Revue québécoise de linguistique

Des déclencheurs des énumérations d'entités nommées sur le Web

Caroline Bush

Volume 32, numéro 2, 2003

URI : https://id.erudit.org/iderudit/017542ar
DOI : https://doi.org/10.7202/017542ar

Aller au sommaire du numéro

Éditeur(s)

Université du Québec à Montréal

ISSN

0710-0167 (imprimé)
1705-4591 (numérique)

Découvrir la revue

Citer cet article

Bush, C. (2003). Des déclencheurs des énumérations d'entités nommées sur le Web. *Revue québécoise de linguistique*, *32*(2), 47–81. https://doi.org/10.7202/017542ar

Résumé de l'article

Le Web est une importante source pour l'acquisition lexicale parce qu'il est continuellement mis à jour. Des énumérations sont particulièrement fréquentes dans les pages Web, parce que ces dernières exigent une structure claire qui facilite la compréhension du lecteur. Nous considérons des énumérations d'entités nommées et nous nous focalisons sur une structure linguistique particulière : le déclencheur – la séquence qui introduit l'énumération. Ayant des fonctions précises, la structure d'un déclencheur est assez limitée. Ce travail a pour but de modéliser cette structure à partir des analyses linguistiques interprétatives et descriptives. Ces modèles contribuent au développement d'un système d'acquisition et de classification d'entités nommées à partir du Web.

Revue québécoise de linguistique, vol. 32, nº 2, 2003, © RQL (UQAM), Montréal

DES DÉCLENCHEURS DES ÉNUMÉRATIONS D'ENTITÉS NOMMÉES SUR LE WEB

Caroline Bush
Laboratoire d'Informatique pour la Mécanique et
les Sciences de l'Ingénieur (LIMSI), CNRS

1. Introduction

Le Web est une importante source pour l'acquisition lexicale (Crimmins et coll. 1999), et surtout pour l'acquisition des entités nommées, parce qu'il est continuellement mis à jour. Amitay 1997 et Hunter 1998 étudient la linguistique du langage et des structures employés dans les documents hypertextuels. Notre travail se situe dans le cadre de ces études, au sein d'un domaine actuellement en plein essor.

Il existe des énumérations dans tous les types de textes, mais elles sont particulièrement fréquentes dans les pages Web, parce que ces dernières exigent une structure claire qui facilite la compréhension du lecteur. Ces énumérations partagent un certain nombre de caractéristiques communes, et nous pensons qu'il faut utiliser ces traits communs pour une exploitation automatique efficace de ces sources d'information. Nous étudions dans ce document un trait particulier : le **déclencheur**, c'est-à-dire la phrase ou la séquence qui introduit l'énumération.

Parce qu'elle a des fonctions précises, la structure d'un déclencheur est assez limitée. Ce travail a pour but de modéliser cette structure à partir des analyses linguistiques interprétatives et descriptives. Ces modèles contribuent au développement d'un système d'acquisition et de classification d'entités nommées (EN) à partir du Web (Jacquemin et Bush 2000).

Dans ce document, nous ne traitons pas en détail le balisage HTML, qui est un trait distinctif des documents sur le Web. Ce balisage fournit de nombreuses façons différentes de construire une énumération sur le Web, donc il n'est pas pratique d'exploiter les balises HTML afin de repérer ces structures. Cependant, une fois qu'on a trouvé une énumération, des connaissances à

propos du balisage HTML permettent d'en extraire des données, exploitables dans l'extraction des EN à partir des énumérations.

1.1 Contribution à l'analyse automatique des déclencheurs des énumérations d'entités nommées

Pour mieux définir une énumération, nous prenons la définition de Pascual 1991 qui dit qu'énumérer, «c'est conférer une égalité d'importance à un ensemble d'objets, et ensuite c'est ordonner ces objets, selon des critères variés». Les articles d'une énumération sont des entités qui ont une fonction commune et qui sont marquées des mêmes traits de formatage. Plusieurs idées dans le travail de Luc et coll. 1999 sur des énumérations standard et non standard nous servent de notions de base dans ce document. Ces auteurs ont noté que toute énumération dans leur corpus possède un élément introducteur, qui peut avoir quatre types de balises : des balises lexicales, syntaxiques, typographiques, et dispositionnelles. C'est cette idée que nous utilisons comme thème central dans ce travail et que nous allons examiner beaucoup plus en détail. Virbel 1985 introduit le terme de *mise en forme matérielle* pour l'ensemble des caractéristiques physiques d'un texte édité, ce qui est très important lorsqu'on considère des textes structurés tels que les énumérations.

Le but de notre travail est d'élargir les connaissances sur l'élément introducteur, le **déclencheur**, en précisant ses propriétés plus en détail. Ces analyses contribuent à la conception d'un outil pour l'identification et l'analyse des séquences qui jouent le rôle de déclencheurs.

Notre corpus est construit de séquences tirées du Web en utilisant deux méthodes différentes. Presque 90 % des séquences ont été collectées automatiquement par un outil d'acquisition d'entités nommées (Jacquemin et Bush 2000). Le reste des déclencheurs dans le corpus a été extrait du Web manuellement en utilisant des requêtes d'entités nommées particulières dans des moteurs de recherche. Le corpus entier contient 650 déclencheurs.

Dans ce document, nous commençons par présenter une analyse linguistique des déclencheurs, d'abord sur le plan sémantique (section 2) et ensuite sur les plans syntaxique et structurel (section 3). Nous proposons une forme canonique pour un déclencheur (section 3.1) et nous en formulons un modèle. Puis, nous traitons des formes réduites de cette structure canonique et nous formalisons les règles qui contraignent les réductions (section 4). Finalement nous décrivons la contribution de ce travail à un système d'acquisition et de classification des entités nommées à partir du Web, et nous faisons une évaluation de nos programmes (section 5). Comme point de départ, nous allons décrire les analyses linguistiques des séquences qui jouent le rôle de déclencheurs.

2. Le rôle sémantique des déclencheurs

Dans cette section, nous allons étudier les déclencheurs du point de vue sémantique, en considérant leur rôle quant aux énumérations qu'ils introduisent. Nous partons de quelques séquences qui jouent le rôle de déclencheurs et que nous analysons au niveau de leurs fonctions et de leurs constituants sémantiques.

2.1 Les constituants sémantiques d'un déclencheur

Afin de bien comprendre la structure d'un déclencheur, il faut d'abord comprendre son rôle. Les quatre fonctions d'un déclencheur sont premièrement de signaler la présence d'une énumération, deuxièmement de décrire la nature de la structure de l'énumération, troisièmement de décrire l'**hyperonyme**, c'est-à-dire la catégorie des articles de l'énumération, et quatrièmement, de décrire les **differentia**, c'est-à-dire de caractériser les articles de l'énumération.

Il est donc logique qu'un déclencheur se compose de quatre éléments, dont certains sont facultatifs, qui correspondent chacun à une de ces fonctions. Nous allons maintenant essayer de les identifier. Prenons un déclencheur typique :

(1) The following is `a list of` *publicly funded* ***universities*** *in Vietnam.*
Ce qui suit est une liste d'universités du Vietnam qui sont financées par l'État.

Dans cet exemple, repris dans la figure 1, nous pouvons détecter les quatre constituants distincts.

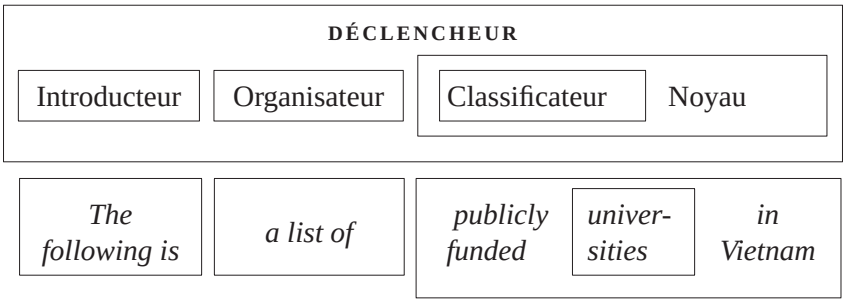


Fig. 1 : Modèle des éléments d'un déclencheur

Au centre, on a le **noyau sémantique** qui précise la nature des articles dans l'énumération (*publicly funded universities in Vietnam*). Il n'est pas possible de déduire cette information en étudiant la structure de l'énumération. Dans

ce noyau, il existe ce qu'on peut appeler un **classificateur** (ici *universities*), c'est-à-dire l'élément qui précise le type de l'article. Nous trouvons ensuite l'élément qui décrit la structure de l'énumération : **l'organisateur**. Dans l'exemple (1), il prend la forme *a list of*. Enfin, il y a **l'introducteur**, l'élément qui introduit la présence et la proximité de l'énumération. Ici, l'introducteur est *The following is*.

Après avoir constaté l'existence des quatre éléments, il faut les étudier plus précisément pour bien les définir. Nous pouvons classer les éléments en deux groupes principaux : ceux qui servent à **localiser** l'énumération et ceux qui font du **typage**, en expliquant soit le type de la structure, soit le type des articles de l'énumération.

2.2 La localisation

C'est l'**introducteur** qui indique la localisation spatiale de l'énumération. Il est toujours composé d'un lexème **déictique** d'un des types suivants :

Tableau 1
Pourcentage des déclencheurs du corpus
utilisant des mots déictiques

MOTS DÉICTIQUES		% DES DÉCLENCHEURS
following	'suivant'	6
this	'ceci'	4
here	'voici'	2,5
below	'ci-dessous'	2,5

Le tableau 1 montre le pourcentage des déclencheurs du corpus qui contiennent chacun des mots déictiques. Ces mots sont une référence qui renvoie à la situation particulière de l'énonciation plutôt qu'aux éléments du monde (Gezundhajt 1999). Placés hors contexte, leur référence manque. Mais dans un texte, ils indiquent que l'énumération se trouve dans le **cotexte**, c'est-à-dire dans l'environnement textuel immédiat (Maingueneau 1996). L'introducteur est donc une **cataphore**, c'est-à-dire une référence déictique qui annonce par un substitut une partie du contexte à venir.

L'emploi de ces mots constitue une **catachrèse** spatio-temporelle. Certains mots acquièrent de nouveaux sens au cours du temps par le phénomène de la catachrèse (Ferrari 1997). Autrement dit, la catachrèse est un écart que certains mots font de leur première signification pour en prendre une autre qui y a rapport. Par exemple, le mot *below* (ci-dessous) exprime une position qui

est au-dessous, donc après dans le texte comme si l'orientation de la page était verticale. De la même façon, le sens littéral du mot *following* (suivant) est *plus tard sur une échelle temporelle*. Ce type de métaphore est très courant dans l'écriture des textes et il est donc devenu lexicalisé.

2.3 Le typage

Les trois autres éléments d'un déclencheur servent à typer soit les articles de l'énumération, soit l'énumération elle-même. Examinons tout d'abord le noyau et le classificateur qui typent les articles.

2.3.1 Le typage des articles

Le noyau est le seul élément obligatoire dans un déclencheur parce qu'il n'y a pas d'autres indices, ni sémantiques ni structurels, qui indiquent ce que l'énumération contient. Il caractérise la nature de chaque article qui compose l'énumération, en décrivant les conditions nécessaires pour qu'un article y apparaisse. Le classificateur est la sous-structure du noyau qui rend explicite le **type large** des articles qui composent l'énumération. Prenons l'énumération suivante :

(2) *The following is a list of universities with field camps.*
- *Georgia State University*
- *Ohio University*
- *University of Tennessee*

Ce qui suit est une liste d'universités avec des campements. […].

Ici, chaque article de l'énumération est une université. Le mot *universities* caractérise le type des articles; il est donc le classificateur. Cet élément est presque toujours au pluriel parce que l'énumération est une collection de plusieurs exemples de ce type, ainsi plusieurs personnes ou plusieurs universités. Si nous disons que l'énumération a *n* articles, le classificateur recouvre les articles 1 à *n*.

Nous pouvons définir le noyau et le classificateur de la manière suivante : le **classificateur** est le nom qui donne le type des articles de l'énumération, tandis que le **noyau** se compose du classificateur et de tous ses modifieurs.

La relation entre ces deux éléments est une **hyponymie**, une relation transitive (Hearst 1998) et paradigmatique (Borillo 1995). L'hyponymie dénote le fait que le sens d'un mot est inclus dans celui d'un autre. L'hyperonyme, le terme le plus général, exprime le **genus** (le type); et l'hyponyme, le terme plus spécifique, est composé du genus ainsi que de ses **differentiae**, qui font

la différence entre ce terme et les autres membres de la classe (Leech 1977). La figure 2 montre que le classificateur est l'hyperonyme du noyau, et que le noyau est l'hyperonyme de chaque article de l'énumération. Tous les articles sont des membres d'un paradigme : ils sont des hyponymes du noyau, et donc des hyponymes du classificateur. Puisque les relations entre leurs articles sont paradigmatiques, les énumérations d'entités nommées peuvent être caractérisées comme des **énumérations paradigmatiques** (Luc et coll. 2000).

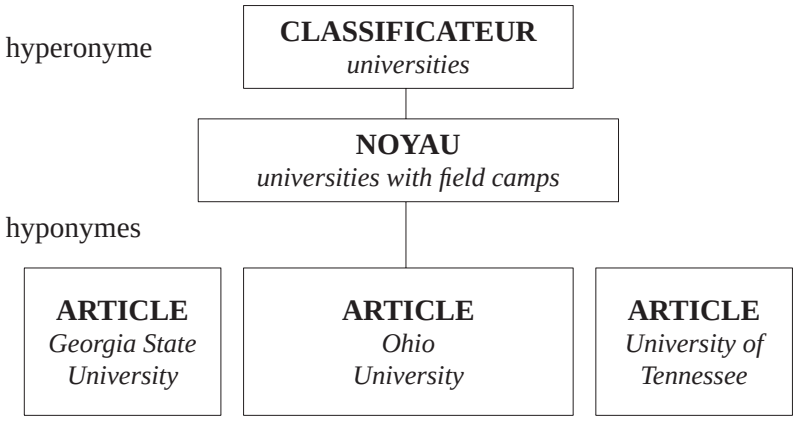


Fig. 2 : L'ontologie du noyau

En fait, le noyau est un exemple de discours **définitoire**, parce qu'il définit les caractéristiques des articles de l'énumération. L'hyponymie est utilisée dans la construction d'un grand nombre de définitions (Borillo 1995). Péry-Woodley 1998 montre qu'une définition est typiquement constituée d'un élément qui donne la classe (l'hyperonyme), ainsi que de modifieurs qui expriment la spécificité.

2.3.2 Le typage de la structure de l'énumération

L'**organisateur** a pour rôle de typer la structure de l'énumération. Il explicite l'organisation des données qui composent l'énumération et déclare le type de la structure, qui peut être soit une liste, soit un tableau.

Dans la plupart des exemples de notre corpus, l'organisateur a la forme suivante (où *ADJ* est un adjectif) :

(3) *(a)* (ADJ)* *list*(*ing*) *of*
par exemple : *a long list of*
'une longue liste de'

2.4 Synthèse

Nous venons de définir les quatre éléments d'un déclencheur du point de vue de leur rôle sémantique. Le **noyau** caractérise des conditions nécessaires pour l'apparition d'un article dans l'énumération. Il est l'élément minimal qui peut fonctionner comme déclencheur, et c'est un hyponyme du classificateur, décrivant la catégorie fine des articles. Il contient le **classificateur** qui rend explicite le type des articles de l'énumération et qui est l'hyperonyme du noyau, décrivant la catégorie de base des articles. L'**organisateur** déclare le type de la structure de l'énumération. L'**introducteur** est une cataphore dont l'antécédent est l'énumération elle-même, qui indique la proximité de l'énumération.

Étudions maintenant les déclencheurs du point de vue de leur structure.

3. La structure syntaxique d'un déclencheur

Dans la section précédente, nous avons défini les constituants d'un déclencheur ainsi que leur rôle sémantique. Dans cette section, nous examinons les déclencheurs du point de vue de leurs structures syntaxiques. Les déclencheurs ont une structure canonique dont nous définissons les variantes acceptables.

Bien qu'elles partagent les éléments communs présentés ci-dessus, les séquences qui jouent le rôle de déclencheurs peuvent avoir plusieurs formes. Les structures différentes peuvent être groupées dans deux classes principales : celles auxquelles il manque un élément syntaxique, ce qu'on va appeler des structures **non saturées** (exemple (4)), et celles qui sont des phrases complètes, ce qu'on appellera des structures **saturées** (exemple (5)).

(4) *Some of the well-known graduates from the La Guardia school include:*
La liste des diplômés célèbres de l'école La Guardia comprend :

(5) *This is a list of international organizations which are searching for volunteers.*
Voici une liste d'organisations internationales qui cherchent des bénévoles.

Cette distinction a déjà été faite chez Luc et coll. 1999, qui parlent d'**amorces incomplètes** et d'**amorces complètes**. Nos constatations sont en accord avec leurs définitions, d'après lesquelles, pour un déclencheur sous la forme d'une phrase incomplète, chaque article de l'énumération remplit le trou dans la

structure syntaxique, permettant ainsi de rendre la phrase complète. Les structures saturées fonctionnent comme des titres qui annoncent l'énumération.

3.1 La forme canonique d'un déclencheur

Un déclencheur saturé est une phrase complète presque indépendante. Séparé de son énumération, il manquerait de référence, mais il serait bien formé sur le plan grammatical. La dénomination anglaise de ce type de séquences données par Luc et coll. 1999, **leading sentences** (les amorces complètes), est particulièrement juste parce qu'il s'agit de phrases qui mènent à l'énumération.

La forme canonique d'un déclencheur est une structure saturée qui rend explicites tous les éléments identifiés dans la section 2.1. Voici un exemple :

(6) *This is a list of American companies with business interests in Latvia.*
Voici une liste de sociétés américaines qui ont des intérêts commerciaux en Lettonie.

Cette séquence comprend une indication de la proximité de l'énumération, l'introducteur (*This is*), une référence au type de la structure, l'organisateur (*a list of*), ainsi qu'une identification du type de chaque article, le classificateur (*companies*). Les caractéristiques des articles sont précisées dans le noyau *American companies with business interests in Latvia.* Le tableau 2 montre plusieurs déclencheurs, soit de la forme canonique (qui est saturée), soit d'une forme réduite, présentés avec leurs différents éléments. Dans ce tableau, les éléments en gras peuvent se déduire des autres éléments du déclencheur (ils ne sont pas explicités dans la séquence), et les références sont co-indexées.

Dans notre corpus, on rencontre plusieurs formes réduites de cette structure canonique. Le phénomène sera traité dans la section 4. Nous allons maintenant décrire notre modèle de la structure syntaxique de la forme canonique d'un déclencheur.

Tableau 2
Les composants des déclencheurs

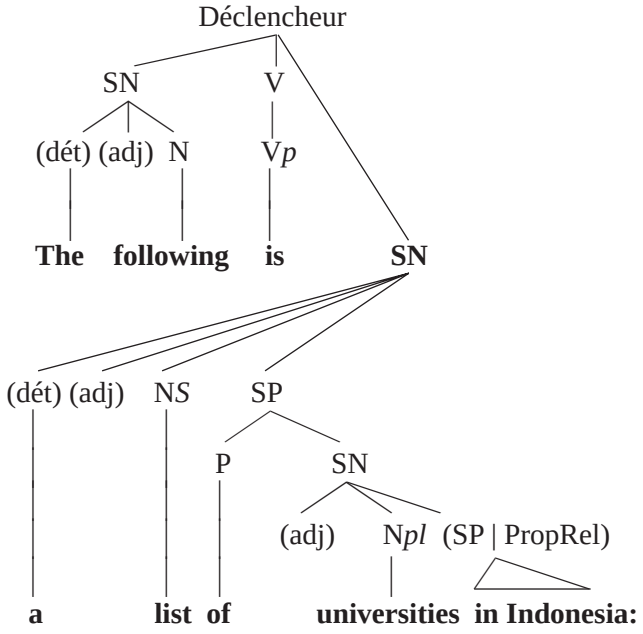
		NOYAU	
INTRODUCTEUR	**ORGANISATEUR**	**CLASSIFICATEUR**	**MODIFIEURS**
Here's	*a list of*	*people*	*who made significant contributions to the creation and growth of the software products industry.*
Voici	une liste de	personnes	qui ont fait des contributions considérables à la création et à la croissance de l'industrie du logiciel.
This list$_2$ contains	***list$_2$***	**[*competitions$_3$*]**	*all important Hungarian orienteering competitions$_3$.*
Cette liste contient			toutes les compétitions hongroises importantes en course d'orientation.
This page gives	*a ranked listing of*	*journals*	*to which the School of Biological Sciences currently subscribes.*
Cette page donne	un listage classé de	revues	auxquelles le Département des Sciences Biologiques est actuellement abonné.
[*following list*]	*List of*	*towns*	*of Zala county*
	Une liste de	villes	de la région Zala
[*following list*]	*complete list of*	resorts	
	Une liste complète de	stations estivales	
[*following list*]	**[*list*]**	**[*places*]**	*Where books have been purchased.*
			Lieux où on a acheté des livres.

Dans notre corpus, on rencontre plusieurs formes réduites de cette structure canonique. Le phénomène sera traité dans la section 4. Nous allons maintenant décrire notre modèle de la structure syntaxique de la forme canonique d'un déclencheur.

3.2 Un modèle de la structure de la forme canonique

La forme canonique d'un déclencheur est composée d'un introducteur, d'un organisateur et d'un noyau qui contient un classificateur. La figure 3 montre les constituants syntaxiques de cette structure, et le tableau 3 donne la structure syntaxique de chacun des éléments, ainsi qu'une récapitulation de leur fonction sémantique respective.

L'introducteur est un syntagme nominal qui peut être modifié par un adjectif. Dans cette forme étendue, il y a toujours un verbe conjugué et ce verbe est souvent une forme de *to be*. L'organisateur est un syntagme nominal qui comprend un syntagme prépositionnel dont la préposition est *of*, par exemple *a list of*. Ce syntagme prépositionnel attend le noyau pour le compléter. Le noyau est toujours un syntagme nominal dont la tête est le classificateur.



dét = déterminant, SN = syntagme nominal, N*s* = nom au singulier, N*pl* = nom au pluriel, adj = adjectif, PropRel = proposition relative, V*p* = verbe au présent, SP = syntagme prépositionnel

Fig. 3 : La structure syntaxique de la forme canonique

Tableau 3
La structure des éléments de la forme canonique

EXEMPLE	*The following*	*is*	*a list of*	*universities in Indonesia*
	Voici		une liste de	universités d'Indonésie
FONCTIONS	**INTRODUCTEUR**		**ORGANISATEUR**	**NOYAU**
SÉMANTIQUE	Cataphore		Structure	Catégorie (hyperonyme + modifieurs)
SYNTAXIQUE	SN	V	SN	SN
	(dét) (adj) N	V_p	(dét) (adj) N_s (SP) of	(adj) N_{pl} (SP|propRel)

Nous allons maintenant étudier des structures saturées qui n'ont pas la forme canonique.

3.3 Structures saturées non canoniques

Un type particulier de structure saturée jouant le rôle de déclencheur a comme classificateur un syntagme nominal qui contient le modifieur *following* (suivant), par exemple *the following universities* (les universités suivantes). Ce modifieur apparaît aussi dans les introducteurs des déclencheurs, par exemple *the following is a list of* (ce qui suit est une liste de) ou *the following list* (la liste suivante). La tête du syntagme modifiée par le mot *following* (suivant) est donc le classificateur sauf si elle est un lexème dénotant une structure énumérative comme *list* (liste), *listing* (listage), *collection* (ensemble), etc. Dans ce cas, elle représente l'organisateur.

Dans l'exemple (7), le syntagme nominal *the following institutions* (les institutions suivantes) est une **cataphore** pour les articles de l'énumération. Nous pouvons remplacer ce syntagme nominal par l'ensemble des articles et obtenir une phrase bien formée. Le seul changement exigé est syntaxique : il faut ajouter des virgules et une conjonction afin de transformer la liste en une chaîne de texte. Ainsi l'énumération suivante :

(7) *Since January 1, 1997, the **following** institutions have become JSTOR participants:*
– *Agnes Scott College*
– *Albion College*
– *Ohio State University*
Depuis le 1er janvier 1997, les institutions suivantes sont devenues des participants de JSTOR : […]

peut se transformer en :

(8) *Since January 1, 1997, **Agnes Scott College, Albion College** and **Ohio State University** have become JSTOR participants.*

La combinaison du déclencheur et son énumération (7) est sémantiquement équivalente à la phrase bien formée (8). Ce type de structure est une variante de la forme canonique d'un déclencheur et n'est pas saturé parce qu'il ne contient pas d'organisateur explicite.

3.4 Les déclencheurs non saturés

L'autre type de déclencheur est une structure non saturée. Il s'agit de phrases syntaxiquement incomplètes dont les constituants manquants sont fournis par les articles de l'énumération (Luc et coll. 1999). Nous les analysons plus en détail dans cette section.

Les déclencheurs qui contiennent des lexèmes tels que *such ... as* (tel que) ou une forme du verbe *include* sont des phrases avec un trou syntaxique. Chaque article de la liste remplit ce trou. Donc, sans la liste, la phrase est incomplète. Par exemple :

(9) *In North America, there are performing arts schools like the one in «FAME – THE MUSICAL" in such cities as:*
– *Baltimore*
– *Boston*
En Amérique du Nord, il y a des écoles d'arts du spectacle comme celle de la comédie musicale FAME dans des villes telles que : […].

(10) *Some of the well-known graduates from the La Guardia school include*:
– *Diahann Carroll*
– *Janis Ian*

La liste des diplômés célèbres de l'école
La Guardia comprend :
[…]

Ces déclencheurs ne sont pas des phrases complètes. Dans le déclencheur de l'exemple (9), le lexème qui introduit les articles de l'énumération est *such as* (tel que), et il exige un syntagme nominal (SN) pour compléter la phrase. Dans la séquence (10), il y a un trou après le verbe *include* (comprendre), parce que ce verbe attend un SN comme complément d'objet :

Some of the well-known graduates from the
La Guardia school include [X]
où *X* = chaque élément de l'énumération.

Les quatre constituants sémantiques (le noyau, le classificateur, l'organisateur et l'introducteur) sont plus difficiles à mettre en évidence dans un déclencheur non saturé, en particulier parce que sa structure est moins régulière.

Dans la séquence (10), la chaîne qui caractérise les articles de l'énumération (le noyau) est *well-known graduates from La Guardia school*. Dans certains cas, une réorganisation de mots est nécessaire pour former une séquence grammaticale, mais ce n'est pas le cas pour cet exemple. Cette chaîne est le plus petit élément qui peut servir à introduire cette énumération. Comme toujours, le classificateur est la tête syntagmatique de ce noyau (ici, *graduates*).

Dans cet exemple, le signe de ponctuation est un **deux-points** qui indique que ce qui suit est lié. Ici c'est la ponctuation, une marque typographique, liée aux exigences syntaxiques du lexème *include*, qui joue le rôle de l'introducteur de l'énumération en indiquant qu'elle se trouve à proximité du déclencheur. Dans cette séquence, l'organisateur est implicite parce que le déclencheur ne décrit pas la structure de l'énumération.

Des expressions telles que *such as* et *include* sont des **indicateurs de l'apposition** (Greenbaum et Quirk 1973). Elles se trouvent entre des appositives et indiquent explicitement l'apposition. Possédant des instructions sémantiques particulières, chaque indicateur correspond à des cas particuliers d'apposition. Toutefois, les séquences suivantes sont presque synonymes, et pourraient toutes introduire la même énumération : *such cities as [TROU]*, *cities such as [TROU]* (des villes telles que) ; *cities include [TROU]* (ces villes comprennent), *cities including [TROU]* (plusieurs villes, dont).

Les indicateurs d'apposition *such as* et *including* doivent précéder la deuxième appositive. Ils exigent donc un élément qui les suit pour former une phrase syntaxiquement complète, comme dans les exemples (9) et (10).

3.5 Transformations en une forme canonique

Dans les sections précédentes, nous avons vu quelques exemples de structures qui jouent le rôle de déclencheurs. En ce qui concerne les structures non saturées, l'identification des constituants sémantiques peut poser un problème. Il est donc peut-être utile de regarder ces déclencheurs du point de vue de leur relation à la forme canonique.

Nous pouvons réécrire les exemples (7) et (10) sous la forme canonique saturée, sans changer le sens global du déclencheur. Les éléments qui ne sont pas explicités dans la forme non saturée sont soulignés ici:

Tableau 4
Réécriture en forme canonique saturée

INTRODUCTEUR	ORGANISATEUR	NOYAU [CLASSIFICATEUR]
Following is	*a list of*	***institutions* that have become JSTOR participants since January 1, 1997:**
Voici	une liste d'	institutions qui sont devenues des participants de JSTOR depuis le 1er janvier 1997 :
This is	*a list of*	*well-known* **graduates from La Guardia school**.
Ceci est	une liste des	diplômés célèbres de l'école La Guardia.

Dans les séquences non saturées, il peut être difficile de reconnaître des constituants sémantiques. Il est possible de changer toutes les variantes de déclencheurs en une forme canonique composée d'un introducteur, d'un organisateur et d'un noyau contenant lui-même un classificateur, la structure expliquée visuellement par la figure 1. Cette structure est la forme la plus explicite d'un déclencheur et elle facilite l'identification des éléments sémantiques. Nous avons étudié la transformation des formes variantes en forme canonique dans Bush 2000, mais nous ne présentons pas dans cet article le mécanisme utilisé, car il est relativement complexe.

3.6 Des structures non traitées

D'autres types de structures sont rarement employés pour introduire des énumérations. Par exemple, il existe des titres qui introduisent une énumération sans donner explicitement ni la présence de l'énumération, ni le type des éléments qui la composent, comme *Sightseeing in Trento* (Le tourisme à Trente). Cet exemple introduit une liste d'attractions à Trente. Dans ce cas, c'est la structure elle-même qui annonce ce dont il s'agit, et le déclencheur ne respecte pas les fonctions décrites à la section 2.1, mais donne plutôt une idée générale du sujet de l'énumération. Il exige du lecteur une reconstruction importante pour qu'il comprenne qu'il s'agit d'une énumération. Un tel exemple fait donc appel à beaucoup d'informations pragmatiques. C'est pourquoi nous ne traitons pas ce type de déclencheurs dans cet article.

4. Des formes réduites de la structure canonique

Nous avons identifié la structure canonique d'un déclencheur dans la section 3.1. La plupart des énumérations ne sont cependant pas introduites par la forme canonique entière : environ 70 % des déclencheurs dans notre corpus sont des formes réduites de la forme canonique. Par exemple :

(11) `[List of]`$_{O}$ [[***towns***]$_{C}$ *in Zala country*]$_{N}$
Une liste de villes de la région de Zala

(12) [[***Colleges and Universities***]$_{C}$ *from all over the world*]$_{N}$
Des collèges et des universités du monde entier

où *I* = Introducteur; *O* = Organisateur; *C* = Classificateur; *N* = Noyau.

La séquence (11) n'a pas d'introducteur et la séquence (12) n'est composée que du noyau et de son classificateur. Toutefois, ces deux expressions peuvent jouer le rôle de déclencheurs et, comme nous allons le voir, l'emploi de ces formes réduites peut être justifié dans certains cas. Dans cette section, nous allons examiner quelques cas spécifiques d'emploi de ces réductions et formaliser les règles concernant les formes réduites qui sont acceptables dans le rôle de déclencheur.

4.1 Justification de l'emploi de formes réduites

Nous avons identifié les quatre constituants sémantiques d'un déclencheur (section 2.1). Il faut toutefois considérer le rôle des caractéristiques physiques du

texte comme des marques typographiques et sémantiques. Elles jouent un rôle aussi important que les marques lexico-sémantiques étudiées jusqu'à présent en introduisant et en caractérisant des énumérations.

Chaque article d'une énumération typique d'entités nommées sur le Web se trouve sur une nouvelle ligne avec les mêmes traits de formatage. Cette forme de texte indique par elle-même qu'il s'agit d'une liste, ce qui permet d'omettre l'organisateur du déclencheur sans nuire à la compréhension. Péry-Woodley 1998 montre comment un texte écrit est un objet visuel dont on exploite les propriétés visuelles pour reconstruire le sens.

Les marques typographiques comme la ponctuation sont importantes pour l'acceptabilité de certaines formes réduites qui jouent le rôle de déclencheurs. La mise en forme matérielle (MFM) d'une chaîne de texte donnée renvoie à certaines contreparties entièrement ou partiellement discursives (Virbel 1985, Virbel 1989). Par exemple, le contraste entre le type de caractères utilisé pour le déclencheur (gras, souligné, couleur différente) et celui employé pour les articles de l'énumération indique le rôle de déclencheur de l'énumération. Comme Virbel (1985) l'a fait pour les définitions, nous pouvons identifier la forme entièrement discursive d'un déclencheur, ainsi que sa contrepartie basée sur la mise en forme :

(13) *Following is a list of lakes in Tillamook County.*
Voici une liste de lacs du comté de Tillamook.

(14) ***Lakes in Tillamook County:***

Dans l'exemple (14), la combinaison des traits de formatage (le texte en gras et souligné) et l'emploi du deux-points (:) prennent la place de l'introducteur en introduisant l'énumération. Sans formatage et sans le deux-points, cette chaîne serait un déclencheur beaucoup moins convaincant. Ainsi, certains phénomènes de MFM peuvent être interprétés comme des équivalents non discursifs de certains énoncés ou même comme des traces de réductions des formes discursives.

Catach 1994 explique que le deux-points constitue «la marque principale et quasi-unique de la prise de distance entre segments majeurs, différents du point de vue de la construction, mais liés au point de vue du sens». Dans le cas des énumérations et de leurs déclencheurs, le deux-points indique que les deux éléments sont liés, mais il sert à séparer les deux types de structure du texte. Un guide de la ponctuation anglaise (Peck 1996) explique de la même manière que le deux-points fixe l'attention du lecteur sur ce qui suit. Il en résulte que ce signe est employé afin d'introduire une idée qui complète l'idée introductrice, comme une liste complète l'idée introduite par le déclencheur.

Souvent, à la fin des déclencheurs de la forme canonique – la forme entièrement discursive – on emploie un point à la place du deux-points. Or, le point est une forme de ponctuation qui n'indique pas de façon explicite qu'il y a un élément lié qui suit. La forme discursive rend explicite la référence à l'énumération en utilisant des marques lexicales telles que *following* (suivant), rendant redondante une notation telle que le deux-points.

La réduction linguistique est présente dans plusieurs aspects de la conception des pages Web. Le pourcentage de déterminants dans les documents hypertextuels est considérablement inférieur à celui de l'anglais écrit standard (Amitay 1999), ce qui indique qu'il y a une véritable tendance à la réduction des mots dont l'absence ne pose pas de problèmes pour la compréhension. De plus, afin de faciliter la navigation et la lisibilité, il y a dans les pages Web beaucoup moins de conjonctions, qui rendraient la structure plus compliquée. Le même principe explique la tendance à omettre des éléments du déclencheur qui sont redondants sur le plan sémantique. On préfère simplifier autant que possible les structures employées.

On peut alors se demander pourquoi utiliser la forme étendue d'un déclencheur si les marques typographiques et dispositionnelles portent les mêmes informations. Hunter 1998 dit qu'afin d'aider la compréhension, il vaut mieux signaler le contexte explicitement, même si cela entraîne des répétitions. C'est surtout important pour les lecteurs dont l'anglais n'est pas la première langue, comme beaucoup d'utilisateurs du Web. Si la structure de surface et les structures conceptuelles sous-jacentes sont parallèles, cela facilite la compréhension.

4.2 Modèle du déclencheur en contexte

Péry-Woodley 1998 suggère qu'il est possible d'élargir l'idée des **marques sémantiques** en intégrant dans la description sémantique d'une définition des traits de formatage et de typographie, pouvant jouer un rôle équivalent aux marques lexicales. De la même manière, dans la section précédente, nous avons vu que certaines des marques dispositionnelles et typographiques de l'énumération sont associées aux fonctions du déclencheur. Il est donc utile de formuler un modèle du déclencheur en contexte afin de prendre en compte ces propriétés.

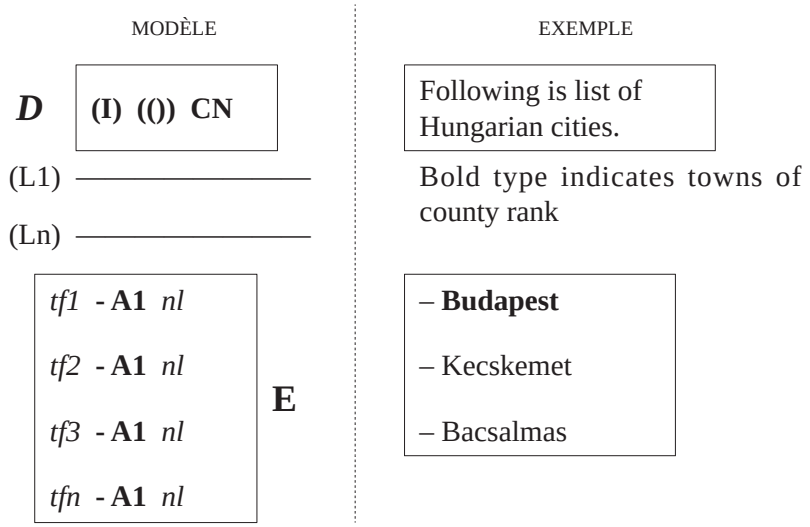


Fig. 4 : Modèle d'une énumération avec son déclencheur

Traduction de l'exemple : *Voici une liste de villes hongroises. Le texte gras indique des villes régionales.*

La figure 4 montre un déclencheur en contexte suivi par son énumération, avec éventuellement quelque lignes de texte entre ces deux éléments. Dans cette figure, nous employons la notation suivante : *A* = Article de l'énumération; *D* = Déclencheur; *L* = Ligne de texte; *E* = Énumération; *tf* = traits de formatage; *nl* = nouvelle ligne.

Avec cette notation, on peut exprimer des règles qui prennent en compte le rôle des propriétés structurelles d'une façon plus formelle :

(15) $(tf_1 = tf_2 = tf_3 = tf_4) + (nl_1 + nl_2 + nl_3 + nl_4) \Rightarrow$ *Structure = Liste (organisateur)*

Cette règle explique que si chaque article se trouve sur une nouvelle ligne et si tous les articles partagent les mêmes traits de formatage, la structure est du type liste.

(16) *D précède E et D est près de E* $\Rightarrow$ *déclencheur introduit énumération*

$(D < E) \cap (D \sim E)$ *introducteur*

Cette règle explique que la proximité entre le déclencheur et l'énumération indique que les deux sont liés, ce qui peut remplacer l'introducteur. Il est possible de réécrire la règle (16) en spécifiant le nombre de lignes de texte maximum permis entre le déclencheur et l'énumération :

(17) $(D < E) \cap (\sum L < 6) \quad \Rightarrow$ *déclencheur introduit énumération*

S'il y a trop de texte entre le déclencheur et l'énumération (nous avons mis six lignes ici, mais il faudrait faire des évaluations systématiques de corpus pour fixer une limite précise), il faut expliciter le fait que ces deux éléments sont liés au moyen d'un introducteur dans le déclencheur.

Il n'est pas possible pour l'instant d'être beaucoup plus précis en ce qui concerne la nature des traits de formatage, parce que, comme l'a constaté Virbel 1985, il n'existe pas en général de conventions pour l'emploi des marques de mise en forme matérielle des textes. Certaines conventions se créent localement pour certains types de textes, mais il n'est globalement pas possible de faire des généralisations systématiquement vérifiées. Il s'agit plutôt de donner un contraste entre des unités de texte ayant des fonctions ou des statuts différents.

Nous examinons maintenant un type particulier de déclencheur qui est une forme réduite de la structure canonique.

4.3 Des formes réduites qui contiennent des mots WH

Les **mots WH** sont les mots anglais *who* (qui), *what* (quoi), *where* (où) et *when* (quand), qui sont des pronoms relatifs ou interrogatifs, selon leur emploi. Des déclencheurs qui contiennent des mots WH comme *where* et *who* sont un cas particulier des déclencheurs de la forme canonique réduite. En voici deux exemples :

(18) *Who Can Benefit*
Des gens qui peuvent profiter de

(19) *Where Books Have Been Purchased*
Où on a acheté des livres

Nous allons appeler des séquences de ce type des **propositions WH**. Des déclencheurs de ce type ressemblent aux propositions relatives qui modifient le classificateur sans que ce dernier soit explicite. Par exemple nous pouvons dire que l'exemple (18) est une forme réduite du déclencheur plus explicite suivant :

(20) *This is a list of people who can benefit*

où le mot WH *who* (qui) est un pronom relatif, et *people* (personnes) est son antécédent.

Les mots *who*, *where* et *when* ont tous la propriété de pouvoir prédire le type de l'antécédent (ici du classificateur) à partir du mot lui-même. Quand il

s'agit du pronom *who* (qui), son antécédent doit être *people* (personnes), ou un mot qui décrit un type de personne, par exemple *actors* (acteurs), *women* (femmes), *teenagers* (adolescents).

Pour le pronom relatif *where* (où), le classificateur sera évidemment toujours un type de lieu, par exemple *places* (lieux), *universities* (universités), *resorts* (stations estivales). Cette notion de typage basé sur le mot WH ou sur d'autres indices dans la séquence relie ce sujet à un certain aspect du travail sur l'analyse des questions (Hirschman et coll. 1999, Ferret et coll. 1999).

Puisque les informations structurelles du texte ne contiennent pas les caractéristiques des articles, le noyau est en général le plus petit élément pouvant jouer seul le rôle de déclencheur. Cependant, dans les déclencheurs (18) et (19), le classificateur n'est pas explicitement présent :

(21) *Who can benefit*
Classificateur = **People**

Dans de tels cas, l'élément minimal qui peut jouer le rôle de déclencheur est effectivement le noyau, avec le classificateur en moins. Cela ne nuit pas à la caractérisation des articles de l'énumération, parce que le classificateur peut se déduire implicitement du mot WH qui introduit la séquence réduite. C'est le cas seulement quand il s'agit des mots *who* (qui), *where* (où) et *when* (quand).

Cependant, il n'est pas toujours acceptable d'employer une **proposition WH** seule comme déclencheur, même quand il s'agit des mots WH *who*, *where* et *when*. Prenons quelques exemples :

(22) *Famous People Who have a First Name*
for a Last Name
Des célébrités qui ont comme nom un prénom

(23) **Who have a First Name for a Last Name*

(24) *Businesses who want to move into the 21st century*
Des entreprises qui veulent passer au 21e siècle

(25) **Who want to move into the 21st century*

La raison principale pour laquelle les formes réduites de ces deux déclencheurs ne sont pas acceptables concerne la syntaxe des chaînes. Les séquences (23) et (25) sont inacceptables grammaticalement. Par contre, les formes suivantes peuvent jouer seules le rôle de déclencheurs :

(26) *Who* **has** *a First Name for a Last Name*

(27) *Who* **wants** *to move into the 21st century*

Dans les **propositions WH** (26) et (27), les verbes sont à la troisième personne du singulier. Dans les séquences (23) et (25), les verbes sont à la troisième personne du pluriel, et on les trouve grammaticalement inacceptables sans leur classificateur respectif. Les séquences qui peuvent exister seules sans un classificateur ressemblent toutes aux questions indirectes. Considérons toutes les **propositions WH** observées jusqu'à maintenant qui sont acceptables seules comme déclencheurs :

(28) *You want to know* ***who can benefit****.*
Vous voulez savoir qui peut en profiter.

(29) *You want to know* ***where books have been purchased****.*
Vous voulez savoir où des livres ont été achetés.

(30) *You want to know* ***who has a first name***
for a last name.
Vous voulez savoir qui a comme nom un prénom.

(31) *You want to know* ***who wants to move into***
the 21st century.
Vous voulez savoir qui veut passer au 21e siècle.

Nous voyons que ces déclencheurs particuliers sont des propositions WH qui forment des questions indirectes, mais qu'ils ressemblent aussi aux propositions relatives. Une question indirecte exige que le verbe soit à la troisième personne du singulier. Dans un déclencheur, le classificateur est toujours au pluriel parce qu'il décrit une collection. Le verbe de la proposition relative qui modifie le classificateur est donc au pluriel, et cette proposition ne ressemble pas à une question indirecte. Cependant, dans certains cas, il n'y a pas de différence en anglais entre la forme du mot singulier et la forme du mot au pluriel. Ainsi, la proposition relative au pluriel et la question indirecte (au singulier) sont identiques. Par exemple, le verbe *can* (pouvoir) ne change pas au singulier et au pluriel (*He* ***can****/They* ***can*** (il peut/ils peuvent)) (exemple (18)). C'est dans ces cas que la proposition WH peut jouer seule le rôle de déclencheur.

Sur le plan sémantique, dans certains déclencheurs, le classificateur porte plus d'informations que l'on peut déduire à partir du mot WH. Par exemple, dans la séquence (22), le classificateur est *famous people* (célébrités). Le mot *who* suggère qu'il s'agit de personnes, mais il n'est pas possible de déduire les propriétés spécifiques de ces personnes (ici, le fait qu'ils sont célèbres). Donc, si un classificateur a des caractéristiques plus précises que celles données par le mot WH, il faut les expliciter.

Il existe encore une structure de déclencheur qui commence par un mot WH. Prenons l'exemple suivant :

(32) *Where to Purchase a Phone Card from the City Program*
Où acheter une carte téléphonique du projet City

Ceci n'est pas une proposition relative parce qu'en anglais il n'est pas possible d'y ajouter un classificateur de la même façon que dans l'exemple (20).

(33) **Places where to Purchase a Phone Card from the City Program*

L'exemple (33) n'est pas acceptable, parce que les propositions relatives en anglais contiennent plutôt un verbe à la forme active qu'un verbe à l'infinitif. Des grammaires de la langue anglaise (Leech et Svartvik 1974, Greenbaum et Quirk 1973) aident à clarifier cette construction. Il existe en anglais un type de question indirecte où la proposition qui est rapportée contient un infinitif avec *to* et commence par un mot WH (Leech et Svartvik 1974). Par exemple :

(34) *I asked him where to Purchase a Phone Card from the City Program.*
Je lui ai demandé où acheter une carte téléphonique du projet City.

(35) *I asked him where **I ought** to Purchase a Phone Card from the City Program.*
Je lui ai demandé où je devais acheter une carte téléphonique du projet City.

Ce type de structure est lié aux ordres rapportés. Un ordre de style indirect en anglais utilise une proposition subordonnée avec un infinitif en *to* (Greenbaum et Quirk 1973).

(36) *He told me where to Purchase a Phone Card from the City Program.*
Il m'a dit où acheter une carte téléphonique du projet City.

Les exemples (34), (35) et (36) sont liés par le fait qu'ils peuvent tous représenter la même situation, celle de trouver des renseignements sur des lieux où l'on peut acheter une carte de téléphone. Cette structure diffère des autres déclencheurs étudiés jusqu'à maintenant parce qu'elle ne contient pas de mention explicite du type des articles, et qu'elle devient fausse sur le plan grammatical si on en ajoute une.

Nous venons de voir un cas exceptionnel de formes réduites de la structure canonique. Pour certaines chaînes qui comprennent un mot WH *who, where* ou *when*, il arrive que le noyau sans son classificateur joue le rôle de déclencheur, parce que ce classificateur peut se déduire du mot WH. Nous allons maintenant formaliser les règles qui contrôlent ces formes réduites.

4.4 Formalisation de l'emploi des formes réduites

Dans les deux sections précédentes, nous avons vu que nous pouvons légitimement choisir d'omettre certains éléments d'un déclencheur de la forme canonique. Par exemple, pour introduire la même énumération, chacune des séquences suivantes peut s'utiliser :

(37) [*who can benefit*]$_{\textbf{N-C}}$

(38) [[***people***]$_{\textbf{C}}$ *who can benefit*]$_{\textbf{N}}$

(39) `[a list of]`$_{\text{O}}$ [[***people***]$_{\text{C}}$ *who can benefit*]$_{\text{N}}$

(40) [This is]$_{\text{I}}$ `[a list of ]`$_{0}$ [[***people***]$_{\text{C}}$
who can benefit]$_{\text{N}}$

Voici une liste de gens qui peuvent en profiter.

Cependant, tous ces éléments ne forment pas nécessairement un déclencheur acceptable. Par exemple, comme variante des exemples (37-40), on ne s'intéressera pas à l'exemple (38), qui est trop générique :

(41) *This is a list of people.*
Voici une liste de gens.

Le noyau est obligatoire parce que c'est lui seul qui précise les caractéristiques des articles. Il n'est pas nécessaire qu'un déclencheur explicite les quatre éléments (l'introducteur, l'organisateur, etc.), mais il doit toujours comprendre au moins le noyau. Nous avons donné un cas particulier (section 4.2) où le classificateur est facultatif. Dans cette section, nous allons préciser plus en détail les combinaisons d'éléments qui sont acceptables, en formulant des règles.

Les exemples (37), (38), (39) et (40) montrent qu'on peut utiliser plusieurs formes réduites du même déclencheur pour introduire la même énumération. Maintenant nous allons examiner cet exemple et quelques autres d'une manière plus systématique. Dans le tableau 5, nous avons décomposé trois déclencheurs :

a) [The following is]$_{I}$ [`a list of`]$_{O}$ [*publicly funded* [***universities***]$_{C}$ *in Vietnam*]$_{N}$
Voici une liste d'universités du Vietnam qui sont financées par l'État.

b) [Here is]$_{I}$ [`a list of`]$_{O}$ [[***lakes***]$_{C}$ *you can fish*]$_{N}$
Voici une liste de lacs où on peut pêcher.

c) [This is]$_{I}$ [`a list of`]$_{O}$ [[***people***]$_{C}$ *who can benefit*]$_{N}$
Voici une liste de gens qui peuvent en profiter.

Tableau 5
Tableau des éléments optionnels

a)	The following is	a list of	publicly funded **universities** in Vietnam		*I + O + C + N*
		A list of	publicly funded **universities** in Vietnam		*O + C + N*
			Publicly funded **universities** in Vietnam		*C + N*
	The following are		*publicly funded* **universities in Vietnam**		***I + C + N***
b)	Here is	a list of	**lakes** you can fish		*I + O + C + N*
		A list of	**lakes** you can fish		*O + C + N*
			Lakes you can fish		*C + N*
	Here are		*(some)* **lakes you can fish**		***I + C + N***
c)	This is	a list of	people	who can benefit	*I + O + C + N*
		A list of	people	who can benefit	*O + C + N*
			People	who can benefit	*C + N*
				Who can benefit	*N*
	This is	a list of		who can benefit	*I + O + N*
		A list of		who can benefit	*O + N*
	This is			who can benefit	*I + N*
	These are		*people*	*who can benefit*	***I + C + N***

I = Introducteur, *O* = Organisateur, *C* = Classificateur, *N* = Noyau moins classificateur

4.4.1 Le cas standard

Les séquences *a*) et *b*) ont des structures standard. Dans ce cas, le classificateur est enchâssé dans le noyau. Utilisant la notation employée dans le tableau 5, nous pouvons donner des règles pour les réductions du type dans les exemples *a*) et *b*) :

(42) *N* est obligatoire

(43) *N* entraîne *C (i.e. CN* est obligatoire*)*

(44) *CN* peut se combiner avec *I* ou *O* ou les deux (*IO*).

(45) L'ordre des éléments est toujours conservé :
I < *O* < *C* < *N*.

(46) Si *CN* se combinent avec I mais sans *O*, le nombre
du verbe de I doit devenir pluriel.

La seule combinaison qui présente un problème est *I + C + N*, comme dans l'exemple (44).

(47) ***This is*** ***a list of*** *publicly funded universities in Vietnam.* *IOCN*
This are *publicly funded* ***universities*** *in Vietnam.* *I CN*

I est l'introducteur et contient un verbe et une cataphore. Si la cataphore est un pronom démonstratif comme *this* (ceci), il s'accorde en nombre avec l'objet du verbe. Si on omet l'élément *O*, l'organisateur, le classificateur devient l'objet du verbe. Le classificateur est toujours au pluriel parce qu'il s'agit d'une collection (section 2.3.1). Donc le démonstratif doit être au pluriel, comme *these* (ceux-ci). Si le verbe est *to be* (être), il faut un changement du même type, parce que ce verbe s'accorde en nombre avec ses arguments.

4.4.2 Le cas particulier

L'exemple *c*) montre le cas particulier où le classificateur et le noyau peuvent fonctionner comme des éléments distincts (section 4.2). Des déclencheurs de ce type ont un comportement différent : ils contiennent une proposition relative, et le classificateur peut se déduire du pronom relatif *who* (qui) ou *where* (où). À l'encontre des exemples *a*) et *b*), dans des exemples comme *c*), il n'y a pas nécessairement un classificateur (*C*), bien que le noyau (*N*) soit toujours obligatoire. Ainsi, *N* **n'entraîne pas** *C*. La règle (40) n'est donc pas vérifiée en *c*). Cela veut dire que *N* peut être présent sans que *C* le soit.

4.4.3 Synthèse

Il est donc possible de formaliser les conditions pour qu'une forme réduite de la structure canonique joue le rôle d'un déclencheur. Ainsi, dans tout déclencheur :

1° *N* est obligatoire ;
2° *CN* peut se combiner avec *I* ou *O*, ou les deux (*IO*);
3° L'ordre des éléments est toujours conservé : $I < O < C < N$;
4° Si *CN* se combine avec *I* sans *O*, le nombre du verbe de *I* doit devenir pluriel.

Puis, pour le cas particulier où le mot WH *who* ou *where* modifie le classificateur :

5° *N* n'entraîne pas *C*.

Et pour tous les autres cas :

6° *N* entraîne *C* (*CN* est obligatoire).

Nous avons vu dans cette section des cas d'emploi des formes réduites et les raisons de leur utilisation. Nous avons alors donné toutes les combinaisons linguistiques acceptables, ainsi que des modèles de ces réductions.

Nous avons formulé dans les sections précédentes un modèle de la forme canonique d'un déclencheur, ainsi qu'un modèle d'un déclencheur en contexte. Ces modèles servent dans une application qui extrait et classe des entités nommées à partir du Web, et que nous allons décrire dans la section suivante.

5. L'acquisition et la classification des entités nommées candidates

Le terme **entités nommées** (EN) recouvre des noms propres, c'est-à-dire des noms de personnes, des noms de lieux, des noms d'organisations, etc. Puisqu'elles sont évolutives, il est toujours nécessaire de mettre à jour les bases de données qui les stockent. Il existe plusieurs applications qui traitent des EN, par exemple des systèmes de repérage (Illouz et coll. 1999) et de désambiguïsation (Wacholder et coll. 1997). Le système FUNES (Coates-Stephens 1993) utilise des articles courts de journaux afin d'acquérir des EN. Les corpus Web sont une autre source riche pour ce type de données (Crimmins et coll. 1999) parce qu'ils sont constamment mis à jour.

Pour notre approche, Jacquemin et Bush 2000 ont développé un outil d'extraction des entités nommées (EN) à partir du Web. Cet outil a pour but de

collecter des EN dans des énumérations sur le Web et de classer chacune de ces entités selon son type. En utilisant nos modèles des déclencheurs (section 5), nous avons créé un programme qui a pour but de réduire le taux d'erreur dans ce processus en identifiant des déclencheurs bien formés. De plus, nous avons développé un analyseur qui identifie les constituants dans le déclencheur de l'énumération afin de raffiner la classification des EN collectées.

5.1 Description du système d'extraction d'EN

Le système d'acquisition de EN décrit dans Jacquemin et Bush 2000 emploie des contextes définitoires tels que *list of universities* (liste d'universités) et *international organisations such as* (des organisations internationales telles que) pour trouver des énumérations de EN candidates à partir d'un moteur de recherche. Ces contextes définitoires font partie des déclencheurs des énumérations. Le mot dans la chaîne définitoire qui décrit le type des articles est le classificateur. Nous avons montré (section 2.3.1) qu'il existe une relation d'hyponymie entre le classificateur du déclencheur et les articles de l'énumération. Hearst 1992 a utilisé des patrons lexico-sémantiques du type *such as* (tels que) afin d'acquérir des exemples de la relation d'hyponymie, et ce système emploie la même méthode afin de trouver et classer des EN.

Les quatre chaînes définitoires suivantes sont employées par le système :

(48) list of **type de EN**	*list of universities* liste d'universités
the following **type de EN**	*the following politicians* lesw hommes politiques suivants
type de EN such as	*actors such as* des acteurs comme
such as **type de EN** as	*such cities as* des villes comme

L'outil analyse les pages Web renvoyées par le moteur de recherche à partir de requêtes avancées de type chaînes formées sur ces quatre types de séquences. Ensuite, on analyse la phrase qui contient cette chaîne, ainsi que l'énumération qui la suit par des techniques d'analyse superficielle. Plusieurs analyseurs identifient et séparent chaque article de cette énumération en utilisant les balises HTML, et les prennent comme des EN candidates. Le système type ces candidats selon la chaîne définitoire. Par exemple :

(49) *This is a **list of American companies** with business*
interests in Latvia
– AM-SARAS Inc.
– American Airlines
– American Express
Voici une liste de sociétés américaines qui ont des
intérêts commerciaux en Lettonie […].

Ici la chaîne définitoire est *list of American companies*, donc chaque EN candidate, *AM-SARAS Inc., American Airlines*, etc., sera typée comme une *American company*.

Ce système est assez efficace, mais il génère des erreurs. En particulier, quand la phrase qui contient la chaîne définitoire n'est pas un déclencheur, mais est suivie par une énumération qui n'y est pas liée, le système prend les articles de l'énumération en les classifiant selon le contenu de la chaîne définitoire.

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

Par exemple, on trouve la chaîne définitoire *list of American companies* dans la séquence (50), mais cette phrase ne joue pas le rôle de déclencheur. Si cette chaîne était suivie sur la page par une énumération, le système prendrait chaque article comme une société américaine. Afin de traiter ce problème, le premier module de notre programme identifie des chaînes qui jouent le rôle de déclencheur.

Le typage fait par le système d'acquisition de EN candidates (Jacquemin et Bush 2000) n'emploie que le classificateur du déclencheur qui fait partie de la chaîne définitoire. Le deuxième module de notre programme est donc un analyseur qui identifie le noyau d'un déclencheur, et donne ainsi une classification plus précise des EN candidates qui apparaissent dans une telle énumération. Nous décrirons ce programme dans les sections suivantes.

5.2 Programme de repérage et d'analyse des déclencheurs

Le programme est composé de deux modules principaux: le repérage des déclencheurs bien formés et l'analyse de ces déclencheurs pour identifier leurs constituants. La figure 5 montre cette structure.

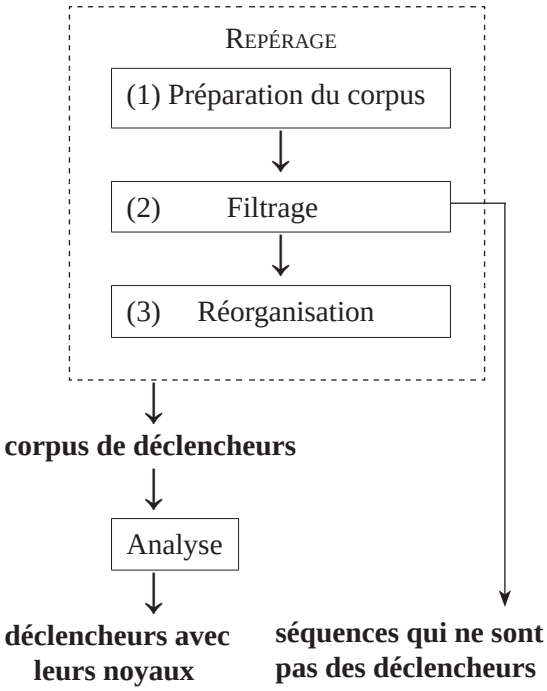


Fig. 5 : Le programme de repérage et d'analyse de déclencheurs

Le premier module est composé de trois sous-modules : la préparation du corpus, le filtrage et la réorganisation de données.

5.2.1 Module de repérage

Ce programme a été développé sur des séquences trouvées sur le Web par le requête *list of EN* de l'outil d'extraction de EN candidates (Jacquemin et Bush 2000). Cette extraction s'appuie sur le formatage et se fait conjointement à l'extraction des EN.

Tous les déclencheurs potentiels sont d'abord étiquetés au préalable par le TreeTagger (Schmid 1999). Le filtrage commence ensuite, fait par onze modules distincts écrits en PERL.

À partir du modèle syntaxique de la forme canonique des déclencheurs (section 5.1), nous avons formulé une expression régulière, construite avec des étiquettes, permettant d'extraire toutes les séquences similaires à cette structure : (*s = au singulier; pl = au pluriel*) .

(51) (*déterminant)? (adjectif*)* nom_s *préposition(adjectif)** *(nom-modifieur)** nom_{pl}

Ce qui est important dans ce patron, c'est la présence des deux noms dans la séquence. Le premier est l'organisateur qui décrit la structure de l'énumération. Comme il n'y a qu'une énumération pour chaque déclencheur, ce nom est toujours au singulier, par exemple *a list* (une liste). Le second nom est le classificateur. Il est toujours au pluriel parce qu'il s'agit de la collection des articles, par exemple : *universitie**s*** (universités), *politician**s*** (hommes politiques). La plupart des erreurs à cette étape sont dues à des erreurs dans l'étiquetage, comme lorsque des noms pluriels sont étiquetés comme des noms singuliers, ce qui provoque le rejet de la chaîne.

Les modules de filtrage suivants traitent des autres mots des séquences, c'est-à-dire ceux qui ne font pas partie de la chaîne de base formulée par la grammaire des déclencheurs canoniques. Ces modules emploient des expressions régulières basées sur les étiquettes ou les lexèmes particuliers, afin de filtrer des chaînes qui ne jouent pas le rôle de déclencheurs. Après ce filtrage, les chaînes qui restent sont triées et des formes identiques sont regroupées dans une seule forme.

5.2.2 L'analyseur

Le système d'acquisition de EN candidates à partir du Web (section 5.1) classe chaque EN trouvée selon le classificateur dans la chaîne de recherche employée (Jacquemin et Bush 2000). Par exemple, toute EN qui suit *list of universities* est classée comme une université. Mais le déclencheur fournit des informations plus précises sur les propriétés des articles en spécifiant des hyponymes du classificateur. Pour l'exploiter, il faut identifier le noyau du déclencheur en faisant une analyse automatique selon la structure mise en évidence dans la section 3.2. Le dernier module de ce programme utilise des expressions régulières afin d'identifier le classificateur et ses modifieurs. Le classificateur est spécifié par la chaîne de recherche utilisée par le système.

(52) `[List of]`$_{O}$ [[***countries***]$_{C}$ *with which Mexico has a free trade agreement.*]$_{N}$
– ***NOYAU:*** [[***countries***]$_{C}$ *with which Mexico has a free trade agreement*]$_{N}$
– ***CLASSIFICATEUR:*** [***countries***]$_{C}$
Une liste de pays avec lesquels le Mexique a un accord de libre-échange.

Cela permet un raffinement de la classification des EN candidates. Par exemple, tous les articles d'une liste qui suivent le déclencheur en (52) pourraient

désormais être classés comme *countries*, mais avec la précision supplémentaire qu'ils sont des *countries with which Mexico has a free trade agreement.*

5.3 Évaluation

L'évaluation suivante des résultats montre que les sorties de ce programme sont assez correctes. Concernant le module de repérage, les sorties de chaque sous-module ont été évaluées manuellement. Pour calculer l'efficacité globale du module, un échantillon aléatoire d'une centaine de séquences a été vérifié. Globalement, la précision est de 76 % et le rappel est de 87 %.

Tableau 6
Exemples des déclencheurs décomposés par l'analyseur et évaluation de la qualité de l'analyse

a.	**List of universities that censored/banned the newsgroup.**		
	N:	*universities that censored/banned the newsgroup*	correct
	C:	*universities*	correct
b.	**A list of countries controlled for national security purposes under Group D :**		
	N:	*countries controlled for national security purposes under Group D*	correct
	C:	*countries*	correct
c.	**A list of lakes known to have exotic plants , including a map**		
	N:	*lakes known to have exotic plants*	correct
	C:	*lakes*	correct
d.	**Below find a list of international organizations of and for women in science.**		
	N:	*international organizations of and for women in science*	correct
	C:	*international organizations*	correct
e.	**Department's list of terrorist groups issued in April 1993.**		
	N:	*terrorist groups issued in April 1993*	trop long
	C:	*terrorist groups*	correct
f.	**List of Universities and Colleges.**		
	N:	*Universities and Colleges*	correct
	C:	*Universities*	trop court
g.	**Following is a list of lakes found in Tillamook County.**		
	N:	*lakes found in Tillamook County*	correct
	C:	*lakes*	correct
h.	**Following is a list of regions accompanied with a map.**		
	N:	*regions accompanied with a map*	trop long
	C:	*regions*	correct

N = Noyau, C = Classificateur

Le tableau 6 montre quelques déclencheurs avec le résultat donné par l'analyseur. Le déclencheur est suivi par son noyau et son classificateur, avec une indication sur la qualité de l'analyse. Il y a trois sorties possibles : soit le constituant identifié est exact, soit l'analyseur identifie un constituant qui est plus long que le noyau ou que le classificateur exact, soit l'analyseur identifie un constituant qui est plus court que le constituant exact. L'évaluation de ces résultats a été faite manuellement par un seul juge.

L'identification du couple noyau/classificateur est basée sur un patron construit avec des étiquettes et nous observons que l'analyse est correcte avec une précision de 82 %. En formulant un modèle des déclencheurs (section 3.2), nous avons considéré la forme canonique en isolation. Il arrive cependant que cette structure canonique soit employée dans une phrase entourée par des éléments qui ont une autre fonction, en plus de celles d'introduire et de caractériser une énumération. Dans ce cas, notre modèle syntaxique n'est pas capable d'identifier la fin du noyau. Par exemple les deux déclencheurs (g) et (h) dans le tableau 6 ont la même structure syntaxique :

(53) *g.* [*Following is*]$_{I}$ [`a list of`]$_{O}$ [[*lakes*]$_{C}$ *found in Tillamook County*.]$_{N}$
Voici une liste de lacs dans le comté de Tillamook.

(54) *h.* [*Following is*]$_{I}$ [`a list of`]$_{O}$ [[*regions*]$_{C}$]$_{N}$
accompanied with a map.
Voici une liste de régions accompagnée d'une carte.
Following is a list of **type de EN**
modifieur (participe passé +)

Toutefois, les modifieurs ont des fonctions différentes. En (*g*), le modifieur modifie le classificateur, donc il fait partie du noyau *lakes found in Tillamook*. Par contre, en (*h*) le modifieur est lié à l'organisateur - c'est la liste qui est *accompanied with a map* et non pas le classificateur. Dans cet exemple, le noyau n'est donc que *regions*, et l'analyse faite par le système n'est pas correcte. Cette ambiguïté syntaxique est très difficile à résoudre automatiquement. C'est pourquoi la plupart des problèmes sont posés par des exemples de ce type.

L'autre problème que nous n'avons pas considéré dans nos modèles concerne des classificateurs coordonnés. Dans tous les exemples analysés jusqu'à présent, chaque déclencheur n'a eu qu'un classificateur. Cependant, si nous examinons l'exemple (*f*) dans le tableau 6, nous voyons qu'il arrive qu'un déclencheur ait deux classificateurs qui sont coordonnés :

(55) [`list of`]$_{O}$ [[***Universities and Colleges***]$_{C}$]$_{N}$
liste d'universités et de collèges

Dans cette séquence, le classificateur est *universities and colleges*, mais l'analyseur ne repère que *universities* comme classificateur. Pour corriger cette source d'erreurs il va falloir reconsidérer nos modèles et modifier le programme afin de traiter ce type de coordination.

Des erreurs dans le module de repérage des déclencheurs bien formés posent des problèmes pour l'analyseur. Des séquences qui ne jouent pas le rôle de déclencheurs ne sont pas analysables de la même manière que les vrais déclencheurs, et ces séquences erronées sont souvent mal analysées par l'analyseur.

Pour cette évaluation, toutes les mesures ont été calculées sur les séquences en dehors de leur contexte. Cela veut dire qu'une séquence est acceptée comme un *bon* déclencheur si elle peut jouer ce rôle. Nous n'avons pas encore vérifié si elles introduisent effectivement des énumérations. Les possibilités du système restent donc limitées, ne permettant d'identifier que les déclencheurs de la forme canonique. Par contre, le niveau de précision de 76 % est très élevé pour une identification de cette forme.

6. Conclusion

Dans ce document, nous avons montré que les séquences qui introduisent des énumérations d'entités nommées partagent certains traits. Nous avons décrit et analysé leurs propriétés sémantiques ainsi que leurs structures syntaxiques, en identifiant une forme canonique de déclencheur. À partir de ces analyses, nous avons proposé quelques modèles initiaux de la structure sémantique et syntaxique.

Ce travail a contribué à l'élaboration d'un système d'acquisition et de classification des entités nommées à partir du Web. Nous avons écrit un programme de base qui identifie des déclencheurs bien formés, ce qui a pour but de réduire le taux d'erreur dans ce système. De plus, nous avons réalisé un analyseur qui décompose des déclencheurs en éléments sémantiques. Cela sert à donner des précisions sur les articles des énumérations qui y sont liées.

Nous allons poursuivre ce travail en améliorant ces programmes afin de les rendre plus efficaces. Nous devrons ensuite étudier le rôle des ancres (ou hyperliens) dans les documents hypertextuels, un sujet qui est abordé chez Amitay 1999. Dans l'état actuel de nos travaux, la plupart de ces séquences ne sont pas acceptées comme des déclencheurs. Elles peuvent toutefois introduire des énumérations de la même manière que les séquences étudiées ici, mais les énumérations associées se trouvent dans un lieu accessible par l'hyperlien et non dans le cotexte.

Le thème des ancres qui introduisent des listes permettra d'exploiter des balises HTML qui sont une caractéristique importante des documents sur le Web. Pour le corpus de déclencheurs étudié dans cet article, ce balisage a été supprimé, donc les modèles n'en prennent par compte. Les balises sont cependant indispensables pour l'extraction des EN par l'outil de Jacquemin et Bush, et c'est une des raisons pour lesquelles le Web est si valable comme source d'informations.

Références

AMITAY, E. 1997 *Hypertext: The Importance of being Different,* Mémoire de MSc, Université d'Edinbourg, Centre for Cognitive Science.

AMITAY, E. 1999 «Anchors in context: A corpus analysis of web pages authoring conventions», dans L. Pemberton, S. Shurville et coll., *Words on the Web – Computer Mediated Communication,* Intellect Books, Exeter, Royaume-Uni, p. 192.

BORILLO, A. 1995 «Exploration automatisée de textes de spécialité: repérage et identification de la relation lexicale d'hyperonymie», LINX 31:113-124.

BUSH, C. 2000 *Analyse des déclencheurs des énumérations d'entités nommées sur le Web,* Rapport Technique 5, LIMSI, Orsay, France.

CATACH, N. 1994 *La Ponctuation,* Paris, Presses Universitaires de France, coll. *Que sais-je?*

COATES-STEPHENS, S. 1993 «The analysis and acquisition of proper names for the understanding of free text», *Computers and the Humanities* 26 : 441-456.

CRIMMINS, F., A.F. SMEATON, T. DKAKI et J. MOTHE 1999 «Tétrafusion: Information discovery on the internet», *IEEE Intelligent Systems and Their Applications* 14-4:55-62.

FERRARI, S. 1997 *Méthode et outils informatiques pour le traitement des métaphores dans les documents écrits,* thèse de doctorat, Université de Paris XI, Orsay, Notes et documents LIMSI N° 97-30.

FERRET, O., B. GRAU, G. ILLOUZ, C. JACQUEMIN et N. MASSON 1999 «QALC – the question-answering program of the language and cognition group at LIMSI-CNRS», dans E.M. Voorhees, D.K. Harman et coll., *Proceedings of the 8th Text REtrieval Conference (TREC 8),* Gaithersburg, Maryland, National Institute of Standards and Technology (NIST), p. 465-474.

GEZUNDHAJT, H. 1999 *Principes généraux de la linguistique énonciative,* SELF, Université de Toronto.

GREENBAUM, S. et R. QUIRK 1973 *A University Grammar of English,* Londres, Longman.

HEARST, M. 1992 «Automatic acquisition of hyponyms from large text corpora», dans *Proceedings of COLING-92,* France, Nantes, p. 539-545.

HEARST, M. 1998 «Automated discovery of wordnet relations», dans Christiane Fellbaum et coll., *Wordnet, an electronic lexical database,* Cambridge (Mass.), MIT Press, p. 131-151.

HIRSCHMAN, L., M. LIGHT, E. BRECK et J. BURGER 1999 «Deep Read: A reading comprehension system», dans *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Université du Maryland, p. 325-332.

HUNTER, L. 1998 «Text nouveau: Visible structure in text presentation», *Computer Assisted Language Learning* 11-4, Pays-Bas, Lisse, Swets & Zeitlinger, p. 363-379.

ILLOUZ, G., C. JACQUEMIN et B. HABERT 1999 *Repérage des entités nommées (REN) dans des transcriptions de documents parole*, Rapport Technique 18, LIMSI, France, Orsay.

JACQUEMIN, C. et C. BUSH 2000 «Fouille de Web pour la collecte d'Entités Nommées», dans TALN 2000, *Actes de la 7e conférence annuelle sur la Traitement Automatique des Langues Naturelles*, ATALA, Lausanne, p. 187-196.

LEECH, G. 1977 *Semantics*, Londres, Penguin Books.

LEECH, G. et J. SVARTVIK 1974 *A Communicative Grammar of English*, Londres, Longman.

LUC, C., C. GARCIA-DEBANC, M. MOJAHID, M-P. PÉRY-WOODLEY et J. VIRBEL 1999 «A linguistic approach to some parameters of layout: A study of enumerations», dans *The AAAI Fall symposium Technical Report*, Mass., North Falmouth, p. 35-44.

LUC, C., M. MOJAHID, M-P. PÉRY-WOODLEY et J. VIRBEL 2000 «Les énumérations : structures visuelles, syntaxiques et rhétoriques», Soumis à CIDE 2000.

MAINGUENEAU, D. 1996 *Les termes clés de l'analyse du discours*, Paris, Seuil.

PASCUAL, E. 1991 *Représentation de l'architecture textuelle et génération de texte*, thèse de doctorat, Université Paul Sabatier, Toulouse III, France.

PECK, FRANCES 1996 *Hypergrammar*, University d'Ottawa, http://www.uottawa.ca/academic/arts/writcent/hypergrammar.

PÉRY-WOODLEY, M-P. 1998 «Textual signalling in written text: a corpus based approach», dans P*roceedings of the Workshop "Discourse Relations and Discourse Markers"*,COLING-98, Université de Montréal.

SCHMID, H. 1999 «Improvements in part-of-speech tagging with an application to German», dans S. Armstrong, K.W. Church, P. Isabelle, S. Manzi, E. Tzoukermann, D. Yarowski et coll., *Natural Language Processing Using Very Large Corpora*, Dordrecht, Kluwer.

VIRBEL, J. 1985 «Langage et métalangage dans le texte du point du vue de l'édition en informatique textuelle», dans *Cahiers de Grammaire* 10:5-72.

VIRBEL, J. 1989 «The contribution of linguistic knowledge to the interpretation of text structures», dans J. André, V. Quint, R. Furtura et coll., *Structured Documents*., Cambridge (Mass.), MIT Press, p. 161-181.

WACHOLDER, N., Y. RAVIN et M. CHOI 1997 «Disambiguation of names in text», dans *Proceedings of the 5th Conference on Applied Natural Language Processing*, ANLP, Washington, p. 202-208.