Traits textuels, structurels et présentationnels pour typer les sites web personnels et marchands

Marie Pasquier (FT R&D, DIH/UCE & MoDyCo Paris X)

Equipe TyPWeb

Valérie Beaudouin, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe, Marie Pasquier

Plan de la présentation

- Projet TyPWeb
  - contexte et hypothèses
- Méthodologie
  - corpus : logiques de constitution et description
  - chaîne de traitement
- Identification de traits et premiers résultats

Projet TyPWeb

Objectifs

- mettre au point une méthodologie de description de sites à l'aide de traits structurels, présentationnels et textuels
- typer les sites afin de catégoriser les parcours sur le web

Comment ?

- transposer et adapter à des corpus web l'architecture de typologie des textes TyPTex (Folch *et al.*, 2000)
- décrire la structure et le contenu des sites web
  - identifier des types de pages et de sites sur la base de traits

Identification de traits multimédias

traits structurels

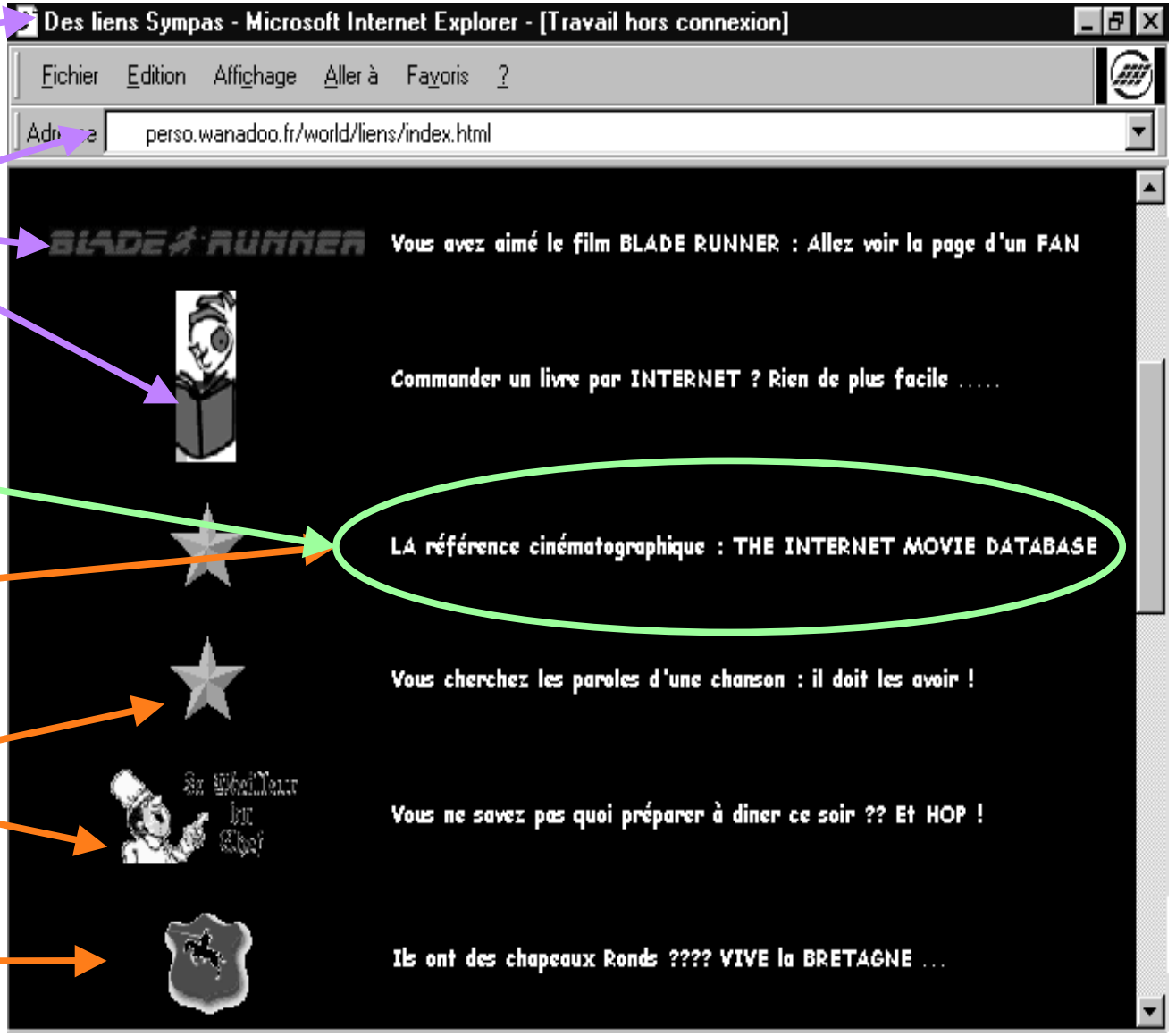
- titre
- liens
  interne et externe

traits textuels

- zone textuelle
  - ↳ contenu :
    LA, référence…
  - ↳ aspect :
    police, taille, couleur

traits présentationnels

- images
  taille, format
- fond d'écran
  couleur, motif,...



Spécificités des corpus web

Corpus web *versus* corpus textuels

- volume et « mouvance » des données à traiter
- articulation texte / non-texte (image, vidéo, structure, etc.)
- texte fortement bruité : usages propres au web, interférence de la mise en forme, changements brutaux de langue, scories, etc.
- français intermédiaire entre oral et écrit
- rôle primordial de la structure et faible place accordée au texte

Analyse multidimensionnelle des traits

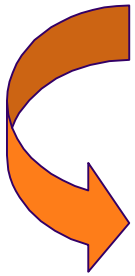
Sélection de sites

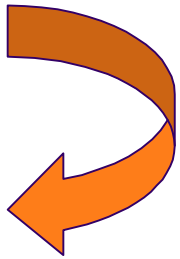
Sites Marchands

- enjeu majeur pour l'économie d'internet
- secteur en évolution permanente
- étude qualitative de grande ampleur menée sur les sites marchands (Licoppe, 2000)

Sites Personnels

- rôle essentiel dans les interactions électroniques (Beaudouin & Velkovska, 1999) : lieu stable de présentation de soi et du réseau de relations
- services d'hébergement de pages personnelles : très visités (données NetValue)
- contenu et forme moins connus





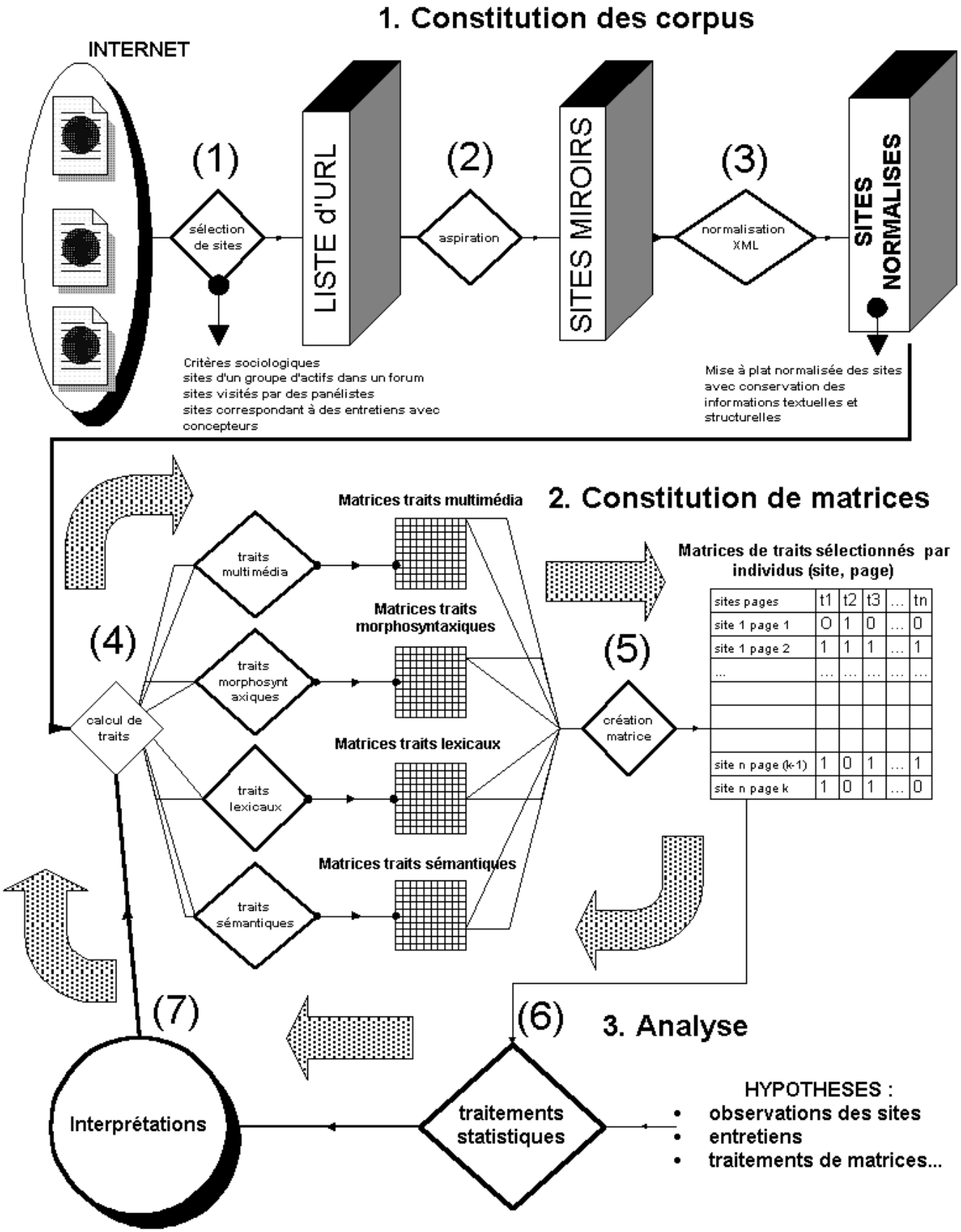
sélection de catégories de sites très contrastées pour mettre à jour leurs spécificités

Corpus

- Corpus de sites et de pages personnels
  - 539 sites wanadoo des membres actifs d'un forum (été 99)
  - 568 sites wanadoo visités en mars 2000 par des internautes du panel NetValue(été 2000)
    - pas de recouvrement avec le premier corpus
  - 12 000 pages visitées par au moins 2 panélistes et hébergées chez 19 serveurs d'hébergement (février 2001)
- Corpus de sites marchands
  - 50 sites marchands
  - 3 états du corpus : 1999/2000, été 2000, mars 2001
  - thématiques des corpus : assurance, voyage, vente par correspondance, informatique, etc.

Descriptif de la chaîne de traitements

- phase 1 : aspiration de sites
- phase 2 : normalisation
- phase 3 : analyses


1. Constitution des corpus
INTERNET
(1)
sélection de sites
LISTE d'URL
(2)
aspiration
SITES MIROIRS
(3)
normalisation XML
SITES NORMALISES
Critères sociologiques
sites d'un groupe d'actifs dans un forum
sites visités par des panélistes
sites correspondant à des entretiens avec concepteurs
Mise à plat normalisée des sites avec conservation des informations textuelles et structurelles
2. Constitution de matrices
Matrices traits multimédia
Matrices traits morphosyntaxiques
Matrices traits lexicaux
Matrices traits sémantiques
traits multimédia
traits morphosynt axiques
traits lexicaux
traits sémantiques
(4)
calcul de traits
(5)
création matrice
Matrices de traits sélectionnés par individus (site, page)
sites pages | t1 | t2 | t3 | ... | tn
site 1 page 1 | 0 | 1 | 0 | ... | 0
site 1 page 2 | 1 | 1 | 1 | ... | 1
site n page (k-1) | 1 | 0 | 1 | ... | 1
site n page k | 1 | 0 | 1 | ... | 0
(6)
3. Analyse
traitements statistiques
HYPOTHESES :
observations des sites
entretiens
traitements de matrices...
(7)
Interprétations

Aspirations des sites

- utilisation d'outils d'aspiration de sites (WinHTTrack, Teleport Pro)
- sauvegardes locales des « sites miroirs »
- problèmes techniques rencontrés
  - temps de traitements très variables
  - difficultés techniques liées aux architectures mises en place pour construire les sites (scripts, programmes, …)
  - stockage des données aspirées

	Nb de pages	Durée aspiration	% de récupération	Nb de page/minute	Taille des fichiers (en Mo)
home-nordnet	**122**	**40 sec.**	**99**	**183**	
free	**3 687**	**24 min.**	**86**	**131**	**13,3**
multimania	**2856**	**54 min.**	**98**	**52**	**21,7**
www.fnac.com (5 niveaux)	**41 302**	**3j 6h 10 min.**	**78**	**0,9**	**42,8**

Exemple de temps d'aspiration de pages et sites web

Normalisation des sites

- construction des éléments suivants :
  - **une base documentaire XML globale**
    - ↪ réorganisation normalisée des éléments (structurels et textuels) qui composent les pages originales
  - **des états statistiques** sur la composition des pages de cette base
    - « mots »
    - éléments HTML : TAG, attributs, valeurs d'attribut…
    - états statistiques sur les sites
  - **des corpus** construits à partir de sélections (texte, TAG HTML,...)

- représentations des corpus soumises à des outils d'analyse
  - statistiques générales : R, SAS
  - statistiques textuelles : Lexico3, Alceste, Tropes
  - étiqueteurs : Cordial, TILT
  - ...

Caractéristiques des parties

	Sites Personnels		Sites Marchands		
	PP-Ete99	PP-mars00	SM-99-00	SM-été00	TOTAL
nb de sites	539	568	29	16	1 162
nb de pages	11 006	24 938	29 199	5 726	96 885
moy. pages/site	20	44	1 007	358	83
nb occurrences	3 878 647	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	

Résistance et complexité des données

- **grand nombre de traits « bruts »**
  - de l'ordre de 200 000 sur l'ensemble des corpus
  - construction de matrice de traits ingérable par les outils disponibles
- **complexité des éléments structurels et présentationnels**
  - exemple : élément **`FONT`** (police) peut être caractérisé par le type de police, par sa taille, sa couleur…
    - ↪ reste à trouver pour chaque attribut de ce type les découpages pertinents
- **complexité des éléments textuels**
  - éclatement local du texte par les balises (lettrines par exemple)
  - « sauts » d'une langue à l'autre (surtout avec l'anglais)
  - faible proportion de texte dans bon nombre de pages

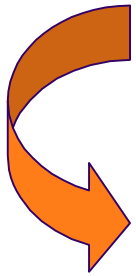
Enjeux et démarches d'analyse

- **profiler des catégories de sites définies *a priori***
  - constitution d'hypothèses sur la base de l'observation des sites et des entretiens
  - identification des traits nécessaires pour tester les hypothèses
  - constitution de matrices avec ces traits
  - analyses statistiques et interprétations

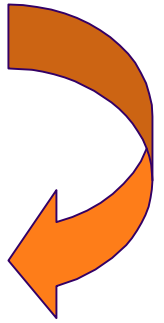
→ **Approche hypothético-déductive**

- **faire émerger des types de catégories de sites**
  - traitements statistiques pour faire apparaître des types de pages ou sites (cf. Biber)
  - sélection de traits pour décrire les pages ou les sites

→ **Approche inductive**



TyPWeb → double angle d'attaque



Des traits « bruts » aux traits jugés pertinents

- traits structurels et présentationnels
  - complexité des sites
  - redondance des liens
- traits textuels
  - répartition des pronoms personnels
  - spécificité de vocabulaire par hébergeur
  - oppositions lexicales

Complexité des sites

- la structure des sites marchands est plus complexe que celle des sites personnels
  - en moyenne 20 fois plus de pages dans les sites marchands
    - donc plus de liens
  - même nombre moyen de liens par page
- la structure des sites personnels visités par un panel d'internautes est plus complexe que celle des sites des «habitués » d'un forum
  - 44 pages par site contre 25
  - 1,9 liens externes par page contre 1,2
  - 10 liens internes contre 6
  - même nombre moyen de liens par page
- corrélation entre taille des sites et audience

Redondance des liens

- contraste important sur les sites marchands entre la page d'accueil et les autres pages
  - nombre de liens

	Page d'accueil	Autres pages
liens internes	11	4
liens externes	3	0,1

  - redondance (R=nombre de liens total divisé par le nombre de liens différents)
    - R=1,25 sur page d'accueil ; R=1,1 sur autres pages
- pas de contraste entre la page d'accueil et les autres pour les pages perso

Répartition des pronoms personnels

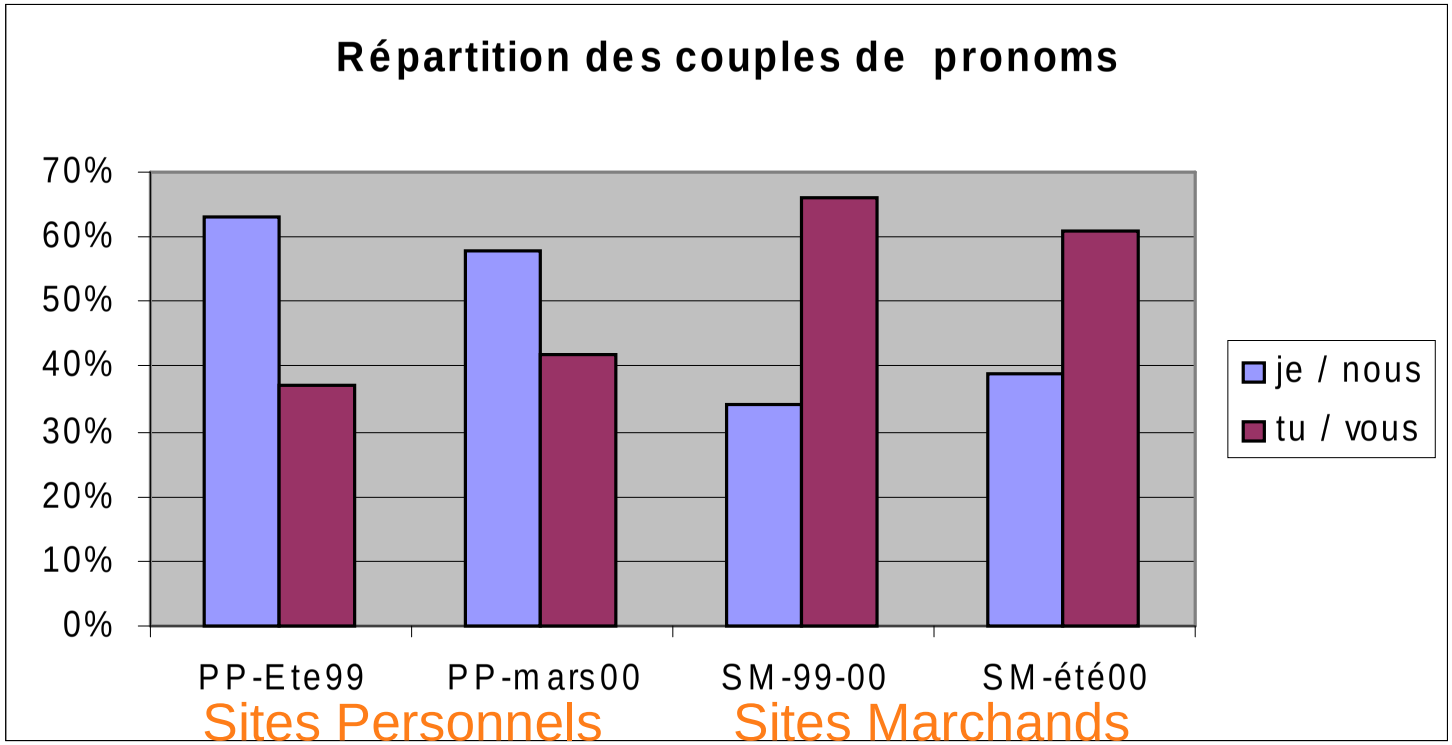
- utilisation très différenciée selon les types de sites
- 6 catégories de pronoms personnels fondées sur le nombre et la personne → choix
  - absence des pronoms *le*, *la*, *les* trop ambigus
  - *on* et *soi* classés dans la 3ème pers. du singulier
  - sites contenant au moins 10 pronoms

	TOTAL (705 sites)	PP-Ete99 (239 sites)	PP-mars00 (430 sites)	SM-99-00 (22 sites)	SM-été00 (14 sites)
je/me/moi	24	26	23	13	14
tu/te/toi	5	5	5	2	3
il/elle/on/lui/leur/soi	34	35	34	25	33
nous	10	10	10	11	11
vous	21	18	20	44	34
ils/elles/leurs	6	6	6	5	5
Total	100	100	100	100	100

Répartition des pronoms personnels (en %)

Répartition de couples de pronoms

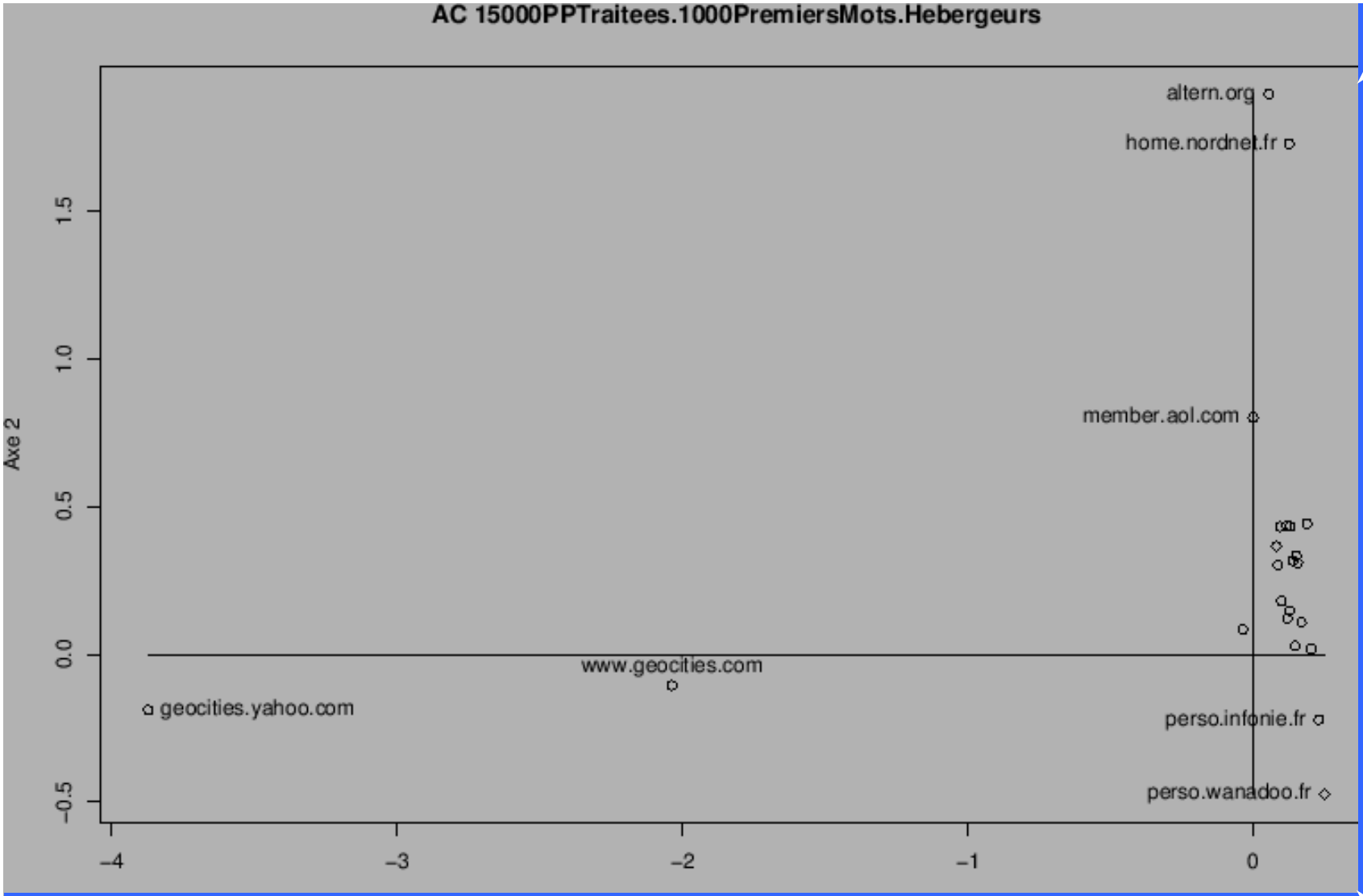
- répartition des couples de pronoms des 1ère et 2ème pers
  - nette rupture entre sites marchands et personnels
  - mise en évidence des spécificités des parties de corpus



Spécificités des hébergeurs de pages personnelles

- **free**
  - INTERDIT / SEXE / LOGICIEL / GRATUITE (reflet du double sens de free)
- **chez**
  - SEXE (plus hard) / SPONSOR (vers la sphère marchande)
- **wanadoo** et **club-internet**
  - MOI-TOI / SE RACONTER / VIE / AMOUR /TRAVAIL
- **multimania**
  - FORMATION / DIVERTISSEMENT (bd, jeux, musique) / CHAT
- **geocities**
  - pages perso en anglais

Oppositions des 1000 premiers mots



Vocabulaire internaute

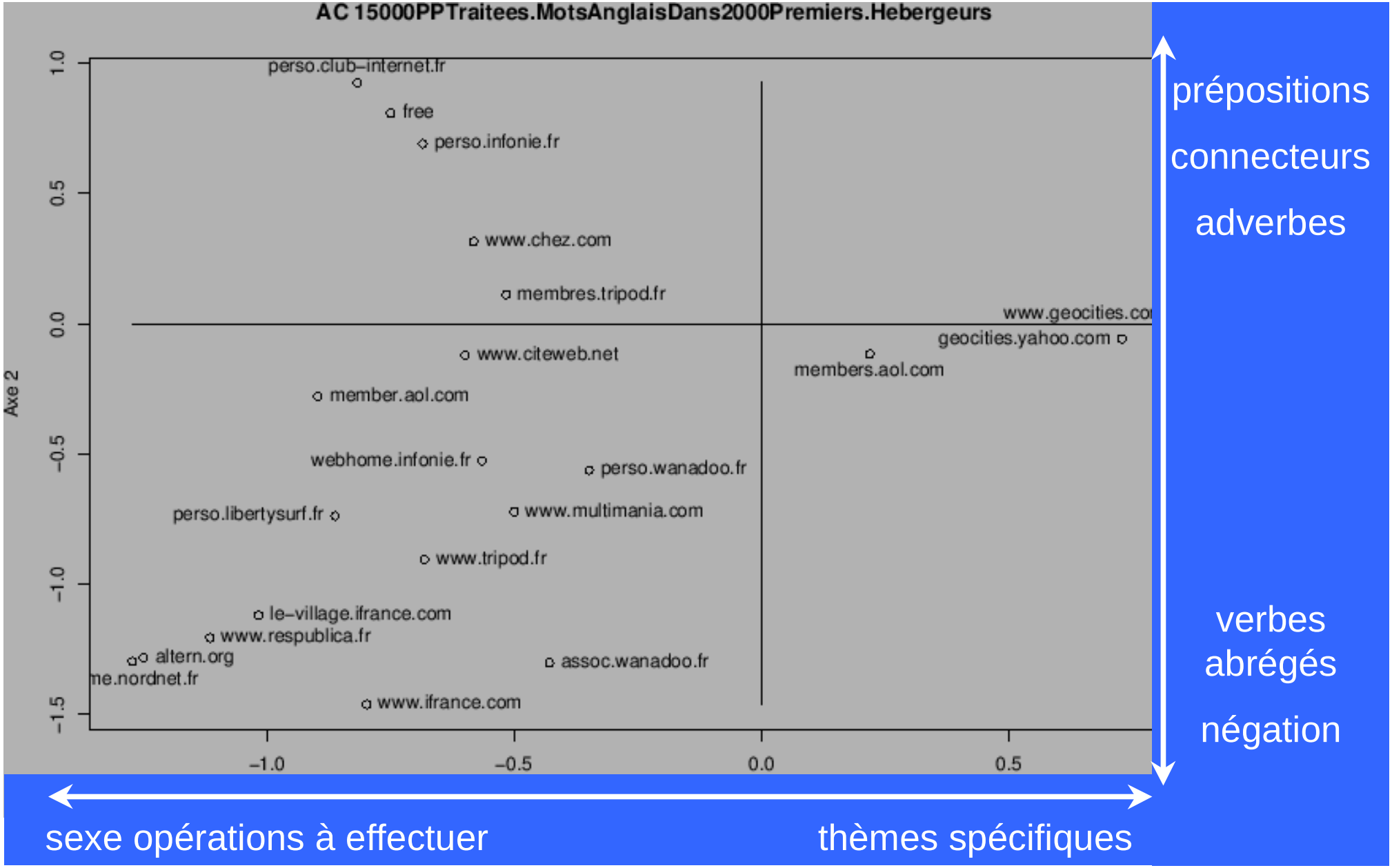
Subjectivité dialogue

Mots organisateurs

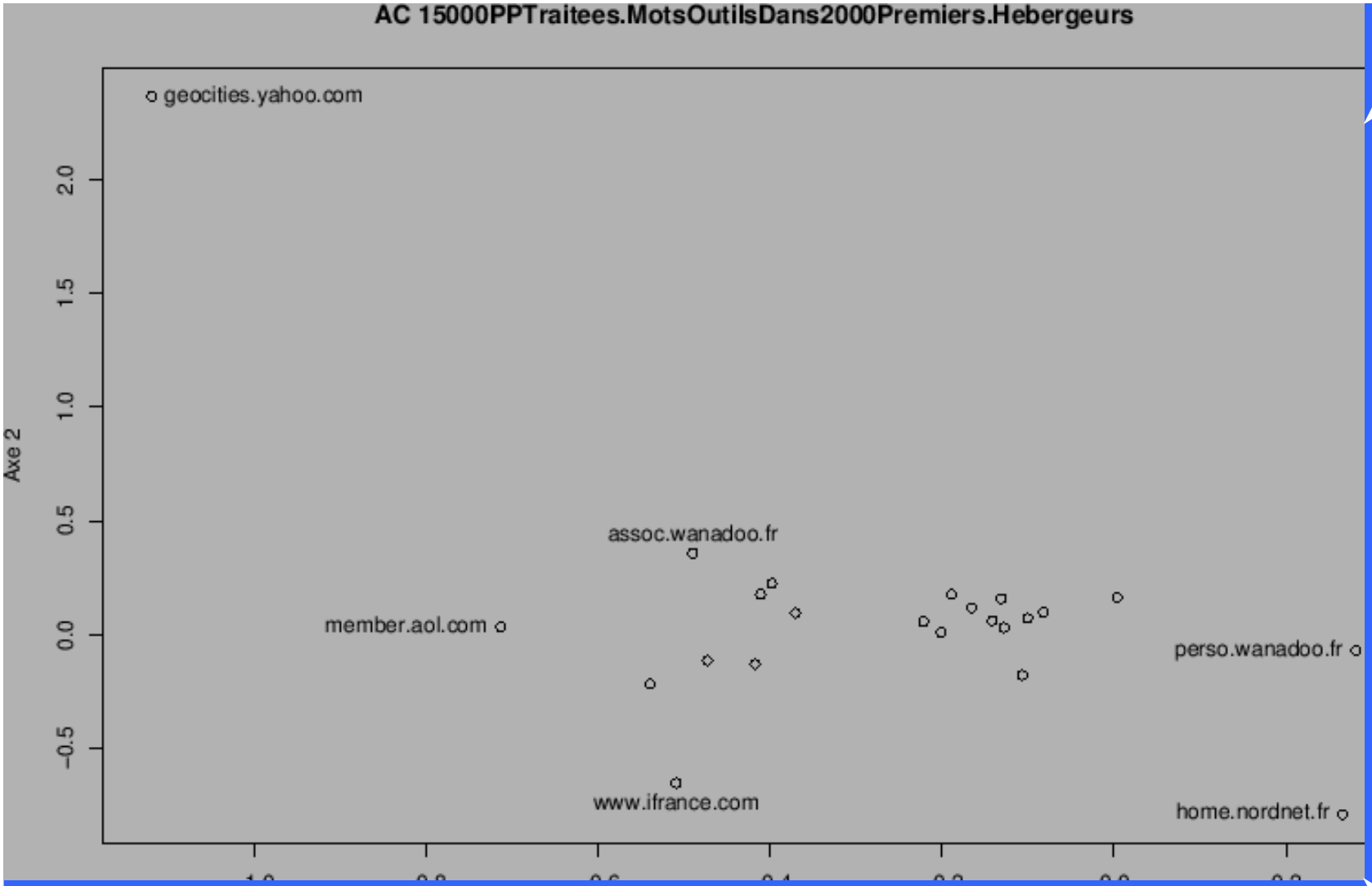
Mots anglais butinage

prénom abréviation discours rapporté

Oppositions des mots anglais



Oppositions des mots « outils »



prépositions
adverbes
conjonctions
et
connecteurs

pronoms et
déterminants
du dialogue

expression non personnelle vousvoiement
éléments structurants de l'énoncé

dialogue échanges personnel
immédiateté de la rédaction

Perspectives

- des traits « bruts » aux traits jugés pertinents
  - approfondissement de l'existant
  - identification des combinaisons de traits pertinentes
    - de nature identique (traits + attributs)
    - de nature différente
- autres traits
  - étiquetage morpho-syntaxique
  - identification de traits sémantiques