

Fouille du Web pour la collecte d'Entités Nommées

Christian Jacquemin¹ et Caroline Bush^{1,2}

¹CNRS-LIMSI, BP 133, F-91403 ORSAY Cedex, FRANCE

²UMIST, Dept of Language Engineering, PO Box 88, Manchester M60 1QD, UK

{jacquemin, caroline}@limsi.fr

Résumé

Cette étude porte sur l'acquisition des Entités Nommées (EN) à partir du Web. L'application présentée se compose d'un moissonneur de pages et de trois analyseurs surfaciques dédiés à des structures spécifiques. Deux évaluations sont proposées : une évaluation de la productivité des moteurs en fonction des types d'EN et une mesure de la précision.

1. Présentation

L'acquisition lexicale à partir de corpus électroniques est désormais considérée comme une technique classique d'enrichissement des dictionnaires (Boguraev & Pustejovsky, 1996). Nous nous intéressons ici à un type d'unités lexicales particulier, les Entités Nommées (EN), une appellation générique pour les noms propres désignant des personnes, des lieux ou des organismes. Alors que les corpus électroniques classiques¹ conviennent bien pour de l'acquisition de noms communs stables, il est nécessaire de faire appel au Web pour retrouver les EN qui sont évolutives — par exemple, de nouvelles entreprises sont créées chaque jour — et temporairement visibles sous la pression des événements et de leurs reflets dans les médias.

Ce travail se situe parmi les applications d'acquisition de connaissances à partir du Web (Crimmins *et al.*, 1999). Comme Hearst (1998) et Morin & Jacquemin (1999) utilisent des contextes énumératifs pour le repérage de liens hyponyme/hyperonyme, il s'appuie sur des amorces de listes d'EN pour détecter les EN. Notre approche se démarque des approches traditionnellement utilisées en linguistique de corpus puisqu'elle exploite des informations sur la structure des documents pour acquérir des données. Elle se rapproche des études sur les *wrappers*, les outils d'analyse des données semi-structurées (Kushmerick *et al.*, 1997).

La technique d'acquisition des EN décrite ici combine un moissonneur (associé à un moteur de recherche sur le Web) et des analyseurs surfaciques de pages HTML dont les grammaires utilisent des informations lexico-syntaxiques et des instructions de formatage au sein d'un même formalisme.

¹Par exemple, le BNC (<http://info.ox.ac.uk/bnc/>) pour l'anglais ou le corpus CLEF (<http://www.biomath.jussieu.fr/CLEF/>) pour le français.

2. Utilisation des contextes définitoires

L'acquisition des EN à partir du Web pose deux problèmes nouveaux en acquisition lexicale :

1. Alors que les corpus textuels classiques peuvent être analysés en totalité, les pages du Web ne sont vues que par le trou de serrure des moteurs de recherche. Pour repérer les EN, il nous faut donc nous concentrer sur les indices linguistiques qui permettent de repérer les pages en ligne contenant des EN typées.
2. Alors que les pages Web ont pleines d'EN, seule une partie d'entre elles correspond à des EN publiques, fraîches et largement connues (le nom d'un perroquet sur une page personnelle ne doit pas être inclus dans un dictionnaire d'EN). En outre, les EN acquises automatiquement ne peuvent être incluses dans des bases lexicales électroniques que si elles sont associées à des types tels que PERSONNE, LIEU ou ORGANISME, voire à des types plus fins.

Le besoin d'indices linguistiques sélectifs et acceptables par les moteurs de recherche actuels nous a conduit à nous focaliser sur les **collections** — un cas particulier de contexte définitoire. Les amorces de ces collections ont une structure remarquable (Péry-Woodley, 1998) : elles contiennent des déclencheurs linguistiques tels que *following* (suivant) ou *such as* (tel que) qui, combinés aux types des EN, forment des requêtes suffisamment précises pour être utilisables par un moteur de recherche. En outre, ces amorces suivent le schéma *genus/differentia* pour définir les EN et fournissent ainsi, par le *genus*, un hyperonyme ou un type de l'EN qu'elles annoncent. Notre étude prolonge le travail de (Hearst, 1998; Morin & Jacquemin, 1999) aux corpus du Web ayant une mise en forme matérielle riche.

3. Architecture et principes

Pour acquérir les EN à partir du Web, nous avons développé un système qui se compose de trois modules séquentiels (voir figure 1) :

1. Un moissonneur chargé de rapatrier les pages ramenées par un moteur de recherche sur les quatre familles de requêtes suivantes
(1.a) *following* ⟨EN⟩ (1.b) *list of* ⟨EN⟩ (1.c) ⟨EN⟩ *such as* (1.d) *such* ⟨EN⟩ *as*
dans lesquelles ⟨EN⟩ représente un hyperonyme typant une famille d'EN tel que *Universities* (Universités), *politicians* (politiciens), ou *car makers* (fabricant de voiture). La liste complète des chaînes représentées par ⟨EN⟩ est donnée section 4.
2. Trois analyseurs surfaciques Ae, Al et Aa qui extraient les entités nommées candidates respectivement à partir des énumérations, des listes et tables et des ancres.
3. Un module de post-filtrage qui épure les EN produites par les analyseurs et éclate les EN coordonnées en EN unitaires.

Moissonnage de corpus Web

Les quatre chaînes données en (1.a-d), composées d'un nom d'hyperonyme et d'un marqueur discursif annonçant la collection, sont utilisées pour construire des requêtes pour les moteurs de recherche. La figure 2 montre cinq exemples prototypiques de collections rencontrées dans les pages HTML rapportées par les requêtes précédentes.²

²Dans les exemples de la figure 2, ⟨EN⟩ est la chaîne *international organizations*; les mises en relief typographiques ont été ajoutées par nous.

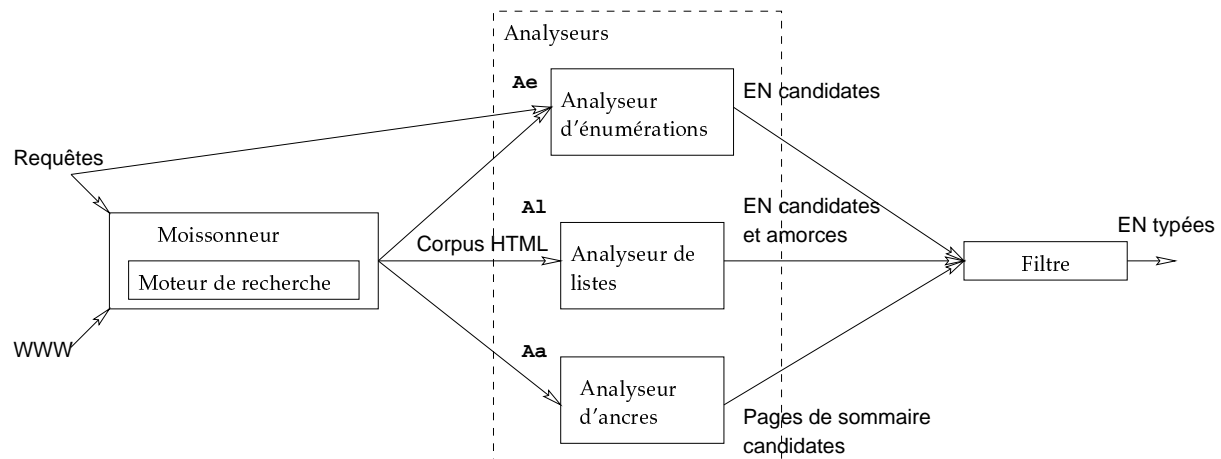


FIG. 1: Architecture de l'acquisition des EN

La première collection est une **énumération** ; elle est formée d'une coordination de trois EN. La deuxième collection est une **liste** organisée en deux sous-listes enchâssées. Chacune des sous-listes est introduite par un hyperonyme. La troisième structure est également une **liste** dont les articles débutent par une marque graphique. De telles listes s'obtiennent en HTML au moyen de tables (notre exemple) ou au moyen des marques d'énumération (`` ou ``). La quatrième liste est également construite au moyen d'une table, mais elle est plus riche que la précédente car chaque article est accompagné d'une image et précédé d'un titre. Dans le cinquième exemple, la liste n'est pas présente dans le document, elle est accessible au moyen d'un hyperlien dont l'ancre contient l'amorce.

Le corpus de pages HTML est récolté au moyen de deux moteurs de recherche avec des capacités différentes : AltaVista (AV) et Northern Light (NL).³ AV propose un mode de recherche avancé de chaînes de caractères alors que NL ne fait la recherche que sur un « sac de mots ». En contrepartie, le nombre de documents accessibles par NL est potentiellement illimité⁴ alors qu'AV ne donne accès qu'aux 200 premiers documents. Ces deux types de moteurs de recherche ont été choisis afin d'évaluer dans quelle mesure un mode de recherche plus pauvre peut être compensé par un accès à un plus grand nombre de documents.

Acquisition d'EN candidates

Trois analyseurs surfaciques parallèles Ae, A1 et Aa servent à extraire les EN des documents récoltés par le moissonneur. Les grammaires de ces analyseurs sont des expressions régulières construites sur les chaînes des requêtes (1.a-d) et sur les marques HTML qui signalent la mise en forme matérielle d'une collection. Les trois types suivants de structures spatio-syntaxiques sont recherchés : les **énumérations** (analyseur Ae, premier exemple de la figure 2), les **listes** (analyseur A1, les trois exemples suivants de la figure 2), et les **ancres** vers une page contenant une collection d'EN (analyseur Aa, dernier exemple de la figure 2).

Ces trois analyseurs combinent des recherches de chaînes (la requête), une analyse syntaxique (les énumérations dans Ae), une analyse des instructions de mise en forme (les listes

³Le moissonneur qui récupère les pages HTML est une combinaison de wget récupérable à l'URL <ftp://sunsite.auc.dk/pub/infosystems/wget/> et de scripts en Perl.

⁴Pour des raisons de temps de calcul et de tailles de corpus, nous nous sommes limités aux 2000 premiers documents sur NL.



<p>It's development is due to the support given by the Ministry of Public Health, aided by international organizations such as the Pan American Health Organization (PAHO), the United Nations Development program, and the Caribbean and Latin American Medical Science Information Center.</p>
<p>7. The session was also attended by observers from the following international organizations:</p> <p>(a) <i>United Nations organs</i></p> <p>International Bank for Reconstruction and Development (World Bank)</p> <p>(b) <i>Intergovernmental organizations</i></p> <p>Asian-African Legal Consultative Committee (AALCC) Inter-American Development Bank International Institute for the Unification of Private Law (UNIDROIT)</p>
<p>International Organizations</p> <p>The following international organizations are collaborating on the Project:</p> <ul style="list-style-type: none"> ▶ International Commission on Non-Ionizing Radiation Protection (ICNIRP) ▶ International Agency for Research on Cancer (IARC) ▶ United Nations Environment Programme (UNEP)
<p>Below is the list of international organizations that we distribute:</p> <hr/> <div style="display: flex; align-items: flex-start;"> <div style="margin-right: 20px;">  </div> <div> <p>EU (European Union) Books, documentation, periodicals on European legislation, economy, agriculture, industry, education, norms, social politics, law. For more information on publications, COM documents and to subscribe to the Official Journal please contact Dünya Infotel.</p> </div> </div> <hr/> <div style="display: flex; align-items: flex-start;"> <div style="margin-right: 20px;">  </div> <div> <p>UN (United Nations) Peace and security, economics, statistics, energy, natural resources, environment, international law, human rights, political affairs and disarmament, social questions. 1997 periodicals include: Development Business, East-West Investment News, Transnational Corporations, Monthly Bulletin of Statistics, etc.</p> </div> </div> <hr/>
<p>An agency may detail or transfer an employee to any organization which the Office of Personnel Management has designated as an international organization (see list of international organizations).</p>

FIG. 2: Cinq différents types de formatages de collections d'EN.

et les tables dans A1) ou un accès aux documents (dans Aa). Les résultats présentés dans cet article ne concernent que les deux premiers analyseurs, le travail sur les ancres est en cours de développement et pose des problèmes spécifiques (Amitay, 1999). Une analyse manuelle préliminaire de 100 ancras montre que 37% d'entre elles réfèrent à une page contenant une collection d'EN du type attendu et 9% d'entre elles pointent vers une page répertoire contenant un ou plusieurs liens vers des pages avec de telles collections. Les cas d'échecs sont constitués par 40% de liens vers des URL inaccessibles et 14% de liens vers des pages qui ne contiennent pas de collection d'EN et ne pointent pas vers une page avec une collection d'EN.

L'analyseur d'énumérations Ae

Une énumération est attendue comme une structure coordonnée suivant la chaîne de requête à l'intérieur de la même phrase. Ae recherche donc les EN de patrons décrits par (3) à l'intérieur d'énumérations décrites par (4).⁵

$$\text{EN_cand} = ([A-Z \ \&][a-zA-Z \ -\']^*)^+ \quad (3)$$

$$\text{Énumération} = (\text{EN_cand},)^* \text{EN_cand} (, ?) (\text{and}|\text{or}) \text{EN_cand} \quad (4)$$

L'analyseur de listes Al

Les listes sont attendues dans les quatre lignes suivant la phrase qui contient la chaîne de recherche. Elles sont extraites au moyen d'un des trois patrons suivants qui correspondent aux trois modes principaux de descriptions de listes alignées verticalement en HTML. La chaîne retenue est la **plus courte** séquence soulignée acceptable (la casse n'est pas prise en compte) :

$$\langle li \rangle \underline{\quad}^* (\langle /li \rangle | \langle li \rangle | \langle /ol \rangle | \langle /ul \rangle) \quad (5)$$

$$\langle br \rangle \underline{\quad}^* \langle /br \rangle \quad (6)$$

$$(\langle td \rangle | \langle th \rangle) \underline{\quad}^* (\langle td \rangle | \langle th \rangle | \langle /td \rangle | \langle /th \rangle | \langle /table \rangle) \quad (7)$$

Après suppression des étiquettes HTML, seules les **plus longues** sous-parties des chaînes acceptées par (3) sont fournies au module de post-filtrage final. Ces patrons ne couvrent pas les listes mises en forme au moyen de textes préformatés (marques `<pre>`) ou au moyens de retours à la ligne par des paragraphes (marques `<p>`). Ces modes de construction sont ignorés parce que, n'étant pas assez typiques des listes, ils produiraient des résultats trop imprécis.

Le module de post-filtrage

Les EN précandidates produites par les analyseurs sont filtrées avant d'être proposées comme EN candidates. Les filtres employés sont peu sélectifs, ce qui explique la bonne productivité des patrons de recherche et leur faible précision. Les rôles des filtres sont, dans cet ordre, de

- supprimer les mots en minuscules terminaux, les déterminants initiaux et les conjonctions de coordinations non initiales et les mots qui les suivent,
- rejeter les précandidats contenant les caractères @, {, #, ~, \$, ! ou ?,
- supprimer les marques d'articles telles que 1., –, * ou a), les marques HTML, les conjonctions de coordinations initiales et les appositions suivant un tiret ou une virgule,
- transformer les mots en majuscules en mots dont seule la première lettre est capitalisée dans les EN candidates qui ne sont pas des organismes (dans l'hypothèse où seuls les noms d'organismes contiennent des sigles en majuscules).

Le post-filtrage s'achève par une suppression des candidats monolexicaux qui sont des mots simples non capitalisés de la base CELEX⁶ et des candidats polylexicaux de plus de 5 mots.

⁵Les patrons effectifs sont plus complexes pour accepter les signes diacritiques et les abréviations.

⁶La base lexicale CELEX pour l'anglais est distribuée par le Consortium for Lexical Resources à l'URL www.ldc.upenn.edu/readme_files/celex.readme.html.

TAB. 1: Tailles des corpus de documents HTML (en Mb) collectés sur les 4 patrons (1.a-d).

	<i>following</i> ⟨EN⟩	<i>list of</i> ⟨EN⟩	⟨EN⟩ <i>such as</i>	<i>such</i> ⟨EN⟩ <i>as</i>
AV	85,9	64,9	150,4	66,3
NL	172,8	1 306,9	652,7	458,1

4. Expériences et évaluations

L'acquisition est faite sur 34 types d'EN choisies arbitrairement parmi les trois sous-types de la typologie des évaluations *Message Understanding Conference* (MUC-6, 1995) pour lesquelles une tâche d'identification des EN a été définie : ORGANISME (*American companies, international organizations, universities, political organizations, international agencies, car makers, terrorist groups, financial institutions, museums, international companies, holdings, sects, et realtors*), PERSONNE (*politicians, VIPs, actors, managers, celebrities, actresses, athletes, authors, film directors, top models, musicians, singers, et journalists*) et LIEU (*countries, regions, states, lakes, cities, rivers, mountains, et islands*).

Chacun de ces 34 types est combiné avec les quatre types de marqueurs discursifs (1.a-d), produisant 136 requêtes pour les deux moteurs de recherche. Chacun des 272 corpus HTML fournis par le moissonneur est constitué d'au plus 200 documents sur AV en mode de recherche par chaîne et d'au plus 2000 documents par NL en mode de recherche standard. Ces documents sont traités par les analyseurs d'énumérations et de listes. Le corpus total collecté fait 2 958Mb (368Mb sont extraits par AV et 2 590 par NL) ; il se sous-divise comme indiqué table 1. Les documents récupérés par NL pour le patron *list of* ⟨EN⟩ représentent plus de la moitié des documents extraits par NL (1 307 Mb). Le patron fournissant le plus de documents pour AV est ⟨EN⟩ *such as* par lequel 41% des documents récupérés par AV sont obtenus (150 Mb).

Mesures quantitatives

44 624 EN candidates sont produites : 17 116 à partir des corpus AV et 34 978 des corpus NL. L'utilisation de corpus plus importants avec NL compense bien sa précision plus faible qu'AV puisque la production à partir des corpus NL est plus importante que celle à partir des corpus AV. Toutefois, le coût (calculatoire et en accès réseau) est plus élevé pour les candidats issus de NL que d'AV puisqu'il faut collecter et analyser plus de données textuelles pour produire le même volume de termes.

En plus du nombre de candidats produits pour chacun des moteurs de recherche et pour chacune des quatre familles de chaînes, la table 2 fournit quatre mesures pour lesquelles les valeurs les plus élevées sont en gras et les valeurs les plus faibles en italiques. La *productivité* est le nombre moyen de candidats produits pour un corpus de 100kb. La productivité est 3,5 fois plus élevée pour AV (46,7) que pour NL (13,5), ce qui signifie que l'utilisation d'un moteur de recherche permettant des requêtes plus précises nécessite 3,5 fois moins de données pour parvenir au même nombre de candidats.

Le *taux d'énumérations par rapport aux listes* mesure le rapport du nombre d'énumérations d'EN rencontrées dans les documents analysés par A_e par rapport au nombre de listes d'EN dans les documents analysés par A_l. Comme l'on pouvait s'y attendre, il y a 11 fois plus d'énumérations que de listes après ⟨EN⟩ *such as* et 18 fois plus après *such* ⟨EN⟩ *as*. La sortie est plus équilibrée avec *list of* ⟨EN⟩ : il y a seulement 1,66 fois plus de listes que d'énumérations. Le plus

TAB. 2: Volumes d'EN candidates acquises à partir des corpus décrits table 1.

Moteur AV	<i>following</i> ⟨EN⟩	<i>list of</i> ⟨EN⟩	⟨EN⟩ <i>such as</i>	<i>such</i> ⟨EN⟩ <i>as</i>
Nb candidates	4 747	3 112	5 738	3 579
Productivité	55,2	48,0	38,2	53,9
Taux énum./listes	0,28	0,83	12,5	43,74
Redondance	2,12	2,15	1,77	1,69

Moteur NL	<i>following</i> ⟨EN⟩	<i>list of</i> ⟨EN⟩	⟨EN⟩ <i>such as</i>	<i>such</i> ⟨EN⟩ <i>as</i>
Nb candidates	5 667	5 176	14 800	9 335
Productivité	32,8	4,0	22,7	20,4
Taux énum./listes	0,31	0,49	10,41	14,72
Redondance	2,12	2,34	2,13	2,20

AV & NL	<i>following</i> ⟨EN⟩	<i>list of</i> ⟨EN⟩	⟨EN⟩ <i>such as</i>	<i>such</i> ⟨EN⟩ <i>as</i>	Total
Nb candidates	8 673	7 380	18 005	10 566	44 624
Recouvrement	16,7%	11,0%	12,3%	18,2%	15,0%

grand nombre d'énumérations suivant cette structure est certainement dû à l'effet de la combinaison des informations linguistiques et des marques de mise en forme dans la construction du sens. Afin de satisfaire la maxime de quantité de Grice (1975), le rédacteur évite d'utiliser le mot *list* (liste) quand le texte est suivi par une liste (physique).

La *redondance* en acquisition d'EN est le taux de candidats acquis plus d'une fois pour un moteur de recherche. La redondance vient de ce qu'une même EN peut être acquise à partir de plusieurs listes ou énumérations distinctes au sein du même corpus. En moyenne la redondance est de 2,09. Chaque candidat est acquis un peu plus de deux fois. En outre, une même EN peut être obtenue indépendamment par les deux moteurs de recherche. Le *recouvrement* est le rapport du nombre de candidats acquis conjointement par AV et NL sur le nombre de candidats total. En raison des modes de recherche différents sur les deux moteurs de recherche, des différentes zones d'indexation sur le Web et des types d'index construits, le recouvrement n'est que de 15%. Dans un tel cadre, les deux moteurs de recherche apparaissent plus comme complémentaires que comme concurrentiels parce qu'ils ramènent des ensembles de documents différents.

Évaluation en précision

Parmi les 44 624 EN produites par le module de post-filtrage, un ensemble de test de 339 candidates est choisi. Pour chaque candidate, on formule une requête avancée sur AV composée de la chaîne de cette EN et on relève les 20 premières pages rapportées par le moteur de recherche. Ces pages sont analysées manuellement afin de vérifier que l'EN candidate désigne sans ambiguïté une personne, un lieu ou un organisme du type attendu. L'analyse manuelle des EN candidates donne une précision de 55% dont les erreurs se répartissent ainsi : 25% de type incorrect, 24% incomplètes, 8% sur-complètes et 43% diverses.

L'analyse du contexte proche fournit parfois un **typage incorrect** des EN. Par exemple *Ash-*

ley Judd (une actrice) est incorrectement extraite comme une athlète de

His clientele includes stars and athletes such as Ashley Judd and Mats Sundin. (Sa clientèle inclut des vedettes et des athlètes tels que Ashley Judd et Mats Sundin.)

L'erreur vient d'une analyse incorrecte de l'amorce : seul *athletes* (athlètes) est vu comme hyperonyme alors que *stars* (vedettes) en fait aussi partie. D'autres typages incorrects proviennent de la polysémie des types comme *actors* (acteurs) qui a un sens métaphorique en informatique.

L'extraction **incomplète** de candidats est due principalement à l'expression partielle des EN dans les listes ou les énumérations où elles apparaissent. Par exemple, le nom de l'auteur *Goffman* est extrait de

Readings are drawn from the work of such authors as Laing, Szasz, Goffman, Sartre, Bateson, and Freud. (Des lectures sont extraites des travaux d'auteurs tels que Laing, Szasz, Goffman, Sartre, Bateson et Freud.)

Cette énumération ne contenant pas les prénoms des auteurs, elle ne permet pas une acquisition désambiguïsée des EN les désignant. En outre, certains noms propres, bien que complets, sont ambigus et sont aussi comptés comme des erreurs car ils conduisent à des reconnaissances incorrectes d'EN. Par exemple, *Lucero* est une compagnie de logiciels, une chanteuse et actrice mexicaine, le nom de famille d'un sculpteur, une revue, etc.

Les cas de **sur-complétude** viennent d'une mauvaise reconnaissance de frontière et, donc, de l'incorporation à l'EN de mots indépendants. Par exemple, *O. Shtrichman Next* est extrait d'une liste de noms d'auteurs où le nom *O. Shtrichman* est suivi d'un hyperlien *Next* (Suivant). D'autres erreurs sont dues à une analyse incorrecte de la page Web et ne peuvent être classées par manque de lien clair entre la candidate et l'entité qu'elles désignent.

5. Raffinement des types d'EN

Le typage des EN candidates décrit jusqu'à présent est basé sur l'hyperonyme de l'EN donné dans les séquences (1.a-d). Or, l'amorce qui précède la collection d'EN à extraire contient des précisions supplémentaires sur le type des EN qui la suivent. Plus précisément, l'amorce a quatre fonctions différentes :

1. signaler la présence d'une énumération, par exemple, *Here is* (Voici) ;
2. décrire la nature de la structure de l'énumération, par exemple, *a list of* (une liste de) ;
3. décrire l'**hyperonyme**, c'est-à-dire la catégorie des articles de l'énumération, par exemple, *universities* (universités) ;
4. décrire les **differentiæ**, c'est-à-dire de caractériser les articles de l'énumération, par exemple, *universities in Vietnam* (universités au Vietnam)

Les requêtes utilisées par le moissonneur sont soit des éléments qui introduisent la collection (par exemple, *the following* (suivant)), soit des éléments qui décrivent la structure (par exemple, *a list of* (une liste de)). En général dans un amorce, il n'est pas obligatoire d'exprimer ces deux fonctions de manière explicite en utilisant des marques lexicales, parce que la structure du texte elle-même indique qu'il s'agit d'une collection de ce type. Les lecteurs exploitent les propriétés visuelles d'un texte écrit pour en construire le sens (Péry-Woodley, 1998).

Il est, par contre, nécessaire de rendre explicite la définition des articles d'une collection, puisque le lecteur ne peut pas déduire ces détails des propriétés structurelles du texte. Les

amorces donnent des spécifications supplémentaires qui précisent les *differentiæ* (soulignés dans cet exemple) :

This is a list of American companies with business interests in Latvia.

(Voici une liste de sociétés américaines ayant des intérêts commerciaux en Lettonie)

Cette séquence est la forme la plus explicite d'une amorce, contenant quatre éléments lexicaux dont chacun correspond à une des fonctions décrites ci-dessus. L'extraction des caractéristiques des articles de la collection n'est pas difficile, parce que l'hyperonyme est modifié soit par une proposition relative, soit par un syntagme prépositionnel, soit par un syntagme adjectival. La figure 3⁷ montre une grammaire détaillée de cette forme d'amorce.

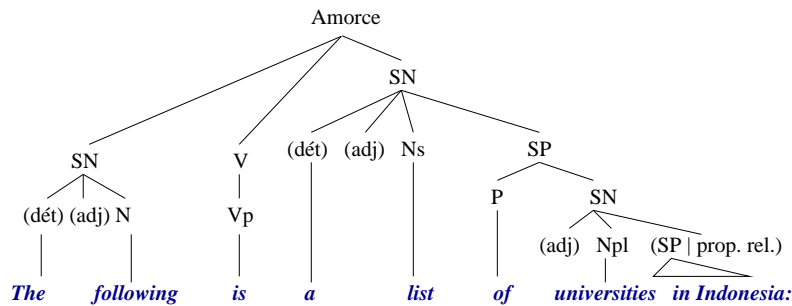


FIG. 3: La structure d'une amorce de base

Avant d'être analysées, les amorces sont étiquetées par le TreeTagger (Schmid, 1999). Les éléments qui expriment les *differentiæ* sont extraits en utilisant des expressions régulières : ils sont toujours les modificateurs du nom au pluriel dans la séquence, ce nom étant l'hyperonyme des articles de la collection.

Les amorces qui contiennent la requête *such as* (tel que) ont un comportement légèrement différent. Elles ne sont pas saturées sur le plan syntaxique et l'élément qui manque est fourni par chaque article de la collection (Virbel *et al.*, 1999). La structure de ces phrases peut prendre plusieurs formes et une analyse syntaxique complexe est nécessaire pour extraire les propriétés de l'hyperonyme. Nous ne l'étudierons pas dans cet article.

Un type d'erreur fréquent dans l'analyse des amorces est due aux paragraphes qui contiennent la chaîne de recherche suivie par une liste qui ne lui est pas liée. Par exemple, le moissonneur reconnaît la séquence

Ask the long list of American companies who have unsuccessfully marketed products in Japan. (Demandez à la longue liste de sociétés américaines qui ont lancé sans succès des produits sur le marché japonais)

comme une amorce, bien qu'elle ne soit pas reliée à une collection. Si cette séquence était suivie sur la page par une collection de n'importe quel type, le système prendrait ses articles en leur attribuant une classe incorrecte, celle fournie par la liste définitoire.

On utilise souvent les mots *list of* (liste de) dans des textes discursifs, donc il est nécessaire de faire un filtrage pour repérer des séquences qui ne jouent pas le rôle d'amorce. Cela nous permet de réduire les erreurs de classifications d'entités. L'analyse des structures syntaxiques a contribué à la construction de plusieurs expressions régulières qui servent à éliminer des *non amorces*, et à ignorer tous les articles qui les suivent.

⁷SP = syntagme prépositionnel, N_s = nom (singulier), N_{pl} = nom (pluriel), adj = adjectif, V_p = verbe au présent, prop. rel. = proposition relative.

Nous avons extrait 1 813 amorces potentielles du corpus de pages HTML trouvées en utilisant AV & NL pour la requête *list of <EN>*. Un analyseur surfacique filtre et analyse les séquences en utilisant des patrons lexico-syntaxiques afin d'identifier des amorces correctes. Il est composé de 14 modules, dont 4 font du préfiltrage afin de préparer et d'étiqueter le corpus. Les 10 autres modules font une analyse syntaxique à grain fin, en rejetant des séquences qui ne jouent pas le rôle d'amorces. Après le filtrage, le corpus est composé de 520 séquences. La précision du processus d'analyse des amorces est 78% et le rappel est 90%.

6. Conclusion

Cette étude est une application démontrant l'utilisabilité du Web comme source de connaissance pour le TAL (voir, par exemple, (Grefenstette, 1999) pour l'utilisation du Web en traduction à base d'exemples). La précision pour l'acquisition des EN n'est que de 55%, nécessitant un filtrage humain qui peut se faire à partir du Web comme cela a été fait pour la validation. Il faut toutefois noter que la valeur de la précision a été établie dans des conditions très exigeantes (rejet des formes partielles et des EN ambiguës). Accepter de telles formes et ne rejeter que les EN candidates incorrectes met la précision à 81%. Quant à la précision de l'analyse des amorces, elle est de 78% et définit donc un typage précis des EN.

Références

- AMITAY E. (1999). Anchors in context: A corpus analysis of web pages authoring conventions. In L. PEMBERTON & S. SHURVILLE, Eds., *Words on the Web - Computer Mediated Communication*, p. 192. UK: Intellect Books.
- B. BOGURAEV & J. PUSTEJOVSKY, Eds. (1996). *Corpus Processing for Lexical Acquisition*. Cambridge, MA: MIT Press.
- CRIMMINS F., SMEATON A., DKAKI T. & MOTHE J. (1999). Tétrafusion: Information discovery on the internet. *IEEE Intelligent Systems and Their Applications*, **14**(4), 55–62.
- GREFENSTETTE G. (1999). The WWW as a resource for example-based MT tasks. In *Proc., ASLIB Translating and the Computer 21 Conference*, London.
- GRICE H. P. (1975). Logic and conversation. In P. COLE & J. MORGAN, Eds., *Speech acts. Syntax and semantics, Vol. 3*, p. 41–58. NY: Academic Press.
- HEARST M. (1998). Automated discovery of WordNet relations. In C. FELLBAUM, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- KUSHMERICK N., WELD D. & DOORENBOS R. (1997). Wrapper induction for information extraction. In *Proc., IJCAI'97*, p. 729–735, Nagoya.
- MORIN E. & JACQUEMIN C. (1999). Projecting corpus-based semantic links on a thesaurus. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 389–396, University of Maryland.
- MUC-6 (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. San Mateo, CA: Morgan Kaufmann.
- PÉRY-WOODLEY M.-P. (1998). Signalling in written text: a corpus based approach. In *Workshop on Discourse Relations and Discourse Markers at COLING-ALC'98*, p. 79–85.
- SCHMID H. (1999). Improvements in part-of-speech tagging with an application to German. In S. ARMSTRONG et al., Ed., *Natural Language Processing Using Very Large Corpora*. Dordrecht: Kluwer.
- VIRBEL J., PÉRY-WOODLEY M.-P., GARCIA-DEBANC C., MOJAHID M. & LUC C. (1999). *A Linguistic Approach to Some Parameters of Layout: A Study of Enumerations*. Rapport interne, AAI.