Conventions d'annotations en Entités Nommées
- ESTER -

Céline Le Meur - Sylvain Galliano - Edouard Geoffrois

30 juillet 2004

Table des matières

1 Préambule

1.1 Remerciements

Ce document s'appuie sur les conventions d'annotation utilisées dans le cadre des évaluations du NIST[1] pour l'anglais, sur un manuel de conventions d'annotations pour le français rédigé par S. Rosset (LIMSI) et M.Vergnes (VECSYS), et sur les différentes remarques apportées par François Yvon, Frédéric Béchet et les membres de la campagne ESTER.

1.2 Objet du document

Ce document décrit les conventions adoptées pour l'annotation manuelle en entités nommées d'enregistrements audio de journaux télévisés et radiophoniques en français, déjà transcrits. Ces annotations serviront au développement de systèmes automatiques de détection d'entités nommées, que ce soit pour leur mise en relief pour le lecteur ou pour leur exploitation dans d'autres systèmes automatiques d'indexation et d'analyse du contenu de documents (comme la veille, les résumés automatiques, etc.).

Les performances des systèmes automatiques dépendent directement de la qualité et de la cohérence des annotations manuelles. Par conséquent, ce document tente de référencer des conventions à la fois les plus complètes possibles, simples à mémoriser et à appliquer.

Cependant, la notion d'entité nommée est un concept riche et complexe, qui n'est pas réellement figé. C'est la raison pour laquelle ce document sera voué à évoluer afin de prendre en compte de nouveaux cas qui ne seraient pas couverts par la version actuelle. Un tag spécifique est prévu à cet effet [ent=unk] (incertain), ce qui nous permettra par la suite de revenir sur les hésitations rencontrées et d'y pallier.

1.3 Organisation des fichiers

Les fichiers à annoter sont des transcriptions (automatiques ou manuelles) d'enregistrements d'émissions télévisées ou radiophoniques. Ils sont fournis sur CD-Rom. Il s'agit de fichiers au format Transcriber.

Après l'annotation en entités nommées d'un fichier original fichier.trs, le fichier sera sauvegardé sous la forme : fichier_NE.trs.

[1] http ://www.ldc.upenn.edu/Projects/ACE/docs/EnglishEDTV4-2-6.PDF

2 Introduction

2.1 Définition d'une entité nommée (EN)

Avant de présenter les différents types d'EN retenus pour annoter les corpus de transcription, il convient de définir la notion d'EN. Même s'il n'existe pas de définition standard, on peut dire que les EN sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme).

Une entité a généralement une existence relativement stable dans le temps, même si cette existence a un début (naissance, fondation, dépot, formation...) et une fin (mort, dissolution, faillite, disparition...) et si l'entité évolue entre temps. Pour appréhender plus simplement cette notion, on s'appuiera de manière générale sur le "principe du catalogue" pour savoir si on a affaire à une EN. Ainsi si on peut aisément imaginer l'EN supposée comme étant une entrée d'un catalogue, annuaire, dictionnaire ou index alors celle-ci sera bien une EN.

Les entités sont au cœur de la problèmatique de l'extraction de l'information d'un document. Par extension, on annotera les dates et les grandeurs physiques.

2.2 Les différentes mentions d'EN

Selon la définition qui vient d'être énoncée, une entité est un terme générique qui cherche à définir certaines unités lexicales. Il est possible de rencontrer au sein d'un document plusieurs mentions qui font référence à une seule et même entité. Voici les différents types de mentions :

1. l'entité de référence, qui est celle que l'on s'attendrait à trouver dans un catalogue, par exemple : **"Jacques Chirac"**, **"France"**, **"SNCF"**, etc.
2. la dénomination partielle (ellipse), par exemple : "**2** ou **3 euros**", où le "**2**" est une ellipse de l'EN "**2 euros**", etc.
3. la dénomination complétée par ajout de précision, par exemple : "le président **Jacques Chirac**", "la fusée **Ariane**", etc
4. le surnom, par exemple : "**Zizou**", "la **Dame de fer**", "le **Baron rouge**", etc.
5. l'expression référentiellement autonome, faisant référence à l'EN sans la nommer, par exemple : "**le président français**", "**le leader des Festina**", etc.

6. l'anaphore, qui est une unité de la langue qui ne peut être interprétée que lorsqu'elle est mise en relation avec un antécédent, présent dans la phrase, et qui lui permet d'acquérir une référence (comme les pronoms personnels, démonstratifs, etc.), par exemple : "**celui-ci**", "**il** a aussi déclaré", etc.

Seuls les cas de figures où l'entité est nommée (cas 1 à 4 ci-dessus) sont annotés. La seule exception à cette règle concerne les dates relatives (cette **année**, **hier**, etc.) car tous les substantifs de ce type (explicites ou non) vont être annotés.

Entités	**Mentions annotées**	**Mentions non annotées**
[Corse]	la **[Corse]**..., l'**[île de beauté]**..., etc.	celle-ci à été le théâtre..., elle est au centre des discussions..., etc.
[Jacques Chirac]	le président la république **[Jacques Chirac]**..., monsieur **[Jacques Chirac]**..., **[Chirac]**..., etc.	ce dernier a dit..., il est aujourd'hui en voyage..., etc.
[22 juin 2004]	depuis **[aujourd'hui]**..., ce **[jour]**..., etc.	aucune

2.3 Représentation hiérarchique des EN

Le repérage des EN (tels que les noms de personnes, de lieux, etc.) pose problème dans la mesure où ces noms constituent un ensemble ouvert, par définition impossible à recenser de manière exhaustive, d'autant qu'ils sont sujets à variation.

Pour limiter cette difficulté et faciliter ce repérage, les EN sont classées par catégories, elles-mêmes divisées en sous-catégories, ceci dans le but de dégager des ensembles suffisament précis, qui ne porteront pas à confusion lors de l'annotation. Ainsi la classification réalisée se présente sous la forme d'une arborescence à hiérarchisation descendante, composée de macro et micro classes (autrement dit de classes et de sous-classes) que nous présentons par la suite.

De façon générale, on utilisera les sous-catégories définies par la suite pour annoter une EN donnée, et ce à chaque fois que cela paraîtra évident (ex. l'[**UMP**] est une organisation politique [org.pol]). En revanche, lors d'hésitation sur le choix d'une sous-catégorie on devra utiliser alors la ca-

tégorie générale de l'EN correspondante : par exemple [**Meteo France**] peut être considérée soit comme une organisation non commerciale ([org.non-profit]) en tant que service public, soit comme une organisation commerciale ([org.com]) puisqu'elle vend certains services, dans ce cas on utilisera la catégorie générale [org]. Enfin on utilisera l'étiquette [ent=unk] si on pense avoir affaire à une EN mais que l'on n'a aucune idée de la catégorie à choisir ou si tout simplement aucune ne correspond.

Quelques soient les catégories retenues et le soin apporté à la définition de leur contour, il restera inévitablement un faible pourcentage de cas limites difficiles à étiqueter et sur lesquels on pourra hésiter (voir section 3.9, page 23). En conséquence, afin de participer à l'amélioration des présentes conventions, nous sollicitons, chers annotateurs, votre collaboration. Il vous suffira pour cela d'indiquer dans un fichier texte les difficultés rencontrées sur une EN donnée, en précisant brièvement les raisons pour lesquelles la représentation choisie ne convient pas, avec le nom du fichier correspondant. Toutes vos remarques seront ensuite analysées et prises en compte dans les futures versions de ce manuel.

2.3.1 Présentation de l'arborescence

Nous distinguons 8 classes d'entités nommées (la dernière classe [ent=unk] n'étant pas considérée comme telle) dont voici une représentation générale :

1. pers
   - pers.hum
   - pers.anim
   - pers.imag
2. org
   - org.pol
   - org.edu
   - org.non-profit
   - org.com
3. gsp
   - gsp.pers
   - gsp.org
   - gsp.loc
4. loc
   - loc.geo
     - loc.geo.line
   - loc.addr
     - loc.addr.post
     - loc.addr.tel
     - loc.addr.elec
5. fac
6. prod
   - prod.vehicule
   - prod.award
   - prod.art
   - prod.printing
7. time
   - time.date
     - time.date.abs
     - time.date.rel
   - time.hour
8. amount
   - amount.phy
     - amount.phy.age
     - amount.phy.dur
     - amount.phy.temp
     - amount.phy.len
     - amount.phy.wei
     - amount.phy.spd
   - amount.cur
9. unk

2.3.2 Présentation des catégories retenues

Le tableau qui va suivre présente les EN que nous avons retenues de façon très succincte. Ceci est réalisé dans le but d'offrir une vision globale des étiquettes à utiliser. Chaque EN sera présentée de façon plus détaillée dans la troisième partie de ce manuel (voir chapitre 3 page 15).

TAB. 1 – Tableau récapitulatif

Étiquettes	**Description**	**Exemples**
pers	Cette classe regroupe tous les noms de personnes et animaux réels ou fictifs. Les surnoms sont annotés si et seulement s'ils ne sont pas ambigus (exemple : **[Zizou]**).	
pers.hum	Les personnes réelles (nom, prénom, surnom, familles, couples, etc.)	**[Richard Virenque]** a gagné..., la famille **[Chirac]**..., etc.
pers.anim	Les animaux réels	**[Mabrouk]**..., mon chien **[Médor]**..., etc.
pers.imag	Les personnes et les animaux fictifs (romans, films, dessins animés, etc.)	**[Superman]**..., **[Rambo]**..., **[Mickey]**..., **[ET]**..., etc.
org	Cette classe regroupe tous les noms, mots, groupes de mots, sigles, etc. faisant référence à des entreprises, des sociétés, des organisations.	
org.pol	Les organisations politiques	le **[parti communiste français]**..., etc.
org.edu	Les organisations éducatives	l'**[ENSAM]**..., l'**[École Polytechnique]**..., etc.
org.non-profit	Les organisations qui cherchent à promouvoir des services, des idées, des produits au profit de l'organisation elle-même et non par rapport à l'argent qu'elle pourrait en tirer	la **[JOC]** qui est la **[Jeunesse Ouvrière Chrétienne]**..., **[la chaîne de l'espoir]**..., l'**[ANPE]**..., l'**[Unedic]**..., l'**[Unicef]**..., etc.

org.com	Les organisations commerciales	des groupes tels **[Les Beatles]**, **[Dire Straits]**..., l'ombre des **[All Blacks]** se profilent..., l'équipe **[Festina]** ..., **[France Inter]**..., aux **[éditions Somogui]**..., **[Le Monde]**..., la **[SNCF]**..., **[Thales]**..., **[SAGEM]**..., etc.
gsp	Cette classe regroupe tous les mots ou groupes de mots employés indifféremment pour référer à la fois à une région administrative, son peuple, son gouvernement.	
gsp.pers	Les groupes géo-socio-politiques lorsqu'ils font référence à leurs habitants, leur population	...les habitants du nord et le reste de la **[France]**...
gsp.org	Les groupes géo-socio-politiques lorsqu'ils sont envisagés en tant qu'organisation	la **[France]** a signé un accord avec les Etats-Unis..., etc.
gsp.loc	Les groupes géo-socio-politiques lorsqu'ils sont perçus comme un lieu	Ils se sont retrouvés en **[France]**..., etc.
loc	Cette classe regroupe tous les noms de lieux naturels (fleuves, montagnes, rivières, continents, systèmes solaires) et bâtis par des humains (chemins ferroviaires, tunnels, ponts, rues et boulevards, etc.), ainsi que les adresses postales, téléphoniques, électroniques et fax.	
loc.geo	Les lieux géographiques naturels	Le **[Mont-Blanc]**..., la **[cordillère des Andes]**..., etc.
loc.geo.line	Les axes de circulation (chemins ferroviaires, tunnels, ponts, noms de rues et boulevards, etc.) lorsqu'ils sont envisagés comme des lieux géographiques	le musette [...] né de la rencontre de ces immigrés italiens avec les Auvergnats , c'était **[rue de Lappe]** à ..., vous pouvez notamment appeler une cabine sur la **[5ème avenue]** à New-York..., un autre accident à l'est de Paris sur l'**[autoroute A4]**..., etc.

loc.addr.post	Les adresses postales	au **[3 avenue de Matignon 75008 Paris]**..., etc.
loc.addr.tel	Les numéros de téléphone et de fax	c'est le **[01 45 65 90 99]**..., etc.
loc.addr.elec	les adresses électroniques	**[www.telecharger.com]**..., etc.
fac	Les entités qui se limitent aux bâtiments et autres constructions humaines (musées, aéroports, hôpitaux, etc.)	...se transfère au **[palais de Chaillot]**..., etc.
prod	Cette classe regroupe tous les mots ou groupes de mots faisant référence à des moyens de transport (voitures, trains, navettes, avions, etc.), des œuvres de nature artistique (peinture, sculpture, danse, programmes télévisés, films, musiques, etc.), des œuvres de nature intellectuelle (romans, bandes dessinées, journaux/magazines, etc.), à des récompenses (oscars, etc.)	
prod.vehicule	Les moyens de transport	la fusée **[Ariane]**..., la navette **[Endeavour]**..., le **[boeing 747]**..., etc.
prod.award	Les récompenses	le **[prix nobel de la paix]** est décerné à ...,
prod.art	Les oeuvres artistiques	l'émission **[le téléphone sonne]**..., la comédie musicale **[Roméo et Juliette]**..., etc.
prod.printing	Les oeuvres littéraires	un article du journal **[Le Monde]**..., **[Les Fleurs du Mal]**..., etc.
time	Cette classe regroupe toutes les expressions faisant référence à une date absolue ou relative.	
time.date.abs	Les dates absolues	Le **[11 septembre 2001]** deux avions..., etc.
time.date.rel	Les dates relatives	c'était **[hier]** à..., etc.
time.hour	Les heures indiquant un instant précis dans le temps et non une durée	sur le coup de **[4 heures du matin]**..., etc.

amount	Cette classe regroupe tous les mots, groupes de mots faisant référence à un montant, une valeur, une mesure de poids, de taille, de distance, de durée, etc.	
amount.phy.age	Les âges	ce monsieur a **[87 ans]**..., etc.
amount.phy.dur	Les durées	pendant **[3 semaines]**..., etc.
amount.phy.temp	Les températures	il fait **[moins 20 degrés]**..., etc.
amount.phy.len	Tout ce qui fait référence à une distance (hauteur, longueur, etc.)	il mesure **[1 mètres 85]**..., il a couru pendant **[25 kilomètres]**..., etc.
amount.phy.wei	Les poids	et pèse **[70 kilos]**..., etc.
amount.phy.spd	Les vitesses	le mistral approchera **[60]** **[80 Km/h]**..., etc.
amount.cur	Les valeurs monétaires	**[7,5 euros]**..., etc.

Les phénomènes de métonymie susceptibles d'être rencontrés ne sont ni représentés dans l'arborescence, ni dans le tableau récapitulatif car ils ne constituent pas une sous-classe des catégories présentées. Cependant, on les mentionnera par la suite, quand on en viendra à préciser les classes distinguées. Il est fort probable que tous les cas de métonymie ne soient pas évoqués. Pour pallier ce manque, vous utiliserez la balise [ent=unk] prévue à cet effet et vous pourrez si vous le souhaitez proposer des solutions dans un fichier texte indépendant.

2.4 Généralités

2.4.1 Format des annotations

Les entités nommées seront représentées de la façon suivante dans l'interface de Transcriber :

[ent= ? ?-]**entité**[-ent= ? ?]

Pour information, le codage au format XML correspondant est :

<Event desc =" ? ?" type="entities" extent="begin">**entity**<Event desc=" ? ?" type="entities" extent="end">

où la valeur de "**desc**" contient une description précise de l'entité (son type)[2].

Remarque :

[2]Pour des raisons de compatibilité logicielle, il a été décidé d'utiliser le champ de description d'un évènement (**desc**) déjà présent dans l'interface de Transcriber afin de préciser le type de l'entité à annoter. Il est prévu à l'avenir de créer un champ spécifique "type" pour introduire les différentes catégories d'EN.

De façon générale et pour des questions de lisibilité, les entités nommées des exemples de ce document seront représentées dans des tableaux, inscrites en **gras** et entre crochets ([]) sans les balises de description. Ces dernières seront représentées indépendamment des EN.

2.4.2 Le champ description (ent)

Ce champ représente le type de l'entité. Sa valeur contient donc des informations concernant la catégorie d'appartenance de l'entité (catégories ou sous-catégories) ces dernières ayant pour but de préciser la catégorie générale. Dans ce cas, les sous-catégories seront séparées par des points ".", par exemple :

desc=pers.hum : représente une personne (catégorie de base) humaine (sous-catégorie de personne) et réelle.

2.4.3 Étendue de l'étiquette

De façon générale, nous pouvons dire que seule l'entité sera prise en compte dans l'étiquette. Toutes les particules (déterminants, articles, adverbes, étiquettes statutaires, etc.) sont exclues de l'annotation (ex. 1) sauf si elles font partie intégrante de l'entité à annoter (ex. 2) :

N°	Exemples
1.	**[Jacques Chirac]**..., Les **[éditions Somogui]**..., Le **[Palais Bourbon]**..., la **[Nouvelle Star]**..., plus de **[20 degrés]**, la marionnette **[Virenque]**, la navette **[Endeavour]**, le frère de **[Richard Virenque]**..., etc.
2.	**[Les Beatles]**..., **[Les Rolling Stones]**..., **[Le Monde]**..., **[Le Figaro]**..., **[La Mecque]**..., **[Le téléphone sonne]**...,**[moins 20 degrés celsius]**..., etc.

Une balise d'évènement peut être greffée à certaines EN. Dans ce cas l'étiquette devra englober l'ensemble composé de l'EN et de sa balise, par exemple :

N°	Exemple	Étiquettes utilisées
1.	la fusée **[Ariane V+[pron=(chiffres romains :) 5]]** a décollé...	[ent=prod.vehicule]

2.4.4 Composition des entités nommées

Des entités composites peuvent être crées à partir de combinaisons d'entités élémentaires :

– EN coordonnées et ellipsées :

N°	Exemples	Commentaires
1.	le **[littoral méditerranéen]** et **[breton]**	mis pour le **[littoral méditerranéen]** et le **[littoral breton]** : ici les deux entités sont coordonnées et la seconde est ellipsée, elles sont annotées de la façon suivante : [ent=loc.geo] puis [ent=loc.geo]. Il s'agit ici d'un cas d'ellipse partielle.
2.	de **[3]** à **[7 degrés]**	mis pour de **[3 degrés]** à **[7 degrés]** : cas similaire à l'exemple 1 avec la première entité ellipsée.
3.	M et Mme **[Chirac]**	mis pour M **[Chirac]** et Mme **[Chirac]**. On a bien deux entités de type personne, le nom de la seconde étant ellipsé. On pourrait donc l'annoter de la même manière que les deux autres exemples cités ci-dessus à savoir : [ent=pers.hum] puis [ent=pers.hum]. Comme les étiquettes statutaires ne sont pas prises en compte, seul le nom propre sera annoté, c'est-à-dire **[Chirac]** (annoté [ent=pers.hum]). Nous considérons ici que l'ellipse est totale.

– imbrication d'EN :

Il y a des cas où des EN de référence (rappel des différentes mentions d'EN possibles, section 2.2) peuvent contenir d'autres EN de référence : il s'agit là d'un phénomène d'imbrication. Pour savoir dans quels cas il faut annoter ou non le type de chacune des EN imbriquées, la règle suivante sera utilisée :

1. si le type de chacune des EN est différent alors chacun des types est annoté, et les EN imbriquées constituent une seule et même entité, possédant un type qui lui est propre (ex. 1) ;
2. en revanche, si le type de chacune des entités est semblable, alors ils ne sont pas tous annotés, les EN imbriquées forment une entité englobante dont le type général correspondra au type des EN qui la constituent (ex. 2).

N°	Annotation correcte	Commentaires
1.	l'**[Université de [Corte]]** est en grève...	l'imbrication est justifiée car le type global de l'EN **[université de Corte]** est ici [ent=org.edu/gsp.pers] (puisque d'aprés le contexte on entre dans un cas de métonymie) et celui de l'EN imbriquée **Corte** est [ent=gsp.loc]) (puisque considérée comme une région administrative).
2.	le **[12 juillet 1998]**...	l'imbrication de **[12 juillet]** et/ou **[1998]** n'est pas justifiée car le type des deux EN est ici identique [ent=time.date.abs], on a donc une seule et même entité englobante de ce type.

2.4.5 Particularités de l'oral

Les fichiers que nous annotons ont la particularité d'appartenir au domaine de l'oral. Aussi, il est possible de rencontrer des traits propres à ce domaine au sein de ces mêmes fichiers (tels que les hésitations, les bégaiements, les répétitions, les reprises, etc.) et ceci peut poser problème lors de l'annotation dans la mesure où l'on peut se demander s'il faut intégrer ces phénomènes dans l'étiquette choisie ou non.

Les règles sont donc les suivantes :

1. toutes les mentions d'EN possédant des phénomènes liés à l'oral sont annotées ;
2. tous les signes particuliers de ces phénomènes sont à intégrer dans l'étiquette. Ainsi la troncature (mentionnée entre parenthèses) d'un mot tronqué (ex. 1 à 4), l'astérisque d'un mot mal prononcée (ex. 5) sont annotés. En revanche, toutes les hésitations (telles que "hum hum", "euh", etc.) ne sont pas à intégrer dans l'étiquette.

N°	Exemples	Étiquettes utilisées
1.	**[Mi ()]** euh je voulais dire **[Michel]**...	[ent=pers.hum] puis [ent=pers.hum]
2.	**[Ja()]** pardon **[Jacques Chirac]** a procédé à une élocution...	[ent=pers.hum] puis [ent=pers.hum]
3	et puis dans le **[neu(vième)] [neuvième arrondissement]** de **[Paris]**...	[ent=gsp.loc] puis [ent=gsp.loc] et [ent=gsp.loc]
4	je vous présente le président de l'**[a(viation)]** de l'**[aviation sans frontières]**...	[ent=org.non-profit] puis [ent=org.non-profit]
5	à **[*Londres]**...	[ent=gsp.loc]

2.4.6 Les phénomènes de métonymie

La métonymie est une figure de style par laquelle on tente d'exprimer un concept au moyen d'un terme désignant un autre concept qui lui est uni par une relation nécessaire : la cause pour l'effet, le contenant pour le contenu, le lieu ou le producteur pour la production, etc.

Ce phénomène se retrouve dans les corpus annotés. Afin de l'indiquer, la catégorie de base à laquelle appartient le mot sera suivi d'un slash "/" et de la catégorie issue du contexte, par exemple :

N°	Exemple	Étiquette correcte	Étiquette incorrecte
1.	plusieurs **[prix Nobel de la Paix]** ont été convié à ...	[ent=prod.award/pers.hum]	[ent=prod/pers]

Attention : Veillez à préciser la sous-catégorie à chaque fois que cela est possible.

3 Description des entités nommées

Nous nous attachons ici à présenter les entités nommées que nous avons retenues et la façon dont nous allons les annoter.

3.1 Les personnes (pers)

Tous les mots ou expressions désignant des personnes ou animaux, qu'ils soient réels ou fictifs (roman, film, bande dessinée, etc.), seront annotés.

On distingue trois sous-classes :

1. **pers.hum** : toutes les personnes réelles (nom, prénom, surmom, couples, familles, ...) (ex. 1 à 6) ;
2. **pers.anim** : tous les animaux réels (ex. 7) ;
3. **pers.imag** : tous les personnages et animaux fictifs (ex. 8 à 10).

N°	**Exemples**	**Étiquettes utilisées**
1.	Le président français **[Jacques Chirac]** sera l'invité d'honneur...	[ent=pers.hum]
2.	le sécrétaire d'état américain **[Madeleine Albright]** sera aussi présent...	[ent=pers.hum]
3.	Moi je vous dis Monsieur **[Fillon]**...	[ent=pers.hum]
4.	Monsieur et Madame **[Chirac]** se sont rendus...	[ent=pers.hum]
5.	une chanson en l'honneur de **[Zizou]**...	[ent=pers.hum]
6.	les **[Bush]** sont en voyage...	[ent=pers.hum]
7.	le célèbre chien **[Mabrouk]**...	[ent=pers.anim]
8.	On peut y référer l'irréalisme de **[Don Quichotte]**...	[ent=pers.imag]
9.	**[Félix le chat]**...	[ent=pers.imag]
10.	**[Atakou]** qui se révèlera être le sauveur et le héros de cette histoire...	[ent=pers.imag]
N°	**Contre-Exemples**	**Étiquettes utilisées**
11.	le préfet de Corse était abattu en pleine rue à Ajaccio...	aucune étiquette
12.	le dalaï-lama a mis fin lui-même à la polémique sur sa présence aux cérémonies du cinquantenaire...	aucune étiquette

13.	les manifestants..., les pompiers..., les SDF...	aucune étiquette

3.2 Les organisations (org)

Toutes les expressions, noms, sigles, etc. faisant référence à une organisation de nature politique, religieuse, culturelle, etc.

On distingue quatre sous-classes :

1. **org.pol** : les organisations à caractère politique, qui traitent des affaires gouvernementales, politiques, ou étatiques (ex. 1 à 3) ;
2. **org.com** : les organisations à caractère commercial, qui offrent des services, des produits en vue de réaliser des bénéfices (ex. 4 à 7) ;
3. **org.edu** : les organisations à visée éducative (ex. 8) ;
4. **org.non-profit** : les organisations qui ne font pas de bénéfices, qui cherchent à promouvoir des services, des idées, des produits au profit de l'organisation elle-même et non par rapport à l'argent qu'elle pourrait en tirer (ex. 9).

N°	**Exemples**	**Étiquettes utilisées**
1.	au sein de son cabinet et du **[RPR]**...	[ent=org.pol]
2.	le **[parti communiste français]** ne deviendra pas...	[ent=org.pol]
3.	en Chine, le **[parti communiste]** ne deviendra pas social démocrate...	[ent=org.pol]
4.	le guide de haute montagne qui accompagnait les enfants un directeur et un accompagnateur de l'**[UCPA]**...	[ent=org.com]
5.	l'ombre des **[All Blacks]** se profilent...	[ent=org.com]
6.	le conservateur du **[Louvre]**...	[ent=org.com]
7.	les syndicats de la **[SNCF]** appellent à la grève...	[ent=org.com]
8.	le **[Conservatoire National des Arts et Métiers]**...	[ent=org.edu]
9.	les **[Restos du cœur]**...	[ent=org.non-profit]

3.3 Les groupes géo-socio-politiques (gsp)

Un même mot peut parfois être employé pour référer à la fois à une région, son peuple et son gouvernement.

Cette classe se subdivise en trois sous-catégories :

1. **gsp.org** : lorsque les régions administratives (pays, ville, région, département, commune, arrondissement...) sont envisagées comme une organisation (ex. 1) ;
2. **gsp.loc** : lorsque les régions administratives (pays, ville, région, département, commune, arrondissement...) sont perçues comme un lieu (ex. 2 et 3) ;
3. **gsp.pers** : lorsque les régions administratives (pays, ville, région, département, commune, arrondissement...) font référence à leurs habitants, leur population (ex. 4).

N°	Exemples	Étiquettes utilisées
1.	Aujourd'hui la **[France]** a signé un traité...	[ent=gsp.org]
2.	je pars pour la **[France]**...	[ent=gsp.loc]
3.	la **[France]** est le pays des droits de l'homme...	[ent=gsp.loc]
4.	la région **[PACA]** prépare ses élections...	[ent=gsp.pers]

Il arrive souvent que des points cardinaux viennent à préciser les régions administratives. Ces derniers ne sont pas à annoter et n'appartiennent donc pas à l'étiquette (ex. 1 et 2) sauf dans certains cas particuliers (ex. 3 et 4).

N°	Exemples	Étiquettes utilisées
1.	l'embouteillage a eu lieu très exactement au sud de **[Metz]**...	[ent=gsp.loc]
2.	dans les régions de l'est et de l'ouest...	aucune
3.	en **[Afrique du Sud]**...	[ent=gsp.loc]
4.	en **[Europe du Nord]**...	[ent=gsp.loc]

3.4 Les lieux (loc)

On distingue les cinq sous-catégorie suivantes :

1. **loc.geo** : les lieux géographique naturels (fleuves, montagnes, rivières, continents, systèmes solaires, etc.) (ex. 1 à 3) ;
2. **loc.geo.line** : les axes de circulation (chemins ferroviaires, tunnels, ponts, rues et boulevards, etc.) lorsqu'ils sont envisagés comme des lieux géographiques (ex. 4 à 6) ;
3. **loc.addr.post** : les adresses postales (ex. 7 à 9) ;

4. **loc.addr.tel** : les numéros de téléphone et de fax (ex. 10 et 11) ;
5. **loc.addr.elec** : les adresses électroniques (mail et URL) (ex. 12 et 13).

Règles d'annotation :

- Si l'adresse mentionnée renvoie à un lieu géographique, alors ce sera l'étiquette **loc.geo.line** qui sera utilisée (ex. 6 à 8).
- En revanche si on donne une adresse ellipsée, la forme de surface peut ne pas être complète (ex. 9), tout en désignant une adresse postale. L'étiquette utilisée sera alors **loc.addr.post** de la même façon que pour les adresses complètes (ex. 7 et 8). Il faut tenir compte de l'intention du locuteur.

N°	Exemples	Étiquettes utilisées
1.	dans la mythologie grecque [...] afin de la porter sur son dos sur l'autre rive du **[Bosphore]** où il tenta de ...	[ent=loc.geo]
2.	sur la majeure partie du pays, à l'exception du **[littoral breton]** et **[méditerranéen]** il gèle.	[ent=loc.geo] et [ent=loc.geo]
3.	puis en ce qui concerne la *planète () la **[planète Mars]** , il y aurait beaucoup moins d'eau glacée sur la **[planète rouge]** que ...	[ent=loc.geo] et [ent=loc.geo]
4.	le musette [...] né de la rencontre de ces immigrés italiens avec les Auvergnats , c'était **[rue de Lappe]** à ...	[ent=loc.geo.line]
5.	vous pouvez notamment appeler une cabine sur la **[5ème avenue]** à New-York...	[ent=loc.geo.line]
6.	un autre accident à l'est de Paris sur l'**[autoroute A4]**...	[ent=loc.geo.line]
7.	doit être envoyé au **[3 avenue de Matignon 75008 Paris]**...	[ent=loc.addr.post]
8.	vous envoyez vos dons à la chaîne de l'espoir **[1 rue de la Cabanis 75 0 14 Paris]**...	[ent=loc.addr.post]
9.	parmi les deux adresses qui viennent de vous être annoncées, c'est bien la première, celle **[boulevard national]**, qu'il faut utiliser pour...	[ent=loc.addr.post]

10.	et donc à composer le **[trente-trois un cinquante-six quarante treize soixante-douze]**...	[ent=loc.addr.tel].
11.	vous pouvez poser vos questions [...] au **[0 1 45 24 7000]** ou par internet...	[ent=loc.addr.tel]
12.	sur le site **[www.telecharger.com]**...	[ent=loc.addr.elec]
13.	sur le site **[rfi.fr]**...	[ent=loc.addr.elec]

3.5 Les constructions humaines (fac)

Cette étiquette est utilisée pour toutes les entités qui se limitent aux bâtiments et autres constructions fonctionnelles permanentes humaines (en anglais "facilities"). Les entités de cette catégorie ne sont pas considérées ici comme des institutions (organisations) mais comme des "bâtiments". Grossièrement, on peut dire qu'il s'agit des lieux confinés, où l'on peut circuler, comme les maisons, les usines, les stades, les entreprises, les prisons, les musées, etc. (ex. 1 à 5). Dans certains cas, une construction pourra être employée de manière animée en faisant directement référence aux personnes la composant, on aura alors affaire à un phénomène de métonymie (ex. 6 et 7).

N°	**Exemples**	**Étiquettes utilisées**
1.	le séjour de l'enfant à l'**[hôpital Broussais]**...	[ent=fac]
2.	la toute jeune ONU se transfère au **[palais de Chaillot]**...	[ent=fac]
3.	et qui se tient au **[musée d'art contemporain]** à l'**[hôtel des Invalides]**...	[ent=fac] et [ent=fac]
4.	c'est au **[théâtre de l'Eldorado]**...	[ent=fac]
5.	au déjeuner offert demain à l'**[Élysée]**...	[ent=fac]
6.	l'**[Elysée]** a organisé...	[ent=fac/gsp.pers]
7.	l'**[hôpital Broussais]** était en grève...	[ent=fac/gsp.pers]

3.6 Les productions humaines (prod)

Cette classe se subdivise en quatre catégories :

1. **prod.award** : tous les types de récompense (ex. 1) ;
2. **prod.vehicule** : tous les moyens de transport (voitures, navettes avions, fusées, etc.) (ex. 2) ;
3. **prod.art** : toutes les œuvres artistiques (peinture, sculpture, danse, musique, programmes télévisés, etc.) (ex. 1 et 3) ;
4. **prod.printing** : toutes les œuvres littéraires (romans, bandes dessinées, magazines/journaux, etc.) (ex. 4).

N°	Exemples	Étiquettes utilisées
1.	son second film **[une affaire de goût]** a reçu hier le **[grand prix Cognac]**...	[ent=prod.art] puis [ent=prod.award]
2.	les astronautes de la **[navette Endeavour]** ont réussi cette nuit...	[ent=prod.vehicule]
3.	avec une oeuvre colossale, un empilement d'automobiles compressées , intitulé "**[520 tonnes]**"...	[ent=prod.art]
4.	un texte qui reprend d'ailleurs les principes de la **[déclaration française des droits de l'homme]**...	[ent=prod.printing]
5.	s'apprête à accueillir plus de 1000 personnalités étrangères et françaises, dont plusieurs **[prix Nobel de la paix]**...	[ent=prod.award/pers.hum].

3.7 Les dates et heures (time)

Cette catégorie regroupe toutes les expressions relatives au temps (dates et heures).

1. Les dates (time.date)

   Tous les mots ou groupes de mots faisant référence aux jours de la semaine, aux mois, aux années, à des évènements calendaires (**Noël**, **la Pentecôte**, etc.), etc. sont annotés, si et seulement si , au vu du contexte, on sait définir à quelle date précise ils se réfèrent. Dans le cas contraire, ils ne seront pas annotés (ex. 6).

   De façon générale, toute notion temporelle qui constitue un point dans le temps est une date.

Cette classe regroupe deux types de date :

(a) **time.date.abs** : les dates absolues (ex. 1 à 3) ;

(b) **time.date.rel** : les dates relatives (ex. 3 à 5).

Règles d'annotation :

- tout ce qui est adverbes de temps (pendant, durant, etc.), conjonctions de coordination (et, ou, etc.), expressions adverbiales (il y a, etc.) n'est pas annoté ;
- pour les dates relatives, nous nous limitons aux unités lexicales suivantes : hier, demain, aujourd'hui, avant-hier, année, an, etc. De façon générale, on ne prend que les mots désignant une période dont la délimitation est standard. Ainsi, "le début des années 60", "cet après-midi"..., étant des périodes subjectives, ne seront pas annotées.

N°	**Exemples**	**Étiquettes utilisées**
1.	c'était le **[vingt-trois janvier quatre-vingt-dix-huit]**...	[ent=time.date.abs]
2.	elle garde espoir pour **[Noël]** prochain...	[ent=time.date.abs]
3.	à partir d'**[aujourd'hui]** et jusqu'à **[vendredi]**...	[ent=time.date.rel] et [ent=time.date.abs]
4.	le quarante-quatrième congrès de la CFDT s'ouvre **[aujourd'hui]** à...	[ent=time.date.rel]
5.	César est mort **[hier]** à Paris...	[ent=time.date.rel]
N°	**Contre-Exemple**	**Étiquette utilisée**
6.	le **[lundi]** est un jour comme les autres...	aucune car ici le **lundi** ne réfère à aucune date en particulier.

2. Les heures (time.hour)

Toutes les heures sont annotées. S'il est fait une distinction entre les horaires du matin, de l'après-midi et du soir et qu'ils constituent un horaire à part entière alors ils sont annotés au sein de l'étiquette (ex. 1) sinon l'éventuelle précision n'est pas à intégrer dans l'étiquette (ex. 2 et 3).

N°	**Exemples**	**Étiquettes utilisées**
1.	il est **[6 heures du matin]**...	[ent=time.hour]

2.	l'alerte a été donnée vers **[4 heures]** ce matin...	[ent=time.hour]
3.	le rendez-vous aura lieu vers **[4 heures]** dans l'après-midi...	[ent=time.hour]

3.8 Les quantifiables (amount)

Cette classe distingue sept catégories de quantifiables :

1. **amount.phy.age** : âge ;
2. **amount.phy.dur** : durée ;
3. **amount.phy.temp** : température ;
4. **amount.phy.len** : hauteur, largeur, distance..., etc. ;
5. **amount.phy.wei** : poids ;
6. **amount.phy.spd** : vitesse ;
7. **amount.cur** : les valeurs monétaires.

N°	Exemples	Étiquettes utilisées
1.	l'hiver dernier il s'est même attaqué à cette petite fille âgée de **[douze ans]**...	[ent=amount.phy.age]
2.	il a couru pendant **[deux heures]**...	[ent=amount.phy.dur]
3.	il a décidé pendant **[3 semaines]**...	[ent=amount.phy.dur]
4.	dans**[2]** ou **[3 ans]**...	[ent=amount.phy.dur] et [ent=amount.phy.dur]
5	changer la serrure puis rester**[2]** ou **[3 jours]** dedans...	[ent=amount.phy.dur] et [ent=amount.phy.dur]
6.	une série de manifestations qui dureront toute la **[semaine]**...	[ent=amount.phy.dur]
7.	elles afficheront cet après-midi autour de **[seize degrés]**...	[ent=amount.phy.temp]
8.	venait se détacher sur une largeur de **[cent mètres]**...	[ent=amount.phy.len]
9.	de **[50]** à **[60 mètres]** environ...	[ent=amount.phy.len] et [ent=amount.phy.len]
10.	souvenez-vous de son pouce agrandi à **[1 mètre 85]**...	[ent=amount.phy.len]
11.	le mistral approchera **[60]** **[80 km/h]**...	[ent=amount.phy.spd] et [ent=amount.phy.spd]

12.	d'un montant de **[vingt millions de francs]**...	[ent=amount.cur]

3.9 Incertain (unk)

Cette classe est utilisée pour les cas non envisagés dans le document et lorsqu'aucune autre des catégories et sous-catégories proposées ne correspond.

N°	**Exemples**	**Étiquettes utilisées**
1.	le **[trône d'Angleterre]**...	[ent=unk]
2.	une grande **[opération "mains propres"]** a été lancée par le préfet Bonnet...	[ent=unk]
3.	l'indice **[Nikkeï]**...	[ent=unk]