# TyPWeb : décrire la Toile pour mieux comprendre les parcours

Valérie Beaudouin*, Serge Fleury[+], Benoît Habert°^, Gabriel Illouz°, Christian Licoppe*, Marie Pasquier*^

* FT R&D/DIH/UCE
[+] SYLED – CLA2T Université Paris 3
° Limsi – CNRS
^ MoDyCo UMR 2329 – Université de Paris X
38-40, rue du Général Leclerc
92794 Issy-les-Moulineaux Cedex 9
{valerie.beaudouin, christian.licoppe, m.pasquier}@francetelecom.com
serge.fleury@univ-paris3.fr
{habert, gabrieli}@limsi.fr

## 1 Résumé

Parallèlement et en interaction avec une analyse des parcours des internautes[1] et des entretiens auprès des concepteurs de sites, est développée une architecture, TyPWeb, permettant l'analyse des sites sur les plans textuel, structurel et hypertextuel. Nous présentons cette architecture, les problèmes qu'elle affronte et les solutions que nous avons retenues. Nous montrons les distinctions qu'elle permet de mettre en évidence entre des échantillons significatifs de sites personnels et de sites marchands, à partir de l'examen des liens au sein des sites et des sites vers l'extérieur, mais aussi à partir de lexiques déterminés (pronoms personnels, mots outils, mots anglais ou français très fréquents).

## 2 Contexte et hypothèses

Une connaissance fine des usages d'internet passe par l'analyse des parcours des internautes à travers les différents services proposés (Web, messagerie, forums…). Or pour décrire finement les types d'activité sur le Web, il est indispensable de pouvoir qualifier les contenus des pages visitées, autrement dit, de donner un sens aux URL visitées, qui sont par elles-mêmes vides de sens. L'étude des parcours est donc indissociable de l'étude des productions de la Toile. C'est dans cette perspective que nous avons mis en place le projet TyPWeb, qui porte sur l'analyse sémantique et structurelle des sites, à la fois pour donner sens aux parcours, et pour montrer comment la structure et les contenus des sites conditionnent les visites.

Le Web est un nouveau moyen d'expression et d'information, qui reprend en les transformant partiellement, des « genres textuels » traditionnels (la lettre personnelle par exemple) et qui ajoute de nouveaux genres (le « chat »). C'est pourquoi on ne peut pas plaquer un inventaire prédéfini de genres importé d'autres modes de communication. Il s'agit plutôt de faire émerger, de manière inductive, des typologies sur la base des corrélations observées entre des indicateurs portant sur l'outillage grammatical et le lexique, sur la structuration textuelle et hypertextuelle, et sur l'aspect multimédia. La construction de ces typologies se nourrit des entretiens menés auprès des concepteurs de sites marchands et de sites personnels comme de l'observation fine de tels sites.

Ce double angle d'attaque nous permet de formuler les hypothèses suivantes :

1. Les sites marchands sont plus nettement complexes (nombre de pages, de liens internes et externes, d'images, etc.) que les sites personnels.

2. L'utilisation et la répartition des pronoms personnels sont très différenciées selon les types de sites: elles constituent un critère opératoire de distinction. La prise en charge par un locuteur de ce qui est dit est clairement affirmée dans les pages personnelles. A l'inverse, dans les sites marchands, le discours est orienté vers le destinataire ou client potentiel : l'utilisation de la première personne est nettement plus rare et relève d'une stratégie de marketing marquée.

3. La redondance, c'est-à-dire la multiplication des moyens d'accès distincts à un même « objet » (par exemple, sur une même page, un texte, une image, une icône, etc., peuvent pointer vers un même document), fonctionne différemment pour les sites marchands, où elle est systématique, et pour les sites personnels. Pour ces derniers, la redondance semble ciblée sur les éléments communicationnels (*mail to*, changement de langue, etc.). Elle est maximale sur la page d'accueil des sites marchands, afin d'autoriser tous les parcours possibles, ce qui n'est pas le cas sur les autres pages.

4. Le serveur d'hébergement joue un rôle marquant dans la forme et dans le contenu des pages des sites personnels : structure de la page d'accueil, présence et organisation de l'index thématique, mise en avant de certaines pages personnelles, apport d'outils et de conseils d'aide à la conception de sites.

Nous constatons une confusion croissante entre sites marchands et personnels, et l'émergence, parmi les sites dits personnels, d'une catégorie de sites intermédiaires : sites d'associations, de collectivités, d'indépendants et de PME. Les traits retenus doivent permettre de discriminer entre les types de site en terme de thématiques ou de profils utilisateurs et/ou concepteurs.

[1] Dans le cadre d'un partenariat entre France Télécom R&D, Wanadoo SA, NetValue et HEC.

## 3 TyPWeb : une adaptation au Web d'une architecture de typologie des textes

Le projet TyPTex[2], qui s'achève en 2001, a pour objectif de classer les textes d'un corpus en fonction des « styles » employés, c'est-à-dire de l'outillage grammatical et des structures privilégiés, pour pouvoir ensuite adapter les outils de traitement à ces styles (Folch et *al*., 2000). En effet, les outils de traitement des textes (étiquetage, analyse syntaxique, segmentation thématique...) ont des performances qui varient significativement selon le style (Illouz 2000). Un texte scientifique par exemple aura ainsi plus volontiers recours aux passifs sans agents (« l'action de la dopamine *a été isolée* par... ») et aux phrases impersonnelles en « on » (« *On* constate une variation sensible... ») qu'un récit d'épreuve des jeux olympiques. Un étiqueteur probabiliste entraîné sur le premier type de texte aura de moins bonnes performances sur le second. En outre, s'il est entraîné sur un mélange de styles, il sera moins précis sur un style donné que s'il avait été entraîné sur un ensemble de textes relevant de ce style uniquement.

Il ne s'agit pas de retrouver des styles pré-définis (narratif, descriptif, explicatif, argumentatif, poétique, etc.) mais de regrouper (*clustering*) des documents (ou des portions de documents) en fonction de l'emploi qu'ils font de l'outillage grammatical (pronoms, temps et modes...) et de certains marqueurs lexicaux spécifiques (par exemple, types sémantiques d'adverbes : négation, possibilité, temps et espace...). La classification des documents se fait donc sur la base de traits linguistiques fins articulant étiquetage grammatical et projection de dictionnaires spécifiques (classes sémantiques d'adverbes ou de conjonctions de subordination, par exemple). Les premiers résultats manifestent une bonne corrélation entre ces « styles » induits et les performances de modules de traitement du langage comme l'étiquetage ou le parsage.

L'exploration qualitative des sites Web, professionnels ou personnels, rencontre cette diversité de styles. Au delà de cette perception immédiate, la reprise de styles pré-existants dans d'autres médias comme l'invention de nouveaux styles adaptés aux possibilités d'Internet participent des stratégies communicationnelles conscientes ou non des développeurs de sites. L'inventaire des styles effectivement employés et des types de sites associés contribue ainsi à une connaissance fine des usages.

Les outils développés dans le cadre de TyPTex (étiquetage morpho-syntaxique et marquage typologique de corpus, passerelles vers des traitements statistiques standard, visualisation des résultats et rétro-projection des styles répertoriés sur les documents de départ...) doivent intégrer les phénomènes propres aux sites Web :

- les textes comportent de nombreuses scories. Ils passent brutalement d'une langue à l'autre, et sont « bruités » par les balisages de mise en forme ou de structure. Les outils de TAL robustes (étiquetage, parsage partiel) nécessitent des pré-traitements pour pouvoir être mis à contribution ;
- le français employé est parfois intermédiaire entre oral et écrit (*smileys*, onomatopées, orthographe « souple », « oralismes »), ce qui contrarie également le travail des outils disponibles, adaptés à un écrit plus contrôlé ;
- la structure joue un rôle primordial dans le mode d'interaction, associée à la dimension multimédia et hypertextuelle, alors que les outils d'ingénierie linguistique travaillent dans le cadre de la phrase et ne sont pas conçus pour les textes en tant que tels.

## 4 Corpus et chaîne de traitement

Nous présentons dans cette section les logiques qui ont présidé à la constitution des corpus et l'architecture de la chaîne TyPWeb.

### *4.1 Logiques de constitution des corpus*

Pour aboutir à des résultats prometteurs et réutilisables à plus grande échelle en un temps limité, nous avons restreint notre étude à deux grandes catégories de sites français : les sites personnels et les sites marchands, sites *a priori* diamétralement opposés. Ils sont choisis comme représentants du Web gratuit et du Web marchand. Gensollen (1999) mobilise cette distinction pour montrer comment le premier par les externalités très fortes qu'il produit donne de la valeur au second.

Nous sommes conscients néanmoins du fait que ce sont deux catégories de sites qui tendent progressivement à nouer des liens entre elles (apparition de bannières commerciales dans les sites personnels, diffusion d'informations ciblées en fonction des attentes des visiteurs sur les sites marchands, participation directe des sites personnels au développement des sites marchands en servant de moyen de rabattage efficace, etc. (Gensollen, 1999)).

ß Méthodologie de constitution de corpus web

Le web est un médium en perpétuel changement. Il importe donc de pouvoir garder une photographie d'une partie de la toile à un instant *t*, pour pouvoir mesurer les évolutions dans le temps. C'est particulièrement crucial pour les sites marchands.

Dans un premier temps, il s'agit d'aspirer les sites web, c'est-à-dire de recopier localement tout ou partie des sites retenus : nous parlons alors de « sites miroirs locaux ». Nous utilisons des logiciels d'aspiration de

[2] Typage et Profilage de Textes, réalisé au LIMSI sous la responsabilité de B. Habert et soutenu financièrement par ELRA (European Language Resources Association) dans le cadre de l'appel d'offres *Contribution à la réalisation de corpus du français contemporain* de la DGLF (Direction Générale de la Langue Française).

sites, ou « aspirateurs », disponibles sur le marché[3]. Dans un second temps, il faut rendre manipulables de manière uniforme les différents composants des pages des sites : texte, structure, liens, images, applets... Le passage au format XML permet en particulier de s'affranchir des incohérences et de la mauvaise qualité du code HTML des pages aspirées.

ß Descriptif des corpus

Ainsi, les différents objectifs retenus pour la constitution de corpus conditionnent la stratégie d'aspiration mise en œuvre, la taille des corpus et leurs mises à jour. Actuellement, nous disposons de plusieurs corpus de sites web :

1. Un premier corpus est constitué par les pages personnelles de participants actifs dans un forum. Une précédente étude sur les interactions électroniques (Beaudouin et Velkovska, 1999) a montré comment les différents supports de communication (pages personnelles, forums, mail, chat…) étaient entrelacés par les pratiques des acteurs et comment les liens d'amitié entre les participants du forum étaient reflétés par les liens entre leurs pages personnelles. En déployant l'aspiration (avec le logiciel WinHTTrack) à partir du site d'une des figures centrales du forum, le « leader », ont été aspirés les sites des habitués du forum. Ainsi, ce premier corpus ne contient que des pages personnelles hébergées par Wanadoo élaborées par les participants actifs dans un forum. Au total, ce corpus, constitué l'été 1999 et noté PPWanadooForum-été99, contient 539 sites personnels.

2. Le second corpus est constitué de sites personnels hébergés chez Wanadoo qui ont été *visités* par des internautes du panel Netvalue[4] en mars 2000. C'est le point de vue de la réception, celui de la visite, qui est ici retenu pour constituer le corpus[5]. Le corpus, aspiré en juillet et août 2000 et noté PPWanadoo-mars00, se compose de 568 sites.

3. La volonté de pouvoir vérifier s'il existe des différences tangibles dans la forme et le contenu des pages selon le type d'hébergeur constitue le critère de constitution du troisième corpus web. Nous avons retenu une vingtaine de serveurs d'hébergement, tels que, `multimania free`, `wanadoo`, `libertysurf`, etc., et aspiré toutes les pages consultées par au moins deux panélistes NetValue entre janvier et juin 2000. Constitué au cours des mois de janvier et février 2001, ce corpus contient 12070 pages (Hébergeurs15000PP).

4. L'étude de la conception et de l'évolution des sites marchands pour lesquels des entretiens approfondis ont été menés auprès des responsables et des prestataires de technologies (Licoppe 2000) nous a permis de créer un corpus de sites marchands (45 sites). Les entretiens sont menés à deux reprises à un an d'intervalle et les sites sont aspirés parallèlement aux entretiens. De ce fait, certains sites marchands ont été aspirés deux fois à un an d'intervalle. Cela nous permet de mesurer leur évolution dans le temps. Nous avons d'ailleurs divisé ce corpus en 3 sous-corpus qui correspondent à des dates d'aspiration différentes[6]. Le premier corpus (Marchands-99-00) contient 29 sites marchands aspirés entre août 1999 et février 2000 ; le second (Marchands-été2000) en contient 16 aspirés pendant l'été 2000 ; et le dernier (Marchands-Voyage) est un sous-corpus thématique (sur le thème du voyage) puisqu'il comporte 10 sites marchands déjà archivés ou nouvellement aspirés.

Pour l'identification de traits spécifiques, nous avons travaillé avec 4 corpus concaténés : les deux premiers corpus de sites personnels (PPWanadoForum-été99 et PPWanadoo-mars00) et deux corpus de sites marchands (Marchands-99-00 et Marchands-été00). Pour se faire une idée de la taille de ce corpus, nous en donnons les principales caractéristiques dans les tableaux 2, 3 et 4.

Afin d'étudier les spécificités des hébergeurs de pages personnelles, seul le corpus Hébergeurs15000PP a été utilisé.

### *4.2 Chaîne de traitement*

Les traitements réalisés se décomposent de la manière suivante (la numérotation utilisée reprend celle de la figure 1 qui résume l'architecture) :

- (1) sélection d'une liste de sites web à analyser ;
- (2) aspiration de sites : construction de sites miroirs, consultables « off-line » ;
- (3) normalisation des sites web miroirs ;
- (4) sélection de traits pertinents pour le travail typologique ;
- (5) construction de matrices croisant les sites et/ou

[3] Nous avons utilisé deux aspirateurs : WinHTTrack (hypertoile: = httrack.free.fr) et Teleport Pro (logiciel commercialisé par la société Tennyson Maxwell Information System, hypertoile: = www.tenmax.com/teleport/home.htm) et, après comparaison, avons opté pour le premier (Maisondieu & Kuncova 2000).

[4] NetValue est un des principaux acteurs de la mesure d'audience sur Internet. Un partenariat a été monté entre France Télécom R&D, Wanadoo SA, Netvalue et HEC pour mener des explorations approfondies sur des thèmes précis, en particulier sur la consultation et les contenus des pages personnelles Les données d'usage proviennent du panel résidentiel français de NetValue.

[5] Aspirations effectuées avec Teleport Pro.

[6] Dans un premier temps, nous aspirions le site dans sa totalité. Toutefois, nous nous sommes aperçus au fil du temps que l'aspiration des sites marchands devenait problématique. En effet, ce sont des sites ayant un contenu de plus en plus étendu et dense, rendant l'étape d'aspiration longue et gourmande en capacité de stockage. De plus, dès qu'un site est très volumineux, la normalisation des sites n'est pas toujours possible. Afin de réduire significativement la taille du corpus sans perdre des informations majeures, nous avons donc décidé de limiter l'aspiration à 5 niveaux de profondeur.

les pages et les traits choisis ;
- (6) traitements statistiques des matrices ;
- (7) interprétation des résultats.

### *4.3 Constitution d'un corpus de sites sur l'hypertoile*

Les sites aspirés sont sauvegardés localement après aspiration. L'aspiration des sites pose les problèmes techniques suivants (Maisondieu & Kuncova 2000) :

- pour gérer cette aspiration : les temps de traitements consacrés à l'aspiration peuvent être relativement longs suivant la taille des sites visités (l'aspiration d'un site marchand peut prendre plusieurs jours) ; cette aspiration peut aussi comporter des difficultés techniques liées aux architectures mises en place pour construire les sites (scripts, programmes…) ;
- pour gérer le stockage des données aspirées, des données modifiées ou générées.

Pour illustrer ces difficultés techniques, nous présentons dans les tableaux qui suivent des indicateurs quantitatifs sur les temps d'aspiration, sur les volumes des données aspirées (nombre de sites, des pages), et sur les volumes des corpus normalisés construits *in fine*.

ß **Aspiration**

Le tableau présenté ci-dessous donne un état des temps d'aspiration réalisés lors de la création des sites miroirs du corpus des pages visitées (Hébergeurs15000PP).

| Hébergeur | Durée aspiration | Fichiers scannés | Fichiers écrits | Taux de transfert (byte/second) |
|---|---|---|---|---|
| Multimania | 53 mn 59 sc | 2856 | 2804 | 7180 |
| chez | 9 mn 21 sc | 1828 | 1826 | 17558 |
| citeweb | 5 mn 7 sc | 202 | 88 | 1812 |
| free | 24 mn 22 sc | 3687 | 3158 | 10602 |
| geocities | 15 mn 32 sc | 644 | 625 | 6103 |
| home-nordnet | 40 sc | 123 | 122 | 13741 |
| ifrance | 4 mn 27 sc | 707 | 706 | 25783 |
| le-village | 32 sc | 13 | 12 | 6161 |
| members-aol | 1 mn 24 sc | 214 | 211 | 19309 |
| perso-club-internet | 2 mn 51 sc | 534 | 532 | 21522 |
| perso-infonie | 1 mn 49 sc | 303 | 301 | 21417 |
| perso-libertysurf | 1 mn 54 sc | 198 | 196 | 9211 |
| perso-wanadoo | 10 mn 36 sc | 1524 | 1521 | 22885 |
| respublica | 19 mn 6 sc | 1060 | 1053 | 4929 |
| tripod | 2 mn 57 sc | 37 | 36 | 3781 |
| webhome-infonie | 1 mn 31 sc | 35 | 25 | 2399 |

***Tableau 1 : temps d'aspiration sur Hébergeurs15000PP***

ß **Tailles des données**

Le tableau 2 fournit des données détaillées pour une sous-partie du corpus PPWanadoo-mars00, le tableau 3 pour les corpus PPWanadoForum-été99, PPWanadoo-mars00, Marchands-99-00 et Marchands-été00 et le tableau 4 pour Hébergeurs15000PP.

| PPWanadoo-mars00 (partie de) | |
|---|---|
| Nb de sites | *534* |
| Nb de pages | *11 129* |
| Nb moyen de pages par site | *20,84* |
| Nb de mots/occurrences | *150 316 / 3 862 198* |
| Nb moyen de mots/occurrences par site | *281 / 7 232* |
| Nb moyen de mots/occurrences par page | *13 / 347* |
| Nb de liens | *69 932* |
| Nb de liens internes (fichiers) | *43 046* |
| Nb de liens externes | *11 433* |

***Tableau 2. PPWanadoo-mars00 (partie de)***

| | PPWanadoo Forum-été99 | PPWanadoo-mars00 | Marchands-99-00 | Marchands été00 | Total |
|---|---|---|---|---|---|
| Nb de sites | *539* | *568* | *29* | *16* | *1 162* |
| Nb de pages | *11 006* | *24 938* | *29 199* | *5 726* | *96 885* |
| Moyenne pages/site | *20* | *44* | *1 007* | *358* | *83* |
| Nb d'occurrences | *3878647* | *10 577 421* | *3 090 399* | *1 284 664* | *18 831 131* |
| Nb de formes | *148 360* | *348 092* | *66 635* | *53 805* | *616 892* |
| Nb d'éléments HTML | | | | | *13 882 836* |
| Nb de formes HTML | | | | | *349* |
| Fichiers XML (en ko) | *292 074* | *1 029 274* | *450 433* | *159 434* | |

***Tableau 3. Caractéristiques des corpus***

| Hébergeurs15000PP | |
|---|---|
| Nb de pages | *11932*[7] |
| Nb d'occurrences | *2 973 692* |
| Nb de formes | *171 983* |
| Nb d'éléments HTML | *2 254 352* |
| Nb de formes HTML | *289* |
| Taille fichiers XML (en Ko) | *372 170* |

***Tableau 4. Caractéristiques du corpus Hébergeurs15000PP***

## *4.4 Normalisation des corpus de sites*

### *4.4.1 Descriptif de la chaîne TyPWeb*

La première phase est le « désossage », c'est-à-dire la description formelle des éléments composant une page HTML. Ce travail réalisé sur un site complet vise à produire une analyse de toutes les pages d'un site et la production d'un rapport détaillant l'ensemble des éléments présents dans chaque page (éléments textuels et éléments structurels). L'outil construit est l'adaptation du programme *webxref035*[8].

La mise à jour de `webXref038` permet de parcourir plusieurs sites, en appliquant aux documents trouvés une fonction qui extrait de manière récursive les éléments HTML et les éléments textuels qu'ils enchâssent. `WebXref038` produit des tableaux contenant les références des documents trouvés, URLs, ancres, images et fichiers[9]. Cette étape de normalisation des sites a conduit à la mise en place d'un modèle de représentation des sites. Ces modèles tiennent compte à la fois de la cartographie interne du site (les arêtes du réseau de liens entre les pages du site) du contenu multimédia des pages traités (les nœuds du réseau). Chaque nœud du réseau (chaque page) est éventuellement lui-même un micro-réseau de liens : une page peut être constituée d'ancres pour faciliter la navigation. Les rapports produits par `WebXref038` sont transcodés au format XML, en tenant compte des éléments propres au site global (structure et contenu) et des éléments propres à chaque page (idem). L'articulation des zones textuelles et des éléments structurels est maintenue dans la normalisation du site traité. Un traitement particulier a été mis en œuvre pour préparer les contenus textuels bruts.

### *4.4.2 Résistance et complexité des données*

Les traits disponibles dans le corpus normalisé sont à la fois très nombreux et très complexes. En ce qui concerne les traits structurels comptabilisés dans les tableaux 2, 3 et 4 précédents, il convient de préciser que les formes comptabilisées (les éléments HTML) sont généralement accompagnées d'une liste d'attributs associés à des valeurs particulières : par exemple un élément `FONT` (police) peut être caractérisé par le type de police, par sa taille, sa couleur… Il reste donc à trouver pour chaque attribut de ce type les découpages pertinents.

En outre, les données textuelles présentes dans les pages HTML ne se laissent pas « saisir » aisément, en raison principalement de l'éclatement local du texte par les balises (lettrines par exemple) mais aussi des « sauts » d'une langue à l'autre (surtout avec l'anglais).

Soulignons enfin la faible proportion de texte dans de nombreuses pages (tableau 2).

---

[7] Le nombre de pages est plus faible que le nombre de pages visitées (15000) car beaucoup ont disparu entre le moment de leur visite et leur aspiration.

[8] *Webxref035*, écrit par Rick Jansen en juin 1995, est distribué par le site PERL (hypertoile : = http://www.perl/com). C'est un programme Perl conçu pour vérifier rapidement un ensemble local de pages HTML et mettre au jour les dysfonctionnements possibles.

[9] Via les algorithmes utilisés dans le script *dissectsite.hts* : utilisable dans un navigateur, il prend en argument une URL (en ligne) et génère des variables stockant l'information sur les en-têtes et les éléments envoyés par le serveur (TEXT objects ou TAG objects). Les résultats produits sont présentés au format HTML dans une nouvelle fenêtre du navigateur (hypertoile : = http://worldwidemart.com/scripts/).

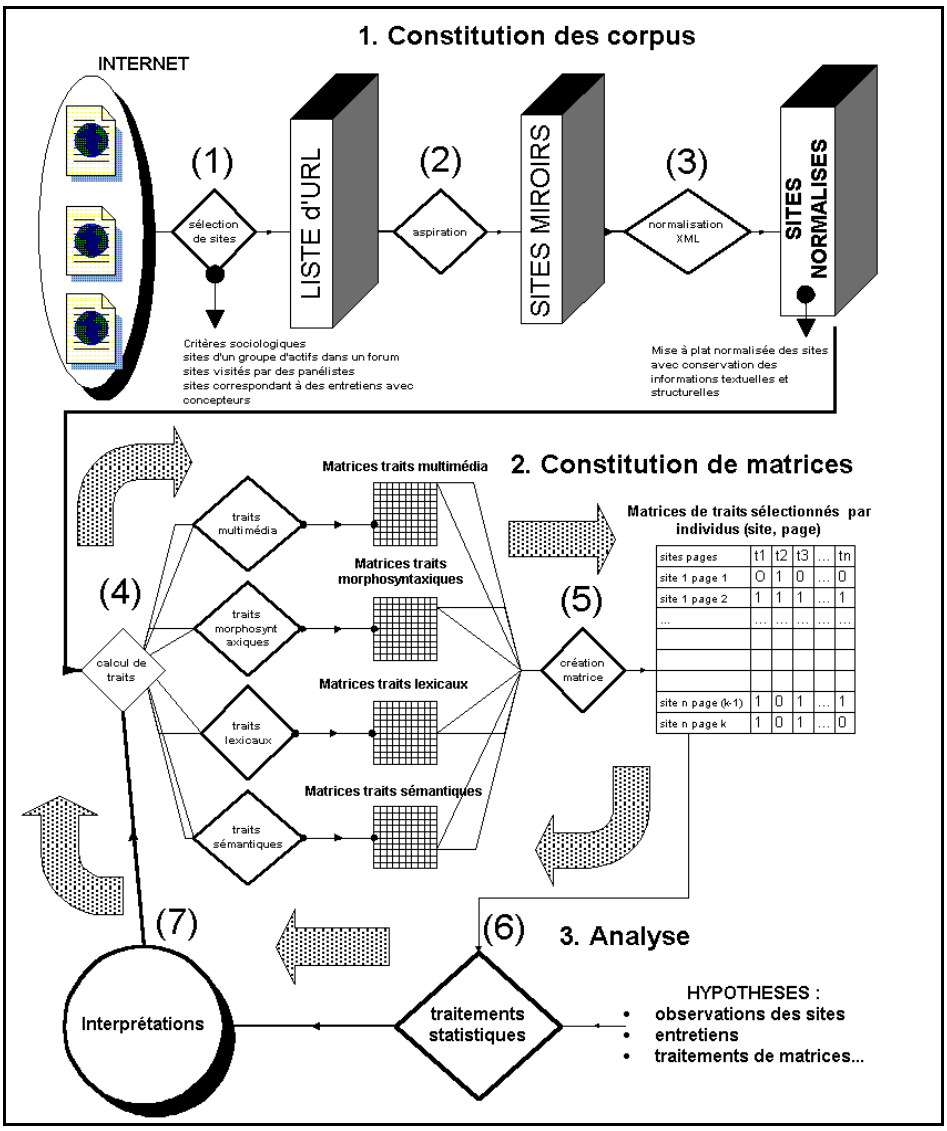


***Figure 1. Architecture TyPWeb***

### *4.5 Corpus de sites analysés*

La phase de normalisation conduit à la construction des éléments suivants :

1. Un corpus XML de référence regroupant l'intégralité des sites normalisés.
2. Des états statistiques sur la composition des pages de ce corpus (comptage de mots, d'éléments HTML : TAG, attributs, valeurs d'attribut…), et des états statistiques sur les sites. A partir du moment où le nombre de traits de ce corpus, de l'ordre de 200 000, aboutirait à une matrice ingérable par les outils statistiques disponibles, nous sélectionnons pour chaque étude particulière un sous-ensemble dans cette liste de traits, comme celui des pronoms personnels dans l'étude *infra*.
3. Des corpus « spécialisés » construits à partir de telles sélections : corpus « textuel » contenant uniquement les parties textuelles des pages des sites normalisés, corpus de TAG HTML contenant uniquement les éléments structurels de ces mêmes pages, etc.

Les représentations de ces corpus sont ensuite soumises, après reformatage, à des outils d'analyse tels que R[10], Lexico3[11], Alceste[12], SAS[13], Cordial[14], Tropes[15]…

[10] Hypertoile : = www.r-project.org

[11] LEXICO3 est un logiciel conçu pour le traitement lexicométrique de textes. Il a d'abord été développé par André Salem (ILPGA - Paris 3). Il est désormais maintenu par l'équipe LEXICO de l'UPRES SYLED (Paris 3). La version LEXICO3 est utilisable sous Windows. (hypertoile: = www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/)

[12] ALCESTE est un logiciel d'analyse des données textuelles. ALCESTE a d'abord été développé par Max Reinert (CNRS,

## 5 Identification de traits spécifiques

L'observation fine des sites et les entretiens nous ont conduit à tester l'architecture TyPWeb pour explorer deux questions : celle de l'utilisation des pronoms personnels et celle de la structure des liens, qui nous semblaient être des facteurs de différenciation des sites.

### *5.1 Utilisation et répartition des pronoms personnels*

Nous avons défini 6 catégories de pronoms fondées sur le nombre (singulier/pluriel) et la personne grammaticale qui est désignée : p1 ={*je/j, me/m, moi*}, p2 ={*tu, t, toi*}, p3 ={*il, elle, on, lui, soi*}, p4 ={*nous*}, p5 ={*vous*}et p6 ={*ils, elles, leurs*}. Nous avons délibérément omis d'ajouter les pronoms *le, la* et *les* trop ambigus. Nous avons choisi de classer les pronoms *on* et *soi* avec les pronoms à la 3ème personne du pluriel. Enfin, pour obtenir des résultats significatifs, nous avons uniquement pris en compte les sites dans lesquels apparaissent au moins 10 pronoms : nous réduisons ainsi notre corpus à 705 sites, soit 61 % du corpus initial. En fonction de ces choix, la répartition des pronoms est donnée dans le tableau 5.

| | corpus total (705 sites) | PPWanadoo Forum-é99 (239 sites) | PPWanadoo-mars00 (430 sites) | Marchands-0999 (22 sites) | Marchands-0800 (14 sites) |
|---|---|---|---|---|---|
| p1 | 24 | 26 | 23 | 13 | 14 |
| p2 | 5 | 5 | 5 | 2 | 3 |
| p3 | 34 | 35 | 34 | 25 | 33 |
| p4 | 10 | 10 | 10 | 11 | 11 |
| p5 | 21 | 18 | 20 | 44 | 34 |
| p6 | 6 | 6 | 6 | 5 | 5 |
| Σ | 100 | 100 | 100 | 100 | 100 |

***Tableau 5. Répartition des pronoms (en %)***

L'examen des pronoms met en évidence une séparation nette entre sites personnels et sites marchands. En effet, l'utilisation de la première personne du singulier (*je, me, moi*) est caractéristique des sites personnels, tandis que la deuxième personne du pluriel (*vous*) est spécifique des sites marchands. Ainsi, un site personnel est un lieu de présentation de soi centré sur un individu (le concepteur du site qui se présente et se raconte). Il existe cependant une forme d'équilibre entre l'émetteur (*moi*) et le destinataire (*toi* ou *vous*) qui montre bien que la page personnelle tend vers l'échange. Au contraire, les sites marchands sont résolument adressés à un visiteur qui peut devenir un acheteur potentiel au fil des consultations : l'émetteur s'efface au profit du destinataire. Nous voyons donc ainsi que l'analyse des occurrences des pronoms est capable de discriminer empiriquement des classes de sites, y compris dans des corpus de sites web très volumineux.

En étudiant uniquement la répartition des pronoms aux 1ères et 2èmes personnes (i.e. les catégories p1, p2, p4 et p5), la rupture observée entre sites personnels et sites marchands est encore plus nette (tableau 6) et permet d'affiner les analyses par type de corpus.

| | corpus total (665 sites) | PPWanadoo Forum-é99 (219 sites) | PPWanadoo-mars00 (410 sites) | Marchands-0999 (22 sites) | Marchands-0800 (14 sites) |
|---|---|---|---|---|---|
| p1 | 41 | 46 | 41 | 18 | 21 |
| p2 | 8 | 8 | 9 | 3 | 9 |
| p4 | 17 | 17 | 17 | 16 | 18 |
| p5 | 34 | 29 | 33 | 63 | 52 |
| Σ | 100 | 100 | 100 | 100 | 100 |

***Tableau 6. Répartition des pronoms aux 1ères et 2èmes personnes (en %)***

Il est intéressant de constater que les sites personnels des participants du forum contiennent légèrement plus de pronoms à la première personne du singulier (46%) que les sites personnels visités (41%) : les concepteurs de ces derniers sites tendent à s'effacer derrière la thématique de leur site (ex. : présentation de sa ville natale, mise en ligne d'astuces informatiques, etc.), ce qui tempère l'apparition directe du *je*. L'observation de la répartition des couples de pronom *je/nous* et *tu/vous* accentue d'autant plus la spécificité des deux corpus de sites personnels : le premier est davantage centré sur le concepteur (63% de *je*/*nous* contre 37% de *tu*/*vous*), tandis que le second présente un équilibre entre la présence du concepteur de site et celle du visiteur (58% de *je*/*nous* contre 42% de *tu*/*vous*).

En parallèle, nous observons la présence massive du *vous* dans les deux corpus de sites marchands (63 et 52%) et du couple *tu*/*vous* (66 et 61% de *tu*/*vous* contre seulement 34 et 39% de *je*/*nous*). Au sein des sites marchands, des distinctions doivent cependant être faites. Ainsi, l'examen de la répartition des pronoms dans le corpus thématique de voyage (Marchands-Voyage) fait apparaître des stratégies de marketing nettement différenciées. La majorité des sites utilisent massivement le *vous* en considérant le visiteur soit comme un acteur (*vous découvrirez, vous traverserez, vous survolerez*), soit comme un spectateur (*nous vous ferons découvrir, nous vous ferons traverser*, etc.). Néanmoins, quelques sites se positionnent différemment en privilégiant l'utilisation des pronoms *je* et *nous* indépendamment ou le couple *je/nous*. Actuellement, nous sommes en train d'interpréter ces résultats plus en détail en les croisant avec les discours des concepteurs des sites.

Université de Toulouse-Le Mirail). ALCESTE est disponible via la société IMAGE, spécialisée en mathématiques appliquées.

[13] Hypertoile : = http://www.sas.com

[14] Hypertoile : = http://www.synapse-fr.com/

[15] Hypertoile : = http://www.acetic.fr

La répartition des pronoms personnels permet donc de faire une première séparation marquée entre sites personnels et sites marchands. Nous continuons à travailler dans cette voie afin d'affiner la granularité des résultats. En outre, cette répartition gagnera à être reliée à l'emploi d'autres traits morphosyntaxiques (mode de verbes : indicatif, impératif ; formes impersonnelles…).

### *5.2 Liens et redondance*

Outre le nombre de pages par site et le nombre de liens par page, nous avons construit les indicateurs suivants : nombre de liens hypertextuels *internes* (pointant vers une autre page du site) et *externes* (pointant vers un autre site), nombre de liens vers des images internes et externes, nombre de liens vers une boîte-aux-lettres ou vers un fichier FTP pour le téléchargement.

La première opposition majeure entre pages personnelles et sites marchands se construit autour du nombre de pages du site : les sites marchands ont en moyenne 20 fois plus de pages que les sites personnels. Ils ont globalement beaucoup plus de liens, mais il n'y a pas de différence très marquée dans le nombre moyen de liens par page entre sites personnels et marchands.

Des différences sensibles existent entre les deux corpus de pages personnelles. Les sites visités par des panélistes sont en moyenne plus gros que les sites des participants au forum (44 pages par site contre 25). Ils ont davantage de liens externes (1,9 par page contre 1,2) et surtout davantage de liens internes (10 par page contre 6). La taille d'un site, l'importance des liens internes qui facilitent la navigation, l'insertion dans un réseau d'interconnexion de sites semblent être des conditions nécessaires pour pouvoir bénéficier de visites. Les pages personnelles du forum sont plus simples et moins ouvertes sur l'extérieur. Nous avons montré (Beaudouin & Velkovska 1999) que le réseau des pages personnelles du forum était le reflet des relations amicales au sein du forum : les pages personnelles constituent alors un réseau dense mais assez fermé.

Sur les sites marchands, on observe un contraste important entre la page d'accueil et les autres pages. Alors qu'il y a en moyenne 4 liens internes par page, il y en a 11 sur la page d'accueil. Pour les liens externes, il y en a 3 sur la page d'accueil contre 0,1 sur les autres. Ce contraste est beaucoup plus faible pour les pages personnelles et signale donc une spécificité des sites marchands.

La page d'accueil du site marchand est donc un concentré de liens hypertextuels internes et externes. De plus, les liens y sont particulièrement redondants : une même rubrique peut être atteinte en cliquant en divers endroits de la page d'accueil : tantôt sur une zone de texte, tantôt sur une petite icône, tantôt sur une bannière… Nous avons constitué un indicateur de redondance (nombre de liens total divisé par le nombre de liens différents) qui vaut 1 si aucun lien n'est redondant et dont la valeur augmente avec la redondance. Cet indicateur vaut 1,25 sur les pages d'accueil contre 1,1 sur les autres pages.

La densité en liens et la redondance des liens dans les pages d'accueil constituent autant de points d'entrée dans le site selon le niveau d'expertise du visiteur (accès pour novice *versus* expert) et le type de parcours autorisé ou non par la structure même du site. Ainsi, la composition des sites se répercute bel et bien sur les parcours.

## 6 Spécificités des hébergeurs de pages personnelles

Sur la base du corpus Hébergeurs15000PP, nous avons constitué un sous-échantillon de 4 000 pages environ que nous avons exploité avec différents outils de statistique (R, Lexico, Alceste), en cherchant à interpréter les spécificités des pages personnelles visitées selon leur serveur d'hébergement, en posant comme hypothèse que le serveur, par son mode de recrutement, par l'image qu'il cherche à véhiculer, par les caractéristiques de son offre formate en partie les contenus et la forme des sites qu'il héberge.

Le tableau ci-dessous présente pour les principaux hébergeurs, le nombre de pages personnelles vues par au moins deux panélistes Netvalue entre janvier et juin 2000, le nombre moyen de pages vues par site personnel et le poids textuel de chaque hébergeur dans le corpus analysé. Ce tableau révèle une forte distorsion entre les URL vues et le sous-corpus constitué, qui tient au fait qu'une grande partie des URL n'est pas accessible au moment de l'aspiration (interdiction d'accès, redirection vers un autre serveur, pages inexistantes), que les pages inaccessibles sont plus ou moins fréquentes selon les hébergeurs et que la richesse textuelle varie aussi.

| Pages personnelles | Nb d'url vues par au moins deux panélistes Netvalue (janvier-juin 2000) | Nb moyen de pages vues par site | % du corpus textuel |
|---|---|---|---|
| Free | 3229 | 10,0 | 14 % |
| Multimania | 2808 | 5,0 | 27 % |
| Wanadoo | 1530 | 5,7 | 19 % |
| Chez | 1834 | 6,5 | 11 % |
| Geocities | 63 | 9,5 | 8 % |
| Club-internet | 535 | 5,7 | 4 % |

***Tableau 7. Pages personnelles vues selon les hébergeurs***

## *6.1 Oppositions lexicales[16]*

Des « vues » distinctes de cet échantillon textuel de 4 000 pages ont été constituées : les 1 000 premiers mots, les mots anglais d'une part et les mots outils d'autre part dans les 2 000 premiers mots. Chacune de ces vues a été soumise à l'analyse des correspondance pour mettre en évidence les contrastes essentiels.

### *6.1.1 1000 premiers mots[17]*

Le premier axe oppose `geocities.yahoo.com` et `www.geocities.com` à l'ensemble : `member.aol.com`,`home.nordnet.fr`,`altern.or`,`perso.wanadoo.fr` et `perso.infonie.fr`. Les mots sous-jacents sont pour le premier groupe des mots anglais : *Alerts, Broadcast, Classifieds, Companion, Invites, PayDirect, Personals, Policy, Quotes, Yahooligans, GeoCities, Add-Ons, Helpful, PageWizard, Techie, Stock, Travel, Address, Calendar, ·, building, area, share, Privacy, learn, Greetings, started, Browse, Site!...*, en particulier du vocabulaire lié au butinage. Pour le second, des mots français, particulièrement des prénoms, des indications de discours rapporté ou de mots mis à distance par les guillemets, et d'abréviations graphiques (flèches) : *heures, Elle, CD, disque, Tu, données, problème, t', audio, =, Qu', À, », «, est-ce, Sarah, disques, monsieur, CD-R, Mélanie, _, Bernadette, Édouard...*

Le second axe oppose surtout `perso.infonie.fr` et `perso.wanadoo.fr` à `altern.org` et `home.nordnet.fr`. Le premier groupe semble lié à l'expression de la subjectivité et du dialogue (*Tu, Moi, J', Je, te, Nous, tu, nous* et prénoms : *Édouard, Bernadette, Mélanie, Sarah, Robert, Paul*) ainsi qu'à des mots décrivant du matériel de manière peu technique (*CD-R, gravure, CD-ROM, graveur, pistes, disques, audio, disque, copie, CD, lecteurs*) et à des mots « organisateurs » (*est-ce, Qu', Cela, raison, parce, Elle, plupart, peut-être, pas., question, =*). Le second correspond à un vocabulaire « internaute » relativement spécialisé (*forum, server, ko, Crack, Ko, unzip, download, crack, rename, Win, homepage*), pas uniquement anglais (*virus, télécharger, antivirus*).

### *6.1.2 Mots outils[18] au sein des 2000 premiers mots[19]*

Le premier axe oppose `geocities.yahoo.com`, `members.aol.com`, `assoc.wanadoo.fr` et `www.ifrance.com` à `perso.wanadoo.fr` et `home.nordnet.fr`. Le vouvoiement et une expression non-personnelle caractérisent le premier pôle : *Vos, Votre, vos, votre, on*. S'ajoutent de nombreux éléments de structuration de l'énoncé : *afin, entièrement, Ensuite, assez, malgré, facilement, ensuite, ci-dessus, ci-dessous, ainsi, ici, dessus, enfin, etc., donc, ET, Très, désormais, ou, sur, grâce* et les déterminants : *LA, LES, Cet, LE, L', chaque, Les, Ce*, en particulier indéfinis : *DES, Toutes, Tous, toute, toutes*. Le second pôle est centré sur le dialogue et le rapport au moment de la rédaction : *moi., ton, Voilà, je, nous, ta, toi, Nous, t', tu, Je, te, J', Moi, est-ce, demain, T', Tu*.

Le deuxième axe oppose `www.ifrance.com` et `home.nordnet.fr` à `assoc.wanadoo.fr` et `geocities.yahoo.com`. Le premier groupe est dominé par les pronoms et déterminants du dialogue, qui regroupent cette fois vouvoiement et tutoiement : *Vos, Votre, vos, Ils, votre, Ça, Tu, T', Nous, Moi, Mon, Je, nos, vous, J', VOUS, t'*. Le deuxième pôle se caractérise par les prépositions, adverbes, conjonctions et connecteurs : *alors, donc, dont, également, Pour, ici., ET, presque, ci-dessus, Très, DE, partout, ici, Donc, après, ensuite, Après, Cependant, Ensuite, notamment, ensemble, Or, entièrement, particulièrement, seulement, selon, puisque, lors, désormais, enfin, ci-dessous, hors, etc., Ainsi, laquelle.*

### *6.1.3 Mots anglais au sein des 2000 premiers mots[20]*

Le premier axe oppose `altern.org`, `perso.club-internet.fr`, `member.aol.com`, `perso.infonie.fr`, `www.respublica.fr`, `home.nordnet.fr` à `www.geocities.com` et `geocites.yahoo.com`. Au premier pôle correspondent des mots renvoyant au sexe (*sex, nude,* et peut-être *pictures* et *picture*) , ainsi que des mots indiquant des opérations à effectuer (*Click, download, don', Submit, rename, Download*). Dans l'autre pôle,

[16] Les figures commentées dans cette partie sont placées en annexe.

[17] Figure 2 en annexe.

[18] Outre les prépositions (*à, de*), les déterminants (articles, possessifs, démonstratifs...), les pronoms, les adverbes (*jamais, également*), les conjonctions (*et, comme*), on a retenu les réalisations des verbes *être* et *avoir*.

[19] Figure 3 en annexe.

[20] Figure 4 en annexe.

figurent des mots renvoyant à des thèmes : *Weather, Auctions, Broadcast, Classifieds, Personals, interests*.

Le deuxième axe oppose `home.nordnet.fr` et `altern.org` et `assoc.wanadoo.fr` à `perso.club-internet.fr`, `free` et `perso.infonie.fr`. Le premier pôle se caractérise par les prépositions (*with, in, of, to, at, To*) et les connecteurs et adverbes (*here, down, or, And*). On trouve dans le second pôle des verbes abrégés (*can', doesn', don'*) qui indiquent le poids de la négation (*not, Not* figurent également), les mots renvoyant au sexe (*sex, nude,* et peut-être *pictures* et *picture*), ainsi que des mots renvoyant au téléchargement (*download, Download*).

### *6.2 Des contenus spécifiques*

Examinons les mots spécifiques des pages personnelles vues selon leur serveur d'hébergement. Dans les pages personnelles de `Free` *visitées,* quelques domaines sémantiques peuvent clairement être identifiés : les messages renvoyés par les serveurs d'interdiction d'accès ou de redirection (*you don't have permission*, *forbidden*, *click here*), le champ sémantique du sexe (y compris les mises en garde pour les visiteurs), celui des logiciels (*cracks*, *download*…) et celui de la gratuité. Chez `Free`, on observe un entrelacement intéressant entre la liberté (sexuelle et logicielle) et la gratuité, portée par le double sens du mot *free.* Dans les sites visités sur le serveur `Chez`, la thématique sexe-porno est sur-représentée, le discours promotionnel y est plus marqué et le positionnement est plus proche de la sphère marchande : la gratuité n'est plus revendiquée, les *sponsors* sont mis en avant.

Chez `Wanadoo`, le contenu des pages visitées a diverses caractéristiques : forte présence des verbes *dire*, *parler*, *penser* ; mise en scène de l'échange (*moi*, *nous* / *toi*, *vous*) ; thèmes du gravage de CD ; le travail (*bureau*, *directeur*, *patron*, *licenciement*…) ; l'amour (*rencontrer*, *regard*, *plaire*) ; la vie (*vieillir*, *mourir*…) et autres préoccupations existentielles. Le site est alors un lieu d'expression intime du moi qui s'adresse à l'autre. Les pages visitées hébergées par `Club-internet` présentent des caractéristiques proches de celles de `Wanadoo`.

Les pages visitées de `Multimania` ont un profil assez différent : le nom de l'hébergeur y apparaît fréquemment (sans doute à cause de l'affichage du bandeau publicitaire) ; les références aux « objets » internet, en particulier aux messageries instantanées (IRC, ICQ, chat…), aux loisirs (bd, musique), aux études (formation, école, bac) sont spécifiques de ces sites. Une étude auprès des concepteurs de ces sites montrerait sans doute qu'ils sont plus jeunes que chez les autres hébergeurs.

Enfin, les pages de `Geocities` ont pour caractéristique d'être en anglais, c'est donc la différence de langue que font ressortir les calculs statistiques.

Les pages personnelles visitées ont donc globalement des tonalités différentes selon leur serveur d'hébergement : le domaine, au sens de terroir, donne un style à ses habitants. Cela est à mettre en relation avec les caractéristiques de la clientèle de chaque hébergeur et/ou fournisseur d'accès, avec l'image que cherche à véhiculer l'hébergeur (campagnes de publicité, portail…), avec l'offre d'hébergement proposée, avec la manière dont sont référencées dans les annuaires les pages en question. Reprenons le cas de `Free`. Ce fournisseur axe son discours sur la gratuité : cette thématique revient dans les sites. Il propose un espace d'hébergement beaucoup plus important que les autres fournisseurs et la possibilité d'installer des scripts sur les pages personnelles, ce qui tend à attirer des sites avec images et vidéos très consommatrices d'espace et des sites très sophistiqués ; enfin, `Free` affiche dès le premier niveau de son annuaire de sites une rubrique « charme ». Il se constitue donc une espèce d'adéquation entre l'idéologie portée par le fournisseur d'accès et ce que mettent en scène les clients dans leurs sites.

## 7 Perspectives

L'architecture TyPWeb permet une étude fine des traits jugés pertinents des sites Web personnels et marchands.

Nous avons exploité ici les traits textuels les plus immédiats : lexique le plus fréquent, pronoms personnels… Nous avons également pris en compte la connexité interne et externe des sites. Ces différents indicateurs montrent à la fois des oppositions tranchées entre l'univers des sites personnels et celui des sites marchands, mais aussi l'apparition de stratégies plus complexes, dans des sites intermédiaires entre ces deux premières catégories, comme au sein des sites marchands. Ce sont ces différenciations en cours que l'architecture TyPWeb a pour objectif de repérer et de mesurer.

Quatre directions de travail s'ouvrent immédiatement :

1. Le « texte » des pages gagnera à être étiqueté au plan morpho-syntaxique (temps, personne, mode et forme canonique des verbes, pour étudier par exemple les différentes attitudes par rapport au lecteur des pages). Cela suppose d'être capable d'isoler les parties en français de celles en anglais, mais aussi d'adapter les outils habituels au « français du web ».
2. Les résultats des traitements statistiques sont à replacer « en contexte ». Il s'agit donc de pouvoir revenir d'un sur-emploi constaté d'un TAG HTML, d'un attribut, d'un mot, aux pages qui exhibent ce sur-emploi de manière prototypique.

3. L'accent mis sur tel ou tel des TAG HTML comme le jeu sur certains attributs ou valeurs d'attributs participe de la caractérisation des sites. Appréhender ce niveau de fonctionnement implique de retraiter les valeurs d'attribut (par exemple, pour saisir les oppositions pertinentes dans le choix des couleurs).
4. Plutôt que de parler d'un accès « sémantique » au web dans son ensemble, notre travail actuel amènerait plutôt à entamer l'apprentissage, supervisé ou non, de classes sémantiques sur des sous-corpus pertinents du web.

Enfin, la chaîne TyPWeb continuera d'être utilisée pour tester de nouvelles hypothèses nées de la consultation des sites et des entretiens menés auprès des concepteurs.

## 8 Bibliographie


**Amitay, E.** (1999) "Anchors in context" in *Words on the Web - Computer Mediated Communication*, Lynn Pemberton & Simon Shurville eds., Intellect Books, UK. (October 1999 / 192 pp. ISBN 1-871516-56-0).

**Beaudouin, V. Velkovska, J.** (1999) "Constitution d'un espace de communication sur Internet (forums, pages personnelles, courrier électronique…", Réseaux, 17(97), pp. 121-177.

**Biber, D.** (1995) Dimensions of register variation : a cross-linguistic comparison, Cambridge University Press, Cambridge.

**Folch, H. Heiden, S. Habert, B. Fleury, S. Lafon, P. Nioche, J. Prévost, S. Illouz, G.** (2000) "TyPTex : Inductive typological text classification analysis for NLP systems tuning/evaluation", in *Second International Conference on Language Resources and Evaluation*, pages 141-148, editors : M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer, volume 1, Athens (Greece), 31 may-2 june.

**Gensollen, M.** (1999) "La création de valeur sur Internet", in P. Flichy, éd., *Internet un nouveau mode de communication ?*, Réseaux 17(97), CNET - HERMES Science Publications, Paris, pp. 15-76.

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

**Illouz, G.** (2000) Typage de données textuelles et adaptation des traitements linguistiques, Thèse de doctorat en informatique, Université Paris-Sud.

**Licoppe, C.** (2000) Commerce électronique, la question de la vente aux particuliers sur l'internet (« Business to Consumers »), Réseaux, 18(100), p. 359-384.

**Maisondieu, A. Kuncova, A.** (2000) "Constitution d'un corpus web dans le cadre du projet TyPWeb", Rapport de stage, Paris 3 & FT R&D.


## 9 Annexes

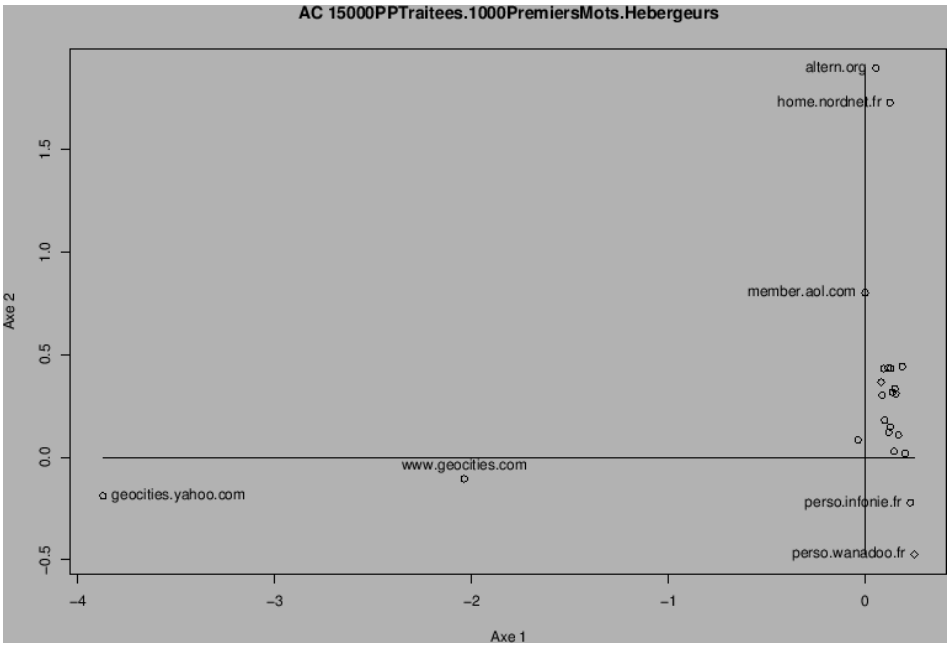


***Figure 2. 1000 premiers mots***

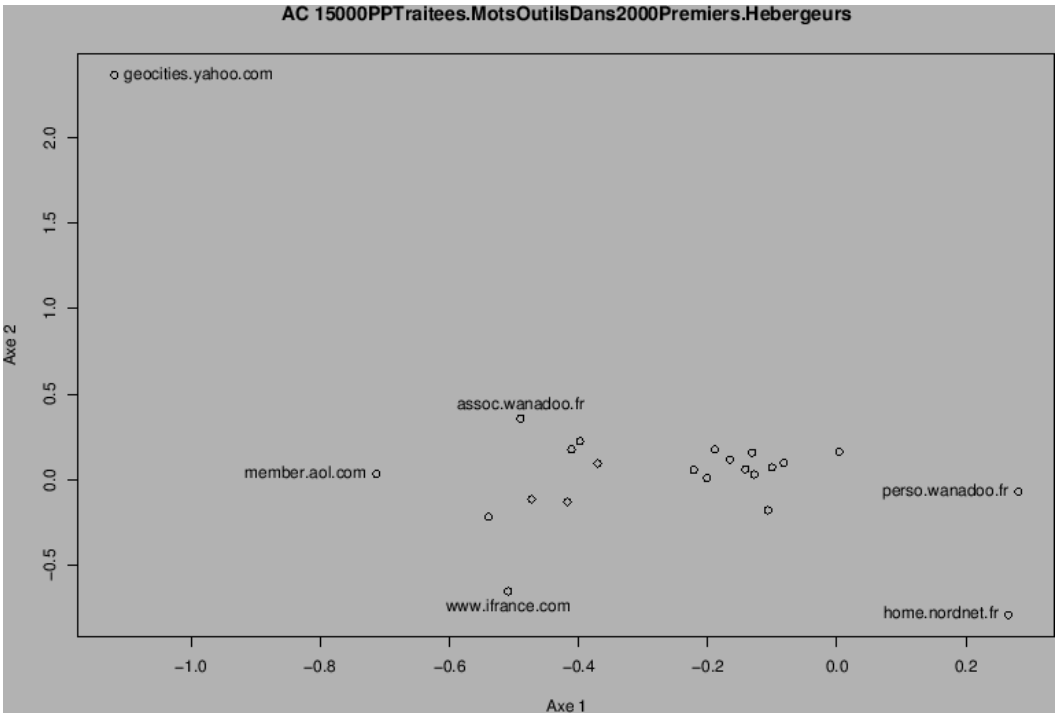


*Figure 3. Mots outils au sein des 2000 premiers mots*

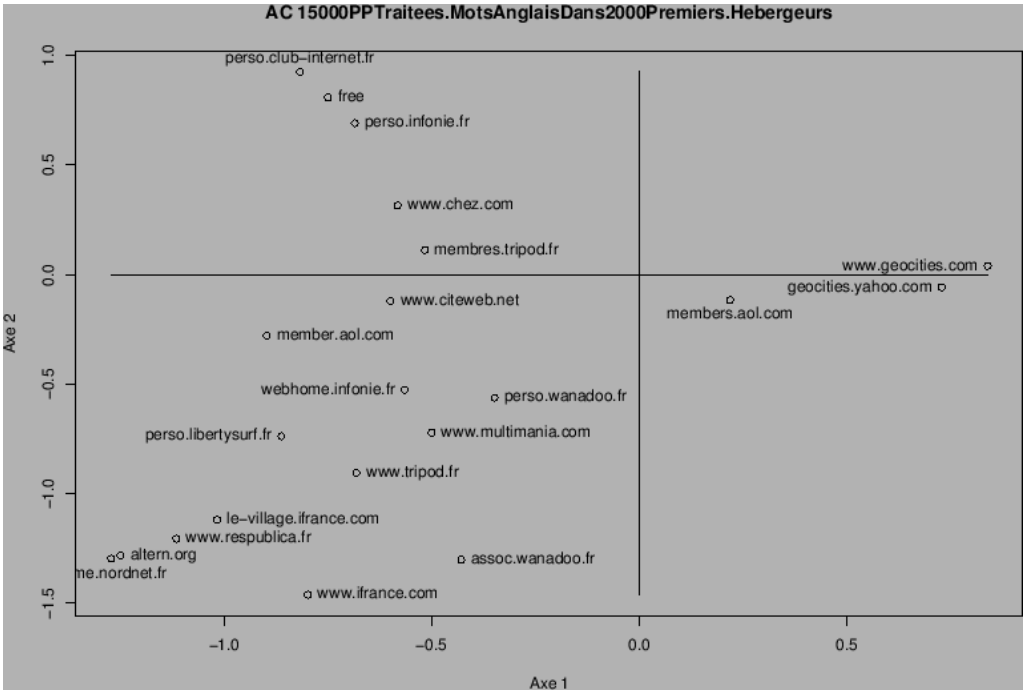


*Figure 4. Mots anglais au sein des 2000 premiers mots*