La glose, le document électronique et l'extraction automatisée.

Rachel Panckhurst

LA GLOSE, LE DOCUMENT ELECTRONIQUE ET L'EXTRACTION AUTOMATISEE.

PANCKHURST Rachel
U.M.R. 5475 C.N.R.S. « Discours, textualité et production de sens », Praxiling,
Université Paul-Valéry Montpellier 3
Route de Mende
34199 Montpellier cedex 5
France
rachel.panckhurst@univ-montp3.fr

Résumé

Dans cet article, nous nous intéressons au repérage et à l'extraction de candidats glose, à partir de données sous format électronique, et ce, au moyen de méthodes en traitement automatique des langues (TAL). En premier lieu, nous décrivons le parcours allant de la constitution du corpus jusqu'à l'étiquetage morpho-syntaxique, en passant par l'épuration des données. Nous réfléchirons ensuite au processus de repérage et d'extraction proprement dit, en nous posant les questions suivantes : que peut-on repérer à l'aide d'un traitement automatisé ? quelles sont les limites actuelles de ce type de démarche ?

1. Situation de l'étude

Nous avons commencé cette recherche dans le cadre du séminaire proposé par Agnès Steuckardt de l'université de Provence (Groupe de travail « Sémantique lexicale et discursive »), et suite à la journée d'études « Le mot et sa glose », qui s'est tenue à l'université de Provence le 24 septembre 2001. Les recherches présentées lors de cette journée et tout au long de l'année universitaire 2001-2002 dans le cadre du séminaire ont permis de mieux définir la notion même de glose. Nous ne reviendrons pas sur ces débats de manière détaillée.

Le mot glose vient du latin *glosa* « mot qui a besoin d'être explicité » et du grec *glôssa* « langue ». Le dictionnaire Petit Robert définit la glose ainsi : « annotation entre les lignes ou en marge d'un texte, pour expliquer un mot difficile, éclaircir un passage obscur ». Authier-Revuz, pour sa part,

parle de « commentaire épi-linguistique ». Nous adoptons la définition de Julia : « [les gloses servent] à la spécification et à l'explicitation du sens au fil du discours » (2001 : p. 11).

Julia a étudié la glose à partir de textes romanesques (Hugo, Proust, Flaubert, etc.), d'essais (Mme de Stäel, Valéry, Gide, etc.), de journaux intimes ou de biographies (G. Sand, Mauriac, Gide, etc.), de correspondances (Mallarmé, Claudel, Hugo, etc.), de textes historiques, de géographie, d'ouvrages d'esthétique et d'histoire de l'art, de manuels de droit… 75% de son corpus émane de Frantext et du TLF[1], et 25% de la presse quotidienne, d'essais et de romans. Elle a relevé environ 500 formes de glose, essentiellement sur le modèle de : *au sens* / *dans l'acception* / *dans tous les sens*, etc. et des locutions adverbiales comme : *littéralement, à proprement parler,* etc.

Dans le cadre de cette recherche, nous avons choisi d'étudier des documents électroniques (en provenance de courriels, forums et chats). Dans cette optique, il sera intéressant de comparer les types de glose qui y figurent et de comparer les occurrences avec les textes du corpus de Julia. Les pourcentages obtenus pour les mots glosés dans le corpus sont rappelés ci-dessous (Julia, 2001, p. 57-58). Une comparaison avec notre corpus électronique est présentée à l'issue de l'analyse des résultats (*cf.* §5.1.).

noms (68,5%)

> La première scène d'Iphigénie est une ouverture au sens musical du mot (H. Bremond, La Poésie Pure, 1926, p. 17)

adjectifs (17,5%)

> […] je le trouve peu humain, au sens psychologique (G. Flaubert, Correspondance, 1853, p. 230)

verbes (12,5%)

> […] mon amour à moi, tout fier, se prévalait et se targue pour t'adorer au sens mystique (P. Verlaine, Œuvres poétiques complètes, 1896, p. 779)

locutions idiomatiques (1,5%) :

> Elle s'arrachait les cheveux, au sens propre, par poignées, comme dans un roman russe. (D. Pennac, La petite marchande de prose, 1989, p. 98)

2. Constitution du corpus

Lorsqu'on souhaite repérer et extraire des informations en langage naturel en vue d'une analyse linguistique, il convient de disposer au préalable d'un échantillon de *données*. Avant de choisir le type d'échantillon requis, et de le constituer, il nous faut définir la notion même

[1] Frantext : `http://frantext.inalf.fr/frantext.htm`
TLF-i : `http://frantext.inalf.fr/tlf.htm`

de *données*. Suite à Panckhurst (1994, p.39-40), nous pensons qu'il existe trois types de données en linguistique : le texte nu, ou *corpus*, la donnée lexicale (qui consiste essentiellement d'entrées lexicales, avec des ensembles attribués de propriétés), la donnée observable en linguistique, ou *l'exemple*, selon Milner (1989, p.31). Pour l'étude présente, nous avons choisi de travailler à partir du premier type de donnée, soit le *corpus*, et ce indépendamment d'une théorie linguistique spécifique. Précisons maintenant quelle définition nous adoptons du terme *corpus*. Crystal (1997, p.414) définit le corpus comme suit : « a representative sample of language, compiled for the purpose of linguistic analysis »[2]. À notre sens le corpus peut être constitué de textes bruts en entier ou d'échantillons partiels. Dans le cas présent, nous avons décidé de travailler à partir de données textuelles électroniques brutes, en provenance de courriels, de forums de discussion et de chats. Contrairement à d'autres chercheurs, œuvrant dans le domaine des linguistiques du corpus ou du TAL, pour lesquels la (grande) taille du corpus peut constituer un atout majeur, nous pensons, avec Crystal que : « An important principle is that all corpora, whatever their size, are inevitably limited in coverage, and always need to be supplemented by data derived from the intuitions of native speakers of the language, through either introspection or experimentation »[3], (Crystal, 1997, p.415).

Pourquoi avons-nous choisi de constituer un corpus à partir de données en provenance de courriels, de forums et de chats ? Dans le cadre de notre projet de recherche actuel concernant la *communication médiée par ordinateur*, (C.M.O., forme de communication entre deux ou plusieurs personnes via des ordinateurs interposés), nous pensons que la C.M.O. modifie le discours et la façon de communiquer avec autrui. Un nouveau « genre » de discours est ainsi induit : le *discours électronique médié* (DEM). Le DEM renferme ses propres marques linguistiques et extra-linguistiques et nous continuons à nous intéresser aux différents moyens de communication (courriel, forum, chat), dans l'optique de l'amélioration de nos connaissances sur ce nouveau type de discours. Par ailleurs, l'avantage des données textuelles de type électronique est lié à la facilité de

---

2 « Un échantillon représentatif du langage, selectionné et organisé à des fins d'analyse linguistique » (notre traduction).

3 « Un principe important est que tout corpus, quelle que soit sa taille, est inévitablement limité quant à sa couverture, et doit être nécessairement complété par des données dérivées d'intuitions de locuteurs natifs, au moyen d'introspection ou d'expérimentation » (Notre traduction).

constitution de corpus : ces données sont quasi-immédiatement exploitables.

Le corpus décrit dans cet article est donc constitué, d'une part, de messages courriels en réception (professionnels et privés), pour la période de janvier 2000 à janvier 2002 (7 728 messages, 2 576 879 mots, un fichier de 19,8 mo), et, d'autre part, de messages en provenance de forums et de chats dans le cadre de deux enseignements pendant une partie de l'année universitaire (d'octobre 2001 à mars 2002) :

> Licence sciences du langage, mention traitement automatique des langues (étudiants en présentiel) ;
>
> Licence information-communication[4] (étudiants en formation continue).

Pour la licence sciences du langage, mention TAL (désormais LTAL), nous avons recueilli 178 messages (4215 mots) sur le forum et 181 messages (1185 mots) sur le chat – séances en présentiel le 9/10/01 de 13h13 à 14h30, et le 15/10/01 de 11h48 à 12h10. Pour la licence information-communication (désormais LIC), 124 messages (7273 mots) proviennent du forum et 1040 messages (9416 mots) proviennent du chat n° 1, intitulé « infocom1 » (parmi ceux-ci 20 messages sont vides car l'émetteur a appuyé sur la touche retour-chariot par inadvertance) : le 28/1/02 de 20h29 à 22h36, et le 4/2/02 de 19h28 à 22 : 13. Les deuxième et troisième fichiers de chat proviennent également du même public LIC, le 25/2/02, de 20h34 à 22h09 (« infocom2 ») — 446 messages et 3 983 mots, et le 11/3/02 de 20h33 à 22h40 (« infocom3 ») — 945 messages et 8463 mots.

À partir de ces données on constate que :

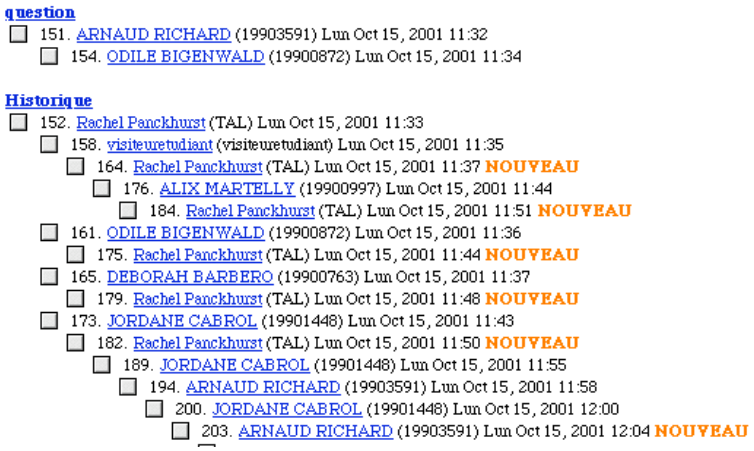
> les messages en provenance du forum LTAL sont plus nombreux que ceux du forum LIC, mais la taille du corpus (en mots) est plus importante en LIC ;
>
> les messages de chat sont très nombreux, mais la longueur de ceux-ci est en général très réduite. Il suffit de comparer le chat et le forum pour s'en apercevoir : 2612 messages au total pour le chat (23 047 mots) contre 302 messages en provenance du forum (11 488 mots).

---

[4] Nous remercions Patricia Jullia, qui a bien voulu nous fournir les données pour les forums et les chats information-communication, dans le cadre de la convention Praxiling-METICE (projet « communication médiée par ordinateur »).

2.1. Présentation graphique du forum

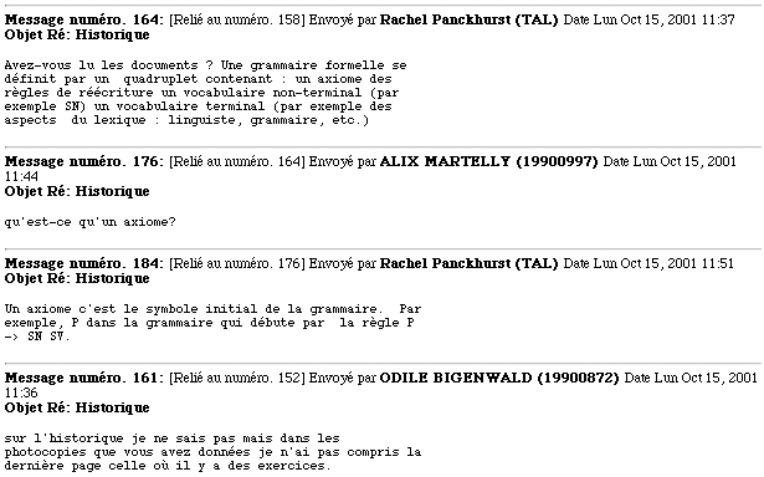
Un forum comme celui existant au sein de la plateforme WebCT (adoptée à l'université Paul-Valéry Montpellier 3 pour les enseignements en «□présentiel assisté□ ou «□à distance□), se présente de la manière suivante :



Les «□fils□ (ou thèmes) de discussion sont présentés sous forme d'emboîtement hiérarchique. Par exemple, *Rachel Panckhurst* envoie un message (n°□152) au forum, intitulé «□Historique□. *Visiteurétudiant* répond (n° 158) ; une série de réponses s'enchâinent, puis, *Odile Bigenwald* répond au message n°□152, etc. Les emboîtements peuvent se situer à plusieurs niveaux (selon le choix de réponse de l'utilisateur)□ il pourrait être intéressant de réflechir sur les possibilités de glose figurant dans un système de questions-réponses, mis en lumière à partir de cette dimension graphique du forum.

2.2. «□Compilation□ de messages

Il est possible de «□compiler□ tous les messages du forum pour pouvoir les lire de manière linéaire, séquentielle□ l'écran. Dans ce cas, les fils sont mentionnés dans l'en-tête□



2.3. Épuration des données

Afin de traiter les données, on les sauvegarde en format texte ; on perd alors les éventuels fichiers annexés. Les données peuvent ensuite être ouvertes dans un éditeur de texte (Alpha, BBEdit, etc.). En vue d'une extraction automatisée, les données doivent être « épurées », sinon, certains aspects statistiques risquent d'être erronés. Dans la photo d'écran ci-dessus (*cf.* § 2.2.), si la ligne contenant « Message numéro X Envoyé par Y Date Z » n'est pas effacée pour chaque message, alors les données verbales seront faussées ; par exemple, le verbe « envoyer » figure à maintes reprises. Pour ce qui concerne les fichiers annexés, leur contenu n'est pas sauvegardé pour le traitement automatique dans le cadre présent[5].

```
------------------------------------------------------------
Avez-vous lu les documents ? Une grammaire formelle se définit par un  quadruplet contenant :
un axiome des règles de réécriture un vocabulaire non-terminal (par exemple SN)
un vocabulaire terminal (par exemple des aspects  du lexique : linguiste, grammaire, etc.)
------------------------------------------------------------
qu'est-ce qu'un axiome?
------------------------------------------------------------
Un axiome c'est le symbole initial de la grammaire.
Par exemple, P dans la grammaire qui débute par  la règle P -> SN SV.
------------------------------------------------------------
sur l'historique je ne sais pas mais dans les photocopies que vous avez données
je n'ai pas compris la dernière page celle où il y a des exercices.
------------------------------------------------------------
```

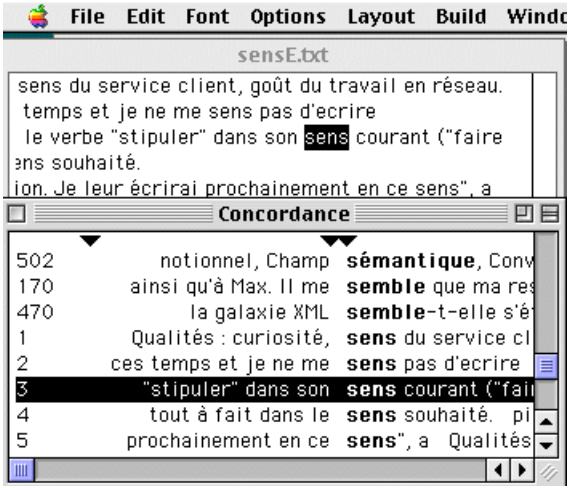
3. Utilisation d'un concordancier

Le traitement par un concordancier peut permettre l'« approche » d'un texte. Les fonctions minimales d'un logiciel concordancier sont de permettre (1) le tri alphabétique en contexte de tous les mots[6] et (2) l'affichage d'un index par nombre d'occurrences (croissant/décroissant). Nous nous sommes servi du logiciel « Conc » afin de cerner l'utilisation du mot « sens ».

---

[5] À notre demande, un de nos étudiants de LTAL, Stéphane Riou, a élaboré un programme PERL permettant d'épurer les données figurant dans les en-têtes des messages en provenance des forums. Il n'a pas été possible de prévoir un effacement automatique des programmes informatiques joints dans le corps du message, en réponse à une de nos demandes pendant le cours en présentiel, car il n'y a aucun marquage distinctif.

[6] La définition d'un mot n'est évidemment pas la même en linguistique et en informatique. Des problèmes de segmentation interviennent à ce stade : *pomme de terre, machine à laver,* etc.



Dès lors, on remarque que les catégories morpho-syntaxiques seraient très utiles car nous ne voulons pas, par exemple, extraire les occurrences verbales de « sens » (*sentir*). Pour ce faire, le recours à un logiciel permettant l'étiquetage morpho-syntaxique s'avère nécessaire.

4. Étiquetage morpho-syntaxique

Dans le cadre de notre recherche, nous avons volontairement limité les recherches de gloses à quelques cas que nous énumérons ci-dessous :

sens (Nom/Verbe/Adjectif (ponctuation) *au sens de, dans le sens de*, etc.)
c'est-à-dire
Nom (ponctuation) ou Nom
Nom/Verbe/Adjectif (parenthèse ouvrante)

Le choix a porté sur ces quatre[7] cas puisqu'un traitement automatisé est envisageable dans la mesure où des marques lexico-syntaxiques sont apparentes.

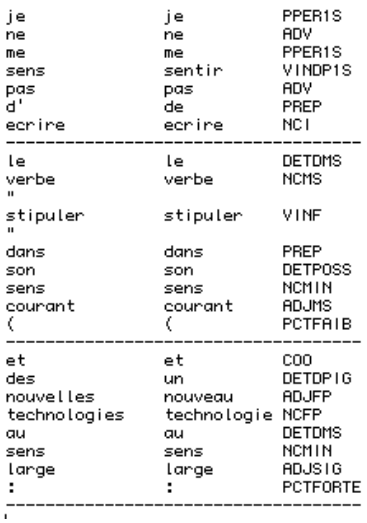
Le logiciel Cordial, disponible sous Windows[8], permet d'affecter un étiquetage morpho-syntaxique aux mots, et, par extension, aux phrases d'un texte. On peut consulter les « occurrences » soit « orthographiques » soit « syntaxiques ». Par exemple, si un verbe comme « vouloir » apparaît dans le texte, le nombre d'occurrences syntaxiques sera fourni, mais si on

[7] L'extraction pour *N/V/Adj (* dans le corpus courriel a été écartée du traitement actuel à cause du volume des réponses (plus de 23 000 ! *Cf.* § 7.)

[8] Étant donné que nous avons travaillé avec un Macintosh, le recours à l'émulateur (pour simuler l'envrionnement Windows) a été nécessaire afin d'effectuer le traitement à l'aide de Cordial. Bien entendu, un transfert d'encodage de caractères (MacOS -> Windows, puis *vice versa* après le traitement) a été fait par un logiciel comme « Text to .txt », disponible sur Internet.

fait ensuite une recherche sur « vouloir » afin de visualiser les occurrences dans le texte, la recherche n'aboutira pas si certaines formes verbales figurent dans le texte sous la forme conjuguée du verbe. Il sera alors important de procéder à une recherche par occurrences « orthographiques » afin de localiser la totalité des occurrences[9].

Dans un premier temps, nous avons effectué des recherches sur le courriel, pour le mot « sens ». Dans la photo d'écran ci-dessous, nous présentons un exemple d'étiquetage pour plusieurs lignes de notre corpus :



je	je	PPER1S
ne	ne	ADV
me	me	PPER1S
sens	sentir	VINDP1S
pas	pas	ADV
d'	de	PREP
ecrire	ecrire	NCI
le	le	DETDMS
verbe	verbe	NCMS
"		
stipuler	stipuler	VINF
"		
dans	dans	PREP
son	son	DETPOSS
sens	sens	NCMIN
courant	courant	ADJMS
(	(	PCTFAIB
et	et	COO
des	un	DETDPIG
nouvelles	nouveau	ADJFP
technologies	technologie	NCFP
au	au	DETDMS
sens	sens	NCMIN
large	large	ADJSIG
:	:	PCTFORTE

Dans les deux colonnes à droite du mot, figurent respectivement : le lemme (forme canonique — en général l'infinitif pour le verbe, le masculin pour un adjectif, le singulier pour un nom, etc.) puis la catégorie retenue par Cordial.

5. Repérage et extraction de données : analyse des résultats[10]

Le parcours qui suit est indiqué afin d'expliquer comment on peut appréhender un texte à partir de l'utilisation d'outils pour le traitement automatique selon des critères morpho-syntaxiques. Nous verrons ultérieurement (*cf.* §6.) les limitations d'un tel traitement.

---

[9] Nous avons choisi les paramétrages suivants pour l'étiquetage dans Cordial : Lemmes, Lemmes fém -> masculin, aucune expression, type grammatical : abrégé en majuscules, codage spécialisé : aucun, traitement des erreurs : ne pas corriger, ne pas signaler.

[10] Les exemples authentiques présentés dans cet article incluent les éventuelles erreurs orthographiques et grammaticales qui figurent dans le corpus.

5.1. Le cas de « sens »

À partir des occurrences retenues (417 par BBEdit, 397 par Conc) pour le mot « sens » au sein du corpus courriel, 362 occurrences correspondent à des noms. Bien entendu, ce premier tri permet d'écarter les formes verbales. Le programme Perl[11] que nous avons élaboré pour le repérage des gloses permet ensuite de diviser les catégories apparaissant à gauche et à droite de la forme étudiée.

Dans un premier temps, nous avons voulu comprendre si le nombre de positions intercalaires entre le nom, le verbe ou l'adjectif et le mot « sens » était un facteur discriminant. Le patron retenu dans un premier temps pour « approcher » la glose était : N ou Adj ou V suivi de « n'importe quelle chaîne (ne contenant pas « sens ») » suivi de « sens ». 329 formes sont extraites pour ce patron. Parmi celles-ci, 202 formes correspondent à une seule position intercalaire (*N/Adj/V au sens*) ; en augmentant à 2 positions, on obtient 299 formes (*N/Adj/V , au sens*) ; 3 ou 4 positions nous fournit 313 formes (*N/Adj/V , dans le sens, N/Adj/V , mais dans le sens*). On peut conclure que la recherche par positions intercalaire ne nous apprend rien de significatif.

En modifiant maintenant notre patron afin d'obtenir les formes *N/V/Adj au sens*, on obtient 21 candidats à glose. À partir des 329 formes initiales, on peut supposer qu'il y aura un problème de « silence[12] ». Effectivement, certaines formes de glose correspondent plutôt au patron : *N/V/Adj (ponct) prep det sens*. Ce deuxième patron nous ramène 86 candidats à glose ; y a-t-il du bruit ? En spécifiant le choix de la préposition (*dans*) et en écartant certains déterminants (par exemple, « de » est codé déterminant démonstratif dans Cordial), on obtient 17 formes. En fin de compte, nous obtenons un total de 38 candidats à glose (21 candidats pour le 1er patron, 17 candidats pour le second). Parmi ces 38 formes, 10 correspondent à des doublons (car nous recevons les courriels envoyés à deux listes électroniques de l'université Paul-Valéry) ; 8 candidats sont

---

[11] La première version du programme que nous avons développé en Perl a été adaptée à partir d'un programme élaboré par Augusta Mela pour le traitement de la glose. Bien entendu, nous l'avons remanié dans le détail afin de répondre à nos besoins précis de repérage et d'extraction.

[12] Dans un sens documentaire, le silence correspond à une absence de réponses alors qu'il existe des réponses valides. Le bruit, au contraire, renvoie certaines réponses superflues par rapport à la demande. En traitement automatique, il est important de repérer un surnombre de candidats (bruit) et d'en éliminer certains par la suite, plutôt que d'être dans la situation de n'avoir pas repéré certains candidats importants, mais le bruit ne doit pas non plus être exagérément élevé.

écartés manuellement. Ainsi, il reste 20 candidats repérés automatiquement. Après un travail de comparaison manuelle avec les 362 candidats d'origine, afin de vérifier le travail automatique, on peut conclure qu'il y a 25 cas réels de glose, parmi lesquels figurent :

14 occurrences « N/V/Adj au sens X » (cf. Annexe (1)) ;
6 occurrences « N/V/Adj dans le sens X » (cf. Annexe (2)) ;
5 occurrences non repérées.

Les occurrences non repérées par le programme, ainsi que celles écartées manuellement sont présentées ci-dessous ; les autres figurent en annexe.

Cas non repérés

Parmi les 5 cas (réels de glose) non repérés par le programme, seul le dernier cas ci-dessous provient d'une faiblesse de traitement du programme de repérage et d'extraction : (1) et (2) correspondent à un problème de guillemets ayant provoqué un étiquetage incorrect. Par conséquent, les phrases n'étaient pas repérables par le programme.

1. Suite aux remarques des camarades juristes (pour eux seul un contrat stipule et non la loi), je précise que j'avais utilisé - dans mon précédent courriEL - le verbe "***stipuler*** " ***dans son sens*** *courant ("faire savoir expressément") et non dans son usage juridique ("énoncer comme condition dans un contrat, un acte")* ! Ah mais! merci Petit Robert... (message syndical, le 27 mai 2000)
2. Je pense comme vous que ceux qui utilisent l'expression « communication médiatisée par ordinateur » utilisent ***« médiatisée » dans le sens de*** *« médiée »*, mais n'osent utiliser le néologisme. (le 18 mars 2001, message envoyé à un collègue, avec double en retour — commentaires sur un passage de sa thèse).

Dans les exemples (3) et (4), les suites « au sens large » et « du terme » respectivement, figurent en début de ligne et une coupure de ligne a provoqué leur non-traitement par le programme :

3. Ce constat pourrait aussi se faire a propos *d'autres groupes* ***culturels, au sens*** *large* (ethniques, sociologiques, culturels). (liste de l'université, 21 novembre 2000)
4. Je crains qu'une personnalité « trop compétente », impatiente, ***ambitieuse dans tous les sens*** *du terme*, ne les heurte et démotive. (envoi à la présidente par un collègue, avec copie pour moi-même, 2 juin 2001).

Le cas suivant n'est pas repéré par l'un des deux patrons, car « sens » est antéposé par rapport au nom glosateur. Ce cas constitue une faiblesse de notre repérage automatique :

5. *Au sens strict du terme, un schéma directeur* impose une étude longue et coûteuse, dont il n'est pas certain que nous ayons actuellement les moyens. (envoi à la présidente avec copie pour un groupe de personnes, par un collègue, 4 octobre 2001).

Cas écartés manuellement

Dans les 7 exemples suivants (n° 6 à 12) la syntaxe présente est identique à des cas réels de glose. La désambiguïsation automatique est donc impossible à partir de critères uniquement syntaxiques.

N|V prep=*dans* DET=*le* N=*sens* V|DET N ADJ DET=*au* N=*sens*

6. Les montées sont simulées par des freins qui obligent à faire un effort supplémentaire pour se mouvoir, alors que les descentes sont facilitées par un moteur électrique qui actionne les ***roues dans le sens*** indiqué par l'ordinateur.
7. D'ailleurs, je sais que je ne me suis pas toujours bien fait voir en disant tout haut ce que tout le monde me répète tout bas, mais de là à laisser ceux «qui lui passe la ***brosse dans le sens*** du poil» accéder au net et pas moi, c'est un peu exagéré. (reçu d'un personnel du METICE)
8. Très rapidement (le sujet mériterait de plus amples développements qui glisseraient vers l'épineux problème des marchés publics) mais pour ***aller dans le sens*** du collègue [X], je dois dire que, d'après ce que j'ai compris, Sauramps n'est pas tout à fait maître du prix des livres importés. Ils doivent négocier cela avec des importateurs et/ou diffuseurs qui peuvent parfois appliquer des tarifs assez délirants.
9. confirmation de la réforme des normes sanremo ***dans un sens*** favorable aux universités de Lettres
10. Les transformations du cadre national des ***diplômes dans le sens*** du développement, de l'ouverture à la société et au monde pourraient permettre d'élargir les voies de la réussite du plus grand nombre de jeunes et d'adultes.
11. Il serait paradoxal qu'un travail mené depuis plus d'un an, auquel a contribué [X], qui débouche sur des mesures de progrès pour la communauté des enseignants, donne lieu à des interprétations rigoureusement ***contraires au sens*** et à la portée que revêtent les mesures prises ou souhaitées.
12. Si nous ne voulons pas que la droite reprenne le pouvoir, il faut changer les méthodes de gestion de l'actuelle direction, retrouver les voies de la concertation, construire les ***projets dans le sens*** du maintien et de l'amélioration du service public, mobiliser la communauté universitaire contre la,libéralisation et la marchandisation des formations.

L'exemple 13 n'est pas correctement repéré à cause d'une erreur d'étiquetage par Cordial.

13. J'ai pris connaisance de votre rapport sur [X] qui me semble aller tout à ***fait dans le sens souhaité***.

Tout	tout	ADV
à	à	PREP
fait	fait	NCMS
dans	dans	PREP
le	le	DETDMS
sens	sens	NCMIN
souhaité	souhaiter	VPARPMS

Si l'on compare les statistiques obtenues par Julia (que nous rappelons dans le tableau ci-dessous) avec les nôtres, on constate qu'il y a une ressemblance au niveau des pourcentages concernant le nom, mais une variation concernant les verbes et les adjectifs.

Catégorie		N° sur 25	%	% Julia
N		13	52 %	68,5 %
Adj | N | V en pos. Adj.		10	40 %	17,5 %
	Adj	8	32 %	
« médiatisée »	V	1	4 %	
« directeur »	N	1	4 %	
V		1	4 %	12,5 %
UVPL		1	4 %	1,5 %

Dans notre corpus de courriel, seuls apparaissent 25 cas de glose de type « sens » pour un corpus de 7 728 messages, soit 2 576 879 mots. Par ailleurs, ces cas proviennent quasi exclusivement de listes, de revues électroniques, de messages commerciaux ou de la direction de l'Université (groupes restreints) ou encore du Ministère. Seulement 3 messages contenant des gloses constituent un envoi d'une seule personne à une autre personne (parmi ceux-ci 2 correspondent à une explication sur une thèse et le troisième constitue un message reçu en privé).

5.2. C'est-à-dire

La deuxième recherche a porté sur « c'est-à-dire ». Le problème majeur pour le repérage automatique a concerné l'orthographe. Outre l'orthographe dictionnairique classique, nous avons rencontré dans le corpus global courriel/forum/chat), les formes suivantes parfois très abrégées, plus insolites : *c'est à dire, cest a dire, cest à dire, c est a dire, cad, càd, c'à dire*.

Pour le corpus courriel, parmi les 30 occurrences repérées, 21 correspondent à la forme canonique (*c'est-à-dire*), 6 formes de *c'est à dire*, 2 occurrences de *cad* et 1 occurrence de *càd*. Se pose alors le problème du prétraitement d'édition avant l'utilisation du logiciel étiqueteur : faut-il rectifier les différentes formes utilisées ? Sinon, on obtient des mots incorrectement étiquetés. Par exemple, Cordial traite *cad* ou *càd* en NCI (nom commun invariant en nombre et en genre). Quant à *c'est à dire,* l'étiquetage proposé est le suivant, car l'absence de traits d'union ne facilite plus le repérage de la locution conjonctive :

```
c'      PDS
est     VINDP3S
à       PREP
dire    VINF
```

Enfin, l'abréviation officielle *c.-à-d.* est incorrectement étiquetée, les points abréviatifs étant repérés en tant que ponctuateurs forts ou laissés sans étiquetage.

```
Phrase testée : Toto c.-à-d. Titi.
mot     lemme Typegram
  ==== DEBUT DE PHRASE ====
Toto    toto    NPMS
c.      c.      NCMIN
  ===== FIN DE PHRASE =====
  ==== DEBUT DE PHRASE ====
-
à       à       PREP
-d      de      PREP
.
Titi    titi    NPMS
.       .       PCTFORTE
  ===== FIN DE PHRASE =====
```

À partir de 30 formes, il reste 25 cas en définitive ; 5 constituent des doublons. Parmi celles-ci (*cf.* (3) en annexe), 3 ne correspondent pas à des gloses, mais à une explicitation apportant des informations complémentaires (par exemple : « Bien sur oui, je compte faire une formation à partir de ce type de fichier, c'est à dire que j'ai mis ce fichier au point conformément à un résultat donné, conformément à une feuille de style donnée »). Les statistiques concernant la répartition des catégories grammaticales et leurs positionnements syntaxiques sont les suivantes :

SN□ 77,27 %, Sprép = 13,6%, V = 4,54 %, participe passé en position Adj.□□4,54 %

5.3. N (ponctuation) ou N

22 candidats à glose apparaissent dans le corpus courriel. 11 doublons peuvent être retirés□ parmi les 11 formes restantes, 2 cas correspondent à des erreurs de codage par Cordial (à cause d'un manque d'accentuation)□ *Les propositions, preparees au moyen du logiciel ProTool, sont imprimees par le coordonnateur, ou **presentees*** (codé en tant que NCI par Cordial) *sur les formulaires papier joints au guide du proposant.* 5 candidats sont ensuite écartés après vérification manuelle (par ex.□ «□Non, je peux, jeudi ou vendredi ***matin, ou vendredi*** à 14h□ ou «□faut-il choisir l'option tour de cederoms, ***jukebox, ou stockage*** sur disque ?□)

En fin de compte, à partir de 22 candidats glose, 4 seulement constituent des cas réels de glose□

1. La représentation écrite de la **parole, ou graphie** présente une grande autonomie par rapport à la parole.
2. Les ingénieurs technico-commerciaux du PAD (division des**Preferred Accounts, ou Grands Comptes**) sont des références techniques sur l'ensemble de l'offre Dell□portables, stations, serveurs, stockage, etc.
3. un peu comme pour le DALF/DELF en **FLE, ou Cambridge** pour l'anglais).
4. Le désabonnement s'obtient en envoyant à l'adresse **SYMPA (ou sympa@univ-montp3.fr)** un message vide contenant seulement, en subject (ou sujet, ou objet) la commande unsub upv-l

Bien entendu, on peut conclure que la forme *N (ponctuation) ou N* ne constitue pas une forme discriminante de glose, mais cette hypothèse doit être vérifiée à la lumière d'une recherche au sein d'autres corpus[13].

5.4. Forums et chats

Nous avons ensuite porté notre attention sur les forums et les chats, afin d'essayer de dégager une typologie des formes employées.

Sens et *c'est-à-dire*

Aucune occurrence de «□sens□ ne figure dans tout le corpus forum. Un seul cas apparaît dans le chat («□nfocom2□)□

1. imho, efficace pour la com interne et par ailleurs, l'image, dans le sens «□de marque□, bien sur

[13] Pour ce qui concerne le quatrième cas de figure (N/V/Adj suivi de parenthèse ouvrante) *cf.* § 6.

2 occurrences de « c'est-à-dire » figurant dans le chat « infocom1 » (sous la forme graphique : *c'est a dire* et *c'à dire*) ne correspondent pas à des gloses : un seul cas de glose est relevé dans le chat (« infocom2 ») et 2 cas dans le forum LTAL (*cad, c'est-à-dire*) :

1. la communication sociale **cad** qui touche aux domainces sociaux association de quartier, humanitaire, et autre - avez vous remarquez des constantes (*chat*)
2. le ta est le traitement automatique des langue **cad** comment on fait pour que l'ordinateur comprenne sans ambiguite et en prenant en compte ce qui peut etre implicite. (*forum*)
3. Il s'agit de la question comment des donnéés linguistiques peuvent être traité dans l'informatique, **c'est-à-dire** dans un système artificielle. (*forum*).

N (ponctuation) ou N

À partir des 19 candidats (15 dans le corpus chat, 4 dans le corpus forum) seuls 3 cas sont retenus, mais ceux-ci sont ambigus et pourraient s'avérer incorrects.

1. Le contexte expressif rencoie quant à lui , aux ***intentions ou projets*** des acteurs. (*forum*)
2. En effet, il donne prix des ***accessoires ou produits*** pour enfants (*chat*).
3. Les clients qui ne peuvent pas assister aus ***defiles ou distributions*** de reves (*chat*).

N/V/A (

9 gloses (8 en provenance d'enseignants, 1 en provenance d'un étudiant), apparaissent au sein de forums :

1. générative veut dire qu' elle doit être capable, à partir des règles de la grammaire décrite, de ***construire (de générer, de créer)*** toutes les phrases pour lesquelles elle a été initalement prévue. (*forum — enseignante*)
2. Les logiciels du ***cours (MAcPerl, Prolog, OUtils suédois sur MAC et 101, Conc sur PC)*** sont accessibles. (*forum — enseignante*)
3. Cependant dreamweaver semble plus puissant et plus ***riche (mise en place d'animation...)*** ce qui peut etre interessant pour cette année (*forum — étudiant*).
4. Imaginer quel peut être le sentiment d'une ***hiérarchie intermédiaire (style « contremaître »)*** quand la politique de management prône « moins de contrôle » . (*forum*).
5. Repérer les groupes d' ***acteurs sociaux (ceux qui ont en gros les mêmes intérèts***). (*forum*).
6. Can you put the letter you sent to me by e-mail on this ***bulletin board (forum de discussion)*** please ? (forum — enseignant)
7. ***Imaginer (mettre en image)*** par exemple, comment Nadia, (d' origine arménienne […]appréhende le monde en général. (*forum — enseignante*)

8. Il s' agit de comparer les ***5 analyses (structuro-expressive, analyse transactionnelle, systémique, phénoménologique et sémio-contextuelle)*** selon les rubriques suivantes […] (*forum — enseignante*)
9. Ce cadrage large permet de saisir le ***système (les actions et rétro-actions entre tous les éléments du système)*** et sa logique de fonctionnement. (*forum — enseignant*)

Pourrait-on conclure ainsi que l'explicitation pédagogique est à l'origine de ce phénomène ? Pourtant, au sein du chat (« infocom2 »), 4 gloses exclusivement en provenance d'étudiants sont identifiées. Les comparaisons doivent donc être approfondies.

1. je parle des ***élections (e-voting)*** cela relancerait le nombre des votants
2. c'est un bon moyen de prendre la ***température (feed-back positif ou négatif desinternautes)***
3. c'est aussi intéressant d'avoir des statistiques sur les ***visites (nombre de connexion, nombre de pages lus,…)***
4. Je vous embrasse tous depuis ***le bout du monde (rep dominicaine)***

6. Limites du traitement automatique

Dans cet article, nous essayons de démontrer comment peut être effectuée une « approche » du texte, afin d'automatiser certains processus qui seraient laborieux manuellement. Nous pouvons dire que l'extraction après étiquetage permet de *cerner* des contextes, mais en aucun cas cela ne permet de se fier totalement au repérage automatique. Un post-traitement manuel est tout à fait souhaitable, voire essentiel. Le danger du traitement automatique est précisément que l'on peut être tenté de se satisfaire de résultats sans « replonger » dans le contexte. Il faut que le TAL permette de repérer des « candidats à être gloses », ou, dans un autre contexte, des unités verbales polylexicales, des synapsies, etc., mais ces candidats seront ensuite vérifiés par un expert humain. Évidemment, il est souhaitable de se trouver dans une situation de *bruit* — au sens documentaire — avec des candidats à écarter, plutôt que dans une situation de *silence*, avec le danger évident de ne pas repérer des candidats potentiellement intéressants.

À ce stade, il est également important de mettre en garde contre une approche de type « patrons » *vs*. une approche plus linguistique, par exemple de type positionnel[14]. Certains logiciels (*cf*. David 1993, Souchard *et al.,* 1997) permettent de repérer des positions syntaxiques et non pas simplement des suites linéaires de catégories. C'est le cas du logiciel

[14] Je remercie S. David pour une discussion fructueuse à propos des exemples émanant de sa thèse sur les unités nominales polylexicales. Les exemples qui figurent dans ce paragraphe sont les siens.

Termino, qui repère les candidats termes (de type nominal) au sein de SN. En partant d'un exemple comme *un système de gestion de bases de données*, le logiciel repère la frontière droite (positionnelle) et permet d'extraire les candidats suivants *bases de données, gestion de bases de données, système de gestion de bases de données,* mais il écarte automatiquement **système de gestion, *gestion de bases, *système de gestion de bases*. Une approche par patrons repère indifféremment (et par conséquent incorrectement) tous les exemples précédents.

Par ailleurs, le logiciel positionnel ne peut pas systématiquement repérer les candidats correctement, mais pour des raisons linguistiquement valables. Par exemple, dans les deux exemples suivants, on constate qu'une information en provenance de la sous-catégorisation verbale (*parler de X à Y/ à Y de X*) doit être incorporée pour reconnaître la possibilité de synapse dans le cas « cause de divorce » et non pas dans le premier exemple « fille de programmation ».

Il a parlé à une fille de programmation
Il a parlé d'une cause de divorce à sa sœur

Certaines ambiguïtés réelles ne peuvent être relevées par un logiciel de type Termino, à cause de l'identité syntaxique entre le candidat terme et une phrase canonique, et pour des raisons liées à la sous-catégorisation verbale. Dans l'exemple *J'ai rapporté un vase de Chine*, on ne sait pas, hors contexte, si le vase provient de Chine ou s'il s'agit d'un vase d'un type particulier (*cf.* par exemple, *j'ai ramené une belle assiette en porcelaine de Limoges de Paris*).

Enfin, certains candidats termes n'en sont pas réellement car l'ambiguïté est réelle entre le terme (par exemple *le mur du son)* et le SN classique (par exemple *Le chat du voisin*).

7. Conclusion et perspectives

Cette première recherche, qui doit être approfondie, semble indiquer qu'il y a un faible nombre de gloses dans le discours électronique (ou « Netspeak », *cf.* Crystal 2001). Lorsque les gloses apparaissent, c'est notamment (mais non pas exclusivement) en situation didactique (l'enseignant explique une consigne, etc.) : c'est le cas pour la forme « N/V/A (» dans le corpus forum.

Le pourcentage des noms (*vs.* les verbes) semble ancrer la glose (au sein du discours électronique) du côté d'autres formes de l'écrit (*vs.* des formes orales). Cette hypothèse doit, bien entendu, être vérifiée (*cf.* Gadet

1996, Haliday 1989). Pour Gadet, à l'oral, il y a un emploi accru de verbes et, à l'écrit, davantage de noms. Pour Haliday, les déterminants, les pronoms, les prépositions, les conjonctions, les adverbes et les verbes fléchis sont moins employés à l'écrit, face aux noms.

La glose est peut-être apparente sous d'autres formes dialogiques ou graphico-dialogiques, notamment au sein des forums et des chats. Par ailleurs, les marques typographiques (comme le deux-points) sont sans doute plus fréquentes que des formes lexicales ou syntaxiques de glose, dans la mesure où l'on a moins de temps (et de place) pour écrire. Mais si le marquage est réduit, voire absent, comment avoir recours à un traitement automatique ? D'autres corpus de forum et de chats doivent donc être étudiés de près, afin de poursuivre les comparaisons et d'approfondir la recherche.

Le repérage et l'extraction automatisés sont utiles dans certains cas, et il serait intéressant d'approfondir la recherche sur d'autres formes que celles étudiées ici[15]. Par exemple, il pourrait être utile de mieux cerner le contexte droit du mot glosateur (en apposition, par exemple). Cependant, dans le cas où la glose n'est pas repérable automatiquement dans des séquences de forme : « N/V/Adj (, il est difficile de songer à un traitement pertinent par machine. Dans le corpus courriel, le nombre total de formes parenthétiques est de 23 565 (dont 19 894 N (82,9%), 1 235 V (5,14%), 2 853 Adj (11,8%)).

Que de pistes pour des recherches ultérieures ! Enfin, et cela va dans le sens de notre projet de « communication médiée par ordinateur », il faut essayer de mieux comprendre quelles sont les marques linguistiques et extra-linguistiques qui caractérisent la glose, et quelles sont les situations de communication dans lesquelles se développe l'utilisation de la glose. Puis, il ne faut pas perdre de vue que la glose demeure fondamentalement « du ressort d'une interprétation », et l'interprétation est une spécialité humaine et non machinale ! :

> « Si les gloses de spécification du sens donnent accès à des représentations de la sémantique lexicale, elles ne peuvent être lues que dans la perspective d'une sémantique textuelle permettant l'assignation d'un sens à des formes certes stipulatrices du sens, mais qui ne font que l'orienter, et demeurent fondamentalement du ressort d'une interprétation ». [Julia, 2001 : p. 278]

---

[15] Je remercie Agnès Steuckardt, Aïno Niklas-Salminen et les autres participants du séminaire de leurs remarques et critiques lors de ma présentation du 28 mars 2002.

Références bibliographiques

Authier-Revuz J. (1994), «□L'énonciateur glosateur de ses mots□ explictiation et interprétation□, *Langue française*, n° 10□□

Crystal D. (1997),*The Cambridge Encyclopedia of Language. Second Edition*, Cambridge University Press.

David S. (1993), *Les unités nominales polylexicales. Éléments de description et reconnaissance automatique,* Thèse de doctorat, Université Paris 7.

Gadet F.(1996),□ Une distinction bien fragile□ oral/écrit□, TRANEL, 25, p. 13-27.

Habert B., Nazarenko A., Salem A. (1997), *Les linguistiques de corpus,* Paris, Armand Colin.

Haliday M.A.K. (1989) *Spoken and written language,* Oxford University Press.

Julia C. (2001), *Fixer le sens□ La sémantique spontanée des gloses de spécification du sens*, Paris□Presses de la Sorbonne Nouvelle.

Lawler J., Aristar Dry H., (ed., 1998), *Using Computers in Linguistics. A Practical Guide*, London, New York,Routledge.

Milner J.-C. (1989), *Introduction à une science du langage*, Paris□ des Travaux/Seuil.

Panckhurst R. (1994), «□A database for linguists□ intelligent querying and increase of data□, *Computing and the Humanities* (*CHUM*), vol. 28, n° 1, p.39-52.

Souchard M. *et al.*(1997), *Le Pen. Les mots. Analyse d'un discours d'extrême-droite,* Paris□Éditions Le Monde.

ANNEXE

(1) N/V/A au sens(catgramm=N) (14)

1. En collaboration avec des partenaires internationaux, vous intervenez dans le domaine des télécoms et des nouvelles ***technologies au sens large*** architecture, déploiement de réseau, planification, test de validation et d'intégration. (Internet Actu, jeudi 8 juin 2000)
2. Les organisations intéressées devront notamment apporter la preuve de leur expérience dans l'organisation de grandes campagnes de communication, dans l'utilisation de médias en ligne, dans les relations ***publiques au sens large***, ainsi que dans la conception, la gestion, la maintenance et l'animation de sites Internet multilingues. (Appel d'offre ministériel, Campus numérique, 1 juillet 2000)
3. Ni [X] ni moi n'avons la moindre ***compétence (au sens*** juridique du terme*)* (direction de l'université, envoi à plusieurs personnes, 7 octobre 2000)
4. Il nous faut exiger une véritable négociation, de nature politique, sur nos projets d'établissement (enseignement, recherche et autres missions), l'***évaluation au sens*** traditionnel du terme devant plus porter sur le bilan que sur les projets. (direction de l'université, envoi à la liste des présidents d'université, 12 décembre 2000)
5. Tout ceci est ***hypothétique au sens*** fort du terme (Rapport d'une commission de l'université, 5 mars 2001)
6. Une enquête approfondie a montré que le Recteur n'avait pas donné d'instructions à l'Agent Comptable (il n'y a donc pas pour le moment de ***« tutelle » au sens*** juridique du terme). (message syndical, 9 mars 2001).
7. Pour moi, dans médiatisation, il y a surtout cette notion d'intermédiaire, ou bien, le « processu*s* ***créateur*** *au sens* philosophique*,* « par lequel on passe d'un terme initial à un terme final ». (message envoyé à un collègue — je m'étais envoyé le message en double — commentaires d'un passage de sa thèse, 12 mars 2001)
8. L'équipement à prévoir serait :

   > un système de sonorisation (avec amplis, enceintes et table de mixage); veiller SVP à ce qu'elles ne cassent pas le*s* ***oreilles au sens*** propre de nos étudiants. (Rapport d'une commission, 12 mars 2001)

9. La Commission a adopté un projet de communication concernant les accords d'importance mineure qui ne restreignent pas sensiblement le jeu de la ***concurrence au sens*** de l'article 81, paragraphe 1, du traité CE ("communication de minimis"). (message commercial, 22 mai 2001)
10. Un autre axe serait d'organiser ce travail de ***production***, ***au sens*** d'industrialisation de la formation, c'est à dire pour chiffrer, evaluer et prevoir l'investissement necessaire, sur une annee ou un diplome par ex. (message d'organisation pour un encadrement de stagiaire, 28 juin 2001)
11. L'évaluation de ces demandes, qui sont considérées comme des mesures d'***accompagnement au sens*** du programme de travail IST 2000, suit la procédure décrite dans le manuel de procédures pour l'évaluation des propositions (message ministériel, 12 juillet 2001)
12. Tout enseignant de l'UPV est rattaché à une composante (département ou section, voire service comme le SUFCO, les R.I. ou **l'IEFE),** ***au sens*** employé dans nos calculs de potentiels et charges. (l-chens, 31 octobre 2001)
13. L'université se veut ***archéologique au sens*** le plus large du terme. (l-chens, 24 novembre, 2001)
14. Quel que soit le moyen employé, le texte, le ***document, au sens*** large, connaît aujourd'hui une modification de sa définition. (description d'un projet de recherche – renvoi avec corrections par un collègue, 31 décembre 2001).

(2) N/V/A (Ponct) prep=dans DET sens X (6)

1. *les relations* ***langage/métalangage/sous-langage, dans le sens*** *de la théorie de Z. Harris* (LN, 19 juin 2000)

2. *Penser la construction démocratique de l'université internationale, par la mobilité physique et* ***mentale, dans le sens*** *non seulement de la tolérance, attitude somme toute* ***frileuse, mais dans le sens*** *du goût de l'altérité culturelle et intellectuelle*. (l-chens, l-iatos, 6 juillet 2000) (**2 cas dans une phrase ici**)

3. Le mieux est que, quand tu as un moment, tu téléphones pour voir si je suis ***libre (dans tous les sens..***). (message privé, 25 septembre 2000)

4. "Bjork" s'affiche des la couverture comme un livre ***objet, dans le sen****s fetiche*. (chapitre.com, 3 septembre 2001)

5. Cette didactique moderne des langues reste largement à construire : elle doit être ***fondamentale, dans le sens*** *fort de l'adjectif* […] (Présentation d'une formation doctorale, 11 septembre 2001).

(3) 22 formes en *c'est-à-dire*, corpus courriel.

Positions syntaxique	N° d'occurrences	
SV (UVPL)	1	En tant que nom d'action, l'activité de représentation vise *à « mettre sous les yeux », c'est-à-dire à actualiser des représentations*, envisagées cette fois comme des comportements langagiers stabilisés, et stockés en mémoire, autant d'actualisations* potentielles, qui vont être renégociées dans l'intersubjectivité de la parole.
SN (sans DET)	2	La stylistique est une branche de la linguistique dont l'objet est l'étude scientifique des faits linguistiques (syntaxe, lexique, rythme, etc.) en tant que *style, c'est-à-dire pratique linguistique individuelle* et/ou sociale de production de sens, le plus souvent littéraire. Option 2 : *traitement de texte avancé càd opérations élémentaires sur les textes (recherche, filtrage, tri, transformation)* […].
SN (avec DET)	12	Son objet est *l'analyse de «□expression de l'affectivité »* dans la parole (*c'est-à-dire*, dans la terminologie de Bally, *son contenu expressif*) […]. L'objet de la stylistique, ainsi redéfinie, n'est pas nécessairement le texte littéraire, mais *toute production textuelle, c'est-à-dire toute parole textualisée.* Pour un ecue, l'étudiant a le choix de prendre ou non *les supports pour cet ecue (cad les K7 audio ou video).* Il permet de changer *le «□theme□* (alias «□skin□, *c'est-à-dire, l'apparence*) du browser OpenSource en Wysiwyg.
SN + *qui*	1	Du point de vue de la praxématique, *une analyse stylistique bien comprise*, *c'est-à-dire qui met en relation la production textuelle, le sujet producteur, le réel* […].
SN + *que*	2	Vous verrez que notre site web -intranet est «□n *chantier□, c'est-à-dire que beaucoup travaillent à son amélioration.*
SPrép	3	La démission de [X], qui est des nôtres, […] ne peut être traitée seulement en termes de crise de confiance, voire de mouvement d'humeur, mais doit être rapportée *aux «□rofondes divergerences,* tant sur la méthode de travail que sur les objectifs□, comme le précise [X] dans sa lettre de démission, *c'est-à-dire à une logique d'échec*.
PP en pos. Adj.	1	Parmi les différentes composantes Grundvig, deux sont des actions *«□entralisées□, c'est à dire gérées* dans leur totalité par la Commission européenne […].