Cascades de transducteurs autour de la reconnaissance des entités nommées

Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol, Damien Nouvel

Cascades de transducteurs autour de la reconnaissance des entités nommées

Denis Maurel[*] — Nathalie Friburger[*] — Jean-Yves Antoine[*, *] — Iris Eshkol-Taravella[**] — Damien Nouvel[*]**

[*]*Université François Rabelais Tours - Laboratoire d'informatique*
{Denis.Maurel, Jean-Yves.Antoine, Nathalie.Friburger, Damien.Nouvel}@univ-tours.fr
[**]*Université d'Orléans - Laboratoire ligérien de linguistique*
iris.eshkol@univ-orleans.fr
[***]*Lab-STICC CNRS*

RÉSUMÉ. *Cet article présente tout d'abord la cascade de transducteurs CasEN pour la reconnaissance des entités nommées. CasEN est implantée sous le logiciel CasSys de la plate-forme Unitex et est librement mise à disposition des utilisateurs sous licence LGPL-LR. Après une discussion sur la typologie des entités nommées qu'elle utilise et une description du fonctionnement de la cascade, nous rapportons son évaluation sur le corpus Eslo 1 (corpus d'Orléans) et les performances qu'elle a montrées au cours de la campagne d'évaluation Ester 2. Nous présentons ensuite deux autres cascades qui utilisent le texte étiqueté par CasEN. La première ajoute des informations sur les locuteurs de l'enquête sociolinguistique Eslo 1 et la seconde met en relation des entités nommées dans un corpus du journal* Le Monde.

ABSTRACT. *This paper presents first the CasEN transducer cascade to recognize French Named Entities. CasEN is implemented with the CasSys software of the Unitex plateform and is put at user free disposal (LGPL-LR license). We discuss about Named Entity typology used and we describe the cascade, before reporting its evaluation from Eslo 1 corpus and evaluation campaign Ester 2 corpus. Second, we present two other cascades using a CasEN annotated text. The first cascade adds speaker information into the Eslo socio-linguitic survey and the second one links named entities in* Le Monde *newspaper corpus.*

MOTS-CLÉS : *cascades de transducteurs, Entités nommées, CasSys, CasEN, Eslo 1, Ester 2.*

KEYWORDS: *transducer cascades, Named entities, CasSys, CasEN, Eslo 1, Ester 2.*

1. Introduction

Dans cet article, nous présentons précisément un système symbolique pour la reconnaissance des entités nommées, CasEN (Maurel *et al.*, 2009). Depuis les conférences MUC (*Message Understanding Conferences*), la recherche des entités nommées (noms propres, dates et heures, monnaies) (Chinchor, 1997) est une tâche à part entière du TAL. Une introduction à ce domaine peut être trouvée dans (Ehrmann, 2008) et un état de l'art sur les différents systèmes de reconnaissance des entités nommées se trouve dans (Nadeau et Sekine, 2009). Comme ailleurs dans le TAL, deux approches sont concurrentes (ou complémentaires dans des systèmes hybrides, comme (Béchet *et al.*, 2011)), celles centrées sur les données et les techniques d'apprentissage, d'une part, et celles, symboliques, à base de règles, d'autre part. D'après la campagne Ester 2, il semble que les approches symboliques sont pour le moment celles qui donnent les meilleurs résultats... surtout si on a les moyens de développer des ressources lexicales et syntaxiques d'envergure, ce qui était le cas pour les deux concurrents arrivés premiers. Parmi les approches symboliques, plusieurs utilisent des transducteurs à nombre fini d'états (Poibeau, 2003), éventuellement passés en cascade (Ait-mokhtar et Chanod, 1997 ; Hobbs *et al.*, 1997 ; Friburger, 2002 ; Bontcheva *et al.*, 2002), ce qui est la technique que nous utilisons aussi.

Le système CasEN est implanté avec le programme de création de cascades de transducteurs à états finis CaSys (Friburger et Maurel, 2004). CasSys est disponible sur la plate-forme Unitex[1] (Paumier, 2003), à partir de la version 2.1, sous licence LGPL [2]. La cascade CasEN est disponible sur le site des projets TAL du LI[3], sous licence LGPL-LR. La section 2 compare la typologie des noms propres que nous avons adoptée pour CasEN à celles de Coates-Stephens (1993), de Paik *et al.* (1996) et de Tran et Maurel (2006), puis décrit la cascade elle-même. Nous présentons ensuite, dans la section 3, son évaluation sur le corpus Eslo 1 (corpus d'Orléans) et les performances qu'elle a montrées au cours de la campagne d'évaluation Ester 2[4].

La détection des entités nommées est une étape d'outillage de l'analyse qui peut servir à des applications plus spécifiques dans le cadre d'une démarche incrémentale. Nous avons complété notre travail, dans un deuxième temps, par de la recherche d'information sur des corpus étiquetés par CasEN, à l'aide de deux cascades dédiées à cette tâche, d'une part, dans le cadre du projet ANR VariLing et, d'autre part, dans celui du projet Feder Entités. La section 4 décrit comment l'ajout ultérieur de cascades de transducteurs peut être utilisé à des fins particulières : la détection des entités dénommantes réalisée sur le corpus Eslo 1 et un travail sur la caractérisation des relations directes entre entités nommées d'un corpus du *Monde*.

1. http ://www-igm.univ-mlv.fr/~unitex/
2. On y accède par le menu *Text/Apply CaSys Cascade*. Il peut aussi, comme les autres programmes Unitex, être utilisé en ligne de commande.
3. http ://tln.li.univ-tours.fr/Tln_CasEN.html/
4. Précisons cependant que la cascade disponible n'est pas exactement celle utilisée pour la campagne Ester 2, des évolution ayant eu lieu depuis, comme cela est expliqué sur le site.

2. CasEN, une cascade de reconnaissance des entités nommées

Avant même de concevoir une cascade de reconnaissance des entités nommées, il faut définir l'objet de notre recherche, c'est-à-dire les entités nommées que nous cherchons à étiqueter. Nous ne nous attarderons pas sur la définition elle-même, renvoyant pour cela le lecteur à (Ehrmann, 2008). Parlons de la typologie adoptée, le choix de cette typologie précède et conduit bien sûr la construction de la cascade.

2.1. *La typologie retenue*

Il existe un grand nombre de typologies des entités nommées. Pour plus d'information, on pourra consulter (Tran, 2006). Citons tout d'abord la plus célèbre, celle de la tâche de reconnaissance des entités nommées de la conférence MUC 7 (Chinchor, 1997), qui comporte sept types répartis en trois classes :

1) *ENAMEX* (Entity names) : Persons, Locations, Organizations ;

2) *TIMEX* (Temporal expressions) : Dates, Hours ;

3) *NUMEX* (Number expressions) : Percentages values, Monetary values.

Cette typologie, comme la plupart, comporte deux niveaux. Citons aussi celle de (Coates-Stephens, 1993), qui l'a précédée : elle comporte huit types, mais un seul niveau hiérarchique :

1) noms de personne ;

2) noms de lieu ;

3) noms d'organisation ;

4) noms d'origine ou gentilés ;

5) noms de législation ;

6) noms de source d'information (média, journaux, etc.) ;

7) noms d'événement (guerres, révolutions, catastrophes, etc.) ;

8) noms d'objet (artefacts, produits, etc.).

Les trois premiers types sont les *Enamex*, les autres sont complètement extérieurs à la classification de MUC 7. Des ambiguïtés surgissent : *Le Monde* est-il une *organisation* ou une *source d'information* ? Cela dépendra du contexte... De même un produit peut porter le nom d'une marque qui, elle-même peut être le nom d'une entreprise, d'où une autre ambiguïté parfois entre les types *Noms d'organisation* et *Noms d'objet* (Petit, 2006)...

À peu près au même moment que la conférence MUC 7, Paik *et al.* (1996) proposent une typologie à deux niveaux (voir la figure 1). Plus importante, elle peux contenir les deux précédentes, à condition par exemple que les *Noms de législation* et les *Noms de source d'information* deviennent des sous-types du type *Document*, et aussi, sans doute, qu'à l'inverse, les types *Equipment* et *Scientific* se confondent dans

le type *Noms d'objet*.... Reste la question des événements non traités, sauf à les typer *Miscelaneous*, ce qui ne semble pas une bonne idée...

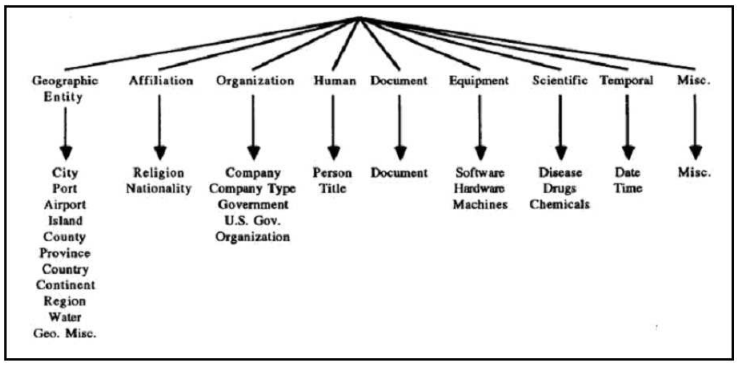

Figure 1. *La typologie de Paik, Liddy, Yu et McKenna*

Laissons là comparaisons et énumérations. Deux constatations sont immédiates : la possibilité d'imbriquer plusieurs niveaux et la précision finale attendue... Avec une difficulté certaine pour le recoupement d'une typologie à une autre... Dans le cadre du projet Prolex, nous avions nous aussi conçu une typologie (Tran et Maurel, 2006), rappelée dans le tableau 1, qui s'inspirait des précédentes, mais aussi de classifications linguistiques, comme celle de Bauer (1985), reprise par Grass (2000). Nous avions souligné à l'époque deux points importants : limiter le nombre final de types pour éviter l'encyclopédisme et jouer sur plusieurs niveaux (et non seulement deux). En effet, une entité peut être difficilement typée *Association* ou *Entreprise*, mais peut être typée *Groupe*, ou même *Anthroponyme collectif*... D'autre part, les ambiguïtés décrites ci-dessus sont en fait des ambiguïtés associables aux types et non aux entités nommées elles-mêmes. Par exemple, tous les toponymes sont susceptibles d'une interprétation comme anthroponyme collectif, toutes les associations ou entreprises peuvent être considérées dans certains contextes comme un toponyme ou un ergonyme, etc. Le tableau 2, lui aussi extrait de (Tran et Maurel, 2006), présente cette ambiguïté inhérante à certains types. Le choix que nous avons fait d'associer à chaque entité un type principal nous a permis de concevoir un système de reconnaissance utilisant des grammaires locales. En effet, nous verrons section 3.4 que ce choix n'a pas été celui de la campagne Ester 2 et que cela nous a posé des problèmes difficiles à résoudre avec notre modèle.

Cependant la nécessité de l'évaluation et surtout l'idée de faire émerger, au moins pour le français, une typologie commune aux différents laboratoires, nous ont conduits à utiliser les types de la campagne Ester 2, tout en ajoutant d'autres types, afin de couvrir l'ensemble de la typologie Prolex. Les sous-types de cette typologie, représentés par des chaînes concaténées séparées par des points (par exemple *org.pol* pour *organisation politique*) ont été éclatés en une série de traits pour s'adapter au formalisme

Nom propre						
Anthroponyme			Ergonyme	Pragmonyme	Toponyme	
Individuel	Collectif					
		Groupe				Territoire
Célébrité	Dynastie	Association	Objet	Catastrophe	Astronyme	Pays
Patronyme	Ethnonyme	Ensemble	Œuvre	Fête	Édifice	Région
Prénom		Entreprise	Pensée	Histoire	Géonyme	Supranational
Pseudo-anthroponyme		Institution	Produit	Manifestation	Hydronyme	
		Organisation	Vaisseau	Météorologie	Ville	
					Voie	

Tableau 1. *La typologie primaire Prolex*

Types	**Types secondaires**
Pays Région Supranational Territoire	Anthroponyme collectif
Ville	Anthroponyme collectif Ergonyme
Édifice Voie	Ergonyme
Fête Histoire Manifestation	
Association Ensemble Entreprise Groupement Institution Organisation	Ergonyme Toponyme
Vaisseau	Toponyme

Tableau 2. *La typologie secondaire Prolex*

d'Unitex (+*org* +*pol*) ; de même le nom de la balise a été lui aussi transformé en trait (+*entity*). Le tableau 3 présente les types hérités de la campagne Ester.

En comparant les deux typologies (Prolex et Ester), nous arrivons à mettre en correspondance certains éléments, comme présenté sur le tableau 4. Il manque un sous-type particulier pour les dynasties (*les Capétiens*), ainsi qu'un type pour les événements (*event*), jugés trop difficiles à annoter dans la campagne.

Personne (+pers)	humain réel ou fictif (+hum) animal réel ou fictif (+anim)	
Fonction (+fonc)	politique (+pol) militaire (+mil) administrative (+admi) religieuse (+rel) aristocratique (+ari)	
Organisation (+org)	politique (+pol) éducative (+edu) commerciale (+com) non commerciale (+non-profit) média et divertissement (+div) géo-socio-administrative (+gsp)	
Lieu (+loc)	géographique naturel (+geo) région administrative (+admi) axe de circulation (+line) construction humaine (+fac) adresse (+addr)	adresse postale (+post) téléphone et fax (+tel) adresse électronique (+elec)
Production humaine (+prod)	moyen de transport (+vehicule) récompense (+award) œuvre artistique (+art) production documentaire (+doc)	
Date et heure (+time)	date (+date) heure (+hour)	date absolue (+abs) date relative (+rel)
Montant (+amount)	valeur physique (+phy) valeur monétaire (+cur)	âge (+age) durée (+dur) température (+ temp) longueur (+ len) surface et aire (+area) volume (+vol) poids (+ wei) vitesse (+spd) autre (+other)

Tableau 3. *Les types hérités de la campagne Ester*

D'autre part, il nous a semblé étrange de ne pas considérer qu'un certain nombre de modifieurs de l'entité n'en faisait pas partie, comme par exemple les nationalités, car cela entraînait des incohérences linguistiques. D'après le guide de la campagne Es-

Prolex	Ester
Anthroponyme/Individuel	+pers
Personne	+pers+hum
Pseudo anthroponyme	+pers+anim / +pers
Anthroponyme/Collectif	+org
Association	? +org+pol / +org+gsp
Ensemble	+org+div
Entreprise	+org+com
Institution	?+org+pol / +org+edu / +org+gsp
Organisation	?+org+gsp / +org+non-profit
Ergonyme	+prod
Œuvre	+prod+art
Produit	?+prod+award / +prod+doc / +prod+vehicule
Vaisseau	?+prod+vehicule
Toponyme	+loc
Astronyme	+loc+geo
Édifice	+loc+fac
Géonyme	+loc+geo
Hydronyme	+loc+geo
Ville	+loc+admi
Voie	+loc+line
Toponyme/Territoire	+loc+admi

Tableau 4. *Correspondance Prolex-Ester*

ter 2, il fallait annoter entièrement *le président du Gabon Omar Bongo* et en deux fois *le président gabonais Omar Bongo.* De même, il fallait en effet annoter uniquement :

– *Sarkozy* dans *Monsieur Sarkozy* ;

– *Raymond Domenech* dans *l'entraîneur Raymond Domenech* ;

– *Nouri al-Maliki* dans *le chiite Nouri al-Maliki*.

Nous avons donc ajouté onze sous-types au type *+pers +hum* pour annoter complètement ces entités nommées. L'ensemble de nos ajouts sont présentés dans le tableau 5. C'est cette typologie (celle d'Ester plus nos ajouts) qui est implantée dans CasEN. Le texte obtenu en sortie contient des *balises Unitex*, c'est-à-dire avec des accolades (voir ci-dessous la section 2.2). Il est donc possible de modifier ce balisage pour le remplacer par un balisage XML ou pour modifier la typologie, à condition que la nouvelle typologie puisse se déduire de la nôtre[5].

5. Sinon, il faudrait modifier la typologie sur les graphes...

+pers +hum +tit	les civilités
+pers +hum +gent	les gentilés et les adjectifs toponymiques
+pers +hum +occ	les professions
+pers +hum +sp	les sports
+pers +hum +art	les artistes
+pers +hum +nat	la nationalité
+pers +hum +rel	la religion
+pers +hum +pol	la politique
+pers +hum +fonc	les titres professionnels
+pers +hum +dynasty	les dynasties
+pers +hum +ethno	les ethnonymes
+event	les événements
+event +hist	l'histoire
+event +manif	les manifestations (sportives, artistiques...)

Tableau 5. *Les types ajoutés à ceux de la campagne Ester 2*

2.2. *Conception de la cascade*

La reconnaissance des entités nommées par la cascade CasEN utilise des ressources lexicales et des descriptions locales de motifs, des transducteurs qui agissent sur le texte par des insertions, remplacements ou suppressions. La plate-forme Unitex permet une écriture et une maintenance faciles de ces transducteurs en les présentant à l'utilisateur sous forme de graphes. Le principe d'une cascade est de pouvoir utiliser dans les descriptions suivantes les motifs déjà détectés ou, au contraire, d'éviter un étiquetage non souhaité pour un motif déjà reconnu. L'ordre de passage de ces transducteurs est donc un paramètre important.

Comme tous les programmes intégrés à la plate-forme Unitex, avant de lancer la cascade CasEN, il faut accepter la préanalyse du texte, puis appliquer les ressources lexicales. Dans le cadre d'un corpus écrit (par exemple, un journal, comme à la section 4.2), cette préanalyse consiste en un découpage en phrases, celui décrit dans (Friburger *et al.*, 2000) et distribué avec Unitex ; dans le cadre d'un corpus oral transcrit (par exemple les corpus Eslo 1 et Ester 2, voir sections 3.1 et 3.4), cette préanalyse suit le découpage en tours de parole, ou, plus exactement, le découpage créé par les balises du logiciel utilisé pour la transcription[6], comme le recommande Dister (2007). Pour les ressources lexicales, en plus du dictionnaire Delas (Courtois et Silberztein, 1990) distribué avec Unitex, la cascade utilise un graphe-dictionnaire des nombres écrits en toutes lettres, le dictionnaire Prolex-Unitex, extraits de Prolexbase (Tran et Maurel, 2006), qui contient des noms propres et des dérivés de noms

6. Nous avons écrit un graphe spécifique pour cette préanalyse : il reconnaît les balises XML, les considère comme une partie du discours spécifique, de type *XML*, et ajoute une balise de segmentation Unitex, {S}.

propres, et un dictionnaire spécifique avec des prénoms, des professions (Gazeau et Maurel, 2006), des noms d'animaux, de sports, de monnaies, etc.

Les graphes de CasEN sont répartis en cinq catégories :

– les graphes de reconnaissance qui étiquettent une catégorie d'entités, leurs noms commencent par cette catégorie : *timeHoraire*, *orgCommerceEtranger*, *persCoordination*, etc. Ces graphes, compilés, forment en général les transducteurs de la cascade et appellent des sous-graphes pour les contextes. Cependant, lorsque des graphes successifs sont tellement spécifiques qu'un passage en cascade est inutile, ils sont compilés en un seul transducteur ; par exemple, le transducteur *amount* appelle presque tous les graphes reconnaissant les différentes mesures (monnaie, température, longueur, etc.) ;

– les graphes outils, qui peuvent, soit être compilés en un transducteur de la cascade, soit constituer un sous-graphe. Leur nom commence par le mot-clé *tool*. Par exemple, lorsqu'on analyse un corpus du *Monde* d'il y a quelques années, on remarque que les sigles étaient écrits avec des points (*C.G.T.* au lieu de *CGT* aujourd'hui). De fait, le premier transducteur de la cascade, *toolChercheSigleAvecPoints* est un graphe outil localisant ce genre de sigles et le deuxième transducteur de la cascade, *toolSupprimePointDansSigle*, est lui aussi un graphe outil qui normalise les sigles en utilisant les informations du premier transducteur ;

– les graphes de listes, dont le nom commence par le mot-clé *list*, qui sont en fait des sous-graphes. Ces listes sont souvent des listes de mots polylexicaux, réduites par la factorisation des chemins. Le troisième transducteur de la cascade, *timeHoraire*, appelle par exemple le sous-graphe *listMinute* qui reconnaît un nombre entre 0 et 59 (en chiffres ou en lettres), suivi du mot *minute*, éventuellement abbrégé en *mn* ou en *min* ;

– les graphes de masques, dont le nom commence par le mot-clé *pattern*, qui sont eux aussi des sous-graphes. Ces graphes décrivent des listes un peu particulières, qui contiennent en général des expressions régulières utilisant des codes Unitex génériques ou des descriptions morphologiques. Le sous-graphe *patternCR* reconnaît par exemple les chiffres romains ;

– les graphes étiqueteurs, dont le nom commence par le mot-clé *tag*, qui sont, soit des listes, soit des masques, mais qui ajoutent des informations sur des éléments internes à une entité nommée. Par exemple, le sous-graphe *tagParti*, appelé par le transducteur *foncPolitique* ajoute une étiquette *org* sur la ville du maire, la région du député, etc. Ceci permet de modifier facilement l'étiquetage : il suffit d'intervenir sur les graphes de reconnaissance et sur les graphes étiqueteurs.

La cascade CasEN commence par les deux transducteurs outils décrits ci-dessus, puis viennent les transducteurs *time* et *amount*. Les transducteurs suivants sont soit des *org*, soit des *fonc*, soit des *pers*. Viennent ensuite des transducteurs *loc*, suivis du transducteur *prod*, puis des transducteurs *loc* et *org*, eux-mêmes suivis du transducteur *event*. La cascade se termine par différents transducteurs *org*, *fonc*, *pers* et *loc*. Cette hétérogénéité s'explique : des entités peuvent être ambiguës entre elles, soit

entièrement (l'entité est totalement ambiguë avec une autre), soit partiellement (l'entité contient des motifs qui peuvent être d'un autre type).

Donnons un exemple simple de balisage : la phrase *Vers le sud, une jetée longue de deux mille mètres s'allongeait comme un bras sur la rade de Suez,* extraite du corpus distribué par Unitex[7] va être transformée par le transducteur *amount*, qui appelle le graphe des longueurs (voir la figure 2), en *Vers le sud, une jetée longue de* {*deux mille mètres,.N+entity+amount+physics+length+grfamountLongueur*} *s'allongeait comme un bras sur la rade de Suez,* ce qui, d'une part, étiquette la séquence *deux mille mètres* comme une longueur et, d'autre part, la fige en une expression polylexicale. Ce balisage peut ensuite être recherché dans Unitex par des masques plus ou moins spécifiques, de <N+entity> à <N+length>. Pour faciliter le débogage, nous ajoutons au balisage le nom du graphe qui l'a inséré (ici *+grfamountLongueur*).

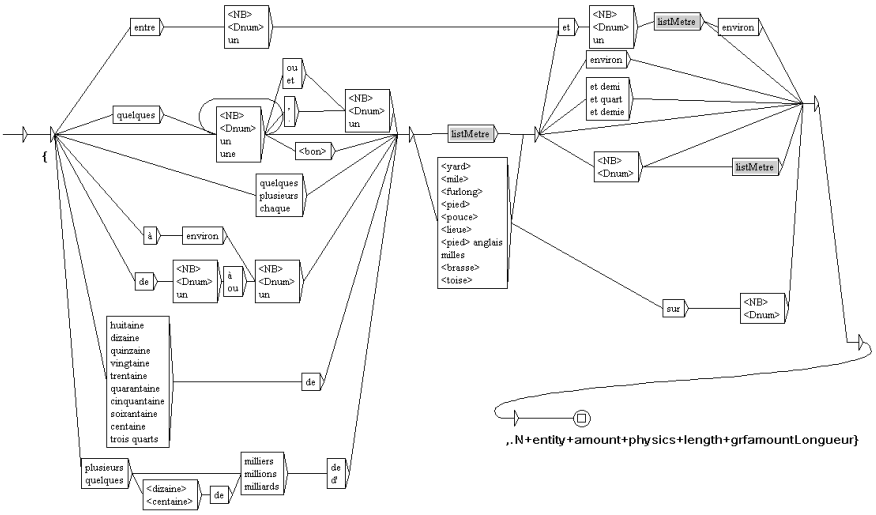

Figure 2. *Le graphe décrivant les longueurs*

La cascade elle-même est constituée à partir des îlots de certitude (Abney, 1996) qu'il est possible de trouver... Par exemple, la phrase *Il est arrivé le 29 février de l'année 2008.* peut être reconnue par plusieurs transducteurs. Par exemple :

– le transducteur *timeAnneesSiecle*, marque *l'année 2008* comme une date absolue ;

– le transducteur *timeDateAbsolue* marque la séquence entière *le 29 février de l'année 2008* comme une date absolue ;

7. *Le Tour du monde en 80 jours*, de Jules Verne.

– le transducteur *timeDateRelative* marque le début de la séquence *le 29 février* comme une date relative.

Il faut obligatoirement passer ces trois transducteurs dans l'ordre *timeDateAbsolue*, *timeAnneesSiecle* et *timeDateRelative*.

Parfois, il ne s'agit pas de concurrence, mais de complément. L'exemple le plus simple est sans doute le graphe des adresses postales qui contient des masques de personnes (<N+pers> pour reconnaître *rue du Général Leclerc*) et de dates (<N+date> pour *rue du 11 novembre 1918*) : les graphes des personnes et celui des dates sont donc placés avant le graphe des adresses. De nombreuses organisations comportent aussi des étiquettes de type *personne*, comme le *Centre Georges Pompidou*, l'*hôpital Henri Mondor* ou... la *menuiserie Gérard Dubois* ! Ces organisations seront donc reconnues après les graphes de personnes. Ainsi, nous constatons que l'ordre des graphes est capital, mais non trivial.

Pour conclure cette description de la cascade CasEN, le tableau 6 donne quelques exemples de reconnaissance sur un corpus du journal *Le Monde*.

« Au pire de la crise, {à l'automne dernier,.entity+time+date+rel+grftimeDateRelative}, nous avons détenu jusqu'à 20 % de liquidités dans notre portefeuille », indique {{{ Denis,.N+Prénom} { Remacle,.N+nom},.entity+pers+hum}, {gérant d'{Amplitude Pacifique,.entity+org+com} ,.entity+job} ,.entity+pers+hum+grfpersPrenomNom}, une sicav de {La Poste,.entity+org+com+grforgDico}.
« C'est à nos clients de décider s'ils souhaitent ou non consacrer une partie de leur patrimoine à l'{Asie,.entity+loc+admi+grflocPays} », souligne {{Pierre,.N+Prénom} {Ciret,.N+nom},.entity+pers+hum+grfpersPrenomNom}, de la {Compagnie financière {{Edmond,.N+Prénom} {de Rothschild,.N+nom} ,.entity+pers+hum} ,.entity+org+com+grforgCommerceGauche}.
Ils ne peuvent pas, en revanche, faire l'impasse sur la {Bourse de {Hongkong , .entity+loc+admi},.entity+org+com+grforgCommerceGauche}, car cette place représente près de la moitié de la capitalisation boursière de la région.{S} Pour sa part, {{ Pierre-Alexis,.N+Prénom} { Dumont,.N+nom},.entity+pers+hum+grfpersPrenomNom}, de {State Street Banque,.entity+org+com+grforgCommerceDroite}, s'est réfugié sur le marché australien, relativement épargné par la tourmente.
{Théâtre Gérard-Philipe,.entity+org+div+grforgDivertissementSorties}, {59, {boulevard Jules-Guesde,.entity+loc+line}, 93000 {Saint-Denis,.entity+loc+ville},.entity+loc+addr+post+grflocAddr}.

Tableau 6. *Exemples de séquences reconnues par CasEN sur un corpus du journal* Le Monde

3. Application de CasEN à deux corpus

3.1. *Présentation du corpus Eslo 1*

Lors de la conception du système CasSys (Friburger, 2002), l'application réalisée portait sur la détection d'entités nommées dans des textes journalistiques. L'ensemble de la cascade a été repris tout d'abord pour le projet ANR Variling (Eshkol *et al.*, 2010), puis dans le cadre de la campagne d'évaluation Ester (voir section 3.4), qui a eu lieu au cours de ce projet. Un des points forts de notre système est sa bonne adaptation à l'oral, à des corpus différents et à des typologies différentes, avec seulement quelques modifications.

Le corpus sur lequel nous avons travaillé est l'Enquête sociolinguistique à Orléans (Eslo 1), qui a été conduite en 1968 par des universitaires britanniques. Cette enquête avait une visée didactique : l'enseignement du français langue étrangère dans le système public d'éducation anglais. Elle comprend environ deux cents interviews en face à face et une gamme d'enregistrements variés (conversations téléphoniques, réunions publiques, transactions commerciales, repas de famille, entretiens médico-pédagogiques, etc.), soit au total 317 heures d'enregistrements, avec environ quatre millions et demi de mots et plusieurs centaines de locuteurs. En particulier, la tâche présentée ici permettra la mise à disposition d'un grand corpus où les entités nommées auront été annotées.

Plus précisément, le corpus dont nous disposions correspond aux cent vingt premières heures transcrites. Il était constitué de cent cinq fichiers Transcriber[8], représentant au total 31 004 Ko. Nous avons travaillé sur quatre-vingt-dix-huit fichiers, soit 29 522 Ko, et réservé sept fichiers pour l'évaluation, soit 1 482 Ko (4,8 %).

Les principales conventions de transcription sont l'absence de ponctuation et de majuscule en début d'énoncé ainsi qu'une transcription orthographique normée (majuscule pour les entités nommées, transcription des chiffres et des dates en toutes lettres avec les traits d'union si nécessaire, termes épelés notés entièrement en majuscules).

Le questionnaire de l'entretien contient tout d'abord des questions préliminaires (*Depuis combien de temps habitez-vous Orléans ?*, *Qu'est-ce qui vous a amené à vivre à Orléans ?*, *Est-ce que vous vous plaisez à Orléans ?*, etc.), puis des questions sur le travail et les loisirs du locuteur et des membres de sa famille, ce qui explique la présence d'un nombre important d'entités nommées dans le corpus. Enfin, sont abordés :

– l'enseignement (Qu'est-ce qu'on devrait apprendre surtout aux enfants à l'école ?, Dans quelles matières aimeriez-vous que vos enfants soient forts ?, etc.) ;

– la politique (Est-ce que, d'après vous, on fait assez pour les habitants d'Orléans ?, Que pensez-vous des événements de mai 68 ?, etc.) ;

– la langue et les habitudes culturelles (Un étranger veut venir en France pour apprendre le français. Dans quelle région est-ce qu'il doit aller d'après vous, dans quelle ville ?, Quelqu'un frappe à la porte de cette pièce. Qu'est-ce que vous lui dites ?, etc.).

3.2. *Évaluation sur le corpus Eslo 1*

Comme annoncé en section 3.1, l'évaluation a été réalisée sur sept fichiers Transcriber, soit 1 482 Ko (4,8 % du corpus).

8. http ://trans.sourceforge.net/

Nous avons réalisé trois évaluations sur le corpus en mesurant tout d'abord la simple détection des entités (trait *+Entity*), en acceptant d'éventuelles erreurs de typage ou de bornage, puis la reconnaissance des entités typées, en acceptant d'éventuelles erreurs de bornage et, enfin, celle des entités typées correctement bornées.

Le corpus de test comprenait 1 305 entités ; 1 227 ont été reconnues , 27 ont généré des erreurs et 51 ont été oubliées. Parmi les erreurs, citons le passage *environ cinq mille livres euh en rayon* qui a été annoté comme correspondant à la monnaie anglaise... Parmis les oublis, plusieurs sont dus à la présence de célébrités citées sans leurs prénoms (*Monet*, *Rabelais*, *Renoir*...)[9].

Sur l'ensemble de ces 1 227 entités reconnues, 1 154 ont été correctement typé. Restaient 73 erreurs de typage, comme *un chef de chorale* qui a reçu le trait *+pol*.

Finalement, 1 142 entités ont été correctement reconnues et bien balisées ; pour la plupart des erreurs (11 sur 12), le balisage était fermé avant la fin de l'entité.

Le tableau 7 présente ces évaluations en termes de rappel, précision et F-mesure (Maurel *et al.*, 2009).

	Entités reconnues	Entités bien typés	Entités bien balisées
Rappel	94,0 %	88,4 %	87,5 %
Précision	97,8 %	92,0 %	91,1 %
F-mesure	95,9 %	90,2 %	89,3 %

Tableau 7. *Évaluation de la cascade CasEN sur le corpus Eslo 1*

3.3. *Balisage du corpus*

Pour permettre la consultation du fichier annoté, nous avons défini un balisage de type XML pour rendre visible les étiquettes en dehors du logiciel Unitex. À partir de l'annotation {*Nicolas Sarkozy.N+Entity+pers+hum*}, nous avons ajouté au texte des balises *<ENT type="pers.hum"> Nicolas Sarkozy </ENT>*. Le tableau 8 présente quelques exemples de balisage. Finalement, les cent douze fichiers ont été relus (et corrigés manuellement) à partir du balisage effectué, c'est-à-dire que les erreurs de rappels ont été ignorées. Ce corpus annoté sera bientôt mis à disposition des chercheurs.

Contrairement à ce que nous pensions *a priori*, le corpus Eslo 1 a révélé une très faible présence des disfluences dans le cadre des entités nommées. Il nous a donc semblé que les erreurs dues aux disfluences ne nécessitaient pas de modifications im-

9. Ce point a été corrigé par des règles d'aliasisation ajoutées à Prolexbase et donc au dictionnaire de noms propres utilisé. Bien sûr, ces règles ne sont pas systématiques sur chaque entrée, mais réservées à des célébrités de grande renommée.

il y a deux ans une euh <ENT type="pers.hum.gent"> anglaise </ENT>
chez moi <ENT type="pers.hum"> Bérénice Nutal </ENT>
dans les <ENT type="org.com"> PTT </ENT>
moi je suis native de <ENT type="loc.admi"> Pithiviers </ENT> j'aime mieux <ENT type="loc.admi"> Orléans </ENT>
oh j'ai une <ENT typr="prod.art"> encyclopédie Quillé </ENT> j'ai le
<ENT type="time.date.abs"> en dix-neuf cent trente-huit </ENT>
je crois que le <ENT type="org.pol"> ministère de l'Education Nationale </ENT>
le <ENT type="org.edu"> lycée Pothier </ENT> et les élèves qui vont au <ENT type="org.edu"> lycée Benjamin Franklin </ENT>
euh passer quelques jours sur la <ENT type="loc.geo"> Côtes d'Azur </ENT>
je suis je travaille à l'<ENT type="loc.fac"> hôpital d'Orléans </ENT> quoi
parce que nous avions un <ENT type="loc.fac"> magasin Phildar </ENT> juste en face de chez nous

Tableau 8. *Exemples de balisage du corpus Eslo 1*

portantes de nos graphes, d'autant plus que les entités sont la plupart du temps localisées, même si la position exacte de l'entité n'est pas correcte. Or, nous souhaitions surtout éviter les erreurs de rappel, puisque, comme nous venons de le dire, le balisage de l'ensemble des entités nommées du corpus Eslo 1 a été révisé. Ces quelques erreurs ont donc été corrigées manuellement. Le tableau 9 présente quelques exemples de disfluences trouvées sur les entités nommées.

3.4. *La campagne Ester*

Le système CasEN a participé à la campagne Ester 2[10], sur la tâche destinée à l'évaluation des systèmes de reconnaissance des entités nommées sur des flux de paroles transcrites manuellement ou automatiquement. Une autre discussion sur cette campagne se trouve dans (Brun et Ehrmann, 2009). Le corpus de la campagne Ester 2 était constitué d'émissions radiophoniques transcrites (fichiers Transcriber). Les émissions enregistrées contenaient des émissions d'information, des dossiers liés à l'actualité du moment et des émissions plus conversationnelles. Les entités nommées détectées devaient être catégorisées selon sept catégories : personnes (pers), lieux (loc), organisations (org), productions humaines (prod), montants (amount), mesures de temps (time) et fonctions (fonc). Cette typologie était sous-divisée en 38 sous-catégories, qui n'ont pas été évaluées (voir section 2.1). Le tableau 10 présente les résultats officiels de la campagne Ester 2, publiés par (Galliano *et al.*, 2009). La mesure des performances était une variante pondérée du *slot error rate* (SER) (Makhoul *et al.*, 1999) ; la précision, le rappel et la F-mesure étaient aussi fournis.

10. http ://www.afcp-parole.org/ester/

Les pauses vides ou remplies :
le faubourg Saint </Turn> <Turn speaker="spk1 spk4" startTime="1671.987" endTime="1672.683">
<Sync time="1671.987"/> <Who nb="1"/> <ENT type="loc.admi"> Vincent </ENT>
<ENT type="time.hour"> trois heures </ENT> euh moins vingt
<ENT type="loc.admi"> La Chapelle </ENT> euh <ENT type="loc.admi"> Saint Mesmin </ENT>
Les répétitions, éventuellement accompagnées d'une pause :
de janvier à <Sync time="1337.908"/> <ENT type="time.date.rel"> à mai </ENT>
<ENT type="loc.admi"> France </ENT> euh <ENT type="org.div"> France Inter </ENT>
Les amorces, éventuellement accompagnées d'une pause :
<ENT type="time.date.rel"> au mois </ENT> de sep- <Sync time="3288.958"/> tembre
<ENT type="time.date.abs"> en mille neuf cent </ENT> cin- <Sync time="563.383"/>
<Sync time="564.101"/> Cinquante
Les autocorrections :
<ENT type="time.date.rel"> une période </ENT> de euh <Sync time="2859.139"/> <Sync time="2859.913"/>
combien <Sync time="2860.444"/> <Sync time="2860.882"/> <ENT type="time.date.rel">
six mois </ENT> ?
<ENT type="time.date.rel"> dans l'année </ENT> ils font <ENT type="amount.phy.wei"> un g
</ENT>- euh un un grand voyage de <ENT type="time.date.rel"> plusieurs jours </ENT>
<ENT type="time.hour">trois heures</ENT> de du matin
<ENT type="pers.hum"> Louis </ENT> trei- <ENT type="pers.hum"> Louis seize </ENT>
vers la <ENT type="loc.admi"> Seine </ENT> -et-Oi- euh la
<ENT type="loc.admi"> Seine-et-Marne </ENT>

Tableau 9. *Exemples de disfluences sur les entités nommées du corpus Eslo 1*

	Transcription			
	humaine			automatique
Participant (approche)	SER	Précision	Rappel	SER
Xerox (syntaxe profonde)	9,8	93,6	91,5	44,6
Synapse (syntaxe profonde)	9,9	93,0	89,3	44,9
LIA (CRF)	23,9	86,4	71,8	43,4
LIMSI (syntaxe surface)	30,9	81,1	70,9	45,3
LI Tours (syntaxe surface)	33,7	79,3	65,8	50,7
LSIS (CRF)	35,0	82,6	73,0	55,3
LINA (syntaxe surface)	37,1	80,7	55,4	54,0

Tableau 10. *Les résultats de la campagne Ester 2*

Sept systèmes (dont le nôtre, CasEN, noté *LI Tours* sur le tableau 10) ont participé à cette campagne, reposant sur des méthodes variées : apprentissage par CRF, systèmes à base de règles, avec analyses syntaxiques de surface ou profondes. Les systèmes centrés sur les connaissances qui intègrent une analyse syntaxique profonde (Xerox, Synapse) obtiennent les meilleurs résultats pour la transcription manuelle. Pour les

transcriptions automatiques, la meilleure approche, celle du LIA, est à base d'apprentissage ; on constate ici que les approches à base de règles voient leurs performances se dégrader progressivement en présence de bruits.

Les performances de CasEN sont proches de celles des autres systèmes non industriels (à l'exception du LIA), ce qui est rassurant pour un système qui, rappelons-le, a été développé initialement pour l'écrit. La faible dégradation des performances de CasEN sur la première des transcriptions automatiques est satisfaisante ; cela est notamment dû à la robustesse des analyses syntaxiques partielles. Nos performances sont plus modestes sur les deux autres car les transcriptions ne comportaient pas de majuscules, dont la présence est utilisée par CasEN. Ces derniers résultats ne sont donc pas réellement significatifs et nous ne les avons pas reportés sur le tableau 10. Nos résultats étaient meilleurs sur le corpus Eslo 1 (voir section 3.2) principalement à cause d'une difficulté supplémentaire dans la campagne d'évaluation : la catégorisation des noms propres ambigus, comme les toponymes qui peuvent désigner un lieu géographique (+loc+admi), une entité gouvernementale (+org+gsp) ou... une équipe sportive (+org+div) ! (voir la présentation de la typologie de CasEN section 2.2). Dans les évaluations présentées section 3.2, ce genre d'erreur n'a pas été pris en compte, ces toponymes étant systématiquement étiquetés +loc+admi, conformément au principe défini dans le tableau 2. Cependant, il faut souligner que le passage entre les deux corpus s'est réalisé sans travaux importants sur la cascade qui montre, de ce fait, une grande stabilité entre les corpus. Il nous a juste fallu supprimer les balises que nous avions ajoutées à la typologie. D'ailleurs les transformations de balises se réalisent facilement, contrairement aux systèmes à base d'apprentissage qui ont besoin d'un nouveau corpus balisé...

Afin d'analyser en détail le comportement du système, chaque erreur de CasEN (1 180 erreurs pour 2 512 entités nommées) a été annotée en précisant les informations suivantes : la localisation, le type d'erreur (suppression, insertion, catégorie erronée, erreur de frontière, etc.) et la règle de la convention Ester 2 concernée (Nouvel *et al.*, 2010a). Au passage, cette annotation a révélé que la référence comportait un nombre non négligeable d'erreurs. En utilisant sept types d'entités nommées, Ester 2 introduit en effet une catégorisation plus fine que celles mise en place lors des campagnes antérieures. Il en résulte des subtilités de catégorisation qui expliquent aussi les difficultés qu'ont rencontrées les annotateurs. Au final, CasEN s'est en tous cas vu compter 43 fausses erreurs d'insertion (entités nommées omises dans la référence, mais correctement détectées par le système). Dans d'autres cas, les annotations de référence se conforment au corpus d'apprentissage plutôt qu'aux règles spécifiées dans le guide d'annotations, ce qui a pu pénaliser les systèmes qui, comme CasEN, ont suivi le guide. Au final, nous observons une réduction de presque $10\ \%$ du SER après avoir nettoyé la référence.

Les instructions de cette référence étaient parfois déconcertantes. Donnons deux exemples extraits de la version 0.3.

– (Règle 1.3.3.1) Dans la phrase *Le Laboratoire de Recherche Informatique de l'université Paris Sud relève du département des Sciences Pour l'Ingénieur*, il fal-

lait reconnaître comme *org.edu* seulement *université Paris Sud* et *département des Sciences Pour l'Ingénieur*, bien que *Laboratoire de Recherche Informatique*, qui comporte des majuscules, nous semble être aussi un nom d'organisation, ce qui est confirmé par ailleurs par l'utilisation d'un sigle (*LRI*).

– (Règle 1.4.6.4) Il fallait *Ne pas étiqueter les lieux personnels, c'est-à-dire tout lieu ou habitation désignée comme appartenant à un particulier*, avec comme exemple *La propriété Saint-Vincent a été rachetée par le Comte de Bourgogne*. Mais, en Touraine, de nombreux châteaux sont à la fois des lieux publics qui se visitent, donc des *loc.fac*, et des habitations privées (certaines parties du château ne se visitent pas). Qui nous dit qu'il n'en est pas de même pour la *propriété Saint-Vincent* ?

La figure 3 présente les performances de CasEN par catégories. Le rappel varie significativement d'une catégorie à une autre. Globalement satisfaisante, la précision est médiocre pour la catégorie des productions humaines. Cette catégorie, très hétérogène, a aussi gêné les autres participants. Nos plus grosses difficultés concernaient les véhicules, qu'il fallait reconnaître même sous une forme incomplète, comme *depuis que j'ai acheté le Zafira je me demande comment j'ai fait pour conduire ma 106*, où il fallait reconnaître *Zafira* et *106*. La reconnaissance des productions artistiques, des prix et des productions documentaires comportait moins d'erreurs. Une des questions soulevées ici concerne les ressources dictionnairiques : faut-il les augmenter, par exemple ici avec des noms africains ou des noms de marque ? Mikheev *et al.* (1999) prônaient l'utilisation de peu de ressources, mais l'extension de la notion d'entités nommées et la possibilité d'utiliser de grandes bases documentaires. Charton et Torres-Moreno (2009) mettent la question à nouveau à l'ordre du jour.

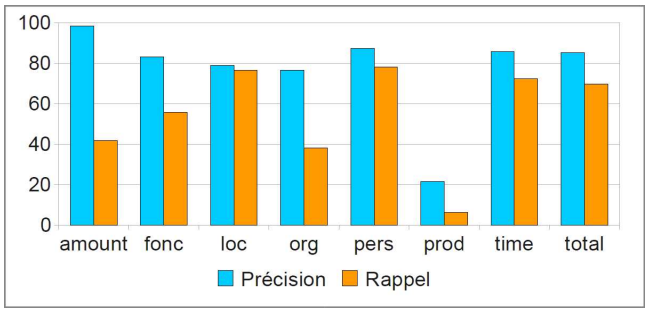

Figure 3. *Résultats de CasEN par catégories*

4. Utilisation de deux cascades successives

Nous présentons deux exemples d'une seconde étape d'annotations, réalisée à partir d'un corpus balisé par CasEN. Il s'agit tout d'abord de la suite de notre travail sur Eslo 1 (section 4.1), puis d'une étude réalisée sur un corpus du *Monde* (section 4.2).

4.1. *Les entités dénommantes du corpus Eslo 1*

4.1.1. *La protection des personnes*

Dans le cadre du projet Variling, le corpus Eslo 1 (déjà présenté section 3.1) sera mis à disposition des chercheurs et, peut-être aussi du public. Or, la mise à disposition d'un corpus oral soulève le problème de la protection des personnes (Baude, 2006). En effet, celles qui ont répondu à cette enquête, l'ont fait avec des garanties d'anonymat, ce qui suppose la disparition, par effacement ou bipage, de toutes les indications personnelles qui permettraient de les reconnaître. Le corpus Eslo est particulièrement touché par cette problématique, car les personnes enregistrées n'ont pas donné leur autorisation pour une exploitation de leurs paroles telle qu'elle est prévue maintenant (diffusion en ligne notamment).

Diffuser le Corpus d'Orléans selon les techniques actuelles, implique donc une démarche fondée sur de *bonnes pratiques* juridiques et éthiques. Ainsi, si pour des analyses scientifiques précises, le corpus brut reste le seul objet d'analyse possible, la diffusion par Internet requiert un corpus anonymisé. L'objectif du projet de construire un portrait sonore de la ville et de ses habitants implique à un haut degré des propos dont la diffusion demande une extrême prudence (informations personnelles, confidences, avis exprimés, etc.). Bien que l'on parle souvent d'anonymisation, la question légale concerne principalement l'assurance qu'il sera impossible d'identifier des personnes. Juridiquement l'anonymisation sert à qualifier l'opération par laquelle se trouve supprimé dans un ensemble de données, recueillies auprès d'un individu ou d'un groupe, tout élément qui permettrait l'identification de ces derniers. Bien sûr il ne s'agit pas de rendre totalement impossible l'identification d'un locuteur (il faudrait alors brouiller la voix sur l'ensemble de l'enregistrement, ce qui rendrait toute analyse linguistique impossible). L'objectif du projet était de repérer des éléments dans le discours du locuteur permettant son identification par un éventuel utilisateur du corpus.

Le processus d'anonymisation ne coïncide pas avec la reconnaissance des entités nommées classiques, car il s'agit aussi du repérage des éléments d'identification hors nom propre (profession, lieu de travail, lien de parenté...). Nous avons appelé ces donnés entités dénommantes puisqu'elles permettent d'identifier le locuteur (Eshkol, 2010). De plus, tous les noms propres ne sont pas à anonymiser : la Loire et Jeanne d'Arc ne sont pas à inclure dans l'effacement, ainsi que les noms de lieu se trouvant dans la réponse à la question *Où parle-t-on bien le français ?* ou encore le nom des animateurs célèbres de l'époque, dans les réponses sur les questions concernant les émissions télévisées ou radiophoniques. Enfin, soulignons que l'entité nommée repérée doit être étiquetée selon le rapport avec le locuteur. Un lieu va devenir un lieu d'habitat ou de travail, etc. Le balisage présenté ci-dessous servira à un annotateur humain qui prendra ou non la décision d'anonymiser.

4.1.2. *La méthodologie adoptée*

Nous avons choisi pour cette tâche de construire une deuxième cascade, qui passe sur le corpus Eslo 1 annoté par la cascade CasEN, pour y effectuer une recherche

d'informations (Hobbs *et al.*, 1997). Cette cascade a pour finalité le repérage des informations personnelles (famille, travail, engagement...). Certaines de ces informations serviront à l'anonymisation du corpus (section 4.1.1) et les autres permettront des études sociologiques sur la vie à Orléans à cette période. Bien sûr, certaines informations, qui sont présentées de manière trop dissemblable pour les reconnaître par un graphe, ont été annotées manuellement à partir de l'annotation des entités nommées.

Nous avons décidé de conserver une certaine homogénéité entre les deux cascades. L'enquête correspond essentiellement à des questions concernant la personne interrogée et sa famille : origine, âge, naissance, arrivée à Orléans, travail et même syndicat. Pour cela nous avons défini une typologie avec six types principaux :

1) le type *personne* permet de repérer les informations concernant la personne intérrogée et celles qu'il donne sur sa famille ;

2) le type *identité* marque des informations précises comme la date de naissance ou la date d'arrivée à Orléans, l'âge de la personne dont on parle, son origine, sa date de mariage, etc. ;

3) le type *travail* étiquette le métier, le secteur d'activité, le lieu de travail ou le nom de l'entreprise de la personne dont on parle ;

4) le type *engagement* concerne la vie associative (y compris syndicale ou parentale) et la vie militaire ;

5) le type *voyage* les différents déplacements car il ne faut pas oublier que ceux-ci étaient plus rares à l'époque du questionnaire qu'aujourd'hui ;

6) le type *études* indique les diplômes, les lieux ou les établissements.

Ces différents types et sous-types sont présentés dans le tableau 11.

Pour la reconnaissance des entités nommées, nous avons considéré que la présence des disfluences était négligeable (voir section 3.3). Il n'en est pas de même pour celle des entités dénommantes car leur reconnaissance se déroule sur plusieurs groupes syntaxiques, qui peuvent même relever de plusieurs locuteurs. Il était donc indispensable de prévoir la présence éventuelle de disfluences et de reprises syntaxiques. Deux sous-graphes spécifiques ont été écrits pour détecter respectivement les pauses simples et les insertions et amorces. Les répétitions ne sont pas traitées et les erreurs commises à ces endroits ont été corrigées manuellement. D'autre part, certains graphes utilisent la question comme amorce. Ils comportent donc une description du découpage XML de transcriber et de la balise {S} qu'utilise Unitex pour segmenter le document analysé (voir section 2.2). Par exemple le graphe de la figure 4 comporte quatre sous-graphes décrivant respectivement une question (sur la date d'arrivée à Orléans), les balises XML Transcriber et la segmentation, une disfluence éventuelle et la réponse à la question.

La figure 5 présente un graphe pour la reconnaissance de l'origine géographique de la personne interrogée.

Personne (+pers)	la personne interrogée (+speaker)
	son conjoint (+spouse)
	ses enfants (+child)
	les autres membres de la famille (+parent)
Identité (+identity)	le nom (+name)
	l'adresse (+addr)
	l'âge (+age)
	le mariage (+wedding)
	l'origine (+origin)
	la naissance (+birth)
	l'arrivée à Orléans (+arrival)
	le nombre d'enfants (+children)
Travail (+work)	métiers (+occupation)
	secteur d'activité (+field)
	lieu de travail (+location)
	entreprise (+business)
Engagement (+involvement)	association (+voluntary)
	militaire (+military)
	scolaire (+school)
	syndical (+tradeunion)
Voyage (+trip)	études (+study)
	vacances (+holiday)
	professionnel (+work)
Études (+study)	lieu (+location)
	diplôme (+degree)
	établissement (+edu)

Tableau 11. *Typologie des entités dénommantes*

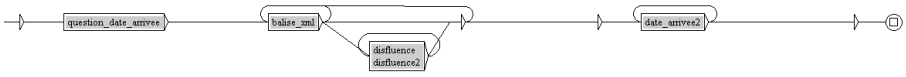

Figure 4. *Un graphe question-réponse sur la date d'arrivée à Orléans*

Comme pour les entités nommées, nous avons transformé le corpus Eslo 1 en un texte balisé. Quelques exemples sont présentés sur le tableau 12.

4.1.3. *Évaluation*

Pour évaluer la cascade des entités dénommantes, nous avons utilisé les mêmes enregistrements que pour celle des entités nommées (voir section 3.2). Les fichiers

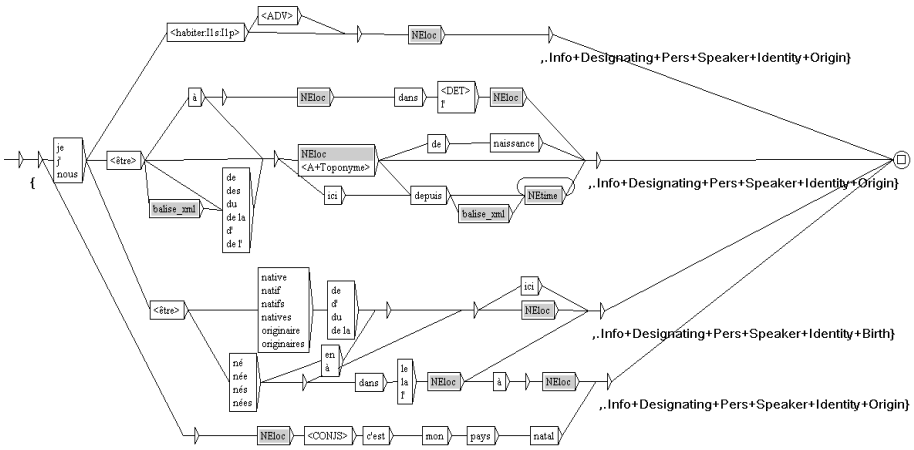

Figure 5. *Un graphe pour la reconnaissance de l'origine géographique*

de test comprenaient 77 entités dénommantes, ce qui est peu, mais nous voulions évaluer ce repérage sur le même corpus d'évaluation pour les deux cascades, or, en mettant de côté ce corpus pour le repérage des entités nommées, nous n'avons pas contrôlé la présence les entités dénommantes sur lesquelles nous n'avions pas encore commencé à travailler... Sur ces 77 entités dénommantes, nous en avons reconnu 69, nous avons fait 4 erreurs et oublié 12 entités. Notre précision est donc de 94,2 % et notre rappel de 84,4 %. Parmi les entités non reconnues, certaines étaient dues aux disfluences (*euh j'habitais dans dans le <ENT type="loc.admi"> Berry </ENT> à <ENT type="loc.admi"> Bourges </ENT>*), d'autres à des oublis dans la cascade des entités nommées (*je travaille actuellement à l'agence financière du <ENT type="loc.geo">bassin Loire-Bretagne</ENT>* où la présence d'*agence financière* aurait dû permettre le balisage d'une organisation et donc celui d'une entreprise où travaille le témoin). Notons une remarque intéressante à laquelle nous n'avions pas pensé de prime abord : certains métiers sont présentés par des prédicats (*euh j'enseignais le français*), ce qui nous a conduits à ajouter quelques transducteurs pour ce genre de présentation professionnelle.

Avec une précision qui doit donc toujours dépasser les 90 %, le gain de cette identification automatique est manifeste dans le processus de production de la ressource envisagée : les experts n'auront que peu de corrections à faire et le bon rappel observé nous garantit que le travail d'annotation manuelle sera au final très sensiblement amélioré.

alors <DE type="pers.speaker"> <DE type="identity.name"> je suis <ENT type="pers.hum">
monsieur Gabrion </ENT></DE></DE>
<DE type="pers.speaker"> je suis <DE type="identity.origin"> parisien </DE> </DE> de naissance
<DE type="pers.speaker"> je suis <DE type="work.occupation"> ingénieur chimiste </DE> </DE>
j'ai <DE type="identity.children" value="quatre"> quatre enfants </DE>
<DE type="pers.spouse"> mon mari est <DE type="work.occupation"> instituteur </DE> </DE>
<DE type="pers.child"> la quatrième souhaite être employée </DE type="work.business"> à la
<ENT type="org"> BNP </ENT></DE> <Sync time="515.129"/> <DE typer="work.location">
de <ENT type="loc.admi"> Paris </ENT> </DE></DE>
<DE type="pers.parent"> <DE type="identity.origin"> mon père était de la
<ENT type="loc.admi"> Sarthe </ENT>et ma mère du
<ENT type="loc.admi"> Berry </ENT></DE></DE>
que <DE type="pers.child"> ma fille qui a <DE type="identity.age">
<ENT type="amount.phy.age"> dix huit ans </ENT></DE></DE> ne parle pas très
ne parle pas bien français
<DE type="pers.speaker"> je suis né à <DE type="identity.birth">
<ENT type="loc.admi"> Orléans </ENT></DE></DE>
ça fait <DE type="pers.speaker"> <DE type="identity.arrival">
<ENT type="time.date.rel">trente-quatre ans</ENT>
que je suis à <ENT type="loc.admi"> Orléans </ENT></DE></DE>
je m'y suis habitué depuis longtemps <DE type="pers.speaker">on est marié <DE type="identity.wedding">
<ENT type="time.date.abs"> depuis mille neuf cent trente et un </ENT></DE></DE>

Tableau 12. *Exemples d'entités dénommantes*

4.2. *Les relations directes entre entités nommées du corpus* Le Monde

4.2.1. *La méthodologie adoptée*

L'idée de ce travail était de voir comment la simple détection des entités nommées pouvait permettre de les mettre en relation. Comme la plate-forme Unitex découpe les textes écrits en phrases, nous avons travaillé sur les phrases contenant au moins deux entités.

Le corpus utilisé est constitué de l'ensemble des journaux du quotidien *Le Monde* sur les années 1998 et 1999. Soit, pour 1998, 312 fichiers (344 Mo) et, pour 1999, 278 fichiers (259 Mo). Six journaux ont été utilisés pour construire une nouvelle cascade à partir de la précédente et trois ont été gardés pour une évaluation des résultats. La cascade CasEN de reconnaissance des entités nommées a été lancée sur ces 590 journaux, puis nous avons sélectionné et étudié les phrases contenant deux entités nommées, en les partageant en quatre groupes : personne-organisation, personne-lieu, personne-personne et personne-temps (tableau 13).

Nous avons donc étudié toutes les phrases de six journaux pour y trouver les relations existantes. Cette étude nous a conduits à reconnaître différents liens traduisant

	Pers_org	Pers_loc	Pers_pers	Pers_time
1998	2 034	7 895	14 473	2 671
1999	1 469	5 839	10 873	1 964

Tableau 13. *Nombre de phrases contenant deux entités nommées*

l'appartenance à un parti ou une autre organisation, le simple contact entre deux entités ou des liens plus identitaires sur la famille, la justice, le métier, la politique, la gouvernance, le sport, etc. Une cascade de transducteurs a ensuite été construite pour le balisage des relations entre ces entités. Ces graphes reconnaissent les entités nommées précédemment annotées par la cascade CasEN, ainsi que des contextes entre ces deux entités. Par exemple, le graphe de la figure 6 reconnaît une relation familiale entre deux personnes.

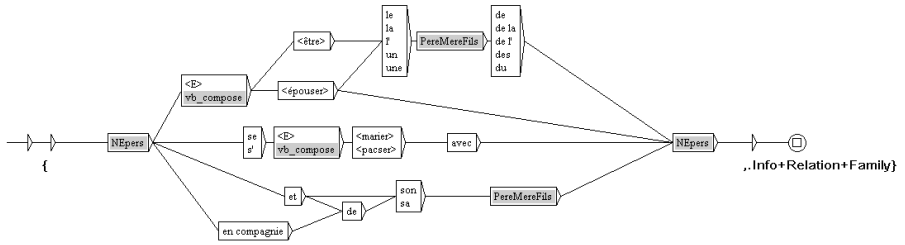

Figure 6. *Un graphe de relation familiale entre deux personnes*

Pour cette nouvelle cascade, nous avons réparti les relations en quatorze catégories, éventuellement sous-catégorisées par un sous-typage identique à celui de la cascade CasEN. Ces catégories sont présentées sur le tableau 14.

4.2.2. *Les résultats*

Bien sûr, ces catégories ne se répartissent pas de la même manière suivant le type des entités nommées en relation (tableau 15).

Comme cela a été dit section 4.2.1, trois journaux ont été utilisés pour l'évaluation, ceux datés des 27 juin 1998, 12 août 1998 et 22 février 1999. Nous avons tout d'abord passé la cascade CasEN sur ces trois journaux, puis la cascade CasRel. Pour la validité de cette dernière, il nous suffisait de localiser la présence des entités, sans tenir compte de la correction du balisage, ce qui nous donne évidemment des résultats surestimés[11], à comparer à la première colonne du tableau 7. Pour la reconnaissance des relations (cascade CasRel), les résultats sont corrects avec, toujours sur l'ensemble des trois

11. Sur l'ensemble des trois journaux, 99,96 % de précision et 98,09 % de rappel, soit une F-mesure de 99,01 %.

+appartenance	+appartenance+org	+appartenance+org+pol +appartenance+org+com +appartenance+org+edu +appartenance+org+non-profit +appartenance+org+div +appartenance+org+gsp
+contact	+contact+org	+contact+org+pol +contact+org+com +contact+org+edu +contact+org+non-profit +contact+org+div +contact+org+gsp
+existence		
+famille		
+identite		
+justice		
+litterature		
+media		
+metier	+metier+org	+metier+org+pol +metier+org+com +metier+org+edu +metier+org+non-profit +metier+org+div +metier+org+gsp
+origine		
+politique		
+presidence	+presidence+org	+presidence+org+pol +presidence+org+com +presidence+org+edu +presidence+org+non-profit +presidence+org+div +presidence+org+gsp
+sport		
+voyage		

Tableau 14. *Les types des balises de CasRel*

journaux, 93,46 % de précision et 90,51 % de rappel, soit une F-mesure de 91,96 %. Nous avons en effet commis quatre erreurs et oublié neuf relations, plus des balisages trop courts ou trop larges.

Le tableau 16 présente quelques exemples de balisages.

	PERS_ORG	PERS_PERS	PERS_TIME	PERS_LOC
+existence	*			*
+div	*			
+metier	*	*	*	
+litterature	*	*		
+sport	*	*	*	*
+president	*			
+politique	*	*	*	*
+appartenance	*			
+famille		*		
+identite		*		
+contact		*		*
+media		*		*
+politique		*		
+candidature		*		
+justice		*	*	
+voyage				*

Tableau 15. *Répartition des balises suivant le type des entités nommées en relation*

<REL type="appartenance.org.non-profit"> <ENT type="pers.hum"> président du conseil régional
Nord-Normandie </ENT> de l'<ENT type="org.non-profit"> Eglise réformée </ENT> </REL>
<REL type="appartenance.org.pol"> <ENT type="pers.hum"> Jean-Michel Alexandre,
directeur de l'évaluation </ENT> à l'<ENT type="org.pol"> Agence du médicament </ENT> </REL>
<REL type="sport"> <ENT type="pers.hum"> Le Brésilien Gustavo Kuerten </ENT>
a gagné le <ENT type="org.div"> tournoi de Rome </ENT> </REL>
<REL type="presidence.org.com"> <ENT type="pers.hum"> Luc Soete </ENT> dirige le
<ENT type="org.com"> Maastricht Economic Research </ENT> </REL>
<REL type="metier.org.com"> <ENT type="pers.hum"> Marie Owens Thomsen </ENT>, économiste chez
<ENT type="org.com"> Merrill Lynch </ENT> </REL>
<REL type="justice"> <ENT type="pers.hum"> Mathieu Filidori </ENT> est condamné à
<ENT type="time.date.rel"> treize ans </ENT> </REL>
<REL type="litterature"> <ENT type="pers.hum"> Nigel Barley </ENT> est l'éditeur de
<ENT type="pers.hum"> Mani </ENT> </REL>

Tableau 16. *Quelques exemples de balisages par CasRel*

5. Conclusion

Dans cet article nous avons présenté plusieurs applications développées à l'aide de cascades de transducteurs à nombre fini d'états autour de tâches liées à la détection des entités nommées. Nous avons cherché à montrer que les techniques à base de connais-

sances restaient adaptées à ce type de tâche, comme l'ont montré, par exemple, nos performances sur la campagne d'évaluation Ester 2. L'indépendance de notre base de connaissance par rapport aux données d'apprentissage permet par ailleurs une adaptation rapide de nos système pour passer d'un domaine d'application (flux audio et vidéo pour Ester et l'ANR EPAC) à un autre (enquête sociolinguistique pour l'ANR VariLing). Il n'en reste pas moins que cette adaptation n'est jamais optimale et que, comme pour tout système travaillant sur le sujet, la détection d'entités nommées requiert le développement de ressources importantes. Nous travaillons actuellement à l'utilisation de techniques de fouille de textes à base de détection de motifs hiérarchiques pour l'évolution semi-automatique de nos bases de connaissances (Nouvel *et al.*, 2010b).

Remerciements

Ce travail a été réalisé grâce au soutien de l'ANR (projet Variling) et du Feder Région Centre (projet Entités).

6. Bibliographie

Abney S., « Partial Parsing via Finite-State Cascades », *Proc. of Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, p. 8-15, 1996.

Ait-mokhtar S., Chanod J.-P., « Incremental Finite-State Parsing », *Applied Natural Language Processing*, p. 72-79, 1997.

Baude O., *Corpus oraux : guide des bonnes pratiques*, CNRS-Éditions et Presses universitaires d'Orléans, 2006.

Bauer G., *Namenkunde des Deutschen*, Bern, Germanistische Lehrbuchsammlung Band 21, 1985.

Béchet F., Sagot B., Stern R., « Coopération de méthodes statistiques et symboliques pour l'adaptation non supervisée d'un système d'étiquetage en entités nommées », TALN 2011, 2011.

Bontcheva K., Dimitrov M., Maynard D., Tablan V., Cunningham H., « Shallow Methods for Named Entity Coreference Resolution », TALN 2002, 2002.

Brun C., Ehrmann M., « Un système de détection d'entités nommées adapté pour la campagne d'évaluation ESTER 2 », TALN 2009, 2009.

Charton E., Torres-Moreno J. M., « Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées », TALN 2009, 2009.

Chinchor N., *Muc-7 Named Entity Task Definition*, 1997.

Coates-Stephens S., *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*, Kluwer Academic Publishers, Hingham, MA, 1993.

Courtois B., Silberztein M., « Dictionnaires électroniques du français », *Langues française*, vol. 87, p. 11-22, 1990.

Dister A., *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL*, thèse de doctorat, Université catholique de Louvain, 2007.

Ehrmann M., *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), 2008.

Eshkol I., *Entrer dans l'anonymat. Étude des entités dénommantes dans un corpus oral*, Narr Francke Attempto Verlag GmbH, Germany, p. 245-266, 2010.

Eshkol I., Maurel D., Friburger N., « Eslo : from transcription to speakers' personal information annotation », *Seventh language resources and evaluation conference (LREC 2010), Valetta, Malte*, 2010.

Friburger N., *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*, thèse de doctorat, Université François-Rabelais de Tours, 2002.

Friburger N., Dister A., Maurel D., « Améliorer le découpage des phrases sous Intex », *Revue Informatique et Statistique dans les Sciences Humaines*, vol. 36, n° 1-4, p. 181-200, 2000.

Friburger N., Maurel D., « Finite-state transducer cascade to extract named entities in texts », *Theoretical Computer Science*, vol. 313, p. 94-104, 2004.

Galliano S., Gravier G., Chaubard L., « The ester 2 evaluation campaign for the rich transcription of french radio broadcasts », *Proceedings of Interspeech'09*, p. 2583-2586, 2009.

Gazeau M.-A., Maurel D., « Un dictionnaire INTEX de noms de professions : quels féminins possibles ? », *Formaliser les langues avec l'ordinateur. De INTEX à Nooj*, Presses universitaires de Franche-Comté, p. 115-127, 2006.

Grass T., « Typologie et traductibilité des noms propres de l'allemand vers le français », *TAL*, vol. 41, n° 3, p. 643-669, 2000.

Hobbs J., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M., « FASTUS : A cascaded finite-state transducer for extracting information from natural-language text », *Finite-State Language Processing*, MIT Press, p. 383-406, 1997.

Makhoul J., Kubala J., Schwartz R., Weischedel R., « Performance measures for information extraction », *Proceedings of DARPA Broadcast News Workshop*, 1999.

Maurel D., Friburger N., Eshkol I., « Who are you, you who speak ? Transducer cascades for information retrieval », *in Proceedings of 4th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland*, p. 220-223, 2009.

Mikheev A., Moens M., Grover C., « Named entity Recognition without Gazetteers », EACL'99, p. 1-8, 1999.

Nadeau N., Sekine S., *A survey of named entity recognition and classification*, Satoshi Sekine and Elisabete Ranchhod, ed., John Benjamins publishing company, p. 3-28, 2009.

Nouvel N., Antoine J.-Y., Friburger N., Maurel D., « An analysis of the performances of the CasEN named entities detection system in the Ester2 evaluation campaign », LREC 2010, 2010a.

Nouvel N., Soulet A., Antoine J.-Y., Friburger N., Maurel D., « Reconnaissance d'entités nommées : enrichissement d'un système à base de connaissances à partir de techniques de fouille de textes », TALN 2010, 2010b.

Paik W., Liddy E. D., Yu E., McKenna M., *Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval*, Massachussetts Institute of Technology, p. 61-73, 1996.

Paumier S., *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, thèse de Doctorat en Informatique, Université de Marne-la-Vallée, 2003.

Petit G., « Le nom de marque déposée : nom propre, nom commun et terme », *META*, vol. 51, n° 4, p. 690-705, 2006.

Poibeau T., *Extraction automatique d'information, du texte brut au web sémantique*, Lavoisier, 2003.

Tran M., *Prolexbase, un dictionnaire relationnel multilingue de noms propres : conception, implémentation et gestion en ligne*, thèse de doctorat en informatique de l'Université François-Rabelais de Tours, 2006.

Tran M., Maurel D., « Prolexbase : un dictionnaire relationnel multilingue de noms propres », *TAL*, vol. 47, n° 3, p. 115-139, 2006.