
Analyse linguistique assistée par ordinateur du courriel

[chapitre dans ouvrage : Anis J. (sous la dir.), *Internet, communication et langue française*, Paris : Hermès, 1999, p.55-70].

Rachel Panckhurst

Praxiling, ESA5475
Discours, textualité et production de sens
Université Paul Valéry - Montpellier III
F-34199 Montpellier cedex 5
rachel@bred.univ-montp3.fr

RÉSUMÉ : Dans cette recherche, nous privilégierons un mode asynchrone de communication sur Internet : le courrier électronique. Nous nous proposons d'effectuer une analyse linguistique de messages électroniques rédigés en français que nous avons reçus dans un cadre universitaire (en provenance d'étudiants — tous cycles confondus — de collègues et d'amis). Cette analyse est effectuée à l'aide de deux outils pour le traitement automatique du langage naturel, Nomino et Cordial.

ABSTRACT : The asynchronous medium, i.e. email via the Internet, has served as a basis for this piece of research ; French email messages received from undergraduate, graduate, doctoral students, colleagues and friends in a University context have undergone linguistic analysis. In order to refine the linguistic analysis of significant corpora of email messages further, I have used two natural language processing tools, Nomino and Cordial.

MOT-CLÉS : communication médiée par ordinateur (CMO), discours électronique médié (DEM), analyse du discours, analyse lexico/morpho-syntaxique, traitement automatique du langage naturel (TALN).

KEY WORDS : computer-mediated communication (CMC), electronic mediated discourse (EMD), discourse analysis, lexical/morpho-syntactic analysis, natural language processing (NLP).

Analyse linguistique assistée par ordinateur du courriel¹

1. Introduction

Le courrier, les listes, les conférences, les forums de discussion électroniques sont des moyens de communication² et/ou de diffusion d'information à travers le réseau Internet. Pour qui s'intéresse aux documents authentiques, et aux corpus [HAB 97], les données qui transitent via ces canaux constituent une richesse indéniable pour des champs d'étude variés. Dans la communauté du traitement automatique du langage naturel (TALN), les recherches à partir de « corpus électroniques » de grande taille foisonnent. Les corpus peuvent être « annotés » ou non, mais en tout état de cause, ces « ressources langagières », mises à disposition des chercheurs, sont un apport primordial qui fournissent un support pour l'avancement de la recherche en sciences du langage en général et en traitement automatique en particulier ; il devient potentiellement possible de tester des hypothèses linguistiques de tous ordres d'une manière plus aisée, puisqu'on accède à des quantités fort importantes de « données » linguistiques. Par ailleurs, on connaît l'enjeu économique et industriel induit par la mise à disposition de ressources de cette nature, surtout depuis l'essor de l'accès au réseau Internet.

Dans le cadre de la recherche présente, une définition stricte du terme « corpus » nous importe peu. Notre souci réside dans l'idée de recueillir (sélectionner et organiser) des « données » linguistiques [PAN 94], formant un « échantillon du langage », [SIN 96 : p. 4]. in [HAB 97 : p. 144], un corpus. Nous souhaitons ensuite l'analyser, le dépouiller dans le but précis d'observer des phénomènes linguistiques, et ce, indépendamment de théories linguistiques précises. Pour notre part, le corpus retenu ne sera pas « annoté » comme il l'est dans certains cadres retenus en linguistique-informatique ; on étudiera le texte grâce à des outils d'analyse automatique lexico-morpho-syntaxique. Autrement dit, le texte électronique en entrée ne subit aucune manipulation manuelle ou semi-manuelle d'annotation avant que n'interviennent les stades de segmentation, de lemmatisation, d'étiquetage et d'analyse automatisés.

Les points clefs qui vont guider cette recherche, sont les suivants :

¹ Nous remercions Sophie David pour une relecture d'une version précédente de cette recherche.

² « À trop prétendre communiquer, à trop faire passer les messages de l'un à l'autre, à trop copier et recopier, peut-être ne reste-t-il plus grand-chose qui tienne de la communication véritable dans l'univers télématique. » [MEL 96 : p.39].

- Y a-t-il des spécificités d'ordre linguistique ou extra-linguistique dans les données apparaissant au sein du courrier électronique ?
- Ce « genre du discours » constitue-t-il une forme nouvelle, qui se distingue de l'écrit (des écrits), de l'oral (des oraux) ?
- Le traitement automatique, à l'aide d'outils de repérage et d'analyse, peut-il nous être utile, et si oui, de quelle manière ?

Avant de répondre à ces questions, un bref historique afin de situer l'étude et le choix du corpus s'avère nécessaire.

1.1. Historique

Les modes électroniques de communication peuvent intéresser les chercheurs en anthropologie, ethnologie, sociologie, philosophie, psychologie, linguistique, information-communication etc., et ce à plusieurs égards. Désormais, sont réunies dans des documents « authentiques » écrits — ou plutôt saisis — les traces de certaines mutations sociales. En effet, grâce à l'essor de l'utilisation du réseau Internet, l'écrit est, en quelque sorte, réhabilité. De prime abord, ceci peut paraître paradoxal, étant donné la disparition annoncée du livre par certains et la montée en puissance du visuel. Toutefois, cet écrit ne répond certainement plus aux critères d'antan. Comme le fait remarquer Gadet, le XXe siècle est « le théâtre de mouvements divergents » en ce qui concerne les relations entre oral et écrit. Depuis les années soixante-dix, avec l'ordinateur, le Minitel et Internet, « on assiste à un retour à une position plus favorable de l'écrit ; mais ceci dans un autre rapport que précédemment, car « écrit » n'implique plus du tout 'langue travaillée et soignée' » [GAD 96 :15-16]. Mais, en tous les cas, cet écrit nous fournit un vaste champ d'exploration multidisciplinaire.

Dans un premier temps, notre attention a porté sur l'appropriation de l'outil que constitue le courrier électronique par un public d'étudiants en premier cycle. Au départ, nous avions un certain nombre d'intuitions — en tant qu'utilisateur chevronné de ce moyen de communication — que nous voulions vérifier. Nos travaux antérieurs ont porté sur deux corpus : le premier [PAN 98a, 99] était constitué de 115 messages de courrier électronique envoyés par 64 étudiants dans le cadre d'un cours de premier cycle, intitulé « Informatique et nouvelles technologies » (DEUG : « Médiation culturelle et communication » et « Sciences du Langage » à l'université Paul-Valéry - Montpellier III) ; le second [PAN 98b] a été élargi afin de prendre en compte un public d'étudiants de 2^e et de 3^e cycles, de collègues et d'amis pour un total de 1285 messages dépouillés.

Nous résumons brièvement ci-dessous quelques résultats (linguistique et extralinguistique) de ces travaux :

- les erreurs sont fréquentes (fautes de saisie et/ou erreurs de type lexico-grammatical) ;
- les abréviations, le style télégraphique sont de mise ;
- les « binettes » (ou les « smileys ») sont parfois introduites ;
- les formules d'ouverture et de clôture sont réduites, voire absentes ;
- les niveaux de langue sont variables (même dans un cadre potentiellement formel constitué par la situation de communication pédagogique) ;
- l'utilisation d'une « simulation » des tours de parole, à l'aide des chevrons « > » (automatiquement apposés en début de ligne au sein d'une réponse), indiquant la parole de l'autre, est répandue.

Cependant, les points évoqués ici ne sont pas toujours discriminants du moyen de communication adopté. Pour avoir effectué une étude comparative entre l'utilisation du courrier électronique, le traitement de texte, et l'écrit manuscrit en situation d'examen [PAN 98b], on remarque que les erreurs (de saisie et d'ordre lexico-grammatical pour les deux premiers et uniquement d'ordre lexico-grammatical pour le troisième) sont relativement fréquentes dans les trois types de documents. Par contre, et cela pourrait surprendre davantage, y compris dans les messages provenant de collègues universitaires, les erreurs sont toutes aussi fréquentes. Pour une part, cela s'explique aisément : le courrier électronique est un moyen rapide de communication ; de ce fait, on se relit peu, et les erreurs (qui peuvent toujours être rejetées sur l'outil informatique) sont, de manière générale, relativement bien tolérées.

Par ailleurs, bien que le courrier électronique (désormais CÉ) soit un moyen asynchrone, « [c']est une forme d'échange asynchrone [...] nourrie par une illusion technique de synchronicité » [MEL 96, p. 34]. Le CÉ tente de se rapprocher du temps réel. D'où d'autres problèmes : on exige une certaine immédiateté dans la réponse, et si le destinataire ne peut répondre rapidement, il le fait généralement savoir à l'expéditeur. Devant l'absence de réponse, un sentiment d'« angoisse de la communication » (Feenberg, in [PER 92 :226]) peut naître : pourquoi le destinataire ne répond-il pas ? N'a-t-il pas reçu le message ? Que représente ce silence ? Cette rapidité exigée peut également mener à des situations d'agressivité et de regret postérieur : le destinataire, devenu expéditeur à son tour, répond rapidement, sans parfois réellement réfléchir à sa réponse. Dans le message suivant, expédié par un collègue à la liste de diffusion de l'université en janvier 1999, on devine sans peine qu'il y a eu, dans un précédent message, ce type de malentendu, de « montée d'agressivité ». Dans le message cité ci-dessous, il envoie ses excuses, ce qui constitue un rattrapage postérieur, phénomène

relativement fréquent par CÉ. Cela étant, bien que l'expéditeur reconnaisse manifestement son erreur, les raisons évoquées sont partiellement rejetées sur la machine « tronqueuse de message » !³

« Chers Collègues,

Je viens de prendre connaissance du message adressé par X. Je l'en remercie vivement et lui présente comme je vous présente toutes mes excuses pour cette erreur. Je vous prie de bien vouloir la mettre au compte de la fatigue, d'une mauvaise lecture par suite d'un message d'abord reçu tronqué et de rien d'autre. Je renouvelle donc publiquement mes excuses à notre collègue [...]. C'est au moins l'avantage du courrier électronique que de permettre ce genre de corrections [...]. »

Par ailleurs, dans [PAN 99], nous avons relevé certains cas où l'étudiant se « raconte »⁴, et ce phénomène nous a intriguée. Alors qu'il ne le fait pas en situation de communication face-à-face avec l'enseignant, il choisit ce moyen de communication qui permet à la fois une certaine proximité avec l'autre, même si cet autre est inclus dans une situation de communication potentiellement formelle. Mais cette proximité est en même temps abstraite. L'autre change quelque peu de rôle : remplit-il alors une fonction de psychologue semi-abstrait ? Par ailleurs, la proximité (illusoire ?) peut s'accompagner d'une certaine « protection », bien que cela puisse paraître contradictoire ; derrière son écran l'on ne peut être atteint ; la distanciation, le recul vis-à-vis de l'autre sont désormais envisageables. Le message suivant, reçu d'une étudiante à la suite d'un échec à un examen du mois de juin 1998, témoigne de cette protection recherchée. Elle n'ose pas demander un rendez-vous à son enseignante ; en effet, l'utilisation du courrier électronique lui semble être une solution moins douloureuse. L'idée de recevoir une réponse négative sur son écran est moins perturbante (dans la mesure où l'autre est précisément absent) que la situation verbale et non-verbale qu'elle aurait dû gérer en face-à-face :

« Si je vous écris, c'est parceque je préfère ne pas être en face de vous pour vous entendre me répondre définitivement "non", au moins le courrier électronique m'aura donné "cet avantage". J'aurais voulu venir vous voir et vous expliquez mon point de vue, mais à quoi bon... je ne sais même pas pourquoi je continue à espérer que pour 0,5 points les choses pourraient changer. [...] »

³ Tous les messages apparaissent tels quels, avec leurs erreurs éventuelles.

⁴ Précisons que les étudiants ne savaient pas que leurs messages allaient servir dans le cadre d'un dépouillement linguistique.

Ces quelques explications qui évoquent les raisons du choix de cette étude étant mentionnées, nous pouvons désormais délimiter le corpus retenu.

1.2. Corpus

Dans le cadre présent, nous avons bâti un corpus à partir de tous les messages électroniques reçus pendant une période de 8 mois (entre mai et décembre 1998). Ceux-ci incluent des courriers en provenance d'étudiants (tous cycles confondus), de collègues et d'amis. Au total, le corpus était initialement constitué de 1676 messages. Nous l'avons ensuite épuré, afin de filtrer tous les messages écrits entièrement en langue étrangère, puisque nous souhaitons procéder à un traitement par analyse morpho-syntaxique automatique en français. Ont été également éliminés, tous les messages envoyés à la liste de diffusion de notre Université (à laquelle nous sommes abonnée), et ce, dans un souci d'étudier un corpus homogène, contenant uniquement des messages de type personne à personne. Suite à cette étape d'épuration, l'on aboutit à un corpus de la taille suivante :

Résultats à partir de Word 98	
Pages	1597
Mots	342092
Caractères (sans espaces)	2189079
Caractères (avec espaces)	2564964
Paragraphe(s)	51621
Lignes	79205
Taille fichier	3,7 mo

Nous l'avons ensuite soumis à deux outils automatiques permettant des analyses linguistiques (Nomino et Cordial), afin de dégager une première répartition concernant les types de catégories syntaxiques employés, avant de procéder à une analyse contextuelle plus fine.

2. Traitement automatique

Avant d'exposer certains résultats de cette analyse (cf. § 3), nous fournissons une explication des types de logiciels employés, et les raisons scientifiques ayant conduit à ce choix.

2.1 *Format texte, segmentation*

Nomino [DUM 96] et Cordial [COR 98] sont des logiciels qui travaillent à partir d'un document électronique en format texte. Ce premier point peut paraître banal, mais il est important à souligner : aucune préparation, aucun prétraitement manuel n'est nécessaire au bon fonctionnement des logiciels. Le texte est fourni tel quel, « brut ». Nomino et Cordial assurent automatiquement le découpage du texte en phrases avant d'en proposer une analyse lexico-syntaxique. À aucun moment, le linguiste-utilisateur n'est obligé d'indiquer quels sont les mots ou les séquences à partir desquels il souhaite voir apparaître des résultats. Ainsi, les hypothèses de départ ne sont pas directement « injectées » au sein du texte, comme cela est parfois le cas avec d'autres logiciels. De la même manière qu'avec le logiciel d'analyse morpho-syntaxique, Termino [DAV 93], on pourra plus aisément « découvrir » ce que le texte contient réellement.

2.2. *Analyse morphologique, étiquetage, concordances*

Par ailleurs, Nomino ne fonctionne pas exclusivement à l'aide d'un dictionnaire électronique. Un analyseur morphologique y est incorporé — qui fonctionne entre autres à partir de règles de terminaison — ce qui permet d'affecter une catégorie syntaxique et des informations morphologiques à un mot non inclus dans le dictionnaire⁵. Autrement dit, la néologie lexicale est permise, et elle n'arrête pas le travail du logiciel. D'autres logiciels existant sur le marché travaillent souvent à partir de dictionnaires de très grande taille, mais ils ne permettent pas systématiquement un travail sur des néologismes.

Il est largement admis par la communauté du traitement automatique du langage naturel (TALN), qu'un analyseur lexico-syntaxique doit incorporer une phase d'étiquetage et de lemmatisation. Un des problèmes qui se posent est qu'un mot (hors contexte) peut recevoir plusieurs catégories (par exemple, « porte » est tantôt verbe, tantôt nom, etc.). Dans une phase d'étiquetage, le logiciel doit donc affecter toutes les valeurs catégorielles possibles et fournir également une indication sur le lemme retenu, et qui sera utilisé au stade de l'analyse syntaxique proprement dite. Par exemple, à partir d'une phrase

⁵ Le fonctionnement de Cordial, qui est un logiciel qui permet à la fois la correction grammaticale et l'analyse de données textuelles, semble différent ; la néologie en serait exclue.

classique comme : *la petite brise la glace*⁶, un étiqueteur (*cf.* à titre d'exemple, celui proposé en ligne par Xerox-France⁷) fournit les résultats que nous résumons dans le tableau suivant :

	lemme	informations morphologiques	catégorie
la	le	Fem+SG+Def	Det
	le	Acc+Fem+SG+P3	PC
	la	Masc+InvPL	Noun
petite	petit	Fem+SG	Noun
	petit	Fem+SG	Adj
brise	briser	SubjP+SG+P3	Verb
	briser	SubjP+SG+P1	Verb
	briser	Imp+SG+P2	Verb
	briser	IndP+SG+P3	Verb
	briser	IndP+SG+P1	Verb
	brise	Fem+SG	Noun
la (<i>cf.</i> ci-dessus)			
glace	glacer	SubjP+SG+P3	Verb
	glacer	SubjP+SG+P1	Verb
	glacer	Imp+SG+P2	Verb
	glacer	IndP+SG+P3	Verb
	glacer	IndP+SG+P1	Verb
	glace	Fem+SG	Noun

Figure 1.

Les résultats de la figure 1 nous indiquent que, hors contexte :

- « la » peut être déterminant, nom (note de musique) ou pronom clitique ;
- « petite » est nom ou adjectif ;
- « brise » est verbe ou nom ;
- « glace » est verbe ou nom.

Par ailleurs, chaque catégorie est ramenée à une forme canonique, à un lemme : l'infinitif pour les verbes, le singulier pour le nom, le masculin singulier pour les autres formes.

⁶ Cf. [FUC 96, p.95].

⁷ <http://www.xrce.xerox.com/research/mltt/Tools/morph.fr.html>

Cet aspect est particulièrement intéressant et présente *de facto* un avantage sur les logiciels dits « concordanciers ».

Un concordancier permet, en effet, de faire un tri rapide de tous les « mots » d'un texte, de les situer en contexte (KWIC - Key word in context), de compter le nombre d'occurrences, etc. (cf. le tableau de la figure 2), mais un logiciel de ce type ne peut pas effectuer un travail linguistique plus poussé⁸, précisément parce qu'il travaille à partir de chaînes de caractères.

	Contexte gauche	Occurrence - KWIC	Contexte droit
Ligne de l'occurrence		Concordance	
849	Nieves ". Après quoi,	abandonnant	ces " Indes
435	monde, elle n'aurait	abandonné	son métier. Elle a tenu
1654	de Seine, lorsqu'elle	abandonne	son arc de cercle pour
1439	grande cité fortifiée	abandonnée	du monde, ou les
2913	de la mosquée	abandonnée	de Bahla (quatorzième
908	aux façades ternies,	abandonnées.	Le long des petites
910	des chantiers	abandonnés	attendent des jours
1696	ses propres tics, ses	abandons	et sa sauvegarde par la

Figure 2.

Par exemple, comment relever et regrouper facilement toutes les occurrences d'un verbe comme « aller » dans un texte, alors qu'elles apparaissent sous diverses formes : *vais, irai, suis allée, alla...* ? Actuellement, certains logiciels de ce type permettent une lemmatisation. À titre d'exemple, le logiciel de requête de la base Frantext⁹ permet une recherche (partielle) de ce type.

2.3. Analyse syntaxique

Dans le cadre présent, le travail de catégorisation et de lemmatisation, couplé avec l'analyse syntaxique dans une étape suivante, est absolument primordial. Revenons à nos résultats de la figure 1 et confrontons-les à une analyse syntaxique :

⁸ La figure 2 présente un résultat d'analyse du logiciel Conc (Concordance Generator for the Macintosh : <http://www.sil.org/computing/conc/>). Un travail permettant de simuler, jusqu'à un certain degré, le travail de lemmatisation, est possible à partir de l'inclusion de « patrons de fouille », mais ceux-ci sont entièrement décrits de manière manuelle pour des syntagmes isolés.

⁹ <http://www.ciril.fr/~mastina/FRANTEXT>

<p>p[_{sn}[_{det}la _{adj}petite _nbrise] _{sv}[_{pc}la _vglace]] interprétation possible = « Le petit souffle de vent lui donne froid »</p>
<p>p[_{sn}[_{det}la _npetite] _{sv}[_vbrise _{det}la _nglace]] interprétation possible = « La petite fille/femme casse le miroir »</p>

Figure 3.

Bien que cette phrase constitue une ambiguïté réelle au niveau syntaxique, certaines informations apportées par l'étiqueteur au stade morphologique auront été écartées : « la » en tant que nom, ou les formes verbales pour « briser » et « glacer » qui concernent les première et deuxième personnes, que ce soit pour l'indicatif ou le subjonctif.

Ce type d'information est précieux dans le cadre d'une analyse automatique sur corpus de manière générale, et d'autant plus dans un domaine de recherche nouveau comme celui constitué par la communication médiée par ordinateur, et ce pour deux raisons : nous pouvons immédiatement avoir accès à une quantification automatique des différents types de constituants — qui auront effectivement été repérés dans leur contexte spécifiquement syntaxique, et qui ont donc un statut linguistique réel — ; grâce à un repérage supplémentaire des expressions nominales polylexicales [DAV 93], (« traitement de textes », « chaise longue » etc.), qui constituent des séquences *a priori* non « listables », des « réseaux de signification » peuvent être tissés :

L'intérêt des expressions pour ce que nous cherchons quand nous analysons des discours est très important. En effet, elles définissent les thèmes dont parlent les discours en formant de grands axes de signification qui permettent de répondre à la question : « De quoi ce discours parle-t-il ? » [SOU 97 :252].

3. Résultats de notre corpus

En fouillant les résultats de notre corpus analysé par Nomino et par Cordial, nous nous sommes constamment posé la question de la place du courrier électronique (ou plutôt le terme que nous préférons, à savoir « discours électronique médié ») vis-à-vis de l'oral (ou des oraux), d'une part, et les autres formes de l'écrit, d'autre part. Nous étions à la recherche de marques explicites qui fassent pencher davantage d'un côté ou de l'autre, ou

bien, au contraire, qui indiquent la naissance d'une forme, d'un « genre de discours » tout à fait nouveau. Pour cela, bien entendu, les intuitions ne suffisent pas ; l'analyse approfondie du texte s'impose, et pour ce faire, l'ordinateur peut nous être d'un secours remarquable, en nous évitant les tâches extrêmement fastidieuses.

3.1. Préliminaires : oral, écrit.

Pour commencer, on ne peut nier le fait que le CÉ corresponde en effet à une forme écrite et non à une forme orale. On saisit le courrier électronique à l'aide d'un clavier, relié à un ordinateur ; il s'agit là d'une évidence. En quoi, donc, le discours électronique médié (DEM) pourrait ressembler à une situation de communication orale ? Rappelons trois divergences entre oral et écrit, à partir de [GAD 96 : 17-19], que nous résumons brièvement ci-dessous :

- hétérogénéité, variabilité de l'oral (face à une relative homogénéité de l'écrit : codifié, fixé, stabilisé, normé) ;
- « scories » : l'oral reste « truffé de pauses, hésitations, reprises, recherches de mots, incomplétudes, redites, anticipations, auto-interruptions... » ;
- intonation (et les facteurs prosodiques).

Du côté du DEM, il semblerait que l'illusion de synchronicité, l'exigence de la rapidité, la pauvreté de la forme textuelle « brute » (absence d'attributs de style : gras, italique, etc.), font que l'on a recours à des types de fonctionnement typiques d'une situation de communication orale, et qui ne figurent pas (ou, en tous les cas, à un degré bien moindre), dans d'autres types d'écrits :

- les « binettes », qui permettent d'introduire des aspects sémiologiques non-verbaux (par ex. le « :-) » qui indique un sourire) ;
- l'utilisation de mots saisis entièrement en majuscules, qui, selon la « netiquette », indiquent un sentiment de colère, ou d'agacement ;
- l'allongement, la répétition de caractères qui peuvent, dans certains cas, simuler l'intonation, et, ainsi, apporter une information para-verbale...

Certaines marques spécifiquement linguistiques tendent vers une situation d'oralité :

- utilisation de verbes de clôture typiques d'une situation de communication orale :
[...] au plaisir de te *reparler* [PAN 99] ;
Je suis pressée donc je te *laisse* ;
- la trace d'un « ratage » :

Je vous envoie un message avant de vous transmettre *mon* nouvelle adresse électronique » [PAN 98a] ;

Dans le premier exemple, il s'agit effectivement d'une situation d'utilisation ultérieure du CÉ, et non d'un autre moyen de communication, car le collègue en question et moi-même ne parlons jamais par téléphone. Le verbe « laisser », dans le deuxième exemple, montre nettement l'inclusion de l'autre, du destinataire, dans la situation pourtant asynchrone. Le troisième exemple (produit par une francophone) montre la trace du « ratage » quant au non-remplacement du déterminant possessif ; à l'oral cette séquence pourrait équivaloir à « mon ad...ma nouvelle adresse ».

En tous les cas, les quelques indices mentionnés ci-dessus montrent peut-être un fonctionnement qui diverge d'autres formes écrites. Anis, pour sa part, suggère que l'écrit communicationnel agit comme « un hybride entre l'écrit et l'oral » [ANI 98 : p. 122].

3.2. Occurrences et catégories

Qu'en est-il des différents types de catégories syntaxiques employés ? Le DEM renferme-t-il une utilisation plus importante de noms ou de verbes, par comparaison à d'autres formes de l'écrit et/ou de l'oral ?

Le tableau de la figure 4 présente des résultats qui sont fournis en sortie de la phase de désambiguïté catégorielle. Nomino, dans sa version actuelle, repère les verbes, les noms, les adjectifs, les adverbes, les unités complexes nominales (de type « pomme de terre ») ; le logiciel affecte tout autre unité à une catégorie « autres ». Cordial, pour sa part, repère également les articles, les pronoms, les prépositions et certaines conjonctions. Pour chaque catégorie syntaxique, nous avons fourni le nombre et le pourcentage d'occurrences repérés par chaque logiciel, et, dans le cas de Nomino, le nombre et le pourcentage de formes différentes. Dans l'analyse qui suit, nous n'évoquerons que les verbes, les noms et les pronoms. Par ailleurs, les variations des résultats entre différents logiciels ne doivent pas surprendre ; aucun logiciel de traitement automatique n'est fiable à 100 % et les différences au niveau de l'analyse sont donc naturelles.

Résultats Nomino	Verbes	Noms	Adj	Adv	Autres				Total
Unités/ occurrences (tokens)	38 920 10,89 %	149 216 41,75 %	16 952 4,74 %	11 037 3,09 %	141 316 39,54 %				357 441 100 %
Formes (types)	2 048 8,67 %	17 877 75,74 %	1 729 7,32 %	369 1,56 %	1 585 6,71 %				23 608 100 %
Résultats Cordial	Verbes	Noms	Adj	Adv	Art	Pro	Prep	Conj	Total
Unités/ occurrences (tokens)	40 163 11,74 %	153 007 44,73 %	20 551 6,00 %	16 040 4,68 %	43 926 12,84 %	23 319 6,81 %	32 558 9,51 %	12 446 3,69 %	342 010 100 %

Figure 4.

3.3. Noms vs. verbes

Les résultats de la figure 4 montrent clairement l'importance de l'utilisation des noms¹⁰, face aux autres formes : entre 41 % et 44 % du nombre d'occurrences au total, par rapport à un taux de 10 % à 11 % pour les verbes. La richesse lexicale nominale est nette : 75 % de formes distinctes, alors que les verbes ne comptent que 8 %.

Ces taux de fréquence élevés semblent concorder avec Gadet [GAD 96], qui confronte l'utilisation des données nominales et verbales entre l'oral et l'écrit. À partir d'une analyse de Halliday [HAL 89] pour l'anglais, elle constate précisément une proportion plus faible de verbes (vis-à-vis des noms) à l'écrit et *vice versa* à l'oral :

L'oral relève d'une complexité d'ordre grammatical (alors que l'on entend souvent dire que l'oral n'a pas de grammaire) et l'écrit d'une complexité d'ordre lexical (qui va de pair avec une certaine monotonie grammaticale). Ceci s'accompagne d'une préférence de l'oral pour les verbes, et de l'écrit pour les noms. [GAD 96 : p. 23]

Selon Halliday, la « densité lexicale » est très élevée à l'écrit, face à l'oral, qui privilégie plutôt une « densité grammaticale ». Les mots grammaticaux (incluant les déterminants,

¹⁰ Dans la mesure où les en-têtes des messages ont été maintenus au sein du corpus, le pourcentage des substantifs est légèrement supérieur à la normale. Par ailleurs, nous n'avons pas inclus les formes nominales complexes (repérées par Nomino) dans la figure 4. Elles représentent environ 12 % des occurrences totales.

les pronoms, la plupart des prépositions, les conjonctions, certaines classes d'adverbes et les verbes fléchis) seraient donc utilisés en plus faible quantité à l'écrit qu'à l'oral :

There is a characteristic difference between spoken and written language. Written language displays a much higher ratio of lexical items to total running words. [HAL 89 : p. 61]

The difference between written and spoken language is one of DENSITY : the density with which the information is presented. Relative to each other, written language is dense, spoken language is sparse. [HAL 89 : p. 62]

Les résultats de la figure 4 montrent que cette densité lexicale est également caractéristique de la forme écrite constituée par le DEM.

3.4. Verbes, pronoms personnels

À quelles conclusions peut-on aboutir face aux autres formes d'écrit ? Un des avantages majeurs apporté par la version 5 de Cordial est de fournir la possibilité d'effectuer des comparaisons avec un corpus de 800 Mo d'ouvrages et plus de 1,5 Go d'articles de presse et de *l'Encyclopædia Universalis*. Les textes représentés appartiennent au domaine littéraire, journalistique, technique, juridique et commercial.

En soumettant notre corpus de messages électroniques à la comparaison de la base de Cordial, et pour les deux seuls exemples des verbes et des pronoms personnels, nous obtenons les résultats suivants :

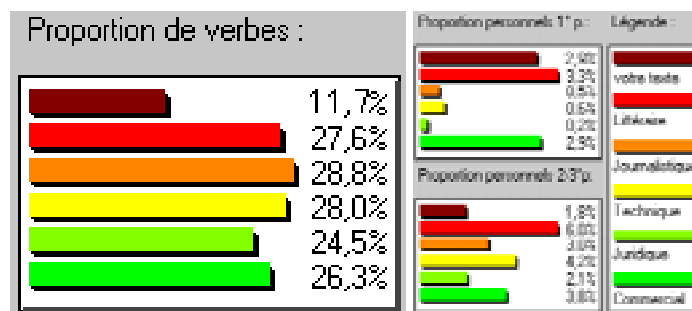


Figure 5.

Non seulement la proportion de verbes est plus faible à l'écrit par rapport à l'oral (cf. § 3.3), mais on voit nettement à partir de la figure 5 que le taux (11,7 %) est très en-dessous d'une utilisation « normale » à l'écrit (plus de 20 %), quel que soit le type de corpus. Par ailleurs, en ce qui concerne les pronoms personnels¹¹, nous avons déjà eu l'occasion [PAN 98a, 98b] de signaler que le DEM renferme un pourcentage inférieur de pronoms personnels de troisième personne que les autres formes de l'écrit (et de l'oral). Bien que les résultats de Cordial englobent à la fois les pronoms personnels de 2^e et de 3^e personnes — ce que l'on peut regretter d'un point de vue linguistique —, le pourcentage global (1,8 %) est encore inférieur au pourcentage des pronoms personnels de 1^{re} personne (2,9 %). De plus, ce résultat montre l'écart significatif vis-à-vis des autres formes de l'écrit (mis à part les documents juridiques).

Parmi les verbes fréquemment employés on remarque les *auxiliaires* (*être, avoir*) et les *semi-auxiliaires de modalité* (*sembler, paraître, devoir, pouvoir*) auxquels peuvent s'ajouter *savoir, croire, vouloir* [LEE 94]. Ces verbes, à eux seuls, représentent plus de 30 % de l'ensemble des occurrences verbales employées :

Cordial			Nomino		
	n° d'occurrences	% de l'ensemble des verbes		n° d'occurrences	% de l'ensemble des verbes
être	5 077	12,6 %	avoir	4 698	12,1 %
avoir	4 473	11,1 %	être	4 312	11,1 %
pouvoir	1 308	3,2 %	pouvoir	1 295	3,3 %
devoir	727	1,9 %	devoir	661	1,7 %
vouloir	356	0,9 %	vouloir	354	0,9 %
savoir	322	0,8 %	savoir	331	0,8 %
croire	114	0,3 %	croire	113	0,3 %
sembler	114	0,3 %	sembler	106	0,2 %
Total	12 491	31,1 %	Total	11 870	30,4 %

Figure 6.

Enfin, parmi les temps verbaux, le présent est nettement privilégié :

¹¹ Le pourcentage de 4,7 % présenté à la figure 5 correspond, bien entendu, à un sous-ensemble du total des pronoms, ici les pronoms personnels. Le pourcentage de 6,81 %, présenté à la figure 4, englobe les autres formes pronominales.

Temps verbaux	%
Présent	63 %
Imparfait et passé	4 %
Futur	8 %
Conditionnel	4 %

Figure 7.

L'utilisation accrue du présent n'étonne guère, dans la mesure où le DEM fonctionne pleinement dans l'immédiateté. Le mode impératif¹² (21 %) montre le style potentiellement concis et direct des communications électroniques¹³.

4. Conclusion

Dans cette recherche — qui ne se prétend absolument pas exhaustive —, nous avons tenté de montrer comment le traitement automatique peut être d'un intérêt tout à fait appréciable pour repérer certains fonctionnements linguistiques. Grâce aux résultats présentés au § 3., l'on comprend que le discours électronique médié se rapproche parfois de l'oral/des oraux et parfois de l'écrit/des écrits¹⁴.

Dans nos recherches antérieures, nous avons signalé deux écarts significatifs vis-à-vis de l'oral : le fonctionnement de l'interrogation et de la négation¹⁵, tous deux se rapprochant d'un genre écrit normé. Alors que l'interrogation en français parlé est très fortement marquée par une forme intonative [GAD 97 : p. 112], le DEM utilise très rarement l'unique signe graphique du point d'interrogation ; l'interrogation directe y est

¹² Comme Leeman [LEE 94], nous estimons que le conditionnel constitue un temps et non pas un mode. Si les résultats de la figure 7 représentent un pourcentage total de 79 % et non pas de 100 %, c'est précisément parce que le mode impératif est également inclus dans les résultats de Cordial pour les temps verbaux (soit 21 % dans le cas présent) ; cela constitue une erreur à nos yeux.

¹³ Le style concis paraît à l'évidence lorsqu'on vérifie les moyennes mots/phrases (4,1 %), mots/paragraphe (6,2 %), et phrases/paragraphe (1,5 %), toutes trois très en-dessous des moyennes pour d'autres formes de l'écrit.

¹⁴ Ce ne sera pas obligatoirement le cas systématiquement. Comme Melançon [MEL 96], nous sommes convaincue que le courrier électronique « n'est pas une nouvelle forme de l'épistolaire ». Dans son essai *Sevigne@Internet*, il confronte ces deux formes d'écrit pour en dégager leurs différences.

¹⁵ Dans le cadre présent, nous écartons une étude de la négation, mais les résultats précédents ont montré que le couple « ne...pas » figure quasi systématiquement ensemble, sauf dans certains cas précis d'ellipse verbale. Cf [PAN 98a, 98b].

généralement signalée par une des formes de l'inversion. Dans le cadre présent, nous avons procédé à une recherche des différentes formes de l'interrogation directe, et ce, de manière manuelle (un concordancier peut être utile, ne serait-ce pour regrouper les phrases interrogatives en un seul bloc, en recherchant le signe graphique « ? »). 89 % des formes interrogatives sont constituées par une forme inversée (pronominale, nominale, ou *ce + être*) avec ou sans élément interrogatif¹⁶ vs. seulement 11 % ne contenant pour seule marque interrogative que le point d'interrogation. Ces situations sont pourtant circonscrites :

- les deux interlocuteurs se connaissent fort bien et se permettent un style très direct (mais un style de cette nature relève d'une utilisation encore relativement rare) ;
- la forme « intonative » est incluse dans un style télégraphique et elliptique général (« Vos avis ? ») ;
- ce type d'interrogation précède ou suit une autre interrogation inversée ou contenant un élément interrogatif (l'écart de style étant parfois grand) :
 - « Comment ça va ? Tu prends des vacances ? »
 - « Gardé-je :-) les comptes ? [...] Je fais pareil cette année ? »
- l'interrogation est brève, en milieu de phrase, et entre parenthèses :
 - « (approprié, non ?) »

Bien entendu, le type d'analyse évoqué ci-dessus n'est pas encore (à notre connaissance) rendu possible par les outils de traitement automatique actuels. Le chercheur doit ensuite « replonger » dans la textualité discursive afin de dégager d'autres indices linguistiques et extra-linguistiques saillants.

¹⁶ Cf. [PAN 90], pour une étude détaillée du système interrogatif français.

Références

- [ANI 98] Anis J., *Texte et ordinateur. L'écriture réinventée ?*, Paris, Bruxelles : DeBoeck, 1998.
- [BAK 84] Bakhtine M., *Esthétique de la création verbale*, Paris : Gallimard, 1979, trad. fr. 1984.
- [BLA 90] Blanche-Benveniste C., *Le français parlé. Études grammaticales*, Paris : Éditions du CNRS, 1990.
- [COR 98] *Cordial 5, Correcteur grammatical et analyseur de la langue française*. Manuel de l'utilisateur, Toulouse : Synapse développement.
- [de FOR 97] de Fornel M., Léon J., « Des questions-échos aux réponses-échos. Une approche séquentielle et prosodique des répétitions dans la conversation », *Cahiers de Praxématique*, 28, 101-126, 1997.
- [DAV 93] David S., *Les Unités nominales polylexicales. Éléments de description et reconnaissance automatique*, thèse de doctorat, université Paris-VII, 1993.
- [DUM 96] Dumas L., Plante A., Plante P., *Nomino*, 1.0., document centre d'ATO, UQAM, 1996.
- [FUC 96] Fuchs C., *Les ambiguïtés du français*, Paris : Ophrys, 1996.
- [GAD 96] Gadet F., « Une distinction bien fragile : oral/écrit », *TRANEL*, 25, 13-27, 1996.
- [GAD 97] Gadet F., *Le français ordinaire*, Paris : Armand Colin, 1997.
- [HAB 97] Habert B., Nazarenko A., Salem A., *Les linguistiques de corpus*, Paris : Armand Colin, 1997.
- [HAL 78] Haliday M.A.K., *Language as Social Semiotic : The Social Interpretation of Language and Meaning*, London : Edward Arnold, 1978.
- [HAL 89] Haliday M.A.K., *Spoken and written language*, Oxford : Oxford University Press, 1989.
- [HER 96] Herring S. C. (éd.), *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*, Amsterdam : J. Benjamins, 1996.
- [KO 96] Ko K.-K., « Structural characteristics of computer-mediated language : a comparative analysis of Interchange discourse », *Electronic Journal of Communication / Revue électronique de communication*, 6, 3, 1996.
- [LEE 96] Leeman-Bouix D., *Grammaire du verbe français. Des formes aux sens*, Paris : Nathan, 1996.
- [LEV 97] Lévy P., *L'Intelligence collective. Pour une anthropologie du cyberspace*, Paris : La Découverte, 1997.
- [MAR 97] Martin F., *Réunions décisionnelles médiatisées par ordinateur en entreprise. Approche interactionnelle*, thèse de doctorat en Sciences du Langage, Université Lumière Lyon II, 1997.
- [MEL 96] Melançon B., *Sévigéné@Internet. Remarques sur le courrier électronique et la lettre*, Montréal : Fides.
- [PAN 90] Panckhurst R., *Description linguistique et implémentation en FX des structures interrogatives (directes) du français*, thèse de doctorat en Sciences du Langage - Linguistique et Informatique, Université Blaise-Pascal, Clermont II, 1990.
- [PAN 94] Panckhurst R., « A database for Linguists: Intelligent Querying and Increase of Data », *Computers and the Humanities*, 28, 39-52, Kluwer Academic Publishers, 1994.

- [PAN 97] Panckhurst R., «La communication "médiatisée" par ordinateur ou la communication "médiée" par ordinateur ? », *Terminologies nouvelles*, 17, 56-58, 1997.
- [PAN 98a] Panckhurst R., « Marques typiques et ratages en communication médiée par ordinateur », Actes du Colloque CIDE '98, Rabat, 15-17/4/98, in Mojahid M., Karczmarczuk J. (éd.), *Document électronique*, Paris : Europia Productions, 31-43, 1998a.
- [PAN 98b] Panckhurst R., « Analyse linguistique du courrier électronique », Actes du colloque *Les relations entre individus médiatisés par les réseaux informatiques*, GRESICO, Vannes, 10-11/9/98, in Guéguen N., Tobin L. (éd.), *Communication, société et internet*, Paris : L'Harmattan, 47-60, 1998b.
- [PAN 99] Panckhurst R., « La Communication médiée par ordinateur : un discours autre ? », in *L'autre en discours*, Bres J., Delamotte-Legrand R., Madray F., Siblot P (éd.), Dyalang-Praxiling, Service des publications de l'Université Paul-Valéry, 307-331, 1999.
- [PER 92] Périn P. & Gensollen M. (éd.), *La Communication plurielle. L'interaction dans les téléconférences*, Paris : La Documentation française, 1992.
- [SCH 85] Schneuwly B. & Bronckart J.-P., *Vygotsky aujourd'hui*, Neuchâtel : Delachaux et Niestlé, 1985.
- [SOU 97] Souchard M., Wahnich S., Cuminal I., Wathier V., *Le Pen. Les mots. Analyse d'un discours d'extrême-droite*, Paris : Éditions Le Monde, 1997.
- [VYG 85] Vygotsky Lev S., *Pensée et Langage*, Paris : Terrains/Éditions Sociales, traduction française par F. Sève, 1985.