

Augusta Mela, Université Montpellier III

Résumé

Linguistes et « talistes » peuvent coopérer : repérage et analyse des gloses

Cet article s'inscrit dans le cadre d'un projet collectif de recherche linguistique sur le mot et sa glose. Comme les définitions, les gloses recueillent « l'expérience parlée du sens ». Cette activité métalinguistique se manifeste dans les textes par des mots tels que *c'est-à-dire*, *ou*, *signifier* qui signent la relation de sémantique lexicale mise en jeu : équivalence avec *c'est-à-dire*, *ou* ; spécification du sens avec *au sens* ; nomination avec *dit*, *baptisé* ; hyponymie avec *en particulier*, *comme* ; hyperonymie avec *et/ou autre(s)*, etc.

Grâce à l'existence de ces marques et des particularités de leurs configurations, un repérage automatique des gloses est envisageable. J'en décris ici la mise en œuvre, en prenant l'exemple des gloses en *ou* telles que « un magazine électronique, *ou webzine* » et d'un environnement informatique « pour linguistes », à savoir la base textuelle Frantext et son interpréteur de langage de requête Stella.

Abstract

Theoretical linguists and computational linguists can work together: finding and analysing the glosses.

This paper is related with a collective linguistic research project about the word and its gloss. Just like definitions, glosses catch “the spoken experience of the meaning”. In French texts, this metalinguistic activity appears in words such as *c'est-à-dire*, *ou*, *signifier*. These signs can clarify the nature of the semantical relationship between two words: specification with *au sens*, equivalency with *ou*, *c'est-à-dire*, nomination with *dit*, *baptisé*, hyponymy with *en particulier*, *comme*, hyperonymy with *et/ou autre(s)*, etc.

Glosses can be automatically located because of both their marks and the features of their configurations. This paper describes an automatical retriever implementation, using “*ou glosses*” as “un magazine électronique, *ou webzine*” and a data-processing environment “for linguists”, namely the textual base Frantext and its Stella query language interpreter.

Linguistes et « talistes¹ » peuvent coopérer : repérage et analyse des gloses

Augusta Mela
Université Montpellier III

Introduction²

Qu'il s'agisse d'évoquer une langue autre, étrangère ou spécialisée (1), de procéder à une explication didactique (2, 3), ou de s'assurer que l'interlocuteur attribue la signification adéquate au mot employé (4), le langage courant fournit de nombreux exemples de commentaires en situation parenthétique qui traduisent, expliquent, régulent le sens des mots en discours.

- (1) Si l'on admet que l'état d'un électron n'est pas entièrement décrit par sa position et sa vitesse de translation dans l'espace, mais qu'il est animé en outre d'un pivotement sur lui-même, ou « *spin* », mouvement essentiellement quantifié : son moment cinétique est d'un demi quantum et crée un moment magnétique égal à un magnéton de Bohr. (Hist. gen. sciences, 1964, t.3, vol.2)³
- (2) Ce sont la dépigmentation, *c'est-à-dire l'absence quasi totale des éléments colorés dermiques qui s'opposent normalement à l'action nocive des rayons ultra-violets d'origine solaire* et l'anophtalmie, ou *réduction de l'appareil oculaire allant le plus souvent jusqu'à sa disparition complète*. (Geze, La spéléologie scientifique, 1965)
- (3) C'est ainsi qu'on parle de normalisation, *non plus au sens que la planification terminologique donne au mot, mais au sens où la communauté d'experts « entérine » des signifiés comme des termes du domaine*. (Bourigault et Slodzian, 1999).
- (4) La « dévotion », *au sens propre du mot, c'est-à-dire un ensemble de pratiques pieuses destinées à honorer l'enfance de Jésus*, ne prendra corps, ne s'organisera qu'après la mort de Bérulle, et sans beaucoup intéresser Condren lui-même. (Bremond, Hist. litt. sent. relig. t. 3, 1921)

Ce que ces trois situations ont en commun, note Authier-Revuz (1995, 12) :

« c'est que le mécanisme communicationnel y est affecté d'un facteur d'hétérogénéité, de non-un : celui de la pluralité des langues, celui de l'inégalité des stades d'appropriation du système, et celui de la différence idio/socio-lectale entre les deux pôles de l'interlocution et que le « recours au métalangage » y est saisi comme une réponse, et même comme un remède à un problème ou un dysfonctionnement de communication. »

À la suite d'Authier-Revuz (1994, 1995) nous appelons *gloses* ces commentaires. Comme les définitions, les gloses recueillent « l'expérience parlée du sens ». Elles constituent un poste idéal d'accès au sens, via ce que le langage dit du mot. Les nombreux travaux dont elles ont fait l'objet en témoignent. Leurs auteurs relèvent des marques de glose telles que *c'est-à-dire*, *à savoir*, *ou*, *signifier* etc., mais les jugent non discriminantes : « l'étude minutieuse de quelques-uns de ces marqueurs nous a rapidement révélé qu'aucun d'eux ne servait exclusivement à la reformulation du sens » concluent Riegel et Tamba (1987, 3). La polysémie, « la multiplicité de leurs « valeurs » est trop connue pour qu'on songe à les identifier à des marques spécifiques de reformulation » (ibid.).

¹ Ce mot est défini en note 4.

² Je remercie Agnès Steuckardt, Sarah Leroy, Benoît Habert, Jean-Marc Sarale ainsi que deux relecteurs anonymes de RFLA pour leurs remarques et critiques.

³ Sauf mention contraire, les exemples cités dans cet article sont extraits de la base textuelle Frantext.

Je montre ici à l'aide de l'outil informatique que ces marques sont opérationnelles si tant est que l'on tire parti des propriétés syntaxico-sémantiques des constructions auxquelles elles sont régulièrement associées.

Ce travail s'inscrit dans un projet collectif de recherche linguistique sur le mot et sa glose⁴ (Steuckardt et Niklas-Salminen (éd.), 2003). La partie I présente les visées du projet. La partie II trace la trajectoire de la réflexion sur la glose, entre définition et reformulation, de Jakobson (1963) à l'actuel projet. La partie III montre, sur l'exemple des gloses en *ou*, la mise en œuvre du repérage automatique dans l'environnement de la base Frantext. La partie IV fait le bilan de l'apport de l'outil informatique pour l'étude linguistique alors que la partie V conclut sur la nécessité d'une synergie entre linguistes et talistes.

Un repérage automatique des gloses est envisageable aujourd'hui

Le sens d'un mot se définit en contexte, par un processus d'intégration sémantique. Par exemple, l'adjectif ou la relative déterminent le sens du mot qu'ils accompagnent mais de façon secondaire, non déclarée. Ce qui caractérise les gloses et les définitions, c'est qu'elles le font de façon déclarée, en parlant du mot. Cette activité métalinguistique se manifeste notamment dans le discours écrit par des marques typographiques comme les guillemets et par des « relateurs »⁵ comme *appelé, c'est-à-dire, signifiant, ou*. Dans ce cas, c'est non seulement la fonction métalinguistique qui est marquée mais aussi la nature même de la relation de sémantique lexicale mise en jeu : équivalence avec *c'est-à-dire, ou* ; spécification du sens avec *au sens* ; équivalence avec *ou, c'est-à-dire* ; nomination avec *dit, baptisé*, hyponymie avec *en particulier, comme*, hyperonymie avec *et/ou autre(s)*, etc. Les relateurs peuvent se combiner à d'autres indices métalinguistiques : des lexèmes comme *terme, mot*, des marques typographiques comme la virgule, la parenthèse, le tiret, les guillemets, à l'écrit, une intonation spécifique ou un geste mimant les guillemets, à l'oral.

Ces marques nous intéressent à un double titre : en tant qu'elles pointent l'explicitation de sens dans les textes et en tant qu'elles explicitent la relation sémantique lexicale mise en jeu.

Grâce à l'existence de ces marques et à la disponibilité d'outils informatiques (corpus numérisés et fonctionnalités de recherche afférentes) un repérage automatique des gloses⁶ peut être envisagé aujourd'hui.

L'outillage informatique optimise les études de la glose sur deux plans :

⁴ Le projet a été lancé par Agnès Steuckardt, Groupe « Sémantique lexicale et discursive », département de Langue française, Université de Provence, Aix-Marseille I. Je participe au projet en tant qu'informaticienne spécialisée dans le domaine du TAL (Traitement Automatique des Langues), ou « taliste ». Les premiers travaux sont publiés (Steuckardt et Niklas-Salminen, 2003). Un second ouvrage sur les marqueurs de glose est en cours de réalisation.

⁵ Le terme est de Tamba-Mecz (1994).

⁶ Les définitions sans marques lexicales, simple juxtaposition ou glose intraphrastique en apposition comme : « Le maïs transgénique, quant à lui, est aussi un hybride résistant à la pyrale, *papillon dont la chenille détruit la tige* » <<http://www.diabetenet.com/L-les-plantes-transgeniques.htm>> sont hors de mon propos ici.

- sur le plan de la linguistique descriptive et exploratoire, le repérage automatique pointe de nouveaux exemples, permet de systématiser la vérification d'hypothèses, induites d'observations « à l'œil nu » et de quantifier des faits ;
- sur le plan de la linguistique appliquée, il sert la terminologie et la néologie en pointant les termes d'un domaine ou les emprunts non codifiés ; il sert la lexicologie synchronique et diachronique⁷ en pointant les différentes acceptions d'un mot dans les textes contemporains ou anciens.

Le groupe de travail « Sémantique lexicale et discursive » s'intéresse aux deux aspects, comme en témoigne l'argument de la première journée d'études, en septembre 2001 :

« La glose introduit dans le discours une explicitation du sens que le locuteur donne au mot qu'il emploie. Commentaire métalinguistique qui traite le mot glosé en autonome ou fugace apposition qualifiante, elle développe une virtualité sémantique du mot. Le lexicographe voit en elle un révélateur ; l'analyste de discours cherche à comprendre les effets de sens qu'elle produit. La journée d'études conjuguera les perspectives des uns et des autres pour s'efforcer d'établir une typologie des gloses et de leur fonctionnement en discours. Elle s'intéressera plus particulièrement aux nouvelles lectures de contextes que permet le traitement automatique de corpus. »

II. La trajectoire de la réflexion sur la glose

Jakobson (1963), le premier, attire l'attention sur la fonction métalinguistique du langage en signalant ses manifestations dans ce qu'il nomme « propositions équationnelles » ; il déclare qu'elles permettent d'accéder au sens de façon *objective* et donne la méthode d'investigation à suivre :

« Est-il dit et dans quels contextes que « A est B », que « B est A » ? Dans la mesure où de telles propositions (équationnelles ici) peuvent être soumises à l'analyse distributionnelle, cette technique s'avère parfaitement applicable aux problèmes de sens tant sur le plan du lexique que sur le plan de la grammaire et il n'est plus permis de considérer les significations comme des « impondérables subjectifs » (ibid., 204).

Le relateur linguistique *est* est assimilé à l'opérateur mathématique « = » et les définitions naturelles à des formules qui rendent deux quantités égales, les équations.

Dans la lignée du programme tracé par Jakobson, Rey-Debove (1985) étudie le métalangage, c'est-à-dire les traces de cette activité métalinguistique, dans les définitions des dictionnaires. Notons que si les définitions des entrées de dictionnaires sont bien des reformulations d'un mot en d'autres mots, il ne s'agit pas de définitions en discours. Fuchs (1982) aborde la reformulation en tant que dispositif expérimental de découverte du sens⁸, « conduite métalinguistique d'identification des sémantismes » dont la reformulation en discours n'est qu'un cas particulier : « il arrive que la reformulation paraphrastique⁹ laisse des traces explicites dans le discours [...]

⁷ Marcello-Nizia (1995, 33) indique l'importance du recours à un contexte explicite pour l'étude des textes dont il n'existe plus de locuteur natif : « on s'efforce d'explicitier le sens en utilisant les indices ou paraphrases donnés par le texte lui-même ».

⁸ Voir (Fuchs, 1982) pour une présentation historique et épistémologique de la paraphrase en tant que dispositif de découverte en linguistique.

⁹ En fait, les huit exemples de (Fuchs, 1982, 32) ne sont pas des reformulations de phrase mais des définitions de mot ou des gloses.

l'étude de telles marques est en soi, intéressante, et éclaire sous un jour particulier le paraphrasage.¹⁰ »

Tamba-Mecz (1994) se recentre sur le discours et élargit le champ d'étude de la définition aux gloses explicatives et de traduction. Il ne s'agit plus seulement d'*équation* ou *identification* mais de *relateurs* linguistiques diversifiés tels que *signifie*, *c'est*, *ou*, *c'est-à-dire*. Comprendre, en effet, c'est mettre en relation, rattacher un concept à un autre. En formulant ces relations, gloses et définitions sont les dépositaires du sens, les « formulaires de l'expérience parlée du sens » :

« La reconnaissance du sens — malaisée d'un point de vue théorique — est chose simple dans la pratique quotidienne du langage, où l'on recourt à ce que G. Granger appelle l'expérience parlée (cf. *Pensée formelle et sciences de l'homme*, Paris, Aubier-Montaigne, 1967, p.32). C'est-à-dire qu'on déclare, à propos d'un objet spécifié, qu'il signifie quelque chose ou rien. On dispose dans les langues d'expressions qui servent précisément à dire l'existence ou l'absence de relation sémantique entre des termes. C'est grâce à de telles expressions que les significations se fixent conventionnellement et, en s'exposant, se prêtent à l'apprentissage et à la réflexion analytique. » (Tamba-Mecz, 1994, 45)

La comparaison de ces opérateurs linguistiques aux opérateurs mathématiques « = », « ⊂ », « ∈ » permet d'en situer les spécificités. Le champ s'étend enfin aux gloses de spécification de sens (3) avec Authier-Revuz (1994, 1995), qui en étudie le fonctionnement sémiotique. À sa suite, Julia (2001) recherche dans les gloses de spécification de sens (3, 4) « les représentations et les conceptions spontanées du sens lexical, c'est-à-dire la composante sémantique de la compétence linguistique ».

La glose en discours dans le champ de la reformulation

À une simplification¹¹ près, le groupe « Sémantique lexicale et discursive » a défini la glose en discours en deux points :

– la glose porte sur le mot (et non sur la phrase ou la proposition) :

« La glose du mot en discours consiste pour le locuteur à apporter un éclairage sur le sens qu'il donne à ce mot. Cette acception de glose se situe dans la continuité du trajet sémantique que suit ce mot. Glossa en effet, qui signifie à l'origine « mot difficile », en vient, par basculement métonymique, à désigner le commentaire explicatif sur un mot difficile, voire le recueil rassemblant les commentaires de glossa, le glossaire. Le Moyen Âge a fait de ce type de textes un genre à part entière, dans lequel les érudits rivalisaient à l'envi. » (Steuckardt, 2003, 9)

– au delà des définitions proprement dites, c'est-à-dire des énoncés dont le propos même est la définition (5), la glose peut être une périphrase définitionnelle intraphrastique et parenthétique (6) :

(5) On appelle script un programme exécuté par le serveur. (exemple construit)

¹⁰ Elle ajoute que ces marques n'ont retenu l'attention des linguistes et théoriciens des discours que récemment, et cite les travaux de Rey-Debove et de Kohler-Chesny.

¹¹ Je simplifie ici en assimilant glose et reformulation alors que les deux phénomènes ne se recouvrent pas totalement : « Il peut y avoir reformulation sans glose. Il en va ainsi lorsque la reformulation est une correction, annulant une première formulation, imparfaite ; le mot de départ, loin d'être expliqué par le suivant, est comme raturé, abandonné. Inversement, il peut y avoir glose sans reformulation. Ainsi dans cette « glose de spécification du sens », relevée par C. Julia : Il n'y a plus, au sens noble du terme, une nation pour l'avenir. [G. Clémenceau], nulle reformulation ne vient préciser en quoi consiste exactement le « sens noble » de nation. » (Steuckardt, 2003, 10)

(6) Pour élaborer ces écrans et disposer les objets texte et graphiques, point n'est besoin de programmer : la création de sous-programmes — *baptisés scripts* — est transparente. (Cité par H.Beciri (Steuckardt et Niklas-Salminen, 2003, 30))

On peut donc considérer la glose ainsi définie comme un sous-ensemble des reformulations en discours parmi lesquelles on distinguera :

- les reformulations de phrases ;
- les reformulations de mots :
 - par des définitions principales ;
 - par des gloses, c'est-à-dire, dans le contexte du projet aixois, des définitions parenthétiques.

Le questionnement sur les marques est au centre du projet. Il est à la fois le moyen du repérage automatique et l'agent catalyseur de réflexions qui englobent la forme et la fonction de la glose en discours : la forme, car si les marques ne sont pas en soi spécifiques des gloses, nous sommes amenés à préciser les propriétés syntaxico-sémantiques des constructions auxquelles elles sont associées ; la fonction, car il s'agit non seulement de chercher les marques de glose mais également d'interroger la correspondance entre classes de marques et rôles des gloses dans le discours. Ainsi, est-ce que les marques *au sens, dans l'acception, littéralement, à proprement parler* signent la spécification du sens ? *ou, qu'on appelle, baptisé*, la nomination ? *en particulier, par exemple*, l'hyponymie ? *et/ou autre(s)*, l'hyperonymie ?

De par leur champ et leur approche, ces études sont complémentaires de certaines études menées ailleurs telles que celle de « *la forme et fonction de la définition en discours* » par Rebeyrolles (2000).

III. Le repérage automatique

Si la glose indique le sens, qu'est-ce qui indique que l'on est devant une glose ? Les premières études linguistiques ont relevé des traces de glose. Ces marques à elles seules ne permettent pas de repérer automatiquement les gloses, mais en examinant leurs configurations, on induit des premiers schémas de description. Il s'agit ensuite de les confronter aux corpus numérisés et de les raffiner au vu des résultats obtenus. Je décris ici la mise en œuvre, de l'accumulation d'indices jusqu'au repérage automatique, en prenant l'exemple des gloses en *ou*, et d'environnements informatiques « pour linguistes ».

Étude de cas : des marques au repérage des gloses en *ou*

Ou n'est pas à lui tout seul une marque de reformulation de sens : il peut apparaître dans un tour du type « des mouvements sans contact, *ou télékinésies* »¹² ou dans une banale disjonction : « une

¹² L'énoncé complet est le suivant : « La question des tables tournantes n'est qu'un des aspects du vaste problème des mouvements sans contact, *ou télékinésies*, déjà décrits dans les maisons hantées ou par les radiesthésistes, que revendiqueront les grands " médiums " à effets physiques » (Amadou, La parapsychologie, 1954)

poire *ou* une pomme ». Cependant une recherche automatique des gloses en *ou* peut être envisagée si l'on tire parti des propriétés syntaxico-sémantiques suivantes¹³ :

- i) dans les tours de ce type, « le terme en mention s'applique métalinguistiquement au premier » (Tamba, 1987, 27). L'application métalinguistique se manifeste par la structure parenthétique du terme, marquée à l'écrit par la ponctuation : virgule le plus souvent, mais aussi tiret ou parenthèse ouvrante.
- ii) le substantif qui suit *ou* est non déterminé. La non répétition du déterminant dans le schéma : « *Déterminant X ou Y* » signale que X et Y désignent la même chose.

Lorsque le « motif » à rechercher est une simple expression textuelle comme « *ou* », « *, ou* », « *, ou plutôt* », la fonction Recherche simple, présente sur la plupart des logiciels convient¹⁴. Pour rechercher en une seule passe, les occurrences de *ou* précédées d'une ponctuation, c'est-à-dire le motif :

(virgule|parenthèse ouvrante|tiret) *ou*

on recourt au langage des *expressions régulières*¹⁵. Ce langage, issu de l'informatique théorique, permet de décrire en intension des motifs recouvrant des réalisations variées. Son vocabulaire plus ou moins étendu lui confère une expressivité variable ; sa syntaxe admet également des variantes (opérateurs pré- ou post-posés, etc.). À ces variations près, ce langage est « compris »¹⁶ par la fonction Recherche de nombreux logiciels. Les caractères génériques proposés dans la fonction de recherche avancée du logiciel Word en sont un avatar. Sous Word, la recherche du motif précédent codé « *[, \ (-) ou* »¹⁷ permet un premier repérage, basé sur la seule propriété i).

Faute de fonction Concordancier, la fonction Remplacer avec caractères génériques de Word est utilisée pour surligner des portions de texte dans le code du document¹⁸. La figure 1 montre comment obtenir le surlignage des gloses introduites par « *, ou* »¹⁹.

¹³ La ponctuation dans ces tours est signalée dans la grammaire de Wagner et Pinchon (Tamba, 1987, 17) ; la propriété ii) est remarquée par Tamba (1987,17).

¹⁴ Sous Word, la fonction Rechercher surligne les occurrences du motif recherché mais n'extrait pas de concordance. Un filtrage est nécessaire (Voir Habert et al., 1998, 23-24). Des éditeurs (gratuits) comme BBEEdit pour MacOS <www.barebones.com> permettent une telle extraction des concordances.

¹⁵ « L'adjectif régulier est employé ici au sens *qui obéit à des règles*. Les expressions régulières sont en effet un outil descriptif obéissant à des règles précises qui sont une série de conventions de notation et de principes de construction servant à décrire abstraitement des éléments textuels » (Desgraupes, 2001, XIII). On attribue la paternité du langage des expressions régulières au mathématicien Stephen Kleene (1909-1994).

¹⁶ On dit « supporté », dans le domaine informatique.

¹⁷ « *[]* » définit une classe de caractères. Le caractère « ** » précédant la parenthèse ouvrante la « déspecialise » : elle est alors interprétée comme simple caractère du texte, au même titre que la virgule et le tiret, et non plus comme caractère du métalangage.

¹⁸ Le surlignage n'est plus seulement concomitant à la fonction Recherche : il est inscrit dans le code du document via l'annotation de mise en forme correspondante qui sera conservée dans le fichier.

¹⁹ Le caractère « *!* » est le caractère d'exclusion. Le motif « *, ou(!, ; \ ?){1 ;}* » signifie donc « *, ou* suivi de caractères autres qu'une ponctuation », c'est-à-dire « *, ou* suivi d'une série de mots terminée par une ponctuation ». Le parenthésage du motif a pour effet de mémoriser dans la variable « *\1* » ce qui est reconnu par le motif. Consulter l'aide en ligne de Word pour une présentation de la Recherche avancée de Word et (Habert et al., 1998, 105-115) pour des applications de cette fonctionnalité.



Figure 1. Word/surlignage de portions de texte décrites par un motif générique

Le résultat obtenu est le suivant :

Les patenostriers, ou **faiseurs de chapelets**, fabriquaient également les dés à jouer. (Allemagne, Récréations et passe-temps, 1904)

La fonction Remplacer permet donc de transformer des portions de texte²⁰. De la même manière que nous avons enrichi les gloses en *ou* de la mise en forme « surlignage », nous pourrions les enrichir d'une annotation structurelle du type :

Les patenostriers [, ou faiseurs de chapelets, *Apposition*] fabriquaient également les dés à jouer. d'une annotation prosodique, ou de toute autre annotation utile pour une recherche ultérieure sur le corpus.

Pour utile qu'il soit dans certaines tâches, le langage de Word montre vite ses limites. Ainsi, la description de variantes telle que « ,ou (encore|pour mieux dire|plutôt) », n'est pas possible.

Tirer parti de la propriété ii) nécessite un corpus annoté grammaticalement²¹ :

Les patenostriers, ou [faiseurs S] de chapelets, fabriquaient également les dés à jouer.

Pour simplifier le motif de recherche, il est alors souhaitable que ce dernier ne soit plus une expression régulière sur les caractères du texte²², mais sur des entités d'un niveau supérieur permettant d'écrire le motif « ,ou S » de la configuration recherchée. Dans ce cas, le vocabulaire du langage de description s'étend à ces entités. Or ces entités dépendent de la stratégie d'étiquetage du corpus (nature, nombre, et choix d'étiquettes), « l'interpréteur » du langage marche en tandem avec l'étiqueteur du corpus. Bien que la technique soit disponible, il n'existe pas de tel outil pour linguistes dans le commerce²³. À ce jour, seule la base textuelle Frantext²⁴

²⁰ Dans le langage informatique, on appelle *transduction régulière* une telle transformation.

²¹ J'annote uniquement le mot *faiseurs* par « S(substantif) », naturellement tous les mots de la phrase seraient annotés de la sorte.

²² En effet l'expression régulière sur les caractères peut se complexifier très vite. Voir Leroy (2001) et ici-même pour un témoignage d'une telle entreprise.

²³ Cela ne signifie pas que ces outils n'existent pas. Voir par exemple l'interpréteur de langage de requête linguistique Yakwa, développé à l'ERSS UMR 5610 par Ludovic Tanguy :<<http://univ->

dans sa version catégorisée, et son interpréteur de langage de requête linguistique Stella propose un tel service. Dans ce langage, le motif « ,ou S » s'écrit « ,ou &e(g=S) »²⁵. La recherche de ce motif dans le tome 3 du volume 2 de « l'histoire générale des sciences », soit 435 385 mots, ramène 38 occurrences, présentées dans leur contexte, sous la forme de page Web :

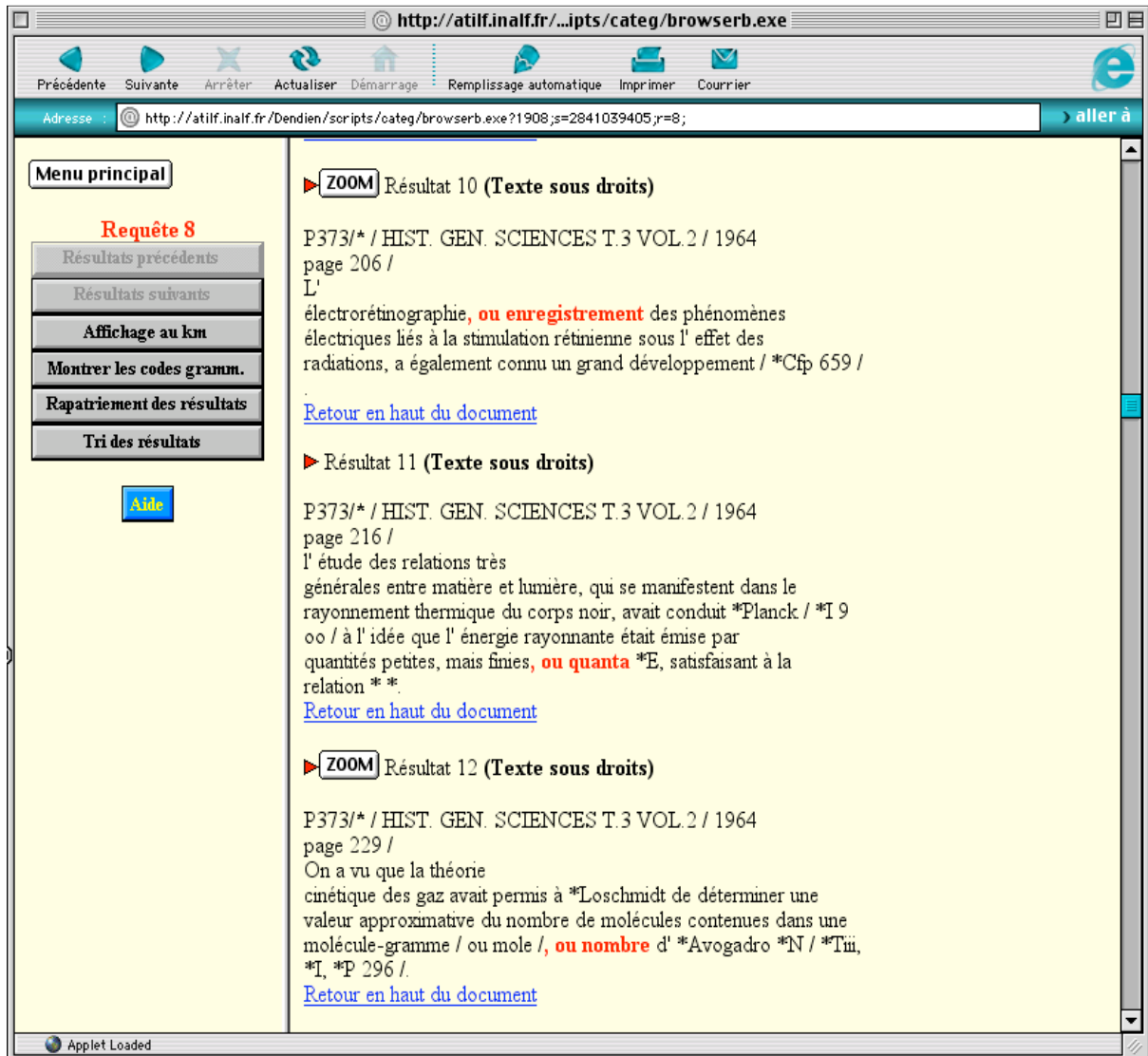


Figure 2. Résultats de la recherche du motif « ,ou &e(g=S) » dans Frantext.

tlse2.fr/erss/membres/tanguy/yakwa.html> qui « interface » l'étiqueteur Cordial Universités de la société Synapse :<http://www.synapse.fr.com>.

²⁴ La base Frantext est accessible par abonnement à l'adresse : <http://www.atilf.inalf.fr/frantext>.

²⁵ « &e(g=S) » signifie entité dont la catégorie est S. Un tutoriel est disponible sur le serveur, à partir du menu, via les liens *À quoi servent les listes/ grammaires*.

Une vue du texte enrichi des annotations morpho-syntaxiques est également possible :

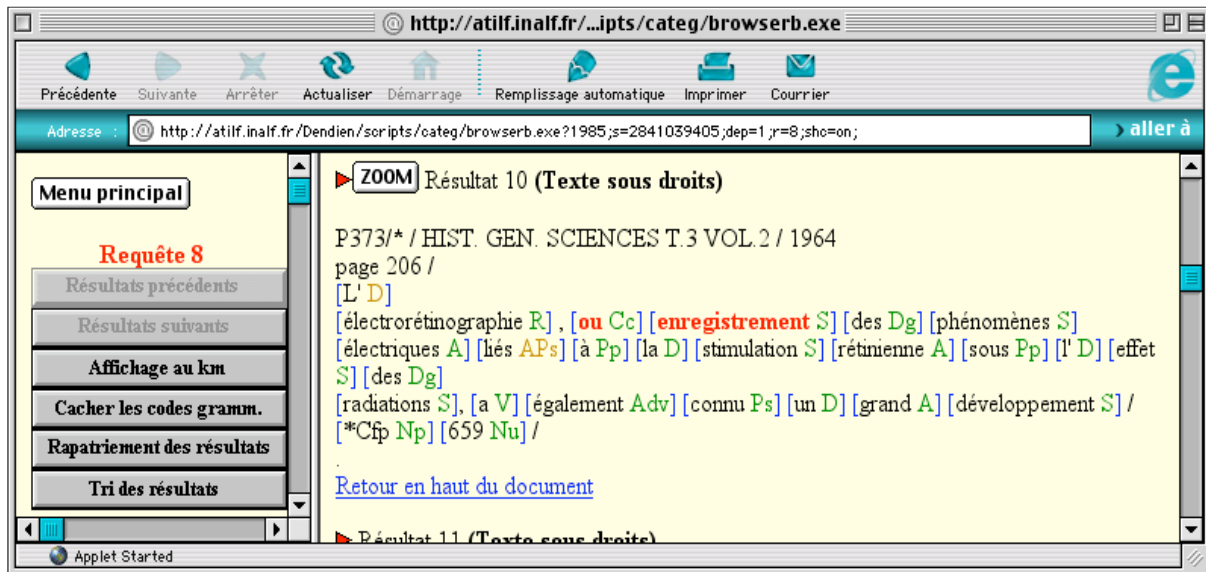


Figure 3. Résultats de la recherche du motif « ,ou &e(g=S) » avec les annotations sous-jacentes.

Les résultats peuvent être « rapatriés », du serveur Frantext vers nos propres machines, au format texte. La fonction Surlignage de Word est alors utile pour visualiser les gloses. 36 résultats sur 38 sont des gloses. L'examen de ces gloses confirme la description du rôle de la glose en *ou* dans le discours consignée dans la conclusion de Tamba (1987, 27-28). Lorsque *ou* joint un terme en usage à un terme en mention, le terme en mention s'applique métalinguistiquement au premier ; il est co-possible. Soit il sert à expliciter le mot glosé peu connu (8) soit inversement, dans une perspective de « prévision didactique » c'est le mot glosé qui sert à expliquer au préalable le terme nouveau introduit dans la glose (7) .

- (7) L'étude des relations très générales entre matière et lumière, qui se manifestent dans le rayonnement thermique du corps noir, avait conduit Planck à l'idée que l'énergie rayonnante était émise par quantités petites, mais finies, *ou quanta E*, satisfaisant à la relation.
- (8) L'électrorétinographie, *ou enregistrement des phénomènes électriques liés à la stimulation rétinienne sous l'effet des radiations*, a également connu un grand développement.

Pour les deux énoncés non pertinents, la cause de « bruit »²⁶ est la même : il s'agit dans les deux cas de coordination multiple de substantifs non déterminés, liés par *ou* :

- (9) Tous ces phénomènes sont caractérisés par un couplage entre d'une part le milieu ambiant/ vivant ou non/ et, d'autre part, soit un individu/ animal, végétal, *ou protiste*, soit une partie de son organisme, soit une association de plusieurs individus.

On peut remédier à cette cause de bruit. Des techniques d'analyse syntaxique partielle peuvent annoter la structure coordonnée de telle sorte que chacun de ses éléments soit affecté de son rang.

²⁶ En recherche documentaire, les notions de bruit, silence, rappel, précision sont utilisées pour mesurer la qualité des résultats d'une recherche : le *bruit* est le nombre de réponses non pertinentes/nombre de réponses données ; le *silence* est le nombre de réponses pertinentes non données/nombre de réponses pertinentes existant ; le *rappel* est le nombre de réponses pertinentes données/nombre de réponses pertinentes existant ; la *précision* est le nombre de réponses pertinentes données/nombre de réponses données.

Ces techniques d'annotation automatique seront disponibles dans un futur proche. En attendant, dans l'environnement Intex²⁷, par exemple, on obtient ces annotations par transduction régulière²⁸. Pour éliminer des énoncés comme (9), il suffit alors de spécifier dans le motif de recherche que l'élément susceptible d'être en mention est nécessairement au deuxième rang :

« , ou &e(g=S+coord_rang=2) »

L'examen de corpus plus gros révèle d'autres types de bruit, c'est-à-dire d'énoncés où l'élément introduit par *ou* tout en étant en seconde position, n'est pas une glose. C'est le cas des coordinations « parenthétiques » (10), ou des coordinations à distance telles que (11) :

(10) Elle ne s'applique plus, dès qu'il y a action organisante, *ou désorganisation*. (Ruyer, *Cybernétique origine inform.*, 1954)

(11) Biot, en 1815, découvre l'existence de liquides produisant la rotation de la vibration lumineuse qui les traverse : *substances pures*, telles que l'essence de térébenthine ou l'essence de citron, *ou solutions* dans un solvant inactif, tel que l'eau, de substances solides, telles que le sucre de canne ou l'acide tartrique. (Hist. gen. sciences, t.3, vol.1, 1961)

Bien que nous n'ayons pas accès au texte entier de façon linéaire, le silence – c'est-à-dire ici les gloses en *ou* présentes dans le corpus mais non ramenées, soit parce que *ou* y est précédé d'une autre ponctuation que la virgule, soit parce que le mot introduit par *ou* est inconnu du logiciel (cas fréquent dans le contexte des gloses), soit parce qu'il est précédé d'un adjectif – peut être mesuré en lançant des recherches plus larges. Le motif « (, | \ (| - | /) ou &e(g=S) », étendu à d'autres ponctuations, ramène effectivement 18 résultats supplémentaires sans augmenter le bruit.

Le motif « (, | \ (| - | /) ou &e(g=S R²⁹) » étendu aux mots inconnus du logiciel, ramène 12 gloses de plus. Le motif « (, | \ (| - | /) ou &?&e(g=A) &e(g=S R) »³⁰, étendu aux adjectifs postposés, ne ramène aucun résultat supplémentaire à partir du texte cité. On reste à 68 gloses en *ou*.

Projeté sur d'autres textes, il ramène des gloses diverses telles que :

(12) Elle est produite par les cellules epsilon, *ou anciennes cellules acidophiles alpha*, qui la sécrètent principalement au cours de la gestation. (Bariety-Coury, *Hist. de la médecine*, 1963)

(13) Cette préparation aboutit à la rédaction d'un projet de lettre, note, *ou autre document*, qui sera soumis à la signature accompagné du document ayant soulevé la question et de divers documents qui ont pu être cités en référence/ pièces en souche. (Lubrano-Lavadera, *Legisl. et administr. milit.*, 1954)

(14) Le mot lui-même est employé en Picardie, tandis qu'en d'autres provinces françaises on se sert des expressions royon, deuve, rive, croc, fraite, heurt, urée, claux, *ou autres mots* voisins de même famille. (Meynier, *Les paysages agraires*, 1958)

²⁷ Voir paragraphe suivant.

²⁸ La transduction remplace les portions de texte correspondant au motif :

(<S>), (<S>), ou (<S>)

en portions de texte annotées :

[{ \$1, .S+coord_rang=1 }, { \$2, .S+coord_rang=2 }, ou { \$3, .S+coord_rang=3 } Coord]

ce qui signifie en clair : lorsqu'une portion de texte correspond au schéma « S, S, ou S », annoter chacun de ses éléments (mémoisés dans les variables \$1, \$2, \$3) par son rang (coord_rang=i) dans la structure coordonnée. Le résultat obtenu sur l'énoncé (9) serait le suivant :

[{ animal, .S+coord_rang=1 }, { végétal, .S+coord_rang=2 }, ou { protiste, .S+coord_rang=3 } Coord].

²⁹ La catégorie R correspond aux mots inconnus du logiciel.

³⁰ « &? » exprime la nature optionnelle de l'expression qui suit.

Avec (13, 14), on sort de la définition pour aller vers la classification : en (13), *lettre* et *note* appartiennent à la classe des *documents*, en (14), *royon*, *deuve*, etc. sont des *mots de même famille*. Il s'agit donc de gloses d'un type différent. Sinon, on retrouve, avec les adjectifs postposés, les mêmes causes de bruit que précédemment, à savoir la coordination multiple de substantifs non déterminés telle que : *universités*, *grands travaux d'hydraulique*, *ou grandes liaisons*, que l'on pourra éliminer *via* une annotation structurelle et le rang dans la structure comme pour l'énoncé (9), et la coordination « parenthétique » telle que :

(15) Il révèle aussi une magnifique pléiade de savants, hommes admirables de courage et de ténacité, hommes d'épopée, quelquefois sublimes, *ou véritables génies*, souvent singuliers ou pittoresques. (Hist. gen. sciences, t.3 vol.1, 1961)

Au final, si l'on souhaite écarter les gloses de « classification » telles que (13, 14), on interdira l'adjectif *autre*. Le motif de recherche devient alors :

(,|\(|-|/) ou &?&e(g=A c!=³¹&mautre) &e(g=S R)

Le motif se complexifiant, on a intérêt à le mémoriser dans un fichier sous la forme d'une « grammaire »³². Il constitue alors l'objet de description de la grammaire :

objet:
(,|\(|-|/) ou &?&e(g=A c!=&mautre) &e(g=S R)

Contenu du fichier « glose_ou »

Sur le texte cité initialement, on obtient 68 résultats avec un bruit faible (2/68). La bonne précision obtenue (66/68) est bien sûr liée à la nature du texte. Les textes littéraires contiennent davantage de coordinations à distance et la précision des résultats diminue d'autant.

On suit la même démarche avec les autres marqueurs de glose.

Les recherches présentées visaient à repérer les gloses d'un type donné. Elles peuvent également viser à repérer les gloses d'un mot donné. Ainsi la recherche du motif :

« cuirassé &robject,glose_ou³³ »

en ramenant l'énoncé suivant :

En 1939 le cuirassé, *ou navire de ligne*, était encore considéré comme le navire principal, le Capital Ship, d'une flotte de guerre. (Le Masson, La marine, 1951)

permet de pointer l'acception du mot *cuirassé* dans le domaine de la marine.

En étendant la grammaire précédente à toutes les gloses (en *ou*, en *c'est-à-dire*, etc.), la recherche d'acceptions de mots est d'autant plus efficace. Pour repérer les substantifs suivis de leur glose, on projette sur les textes le motif suivant :

« &e(g=S) &robject,glose »

où « &robject » est l'objet décrit par une telle grammaire, nommée ici « glose ».

³¹ « &e(c !=expression) » est une entité n'ayant pas le contenu indiqué par *expression*. « &mX » désigne les différentes formes du lemme X.

³² Il s'agit de grammaire au sens de la théorie des langages : les grammaires de Stella généralisent en fait la notion de motif. Voir le tutoriel sous Frantext.

Au delà de Stella : Intex

L'environnement Intex³⁴ permet de faire des recherches analogues à celles proposées par Stella, sur nos propres corpus. La philosophie d'Intex est sensiblement différente : la désambiguïsation des entités de catégorie morpho-syntaxique ambiguë (comme *le*, etc.) ne recourt pas à des techniques probabilistes comme c'est le cas, même de façon partielle, en annotation automatique, mais à des descriptions linguistiques appelées « grammaires de désambiguïsation locales », dont la mise au point nécessite des études linguistiques approfondies. On ne dispose donc pas d'étiqueteur automatique prêt à l'emploi (encore que de nombreuses grammaires de désambiguïsation locales le sont) mais d'un langage de requête et d'étiquetage puissant pour implémenter ces grammaires.

IV. L'outil informatique renouvelle l'intérêt des études de la glose et du sens

Les éléments du renouvellement sont multiples : l'informatique permet de vérifier l'opérationnalité des marques pressenties, de mesurer leur fréquence dans les différents types de textes mais aussi « d'apprendre » de nouvelles marques de gloses à partir des textes ; elle déplace la notion de « marque linguistique » du niveau lexical vers le niveau syntaxico-sémantique ; elle amplifie les retombées pratiques de l'étude dans les domaines de la terminologie, de la lexicologie, de la typologie des textes ; ces applications pratiques ne manquant pas de rejaillir en linguistique théorique comme l'illustre l'impact des travaux de Rey-Debove (1997) pour ne citer qu'eux. Enfin, elle accompagne et amplifie le virage méthodologique de l'étude du sens à partir des textes. Je commente brièvement ces différents points.

La vérification d'hypothèses

J'ai montré en III comment l'outil informatique vérifie que les marques de gloses sont opérationnelles sous certaines conditions.

La quantification

Toute règle linguistique suppose une appréhension du quantitatif. Dans le discours des linguistes, la quantification se traduit par les termes : à *profusion*, *massivement* :

« Le vocabulaire spécialisé, introduit à *profusion* dans les textes de vulgarisation scientifique, y est, en effet, très rarement l'objet d'une définition homogène du type « on appelle S, x » ou « S est x » dans laquelle x serait construit avec des éléments du même sous système que S. Ce que l'on relève, au contraire, *massivement*, ce sont des suites hétérogènes juxtaposant, à la façon d'un dictionnaire bilingue, deux éléments S et Q posés comme équivalents par une opération locale de traduction. » (Authier-Revuz, 1982, 40-41)

Mais ce type d'observation n'est opératoire que si les résultats sont quantifiés. Dès lors que les repérages automatiques des définitions³⁵ et des gloses sont au point, on peut les compter et comparer leur fréquence sur différents types de textes, ce que le vocabulaire à *profusion*, *massivement* laisse dans le vague.

³³ « &robjet, glose_ou » invoque la règle *objet* du fichier *glose_ou*.

³⁴ Intex est un environnement de développement linguistique gratuit sous certaines conditions d'utilisation et téléchargeable à l'adresse <<http://grelis.univ-fcomte.fr/intex/overview.html>>.

³⁵ Cf. (Rebeyrolles, 2000).

La notion de « marque linguistique » se déplace du niveau lexical vers le niveau syntaxico-sémantique

L'outillage informatique fait évoluer le concept de marque linguistique. Pour les linguistes, qu'elle soit lexicale (*c'est-à-dire, ou*), ou typographique (guillemets), la marque, c'est ce qui existe dans le texte « nu » ; s'ils notent une propriété syntaxico-sémantique comme « l'élément en mention n'est pas déterminé », ils ne possèdent pas le langage pour exprimer ce qui fait de cette propriété une marque, à savoir : « ,ou &e(g=S) » ou « ,ou &e(g!=D) »³⁶. La propriété est indicible au niveau lexical. Il faut passer à un niveau abstrait des catégories grammaticales et savoir qu'une « prise » à ce niveau est possible avec les corpus annotés. Un saut conceptuel supplémentaire sera nécessaire lorsqu'il s'agira de passer des catégories morpho-syntaxiques du mot aux catégories structurelles.

Pour le linguiste outillé, la marque c'est ce qui existe dans le texte annoté, voire ce qui pourra exister dans le futur texte annoté³⁷, lorsque de nouvelles techniques d'annotation de corpus seront opérationnelles³⁸. De plus, dans un contexte applicatif, l'appréciation de la marque est relative : elle dépend du temps de traitement (lié à la complexité du calcul) de celle-ci et de la qualité des résultats mesurée en termes de bruit, silence, etc. Une marque peut ne pas être discriminante à 100%, mais être opératoire pour une tâche donnée. Cela ne signifie pas que les 10% restant soient négligeables ; au contraire, il faudra les analyser de manière à améliorer le rendement si c'est possible et à diagnostiquer les limites de l'opérationnalité de la marque.

L'outil informatique permet de « capitaliser » l'étude de la glose dans les domaines applicatifs

Les perspectives d'application de l'étude de la glose sont nombreuses. Nous en avons mentionné quelques unes, en partie I, dans les domaines de la terminologie, de la lexicologie. La glose peut aussi servir à suivre les différents stades de la codification d'un emprunt, manifestés par un continuum des marques. « Quand l'emprunt n'est plus mis en relation avec le mot français gloseur, l'énoncé est jugé compréhensible par l'encodeur : à ce stade, on estime que l'emprunt est codifié » note Niklas-Salminen (2003, 69). La glose pointe les néologismes, qu'elle reformule par des unités conventionnelles. Il arrive, remarque Sablayrolles (2003, 39), que le néologisme apparaisse dans le relateur même de la glose, comme dans :

« Faute de pouvoir traiter une maladie nouvelle, la médecine lui donne un joli nom. Ainsi a-t-on d'abord parlé de polyalgie diffuse, *ce qui est une façon gréco-administrative de dire j'ai mal partout.* » (Meyer, *Nous vivons une époque moderne*, 1991, cité par Sablayrolles (2003, 39))

Parfois, la glose explicite les étapes de fabrication d'un néologisme au statut morphologique ambigu :

« J'avais l'habitude, *je veux dire* j'habitais l'attitude, *je veux dire* j'habitais de monter long » (Simon, *La route des Flandres*, 1960, cité par Sablayrolles (2003, 39))

D'autres pistes sont à explorer en classification de textes, par superposition des listes des termes pointés par la glose et de celles des terminologies existantes ; en typologie des textes : quelles

³⁶ « &e(g!=D) » est une entité autre qu'un déterminant.

³⁷ Cf. en III, la remarque au sujet de l'énoncé (9).

³⁸ Actuellement, la technique de l'annotation structurelle est encore au stade de prototype. Cf. (Véronis, 2000) pour un bilan des techniques d'annotation automatique de corpus.

formes de glose pour quel type de média³⁹, quel type de situation de communication (vulgarisation, pédagogie) ? et comment les gloses se manifestent suivant le média⁴⁰ ?

Ces retombées « économiques » stimulent en soi l'étude linguistique, et génèrent un « retour sur investissement » en linguistique « pure ». Citons enfin, au titre des retombées de la linguistique appliquée, en l'occurrence du TAL, la possibilité, d'« apprendre » de nouvelles marques de gloses à partir des textes. En effet, plutôt que de mettre au point des motifs de repérage, ce qui demande un temps de mise au point important, la tendance est de chercher à les acquérir à partir des textes. Ainsi, plutôt que de décrire la grammaire de la relation d'hyponymie, le système Prométhée (Morin, 1999), part d'une liste préalable de couples de termes qui vérifient la relation, et cherche dans les textes les énoncés qui contiennent des occurrences de ces couples. À partir de ceux-ci, des descriptions génériques sont induites, qui, à leur tour projetées sur les textes, permettent de recueillir de nouveaux couples. Ces nouveaux couples, une fois qu'ils seront validés par l'expert du domaine, serviront à leur tour d'accroches pour repérer d'autres configurations et ainsi de suite jusqu'à stabilisation du nombre de couples et de configurations. Il faut bien sûr disposer de gros corpus ou de textes suffisamment spécialisés pour que les termes en relation soient mis en scène plusieurs fois et puissent signaler de nouvelles manifestations de leur relation.

L'outil informatique amplifie le virage méthodologique vers la sémantique textuelle

De même qu'en terminologie, il ne s'agit plus pour les experts d'un domaine de décider quels termes doivent être utilisés pour tels concepts, mais de valider comme termes du domaine des signifiés induits à partir de leur usage dans les textes (Bourigault et Slodzian, 1999)⁴¹, de même en matière de langue ordinaire, en dépit de tendances politiques dirigistes, c'est l'usage qui fait loi. Par conséquent, la réalité des mots et de leur sens est à chercher non dans les prescriptions des observatoires de la langue (Délégation à la langue française ou autre) mais dans les textes. Si tant est que le linguiste soit descripteur et non décideur. Or puisque, dans la recherche du sens linguistique, on n'a encore rien fait de mieux que la reformulation, et que les reformulations naturelles sont à notre disposition, et en nombre, dans les corpus, il convient de rechercher le sens dans les textes, et notamment via les gloses et les définitions. En cela, et comme pour la terminologie textuelle, l'outil informatique accompagne, amplifie et systématise le virage méthodologique vers une sémantique lexicale textuelle.

À moyen terme, un moteur de recherche pour linguistes

La disponibilité de corpus est cruciale. Le retard de la France en ce domaine est connu. Plutôt que de chercher à rattraper le retard en constituant des corpus de plus en plus gros, mais qui resteront limités en taille et en datation, la tendance actuelle est d'utiliser le Web comme corpus. De nombreux travaux en linguistique et en TAL exploitent les données du Web. Le nombre de mots en alphabet latin étant estimé en à 2 000 milliards de mots, et la présence du français, à 1,8%, cela

³⁹ Plusieurs d'entre nous ont remarqué l'absence de gloses en poésie. Panckhurst (2003) « remarque un faible nombre de gloses dans le discours électronique (ou « netspeak »). Lorsque les gloses apparaissent, c'est notamment (mais non pas exclusivement) en situation didactique (l'enseignant explique une consigne, etc.). »

⁴⁰ Reboul (2003) remarque que le mot glosé sur le Web prend la forme d'un lien hypertextuel vers la définition de ce mot.

porte le nombre de mots français à 36 milliards de mots (200 fois plus que la base Frantext non catégorisée (~ 217 millions de mots)⁴².

Les moteurs de recherche dans le Web sont très performants mais ils sont conçus pour la recherche documentaire et n'offrent pas les fonctionnalités nécessaires à une recherche linguistique. Le nombre d'occurrences ramenées est limité (2 000 pour Google) et surtout les « poignées » sur les textes offertes par la lemmatisation et les annotations morpho-syntaxiques, ne sont pas accessibles. Le méta-moteur Webcorp⁴³ exploite les résultats des moteurs de recherche documentaire habituels et supporte un langage de description de motif de recherche mais ce langage est rudimentaire et les délais de recherche prohibitifs.

Les linguistes ne constituant pas un lobby suffisamment puissant pour influencer les réalisations informatiques, des linguistes américains (Kilgarriff et Grefensette, 2003) construisent actuellement un moteur de recherche pour linguistes (Linguistic Search Engine). Ce moteur relèvera les textes du Web deux fois par an, les annotera et interprètera un langage de requête linguistique. Le Web comme corpus sera à la disposition des linguistes. À ce stade, l'ordinateur ne sera plus seulement un outil pour les linguistes, ce sera leur élément naturel :

« La linguistique venant à maturité, les méthodes qu'elle utilise deviennent empiriques. Il n'est plus temps de se livrer à l'introspection pour glaner des intuitions linguistiques. Nous avons besoin de données. En linguistique, le processus de maturation est intimement lié au développement de l'ordinateur. Si, en biologie ou en chimie, l'ordinateur est seulement un outil pour ranger et traiter des données, en linguistique, l'objet d'étude lui-même (dans une de ses formes primaires, la seconde étant acoustique) est dans l'ordinateur : c'est le texte, et le disque dur de l'ordinateur en est un support tout aussi valide que le papier ou un autre média » (traduction d'un extrait de (Kilgarriff, 2003, 53))

V. Une coopération entre linguistes et talistes

Si le développement des applications constitue un moteur de la recherche en linguistique théorique, ce moteur est optimisé par l'informatique. Inversement, les applications liées à la linguistique doivent s'appuyer sur une recherche fondamentale. Une coopération linguistes et talistes est souhaitable. Elle peut se faire par le biais des communications écrites mais peu de linguistes lisent des articles de TAL, hermétiques. De même, les informaticiens travaillant sur le « matériau » linguistique peuvent ignorer les questions débattues en sciences du langage depuis des décennies et ... réinventer la roue, voire fonder leur système sur des bases suspectes. Cette situation est spécifique à la linguistique car si un informaticien travaillant en biologie moléculaire doit passer par la case « biologistes », l'illusion de « connaître » la langue peut amener l'informaticien à faire l'économie d'une culture en sciences du langage. C'est ainsi que les promesses non tenues de la traduction automatique des années cinquante ont été sanctionnées par le rapport ALPAC⁴⁴.

Les coopérations linguistes-informaticiens ne sont pas rares en France mais elles ne vont pas de soi. En dehors d'une résistance générale au changement de pratiques, et du problème des jargons

⁴¹ Cf. l'exemple (3) ici-même.

⁴² Ces estimations datent de janvier 2003 (Kilgarriff et Grefensette, 2003, 337).

⁴³ Le méta-moteur WebCorp est accessible en ligne à l'adresse : <<http://webcorp.org.uk/>>.

⁴⁴ Les conclusions du rapport ALPAC (Automatic Language Processing Advisory Committee), en 1966, ont eu pour conséquence la coupure immédiate des budgets consacrés au domaine du TAL et de la TA (traduction automatique) et un ralentissement des recherches.

spécifiques aux disciplines, lots communs à tout rapport interdisciplinaire, les blocages de la communication linguistes-informaticiens sont avérés : du côté des informaticiens, un éventuel désintérêt pour les travaux en sciences du langage, le désir obscur de ne pas partager le pouvoir du savoir informatique, la crainte (parfois fondée) d'être le technicien au service des linguistes ; du côté des linguistes, la peur du formel, de la technocratie, de la perte d'autonomie, la suspicion vis-à-vis du réductionnisme des informaticiens travaillant sur le matériau linguistique, et enfin un manque d'intérêt pour les applications. Il est probable que les linguistes motivés par les applications de leurs travaux sont plus favorables à l'outil informatique. Mais ceux-là même hésitent avant d'investir du temps pour la maîtrise d'outils ou d'un langage de requête linguistique. Et à juste titre. Cela représente un coût pour des résultats dont l'analyse ne leur est pas toujours accessible. En effet, il ne suffit d'apprendre un langage, il faut aussi, pour analyser des résultats inattendus, comprendre comment la machine interprète ce langage. Cela suppose une connaissance des processus qui sont en amont : codage de caractères, stratégies d'étiquetage des textes et en aval : stratégies d'analyse de la requête (incidence des quantificateurs « +,* » dans la recherche des motifs, etc.).

Le changement viendra sans doute des jeunes linguistes mais encore faut-il que leur soient enseignées ces techniques et qu'ils soient encouragés à les intégrer dans leur pratique⁴⁵. Or les résultats d'une enquête récente menée par l'ATALA⁴⁶ (Association pour le Traitement Automatique des Langues : <<http://www.atala.org/>>) sur l'enseignement du TAL en France indiquent les faibles effectifs des formations en TAL. Ce fait est à rapprocher des blocages indiqués ci-dessus. Espérons que les progrès de l'informatique appliquée à la langue, annotations aux divers niveaux linguistiques et moteur de recherche pour linguistes, rapprocheront linguistes et informaticiens et amplifieront d'autant les progrès dans le domaine des sciences du langage et de leurs applications.

Bibliographie

- Authier-Revuz, J. (1994) : L'énonciateur glosateur de ses mots : explicitation et interprétation. *Langue française*, 103, 91-102.
- Authier-Revuz, J. (1995) : *Ces mots qui ne vont pas de soi*. Paris, Larousse.
- Authier-Revuz, J. (1982) : La mise en scène de la communication dans les discours de vulgarisation scientifique. *Langue française*, 53, 34-47.
- Bourigault, D. et Slodzian, M. (1999) : Pour une terminologie textuelle. *Terminologies nouvelles*, 19, 29-32. <<http://www.cfwb.be/franca/termin/charger/rint19.pdf>>
- Desgraupes, B. (2001) : *Introduction aux expressions régulières*. Paris, Vuibert Informatique.
- Fuchs, C. (1982) : La paraphrase entre langue et discours. *Langue française*, 53, 22-33.
- Habert, H., Fabre, C., Issac, F. (1998) : *De l'écrit au numérique*. Paris, InterEditions.

⁴⁵ À l'université Paul-Valéry, les étudiants en sciences du langage commencent à s'intéresser à l'outillage informatique en DEA ou en cours de doctorat, mais à ce stade aucune formation ne leur est proposée. En licence Sciences du langage mention TAL, le langage des expressions régulières est enseigné dans l'environnement Frantext et Intex mais le nombre d'étudiants inscrits est faible.

⁴⁶ Cette enquête a été menée par N. Gasiglia et J. Véronis.

- Habert, B. (2002) : Outiller les linguistes/ outiller la linguistique : par où, par qui commencer ?. Intervention à la table ronde TAL et enseignement, TALN'02.
<http://www.limsi.fr/Individu/habert/Cours/index.html>
- Jakobson R. (1963) : *Essais de linguistique générale I*. Paris, Minuit.
- Julia, C. (2001) : *Fixer le sens ? La sémantique spontanée des gloses de spécification du sens*. Paris, Presses de la Sorbonne Nouvelle.
- Kilgarriff, A. et Grefenstette, G. (2003) : Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, vol. 29/3, 333-347. <http://mitpress.mit.edu/>
- Kilgarriff, A. (2003) Linguistic search engine. In Kiril-Simov, (ed.), *Shallow Processing of Large Corpora : workshop tenu conjointement à Corpus Linguistics 2003*, Lancaster, England, 53-58.
<http://citeseer.nj.nec.com/568007.html>
- Leroy, S. (2001) : *Entre identification et catégorisation, l'antonomase du nom propre en français*. Doctorat en Sciences du langage, Université Paul-Valéry, Montpellier III.
- Marcello-Nizia, C. (1995) : *L'évolution du français*. Paris, Armand Colin.
- Morin, E. (1999) : *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Doctorat en Informatique, Université de Nantes.
<http://www.sciences.univ-nantes.fr/info/perso/permanents/morin/publications.html>
- Mortureux, M.-F. (éd.) (1982) : La vulgarisation. *Langue française*, 53.
- Niklas-Salminen, A. (2003) : Les emprunts et la glose. *Langues et langage*, 9, 57-72.
- Panckhurst, R. (2003) : La glose, le document électronique et l'extraction automatisée. *Langues et langage*, 9, 273-294.
- Pearson, J.(1999) : Comment accéder aux éléments définitoires dans les textes spécialisés ?. *Terminologies nouvelles*, 19. <http://www.rifal.org/>
- Rebeyrolles, J. (2000) : *Forme et fonction de la définition en discours*. Doctorat en Sciences du langage, Université de Toulouse-le-Mirail, Toulouse II.
<http://www.univtlse2.fr/erss/membres/rebeyrol/>
- Reboul-Touré, S. (2003) : La glose entre langue et discours. *Langues et langage*, 9, 75-92.
- Rey-Debove, J. (1997) : *Le métalangage*. Paris, Armand Colin.
- Riegel, M. et Tamba, I. (éd.) (1987) : La reformulation du sens dans le discours. *Langue française*, 73.
- Riegel, M. et Tamba, I. (1987) : Présentation. *Langue française*, 73, 3-4.
- Steuckardt, A. et Niklas-Salminen, A. (éd.) (2003) : Le Mot et sa Glose. *Langues et langage*, 9, Aix-en-Provence, Publications de l'Université de Provence.
- Steuckardt, A. (2003) : Présentation. *Langues et langage*, 9, 5-17.
- Sablayrolles, J.F. (2003) : Néologismes et gloses. *Langues et langage*, 9, 23-40.
- Tamba, I. (1987) : «Ou » dans les tours du type : » un bienfaiteur public ou évergète », *Langue française*, 73, 16-28.
- Tamba-Mecz, I. (1994) : *La sémantique*. Paris, Presses Universitaires de France.
- Véronis, J. (2000) : Annotation automatique de corpus : panorama et état de la technique. In *Ingénierie des langues*, Pierrel, J.-M. (éd.), Paris, Hermès, 111-129.

<www.up.univ-mrs.fr/~veronis/pdf/2000hermes4.pdf>

Augusta.Mela@univ-montp3.fr
Université Paul-Valéry Montpellier III
Route de Mende
34199 Montpellier Cedex 5 France
tel : (+33) 4 67 14 22 25