Un outil d'extraction terminologique endogène et multilingue

Jacques Vergne

GREYC - UMR 6072
campus II - BP 5186
Université de Caen
F-14032 CAEN cedex
FRANCE
www.info.unicaen.fr/~jvergne
e-mail : Jacques.Vergne@info.unicaen.fr
tél. : 02 31 56 73 36

Résumé – Abstract

Dans cet article, nous présentons un outil d'extraction terminologique "endogène" à partir d'un corpus multilingue. Cet outil est qualifié d'endogène car, sans autre ressource que le corpus dont il doit extraire les termes, il calcule les mots vides à partir de ce corpus pour centrer les termes candidats sur des mots pleins. Il est placé dans le cadre d'un système de constitution automatique de revue de presse à partir de sites de presse présents sur l'internet[1]. Il s'agit de répondre à des questions telles que : "de qui, de quoi est-il question aujourd'hui dans la presse de tel espace géographique ou linguistique ?". Le corpus est constitué des textes des hyperliens des "Unes" des sites de presse de langues inconnues a priori. Il est renouvelé quotidiennement, et sa taille est d'environ 100 Ko (débalisé). La méthode est fondée sur l'analyse distributionnelle, et utilise des différences entre mots contigus : les différences de longueur et d'effectif.

In this paper, we present an "endogenous" terminology mining tool, from a multilingual corpus. This tool is described as endogenous because, without any other resource than the

[1] Une démonstration est accessible sur :
`www.info.unicaen.fr/~jvergne/demoRevueDePresse/index.html`

corpus from which it has to extract terms, it computes function words from this corpus to focus candidate terms on content terms. It is used inside an automatic news review system from news web sites. The system is able to answer questions as : "who, what are newspapers speaking about today in a given geographic or linguistic search space?". The corpus is made of hyperlinks texts of news web site front-pages in unknown languages. It is daily downloaded, and its size is about 100 Kbytes (untagged). The method is based on distributional analysis, and uses differences between contiguous words : differences of length and of frequency.

Mots Clés – Keywords

extraction terminologique, endogène, multilingue, internet, fouille de texte.

terminology mining, endogenous, multilingual, internet, web mining, text mining.

1 Introduction

Le présent travail s'inscrit à l'intersection des TALN et du "web mining" car il applique des concepts issus de l'informatique linguistique à la problématique du web mining. Ce domaine de recherches mixte se développe actuellement, à l'exemple du groupe CLAIR : Computational Linguistics And Information Retrieval group at the University of Michigan (`perun.si.umich.edu/clair/`) dont les thèmes de recherches principaux concernent les systèmes questions - réponses et le résumé automatique. Cette problématique impose des contraintes qui lui sont propres : multilinguisme généralisé, production quotidienne et massive d'informations surtout textuelles, et besoins nouveaux des utilisateurs d'accès toujours plus efficace et plus précis à ces informations. Cette problématique nous donne la possibilité de concevoir de nouvelles tâches qui constituent des enjeux très intéressants pour les TALN. Nous allons d'abord présenter l'application cadre (section 2), puis les spécifications de l'outil (section 3). Puis nous proposons une manière originale de poser le problème comme un calcul de différences entre mots (section 4). Ensuite une solution est proposée et décrite (section 5). Nous présentons enfin des résultats et leur évaluation (section 6) puis une discussion sur la méthode (section 7).

2 L'application cadre

L'application cadre est un système de constitution automatique de revue de presse à partir de sites de presse présents sur l'internet, pour des utilisateurs qui se demandent de qui et de quoi il est question aujourd'hui dans la presse de tel espace géographique ou linguistique. Ce système inverse la problématique des moteurs de recherche : au lieu de rechercher des

documents à partir de mots-clés qui représentent des thèmes, il s'agit de produire **en sortie** les thèmes principaux de l'actualité, et de donner accès aux articles concernés par ces thèmes (à la manière de Google News : `news.google.fr`). Pour chaque site, **un seul document** est téléchargé : le document du point d'entrée de chaque site de presse, c'est-à-dire sa "Une". De ce document, sont extraits les hyperliens : les URL et le code source des "textes" de liens. Ces codes source de "texte" de liens sont composés de titres ou de résumés d'articles (avec leur mise en forme), et d'URL vers des images, des photographies ou des icones. C'est de ces textes de liens (leur code source débalisé) que sont extraits les termes candidats. Ne sont retenus comme termes que les termes candidats **présents sur plusieurs sites**. Les URL des articles ne servent qu'en sortie, pour donner accès à un article, si l'utilisateur le décide. Le système ne se sert pas des articles eux-mêmes. Cette économie de traitement s'appuie sur le fait que la rédaction d'un texte de lien est un choix éditorial des journalistes des sites de presse. Le système calcule un graphe de termes dans lequel les nœuds sont les termes et les arcs sont les relations entre termes, relations définies par la co-occurrence de deux termes dans un même texte de lien. L'utilisateur peut naviguer dans ce graphe pour accéder à des termes liés et à des articles (à la manière de Kartoo : `www.kartoo.com`).

3 Spécifications de l'outil

Les spécifications de l'outil viennent de l'application cadre : il s'agit d'extraire des termes de petits corpus multilingues : les **corpus des textes d'hyperliens** collectés quotidiennement sur les Unes de sites de presse. Ces corpus sont thématiquement variés, et relativement petits (environ 80 à 170 Ko, 15000 à 30000 mots). Les langues sont alphabétiques, inconnues a priori, mélangées dans le corpus, et non diagnostiquées dans les calculs. Les calculs sont indépendants des langues, et donc insensibles à l'ajout d'une nouvelle langue et aux proportions entre langues différentes. Les tests ont été faits sur des corpus comprenant surtout du français, de l'anglais, de l'allemand, de l'italien, et de l'espagnol (un nouveau corpus acquis chaque jour). La tâche revient à distinguer les mots vides des mots pleins[2], pour centrer la construction des termes candidats sur les mots pleins. La méthode ne doit pas utiliser de ressources propres à une langue, pour ne pas avoir à faire un travail de préparation de ressources linguistiques à chaque nouvelle langue traitée (ouverture dans chaque langue, et ouverture à d'autres langues).

---

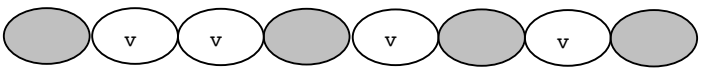
[2] Nous avons choisi dans cet article les termes : "mot vide" - "mot plein", synonymes de "mot grammatical" - "mot lexical" et de "function word" - "content word", à la suite de Lucien Tesnière et de Fathi Debili, et aussi dans la tradition de l'informatique documentaire, où un "mot vide" est un mot qui ne doit pas être indexé, qu'il soit mot grammatical ou mot lexical non discriminant. Les mots vides sont alors souvent regroupés dans une "stoplist" ou "ante-dictionnaire".

4 Comment poser le problème

Plusieurs méthodes ont été explorées, en nous imposant la contrainte de trouver une méthode n'utilisant aucune autre ressource linguistique que le corpus traité lui-même (méthode appelée pour cela "endogène") : la recherche des motifs répétés par l'algorithme glouton (recherche des n-grammes à partir des n-1-grammes) a été expérimentée, en excluant les mots vides par leur fréquence (test de Zipf). Les résultats étaient corrects, mais le départage entre les mots vides rares et les mots pleins très fréquents n'était pas possible (tels que *guerre*, *war*). La fréquence des mots n'est pas un indice suffisant : il faut s'intéresser aux fréquences des mots **à certaines positions**. D'où une autre manière de poser le problème : comment distinguer les mots pleins des mots vides à partir du corpus ? Notre direction de travail est d'utiliser à la fois les formes et leurs positions, ou plus abstraitement le concept de différence (ou de valeur relative), une constante du Thème Syntaxe et Rhétorique du GREYC, dont Hervé Déjean (Déjean, 1998), Nadine Lucas (Lucas, 2001), alors que le test de Zipf n'utilise que les fréquences des formes, sans exploiter leurs positions relatives. Or une observation fondamentale de Zipf lui-même est que les mots vides sont fréquents et courts et que les mots pleins sont plus rares et plus longs (ce qui est d'usage fréquent est court : c'est la loi de l'économie d'effort dans l'usage d'un code, caractérisée par Zipf (Zipf, 1949), et observable aussi dans les langages de programmation). D'où l'idée d'utiliser les différences entre mots[3] : non plus seulement les différences de fréquence, mais aussi les différences de longueur. Nous allons donc considérer le texte comme une suite de mots vides et de mots pleins, plus précisément comme une suite de mots vides et de mots non vides, car tout mot non vide est considéré comme plein. Voici un segment de texte (La Stampa du 15 mars 2003) :

Manifestazioni per la pace in tutto il mondo

L'outil donne **en sortie** le résultat suivant :



Manifestazioni** per la **pace** in **tutto** il **mondo

où chaque mot est symbolisé par un ovale blanc pour les mots vides, et un ovale noir pour les mots non vides. Notre problème devient alors un problème de détection de frontière. Les frontières sont caractérisées par une **différence orientée** sur l'axe syntagmatique : noir - blanc, ou blanc - noir. On en vient donc plus précisément à une détection de différence(s). Voici les critères de différence entre 2 mots contigus :

---

[3] En application du principe bien connu de Saussure :"dans la langue il n'y a que des différences." (Saussure, CLG, p.166).

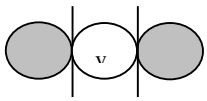
- différence de longueur en nombre de lettres : ***Manifestazioni*** *per* (14 lettres - 3 lettres)
- différence d'effectif dans le corpus : *il* ***mondo*** (19 occurrences - 3 occurrences)

Nous allons utiliser plusieurs différences sur plusieurs mots contigus, ce qui revient à étudier conjointement les dérivées des fonctions "longueur des mots" et "effectif des mots" selon leur position sur l'axe syntagmatique.

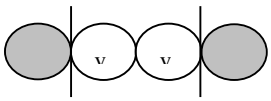
5 Solution proposée

5.1 Principes et propriétés

Sans autre donnée que les graphies elles-mêmes, nous allons détecter deux types de séquences de mots, où un ou deux mots vides sont encadrés par deux mot non vides (d'où un couple de 2 frontières), ce que nous illustrons par deux exemples de séquences à détecter :



tutto *il* ***mondo***



Manifestazioni *per la* ***pace***

profil des longueurs :	long - court - long	long - court - court - long
profil des effectifs :	rare - fréquent - rare	rare - fréquent - fréquent - rare

On détecte la séquence `PvP` par son profil caractéristique long - court - long **et** rare - fréquent - rare sur les fonctions longueur et effectif, et la séquence `PvvP` par son profil caractéristique long - court - court - long **et** rare - fréquent - fréquent - rare. Les deux types de séquence sont détectés sur des profils convergents des deux fonctions longueur et effectif (& booléen sur les deux critères). L'utilisation de deux critères améliore la robustesse des calculs, mais chacun pris séparément donne déjà de bons résultats. Le plus souvent, les résultats selon les deux critères sont convergents, ce qui corrobore la propriété énoncée par Zipf.

5.2 Processus et algorithme

Le processus comporte 2 étapes : étudier le corpus pour en extraire les mots vides, et générer les termes candidats.

5.2.1 Étudier le corpus pour en extraire les mots vides

Le corpus est segmenté sur les limites de textes de liens et sur les ponctuations, d'où des segments physiques délimités des ponctuations, que nous nommons "virgulots". Pour chaque virgulot, on passe les 2 filtres selon les 2 séquences.

Pour détecter une séquence `PvP`, on vérifie les profils long - court - long **et** rare - fréquent - rare, c'est-à-dire que la **longueur** du mot central est **inférieure** aux longueurs des 2 mots qui l'entourent, **et** que son **effectif** est **supérieur** à leurs effectifs. Pour détecter une séquence `PvvP`, on vérifie de la même manière les profils long - court - court - long **et** rare - fréquent - fréquent - rare. Voici un exemple de détection des deux séquences dans le même virgulot :

	Manifestazioni	*per*	*la*	***pace***	*in*	***tutto***	*il*	***mondo***
longueurs	14	3	2	4	2	5	2	5
profils	long	court	court	long		long	court	long
effectifs	1	10	207	2	62	3	19	3
profils	rare	fréquent	fréquent	rare		rare	fréquent	rare
déductions		mot vide	mot vide				mot vide	

Les mots sont catégorisés mots vides sous les deux formes graphiques : tout en minuscules, et avec une initiale majuscule. On trouve en section 6 les mots vides les plus fréquents extraits le 15 mars 2003 des deux espaces de recherche servant aux tests.

5.2.2 Générer les termes candidats

Les mots non vides sont considérés comme des mots pleins et peuvent devenir terme candidat. Les termes candidats sont extraits de chaque virgulot, dont les mots sont étiquetés vides ou non vides, pour générer des termes selon les motifs suivants :

`P+`	***Manifestazioni*** ***pace*** ***tutto*** ***mondo***
`P+ v+P+`	***Manifestazioni*** *per la* ***pace*** ***pace*** *in* ***tutto*** ***tutto*** *il* ***mondo***
`P+ v+P+ v+P+`	***Manifestazioni*** *per la* ***pace*** *in* ***tutto*** ***pace*** *in* ***tutto*** *il* ***mondo***

6 Résultats et évaluation

Voici les résultats des traitements des corpus téléchargés le 15 mars 2003 à partir des deux espaces de recherche servant aux tests, avec les termes candidats les plus fréquents (avant la suppression des termes présents sur un seul site) :

	espace de recherche 1	**espace de recherche 2**
espace de recherche	22 sites de la presse française nationale et régionale, 17 sites de la presse européenne (Suisse, Belgique, Allemagne, Italie, Espagne, UK, Irlande), et 4 sites de presse nord-américaine, chaque langue étant représentée par au moins deux sites	une centaine de sites publiés par Google News, environ la moitié étant des sites nord-américains, le reste du monde entier (`news.google.fr/news/`)

corpus	84 Ko, 14 800 mots				163 Ko, 28 500 mots			
termes candidats	1566 occurrences de 584 termes candidats (de 42 à 2 occ. / terme)				2435 occurrences de 820 termes candidats (de 47 à 2 occ. / terme)			
termes candidats les plus fréquents	article : 42 guerre : 21 Jean-Luc Lagardère : 17 monde : 12 Açores : 11	Weitere Artikel : 10 mort : 10 Bagdad : 8 empire : 8 semaine : 8 Lettre : 7	Plan : 7 fin : 7 guerra : 7 procès : 7 réforme : 7 sommet : 7 Echos : 6		Läs mer: 47 ÉÑ Ä : 29 Laden : 24 war : 22 Kabul : 20 Qaeda : 20 China : 18	Statement : 17 Sep 12 : 15 Pak : 14 Press Secretary : 13	Sep 11 : 13 Northern Alliance: 12 guerra : 12 Irak : 11 Kandahar : 11	
mots vides les plus fréquents	de : 340 la : 207 l' : 153 le : 113 d' : 107 à : 107	du : 103 et : 99 des : 88 en : 87 les : 84 a : 82	un : 80 Le : 74 La : 72 L' : 62 in : 62 une : 56	Les : 55 's : 55 to : 53 pour: 43 au : 41 sur : 41	to : 327 in : 280 of : 237 the : 230 's : 166 de : 154	for : 144 on : 143 and: 138 a : 126 The:118 en : 76	la : 75 by : 55 Al : 53 with: 52 is : 41 A : 38	from: 36 at : 34 i : 34 't : 32 un : 31 à : 31
termes candidats extraits à tort : bruit causé par un silence de la détection des mots vides	Was : 5 Tutti : 4 vous : 3 About: 2 Alors : 2 Ein : 2 Have : 2	If : 2 Mais : 2 Qu' : 2 Wie : 2 Wo : 2 avant : 2 contra: 2	could : 2 depuis:2 encore:2 faut : 2 mieux: 2 nous : 2 now : 2	plusieurs : 2 that : 2 tout : 2 tutto : 2	This: 12 How : 7 Don' : 6 It : 6 Most : 4 contra: 4	won' : 4 Alla : 3 My : 3 auf : 3 One : 2 Wer : 2	Where:2 Why : 2 après : 2 down : 2 einer : 2	enough: 2 only : 2 they : 2 when : 2 which: 2
	25/584 = 4,3% des 584 termes candidats extraits				22/820 = 2,7% des 820 termes candidats extraits (résultats sous-évalués à cause de quelques langues inconnues)			
termes candidats non extraits : silence causé par un bruit de la détection des mots vides	War : 9 paix : 7 soir : 7 war : 7 aide : 4	dimanche: 4 Photo : 3 baisse : 3 Aide : 2 Groupe : 2	attendu : 2 home : 2 turn : 2 voie : 2 world : 2		News : 77 New: 43 news : 23 killed : 18 Home : 17	Help : 16 Free : 10 Global : 9 Air : 8 help : 8	make : 8 First : 7 Get : 7 get : 7 groups : 7 ...	
	15/584 = 2,6% des 584 termes candidats extraits				88/820 = 10,7% des 820 termes candidats extraits			
termes retenus les plus fréquents (nb de sites - nb d'articles)	guerre (12-24) Lagardère (11-16) Jean-Luc Lagardère (9-12) monde (8-13) 15 (7-10) 16 (7-9) Aznar (7-8) Açores (7-10) empire (7-8)		semaine (7-8) Chirac (6-6) Premier ministre (6-7) fin (6-9) français (6-9) mort (6-10) pays (6-10) site (6-8) sommet (6-6) ...		Policy (19-23) U.S. (18-39) China (14-29) war (14-71) Special (12-24) This (12-24) United (12-18) Privacy Policy (11-11) Week (11-14)		East (10-12) American (9-14) Information (9-13) Press (9-25) Saddam (9-13) Azores (8-8) How (8-10) Index (8-8) Middle East (8-8) Money (8-8) ...	

N.B. au sujet de l'évaluation : comme les termes candidats sont construits sur les mots non vides, silence et bruit sur la détection des mots vides entraînent respectivement bruit et silence sur la génération des termes candidats.

Revenons à un de nos objectifs de départ : les mots vides rares et les mots pleins très fréquents sont-ils correctement repérés ? La méthode proposée, parce qu'elle utilise un calcul fondé sur des différences entre mots et non des valeurs absolues, rend la détection des mots vides presque indépendante de leur effectif. Par exemple, les mots *article* (42 occurrences), *guerre* (21), *monde* (12), *mort* (10), *guerra* (7) sont très fréquents et correctement détectés mots pleins, alors que *della* (6), *sous* (5), *bei* (4), *our* (3), *eines* (2), *Vers* (1) sont peu fréquents et correctement détectés mots vides : il suffit qu'un seul contexte ait pu les détecter au moyen des différences adéquates.

7 Discussion

L'idée d'utiliser le matériau linguistique traité lui-même pour en extraire des ressources nécessaires à ce traitement est déjà ancienne : on la trouvait dès 1982 chez Fathi Debili (Debili, 1982), puis chez Didier Bourigault dans LEXTER (Bourigault, 1994, pp. 63-78) et dans SYNTEX (Bourigault, 2002), et aussi chez François Rousselot (Frath, Oueslati, Rousselot, 2000). Cependant, il faut préciser quelles ressources sont extraites, et pour réaliser quel traitement : Bourigault et Rousselot utilisent des régularités distributionnelles lexicales pour extraire des termes candidats d'un corpus, et calculer les rattachements des groupes prépositionnels, à partir d'un vaste corpus monolingue et très cohérent thématiquement. Dans le travail ici présenté, nous utilisons aussi des régularités distributionnelles, mais la tâche est différente : il s'agit plus simplement de discriminer les mots vides pour construire les termes candidats à partir des mots pleins. Les prétraitements sont aussi différents : Debili et Bourigault font d'abord une analyse syntaxique, Rousselot et Helena Ahonen (Ahonen-Myka, 1999) recherchent les motifs répétés (algorithmes extrapolés de l'algorithme glouton) et mettent en entrée les mots vides pour éviter de les prendre comme termes (stopword-list), ce qui dans les deux cas nécessite la connaissance de la langue unique traitée, et la constitution manuelle des ressources propres à cette langue. Mais l'outil présenté ne nécessite **aucun prétraitement** ni **aucune ressource** préalable. Il traite un **corpus multilingue**, c'est-à-dire où les langues sont mélangées, inconnues, et qui ne sont à aucun moment diagnostiquées. Cet outil se situe dans la lignée des travaux d'Hervé Déjean (Déjean, 1998), qui a proposé une méthode de "découverte des structures formelles des langues", sur corpus bruts monolingues, de langues très variées, sans prétraitement ni ressources préalables.

Nous avons ainsi repris le terme "**endogène**" proposé par Didier Bourigault, dans le même sens générique, mais dans un sens spécifique différent.

8 Conclusion et perspectives

Nous avons présenté un outil d'extraction terminologique utilisant une méthode générique sur la dimension des langues, n'utilisant pas d'analyse syntaxique, ni de dictionnaire, ni de stoplist et capable de repérer à la fois les mots vides rares et les mots pleins très fréquents, dans un corpus multilingue, de langues alphabétiques, inconnues a priori, mélangées dans le corpus, et non diagnostiquées dans les calculs. Les calculs sont indépendants des langues, et ne sont donc sensibles ni à l'ajout d'une nouvelle langue, ni aux proportions entre langues différentes. La bonne qualité des résultats et l'adéquation de la méthode à la tâche nous montre que des propriétés linguistiques très générales sont exploitées : principalement les différences (ou valeurs relatives). Notons qu'on ne s'est pas intéressé à la distinction nominal - verbal, cette distinction étant inutile dans la tâche. Le fait de ne pas s'occuper de la distinction verbo-nominale ne perturbe pas l'exécution de la tâche, car on ne retient que les termes présents au moins deux fois dans le corpus (ensuite sur deux sites différents); d'autre part les groupes nominaux sont fréquents et répétés, les groupes verbaux sont plus rares et plus variés, d'où très peu de groupes verbaux en sortie.

De nombreuses directions restent à explorer. Au sujet du calcul des différences entre mots, peut-on définir d'autres critères ? Pour mieux traiter les mots vides rares, nous envisageons deux passes : une pour les mots pour lesquels il n'y a pas eu de contradiction entre critères de différence (traitement actuel), la deuxième les utilisant pour catégoriser les autres mots; une solution complémentaire serait de mémoriser les déductions faites la veille. Enfin, la méthode a été confrontée à un but opératoire, mais elle pourrait être approfondie en tant qu'exploration linguistique indépendante des langues, par exemple en direction de la catégorisation verbo-nominale.

Références

Ahonen-Myka H.,Heinonen O., Klemettinen M., Verkamo A. I. (1999), Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery, Proceedings of *16th International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Text Mining : Foundations, Techniques and Applications*, ed. R. Feldman, 1-9.

`www.cs.helsinki.fi/u/hahonen/ham_ijcai99.ps`

Bourigault, D. (1994), *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*, Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Bourigault, D. (2002), Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de *la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, Nancy, 75-84.
www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN02-Bourigault.doc

CLAIR : Computational Linguistics And Information Retrieval group, University of Michigan (2003).
perun.si.umich.edu/clair/

Debili, F. (1982), *Analyse syntactico-sémantique fondée sur une acquisition automatique de relations lexicales-sémantiques*, Thèse de doctorat d'état, Université de Paris XI, Centre d'Orsay.

Déjean H. (1998), *Concepts et algorithmes pour la découverte des structures formelles des langues*, thèse de l'Université de Caen.

Frath P., Oueslati R., Rousselot F. (2000), Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques, In *J.Charlet, M.Zacklad, G.Kassel & D.Bourigault, eds, Ingénierie des Connaissances, évolutions récentes et nouveaux défis*, Paris, Eyrolles.
www-ensais.u-strasbg.fr/liia/publications/for00.ps

Lucas N. (2001), Étude et modélisation de l'explication dans les textes, Actes du Colloque "*L'explication: enjeux cognitifs et communicationnels*", Paris.

Rousselot F. (2002), LIKES (LInguistic and Knowlege Engineering Station) : outil de traitement de corpus et d'aide à la construction d'ontologies.
www-ensais.u-strasbg.fr/liia/likes/likes.htm

Saussure F. de (éd. 1974), *Cours de Linguistique Générale*, Paris, Payot.

Vergne J. (2000), *Trends in Robust Parsing*, tutoriel du CoLing 2000, Nancy, Sarrebrück.
www.info.unicaen.fr/~jvergne/tutorialColing2000.html

Vergne J. (2001), Analyse syntaxique automatique de langues : du combinatoire au calculatoire (conférence invitée), Actes de *TALN 2001*, 15-29.
www.info.unicaen.fr/~jvergne/TALN2001_JV.ppt.zip

Vergne J. (2002), Une méthode pour l'analyse descendante et calculatoire de corpus multilingues : application au calcul des relations sujet-verbe, Actes de *TALN 2002*, 63-74.
www.info.unicaen.fr/~jvergne/TALN_2002/TALN2002_JVergne.doc.pdf

Zipf G. K. (1949), *Human Behavior and the Principle of Least Effort*, New York, Harper, réédition 1966.