Beaudouin V., Fleury S., Velkovska J. (2000). " Études des échanges électroniques sur internet et intranet : forums et courriers électroniques ". In M. Rajman & J.-C. Chappelier (éd), JADT 2000. 5emes Journées internationales d'Analyse statistique des Données Textuelles, 9-11 mars 2000, EPFL, p. 17-26.

Etudes des échanges électroniques sur internet et intranet : forums et courriers électroniques

Valérie Beaudouin, Serge Fleury, Julia Velkovska

CNET, DIH/UCE, 38-40 rue du Général Leclerc, 92794 Issy Les Moulineaux Cedex 9

{valerie.beaudouin, serge.fleury, julia.velkovska}@cnet.francetelecom.fr

Abstract

This paper presents a study of electronic interactions in several newsgroups on the internet and on a corporate intranet, as well as by e-mail. We describe our methodological approach and the first results obtained from the analysis of our corpora with text-mining tools. These elements were completed by interviews, electronic questionnaires and conversation analysis of the discussion threads in the newsgroups. Finally, we state that in order to understand the electronic interactions, we have to study the global communication space, constructed by interconnected communication tools (homepages, newsgroups, e-mail) and we discuss somme specific issues related to the analysis of homepages.

Résumé

Ce travail propose une étude des échanges électroniques sur le réseau. Nous présentons ici la démarche suivie pour l'étude conjointe d'un corpus d'échanges sur des forums publics et sur des forums d'entreprise et d'un corpus de courriers électroniques. Cette étude est complétée par des enquêtes menées auprès des acteurs de ces échanges. Nous esquissons également une démarche d'analyse globale de l'espace de communication sous-jacent qui prend en compte l'analyse de sites web.

Mots clés : linguistique de corpus, analyse socio-linguistique, espace de communication électronique, courrier électronique, forum, page personnelle.

1. Problématique

Nos études visent à comprendre les modes de sociabilité et de coopération sur les réseaux électroniques. Nous avons cherché à étudier les messages électroniques échangés par les acteurs dans les forums ou par messagerie dans deux univers distincts :

- **L'entreprise** : nous avons étudié les échanges sur une centaine de forums professionnels ouverts à tous dans une entreprise ;

- **Le grand public** : nous nous sommes concentrés d'une part sur les échanges dans un forum grand public et d'autre part sur les messages échangés entre les abonnés à un fournisseur d'accès et ce fournisseur.

Internet donne dans un premier temps l'illusion de la transparence : comme les échanges dans les forums utilisent principalement l'écrit, ils semblent être immédiatement disponibles pour être analysés, contrairement aux échanges par téléphone ou en face-à-face. Mais, dans la pratique, la constitution des corpus s'avère complexe, car chacun des supports suppose des formats spécifiques et les outils existants ne sont pas toujours adaptés pour l'analyse de ce type de données.

2. Constitution et préparation des corpus

Les données sur internet évoluent rapidement : par exemple, sur les forums, la durée de vie d'un message ne dépasse pas quelques semaines. Cette variation permanente des données nous a contraints à mettre en place des dispositifs dynamiques d'archivage de l'information pour constituer nos corpus de travail. Ainsi, avons-nous constitué un certain nombre de corpus bruts correspondant à chacun des supports :

- Pour les forums internet, nous avons recueilli l'ensemble des messages échangés sur cinq forums NNTP réservés aux abonnés sur la période novembre 98-avril 99, soit environ 64 000 messages. Sur l'intranet étudié, nous avons recueilli les messages sur une centaine de forums au format NNTP, soit environ 25 000 messages entre mars 99 et juin 99.
- Pour les messages électroniques, le fournisseur d'accès nous a transmis les bases d'archives des messages sur la période juin 98 à novembre 98, soit 120 000 messages. Nous avons constitué un corpus regroupant les questions des abonnés au fournisseur, et un autre recréant les liens entre les questions des abonnés et les réponses du fournisseur. En effet, dans le corpus brut initial, ce lien n'est pas visible en raison de la concaténation séquentielle des messages lors de leur archivage.

Un message électronique se présente de la manière suivante, ce qui impose des contraintes pour la constitution des corpus.:

En-tête	Newsgroups: cyberiaabonnes.entraide Subject: Re: mon num 06-83-77-55-52 From: Hervé.Sique@cyberia.fr
Phrase introduisant la reprise	Thu, 21 Jan 1999 11:46:50 +0100, "Silence" <Ziglotron@cyberiafr> Tu as écrit :
Reprise de message	>Ciel ! >Je poste maintenant des messages que je n'envoie pas, contenant des petits >fichiers exe.
Message de réponse	Couic On voit bien que tu es nouveau sur le net et en particulier sur ces forums,car depuis des lustres les savants d'ici avertissent de ne jamais ouvrir un .exe de provenance inconnue. Bon courage.
Signature	A+ Amicalement Cyberpapy Hervé. TROMBINOSCOPE des CYBERIEN(NE)S http://perso.cyberia.fr/herve.sique

Figure 1 : Exemple de message électronique

A partir des corpus bruts recueillis, nous avons constitué des corpus de référence qui nous ont servi de base de travail pour réaliser des corpus adaptés aux outils d'analyse utilisés. Ces corpus de référence sont nettoyés et balisés selon les étapes suivantes :

- Balisage du corpus : chaque élément d'un message est balisé et certains champs d'en-tête sont supprimés. On donne ci-dessous un exemple de message de forum avant et après l'opération de balisage ;

Avant : Message brut	Après : Message brut balisé
Path: 193.248.15.195!not-for-mail From: anonymous <anonymous @entreprise.fr> Newsgroups: albi.arp128 Subject: Bienvenue sur le forum Date: Thu, 25 Mar 1999 14:25:57 +0100 Organization: Entreprise Lines: 2 Message-ID: <36FA3965.63FBEAD5@entreprise.fr> NNTP-Posting-Host: 192.144.39.115 Mime-Version: 1.0 Content-Type: text/plain; charset=iso-8859-1 Content-Transfer-Encoding: quoted-printable X-Mailer: Mozilla 4.05 [fr]C-INTRANET-1 (Win95; I) CC: ……..@entreprise.fr Ce site est =E0 votre disposition =E0 partir de ce jour. *signature*	 <from> anonymous</from> <number>1</number> <newsgroups> <news1>albi.arp128 </news1> </newsgroups> <sujet> Bienvenue sur le forum </sujet> <date> Thu, 25 Mar 1999 14:25:57 +0100 </date> <lignes> 2 </lignes> <messageID> anonymouscod. BEAD5@netcompagnie.fr </messageID> <x-mailer_or_x-nexsreader> Mozilla 4.05 [fr]C-INTRANET-1(Win95; I)</x-mailer_or_x-nexsreader> <message> Ce site est =E0 votre disposition =E0 partir de ce jour. *signature* </message>

Figure 2 : Balisage d'un message

- Suppression des messages repris (précédés par des chevrons) et des phrases introduisant les reprises (du type : « un tel a écrit ») ;
- Suppression des pièces jointes (documents attachés, images, sons, etc.) ;
- Suppression de tout format de texte non brut ;
- Harmonisation typographique : corrections d'accents, nettoyage de caractères parasites, etc. ;

Voici un exemple de message avant et après les étapes décrites ci-dessus :

Avant : Message brut	Après : Message brut balisé et nettoyé
Je me demande... Car comment peut-on trouver des roches ayant 4,8 milliards d'ann=E9es d'existence dans un univers ag=E9 de : 1,7 + 2 + 0,864 + 0,432 millions d'ann=E9es = =3D 3,996 millions d'ann=E9es ? L'erreur n'est que de 99,99% , une paille ... Bruce a =E9crit: > - les indiens connaissaient donc aussi l'age de l'Univers depuis le Big= > Bang avec une relative bonne pr=E9cision. Etonnant non ? > jean marc a =E9crit: > > quatre =E2ges: le Krutayuga, qui a dur=E9 1,7 million d'ann=E9es, > > le Tretayuga, qui s'est =E9tendu sur 2 millions d'ann=E9es, le Dwapar= ayuga, > > qui a dur=E9 864.000 ans et le Kaliyuga, qui doit durer 432.000 ans a= vant > > que le monde ne se d=E9truise.	<message> Je me demande... Car comment peut-on trouver des roches ayant 4,8 milliards d'années d'existence dans un univers agé de : 1,7 + 2 + 0,864 + 0,432 millions d'années 3,996 millions d'années ? L'erreur n'est que de 99,99% , une paille ... </message>

Figure 3 : Balisage et nettoyage d'un message

Notons que nous ne sommes pas parvenus à distinguer, entre les balises <message> et </message>, le corps du message de la signature. En effet, aucun signe ne marque la frontière entre le message et la signature. Les noms et les coordonnées des intervenants interviennent dans l'identification des thèmes de discussion, ce qui met en évidence des associations entre certains thèmes et certains acteurs.

Différentes expériences en statistique textuelle et fouille de textes ont montré que la qualité des résultats était sensiblement meilleure quand le corpus était au préalable soumis à des traitements linguistiques, lesquels permettent principalement :

- de distinguer les mots pleins (noms, verbes, adjectifs et adverbes) des mots grammaticaux (articles, pronoms, prépositions…) ;
- de lemmatiser les formes fléchies, c'est-à-dire de les ramener à leur entrée de dictionnaire (infinitif pour les verbes, singulier pour les noms, masculin singulier pour les adjectifs).

Nous avons donc, avant analyse, soumis le corpus de référence à des pré-traitements linguistiques (catégorisation syntaxique et lemmatisation) via les programmes de Traitement Linguistique des Textes (TLT) développés par l'équipe « Langage Naturel » du CNET à Lannion (CNET/DSM/GRI).

Ce corpus étant constitué, la deuxième phase a consisté à établir des filtres pour extraire des sous-corpus à partir du corpus de référence. Les corpus extraits ont été formatés pour être compatibles avec les outils d'analyse utilisés. Par exemple, pour analyser les champs d'en-têtes (la répartition horaire des interventions ou la répartition des intervenants), nous avons extrait les balises correspondantes (heure d'envoi <date>, expéditeur <From>). Pour traiter les contenus textuels des messages, nous avons retenu ce qui se situe entre les balises <message> et </message>.

Entretiens et enquêtes.

En plus de cette masse de données textuelles, nous avons souhaité compléter ces informations par le récit des pratiques. Dans les deux univers étudiés (entreprise et grand public), nous avons mené des entretiens semi-directifs avec des participants et des enquêtes en ligne. Ces méthodes qui permettent d'avoir le point de vue de l'acteur (ses motivations, ses perceptions, le sens qu'il donne à ses pratiques…) se révèlent être des appuis indispensables pour l'interprétation.

3. Exploration et traitement des corpus

3.1. Problèmes

Aucun outil n'est capable de traiter globalement ce type de corpus complexe, pour répondre à nos interrogations portant sur les thèmes des échanges, les modes d'interactions et de participation et la constitution de groupes. Nous avons été amenés à utiliser des fonctions de différentes logiciels pour traiter nos données : outils de statistique classiques pour les en-têtes, outils de statistique textuelle pour les corps des messages, analyse manuelle pour étudier les fils de discussion… Sur ce dernier point, aucun outil informatique n'est adapté à l'analyse des interactions (lien question/réponse et analyse des fils de discussion).

Chaque traitement produit des résultats spécifiques qui doivent être interprétés localement. Une interprétation globale doit tenir compte de chacune de ces interprétations partielles et doit être enrichie par une analyse manuelle des échanges et par tout le matériel d'enquête.

Les outils de linguistique informatique sont confrontés aux particularités langagières des échanges électroniques, sans y avoir été préparés : comment traiter les sous-langages, le caractère parfois phonétique de l'orthographe des messages, les « smileys », les abréviations et les codes locaux partagés ? Enfin, que faire des aspects multimédia des corpus étudiés ici. Les messages électroniques peuvent en effet comporter des éléments multimédia de différente nature (son, image, vidéo…). Un traitement couvrant aussi bien l'analyse des textes, des images et des sons que leurs interactions est loin d'aller de soi.

Pour finir, soulignons que la plupart des logiciels rencontrent quelques difficultés à traiter des corpus aussi volumineux que ceux que nous avons constitués. Seuls les outils d'IBM étaient en mesure de traiter le corpus des 120 000 messages électroniques.

3.2. Résultats produits sur les corpus traités

Les en-têtes des messages de forum.

Le traitement statistique des en-têtes, réalisé avec le progiciel SAS (SAS Institute Inc., 1990), permet de construire des indicateurs quantitatifs : nombre d'intervenants, répartition des intervenants selon le nombre de messages postés, nombre de sujets de discussion, distribution temporelle des interventions. Dans les forums publics et professionnels, près de 40% des messages n'obtiennent pas de réponse, ce qui montre qu'un certain nombre de messages (en raison de leur formulation, de leur sujet, du statut de leur auteur, de l'heure de l'envoi…) sont perçus comme non pertinents par les autres participants. Les degrés d'engagement des acteurs varient selon les forums. Par exemple, sur le forum public étudié 15 intervenants (sur 6 000) produisent un quart des messages, tandis que près de 60% des participants n'interviennent qu'une fois en six mois. Sur les forums d'entreprise, 35 auteurs (sur 2606) postent un quart des messages et 38% n'interviennent qu'une fois. Dans un espace qui *a priori* est ouvert à tous, se mettent ainsi en place des structurations sociales hiérarchisées. Sur chacun des espaces étudiés émerge une minorité active qui, dans certains cas, œuvre pour le bon fonctionnement du forum, tandis qu'elle peut jouer le rôle de perturbateur dans d'autres.

Le corps de messages.

Nous avons utilisé différents outils pour traite les contenus des messages : les outils d'IBM (Gouttas et Granier, 1995), Alceste (Reinert, 1993) et Lexico (Lebart et Salem, 1994). Nous avons ainsi obtenu des typologies de messages, de forums, et différents indicateurs de vocabulaire. Les analyses sont morcelées : une vision globale implique en effet un travail d'interprétation manuelle de ces différentes analyses et doit être enrichie par la prise en compte de tout le matériau recueilli lors des entretiens et des enquêtes.

- **Sur le forum grand public**, une typologie des messages a permis de valider les types d'activité (échange technique, rappel des règles de conduite et surenchères humoristiques) que nous avions identifiés à la main, puis de les quantifier. Les outils existants ne permettent pas

l'analyse des enchaînements de messages. Nous avons donc réalisé une analyse manuelle de certains fils de discussion qui a permis de dégager la structure conversationnelle de chaque type d'activité. Le croisement de cette typologie avec les en-têtes des messages a permis d'identifier des groupes d'interlocuteurs spécialisés sur certains thèmes. Ces groupes se caractérisent par des pratiques langagières spécifiques (maîtrise de la langue, de l'argumentation et des subtilités des jeux de langage). Dans ce type d'interaction médiatisée, l'écrit concentre de manière forte les traces des déterminants sociaux classiques.

- **Sur les forums d'entreprise**, nous avons identifié grâce à l'analyse des contenus des messages sur une centaine de forums quatre grandes catégories de forums décrites ci-dessous :

	Nb de forums	Nb de messages 15 /03-30/06/99	Longueur moyenne d'un fil	% de messages initiaux sans réponse	Nb moyen de messages par intervenant	% de personnes intervenues une seule fois
Forums professionnels	89	4030 (16%)	3,1	45%	4,1	48%
Forums informatiques	3	6428 (25%)	5,0	24%	9,3	31%
Forums de petites-annonces	4	12310 (48%)	4,4	43%	7,9	38%
Forums de discussions informelles	2	2787 (11%)	7,8	28%	8,7	39%

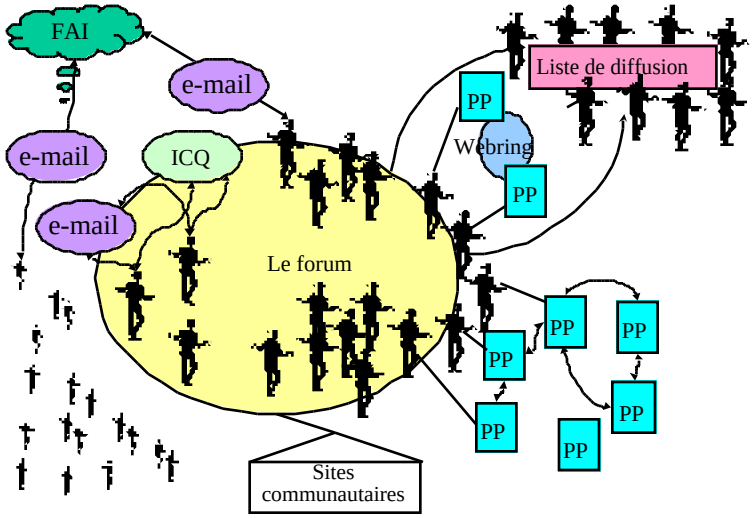
Figure 4 : Indicateurs par catégorie de forums

Ce tableau donne pour chaque catégorie de forums quelques indicateurs formels, construits à partir des champs d'en-tête, qui semblent spécifiques de ces types de forums. Comme nous sommes dans le cadre d'un intranet, il y a une très grande proportion de forums exclusivement professionnels (89 sur 98), mais ils ont en moyenne des niveaux d'activité très faibles, comparés aux quelques forums de petites annonces ou de discussion libre (6 forums qui recueillent 60 % des messages). Nous pouvons considérer comme un indicateur de la qualité des interactions le taux de messages sans suite. Hormis certains messages de type " annonce " qui n'impliquent pas de suite, les messages sans suite correspondent soit à des questions qui n'obtiennent pas de réponse, soit à des ouvertures de discussion/débat qui échouent. Ce taux est particulièrement bas (25 %) dans les forums informatique, où le niveau d'entraide est donc assez élevé. Il est également bas dans les forums de discussion informelle. En revanche, il est très élevé dans les forums professionnels ainsi que le pourcentage de personnes n'étant intervenues qu'une fois, ce qui indique des difficultés d'échange dans ce type de forum.

- **Sur les mails**, la typologie construite rend compte des différents types de problèmes auxquels sont confrontés les abonnés (par exemple, problème pour l'installation de logiciels). Elle permet d'envisager dans un premier temps un classement des messages des abonnés en fonction de la typologie construite. Dans un deuxième temps, elle doit conduire à la mise en place d'applications transactionnelles pour faciliter la résolution de ces problèmes (pour suivre l'exemple précédent : mise en place d'un dispositif en ligne d'installation de logiciel).

4. Perspectives : analyse globale d'un espace de communication

Un certain nombre d'entretiens et d'observations sur les pratiques de communication sur internet nous ont permis d'émettre l'hypothèse que les différents supports d'interaction (pages personnelles, forums, messagerie, conversation en direct...) sont interconnectés et forment un espace de communication multiforme :



Guide de lecture : PP : page personnelle ; FAI : fournisseur d'accès à internet ; les figures à l'extérieur du forum représentent les lecteurs du forum qui n'interviennent pas et des abonnés qui ne vont pas sur les forums.

Figure 5 : Espace de communication sur internet

La compréhension des interactions nécessite donc la prise en compte des pratiques sur l'ensemble des supports, et donc l'intégration de l'analyse des pages personnelles.

A la différence des forums organisés par thèmes, les pages personnelles forment un espace non structuré. Pour constituer des corpus de pages il est donc nécessaire de définir des critères de sélection et d'organisation, en fonction des axes de recherche. Nous en avons retenu deux : l'étude des pratiques d'une minorité active (participants à un forum ayant créé leur page) et la construction d'une typologie de la totalité des pages sur un serveur (Illouz et *al*, 1999). En conséquence, nous adoptons la démarche suivante :

- Extraction des URL des pages personnelles présentes dans les signatures de tous les messages postés dans le forum pendant un mois (150 sites environ). Nous étudions ainsi le groupe des participants aux forums ayant créé leur site personnel. .
- Aspiration à partir d'un site " connecteur " : nous avons observé que les liens hypertextuels entre les pages reflètaient les réseaux de sociabilité dans le forum. L'analyse du forum a permis d'identifier le " leader ", l'individu reconnu comme " guide " par les autres. Son site nous a servi de point d'entrée pour l'aspiration de tous les sites interconnectés sur le serveur, soit 345 sites. Ces sites recouvrent la liste des URL extraites à partir du forum, ce qui prouve que les relations interpersonnelles existant dans ces forums coïncident avec les liens hypertextuels définis dans les pages web.
- Aspiration d'un échantillon représentatif des sites hébergés chez ce fournisseur d'accès ;
- Aspiration des sites personnels visités par un panel d'internautes.

Une observation des premiers sites a permis de dégager quelques éléments récurrents sur les pages (sommaire, accès à la boîte aux lettres, livre d'or, rubrique de liens vers d'autres sites). Cette première observation a aussi permis d'esquisser une première ébauche de typologie des sites :

sites de professionnels, sites thématiques et sites intimistes. La deuxième catégorie prédomine chez le fournisseur d'accès étudié : la présentation de soi sur ces pages se construit principalement autour de centres d'intérêt, de passions et/ou de la description de la ville ou de la région de l'auteur et non pas autour d'une exposition de soi intimiste.

Pour le moment, nous avons constitué par aspiration un premier état du corpus des pages personnelles (au format HTML), à partir d'un site connecteur. Les pages personnelles, en raison de leur structure multimédia et hypertextuelle, imposent de nouveaux types de contraintes pour la constitution d'un corpus de référence. Cette démarche doit conduire à l'élaboration d'outils pour l'analyse des pages : représentation graphique de la structure des sites, élaboration d'indicateurs formels décrivant l'organisation multimédia des pages, construction d'une typologie des sites sur la base des contenus textuels et analyse du rapport entre la structure et le contenu.

Ce travail met en avant la nécessité de confronter les corpus traités pour mieux appréhender l'espace de communication étudié (par exemple, étudier le lien entre la présentation de soi dans la page personnelle et l'identité dans les forums). Une étude comparative de différents espaces de communication, comme les sites communautaires (Geocities, Tripod, Multimania, etc.), devra enrichir ce travail.

Références

Beaudouin V., Velkovska J. (1999). Constitution d'un espace de communication sur internet, *Réseaux*, n°97, Hermès, Paris, pp. 121-177.

Gouttas C., Granier J.-M. (1995). Les mots de l'entreprise : analyse textuelle automatique et sémiotique. In : Bolasco, Sergio, Lebart, Ludovic, Salem, André (eds), *JADT 1995 : III Giornate internazionali di Analisi Statistica dei Dati Testuali, Rome: CISU,* vol. II , pp. 127-133.

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

Illouz G., Habert B., Fleury S., Folch H., Heiden S., Lafon P. (1999). Maîtriser les déluges de données hétérogènes, Actes de *TALN'99*, Atelier Corpus et Traitement Automatique des Langues : pour une réflexion méthodologique, 12-17 juillet 1999, Cargèse.

Lebart L., Salem A. (1994). *Statistique textuelle*, Paris, Dunod.

Reinert M. (1993). Les "mondes lexicaux" et leur logique. *Langage et société,* Paris, Maison des Sciences de l'Homme, n°66, p. 5-39.

SAS Institute Inc. (1990). *SAS language : Reference*, Version 6, First Edition, Cary, NC : SAS Institute Inc.