# Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées

Benoît Sagot, Marion Richard, Rosa Stern

# Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées

Benoît Sagot[1] Marion Richard[1,2] Rosa Stern[1,3]
(1) Alpage, INRIA Paris–Rocquencourt & Université Paris Diderot, 175 rue du Chevaleret, 75013 Paris
(2) ISHA, Université Paris Sorbonne, 7 rue Victor Cousin, 75006 Paris
(3) AFP MediaLab, 2 place de la Bourse, 75002 Paris
benoit.sagot@inria.fr, ma.rih.on75@gmail.com, rosa.stern@afp.com

RÉSUMÉ

Le Corpus Arboré de Paris 7 (ou French TreeBank) est le corpus de référence pour le français aux niveaux morphosyntaxique et syntaxique. Toutefois, il ne contient pas d'annotations explicites en entités nommées. Ces dernières sont pourtant parmi les informations les plus utiles pour de nombreuses tâches en traitement automatique des langues et de nombreuses applications. De plus, aucun corpus du français annoté en entités nommées et de taille importante ne contient d'annotation référentielle, qui complète les informations de typage et d'empan sur chaque mention par l'indication de l'entité à laquelle elle réfère. Nous avons annoté manuellement avec ce type d'informations, après pré-annotation automatique, le Corpus Arboré de Paris 7. Nous décrivons les grandes lignes du guide d'annotation sous-jacent et nous donnons quelques informations quantitatives sur les annotations obtenues.

ABSTRACT

**Referential named entity annotation of the Paris 7 French TreeBank**

The French TreeBank developed at the University Paris 7 is the main source of morphosyntactic and syntactic annotations for French. However, it does not include explicit information related to named entities, which are among the most useful information for several natural language processing tasks and applications. Moreover, no large-scale French corpus with named entity annotations contain referential information, which complement the type and the span of each mention with an indication of the entity it refers to. We have manually annotated the French TreeBank with such information, after an automatic pre-annotation step. We sketch the underlying annotation guidelines and we provide a few figures about the resulting annotations.

MOTS-CLÉS : Résolution d'entités nommées, Corpus annoté, Corpus arboré de Paris 7.

KEYWORDS: Named entity resolution, Annotated corpus, French TreeBank.

## 1 Introduction et état de l'art

La notion d'entité nommée (EN) est au cœur d'un nombre considérable de travaux en traitement automatique des langues depuis plusieurs décennies. Elle a notamment fait l'objet des conférences MUC (Marsh et Perzanowski, 1998) et des campagnes associées, puis de campagnes CoNNL (Sang et Meulder, 2003) et ACE (Doddington *et al.*, 2004). Traditionnellement, et notamment dans les campagnes MUC, les EN sont classées en noms de personnes, noms de lieux, noms

d'organisations, et parfois « autres noms propres ». Au-delà de cette définition simple et restrictive, l'usage s'accorde de plus en plus à étendre la notion d'EN à d'autres types, comme les noms de marques et de produits, qui sont également le plus souvent des noms propres, mais également les noms d'œuvres, les dates, les montants, voire les adresses, les URLs, les nombres, voire tout token [1] ou séquence de tokens qui n'a pas vocation à faire partie d'un lexique et qui respecte une grammaire dite locale, spécifique à la nature de ce qu'elle dénote (Sekine et Nobata, 2004; Grouin *et al.*, 2011). Ce que tous ces types d'EN ont en commun est leur caractère référentiel, qui se décline toutefois de façon différente suivant les types, voire suivant les corpus. C'est ainsi qu'Ehrmann (2008) définit une entité nommée comme suit : *Étant donné un modèle applicatif et un corpus, on appelle EN toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.*

La tâche de **reconnaissance d'EN** est généralement conçue comme ayant pour objectif d'identifier automatiquement en corpus les *mentions* d'EN, c'est-à-dire les séquences de tokens qui réfèrent à une entité, mais également de typer ces mentions, par exemple avec des catégories telles que `Person`, `Location`, `Organization`. De très nombreux travaux ont été publiés depuis longtemps sur cette tâche, qui a été abordée entre autres avec des méthodes à base de règles et, souvent, de ressources lexicales (« gazetteers », ou lexiques de mentions d'entités associées à des types) (Sekine et Nobata, 2004; Rosset *et al.*, 2005; Stern et Sagot, 2010), des techniques statistiques d'apprentissage reposant sur des corpus annotés (Finkel *et al.*, 2005; Bechet et Charton, 2010), et des techniques hybrides. Ces travaux accordent souvent, sous une forme ou sous une autre, une importance particulière au problème du typage en lien avec des phénomènes tels que la métonymie : en effet, on peut distinguer deux façons de typer une entité, soit de façon intrinsèque (*la France* dénote un lieu), soit en contexte (dans *La France a signé le traité*, *France* peut être typé comme une organisation).

Qu'il s'agisse de permettre l'évaluation de ces systèmes, ou leur entraînement dans le cas de systèmes statistiques, des corpus annotés ont été développés pour diverses langues. Pour le français, on peut citer notamment les corpus ESTER et ESTER2 (60 plus 150 heures d'émissions transcrites orthographiquement et annotées en EN) (Galliano *et al.*, 2009), ainsi que le corpus Quaero (Grouin *et al.*, 2011). Tous ces corpus reposent sur des données orales transcrites (émissions de radio). On notera que le corpus Quaero repose sur une définition originale, très riche et structurée de la notion d'EN (Rosset *et al.*, 2011). De plus, il contient des annotations de typage à la fois en termes de type absolu et de type en contexte, contrairement aux corpus ESTER qui ne contiennent que le type en contexte.

Toutefois, la seule tâche de reconnaissance des EN ne suffit pas à extraire les informations nécessaires pour des applications comme l'extraction d'informations. Seul un système permettant d'associer à ces mentions un référent extra-linguistique permet d'exploiter le résultat de la détection grâce au sens qui lui est ainsi conféré (Blume, 2005). Cette tâche dite de **résolution des EN** (REN, *entity linking*) consiste ainsi à associer à chaque mention d'EN l'entrée adéquate dans une base d'entités qui sert de référence, en traitant notamment les cas d'homonymie : une mention de *Michael Jordan* réfère-t-elle au footballeur, au joueur de basket ou à l'économiste ? Une mention d'*Orange* réfère-t-elle à une ville (laquelle ?) ou à l'entreprise ? Outre les difficultés liées à l'homonymie des mentions, mais également aux phénomènes de métonymie, on est également confronté à la diversité des mentions pour une même entité (variantes graphiques : *Jacques Chirac* et *J. Chirac*, surnoms : *Ali le chimique* pour *Ali Hassan al-Majid*).

1. Un token, contrairement à une (occurrence de) forme, est une unité typographique (Sagot et Boullier, 2008).

La résolution automatique d'EN fait l'objet de travaux depuis quelques années sur l'anglais (Bunescu et Pasca, 2006; Cucerzan, 2007; McNamee et Dang, 2009). C'est souvent Wikipedia qui est utilisé comme base d'entités de référence. Pour le français, Stern et Sagot (2010) proposent un système à base de règles intégré à la chaîne de traitements de surface SxPipe(Sagot et Boullier, 2008), nommé NP. Ce système repose sur la base d'entités Aleda (Sagot et Stern, 2012), extraite automatiquement à partir de Wikipedia et de la base de noms de lieux Geonames. Cependant, peu de corpus annotés en référence, associés à une base d'entités, sont disponibles pour évaluer ou entraîner de tels systèmes. Pour l'anglais, on peut citer le corpus rendu disponible pour la tâche de peuplement de base de données de la campagne TAC 2009 (McNamee et Dang, 2009). Pour le français, le seul corpus disponible est constitué de 100 dépêches de l'Agence France-Presse de 300 mots chacune en moyenne, qui utilise comme référence la base Aleda (Stern et Sagot, 2010). Outre les informations référentielles, ce corpus inclut naturellement pour chaque mention annotée les informations d'empan et de type, plus précisément de type absolu, en cohérence avec le type de l'entité tel qu'il est indiqué (ou devrait l'être) dans Aleda.

Dans cet article, nous décrivons un travail d'annotation des entités nommées du Corpus Arboré de Paris 7, ou French TreeBank (Abeillé *et al.*, 2003), avec les mêmes principes que le petit corpus de Stern et Sagot (2010). L'objectif est triple :

– Préciser les conventions d'annotation en les confrontant à un corpus plus important et moins contemporain des ressources utilisées pour construire Aleda ;
– Fournir à la communauté un corpus écrit de taille importante dont les EN soient annotées en empan, type et référence ;
– Ajouter une couche d'annotation à un corpus pour lequel d'autres niveaux d'annotations sont disponibles (à ce jour, annotations morphosyntaxiques, arbres de constituance et fonctions syntaxiques ; à terme, annotations de type FrameNet et annotations discursives).

Nous décrivons donc succinctement les conventions d'annotations utilisées (section 2), le processus d'annotation guidé par une pré-annotation effectuée à l'aide du système NP mentionné ci-dessus (section 3) et les résultats obtenus (section 4).

# 2 Conventions d'annotation

## 2.1 Principes généraux

Nous avons annoté le Corpus Arboré de Paris 7 en indiquant l'empan, le type absolu [2] parfois complété d'un sous-type et l'identifiant du référent dans la base Aleda de toute mention d'EN sous forme de nom propre, à l'exclusion de tout autre type d'expression référentielle (descriptions définies, pronoms...). Nous nous sommes restreints aux noms de personnes, de lieux, d'organisations, d'entreprises, et à certains noms de produits. Une annotation systématique des noms de produits et des noms d'œuvres sera à effectuer ultérieurement. En revanche, nous ne nous sommes intéressés ni aux EN moins standard (URL, pourcentages...) ni aux expressions temporelles. Enfin, nous n'avons annoté aucune EN enchâssée dans une autre.

Plus précisément, nous avons fait usage de 7 types de base : `Person`, `Location`,

2. L'entité *France* sera donc toujours une entité `Location` avec pour sous type `Pays`, comme indiqué dans Aleda pour le référent correspondant, même si le sens contextuel réfère à l'organisation politique, au peuple français, à l'équipe de foot, etc.

`Organization`, `Company`, `Product`, `POI` (Point of Interest) et `FictionChar` (personnage de fiction). Ces types sont parfois précisées par un sous-type. Les types et sous-types ont été organisés via une ontologie, dans laquelle les types sont des classes qui sont à la tête d'une hiérarchie de sous-classes qui correspondent à des sous- types. Pour sous-typer une entité il n'est donc pas nécessaire de préciser la totalité des sous-classes qui lui correspondent.

Notre définition de ce qu'est une EN conduit à des cas limites, notamment pour les mentions qui n'ont pas de référent autonome en soi, mais qui en acquièrent un en contexte, comme par exemple *banque centrale*. Dans ce type de cas, nous avons considéré qu'il y avait bien mention d'EN, et nous avons donc annoté, pour peu que le contexte donné permette d'établir quelle est la banque précise dont il est question, à la condition (arbitraire) supplémentaire que la mention commence par une majuscule. Ainsi, une mention comme *banque centrale* sera systématiquement ignorée. En revanche, les mentions primaires d'entités qui ne dépendent pas du contexte sont annotées qu'elles aient ou non des majuscules, comme par exemple *banque mondiale*. Cette situation se retrouve par exemple également dans le cas de l'annotation des noms d'universités. Nous considérons ainsi qu'*université de Nantes* dénote une université située à Nantes, et nous n'annotons que la ville de Nantes, alors qu'*Université de Nantes* fait directement référence à l'organisation qu'est cette université, et nous annotons donc l'ensemble comme une organisation. Il en va de même, par exemple, pour *Université de Montpellier*, puisqu'il n'existe pas d'organisation unique qui corresponde à ce terme : dans ce cas, seul *Montpellier* est annoté, en tant que ville.

Les mentions annotées peuvent correspondre au nom normalisé (*Jacques Chirac*), à une variante (*Chirac* dans *M. Chirac*, cf. plus bas concernant *M.*) ou à un surnom (comme dans *l'Hexagone*. Ainsi, la description définie *l'avocat de M. Chirac* entraînera une annotation de la mention *Chirac*, en ignorant la référence à l'avocat de ce dernier. Par ailleurs, les mots grammaticaux ou contextuels entourant la mention de l'entité sont ignorés. Ainsi les déterminants ne sont pas pris en compte, ni les titres, professions ou adjectifs pouvant apparaître pour qualifier l'entité. Ainsi, dans *Chine méridionale*, seul *Chine* est annoté comme un nom de lieu, et dans *M. Bill Clinton* seul *Bill Clinton* est annoté comme un nom de personne.

Les balises utilisées pour l'annotation contiennent les informations suivantes :
– l'identifiant de l'EN dans Aleda (attribut `eid`) ; dans le cas d'une entité non présente dans la base l'identifiant est marqué `null`.
– le nom normalisé de l'entité, tel qu'indiqué dans Aleda ; pour les lieux il s'agit du nom donné dans GeoNames et pour les autres entités du titre de l'article dans la Wikipedia française.
– un type, ainsi qu'un sous-type dans le cas où l'entité entre dans une catégorie sous-typée (cf. section ci-dessous).

Voici deux exemples d'annotation :

```
<ENAMEX type="Organization" eid="1000000000016778" name="Confédération
française démocratique du travail">CFDT</ENAMEX>
<ENAMEX type="Location" sub_type="Country" eid="2000000001861060"
name="Japan">Japon</ENAMEX>
```

Dans certaines balises se trouve une information supplémentaire pour les cas de fusions d'entreprises, de changement de nom d'une entreprise et de lieux géographiques qui n'existent plus en tant que tels :
– `current_eid` est l'identifiant dans Aleda d'une entité actuelle correspondant à une entité qui n'existe plus mais qui est le référent réel (par exemple en cas de rachat d'une entreprise par une autre) ;
– `current_name` est le nom normalisé de l'entité `current_eid`.

Si une entreprise a changé de nom sans que l'on puisse considérer qu'il y a eu changement de référent, on ajoute un attribut `former_name` qui permet d'indiquer que la mention de l'entité réfère à l'ancien nom de l'entreprise. Enfin un attribut `former_location` initialisé à `True` permet d'annoter un lieu géographique disparu.

## 2.2 Types et sous-types

L'annotation des **noms de personnes** ne pose pas de problèmes particuliers. Deux précisions toutefois concernant deux cas :
- Les groupes de personnes : Les références à des groupes de personnes, telles que *la famille Agnelli* ou *les frères Maxwell*, ne sont pas annotées. Il en est de même pour les populations : le groupe de personnes désigné par Les Français n'est pas pris en compte dans l'annotation.
- Les fonctions ou titres : Les fonctions ne sont pas annotées. Une expression telle que *le Premier ministre M. Fillon* permettra d'annoter *Fillon* mais ne tiendra pas compte de la mention *Premier ministre* en tant qu'EN (cela suit le fait que nous ne retenons pas les mentions sous forme de description définie). Dans la lignée, l'expression *Général de Gaulle* donnera lieu à l'annotation de la mention *de Gaulle*, mais *Général* ne sera pas inclus dans l'empan.

Le type `FictionChar` permet d'annoter toutes les mentions qui font référence à des **noms de personnes ou d'animaux fictifs**, telles que *McGyver* ou *Zorro*.

Parmi les **noms de lieux**, les sous-types utilisés sont les suivants :
- Sont sous-typés `Country` les états indépendants ;
- Les divisions territoriales des pays sont annotées `CountryDivision` (chaque pays ayant sa propre gestion du territoire, des sous-types tels que *département*, *etat fédéral*, *canton*, *district*, *comté* ne semblaient pas pertinents) ;
- le sous-type `Region` permet d'annoter des parties du monde non liées à une gestion politique. Ce sous-type regroupe les continents et parties de continents, ainsi que des lieux géographiques sans frontières mais dans le vocabulaire courant tel que la région du Golfe ;

Le type « Point Of Interest » (`POI`) est utilisé pour les entités telles que les ports, les salles de spectacle, les stades, les quartiers, etc.

Nous considérons comme des **noms d'organisations** et typons `Organization` toutes les références à des organisations qu'elles soient politique, éducative, économique, etc, à l'exclusion des noms d'entreprises. Un seul sous-type est utilisé, `PoliticalGroup` pour les organisations politiques. Enfin, les **noms d'entreprises**, typés `Company`, ne sont pas sous-typés.

## 2.3 Difficultés génériques

L'annotation est parfois rendue difficile du fait de l'ambiguïté de l'entité ou de sa catégorisation. En contexte journalistique, les principaux cas de difficulté sont liés à la temporalité du corpus. En effet les informations présentes dans un corpus journalistique dépendent essentiellement du contexte temporel. Ainsi, en 2011, année dont datent les informations ayant servi à construire la base de référence Aleda, la Tchécoslovaquie, l'URSS ou l'entreprise Thomson CSF n'existent plus, mais en 1990, date de rédaction des textes constituant le corpus, ces pays existaient encore. Faire l'impasse sur ces entités serait faire l'impasse sur une partie de l'information passée, nous avons donc défini des règles d'annotation pour ces entités particulières.

Les mentions d'entreprises peuvent ne plus avoir de référence actuelle. Si l'entreprise a simplement disparu, son référent est annoté `null`. Mais cela peut aussi être dû à un changement de nom, à un rachat ou à une fusion avec une autre entreprise. Dans les cas d'un changement de nom, nous avons inclus l'ancien nom dans un attribut `former_name`.

```
<ENAMEX type="Organization" eid="1000000000036708" name="France 2"
former_name="Antenne 2">Antenne 2</ENAMEX>
```

Dans le cas du rachat d'une entreprise ou d'une fusion, nous ajoutons deux sous-types pour permettre d'identifier la référence actuelle de l'entreprise en opposition à la référence du texte, (datant ici des années 90). Ainsi nous conservons la valeur temporelle du texte dans l'annotation tout en actualisant la référence. Par exemple à l'époque de la rédaction des dépêches du corpus le groupe UAP n'avait pas encore fusionné avec le groupe AXA, les attributs `current_eid` et `current_name` permettent d'actualiser la référence en donnant l'information que le groupe s'appelle désormais AXA.

```
<ENAMEX type="Company" eid="null" name="Union des assurances de Paris"
current_eid="1000000000201762" current_name="AXA">UAP </ENAMEX>
```

Quant aux filiales, elles sont annotées seules c'est à dire sans référence à leur entreprise mère. Lorsqu'il s'agit de petites filiales qui n'ont pas de référent, comme c'est le cas de nombreuses fois, l'identifiant sera annoté `null`. Les filiales étrangères connues d'Aleda des grosses entreprises sont annotées en tant que telles.

Plusieurs mentions faisant référence à une entreprise ou une usine sont citées via leur marque (ex. : *Mamie nova*). Dans ce cas, l'annotation est faite sur la marque. Dans une expression comme *un tracteur John Deer*, l'entité *John Deer* est annotée en tant qu'entreprise, pour rester cohérent avec l'idée de départ qui est de considérer les EN dans leur sens absolu. Le choix d'annoter une EN présentant une ambiguïté entre nom de produit ou d'entreprise est guidé par le type associé à l'entité dans la base de données Aleda (par exemple, *Nike* et *Adidas* sont annotés `Company`, mais *Evian* ou *Lessieur* sont annotés `Product`).

Pour les dénominations géographiques qui ne sont plus d'actualité à la date d'extraction d'Aleda mais qui sont utilisées dans le corpus, nous ajoutons un sous-type `former-location` de type `boolean` qui est annoté `true` par défaut. Ces lieux ne possédant pas d'identifiant dans la base de données ils sont identifiés `null`, mais cela permet leur annotation et reconnaissance dans le corpus.

```
<ENAMEX type="Location" sub_type="Country" eid="null" former_location="true"
name="Tchécoslovaquie">Tchécoslovaquie </ENAMEX>
```

Outre ces difficultés relativement générales, certaines entités ou mentions ont posé des problèmes spécifiques sur lesquels il a fallu faire des choix. C'est le cas de *Trésor (public)* [3], *Hachette* [4] et *Thomson* [5] qui correspondent à des organisations qui ont fortement évolué entre la date de rédaction du corpus et la date de création d'Aleda.

3. Certains pays disposant de plusieurs organismes se divisant les tâches associées à la notion générale de trésor public. Dans ce cas, nous choisissons le ministère des finances comme référent de la mention *Trésor (public)*.

4. Toutes les occurrences de *Hachette* sont annotées sous la référence au Groupe Lagardère, propriétaire de *Hachette livres* et de *Hachette Filippachi Médias*. *Hachette* ne correspond plus aujourd'hui à une entité unique.

5. À l'époque de rédaction du corpus, la mention *Thomson* pouvait faire référence à deux entreprises : Thomson CSF aujourd'hui devenu Thalès et Thomson SA aujourd'hui Technicolor. Dans les cas les plus simples la mention est suivi de CSF ou SA, et on donne donc le référent Thalès ou Technicolor, en indiquant le `former_name`. Les occurrences de *Thomson* seul sont désambiguïsées en contexte.

## 3 Processus d'annotation

L'intégralité du Corpus Arboré de Paris 7 dans sa version 2007 (à l'exclusion des phrases n'ayant pas reçu d'annotations fonctionnelles), soit 12 351 phrases contenant 350 931 tokens, a été annoté en EN (empan, type, référence) conformément aux directives esquissées ci-dessus. L'annotation a consisté en une validation ou correction manuelle dans un éditeur XML du résultat de SxPipe/NP utilisé comme pré-annotateur. Le résultat de cette campagne d'annotation est à considérer comme préliminaire, puisqu'une seule personne a annoté le corpus. Nous sommes bien conscients des problématiques liées à toute tâche d'annotation manuelle, et en particulier à l'annotation en EN, problématiques analysées par exemple par Fort *et al.* (2009). L'effort d'annotation devrait donc être poursuivi, notamment en obtenant une deuxième annotation, qui permettrait la mesure d'accords inter-annotateurs ainsi qu'une adjudication des cas de désaccord.

## 4 Résultats et perspectives

Au total, 5 890 des 12 351 phrases contiennent au moins une mention d'EN. Au total, 11 636 mentions ont été annotées, qui se répartissent en 3761 noms de lieux, 3357 noms d'entreprises, 2381 noms d'organisations, 2025 noms de personnes, 67 noms de produits, 29 noms de personnages de fiction et 15 POI.

Outre une amélioration de la qualité de l'annotation, comme évoqué ci-dessus, nous prévoyons d'utiliser ce nouveau corpus annoté de multiples façons. Tout d'abord, il pourra être utilisé pour entraîner et évaluer des systèmes de reconnaissance et/ou de résolution d'entités nommées. Une comparaison avec le petit corpus de Stern et Sagot (2010), composé de textes relevant d'un domaine proche mais distinct (dépêches d'agence), sera en ce sens utile. Mais la disponibilité d'informations morphosyntaxiques et syntaxiques permettra des expériences intéressantes impliquant par exemple des modèles joints de reconnaissance d'EN et d'étiquetage morphosyntaxique, ou de reconnaissance d'EN et d'analyse syntaxique.

## Remerciements

Ce travail a été financé par le projet ANR EDyLex (ANR-009-CORD-08).

## Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for French. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Bechet, F. et Charton, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. *In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

BLUME, M. (2005). Automatic entity disambiguation : Benefits to ner, relation extraction, link analysis, and inference. *International Conference on Intelligence Analysis*.

BUNESCU, R. et PASCA, M. (2006). Using encyclopedic knowledge for named entity disambiguation. *In Proceedings of EACL*, volume 6, pages 9–16.

CUCERZAN, S. (2007). Large-scale named entity disambiguation based on wikipedia data. *In Proceedings of EMNLP-CoNLL*, volume 2007, pages 708–716.

DODDINGTON, G., MITCHELL, A., PRZYBOCKI, M., RAMSHAW, L., STRASSEL, S. et WEISCHEDEL, R. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. *In Proceedings of LREC - Volume 4*, pages 837–840.

EHRMANN, M. (2008). *Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation*. Thèse de doctorat, Université Paris 7 Denis Diderot.

FINKEL, J. R., GRENAGER, T. et MANNING, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

FORT, K., EHRMANN, M. et NAZARENKO, A. (2009). Towards a methodology for named entities annotation. *In Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 142–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The Ester 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. *In Interspeech 2009*.

GROUIN, C., ROSSET, S., ZWEIGENBAUM, P., FORT, K., GALIBERT, O. et QUINTARD, L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. *In Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, OR. Association for Computational Linguistics.

MARSH, E. et PERZANOWSKI, D. (1998). Muc-7 evaluation of ie technology : Overview of results. *In Proceedings of the Seventh Message Understanding Conference (MUC-7) - Volume 20*.

MCNAMEE, P. et DANG, H. (2009). Overview of the tac 2009 knowledge base population track. *In Text Analysis Conference (TAC)*.

ROSSET, S., GROUIN, C. et ZWEIGENBAUM, P. (2011). Entités nommées structurées : guide d'annotation Quaero. Notes et Documents 2011-04, LIMSI, Orsay, France.

ROSSET, S., ILLOUZ, G. et MAX, A. (2005). Interaction et recherche d'information : le projet Ritel. *Traitement Automatique des Langues*, 46(3):155–179.

SAGOT, B. et BOULLIER, P. (2008). SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, 49(2):155–188.

SAGOT, B. et STERN, R. (2012). Aleda, a free large-scale entity database for French. *In Proceedings of LREC*. To appear.

SANG, E. F. T. K. et MEULDER, F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *In Proceedings of CoNLL-2003*, pages pp. 142–147.

SEKINE, S. et NOBATA, C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *In Proceedings of LREC 2004*, Lisbon, Portugal.

STERN, R. et SAGOT, B. (2010). Resources for named entity recognition and resolution in news wires. *In Proceedings of LREC 2010 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, La Valette, Malte.