects	2	1	manquant	2
Partiellement corrects	0	0	manquant	0
Manquants	0	14	manquant	0
Faux positif	0	14	manquant	0

TLINK

	corpus_exemple	corpusGALevt1	corpusGALevt2	corpusGALevt2013
Rappel	manquant	0.1714	0.7953	0.5939
Précision	manquant	0.172	0.9528	0.6858
F-mesure	manquant	0.1717	0.867	0.6366
Corrects	manquant	48	101	155
Partiellement corrects	manquant	0	0	0
Manquants	manquant	232	26	106
Faux positif	manquant	231	5	71

4.4 Perspectives pour la ré-évaluation

Il existe plusieurs pistes pour la création d'un corpus de référence sans biais.

Une première piste possible serait de pré-annoter les textes à l'aide de la grammaire pour que les annotateurs, experts ou non, n'aient plus qu'à vérifier s'il n'y a pas de silence ou de bruit, et si les annotations existantes sont correctes au niveau des attributs.

Un travail similaire a été fait par (Bittar, 2010a) et a démontré que la pré-annotation permettait de grandement réduire le temps d'annotation.

Une deuxième piste serait de recruter des annotateurs extérieurs au laboratoire qui se consacreraient à l'annotation des textes. Cette solution écarterait les problèmes liés au temps de prise en main du logiciel d'annotation manuelle et du guide d'annotation et amènerait à une meilleure gestion de la quantité et de la qualité des annotations, ainsi qu'à une possible évolution du guide d'annotation.

Cependant, elle a le désavantage d'être coûteuse en main d'œuvre autant pour les annotateurs que pour le gestionnaire de la campagne d'évaluation. Cette solution est peu utilisé dans les campagnes d'annotation traditionnelles en valeur du *crowdsourcing*[2], comme l'explique K. Fort dans sa thèse (Fort, 2012).

2. "Production participative" permettant de faire appel au savoir-faire et au raisonnement d'une main d'œuvre importante, sans rémunération obligatoire.

Une dernière piste serait donc développé un jeu sérieux pour rendre la tâche d'annotation plus ludique et attrayante. Cela permettrait également d'avoir recours au *crowdsourcing* qui ôterait l'aspect financier du recrutement d'annotateurs dédiés à la tâche, et permettrait l'obtention d'une large quantité d'annotation si le jeu est bien conçu et diffusé.

Des exemples de ce type de jeu sérieux sont JeuxDeMots développé par le laboratoire LIRMM de l'université de Montpellier pour alimenter un réseau lexical (Lafourcade and Jouvert, 2008), et Zombilingo développé par le laboratoire LORIA de l'université de Lorraine afin d'annoter des corpus en syntaxe de dépendances (Fort et al., 2014b).

Les points négatifs de cette méthode sont, dans un premier temps, la difficulté de la mise en place du jeu. En effet, les annotations attendues sont particulièrement détaillées et il faut trouver un scénario qui permette de les obtenir. Cela implique aussi d'avoir une personne pour concevoir et maintenir le jeu. C'est pour cela que des contacts avec l'équipe du laboratoire LORIA ont été pris pour le projet TourinFlux.

De plus, étant donné que n'importe quelle personne pourrait produire des annotations, il faudrait donc tout de même vérifier la qualité de celles-ci. Cela a aussi le désavantage de rendre plus difficile, voir impossible, l'évolution du guide d'annotation puisqu'il faut pour cela que les annotateurs se mettent d'accord sur leurs annotations respectives. Et pour finir, il existe un problème éthique induit par le fait de faire travailler indirectement des personnes et d'utiliser leurs résultats sans les rémunérer (Fort et al., 2014a).

Conclusion

Nous avons pu voir l'importance de l'étude linguistique lors de ce stage.

En effet, c'est grâce à cette étude que nous avons relevé certaines particularités des expressions temporelles dans le domaine du tourisme. Toutes les expressions ne possèdent pas une syntaxe simple comme celle que l'on peut trouver en périodes et horaires d'ouverture sur des tracts par exemple. Beaucoup d'expressions temporelles n'expriment pas clairement leur valeur exacte et nécessite un raisonnement pour la déduire. Ce raisonnement ne pose généralement pas de problème pour un humain, mais il est en revanche plus difficile à faire réaliser à une machine.

En partant de cette étude, nous avons également pu choisir le schéma d'annotation adéquat pour traiter les expressions temporelles dans des corpus touristiques, même si celui-ci a nécessité des apports pour s'adapter pleinement à notre tâche.

Le développement de la grammaire d'extraction et annotation s'est basé à la fois sur l'étude linguistique et le schéma d'annotation, tant pour les traitements marginaux à faire pour prendre en charge certaines ambiguïtés, que pour structurer l'ensemble des graphes la composant.

Nous avons relevé quelques désavantages sur l'utilisation d'une grammaire hors-contexte pour la tâche d'annotation, cela reste tout de même un bon outil pour la détection et l'annotation des expressions temporelles dans les corpus touristiques.

Le point principal à garder à l'esprit lors du développement d'une grammaire à l'aide de transducteurs est d'assurer au maximum leur maintenabilité pour que leur développeur ne soit pas l'unique personne à pouvoir gérer la grammaire.

Néanmoins, même en cas de changement de technologie pour l'extraction et l'annotation et/ou un changement du schéma d'annotation au cours du projet, l'étude linguistique resterait toujours valide, elle est donc bien l'étape primordiale pour la détection et l'annotation des expressions temporelles.

Sur le plan personnel, ce stage m'a permis d'approfondir des connaissances sur les grammaires hors-contexte acquises lors du master ainsi qu'un outil qui permet leur développement, c'est-à-dire Unitex/GramLab.

J'ai également pu aborder les domaines de l'annotation et de l'évaluation lors du travail sur le guide et la campagne d'annotation, qui prennent une part importante dans le domaine du Traitement Automatique de la Langue.

La rédaction du guide d'annotation sous forme d'un document scientifique m'a donné l'occasion d'utiliser pour la première fois le logiciel de composition LaTeX.

Le déroulement du stage au sein d'un laboratoire m'a permis d'avoir un aperçu de l'organisation du travail au sein d'un laboratoire de recherche, qui n'est pas si différente qu'en entreprise.

Bibliographie

Bittar, A. (2010a). *Building a TimeBank for French : A Reference Corpus Annotated According to the ISO-TimeML Standard*. PhD thesis, Université Paris Diderot.

Bittar, A. (2010b). *ISO-TimeML Annotation Guidelines for French Version 1.0*.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2014). *Developing Langage Processing Components with GATE Version 8 (a User Guide)*.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2003). *TIDES 2003 Standard for the Annotation of Temporal Expressions*.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. PhD thesis, Université Paris 13 - Sorbonne Paris Cité.

Fort, K., Adda, G., Sagot, B., Mariani, J., and Couillault, A. (2014a). Crowdsourcing for language resource development : Criticisms about amazon mechanical turk overpowering use. In Vetulani, Z. and Mariani, J., editors, *Human Language Technology Challenges for Computer Science and Linguistics*, pages 303–314. Springer International Publishing.

Fort, K., Guillaume, B., and Chastant, H. (2014b). Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Gamification for Information Retrieval (GamifIR'14) Workshop*, Amsterdam, Pays-Bas.

Fortin, J., Carloni, O., Leclère, M., and Weiser, S. (2009). Extraction et exploitation de données temporelles pour un portail d'e-tourisme. *Fouille de Données Temporelles - Analyses de Flux de Données*.

Friburger, N. and Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*.

Lafourcade, M. and Jouvert, A. (2008). Jeuxdemots : un prototype ludique pour l'émergence de relations entre termes. *Journées internationales d'Analyse statistiques des Données Textuelles*.

Paumier, S. (2006). *Unitex - Manuel d'utilisation.*

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). *TimeML Annotation Guidelines Version 1.2.1.*

Saurí, R., Saquete, E., and Pustejovsky, J. (2010). *Annotating Time Expressions in Spanish : TimeML Annotation Guidelines (Version TempEval-2010).*

Teissèdre, C. (2012). *Analyse sémantique automatique des adverbiaux de localisation temporelle : application a la recherche d'information et a l'acquisition de connaissances.* PhD thesis, Université Paris Ouest-Nanterre La Défense.

Université de La Rochelle (2014). Site du laboratoire de recherche l3i. Repéré le 27 août 2014 à `http://l3i.univ-larochelle.fr/`.

Weiser, S. (2010). *Repérage et typage d'expressions temporelles pour l'annotation sémantique automatique de pages Web - Application au e-tourisme.* PhD thesis, Université Paris Ouest Nanterre La Défense.

Résumé

Ce mémoire professionnel présente le travail effectué au sein du laboratoire L3i de l'université de La Rochelle. Ce stage a prit part au projet TourinFlux dont le but est la création d'un tableau de bord réunissant les informations disponibles sur les différents territoires de France pour les acteurs du tourisme.

La contribution au projet était le développement d'une grammaire hors-contexte pour la détection et l'annotation des expressions temporelles dans les documents touristiques.

Le travail effectué a permis également de souligner l'importance de l'étude linguistique dans le choix du schéma d'annotation, de l'écriture du guide d'annotation ainsi que du développement de l'annotateur automatique, peu importe la méthode choisie pour cela.

Mots-clefs : Extraction automatique d'information, grammaire hors-contexte, expressions temporelles, corpus touristiques, schéma et guide d'annotation

Abstract

This professional report presents the work done within the laboratory L3i of the university of La Rochelle. This internship has took part in the TourinFlux project, the purpose of which is the creation of a dashboard gathering information available on the French territory for various stakeholders in tourism.

The contribution to this project has been the development of a context-free grammar for the detection and annotation of the temporal expressions in tourism documents.

The work achieved has also pointed out the importance of the linguistic study to choose the annotation schema, to write the annotation guidelines, and to develop the automatic annotator, whatever the technology chosen to do so.

Keywords : automatic information retrieval, context-free grammar, temporal expressions, tourist corpora, annotation schema and guidelines

ANNEXES
Projet TourInFlux
Annotation des expressions temporelles

Drat Lucie

UFR LLASIC

Mémoire de master 2 professionnel – Sciences du langage – Industrie de la langue

Parcours : Traitement automatique de la langue écrite et parlée

Sous la direction de : Agnès Tutin – Université Stendhal Grenoble III
Mickaël Coustaty – Université La Rochelle – L3i
Alain Couillault – Université La Rochelle – L3i

Année universitaire 2013-2014

Annexe A

Schéma XML

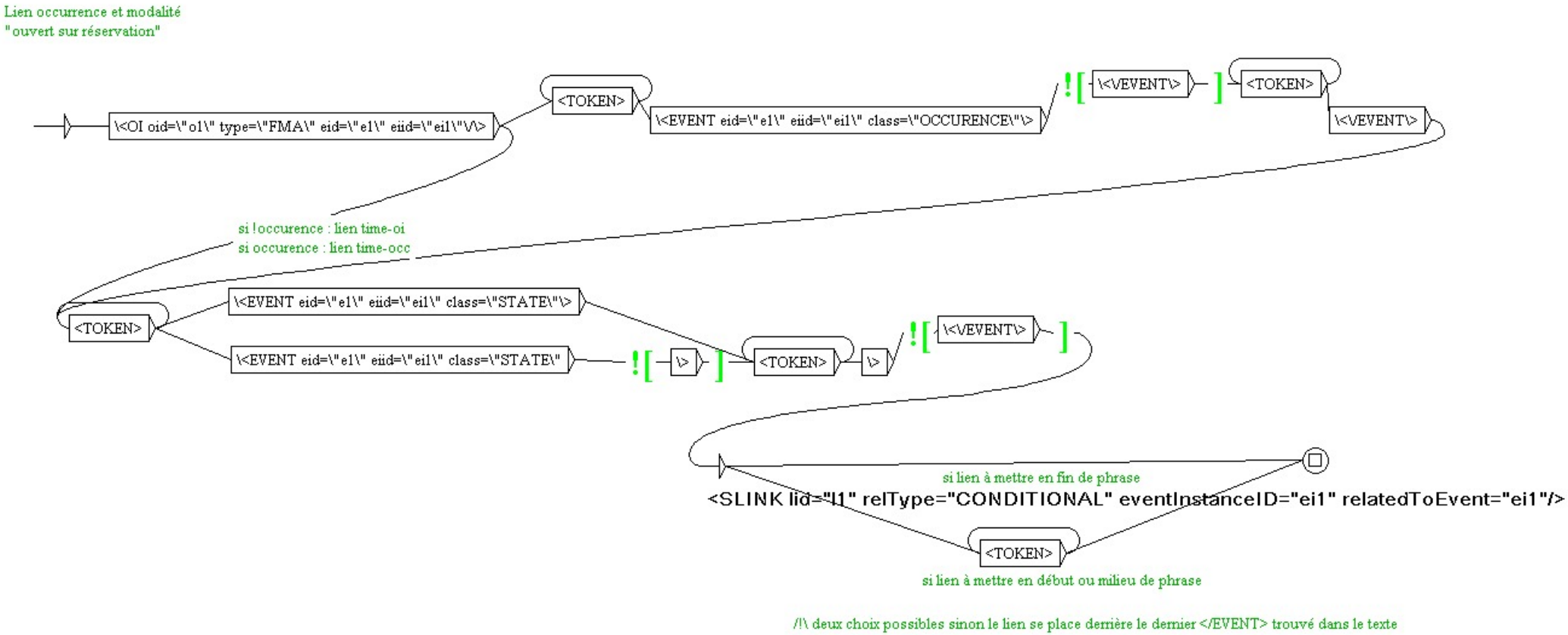
Voici un exemple de schéma XML importable dans GATE qui permet d'annoter un lien temporel (TLINK) entre deux éléments en TimeML.

```
<?xml version="1.0"?>
- <schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  - <element name="TLINK">
    - <complexType>
        <attribute name="lid" use="required" type="LinkID"/>
        <attribute name="eventInstanceID" type="EventInstanceID"/>
        <attribute name="timeID" type="TimeID"/>
        <attribute name="signalID" type="SignalID"/>
        <attribute name="relatedToTime" type="TimeID"/>
        <attribute name="relatedToEventInstance" type="EventInstanceID"/>
      - <attribute name="relType" use="required">
        - <simpleType>
          - <restriction base="string">
              <enumeration value="BEFORE"/>
              <enumeration value="AFTER"/>
              <enumeration value="INCLUDES"/>
              <enumeration value="IS_INCLUDED"/>
              <enumeration value="SIMULTANEOUS"/>
              <enumeration value="IAFTER"/>
              <enumeration value="IBEFORE"/>
              <enumeration value="IDENTITY"/>
              <enumeration value="BEGINS"/>
              <enumeration value="ENDS"/>
              <enumeration value="BEGUN_BY"/>
              <enumeration value="ENDED_BY"/>
              <enumeration value="IS_EXCLUDED"/>
              <enumeration value="XOR"/>
              <enumeration value="AROUND"/>
            </restriction>
          </simpleType>
        </attribute>
      </complexType>
    </element>
</schema>
```

Annexe B

Graphe de lien

Voici un exemple de graphe créant un lien de condition entre une occurrence et une modalité (*"ouvert sur rendez-vous"*.


Lien occurrence et modalité
"ouvert sur réservation"
\<OI oid=\"o1\" type=\"FMA\" eid=\"e1\" eiid=\"ei1\"\/\>
<TOKEN>
\<EVENT eid=\"e1\" eiid=\"ei1\" class=\"OCCURENCE\"\>
\<\/EVENT\>
<TOKEN>
\<\/EVENT\>
si !occurence : lien time-oi
si occurence : lien time-occ
<TOKEN>
\<EVENT eid=\"e1\" eiid=\"ei1\" class=\"STATE\"\>
\<EVENT eid=\"e1\" eiid=\"ei1\" class=\"STATE\"
\>
<TOKEN>
\>
\<\/EVENT\>
si lien à mettre en fin de phrase
<SLINK lid="l1" relType="CONDITIONAL" eventInstanceID="ei1" relatedToEvent="ei1"/>
<TOKEN>
si lien à mettre en début ou milieu de phrase
/!\ deux choix possibles sinon le lien se place derrière le dernier </EVENT> trouvé dans le texte

Annexe C

Inventaire des graphes

Voici l'inventaire des graphes qui fait partie, avec la carte (annexe D), des documents créés pour aider à la lisibilité et la maintenance de la grammaire hors-contexte développée.

Inventaire des graphes

Drat Lucie*

24 juillet 2014

Résumé

Inventaire et descriptif rapide des graphes de la grammaire.

1 Passe 1 : Annotation des expressions temporelles

1.1 Graphe principal : main.grf

- Annotation : Expressions temporelles, occurrences et modalités
- Contenu dans : Aucun
- Contient :
  TIMEX3_DATE.grf
  TIMEX3_TIME.grf
  TIMEX3_INTER.grf
  TIMEX3_SET.grf
  TIMEX3_DURATION.grf
  EVENT_InfoModa.grf
  DESAMB.grf
- Appel au dictionnaire : Aucun

1.2 TIMEX3_DATE.grf

- Annotation : Dates
- Contenu dans :
  main.grf
  TIMEX3_DURATION.grf
  TIMEX3_INTER_DATEg
  TIMEX3_INTER_DATEd.grf

*Université Stendhal Grenoble III/Laboratoire L3I, lucie.drat@gmail.com

- Contient :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Num.grf
  TIMEX3_DATE_Enum.grf
  TIMEX3_DATE_Ancre.grf
- Appel au dictionnaire : Aucun

1.3 TIMEX3_TIME.grf

- Annotation : Heures
- Contenu dans :
  main.grf
  TIMEX3_INTER_TIMEc.grf
  TIMEX3_INTER_TIMEg.grf
  TIMEX3_INTER_TIMEd.grf
- Contient :
  TIMEX3_TIME_HH
  TIMEX3_TIME_HP
- Appel au dictionnaire : Aucun

1.4 TIMEX3_INTER.grf

- Annotation : Intervalles de dates et intervalles d'heures
- Contenu dans : main.grf
- Contient :
  TIMEX3_INTER_DATE
  TIMEX3_INTER_TIME
  V_Alph_Periodes
- Appel au dictionnaire : Aucun

1.5 TIMEX3_SET.grf

- Annotation : Ensembles de dates et ensembles d'heures
- Contenu dans : main.grf
- Contient :
  TIMEX3_SET_DATE
  TIMEX3_SET_TIME
- Appel au dictionnaire : Aucun

1.6 TIMEX3_DURATION.grf

- Annotation : Durées (*Toute l'année*) et durées ancrées à des dates ou des heures (*deux jours avant Noël*)

- Contenu dans : main.grf
- Contient :
  TIMEX3_DATE
  TIMEX3_TIME
  V_Duree.grf
  V_Alph_Annee.grf
  V_Modifieurs.grf
- Appel au dictionnaire :
  <BEFORE> (dico)
  <AFTER> (dico)
  <NEAR> (dico)
  <AV> (dico)
  <AP> (dico)

1.7 EVENT_InfoModa.grf

- Annotation : Occurrences et modalités (*ouverture à ..., sur réservation*)
- Contenu dans : main.grf
- Contient : Aucun
- Appel au dictionnaire : <A>

1.8 DESAMB.grf

- But : Détection de certains éléments pour empêcher leur annotation
- Contenu dans : main.grf
- Contient : Aucun
- Appel au dictionnaire : cf. graphe
- Note : Annotation de l'abréviation "jeu" pour différencier le jour du nom

1.9 TIMEX3_DATE_Alph.grf

- Annotation : Dates alphabétiques (*22 juillet 2014*)
- Contenu dans : TIMEX3_DATE.grf
- Contient :
  TIMEX3_DATE_Alph_SSAAAA.grf
  TIMEX3_DATE_Alph_SS.grf
  TIMEX3_DATE_Alph_MMAAAA.grf
  TIMEX3_DATE_Alph_MM.grf
  TIMEX3_DATE_Alph_WE.grf
  TIMEX3_DATE_Alph_DDMMAAAA.grf

TIMEX3_DATE_Alph_DDMM.grf
TIMEX3_DATE_Alph_DD.grf
TIMEX3_DATE_Alph_JJ.grf
TIMEX3_DATE_Alph_FF.grf
TIMEX3_DATE_Alph_OD.grf
V_Modifieurs.grf
- Appel au dictionnaire :
<JOUR> (dico)
<N>
- Note : Présence d'un contexte négatif droit

1.10 TIMEX3_DATE_Num.grf

- Annotation : Dates numérique (*22/07/14*)
- Contenu dans : TIMEX3_DATE.grf
- Contient :
TIMEX3_DATE_Num_DDMMAAAAs.grf
TIMEX3_DATE_Num_DDMMAAAAt.grf
TIMEX3_DATE_Num_DDMMs.grf
TIMEX3_DATE_Num_DDMMt.grf
- Appel au dictionnaire : Aucun
- Note : Pas de "Mois Année" à cause d'un risque de conflit avec "Date Mois" si l'année est égale ou inférieure à 12 (= mois de décembre et année 2012)

1.11 TIMEX3_DATE_Enum.grf

- Annotation : Énumération de dates (*le 21, 22 et 23 juillet*)
- Contenu dans : TIMEX3_DATE.grf
- Contient :
TIMEX3_DATE_Alph_DDMMAAAA.grf
TIMEX3_DATE_Alph_DDMM.grf
TIMEX3_DATE_Alph_DD.grf
TIMEX3_DATE_Alph_OD.grf
TIMEX3_DATE_Num_DDMMAAAAs.grf
TIMEX3_DATE_Num_DDMMAAAAt.grf
TIMEX3_DATE_Num_DDMMs.grf
TIMEX3_DATE_Num_DDMMt.grf
V_Ordinaux.grf
- Appel au dictionnaire : <JOUR> (dico)
- Note : "&" est remplacé par "&" en preprocessing

1.12 TIMEX3_DATE_Ancre.grf

- Annotation : Dates ancrées (*le premier dimanche avant Noël*) avec le marqueur associé
- Contenu dans : TIMEX3_DATE.grf
- Contient :
  TIMEX3_DATE_Alph_DDMMAAAA.grf
  TIMEX3_DATE_Alph_DDMM.grf
  TIMEX3_DATE_Alph_DD.grf
  TIMEX3_DATE_Alph_OD.grf
  TIMEX3_DATE_Num_DDMMAAAAs.grf
  TIMEX3_DATE_Num_DDMMAAAAt.grf
  TIMEX3_DATE_Num_DDMMs.grf
  TIMEX3_DATE_Num_DDMMt.grf
  V_Ordinaux.grf
- Appel au dictionnaire :
  <BEFORE> (dico)
  <AFTER> (dico)
  <NEAR> (dico)
  <AV> (dico)
  <AP> (dico)

1.13 TIMEX3_DATE_Alph_SSAAAA.grf

- Annotation : Saison Année (*Été 2014*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  Inter_anchor.grf
  Inter.grf
  TIMEX3_SET_DATE.grf
- Contient : V_Alph_Annee
- Appel au dictionnaire : <SAISON> (dico)

1.14 TIMEX3_DATE_Alph_SS.grf

- Annotation : Saison (*Eté*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  Inter_anchor.grf
  Inter.grf
  TIMEX3_SET_DATE.grf
- Contient : Aucun
- Appel au dictionnaire : <SAISON> (dico)

1.15 TIMEX3_DATE_Alph_MMAAAA.grf

- Annotation : Mois Année (*juillet 2014*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Alph_OD.grf
  Inter_anchor.grf
  Inter.grf
- Contient : V_Alph_Annee
- Appel au dictionnaire : Aucun

1.16 TIMEX3_DATE_Alph_MM.grf

- Annotation : Mois (*juillet*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Alph_OD.grf
  Inter_anchor.grf
  Inter.grf
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.17 TIMEX3_DATE_Alph_WE.grf

- Annotation : Semaine et week-end (*en semaine*, *le weekend*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Alph_OD.grf
  V_freq
- Contient : Aucun
- Appel au dictionnaire :
  <SE> (dico)
  <WE>(dico)

1.18 TIMEX3_DATE_Alph_DDMMAAAA.grf

- Annotation : Date Mois Année (*22 juillet 2014*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Enum.grf
  TIMEX3_DATE_Ancre.grf
  Inter_anchor.grf
  Inter.grf

- Contient :
  V_Alph_Date
  V_Alph_Annee
- Appel au dictionnaire : <JOUR> (dico)

1.19 TIMEX3_DATE_Alph_DDMM.grf

- Annotation : Date Mois (*22 juillet*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Enum.grf
  TIMEX3_DATE_Ancre.grf
  Inter_anchor.grf
  Inter.grf
  TIMEX3_SET_DATE.grf
- Contient : V_Alph_Date
- Appel au dictionnaire : <JOUR> (dico)

1.20 TIMEX3_DATE_Alph_DD.grf

- Annotation : Date (*le 22*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Enum.grf
  TIMEX3_DATE_Ancre.grf
  Inter_anchor.grf
  Inter.grf
  TIMEX3_SET_DATE.grf
- Contient : V_Alph_Date
- Appel au dictionnaire : Aucun
- Note : Présence d'un contexte négatif droit

1.21 TIMEX3_DATE_Alph_JJ.grf

- Annotation : Jour (*22 juillet 2014*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Alph_OD.grf
  Inter.grf
- Contient : Aucun
- Appel au dictionnaire :
  <lundi> (dico)
  <mardi> (dico)

<mercredi> (dico)
<jeudi> (dico)
<vendredi> (dico)
<samedi> (dico)
<dimanche> (dico)

- Note : L'annotation de l'abréviation "jeu" se fait dans DESAMBIG pour différencier le nom du jour

1.22 TIMEX3_DATE_Alph_FF.grf

- Annotation : Fêtes et jours fériés (*Pentecôte*)
- Contenu dans :
  TIMEX3_DATE_Alph
  TIMEX3_DATE_Alph_OD
  Inter.grf
  TIMEX3_SET_DATE
- Contient : V_AlphNum_Chiffres
- Appel au dictionnaire :
  <SE>
  <WE>
  <JOUR>

1.23 TIMEX3_DATE_Alph_OD.grf

- Annotation : Date ordinale (*première semaine de juillet*)
- Contenu dans :
  TIMEX3_DATE_Alph.grf
  TIMEX3_DATE_Enum.grf
  TIMEX3_DATE_Ancre.grf
- Contient :
  TIMEX3_DATE_Alph_JJ
  TIMEX3_DATE_Alph_WE
  TIMEX3_DATE_Alph_MMAAAA
  TIMEX3_DATE_Alph_MM.grf
  TIMEX3_DATE_Alph_FF.grf
  V_Ordinaux
- Appel au dictionnaire : Aucun

1.24 TIMEX3_DATE_Num_DDMMAAAAs.grf

- Annotation : Date Mois Année (*22/07/14*)
- Contenu dans :
  TIMEX3_DATE_Num.grf

TIMEX3_DATE_Enum.grf
TIMEX3_DATE_Ancre.grf
Inter_anchor.grf
Inter.grf
- Contient : Aucun
- Appel au dictionnaire :
<JOUR> (dico)

1.25 TIMEX3_DATE_Num_DDMMAAAAt.grf

- Annotation : Date Mois Année (*22-07-14*)
- Contenu dans :
TIMEX3_DATE_Num.grf
TIMEX3_DATE_Enum.grf
TIMEX3_DATE_Ancre.grf
Inter_anchor.grf
Inter.grf
- Contient : Aucun
- Appel au dictionnaire :
<JOUR> (dico)

1.26 TIMEX3_DATE_Num_DDMMs.grf

- Annotation : Date Mois (*22/07*)
- Contenu dans :
TIMEX3_DATE_Num.grf
TIMEX3_DATE_Enum.grf
TIMEX3_DATE_Ancre.grf
Inter_anchor.grf
Inter.grf
- Contient : Aucun
- Appel au dictionnaire :
<JOUR> (dico)

1.27 TIMEX3_DATE_Num_DDMMt.grf

- Annotation : Date Mois (*22-07*)
- Contenu dans :
TIMEX3_DATE_Num.grf
TIMEX3_DATE_Enum.grf
TIMEX3_DATE_Ancre.grf
Inter_anchor.grf
Inter.grf

- Contient : Aucun
- Appel au dictionnaire :
  <JOUR> (dico)

1.28 TIMEX3_TIME_HH.grf

- Annotation : Heure (*14h30*)
- Contenu dans : TIMEX3_TIME.grf
- Contient : V_Heure
- Appel au dictionnaire : Aucun
- Note : Différenciation "à midi/le midi"

1.29 TIMEX3_TIME_HP.grf

- Annotation : Période horaire (*l'après-midi*)
- Contenu dans : TIMEX3_TIME.grf
- Contient : V_Modifieurs
- Appel au dictionnaire :
  <matinée> (dico)
  <matin> (dico)
  <soirée> (dico)
  <soir> (dico)
  <journée> (dico)
  <après midi> (dico)
  <nuit> (dico)
  <midi> (dico)
- Note : Différenciation "à midi/le midi"

1.30 TIMEX3_INTER_DATE.grf

- Annotation : Intervalle de dates (*du lundi au mercredi*)
- Contenu dans : TIMEX3_INTER.grf
- Contient :
  TIMEX3_INTER_DATEc.grf
  TIMEX3_INTER_DATEg.grf
  TIMEX3_INTER_DATEd.grf
- Appel au dictionnaire : Aucun

1.31 TIMEX3_INTER_TIME.grf

- Annotation : Intervalle d'heures (*de 14h à 18h*)
- Contenu dans : TIMEX3_INTER.grf

- Contient :
  TIMEX3_INTER_TIMEc.grf
  TIMEX3_INTER_TIMEg.grf
  TIMEX3_INTER_TIMEd.grf
- Appel au dictionnaire : Aucun

1.32 V_Alph_Periodes.grf

- Annotation : Intervalle non-borné (*pendant les vacances scolaires*)
- Contenu dans :
  TIMEX3_INTER.grf
  TIMEX3_INTER_DATEg.grf
  TIMEX3_INTER_DATEd.grf
- Contient : V_Modifieurs.grf
- Appel au dictionnaire :
  <FF> (dico)
  <SAISON> (dico)

1.33 TIMEX3_INTER_DATEc.grf

- Annotation : Intervalle de dates complet (*du lundi au mercredi*)
- Contenu dans : TIMEX3_INTER_DATE.grf
- Contient :
  Inter_anchor.grf
  Inter.grf
- Appel au dictionnaire : Aucun

1.34 TIMEX3_INTER_DATEg.grf

- Annotation : Intervalle de dates borné à gauche uniquement (*à partir du 22 juillet*)
- Contenu dans : TIMEX3_INTER_DATE.grf
- Contient :
  TIMEX3_DATE
  V_Alph_Periodes
- Appel au dictionnaire : Aucun

1.35 TIMEX3_INTER_DATEd.grf

- Annotation : Intervalle de dates borné à droite uniquement (*jusqu'au 22 juillet*)
- Contenu dans : TIMEX3_INTER_DATE.grf

- Contient :
  TIMEX3_DATE
  V_Alph_Periodes
- Appel au dictionnaire : Aucun

1.36 TIMEX3_INTER_TIMEc.grf

- Annotation : Intervalle d'heures complet (*14h-18h*)
- Contenu dans : TIMEX3_INTER_TIME.grf
- Contient :
  TIMEX3_TIME.grf
  V_BorneG.grf
  V_BorneD.grf
- Appel au dictionnaire : Aucun

1.37 TIMEX3_INTER_TIMEg.grf

- Annotation : Intervalle d'heures borné à gauche uniquement (*à partir de 14h*)
- Contenu dans : TIMEX3_INTER_TIME.grf
- Contient :
  TIMEX3_TIME.grf
- Appel au dictionnaire : Aucun

1.38 TIMEX3_INTER_TIMEd.grf

- Annotation : Intervalle d'heures borné à droite uniquement (*jusqu'à 18h*)
- Contenu dans : TIMEX3_INTER_TIME.grf
- Contient :
  TIMEX3_TIME.grf
- Appel au dictionnaire : Aucun

1.39 Inter_anchor.grf

- Annotation : Intervalle de dates complet avec un anchorID [1] (*du 22 au 23 juillet 2014*)
- Contenu dans : TIMEX3_INTER_DATEc.grf
- Contient :
  TIMEX3_DATE_Alph_SSAAAA
  TIMEX3_DATE_Alph_SS

1. Attribut TimeML : référence vers une date plus complète pour complétion avec une temporalFunction

TIMEX3_DATE_Alph_MMAAAA
TIMEX3_DATE_Alph_MM
TIMEX3_DATE_Alph_DDMMAAAA
TIMEX3_DATE_Alph_DDMM
TIMEX3_DATE_Alph_DD
TIMEX3_DATE_Num_DDMMAAAAs
TIMEX3_DATE_Num_DDMMAAAAt
TIMEX3_DATE_Num_DDMMs
TIMEX3_DATE_Num_DDMMt
V_BorneG.grf
V_BorneD.grf
V_Modifieurs
- Appel au dictionnaire : <JOUR> (dico)

1.40 Inter.grf

- Annotation : Intervalle de dates complet sans un anchorID (*du 22 juillet au 15 août*)
- Contenu dans : TIMEX3_INTER_DATEc.grf
- Contient :
  TIMEX3_DATE_Alph_SSAAAA
  TIMEX3_DATE_Alph_SS
  TIMEX3_DATE_Alph_MMAAAA
  TIMEX3_DATE_Alph_MM
  TIMEX3_DATE_Alph_FF
  TIMEX3_DATE_Alph_DDMMAAAA
  TIMEX3_DATE_Alph_DDMM
  TIMEX3_DATE_Alph_DD
  TIMEX3_DATE_Alph_JJ
  TIMEX3_DATE_Num_DDMMAAAAs
  TIMEX3_DATE_Num_DDMMAAAAt
  TIMEX3_DATE_Num_DDMMs
  TIMEX3_DATE_Num_DDMMt
  V_Alph_Periodes
  V_BorneG.grf
  V_BorneD.grf
  V_Modifieurs
- Appel au dictionnaire : <JOUR> (dico)

1.41 TIMEX3_SET_DATE.grf

- Annotation : Ensemble de dates (*tous les 8 mai*)
- Contenu dans : TIMEX3_SET.grf

- Contient :
  TIMEX3_DATE_Alph_SSAAAA
  TIMEX3_SET_SS
  TIMEX3_SET_MM
  TIMEX3_SET_WE
  TIMEX3_DATE_Alph_FF
  TIMEX3_DATE_Alph_DDMM
  TIMEX3_DATE_Alph_DD
  TIMEX3_DATE_Alph_OD
  V_Quant.grf
  V_Freq.grf
- Appel au dictionnaire : <N>
- Note : Présence d'un contexte négatif droit

1.42 TIMEX3_SET_TIME.grf

- Annotation : Ensemble d'heures (*tous les soirs*)
- Contenu dans : TIMEX3_SET.grf
- Contient :
  V_Quant.grf
  V_Freq.grf
  V_Modifieurs
- Appel au dictionnaire :
  <matin>
  <midi>
  <après midi>
  <soir>
  <nuit>

1.43 TIMEX3_SET_SS.grf

- Annotation : Ensemble de saison (*tous les étés*)
- Contenu dans : TIMEX3_SET_DATE.grf
- Contient :
  TIMEX3_DATE_Alph_SS
  V_Quant.grf
  V_Freq.grf
  V_Modifieurs.grf
- Appel au dictionnaire : Aucun

1.44 TIMEX3_SET_MM.grf

- Annotation : Ensemble de mois (*tous les mois de juillet*)

- Contenu dans : TIMEX3_SET_DATE.grf
- Contient :
  TIMEX3_DATE_Alph_MM
  V_Quant.grf
  V_Freq.grf
  V_Modifieurs.grf
- Appel au dictionnaire : Aucun

1.45 TIMEX3_SET_WE.grf

- Annotation : Ensemble de semaines, week-ends, ou jours (*tous les samedis*)
- Contenu dans : TIMEX3_SET_DATE.grf
- Contient :
  TIMEX3_DATE_Alph_WE
  TIMEX3_SET_JJ
  V_Quant.grf
  V_Freq.grf
  V_Modifieurs.grf
- Appel au dictionnaire : Aucun

1.46 V_Modifieurs.grf

- Annotation : Marqueurs modifiant une expression temporelle
- Contenu dans :
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.47 V_BorneG.grf

- Annotation : Aucune
- But : Marqueurs de borne gauche dans un intervalle
- Contenu dans :
  TIMEX3_INTER_TIMEc.grf
  Inter_anchor.grf
  Inter.grf
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.48 V_BorneD.grf

- Annotation : Aucune
- But : Marqueurs de borne droite dans un intervalle

- Contenu dans :
  TIMEX3_INTER_TIMEc.grf
  Inter_anchor.grf
  Inter.grf
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.49 V_Quant.grf

- Annotation : Marqueurs de quantification pour les ensembles (*tous les ...*, *chaque ...*)
- Contenu dans :
  TIMEX3_SET_DATE
  TIMEX3_SET_TIME
  TIMEX3_SET_SS
  TIMEX3_SET_MM
  TIMEX3_SET_WE
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.50 V_Freq.grf

- Annotation : Marqueurs de fréquence pour les ensembles (*2 fois par ...*)
- Contenu dans :
  TIMEX3_SET_DATE
  TIMEX3_SET_TIME
  TIMEX3_SET_SS
  TIMEX3_SET_MM
  TIMEX3_SET_WE
- Contient :
  TIMEX3_DATE_Alph_WE
  TIMEX3_DATE_Alph_JJ
  V_AlphNum_Chiffres
- Appel au dictionnaire : Aucun

1.51 V_Ordinaux.grf

- Annotation : Aucune
- But : Ordinaux (*2ème ...*)
- Contenu dans :
  TIMEX3_DATE_Enum.grf
  TIMEX3_DATE_Ancre.grf

TIMEX3_DATE_Alph_OD.grf
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.52 V_Alph_Date.grf

- Annotation : Aucune
- But : Date (*30 ...*)
- Contenu dans :
  TIMEX3_DATE_Alph_DDMMAAAA.grf
  TIMEX3_DATE_Alph_DDMM.grf
  TIMEX3_DATE_Alph_DD
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.53 V_Alph_Annee.grf

- Annotation : Aucune
- But : Date (*... 2014*)
- Contenu dans :
  TIMEX3_DURATION.grf
  TIMEX3_DATE_Alph_SSAAAA
  TIMEX3_DATE_Alph_MMAAAA
  TIMEX3_DATE_Alph_DDMMAAAA.grf
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.54 V_Heure.grf

- Annotation : Aucune
- But : Heure (*15 :15*)
- Contenu dans : TIMEX3_TIME_HH.grf
- Contient : Aucun
- Appel au dictionnaire : Aucun

1.55 V_Duree.grf

- Annotation : Durée (*2 jours ...*)
- Contenu dans : TIMEX3_DURATION.grf
- Contient : V_AlphNum_Chiffres
- Appel au dictionnaire :
  <SE>
  <WE>
  <JOUR>

2 Passe 2 : Ajout des liens

- Présence de contextes négatifs gauche et droite
- Pas d'appel au dictionnaire
- Pas d'appel de sous-graphe

2.1 1_date-occ-n.grf

- Premier élément : Date, intervalle de dates ou ensemble de dates
- Deuxième élément : Occurence
- Type de lien : IDENTITY

2.2 2_date-time-n.grf

- Premier élément : Date, intervalle de dates ou ensemble de dates
- Deuxième élément : Heure, intervalle d'heures ou ensemble d'heures
- Type de lien : IS_INCLUDED

2.3 3_occ-time-n.grf

- Premier élément : Occurence ou objet touristique
- Deuxième élément : Heure, intervalle d'heures ou ensemble d'heures
- Type de lien :
  BEGUN_BY pour une heure
  IDENTITY pour un intervalle d'heures ou un ensemble d'heures

2.4 4_date-date_xor-n.grf

- Premier élément : Date ou intervalle de dates
- Deuxième élément : Date ou intervalle de dates
- Type de lien : XOR

2.5 5_mod.grf

- Premier élément : Occurrence ou objet touristique
- Deuxième élément : Modalité
- Type de lien : CONDITIONAL

2.6 5_mod-mod.grf

- Premier élément : Modalité
- Deuxième élément : Modalité
- Type de lien : CONDITIONAL

2.7 5_time-mod-n.grf

- Premier élément : Heure, intervalle d'heures ou ensemble d'heures
- Deuxième élément : Modalité
- Type de lien : IDENTITY

2.8 6_date-date_ancre-n.grf

- Premier élément : Date
- Deuxième élément : Date
- Type de lien : BEFORE, AFTER, ou AROUND

2.9 6_duree-time_ancre-n.grf

- Premier élément : Durée
- Deuxième élément : Heure
- Type de lien : BEFORE, AFTER, ou AROUND

2.10 7_date-date_excluded-n.grf

- Premier élément : Date, intervalle de dates ou ensemble de dates
- Deuxième élément : Date ou heure
- Type de lien : IS_EXCLUDED

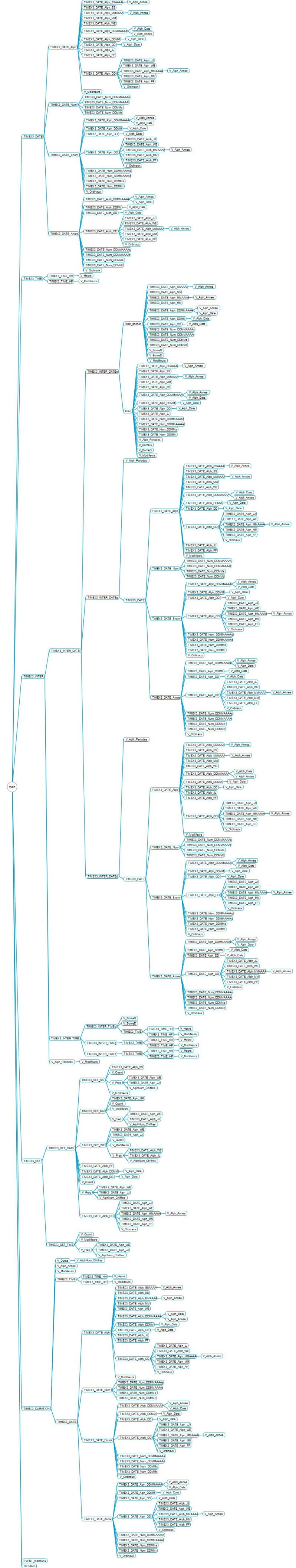
2.11 8_oi-event.grf

- Premier élément : Objet touristique
- Deuxième élément : Occurrence
- Type de lien : IDENTITY

Annexe D

Carte des graphes

Voici la carte des graphes composant la grammaire. Elle permet de visualiser les relations entre les différents graphes.
De part sa taille, elle n'est lisible que sur un écran. Cependant, cela permet de visualiser l'étendue que la grammaire a atteinte à l'issue du stage.



Annexe E

Résultats faussés sous Diff Tool

Les trois exemples d'annotations ci-dessous sont tous corrects mais nous pouvons voir que la différence dans les index de caractères peut fausser l'interprétation de Diff Tool. Le premier exemple est considéré comme correct, le second comme partiellement correct et le troisième comme incorrect avec une annotation manquante (couleur rouge) et l'autre faux-positif (couleur jaune)

6672	6677	matin	{value=TMO, type=TIME, tid=t1}	=	6672	6677	matin	{value=TMO, type=TIME, tid=t1}
27455	27467	18·septembre	{valueFromFunction=tf1, temporalFunction=true, value=XXXX-XX-XX, type=DATE, tid=t1}	~	27458	27470	18·septembre	{valueFromFunction=tf1, temporalFunction=true, value=XXXX-XX-XX, type=DATE, tid=t1}
39004	39006	7h	{value=TXX:XX, type=TIME, tid=t1}	-?				
				?-	39010	39012	7h	{value=TXX:XX, type=TIME, tid=t1}