Corela
Cognition, représentation, langage

14-2 | 2016
Vol.14, n°2

Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI

Solenn Le Pevedic et Denis Maurel

Édition électronique
URL : http://journals.openedition.org/corela/4644
DOI : 10.4000/corela.4644
ISSN : 1638-573X

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique
Solenn Le Pevedic et Denis Maurel, « Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI », *Corela* [En ligne], 14-2 | 2016, mis en ligne le 08 décembre 2016, consulté le 30 avril 2019. URL : http://journals.openedition.org/corela/4644 ; DOI : 10.4000/corela.4644

Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI

Solenn Le Pevedic et Denis Maurel

Introduction

1 La notion d'entité nommée est apparue lors des conférences Muc (Message Understanding Conference) aux États-Unis. Le but de ces conférences était d'évaluer les systèmes d'extraction d'informations. La définition est donnée par (Chinchor, 1997) pour la conférence Muc-7 : On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts. Cette notion a été étudiée et affinée du point de vue linguistique par (Ehrmann, 2008) : Étant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. Le lien qu'elle fait entre modèle applicatif et corpus introduit la question abordée dans cet article, à savoir la portabilité des annotations en entités nommées.

2 Pour comparer les résultats obtenus par différentes équipes de Tal, l'idée est de définir des annotations à insérer dans un corpus et d'évaluer les résultats de chaque équipe, à la fois en terme de précision (a-t-on annoté correctement une entité ?) et de rappel (en a-t-on oublié certaines ?).

3 Une campagne d'évaluation propose donc des annotations à insérer dans un corpus sur lequel tous les systèmes participants seront évalués en terme de rappel et précision (et/ou F-mesure). Un guide d'annotation précise le résultat attendu. Il est accompagné d'un corpus d'entraînement et de test, notamment pour les systèmes à base d'apprentissage.

Cette annotation manuelle est elle-même contrôlée par des mesures d'accord inter-annotateurs (Artstein et Poesio, 2008) (Fort et al., 2012). Depuis la conférence Muc-7 qui portait sur l'anglais, d'autres campagnes ont eu lieu sur d'autres langues, comme, entre autres, Met (Merchant et al., 1996) pour le japonais, le chinois et l'espagnol, Irex (Sekine et Eriguchi, 2000) pour le japonais, Harem (Santos et al., 2006) pour le portugais, etc. Nous tenons aussi à rappeler les campagnes Language-Independent Named Entity Recognition des conférences CoNLL-2002[1] et CoNLL-2003[2], entièrement basées sur l'apprentissage.

4 Pour le français, deux[3] campagnes ont eu lieu sur des corpus oraux transcrits, soit manuellement, soit automatiquement. Ces campagnes, organisées sous l'égide de l'association francophone de communication parlée (AFCP), ont eu lieu en 2008 pour *Ester-2* et en 2012 pour *Etape*. La campagne Ester-2 utilisait un guide d'annotation (Ester-2, 2007), testé lors de la campagne Ester-1, puis revu et disponible en ligne. La campagne Etape s'est appuyée sur le guide d'annotation défini dans le cadre du projet Quaero (Rosset et al., 2011).

5 Le principal inconvénient des annotations préconisées dans ces deux campagnes est leur manque de portabilité. Il est nécessaire de refaire les corpus d'apprentissage et d'adapter les systèmes à base de règles. La présentation des résultats et des exemples n'est pas simple en dehors du cercle restreint des participants ou, du moins, en comparaison avec d'autres systèmes sur d'autres langues.

6 Parallèlement aux guides d'annotation sur les entités nommées, un consortium, la *TEI* (*Text Encoding Initiative*[4]) issu de différentes disciplines en sciences humaines a travaillé, dès la fin des années 1980, sur un projet de définition d'un balisage standard des documents dans un format électronique. Les documents concernés pouvant tout aussi bien être du texte, de l'image, de la vidéo, du son, etc. avaient pour but de faciliter les échanges internationaux de documents. Des années de recherche ont été nécessaires avant de pouvoir publier le premier guide en 1994. Depuis cette date, celui-ci n'a pas cessé d'évoluer. La version actuelle en est la cinquième version couramment appelée *TEI P5* (Text Encoding Initiative Consortium, 2015). À l'origine, la TEI suivait la norme SGML, mais aujourd'hui elle utilise le langage XML. Le balisage ne concerne pas seulement le texte lui-même, car la TEI préconise aussi le balisage des informations liminaires que l'on peut obtenir sur le document à encoder (titre, auteurs, éditeurs, résumé, pagination, paragraphes, etc.). La TEI permet aussi de baliser la structure du texte lui-même (paragraphes, phrases, mots...).

7 L'annotation temporelle existe dans la TEI, mais sur ce point précis, d'autres travaux ont proposé des guides de balisage, citons la thèse d'Andrea Setzer (Setzer, 2001), le projet *Tides* (Ferro et. al., 2001) et, pour finir, *TimeML* (Pustejovsky, 2003). Ce dernier est devenu une norme ISO, définie en mars 2009. *TimeML* concerne non seulement les expressions temporelles, souvent considérées comme des entités nommées, mais aussi les évènements portés par les prédicats, ce qui sort du contexte de cet article[5].

8 L'article s'articule autour de trois parties : en section 1, nous faisons un retour sur les conventions des campagnes Ester et Etape. Dans la section 2, nous présentons les balises TEI concernant les entités nommées. Enfin, dans la section 3 nous préconisons des solutions pour remplacer les balises du projet Quaero par des balises TEI.

1. Retour sur les campagnes Ester et Etape

1.1. Les campagnes Ester-1 et Ester-2

9 Les campagnes pour l'évaluation des systèmes de transcription enrichies d'émissions radiophoniques (plus communément appelée Ester) avaient pour but d'évaluer la performance des systèmes sur différents types de tâches, essentiellement la transcription et l'indexation d'émissions de radio. La campagne se divise en deux phases successives. La première phase courait de 2003 à 2005, la seconde de 2006 à 2008. Plusieurs tâches sont effectuées lors de ces campagnes, telles que la segmentation et la transcription orthographique, mais, également, ce qui nous intéresse ici, la reconnaissance des entités nommées.

10 La première phase n'exploite pas totalement cette tâche contrairement à la seconde dont il s'agit d'un des thèmes principaux. Elle est principalement soutenue par l'Association francophone de la communication parlée (AFCP), mais aussi par le Centre d'expertise parisien de la Délégation générale pour l'armement (DGA/CEP) et par l'European Language Ressources Association (ELRA) qui permet, quant à elle, la pérennité du projet par la diffusion de corpus annotés.

11 Un corpus conséquent a donc été établi. Celui-ci contient des articles de presse tirés du journal « Le Monde », des textes tirés des débats du Conseil européen, et, surtout, d'une centaine d'heures d'émissions de radio (principalement France Inter) transcrites pour et par cette tâche. Le corpus entre la première et la seconde phase a été légèrement remanié. Le nombre d'heures de ressources orales transcrites passent de 78h55 à 103h08. Pour ce corpus, on recense environ 150 000 entités nommées pour ce corpus.

1.2. Retours sur le guide de la campagne Ester-2

12 Le guide d'Ester-2 qui est une version revue et augmentée de celui de la campagne Ester-1, est disponible en ligne sur le site de l'AFCP. Pour ce guide, les entités nommées sont au cœur de la problématique de l'extraction de l'information d'un document. Ce sont des types particuliers d'unités lexicales (groupes de mots) 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). S'ajoutent à la tâche de reconnaissance des entités nommées celle des informations « temps » et « montant ».

13 En dehors des personnes, lieux et organisations déjà présents dans la campagne Muc-7, on retrouve deux nouveaux types : les fonctions (politiques, militaires, administratives, religieuses et aristocratiques) et les productions humaines (moyens de transport, récompenses, œuvres artistiques et productions documentaires). Ces deux types supplémentaires semblent contredire la précision donnée en introduction du guide et rappelée ci-dessus, qui stipule que les entités nommées correspondent à des entités ayant un nom, même si ce sont bien des entités du monde concret. Pour les fonctions, comme dans l'exemple [6], il s'agit bien de personnes qui sont ainsi désignées.

14 1. Le <ent type ="fonc.admi">vice-directeur</ent> est mis en examen suite à l'affaire...

15 Les productions humaines posent la question de la définition d'un nom de marque ou de produit : est-ce un nom propre ou un nom commun ? (Fèvre-Pernet, Roché, 2005) ou une classe d'objets ? (Gross, 2012). L'exemple semble désigner la classe, alors que l'exemple semble plutôt correspondre à un objet particulier.

16 2. Depuis que j'ai acheté la <ent type ="prod.vehicule">Suzuki GSXR 600</ent>, je gagne...

17 3. J'ai acheté une <ent type ="prod.vehicule">Renault</ent>.

18 Signalons une petite incohérence[7]. Dans l'exemple , la précision *UMP* est placée dans la balise *fonc.pol*, alors que, dans l'exemple , *ashkénaze* n'est pas dans la balise *fonc.rel*, bien que les deux fonctions (*député/grand rabbin*) sont associées à (*UMP/ashkénaze*) qui précisent le type de député et le type de grand rabbin.

19 4. <ent type ="pers.hum"><ent type ="pers.hum">Pierre Lellouche</ent> le <ent type ="fonc.pol">député UMP de <ent type ="loc.admi">Paris</ent></ent></ent>

20 5. le <ent type ="pers.hum"><ent type ="fonc.rel">grand rabbin</ent> ashkénaze <ent type ="pers.hum">Avraham Shapira</ent></ent>

21 Il nous semble qu'il faudrait plutôt écrire <ent type ="fonc.rel">grand rabbin ashkénaze</ent>.

22 Ajoutons à cela une remarque sur des difficultés quasiment insurmontables (sans connaissances pragmatiques ou encyclopédiques). Si il est judicieux de séparer, pour un même nom, l'aspect lieu de l'aspect institution, comme les exemples et , car le contexte permet En dehors des personnes, lieux et organisations déjà présents dans la campagne Muc-7, on retrouve deux nouveaux types : les fonctions (politiques, militaires, administratives, religieuses et aristocratiques) et les productions humaines (moyens de transport, récompenses, oeuvres artistiques et productions documentaires).la recherche d'entité nommée par ce qu'on appelle aujourd'hui l'*entity linking* (McNamee et al., 2010).

23 6. Je suis stationné à côté de la <ent type ="loc.fac">mairie de Paris</ent>.

24 7. La course à la <ent type ="org.pol">mairie de Paris</ent> a commencé.

25 8. Je me suis fait opérer à l'<ent type ="org.non-profit">hôpital Necker</ent>.

26 Enfin terminons par la décision de *ne pas étiqueter les lieux personnels, c'est-à-dire tout lieu ou habitation désignée comme appartenant à un particulier*. Comment savoir si, dans l'exemple , la dite propriété est réellement privée ou est un lieu public, même si le propriétaire est privé ? Citons l'exemple du château de Villandry qui est une propriété privée ouverte à la visite, comme un musée, et dans laquelle le propriétaire habite !

27 9. La propriété Saint Vincent a été rachetée par le Comte de Bourgogne

1.3. La campagne Etape

28 La campagne Etape émane du projet européen Quaero[8] qui s'est déroulé sur une période allant de 2008 à fin 2013. Elle se proposait d'évaluer les résultats de ce projet en termes de recherche d'entités nommées. Cependant, cette campagne d'évaluation a été ouverte largement et des équipes ne participant pas au projet Quaero ont pu y contribuer.

29 C'est dans le cadre de ce projet qu'un guide d'annotation a été réalisé (Rosset et al., 2011). La complexité des annotations a largement augmenté (Table 1).

Table : Nombre d'étiquettes et de pages des guides d'annotation

Campagne	Nombre d'étiquettes	Nombre de pages
Muc-7	7	23
Ester-2	37	25
Etape	54	86

1.4. Retours sur le guide de la campagne Etape

30 Comme cela était évoqué dans les précédents guides, la notion d'*entités nommées* recouvre (entre autres) celle de *noms propres*. Le guide Quaero s'affirme *en cohérence avec la définition des entités nommées* de la campagne Ester-2. Il adopte :

– la définition des entités nommées de la thèse de Maud Ehrmann (2008) ;

– la définition des entités temporelles du projet Time ML (Pustejovsky, 2003) ;

– la taxinomie de la thèse de Mickaël Tran (Tran, 2006 ; Tran et Maurel, 2006) ;

– la hiérarchie des entités nommées étendues proposée par Satoshi Sekine (Sekine,2004, Nadeau and Sekine, 2007).

31 Comme le projet Quaero se place dans le contexte d'extraction d'informations, il demande l'inclusion d'expressions ne contenant pas de nom propre et étend l'annotation à des expressions construites autour de noms communs, comme sur l'exemple .

32 10. <func.ind><extractor>un </extractor>des <kind>pompiers</kind></func.ind>

33 Le guide est très complet et comporte de nombreux exemples. L'annotation n'est plus uniquement centrée sur les entités, mais aussi sur les composants de ces entités, comme on le voit sur ce même exemple .

34 Les composants principaux sont *<name>* pour le nom de l'entité et *<kind>* ou *<qualifier>* pour les noms communs et les qualifieurs constituant l'entité étendue.

35 Nous ne ferons qu'une remarque sur ces annotations internes, concernant la distinction entre les qualifieurs suivant qu'ils sont intégrés ou non à des expressions temporelles, comme sur les exemples et .

36 11. le <prod.rule><qualifier>dernier</qualifier> <kind>plan de paix</kind><qualifier>international</qualifier> </prod.rule>

37 12. <time.hour.abs><time-modifier>aux alentours de</time-modifier> <val>quinze</val> <unit>heures</unit> <val>trente</val></time.hour.abs>

38 Or *dernier* sera annoté *time-modifier* dans *le dernier jeudi du mois* et *aux alentours de* sera annoté *qualifier* dans *aux alentours de 15 kg*. Bien sûr, cela vient probablement du fait que certains qualifieurs ne peuvent concerner les expressions temporelles, comme *socialiste* dans l'exemple , et inversement, certains modifieurs ne concernent que celles-ci, comme

matin dans l'exemple . Il nous semble cependant que ces deux étiquettes auraient dû être confondues.

39 13. Le <pers.ind><qualifier>socialiste</qualifier><name.first>Bertrand</name.first><name.last>Delanoë </name.last> </pers.ind>

40 14. <time.date.rel><name>hier</name> <time-modifier>matin</time-modifier></time.date.rel>

41 D'autre part, nous avons du mal à comprendre le choix qui a été fait pour les dates absolues et relatives. Dans les campagnes Ester, une date est considérée absolue lorsque la date du document annoté n'est pas nécessaire pour la déterminer précisément et, a contrario, une date est considérée comme relative lorsque la date du document annoté est indispensable pour la déterminer précisément. Dans le guide Quaero, deux sortes de dates absolues sont considérées, les dates absolues précises (celles de la campagne Ester) et les dates absolues imprécises dont le point de référence dans le calendrier n'est pas connu avec certitude mais dont les composants sont absolus. Par contre, dès qu'un marqueur explicite de relativité (par exemple, prochain) est spécifié (rendez-vous mardi prochain), on a une date relative.

42 Ainsi les exemples et sont censés[9] illustrer ce propos. Pourtant quelle différence à déclarer un rendez-vous mardi ou mardi prochain ? Et le changement de déterminant dans un rendez-vous le 25 janvier (date absolue) et un rendez-vous ce 25 janvier (date relative) crée-t-il réellement une différence sémantique ?

43 15. rendez-vous <time.date.abs><week>mardi</week></time.date.abs>

44 16. <week>lundi</week> <time-modifier>prochain</time-modifier></time.date.rel>

2. Les entités nommées dans la TEI

2.1. Les éléments les plus généraux

45 L'élément le plus général qu'il est possible d'utiliser est la balise *referencing string* qui indique le renvoi à un référent. De ce fait, cet élément peut baliser un nom propre, mais aussi toute référence anaphorique y compris pronominale.

46 17. My dear <rs type="person">Mr. Bennet</rs>, said <rs type="person">his lady</rs> to him one day

47 Plus spécifiquement, mais de manière encore très générale, l'élément *name* permet le balisage des noms propres .

48 18. <name type="person">Mr. Bennet</name>

49 19. <name type="place">Netherfield Park</name>

50 Ces balises peuvent contenir des attributs permettant de lier des entités (*id*, pour un identifiant unique, *key*, pour une clé et *ref* pour le référent), mais ceci n'était pas considéré par le guide Quaero.

51 Le terme *entité* apparaît dans le guide TEI dans la sous-classe d'attributs *att.naming*. On y trouve par exemple l'attribut *role* qui donne des informations spécifiques sur l'entité encodée : *@role may be used to specify further information about the entity referenced by this name, for exemple the occupation of a person, or the status of a place*.

52 20. <name role="politician" type="person">David Paul Brown</name>

2.2. Personnes, lieux et organisations

53 Le guide TEI introduit aussi des éléments plus spécifiques concernant les personnes, lieux et organisations.

54 L'élément *persName* permet le balisage des noms de personne. Il est accompagné de plusieurs autres éléments permettant de spécifier le prénom (*forename*) ou le nom (*surname*), mais aussi les titres (*roleName*), les surnoms (*addName*), les générations (*genName*) ou même les particules (*nameLink*) ou la notion d'abréviation (*full*)... .

55 21. <persName><roleName type="nobility">Princess</roleName> <forename>Grace</ forename> </persName>

56 22. <persName><roleName type="office">Governor</roleName> <forename>Edmund</ forename> <forename full="init">G.</forename> <addName type="nick">Jerry</ addName> <addName type="epithet">Moonbeam</addName> <surname>Brown</ surname> <genName full="abb">Jr</genName>.</persName>

57 Les lieux sont divisés en deux, les lieux administratifs (*placeName*) et les lieux géographiques (*geogName*) . Les lieux géographiques sont spécifiés par l'élément *geogFeat*.

58 23. <placeName>New York</placeName>

59 <geogName type="river"><name>Mississippi</name>

60 24. <geogFeat>River</geogFeat> </geogName>

61 La TEI introduit ici la notion de lieu relatif (*relative place*), c'est-à-dire un lieu qui est défini par sa situation relativement à un autre lieu, par deux éléments, *offset* et *measure*. Le premier va marquer une direction (nord, sud...) et le second une distance .

62 25. <placeName><measure>20 km</measure> <offset>north of</offset> <settlement type="city">Paris</settlement></placeName>

63 Enfin, il existe aussi dans le guide, une description très complète des adresses postales, avec l'élément *address* qui peut contenir divers autres éléments, comme *street*, *postCode*, *settlement*, *country*, etc. . Cependant il est aussi possible de ne pas spécifier ces éléments internes et de les remplacer par un seul élément, *addrLine* .

64 26. <address><street>via Marsala 24</street> <postCode>40126</postCode> <settlement type="city">Bologna</settlement> <country>Italy</country></address>

65 27. <address><addrLine>Computing Center, MC 135</addrLine> <addrLine>P.O. Box 6998</addrLine> <addrLine>Chicago, IL 60680</addrLine> <addrLine>USA</addrLine></ address>

66 La notion d'adresse électronique est aussi présente, avec l'élément *email* .

67 28. <email>info@tei-c.org</email>

68 Quant aux organisations , c'est l'élément orgName qui est préconisé par le guide TEI : *We use the term « organization » for any named collection of people regarded as single unit*.

69 29. <orgName type="voluntary">Pennsyla. Abolition Society</orgName>

2.3. Dates, temps, mesures

70 Les valeurs numériques de toute sorte peuvent être encodées : pourcentages, fractions, nombres cardinaux, romains, chinois... éventuellement avec un attribut value .

71 30. <num type="fraction" value="0.5">one half</num>

72 Les éléments de base pour les dates et le temps sont *date* et *time*. Ils peuvent, l'un et l'autre, encoder n'importe quel format de date ou de temps interne à une date (*a phrase defining a time of day in any format*). Des attributs supplémentaires sont nécessaires afin d'avoir une normalisation des éléments encodés, de la même manière que les chiffres. Ceci est particulièrement utile dans la mesure où, dans les différentes cultures, les dates et la notion de temps (comme les heures) sont représentées de différentes manières. Certains attributs sont spécifiques à cette catégorie. Ainsi, l'attribut *when* va permettre de normaliser *date* et *time* par un format standard basé sur le calendrier grégorien et l'attribut *calendar* indique quant à lui le calendrier auquel appartient la forme encodée. On peut également marquer un intervalle à l'aide d'attributs spécifiques ou utiliser d'autres attributs comme *notBefore* ou *notAfter* .

73 31. <date when="2001-09">September 2001</date>

74 32. <time when="08:48:00">8:48</time>

75 33. <date notBefore="1250" notAfter="1300">the second half of the thirteenth century</date>

76 Remarquons que, dans les projets Timex, Tides et TimeML, *date* et *time* sont remplacés respectivement par les éléments TIMEX, TIMEX2 et TIMEX3, qui sont en fait un seul et même élément, mais avec des attributs possibles en extension. D'après (Pustejovsky, 2003:7), le projet TimeML prend en considération les *(a) Fully Specified Temporal Expressions, June 11, 1989, Summer, 2002; (b) Underspecified Temporal Expressions, Monday, Next month, Last year, Two days ago; (c) Durations, Three months, Two years* .

77 34. John taught <TIMEX3 tid="t1" type="DURATION" value="PT20M">20 minutes</TIMEX3> <SIGNAL sid="s1">every</SIGNAL> <TIMEX3 tid="t2" type="DATE" value="XXXX-WXX-1">Monday</TIMEX3>

3. Une équivalence est-elle possible entre le guide Quaero et la TEI ?

78 Dans cette section, en partant des exemples du chapitre 2 du guide Quaero, nous allons chercher si il existe ou non une correspondance dans la TEI. Les exemples reprennent dans l'ordre un grand nombre[10] de ceux du guide Quaero.

3.1. Personnes

79 À quelques différences près sur le balisage interne , les exemples de Quaero se transposent bien en TEI. Les éléments *function*, *qualifier* et *extractor* peuvent être remplacée respectivement par les éléments *roleName* , *addName*, complété par l'attribut *epithet* , et *num*, complété par l'attribut *ordinal* .

80 35. <pers.ind><title>Sa Majesté le <func.ind>roi</func.ind></title> <name.first>Mohamed </name.first> <qualifier>VI</qualifier></pers.ind>

→

<persName><addName type ="honorific">Sa Majesté</addName>le<roleName

type ="nobility"> roi</roleName><forename>Mohamed</forename><genName>VI</genName></persName>

81 36. <pers.ind><title>Son Altesse Royale le <func.ind>prince</func.ind></title> <name.first>Rainier </name.first></pers.ind>

→

<persName><addName type="honorific">Son Altesse Royale</addName> le <roleName type="nobility">prince</roleName> <forename>Rainier</forename></persName>

82 37. Le <pers.ind><qualifier>socialiste</qualifier> <name.first>Bertrand</name.first> <name.last>Delanoë </name.last></pers.ind>

→

le <persName><addName type="epithet">socialiste</addName> <forename>Bertrand</forename> <surname>Delanoë</surname></persName>

83 38. le <pers.ind><extractor>cinquième</extractor> <name>Beatles </name></pers.ind>

→

le <person><affiliation><num type="ordinal" value="5">cinquième</num> <org>Beatles </org></affiliation></person>

84 Lorsqu'il n'y a pas de nom, ou lors d'une énumération, les éléments *pers.ind* et *pers.coll* peuvent être remplacés par les éléments *person* , *listPerson* et *personGrp* .

85 39. le <pers.ind><demonym>français</demonym></pers.ind> s'est classé quatrième

→

le <person><nationality>français</nationality></person> s'est classé quatrième

86 40. <pers.ind><name>Asterix</name></pers.ind> et <pers.ind><name>Obelix</name></pers.ind> ont apporté un sanglier

→

<listPerson><person><persName>Asterix</persName></person> et <person><persName>Obelix</persName></person></listPerson> ont apporté un sanglier

87 41. <pers.coll><title>Monsieur</title> et <title>Madame</title> <name.last>Chirac</name.last> </pers.coll>

→

<listPerson><person><roleName type="honorific">Monsieur</roleName></person> et <person><roleName type="honorific">Madame</roleName> <persName>Chirac</persName> </person></listPerson>

88 42. les <pers.coll><demonym>Français</demonym></pers.coll>

→

les <personGrp><nationality>Français</nationality></personGrp>

3.2. Fonctions

89 Comme dans l'exemple , les fonctions sont balisées par l'élément *roleName* ; lorsque la fonction apparaît seule, sans nom propre, cet élément est intégré dans un élément *person* ou *personGrp* .

90 43. l'<func.ind><kind>ambassadeur</kind> de <loc.adm.nat>Turquie</loc.adm.nat> en <loc.adm.nat> France</loc.adm.nat></func.ind>

→

l'<person><roleName type="office">ambassadeur</roleName> de <country>Turquie</country> en <country>France</country></person>

91 44. les <func.coll><kind>maires</kind> de <loc.adm.nat>France</loc.adm.nat></func.coll>

→

les <personGrp><roleName type="office">maires</roleName> de <country>France </country></personGrp>

3.3. Organisations

92 L'élément *orgName* peut contenir d'autres éléments, comme un *name* de type *role* , un *surname* , un autre *orgName*, etc. On retrouve donc une équivalence comme précédemment.

93 45. la <org.ent><kind>police</kind> <demonym>française</demonym></org.ent>

→

<org><name type ="role">police</name> <nationality>française</nationality></org>

94 46. la <org.ent><kind>société</kind> <name>Peugeot</name></org.ent>

→

la <orgName><name type ="role">société</name> <surname>Peugeot</surname></orgName>

3.4. Localisations

95 Pour les lieux, il en est de même avec différents types ou .

96 47. <loc.adm.town>La Bolline</loc.adm.town>

→

<placeName type ="city">La Bolline</placeName>

97 48. <loc.adm.town>Maison Blanche</loc.adm.town>[11]

→

<placeName type ="district">Maison Blanche</placeName>

98 49. Le <loc.adm.town><name>13e</name> <kind>arrondissement</kind></loc.adm.town>

→

Le <placeName><district><num type="ordinal" value="13">13e</num>arrondissement </district></placeName>

99 Les lieux relatifs donnent aussi lieu à l'annotation .

100 50. <loc.adm.reg><qualifier>au sud </qualifier> d'<name><loc.adm.nat>Israël </loc.adm.nat></name></loc.adm.reg>

→

Au <placeName><offset>sud d'</offset><country>Israël</country></placeName>

101 Tout comme les lieux géographiques .

102 51. la<loc.adm.sup><kind> région </kind>de l'<name>Atlas</name></loc.adm.sup>

→

la <placeName><offset>région de</offset> <geogName>l'Atlas</geogName></placeName>

103 52. le <loc.phys.geo><kind>désert</kind> de <name>Gobi</name></loc.phys.geo>

→

Le <geogName><geogFeat>désert</geogFeat> de Gobi</geogName>

104 53. la <loc.phys.hydro>Seine</loc.phys.hydro>

→

la <geogName type ="river">Seine</geogName>

105 Pour les voies de circulation, il n'est pas prévu de descripteur spécifique. Le plus simple serait d'utiliser un attribut *oronym* dans un élément *placeName* . C'est d'ailleurs ce qui est proposé par la TEI pour les bâtiments avec l'attribut *building*, éventuellement complété par un élément addName avec l'attribut epithet .

106 54. l'<loc.oro><kind>autoroute</kind> <name>A6</name></loc.oro>

→

l'<placeName type="oronym">autoroute A6</placeName>

107 55. l'<loc.fac>ancienne <kind>gare</kind> de <name>Rungis</name></loc.fac>

→

l'<placeName type="building"><addName type="epithet">ancienne</addName> gare de <settlement>Rungis</settlement></placeName>

108 Pour les adresses physiques, la correspondance existe, même dans des adresses complexes, avec les éléments *address* et *addrLine* . C'est ce dernier élément que l'on utilise seul pour les numéros de téléphone . Les adresses électroniques correspondent à l'élément *email* et les identifiants pourraient être balisés par l'élément *ident* . En revanche, rien ne concerne les fréquences radio qu'on pourrait considérer à l'instar des numéros de téléphone, avec l'élément addrLine .

109 56. J'habite <loc.add.phys><address-number>15</address-number> <loc.oro><kind>rue</kind> de <name>Vaugirard</name></loc.oro> <other-address-component>escalier 2</other-address-component></loc.add.phys>

→

<address><addrLine>15 rue de Vaugirard</addrLine> <addrLine>escalier-2</addrLine></address>

110 57. mon numéro est le <loc.add.elec><name>01 69 85 80 02</name></loc.add.elec>

→

mon numéro est le <addrLine>01 69 85 80 02</addrLine>

111 58. mon identifiant skype est <loc.add.elec><name>jean.dupont</name></loc.add.elec>

→

mon identifiant skype est <ident type = "skype">jean.dupont</ident>

112 59. <prod.media>Radio Bleue</prod.media> sur <loc.add.elec><name>98.8 MHz</name></loc.add.elec>

→

<orgName>Radio Bleue</orgName> sur <addrLine>98.8 MHz</addrLine>

3.5. Productions humaines

113 Comme nous le voyons dans l'exemple , il n'existe pas dans la TEI de description des productions humaines en général. La solution serait d'utiliser l'élément *name* complété par un attribut *type* et éventuellement un élément *role.* Pour les objets, le type pourrait prendre la valeur *object* [12].

114 60. j'ai piloté la <prod.object><kind>navette</kind> <name>Endeavour</name></prod.object>

→

j'ai piloté la <name type ="object"><role>navette</role> Endeavour</name>

61. le <prod.other><name>Monopoly</name></prod.other>

→

le <name type="object">Monopoly</name>

De même, pour les autres produits cet attribut type pourrait prendre comme valeur *finance* , *soft* , *award* , *service* , *doctrine* ou *rule* .

62. la <func.ind>gérante</func.ind> du <prod.fin><kind>fonds</kind> <name>DWS signature court terme</name></prod.fin> chez <org.ent>DWS Investments</org.ent> en <loc.adm.nat>France</loc.adm.nat>

→

la <person><roleName type ="occupation">gérante</roleName></person> du <name type ="finance"><role>fonds</role> DWS signature court terme</name> chez <orgName>DWS Investments</orgName> en <placeName type ="country">France</placeName>

63. le <prod.soft><kind>système d'exploitation</kind><name> Unix</name></prod.soft>

→

le <name type ="soft"><role>système d'exploitation</role> Unix</name>

64. le <prod.award><name>Prix Nobel</name> de <award-cat>chimie</award-cat></prod.award>

→

le <name type="award">Prix Nobel <addName>de chimie</addName></name>

65. le <prod.serv><kind>RER</kind> <name>B</name></prod.serv>

→

le <name type="service"><role>RER</role> B</name>

66. le <prod.doctr><name>socialisme</name></prod.doctr>

→

le <name type ="doctrine">socialisme</name>

67. le <prod.rule><kind>traité</kind> de <name><loc.adm.town>Versailles</loc.adm.town> </name></prod.rule>

→

le <name type ="rule"><role>traité</role> de <settelment type ="city">Versailles</settelment></name>

123 Par contre, dans la catégorie artistique de Quaero, on pourrait plutôt utiliser l'attribut media qui permettrait de préciser le type de media utilisé, ainsi que les éléments *title* et *author*. L'élément *title* peut être précisé par un attribut *level* qui prend les valeurs *m* (monographie), *j* (journal), *s* (série) ou *u* (non publié) .

124 68. <prod.art><name>Le Baiser</name></prod.art> de <pers.ind><name.last>Klimt </name.last></pers.ind>

→

<name media="art"><title>Le Baiser</title> de <author>Klimt</author></name>

125 69. <pers.ind><name.first>Mickaël</name.first> <name.last>Vendetta</name.last></pers.ind> veut quitter <prod.media><name>La Ferme des Célébrités</name></prod.media>

→

<persName><forename>Mickaël</forename> <surname>Vendetta</surname></persName> veut quitter <name media="TV"><title level="s">La Ferme des Célébrités</title></name>

126 70. <prod.media><name>Le Monde</name></prod.media> a publié un article sur <prod.art><name>1984</name></prod.art> de <pers.ind><name.first>George</name.first> <name.last>Orwell</name.last></pers.ind>

→

<orgName>Le Monde</orgName> a publié un article sur <name media="book"><title level="m">1984</title> de <author>George Orwell</author></name>

127 En revanche, pour l'exemple , qui n'est pas une entité nommée, il est impossible d'utiliser l'élément *name*. En supposant qu'il s'agisse d'une anaphore, on pourrait éventuellement utiliser l'élément *rs* avec un attribut *type* de valeur *rule* .

128 71. le <prod.rule><qualifier>dernier</qualifier> <kind>plan de paix</kind> <qualifier>international </qualifier></prod.rule>

→

le <rs type ="rule"><epithet>dernier</epiteth> plan de paix <epiteth>international</epiteth></rs>

3.6. Quantités

129 Les quantités sont bien couvertes par la TEI, avec l'élément *measure*. Cet élément est enrichi par les attributs *quantity* et *unit*. Lorsqu'on compte des objets, la TEI propose d'utiliser *commodity*... Cependant, contrairement à la proposition de Quaero, nous

proposons de distinguer les marchandises des personnes, en utilisant les éléments *person* et *personGrp*, comme dans l'exemple .

130 72. <amount><val>trois </val> <object>pompiers</object></amount>

→

<measure quantity ="3" unit ="count">trois <person>pompiers</person></measure>

131 Pour les quantités approximatives, la TEI propose d'utiliser l'élément *extent* et de placer l'élément *measure* à l'intérieur de celui-ci .

132 73. <amount><val>quelques</val> <unit>mètres</unit></amount>

→

<extent><measure unit ="m">quelques mètres</measure></extent>

133 L'exemple utilise l'élément *commodity* dont nous avons parlé ci-dessus.

134 74./ <amount><val>une douzaine</val> de <object>tomates</object></amount>

→

<measure quantity ="12" unit ="count" commodity ="tomates">une douzaine de tomates</measure>

135 Pour l'exemple , on pourraît définir une unité de valeur *container* applicable aussi aux cuillérées et aux cars .

136 75. <amount><val>une </val><unit>boîte</unit> de <object>tomates</object></amount>

→

<measure quantity ="1" unit ="container" commodity ="tomates">une boîte de tomates ></measure>

137 76. <amount><val>trois</val> <unit>cuillerées</unit> de <object>farine de châtaignes</ object></amount>

→

<measure quantity ="3" unit ="container" commodity ="farine de châtaignes">trois cuillerées de farine de châtaignes</measure>

138 77. <amount><val>un</val> <unit>car</unit> de <object>CRS</object></amount>

→

<measure quantity ="1" unit ="container">un car de <person>CRS</person></measure>

139 L'élément *extent* permet non seulement de marquer les valeurs approximatives, comme signalé ci-dessus, mais aussi des intervalles .

140 78. <amount><range-mark>entre</range-mark> <val>deux</val> <range-mark>et</range-mark> <val>trois</val> <unit>kilomètres</unit></amount>

→

<extent>entre <measure quantity ="2" unit ="km">deux</measure> et <measure quantity ="3" unit ="km">trois kilomètres</measure></extent>

141 L'exemple va être traité différemment. Les deux éléments *amount* du guide Quaero devront être remplacés par l'élément *unit* pour la première partie et l'élément *measure* pour la deuxième.

142 79. la distance en <amount><unit>kilomètres</unit></amount> c'est <amount><val>trois</val></amount>

→

la distance en <unit>kilomètres</unit> c'est <measure quantity ="3" unit ="km">trois</measure>

143 Pour l'âge d'une personne, la TEI propose un élément spécifique, *age* .

144 80. j'ai <amount><val>10</val> <unit>ans</unit></amount>

→

j'ai <age>10 ans</age>

145 Il est aussi possible d'utiliser la valeur *fraction* de l'attribut *unit* .

146 81. <amount><val>60</val><unit> fédérations départementales </unit></amount> sur <amount><val>95</val></amount>

→

<measure quantity ="60/95" unit ="fraction">60 <personGrp>fédérations départementales</personGrp> sur 95</measure>

147 82. <amount><extractor>une</extractor> des <val>trois</val> <object>filles</object></amount>

→

<measure quantity ="1/3" unit ="fraction">une des trois <person>filles</person></measure>

3.7. Point dans le temps

148 Pour la plupart des exemples, la correspondance entre le guide Quaero et la TEI se fait bien, mais avec deux différences notables. Même si il existe des éléments *month* et *year* dans la TEI (mais non dans TimeML), il est recommandé d'utiliser l'attribut *when* de normalisation[13] : *use the when attribute to provide a normalized equivalent* .

149 83. <time.date.abs><week>lundi</week> <day>25</day> <month>janvier</month> <year>2010</year></time.date.abs>

→

<date when="2010-01-25">lundi 25 janvier 2010</date>

150 C'est l'attribut *when* qui permet la distinction entre dates relatives et dates absolues, soit par omission , soit par incomplétude de la forme normalisée . Comme ces deux notions varient selon la campagne d'évaluation et nous semblent étrangement définies, nous proposons de ne pas introduire un attribut *type* pour les distinguer, ce qui serait une possibilité.

151 84. rendez-vous <time.date.rel><week>mardi</week></time.date.rel>

→

rendez-vous <date>mardi</date>

152 85. <time.date.abs><time-modifier>le</time-modifier> <day>25</day> <month>janvier</month> </time.date.abs>

→

<date when="--01-25">le 25 janvier</date>

153 86. <time.date.abs><time-modifier>en</time-modifier> <year>300</year> <time-modifier>avant JC</time-modifier></time.date.abs>

→

<date when="B0300">en 300 avant JC</date>

154 L'attribut *when* peut être remplacé par les attributs *notBefore*, *notAfter*, *from* et *to* .

155 87. le <time.date.abs><century>21e siècle</century></time.date.abs>

→

le <date from ="2001" to ="2100">21è siècle</date>

156 Les noms de période historique sont traités de même par la TEI .

157 88. le <time.date.abs><reference-era>second empire</reference-era></time.date.abs>

→

le <date notBefore ="1852-12-02" notAfter ="1870-09-04">second empire</date>

158 Lorsqu'un modifieur précise une date par rapport à une autre, on utilise, comme pour les lieux, l'élément *offset* à l'intérieur d'un élément *date* . Bien que les adverbes de temps n'apparaissent pas dans les exemples de la TEI, il nous semble possible de les baliser de la même manière [14].

159 89. <time.date.rel><week>lundi</week> <time-modifier>prochain</time-modifier></time.date.rel>

→

<date><date>lundi</date> <offset>prochain</offset></date>

160 90. <time.date.rel><name>hier</name> <time-modifier>matin</time-modifier></time.date.rel>

→

<date><date>hier</date> <offset>matin</offset></date>

161 Pour les heures, le système est identique en remplaçant l'élément *date* par l'élément *time* .

162 91. <time.hour.abs><val>quinze</val> <unit>heures</unit> <val>trente</val></time.hour.abs>

→

<time when ="15 :30 :00">quinze heures trente</time>

163 92. <time.hour.abs><val>3</val> <unit>heures</unit> <time-modifier>du matin</time-modifier></time.hour.abs>

→

<time when ="03 :00 :00">3 heures du matin</time>

164 93. <time.hour.rel><time-modifier>ce</time-modifier> <name>matin</name></time.hour.rel>

→

<time notBefore="00:00:00" notAfter="12:00:00"><offset>ce</offset> <time>matin</time></time>

165 Comme pour les mesures, un intervalle est balisé par un élément *extent* contenant deux éléments, *date* ou *time* .

166 94. la réunion commencera <time.hour.abs><range-mark>entre</range-mark> <val>3</val> <range-mark>et</range-mark> <val>5</val> <unit>heures</unit></time.hour.abs>

→

la réunion commencera <extent>entre <time when ="03 :00 :00">3</time> et <time when ="05 :00 :00">5 heures</time></extent>

3.8. Événements

167 Dans la TEI, comme dans le guide Quaero, les évènements sont balisés par l'élément *event* ; dans la TEI cet élément peut éventuellement être accompagné d'un attribut *when*.

168 95. Le <event>Tour de <loc.adm.nat>France</loc.adm.nat></event>

→

Le <event>Tour de <country>France</country></event>

169 96. le <event>44e congrès de la <org.ent>CFDT</org.ent></event>

→

le <event><num type ="ordinal" value ="44">44e</num> congrès de la <orgName>CFDT</orgName></event>

170 L'élément *event* de la TEI (ou du guide Quaero) est ambigu – à une différence de casse près – avec l'élément du même nom, *EVENT*, utilisé par TimeML pour baliser les évènements, non au sens des entités nommées, mais au sens syntactico-sémantique. Cette différence de casse permet néanmoins son utilisation.

Conclusion

171 Dans cet article nous avons présenté les différents balisages des entités nommées utilisés dans les campagnes d'évaluation Ester et Etape, ainsi que les normes de la TEI et de TimeML. En comparant l'ensemble des exemples proposés par le guide Quaero avec la TEI, nous avons constaté que le balisage peut être effectué de manière quasiment similaire en utilisant seulement la norme TEI. Afin de stabiliser les évaluations sur le français et de permettre aussi d'utiliser les mêmes balises sur des textes d'autres langues, nous préconisons d'utiliser désormais la TEI pour l'annotation des entités nommées.

172 Notre système de reconnaissance et de balisage des entités nommées, CasEN, disponible en ligne[15], va être modifié dans ce sens.

Remerciements

173 Les auteurs remercient Enza Morale de l'Inist pour sa relecture attentive de la section , ainsi que les relecteurs de la revue Corela pour leurs remarques et suggestions.

BIBLIOGRAPHIE

Artstein R., Poesio M. (2008). Inter-Coder Agreement for Computational Linguistics Computational Linguistics. MIT Press, 34:555-596.

Chinchor N. (1997). Muc-7 Named Entity Task Definition[16].

Ehrmann M. (2008). Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. Thèse de doctorat, Université Paris 7 - Centre de recherche Xerox, Grenoble (XRCE).

Ester-2 (2007). Entités Nommées, Dates, heures et montants. Convention d'annotation, version 0.1.[17].

Ferro L., Mani I., Sundheim B., Wilson G. (2001). TIDES Temporal Annotation Guidelines. Version 1.0.2 MITRE Technical Report, MTR01W0000041.

Fèvre-Pernet C., Roché M. (2005). Quel traitement lexicographique de l'onomastique commerciale ? Pour une distinction Nom de marque/Nom de produit. Revue *Corela*.

Fort K., Nazarenko A., Rosset S. (2012). Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis. Proceedings of the International Conference on Computational Linguistics (COLING 2012).

Gross G. (2012), *Manuel d'analyse linguistique*. Villeneuve-d'Ascq, Presses universitaires du. Septentrion.

Lefeuvre A., Antoine J.-Y., Savary A., Schang E., Abouda L., Maurel D., Eshkol I. (2014). Annotation de la temporalité en corpus : contribution à l'amélioration de la norme TimeML. *TALN 2014*, Marseille.

McNamee P., Dang H. T., Simpson H., Schone P., Strassel S. M. (2010). An evaluation of technologies for knowledge base population, LREC 2010, 369–372.

Merchant R., Okurowski M., Chinchor N. (1996). The multilingual entity task (MET) overview. In Proceedings of a workshop on held at Vienna, Virginia, pages 445–447, Morristown, NJ, USA.

Nadeau D., Sekine S. (2007). A survey of named entity recognition and classification. *Linguisticæ Investigationes*, 30(1).

Pustejovsky J., Castaño J., Ingria R., Saurí R., Gaizauskas R., Setzer A., Katz G., Radev G. (2003). TimeML: A Specification Language for Temporal and Event Expressions. In Proceedings of *the International Workshop of Computational Semantics.*

Rosset S., Grouin C., Zweigenbaum P. (2011).Entités nommées structurées : guide d'annotation Quaero. Notes et documents Limsi :2011-04.

Santos D., Seco N., Cardoso N., Vilela R. (2006). HAREM: An Advanced NER Evaluation Contest for Portuguese. LREC 2006:1986–1991.

Sekine S. (2004). Definition, dictionaries and tagger of extended named entity hierarchy. LREC 2004, Lisbon, Portugal.

Sekine S., Eriguchi Y. (2000). Japanese named entity extraction evaluation: analysis of results. In Proceedings of the 18th conference on Computational Linguistics, Morristown, NJ, USA.

Setzer A. (2001). Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study, PhD dissertation, University of Sheffield.

Text Encoding Initiative Consortium (2012). TEI Lite: Encoding for Interchange: an introduction to the TEI Final revised edition for TEI P5[18].

Text Encoding Initiative Consortium (2015). TEI P5: Guidelines for Electronic Text Encoding and Interchange[19].

Tran M. (2006). Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception, implantation et gestion en ligne. Thèse de doctorat, Université François-Rabelais de Tours.

Tran M., Maurel D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, *Traitement automatique des langues*, vol. 47(3) :115-139.

NOTES

1. http://www.cnts.ua.ac.be/conll2002/ner/
2. http://www.cnts.ua.ac.be/conll2003/ner/
3. Même trois si l'on compte la campagne Ester-1 pour laquelle la tâche d'extraction d'entités nommées est restée à diffusion restreinte.
4. http://www.tei-c.org/
5. Voir cependant (Lefeuvre, 2014) pour une discussion sur la norme TimeML.
6. Tous les exemples cités dans cet article sont extraits des guides d'annotations concernés.
7. Les auteurs de l'article ont bien conscience de la difficulté que représente la constitution d'un tel guide. Ces remarques doivent être prises comme des critiques constructives pour des annotations futures.
8. http://www.quaero.org/
9. L'exemple du guide pour *rendez-vous mardi prochain* comporte les balises *time.date.rel*, contrairement au contexte, ce qui nous conforte dans l'idée que cette règle est difficile à soutenir.
10. Lorsqu'un exemple semblait redondant étant donné notre propos, nous l'avons omis pour ne pas rallonger inconsidérément cet article.
11. Une note du guide Quaero précise qu'il s'agit d'un quartier de Paris.
12. Il semble difficile de penser que le Monopoly n'est pas un objet. La catégorie *prod.other* est étrange ici. Comme le guide Quaero précise, sans les exiger et donc sans les mettre dans les exemples, qu'il est possible de créer des sous-types *véhicule*, *aliment*, *vêtement*, *médicament* et *arme*, cette catégorie *prod.other* devrait plutôt correspondre à une catégorie *prod.object.other*, ce qui justifie notre choix.
13. Cette normalisation suit l'ISO 8601.
14. Ce qui serait plus logique que de baliser un adverbe par l'élément *name* comme le demande le guide Quaero.
15. http://tln.li.univ-tours.fr/Tln_CasEN.html
16. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html
17. http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf
18. http://www.tei-c.org/release/doc/tei-p5-exemplars/pdf/tei_lite.doc.pdf
19. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

RÉSUMÉS

Dans cet article, nous comparons les campagnes existantes d'annotation d'entités nommées avec la TEI. Après une analyse détaillée, nous préconisons une utilisation de la norme TEI afin d'uniformiser et d'adapter les méthodes d'annotation des entités nommées utilisées, notamment celles des campagnes Etape et Ester, menées dans le contexte francophone.

This paper deals with French Named Entity shared tasks. We compare the elements defined in the different guidelines with those defined by the TEI Consortium. We suggest to use the TEI guideline, which is shared by a large community in the world, in future Named Entity tasks.

INDEX

Keywords : Named Entity, tags, evaluation campaign, Etape, Ester, TEI
Mots-clés : Entités nommées, annotations, campagnes d'évaluation, Etape, Ester, TEI

AUTEURS

SOLENN LE PEVEDIC

Université François Rabelais de Tours

Laboratoire d'Informatique

DENIS MAUREL

Université François Rabelais de Tours

Laboratoire d'Informatique