TypWeb : profilage de sites Web

Part 3
Corpus 100000

Document de travail

Equipe Typweb

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

**France Télécom R&D (*), Université Paris X (**),
Université Paris 3 (CLA2T) (***), LIMSI (****)**

1 Sommaire

2 Préambule

Travail à partir d'un corpus de pages visitées par un panel d'internautes. Sur la base d'une liste de plus de 100.000 pages visitées nous avons constitué différents corpus de travail. Ces pages ont été aspirées puis traitées par la chaîne Typweb.

Deux chaînes de traitement ont été mises en œuvre pour la constitution des corpus :

1. La chaîne webxref : normalisation des sorties pour constituer différents corpus d'analyse
2. La chaîne Lynx : prise en compte du texte seul des pages via la commande Lynx

Ces deux chaînes sont décrites infra.

3 Descriptif des corpus

3.1 Chaîne WebXref

Liste initiale : 101 427 pages
Altern : 1895 pages non retenues
Différence : 99 532

3.1.1 Corpus 100000 avec webxref

Balise SITE corp100000-1.xml:10096 corp100000-10.xml:11452 corp100000-11.xml:11346 corp100000-2.xml:5545 corp100000-3.xml:4272 corp100000-4.xml:4666 corp100000-5.xml:5405 corp100000-6.xml:3453 corp100000-7.xml:7015 corp100000-8.xml:2279 corp100000-9.xml:6594 Total : 72123	**Balise SITE F** corp100000-1.xml:10075 corp100000-10.xml:11425 corp100000-11.xml:11337 corp100000-2.xml:5523 corp100000-3.xml:4269 corp100000-4.xml:4662 corp100000-5.xml:5400 corp100000-6.xml:3451 corp100000-7.xml:7004 corp100000-8.xml:2276 corp100000-9.xml:6586 Total : 72008
Balise SITEName corp100000-1.xml:10096 corp100000-10.xml:11452 corp100000-11.xml:11346 corp100000-2.xml:5545 corp100000-3.xml:4272 corp100000-4.xml:4666 corp100000-5.xml:5405 corp100000-6.xml:3453 corp100000-7.xml:7015 corp100000-8.xml:2279 corp100000-9.xml:6594 Total : 72123	**Balise SITEName F** corp100000-1.xml:10096 corp100000-10.xml:11452 corp100000-11.xml:11346 corp100000-2.xml:5545 corp100000-3.xml:4272 corp100000-4.xml:4666 corp100000-5.xml:5405 corp100000-6.xml:3453 corp100000-7.xml:7015 corp100000-8.xml:2279 corp100000-9.xml:6594 Total : 72123
Balise DumpLynx corp100000-1.xml:10075 corp100000-10.xml:11425 corp100000-11.xml:11337 corp100000-2.xml:5523 corp100000-3.xml:4269 corp100000-4.xml:4662 corp100000-5.xml:5400 corp100000-6.xml:3451 corp100000-7.xml:7004 corp100000-8.xml:2276 corp100000-9.xml:6586 Total : 72008	**Balise DumpLynx F** corp100000-1.xml:10075 corp100000-10.xml:11425 corp100000-11.xml:11337 corp100000-2.xml:5523 corp100000-3.xml:4269 corp100000-4.xml:4662 corp100000-5.xml:5400 corp100000-6.xml:3451 corp100000-7.xml:7004 corp100000-8.xml:2276 corp100000-9.xml:6586 Total : 72008
Balise DumpText corp100000-1.xml:9987 corp100000-10.xml:11397 corp100000-11.xml:11296 corp100000-2.xml:5484 corp100000-3.xml:4230 corp100000-4.xml:4662 corp100000-5.xml:5397 corp100000-6.xml:3444 corp100000-7.xml:6944 corp100000-8.xml:2252 corp100000-9.xml:6566 Total : 71659	**Balise DumpText F** corp100000-1.xml:9987 corp100000-10.xml:11397 corp100000-11.xml:11296 corp100000-2.xml:5484 corp100000-3.xml:4230 corp100000-4.xml:4662 corp100000-5.xml:5397 corp100000-6.xml:3444 corp100000-7.xml:6944 corp100000-8.xml:2252 corp100000-9.xml:6566 Total : 71659

3.1.2 Corpus 100000 avec webxref : Recomposition du corpus

Nettoyage du corpus webxref en ne conservant que les parties du corpus possédant une balise <SITE> ouvrante et fermante

Nombre de sites : 72008

Balise SITE CORP100000CLEAN-WXF-1.txt:10075 CORP100000CLEAN-WXF-10.txt:11425 CORP100000CLEAN-WXF-11.txt:11337 CORP100000CLEAN-WXF-2.txt:5523 CORP100000CLEAN-WXF-3.txt:4269 CORP100000CLEAN-WXF-4.txt:4662 CORP100000CLEAN-WXF-5.txt:5400 CORP100000CLEAN-WXF-6.txt:3451 CORP100000CLEAN-WXF-7.txt:7004 CORP100000CLEAN-WXF-8.txt:2276 CORP100000CLEAN-WXF-9.txt:6586 **Total : 72008**	**Balise SITE F** CORP100000CLEAN-WXF-1.txt:10075 CORP100000CLEAN-WXF-10.txt:11425 CORP100000CLEAN-WXF-11.txt:11337 CORP100000CLEAN-WXF-2.txt:5523 CORP100000CLEAN-WXF-3.txt:4269 CORP100000CLEAN-WXF-4.txt:4662 CORP100000CLEAN-WXF-5.txt:5400 CORP100000CLEAN-WXF-6.txt:3451 CORP100000CLEAN-WXF-7.txt:7004 CORP100000CLEAN-WXF-8.txt:2276 CORP100000CLEAN-WXF-9.txt:6586 **Total : 72008**
Balise DumpLynx CORP100000CLEAN-WXF-1.txt:10075 CORP100000CLEAN-WXF-10.txt:11425 CORP100000CLEAN-WXF-11.txt:11337 CORP100000CLEAN-WXF-2.txt:5523 CORP100000CLEAN-WXF-3.txt:4269 CORP100000CLEAN-WXF-4.txt:4662 CORP100000CLEAN-WXF-5.txt:5400 CORP100000CLEAN-WXF-6.txt:3451 CORP100000CLEAN-WXF-7.txt:7004 CORP100000CLEAN-WXF-8.txt:2276 CORP100000CLEAN-WXF-9.txt:6586 **Total : 72008**	**Balise DumpLynx F** CORP100000CLEAN-WXF-1.txt:10075 CORP100000CLEAN-WXF-10.txt:11425 CORP100000CLEAN-WXF-11.txt:11337 CORP100000CLEAN-WXF-2.txt:5523 CORP100000CLEAN-WXF-3.txt:4269 CORP100000CLEAN-WXF-4.txt:4662 CORP100000CLEAN-WXF-5.txt:5400 CORP100000CLEAN-WXF-6.txt:3451 CORP100000CLEAN-WXF-7.txt:7004 CORP100000CLEAN-WXF-8.txt:2276 CORP100000CLEAN-WXF-9.txt:6586 **Total : 72008**
Balise DumpText CORP100000CLEAN-WXF-1.txt:9987 CORP100000CLEAN-WXF-10.txt:11397 CORP100000CLEAN-WXF-11.txt:11296 CORP100000CLEAN-WXF-2.txt:5484 CORP100000CLEAN-WXF-3.txt:4230 CORP100000CLEAN-WXF-4.txt:4662 CORP100000CLEAN-WXF-5.txt:5397 CORP100000CLEAN-WXF-6.txt:3444 CORP100000CLEAN-WXF-7.txt:6944 CORP100000CLEAN-WXF-8.txt:2252 CORP100000CLEAN-WXF-9.txt:6566 **Total : 71659**	**Balise DumpText F** CORP100000CLEAN-WXF-1.txt:9987 CORP100000CLEAN-WXF-10.txt:11397 CORP100000CLEAN-WXF-11.txt:11296 CORP100000CLEAN-WXF-2.txt:5484 CORP100000CLEAN-WXF-3.txt:4230 CORP100000CLEAN-WXF-4.txt:4662 CORP100000CLEAN-WXF-5.txt:5397 CORP100000CLEAN-WXF-6.txt:3444 CORP100000CLEAN-WXF-7.txt:6944 CORP100000CLEAN-WXF-8.txt:2252 CORP100000CLEAN-WXF-9.txt:6566 **Total : 71659**

Post-traitement du corpus
PB : un certain nombre de caractères pourris. Impossible de post-traiter les corpus tels quels.

3.1.3 Corpus 100000 avec webxref : Nettoyage du corpus

Nettoyage du corpus précédent en supprimant tous les caractères "pourris"

Nombre de sites : 72008

Balise SITE **Total : 72008**	**Balise SITE F** **Total : 72008**
Balise DumpLynx **Total : 72008**	**Balise DumpLynx F** **Total : 72008**
Balise DumpText **Total : 71659**	**Balise DumpText F** **Total : 71659**

3.1.4 Corpus 100000 avec webxref : Corpus construits

3.1.4.1 Corpus Textuel

Sur la base des corpus précédents, construction des corpus pour outils d'analyse.
Pour les corpus textuels reconstruits (Lannion et Lexico), on ne conserve que les sites (page) qui contiennent du texte (i.e. ceux pour lesquels on dispose d'une balise DUMPTEXT dans le corpus précédent).
Pour les corpus (liens et TAGs), on travaille sur le corpus complet précédent.

Corpus Lannion : CORP100000CLEAN2-WXF.lannion

En entrée : 72008 sites dans le corpus initial CORP100000XMLCLEAN-WXF.xml
En sortie : 71649 sites
La structure du fichier construit est la suivante :

```
(<SITE>
<HEBERGEUR>$hebergeur</HEBERGEUR>
<PAGE>$page</PAGE>
<DUMPTEXT>
$text
</DUMPTEXT>
</SITE>)+
```

Corpus Lexico : CORP100000CLEAN2-WXF.lexico

En entrée : 72008 sites dans le corpus initial CORP100000XMLCLEAN-WXF.xml
En sortie : un corpus avec 71649 sites

La structure du fichier construit est la suivante :

```
(<HEBERGEUR=$hebergeur >
<PAGE=$page >
$text)+
```

Corpus de lien

En entrée : 72008 sites dans le corpus initial CORP100000XMLCLEAN-WXF.xml
En sortie : 71201 entrées pour la matrice de liens, 71906 pages pour lesquelles un comptage global des liens a été effectué

Corpus de TAG

En entrée : 72008 sites dans le corpus initial CORP100000XMLCLEAN-WXF.xml
En sortie : 71016 sites
Liste des balises sélectionnées :
`BODY SCRIPT META FONT IMG TABLE APPLET`

3.1.4.2 Corpus de Tag

A partir du corpus XML complet, constitution de corpus de TAG HTML.

Traits sélectionnés : META, TABLE, BODY… (liste à affiner)

3.1.4.3 Corpus de liens

A partir du corpus XML complet, constitution de fichier des liens présents dans les pages à analyser
=> matrice de liens à analyser (cf travail fait sur corpus 15000 avec Valérie).

3.2 Chaîne Lynx

3.2.1 Corpus 100000 avec chaîne Lynx

75365 pages traitées

HEBERGEUR : perso.libertysurf.fr	872
HEBERGEUR : webhome.infonie.fr	184
HEBERGEUR : www.respublica.fr	575
HEBERGEUR : www.multimania.com	19740
HEBERGEUR : www.levillage.org	1
HEBERGEUR : le-village.ifrance.com	1058
HEBERGEUR : free_fr	10911
HEBERGEUR : www.chez.com	8071
HEBERGEUR : perso.respublica.fr	3
HEBERGEUR : assoc.wanadoo.fr	220
HEBERGEUR : members.aol.com	1517
HEBERGEUR : www.multimania.fr	2
HEBERGEUR : www.celebrities-web.com	1
HEBERGEUR : nounouvevette.jackreation.com	1
HEBERGEUR : perso.club-internet.fr	3249
HEBERGEUR : membres.tripod.fr	1298
HEBERGEUR : perso.wanadoo.fr	16366
HEBERGEUR : home.nordnet.fr	732
HEBERGEUR : www.geocities.com	6501
HEBERGEUR : perso.infonie.fr	124
HEBERGEUR : www.ifrance.com	3701
HEBERGEUR : site.voila.fr	238
TOTAL	**75365**

La structure du fichier construit est la suivante :

```
<SITE Number="1">
<HEBERGEUR>assoc.wanadoo.fr</HEBERGEUR>
<PAGE>yves.dautelle/dijon/QDN.html</PAGE>
<LYNXTEXT>

                Quoi de neuf au cours des 4 derniers mois ?


   (L'actualisation de cette rubrique a lieu 3 fois par an, calquée sur
…
</LYNXTEXT>
</SITE>
```

4 Travail à faire

4.1.1 Chaîne Webxref

4.1.1.1 Corpus textuel

4.1.1.1.1 Traitements TLT

Envoi à Lannion d'un état complet du corpus : définir ce qu'on attend d'eux sur la base de leur première tentative sur le corpus 15000

4.1.1.1.2 Traitements avec Lexico

Analyse avec Lexico : voir ce qu'il est possible de faire

4.1.1.1.2.1 Etat du corpus lu sous Lexico

CORP100000-lexico-clean-WXF

```
45715258 45715258 855198 23881243 784505 678346 431872 5000000 18 2 70674 0
0

*** Résultat de la segmentation du fichier: CORP100000-lexico-clean-WXF.TXT
***
Délimiteurs .,:;!?/_-\"'()[]{}§$

      nombre des occurrences :    23881243
      nombre des formes      :      784505
      frequence maximale     :      678346
      nombre des hapax       :      431872
      nombre des clés(type)  :           2
      nombre des clés(ctnu)  :       70674

*** Fin de la segmentation du fichier:      CORP100000-lexico-clean-WXF.TXT
***
```

Principales caractéristiques de la partition : HEBERGEUR

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
assoc.wanadoo.fr	74229	12674	7027	3131	de
free_fr	3074748	164927	89352	72991	de
home.nordnet.fr	380076	42908	21929	10184	de
le-village.ifrance.com	259453	30705	16377	9365	de
members.aol.com	880337	83420	43332	23386	de
membres.tripod.fr	326735	44675	25384	9750	de
perso.club-internet.fr	1475121	116772	62860	50508	de
perso.infonie.fr	35485	7927	4669	969	de
perso.libertysurf.fr	226880	29830	15465	4420	de
perso.wanadoo.fr	4955756	240849	122780	166811	de
site.voila.fr	105568	13976	7788	3909	de
webhome.infonie.fr	69052	16078	9947	2410	de
www.chez.com	3335698	185039	94353	89672	de
www.geocities.com	2081818	175464	100703	31908	the
www.ifrance.com	851372	80196	43560	28729	de
www.multimania.com	5740606	304568	167381	173821	de
www.respublica.fr	8309	853	613	564	fr

Vocabulaire fréquent (extrait)

Formes (ordre lexicométrique)	Fréquence
de	678346
a	390043
la	326671
et	279996
le	262227
gif	239348
1	230491
les	207006
l	202058
des	200686
d	191260
en	173174
un	159511
est	157482
2	149235
du	138917
une	115926
*	113202
pour	108060
sur	107485
jpg	106273
3	99993
que	98632
dans	89766
vous	85393
qui	82455
4	78773
s	78339
par	78010
pas	75622
au	72719
5	71212
the	70345
il	70079
ou	67511
Le	66204
n	64946
0	64499
plus	58419

Formes (ordre lexicométrique)	Fréquence
plus	58419
ne	57445
avec	57040
6	55399
ce	55206
Les	54973
La	53068
10	50508
France	49155
7	49150
on	45373
8	45113
se	44790
qu	44522
11	42148
sont	41889
12	41272
to	40988
site	40735
L	40037
of	39994
A	39869
9	39201
Marie	38710
page	38285
and	38126
I	36311
c	36203
Jean	35433
je	35350
+	33405
son	32721
Il	32278
15	31301
Naissance	30895
14	30679
13	30605
nous	29737
http	29575

Formes (ordre lexicométrique)	Fréquence
http	29575
mais	29436
C	28973
16	28605
20	28057
18	27443
17	27126
aux	27006
com	26794
y	26470
votre	26414
tout	26239
www	25914
21	25660
D	25482
fr	25240
etre	25188
in	25079
Date	24696
cette	24609
me	24259
Lieu	24149
2000	23837
Sexe	23737
cb1	23677
cf0	23677
octets	23649
19	23201
you	23076
fs16	22803
Retour	22555
22	22530
bien	22298
25	22199
m	22033
Deces	21999
si	21895
faire	21716
meme	21708

Formes (ordre lexicométrique)	Fréquence
meme	21708
Pour	21676
x	21658
comme	21476
23	21335
elle	21254
01	21144
The	21046
Je	21023
fait	21007
for	20304
24	20116
Pierre	19973
sa	19899
tres	19835
30	19811
F	19671
peut	19123
f4	18414
html	18300
26	18196
Un	18161
02	17588
29	17513
t	17482
J	17479
lui	17451
50	17442
27	17298
ses	17263
j	17198
28	17061
DE	16922
Vous	16843
ont	16804
ai	16683
Saint	16647
deux	16576
En	16285

4.1.1.2 Corpus de liens

1. Analyse de la matrice de liens : cf travail sur corpus 15000

4.1.1.3 Corpus de TAG

4.1.1.3.1 Balise META

Analyse d'un corpus qui ne contient que les traits META (avec attributs).

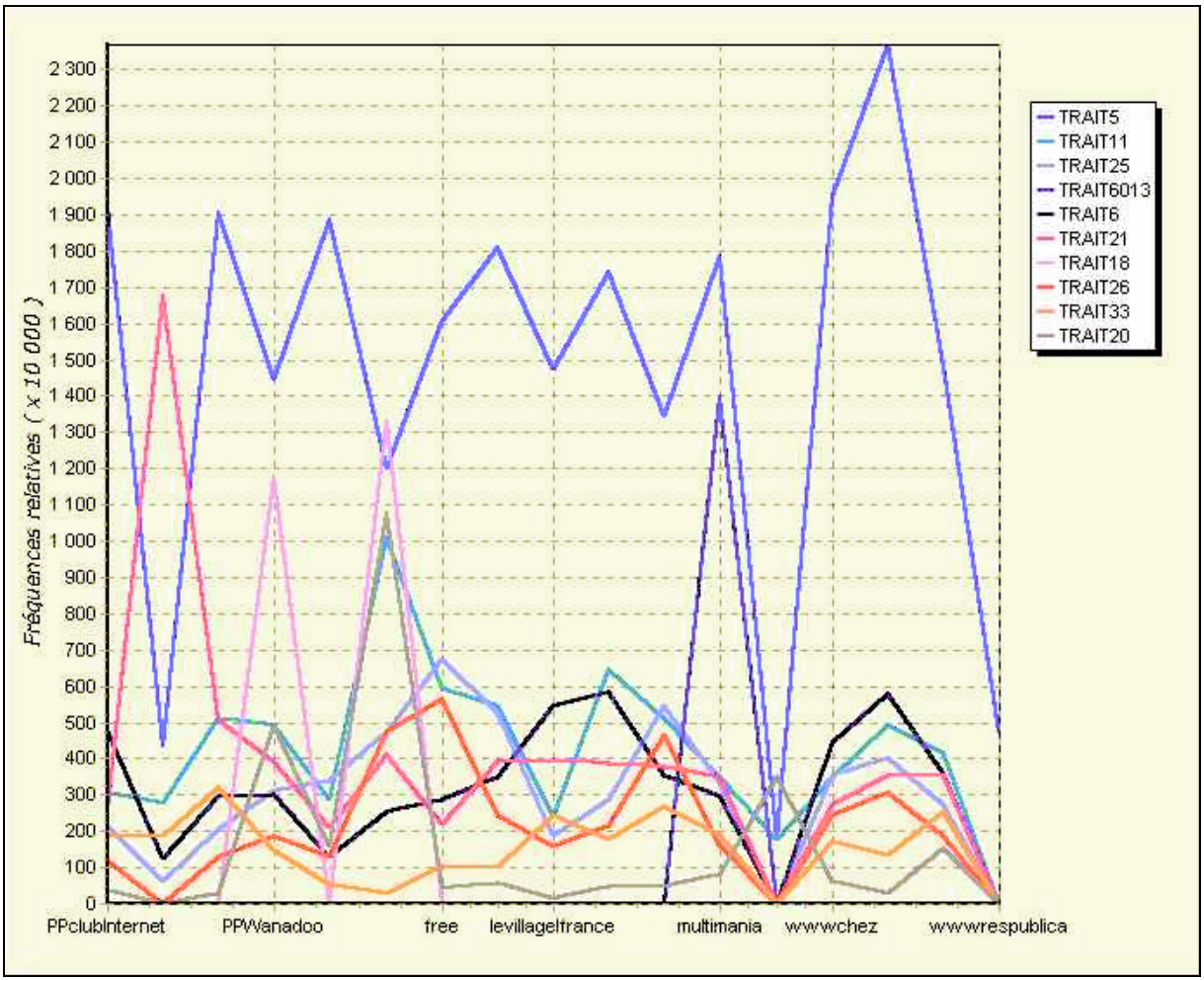
4.1.1.3.1.1 Etat du corpus

```
META-HEB2-CLEAN-CP100000
669607 669607 117248 121850 9135 20950 4444 5000000 3 3 108108 0 0

*** Résultat de la segmentation du fichier: META-HEB2-CLEAN-CP100000.TXT
***
Délimiteurs .,:;!?/_-\"'()[]{}§$£

      nombre des occurrences :      121850
      nombre des formes      :        9135
      frequence maximale     :       20950
      nombre des hapax       :        4444
      nombre des clés(type)  :           3
      nombre des clés(ctnu)  :      108108

*** Fin de la segmentation du fichier:      META-HEB2-CLEAN-CP100000.TXT
***
```

4.1.1.3.1.2 Principales caractéristiques de la partition : HEBERGEUR

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
PPclubInternet	7223	788	377	1390	TRAIT5
PPInfonie	322	56	23	54	TRAIT21
PPLibertysurf	1401	241	106	267	TRAIT5
PPWanadoo	26429	2163	1085	3817	TRAIT5
WHInfonie	382	85	30	72	TRAIT5
assocWanadoo	316	49	19	42	TRAIT18
free	18540	1403	546	2984	TRAIT5
homeNordnet	1194	173	75	216	TRAIT5
levillageIfrance	1505	308	192	222	TRAIT5
membAol	2201	332	155	384	TRAIT5
membTripod	2236	265	119	301	TRAIT5
multimania	32402	2680	1240	5788	TRAIT5
siteVoila	57	19	11	10	TRAIT27
wwwchez	13294	1467	659	2594	TRAIT5
wwwgeocities	8303	936	457	1967	TRAIT5
wwwifrance	6024	890	516	894	TRAIT5
wwwrespublica	21	16	11	2	TRAIT14662

4.1.1.3.1.3 Répartition de traits META dont la fréquence est supérieure à 2000



TRAIT5

- Fréquence TOTALE : 20950
- Valeur : META#TYPE#CHARSET#VALUE#ISO88591#TYPE#CONTENT#VALUE#TEXTHTML#TYPE#HTTPEQUIV#VALUE#

TRAIT11

- Fréquence TOTALE : 5357
- Valeur : META#TYPE#CHARSET#VALUE#WINDOWS1252#TYPE#CONTENT#VALUE#TEXTHTML#TYPE#HTTPEQUIV#VALUE#

TRAIT25

- Fréquence TOTALE : 4619
- Valeur : META#TYPE#CONTENT#VALUE#MICROSOFTFRONTPAGE40#TYPE#NAME#VALUE#GENERATOR#

TRAIT6013

- Fréquence TOTALE : 4528
- Valeur : META#TYPE#CONTENT#VALUE#0#TYPE#HTTPEQUIV#VALUE##TYPE#URL#VALUE#HTTPWWWMULTIMANIAFRCOMMONERREUR404MULTIMANIACOM404HTML#

Ce trait est présent uniquement chez Multimania

TRAIT6

- Fréquence TOTALE : 4327

- Valeur : META#TYPE#CONTENT#VALUE#MICROSOFTFRONTPAGEEXPRESS20#TYPE#NAME#VALUE#GENERATOR#

TRAIT21

- Fréquence TOTALE : 4064
- Valeur : META#TYPE#CONTENT#VALUE#MICROSOFTFRONTPAGE30#TYPE#NAME#VALUE#GENERATOR#

Ce trait est très spécifique chez PPInfonie

TRAIT18

- Fréquence TOTALE : 3149
- Valeur : META#TYPE#CONTENT#VALUE#0#TYPE#HTTPEQUIV#VALUE##TYPE#URL#VALUE#HTTPWWWWANADOOFRPAGESPERSOHTMLEXCINCONNUHTML#

Ce trait est présent uniquement chez Wanadoo

TRAIT26

- Fréquence TOTALE : 3093
- Valeur : META#TYPE#CONTENT#VALUE#FRONTPAGEEDITORDOCUMENT#TYPE#NAME#VALUE#PROGID#

TRAIT33

- Fréquence TOTALE : 2046
- Valeur : META#TYPE#NAME#VALUE#KEYWORDS#

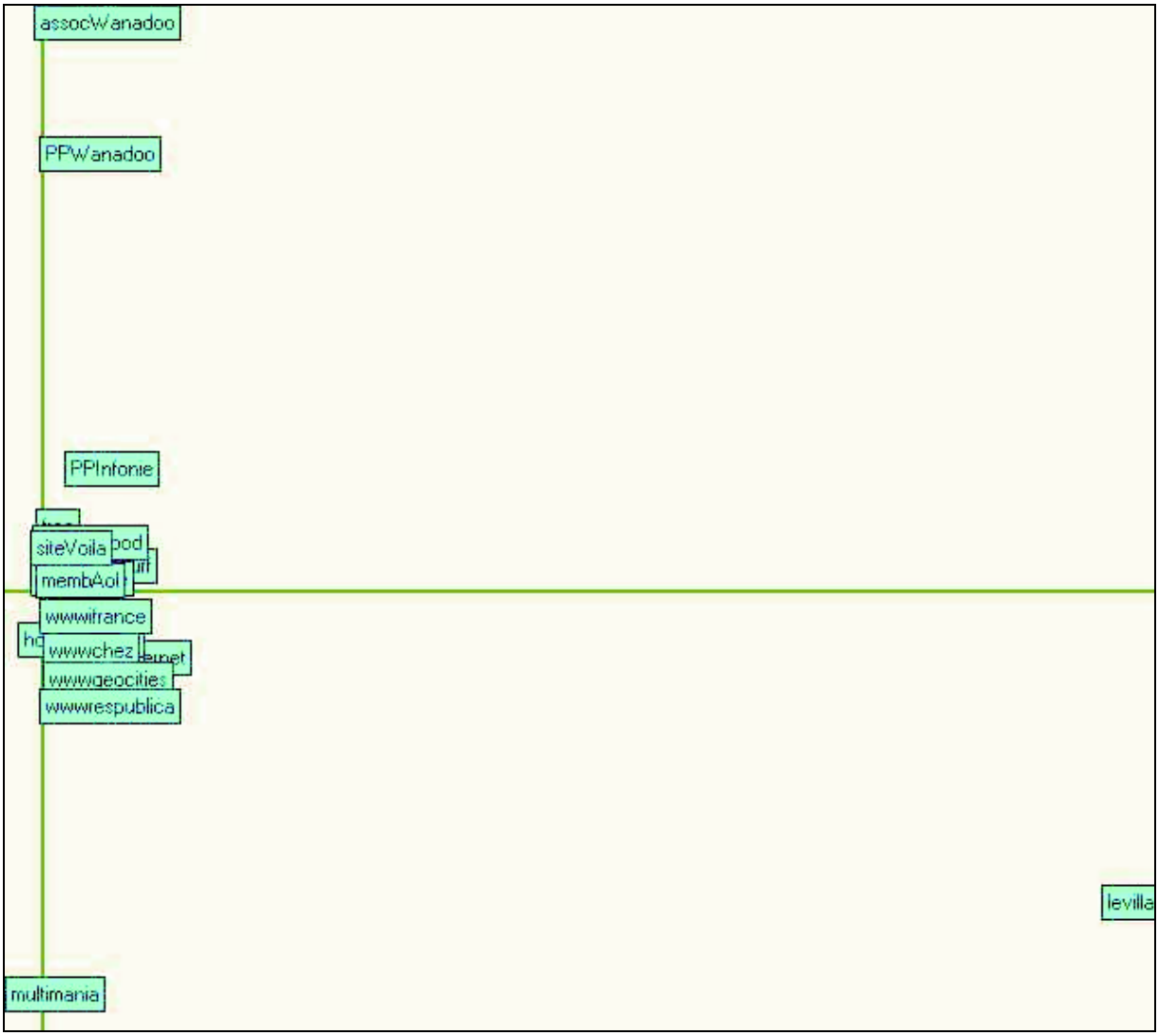
TRAIT20

- Fréquence TOTALE : 1948
- Valeur : META#TYPE#NAME#VALUE#GENERATOR#TYPE#CONTENT#VALUE#CLARISHOMEPAGE20#

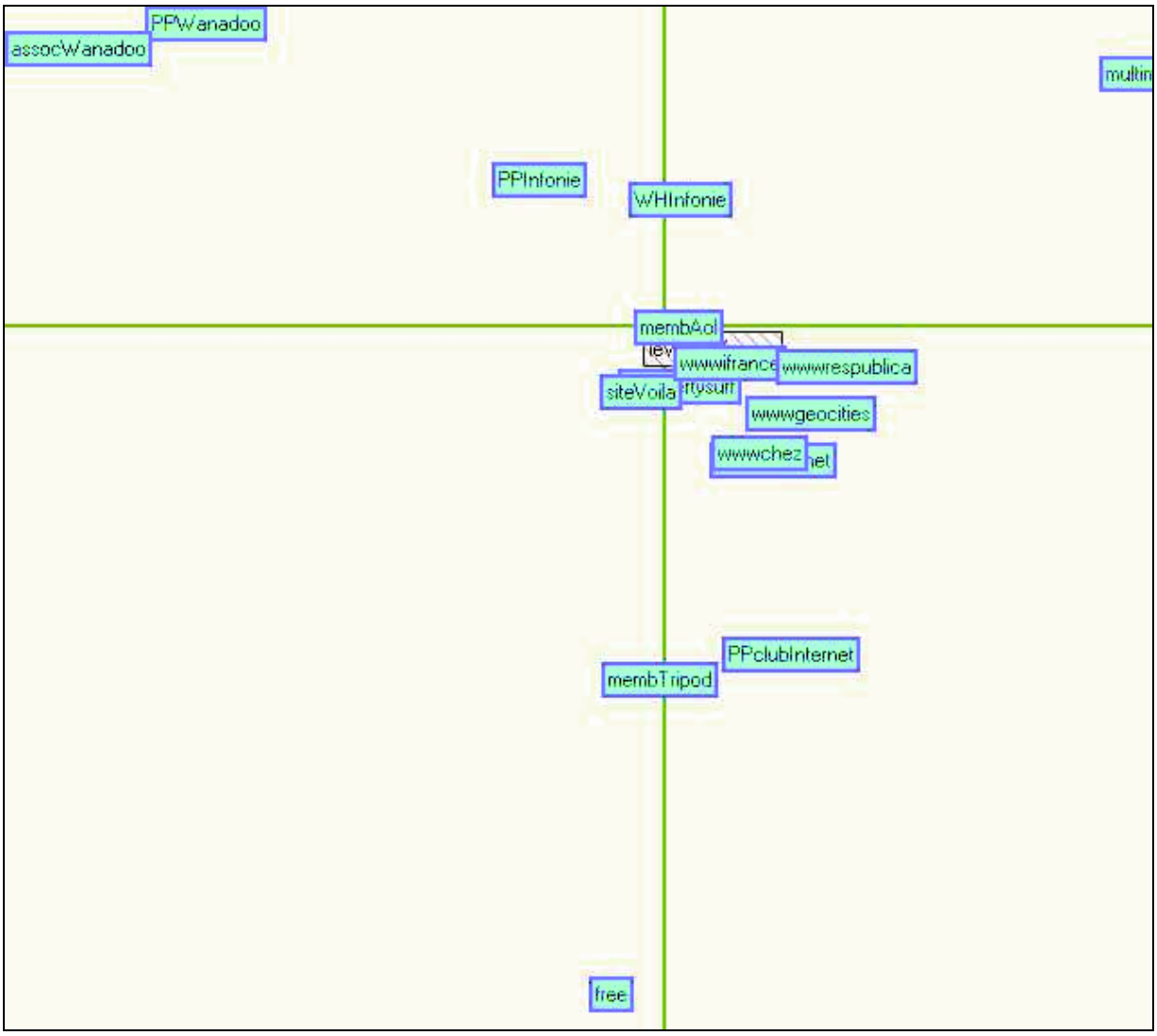
La plupart des traits précédents (très fréquents) contiennent une identification du générateur de la page.

4.1.1.3.1.4 AFC

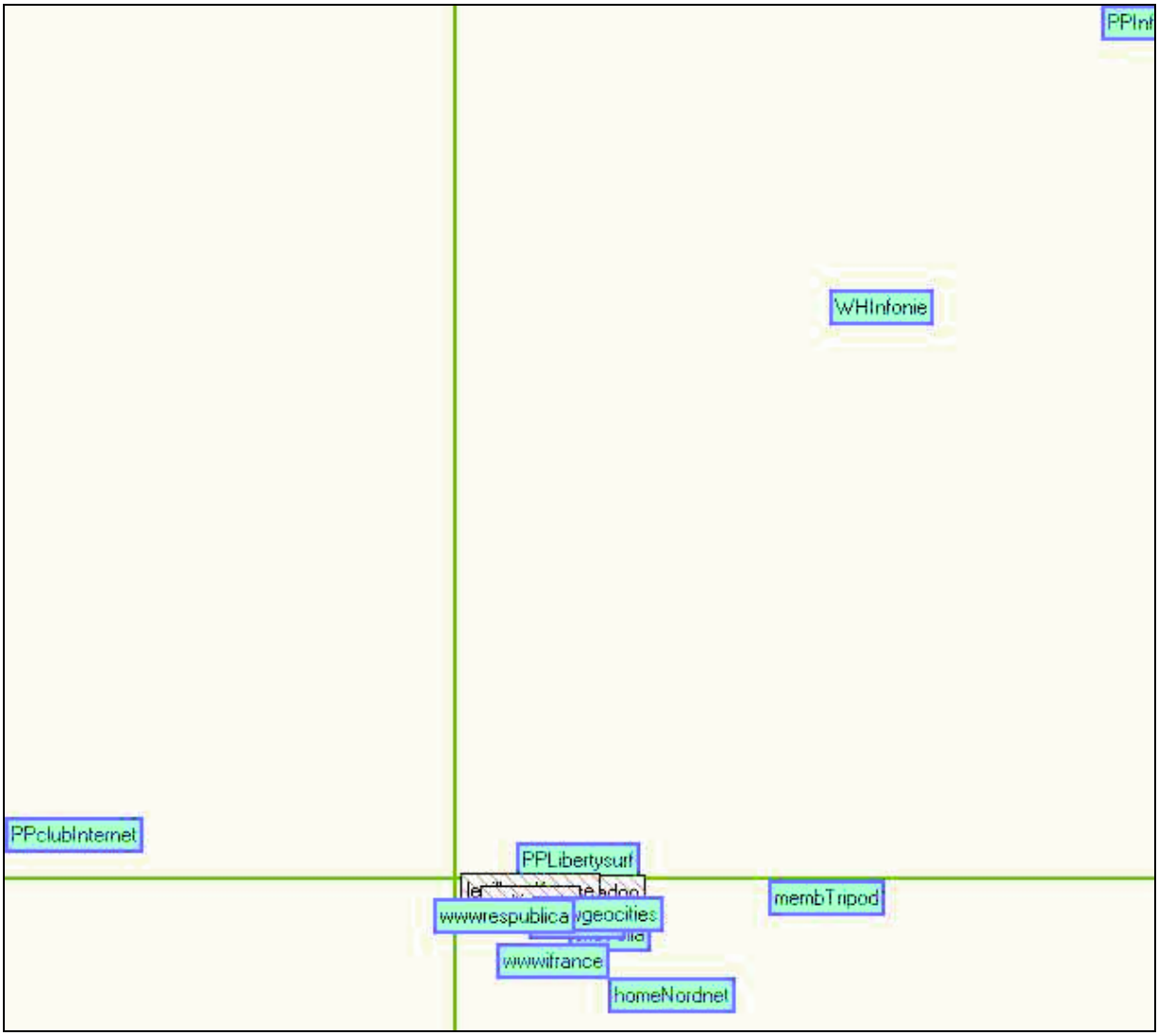
4.1.1.3.1.4.1 N° 1



4.1.1.3.1.4.2 N° 2


PPWanadoo
assocWanadoo
multin
PPInfonie
WHInfonie
membAol
wwwifrance
wwwrespublica
siteVoila
wwwgeocities
wwwchez
PPclubInternet
membTripod
free

4.1.1.3.1.4.3 N° 3


PPInf
WHInfonie
PPclubInternet
PPLibertysurf
membTripod
wwwrespublica
vgeocities
wwwifrance
homeNordnet

4.1.1.3.1.5 Segments répétés (frequence > 100)

Corpus :META-HEB2-CLEAN-CP100000 Date :samedi 12 mai 2001 - 22:52
Segments répétés

Longueur	Segment	Fréquence
2	TRAIT5 TRAIT1579	105
2	TRAIT5 TRAIT21	142
2	TRAIT5 TRAIT21	398
2	TRAIT5 TRAIT21	107
2	TRAIT5 TRAIT21	102
2	TRAIT5 TRAIT2257	108
2	TRAIT5 TRAIT245	154
2	TRAIT5 TRAIT25	130
2	TRAIT5 TRAIT25	103
2	TRAIT5 TRAIT25	470
3	TRAIT5 TRAIT2553 TRAIT6	114
2	TRAIT5 TRAIT409	479
2	TRAIT5 TRAIT55	139
2	TRAIT5 TRAIT6	275
2	TRAIT5 TRAIT6	625
2	TRAIT5 TRAIT6	2292
2	TRAIT5 TRAIT8	125
2	TRAIT5 TRAIT925	114
2	TRAIT5 TRAIT96	179
4	TRAIT11 TRAIT12 TRAIT13 TRAIT14	190
2	TRAIT11 TRAIT12	200
3	TRAIT11 TRAIT188 TRAIT416	330
5	TRAIT11 TRAIT188 TRAIT679 TRAIT680 TRAIT416	156
4	TRAIT11 TRAIT188 TRAIT679 TRAIT680	239
2	TRAIT11 TRAIT188	1016
2	TRAIT11 TRAIT25	2061
3	TRAIT11 TRAIT25 TRAIT26	2055
4	TRAIT11 TRAIT27 TRAIT25 TRAIT26	176
3	TRAIT11 TRAIT27 TRAIT25	178
2	TRAIT11 TRAIT27	195
2	TRAIT11 TRAIT5242	117
2	TRAIT11 TRAIT83	383
2	TRAIT25 TRAIT22	153
2	TRAIT25 TRAIT26	1596
2	TRAIT25 TRAIT26	1234
2	TRAIT21 TRAIT22	233
2	TRAIT21 TRAIT22	110
2	TRAIT21 TRAIT29	133
3	TRAIT21 TRAIT29 TRAIT22	100
2	TRAIT21 TRAIT590	206
2	TRAIT33 TRAIT294	116
2	TRAIT32 TRAIT33	382
3	TRAIT32 TRAIT33 TRAIT294	110

2	TRAIT32 TRAIT33	485
2	TRAIT32 TRAIT33	587
2	TRAIT139 TRAIT259	516
2	TRAIT139 TRAIT4878	106
4	TRAIT27 TRAIT11 TRAIT25 TRAIT26	495
2	TRAIT27 TRAIT11	507
3	TRAIT27 TRAIT25 TRAIT26	102
2	TRAIT27 TRAIT25	104
2	TRAIT188 TRAIT416	358
4	TRAIT188 TRAIT679 TRAIT680 TRAIT416	164
3	TRAIT188 TRAIT679 TRAIT680	247
2	TRAIT96 TRAIT6	112
2	TRAIT29 TRAIT22	143
2	TRAIT590 TRAIT47	103
3	TRAIT679 TRAIT680 TRAIT416	201
2	TRAIT679 TRAIT680	301
3	TRAIT12 TRAIT13 TRAIT14	235
2	TRAIT16574 TRAIT16575	176
2	TRAIT94 TRAIT95	101

Equivalence des traits :

TRAIT1579	**META#TYPE#NAME#VALUE#GENERATOR#TYPE#CONTENT#VALUE#MOZILLA403[FR](WIN95I)[NETSCAPE]#**
TRAIT2257	**META#TYPE#NAME#VALUE#GENERATOR#TYPE#CONTENT#VALUE#MOZILLA404[EN](WIN95I)[NETSCAPE]#**
TRAIT245	**META#TYPE#CONTENT#VALUE#MOZILLA47[FR](WIN98I)[NETSCAPE]#TYPE#NAME#VALUE#GENERATOR#**
TRAIT2553	**META#TYPE#CONTENT#VALUE#JPL#TYPE#NAME#VALUE#AUTHOR#**
TRAIT409	**META#TYPE#CONTENT#VALUE#MICROSOFTFRONTPAGE20#TYPE#NAME#VALUE#GENERATOR#**
TRAIT55	**META#TYPE#CONTENT#VALUE#MICROSOFTPUBLISHER2000#TYPE#NAME#VALUE#GENERATOR#**

4.1.1.3.2 Balises FONT, FRAME, TABLE, BODY

Analyse d'un corpus qui ne contient que les traits ci-dessus (avec attributs).

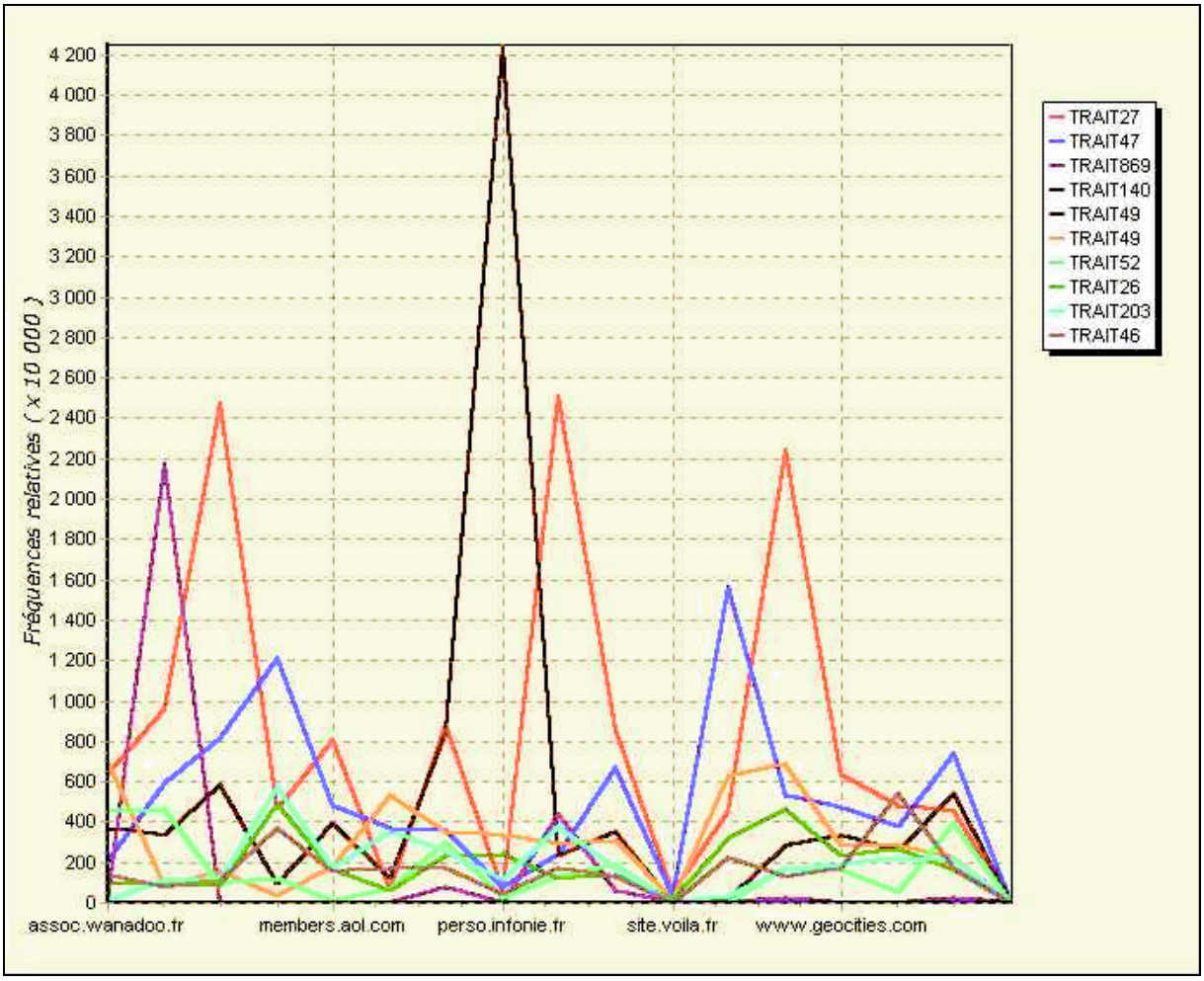
4.1.1.3.2.1 Etat du corpus

```
FOURFEATURES-HEB-CLEAN2-CP100000
2713060 2713060 163515 1700418 31295 166855 10763 5000000 3 3 132215 0 0

*** Résultat de la segmentation du fichier: FOURFEATURES-HEB-CLEAN2-
CP100000.TXT ***
Délimiteurs .,:;!?/_-\"'()[]{}§$£

      nombre des occurrences :     1700418
      nombre des formes      :       31295
      frequence maximale     :      166855
      nombre des hapax       :       10763
      nombre des clés(type)  :           3
      nombre des clés(ctnu)  :      132215

*** Fin de la segmentation du fichier:      FOURFEATURES-HEB-CLEAN2-
CP100000.TXT ***
```

4.1.1.3.2.2 Principales caractéristiques de la partition : HEBERGEUR

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
assoc.wanadoo.fr	4088	290	98	304	TRAIT490
free_fr	328660	5385	1785	71406	TRAIT869
home.nordnet.fr	17814	833	333	4400	TRAIT27
le-village.ifrance.com	15078	797	314	1829	TRAIT47
members.aol.com	37260	1972	771	3001	TRAIT27
membres.tripod.fr	22100	1304	509	1702	TRAIT5
perso.club-internet.fr	98647	3425	1303	8762	TRAIT27
perso.infonie.fr	2016	146	54	857	TRAIT140
perso.libertysurf.fr	16791	1023	420	4203	TRAIT27
perso.wanadoo.fr	333895	10491	3880	28720	TRAIT27
site.voila.fr	4617	138	46	804	TRAIT69002
webhome.infonie.fr	4024	345	150	628	TRAIT47
www.chez.com	253372	5930	2300	56773	TRAIT27
www.geocities.com	118619	5416	2080	7520	TRAIT27
www.ifrance.com	57378	2297	823	3328	TRAIT14831
www.multimania.com	385495	10300	3557	28541	TRAIT47
www.respublica.fr	564	7	4	553	TRAIT10

4.1.1.3.2.3 Répartition de traits dont la fréquence est supérieure à 25000



TRAIT27
- Fréquence TOTALE : 166855
- Valeur : FONT#TYPE#SIZE#VALUE#1#

TRAIT47
- Fréquence TOTALE : 102157
- Valeur : FONT#TYPE#SIZE#VALUE#2#

TRAIT869
- Fréquence TOTALE : 76481
- Valeur : FONT#TYPE#FACE#VALUE#TIMESNEWROMAN#TYPE#SIZE#VALUE#2#TYPE#COLOR#VALUE#000000#

Visiblement, trait spécifique à Free.fr

TRAIT140
- Fréquence TOTALE : 68973
- Valeur : FONT#TYPE#FACE#VALUE#ARIAL#

Visiblement, trait spécifique à Perso.infonie

TRAIT49
- Fréquence TOTALE : 51095

- Valeur : FONT#TYPE#SIZE#VALUE#+1#

TRAIT52
- Fréquence TOTALE : 47184
- Valeur : FONT#TYPE#FACE#VALUE#COMICSANSMS#

TRAIT26
- Fréquence TOTALE : 34456
- Valeur : FONT#TYPE#COLOR#VALUE#000000#

TRAIT46
- Fréquence TOTALE : 25411
- Valeur : FONT#TYPE#COLOR#VALUE#FF0000#

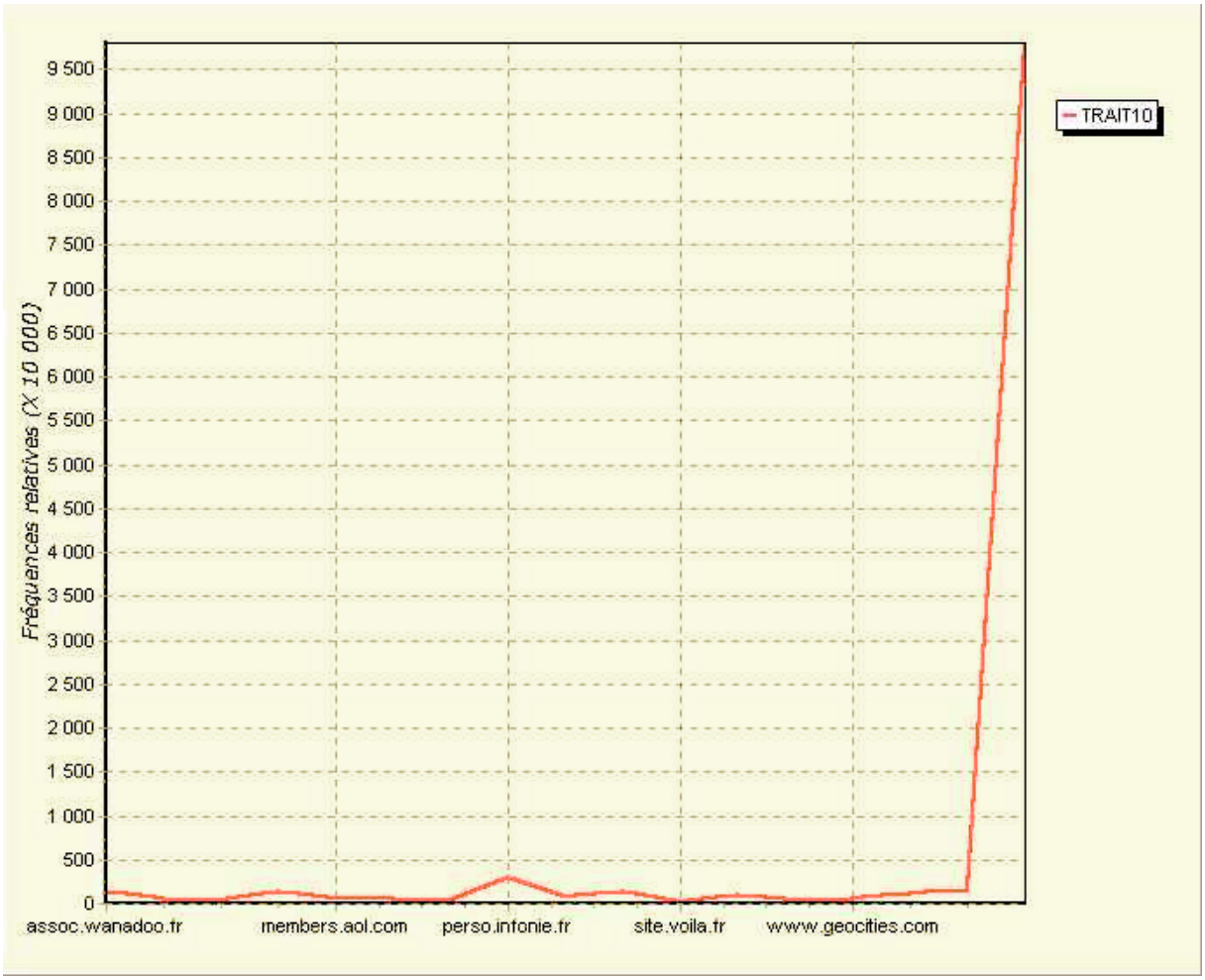
Tous ces attributs sont relatifs à la définition de police de caractères

4.1.1.3.2.4 Examen du trait BODY (occurrence > 100)

BODY	17253
BODY#TYPE#BGCOLOR#VALUE#FFFFFF#	2810
BODY#TYPE#BGCOLOR#VALUE#000000#	1622
BODY#TYPE#TEXT#VALUE#FFFFFF#TYPE#BGCOLOR#VALUE#000000#	610
BODY#TYPE#BACKGROUND#VALUE#BORDBOISJPG#	427
BODY#TYPE#TEXT#VALUE#000000#TYPE#BGCOLOR#VALUE#FFFFFF#	363
BODY#TYPE#BACKGROUND#VALUE#FONDGIF#	285
BODY#TYPE#BGCOLOR#VALUE#BLACK#	260
BODY#TYPE#BGCOLOR#VALUE#WHITE#	235
BODY#TYPE#BACKGROUND#VALUE#IMAGESFONDJPG#	232
BODY#TYPE#ONUNLOAD#VALUE#RELPUBLOC()#	199
BODY#TYPE#BACKGROUND#VALUE#FONDJPG#	183
BODY#TYPE#BGCOLOR#VALUE#FFFFCC#	167
BODY#TYPE#LINK#VALUE#0000FF#TYPE#VLINK#VALUE#800000#TYPE#TEXT#VALUE#000000#TYPE#BGCOLOR#VALUE#F0F0F0#	146
BODY#TYPE#ONLOAD#VALUE#TRAP()#	142
BODY#TYPE#BACKGROUND#VALUE#FOND1JPG#	132
BODY#TYPE#LINK#VALUE#0000FF#TYPE#VLINK#VALUE#800080#	123
BODY#TYPE#BACKGROUND#VALUE#FOND1GIF#	114

Le trait BODY est utilisé très fréquemment sans attribut. (Voir aussi corpus suivant pour complément)

En examinant la ventilation de ce trait, on s'aperçoit en fait qu'il est très fréquent (relativement) chez un hébergeur particulier :



4.1.1.3.2.5 Examen du trait TABLE (occurrence > 100)

TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	8644
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#	6959
TABLE#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#100%#	5884
TABLE#TYPE#BORDER#VALUE#0#	5616
TABLE	**4998**
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#	4944
TABLE#TYPE#RULES#VALUE#ALL#TYPE#FRAME#VALUE#BOX#TYPE#BORDER#VALUE#BORDER#TYPE#WIDTH#VALUE#97%#	3381
TABLE#TYPE#WIDTH#VALUE#85%#	3105
TABLE#TYPE#WIDTH#VALUE#100%#	2080
TABLE#TYPE#BORDER#VALUE#1#	1817
TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#0#	1133
TABLE#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#600#	960
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#TYPE#WIDTH#VALUE#100%#	838
TABLE#TYPE#BORDER#VALUE#2#	725
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#0#TYPE#WIDTH#VALUE#468#TYPE#CELLSPACING#VALUE#0#	583
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#1#TYPE#WIDTH#VALUE#468#TYPE#BGCOLOR#VALUE#666666#TYPE#CELLSPACING#VALUE#0#	561

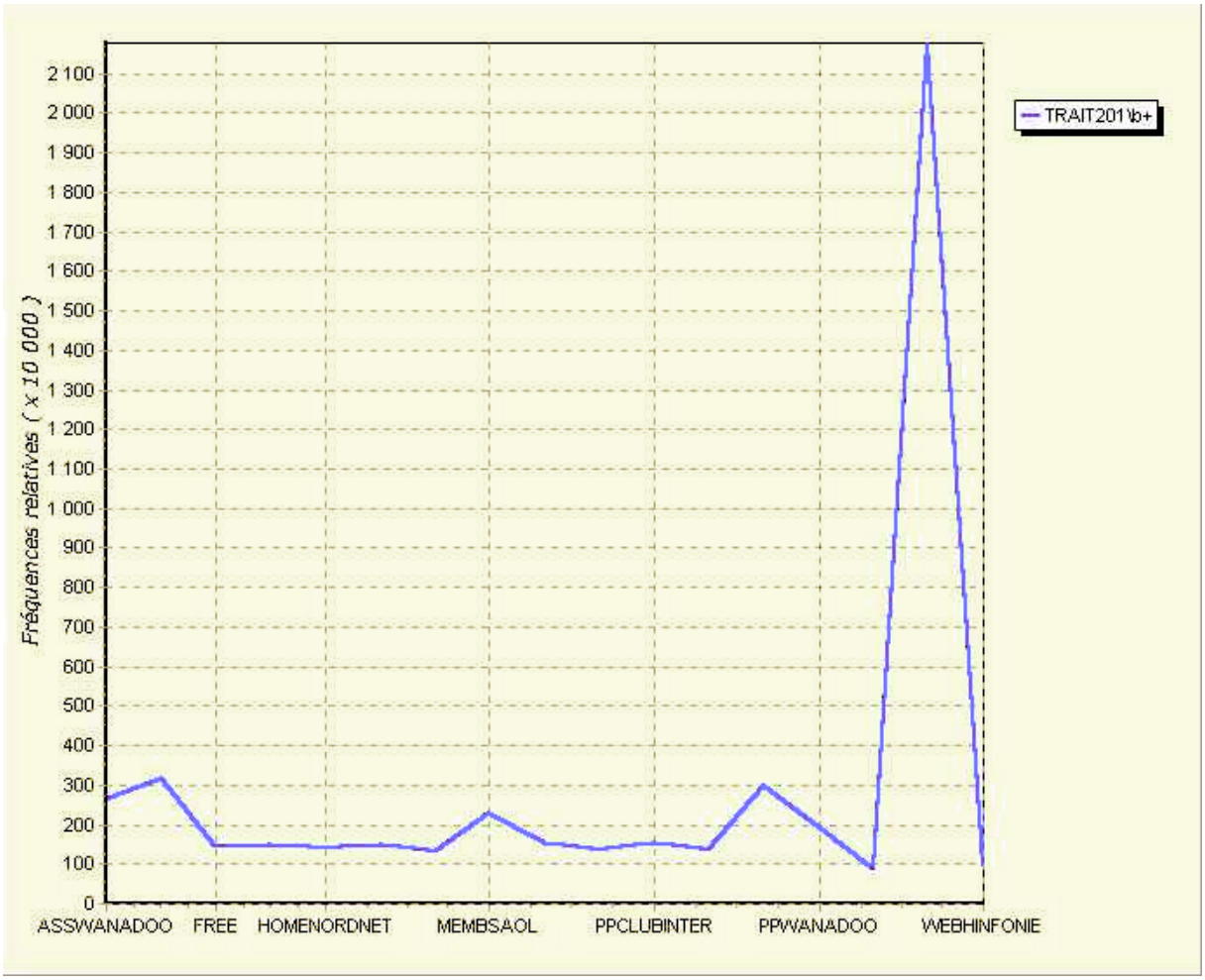
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#2#TYPE#WIDTH#VALUE#100%#TYPE#BGCOLOR#VALUE#EEEEEE#TYPE#CELLSPACING#VALUE#0#	561
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#7#	542
TABLE#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#95%#	535
TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLPADDING#VALUE#0#	427
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#576#TYPE#CELLSPACING#VALUE#0#	416
TABLE#TYPE#ALIGN#VALUE#CENTER#	412
TABLE#TYPE#BGCOLOR#VALUE#A0A0A4#TYPE#WIDTH#VALUE#100%#	384
TABLE#TYPE#CELLPADDING#VALUE#2#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#2#	379
TABLE#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#90%#TYPE#CELLSPACING#VALUE#1#	378
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#TYPE#CELLPADDING#VALUE#2#TYPE#WIDTH#VALUE#762#	335
TABLE#TYPE#RULES#VALUE#ALL#TYPE#FRAME#VALUE#BOX#TYPE#BORDER#VALUE#BORDER#TYPE#WIDTH#VALUE#100%#	321
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#0#TYPE#CELLSPACING#VALUE#0#	310
TABLE#TYPE#CELLPADDING#VALUE#2#TYPE#WIDTH#VALUE#100%#TYPE#ALIGN#VALUE#CENTER#TYPE#BORDER#VALUE#1#TYPE#CELLSPACING#VALUE#0#	291
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#90%#	286
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#600#	276
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	265
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	263
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#75%#	261
TABLE#TYPE#CELLPADDING#VALUE#2#	240
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#600#TYPE#CELLSPACING#VALUE#0#	236
TABLE#TYPE#BORDER#VALUE#0#TYPE#ALIGN#VALUE#CENTER#	235
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#	227
TABLE#TYPE#BORDER#VALUE#BORDER#TYPE#WIDTH#VALUE#100%#	219
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#TYPE#CELLPADDING#VALUE#2#TYPE#WIDTH#VALUE#100%#	213
TABLE#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#75%#	213
TABLE#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#88%#	211
TABLE#TYPE#BORDER#VALUE#BORDER#	209
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#	208
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#780#TYPE#CELLSPACING#VALUE#0#	193
TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLPADDING#VALUE#0#	187
TABLE#TYPE#BORDER#VALUE#1#TYPE#WIDTH#VALUE#18%#	184
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#95%#	180
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#BGCOLOR#VALUE#BLUE#	178
TABLE#TYPE#HSPACE#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#TYPE#VSPACE#VALUE#0#TYPE#CELLPADDING#VALUE#0#	175
TABLE#TYPE#CELLPADDING#VALUE#5#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	175
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#TYPE#CELLPADDING#VALUE#0#	174
TABLE#TYPE#HSPACE#VALUE#0#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#VSPACE#VALUE#0#TYPE#CELLSPACING#VALUE#0#	173

TABLE#TYPE#WIDTH#VALUE#97%#	166
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#TYPE#WIDTH#VALUE#100%#	164
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#4#	163
TABLE#TYPE#MSIMAGELIST#VALUE#MSIMAGELIST#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	163
TABLE#TYPE#CELLPADDING#VALUE#2#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	162
TABLE#TYPE#WIDTH#VALUE#90%#	156
TABLE#TYPE#BORDERCOLOR#VALUE#0000FF#TYPE#CELLPADDING#VALUE#3#TYPE#BORDER#VALUE#1#TYPE#BORDERCOLORLIGHT#VALUE#00FFFF#	150
TABLE#TYPE#CELLSPACING#VALUE#10#	150
TABLE#TYPE#BORDER#VALUE#0#TYPE#COLS#VALUE#2#TYPE#WIDTH#VALUE#100%#	147
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#CELLSPACING#VALUE#0#	146
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#80%#	146
TABLE#TYPE#WIDTH#VALUE#87%#	144
TABLE#TYPE#CELLPADDING#VALUE#2#TYPE#BORDER#VALUE#0#	142
TABLE#TYPE#COLS#VALUE#2#TYPE#WIDTH#VALUE#100%#	142
TABLE#TYPE#WIDTH#VALUE#50%#	141
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#80%#TYPE#CELLSPACING#VALUE#0#	137
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#90%#TYPE#CELLSPACING#VALUE#0#	137
TABLE#TYPE#BORDER#VALUE#5#	136
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#10#	134
TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#WIDTH#VALUE#100%#	132
TABLE#TYPE#BORDER#VALUE#BORDER#TYPE#CELLPADDING#VALUE#2#	129
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#	127
TABLE#TYPE#CELLPADDING#VALUE#1#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#BGCOLOR#VALUE#FFFFFF#TYPE#CELLSPACING#VALUE#0#	125
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#750#	124
TABLE#TYPE#BORDER#VALUE#2#TYPE#WIDTH#VALUE#100%#	123
TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#0#TYPE#WIDTH#VALUE#100%#	123
TABLE#TYPE#BORDER#VALUE#3#	122
TABLE#TYPE#WIDTH#VALUE#88%#	121
TABLE#TYPE#HEIGHT#VALUE#45#TYPE#ALIGN#VALUE#CENTER#TYPE#WIDTH#VALUE#700#	121
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLPADDING#VALUE#0#	120
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#620#TYPE#CELLSPACING#VALUE#0#	119
TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#CELLPADDING#VALUE#0#	118
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	117
TABLE#TYPE#BORDER#VALUE#4#	116
TABLE#TYPE#BORDER#VALUE#0#TYPE#CELLPADDING#VALUE#2#	115
TABLE#TYPE#CELLPADDING#VALUE#2#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#	115
TABLE#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#500#	114
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#750#TYPE#CELLSPACING#VALUE#0#	114
TABLE#TYPE#CELLPADDING#VALUE#1#TYPE#WIDTH#VALUE#95%#TYPE#ALIGN#VALUE#CENTER#TYPE#BGCOLOR#VALUE#WHITE#TYPE#CELLSPACING#VALUE#0#	113
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#CELLSPACING#VALUE#5#	113

TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#ALIGN#VALUE#CENTER#TYPE#CELLPADDING#VALUE#0#	112
TABLE#TYPE#BORDER#VALUE#BORDER#TYPE#COLS#VALUE#2#TYPE#WIDTH#VALUE#100%#	112
TABLE#TYPE#CELLSPACING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#200#TYPE#CELLPADDING#VALUE#0#	112
TABLE#TYPE#HEIGHT#VALUE#100%#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#	110
TABLE#TYPE#CELLPADDING#VALUE#1#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#BGCOLOR#VALUE#FFCC00#TYPE#CELLSPACING#VALUE#0#	108
TABLE#TYPE#CELLPADDING#VALUE#1#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#150#TYPE#CELLSPACING#VALUE#0#	108
TABLE#TYPE#WIDTH#VALUE#80%#	107
TABLE#TYPE#CELLPADDING#VALUE#3#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	106
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#0#TYPE#CELLSPACING#VALUE#0#	106
TABLE#TYPE#CELLPADDING#VALUE#5#	102
TABLE#TYPE#CELLSPACING#VALUE#2#TYPE#BORDER#VALUE#1#TYPE#CELLPADDING#VALUE#0#TYPE#WIDTH#VALUE#450#	101
TABLE#TYPE#BORDER#VALUE#BORDER#TYPE#COLS#VALUE#1#TYPE#WIDTH#VALUE#100%#	101
TABLE#TYPE#BORDER#VALUE#10#	101
TABLE#TYPE#CELLSPACING#VALUE#1#TYPE#BORDER#VALUE#BORDER#TYPE#WIDTH#VALUE#450#TYPE#CELLPADDING#VALUE#4#	100

A la différence de la page (via le trait BODY), la définition d'un tableau est plus fortement contrainte.

Ventilation du trait TABLE (les 5 premiers)



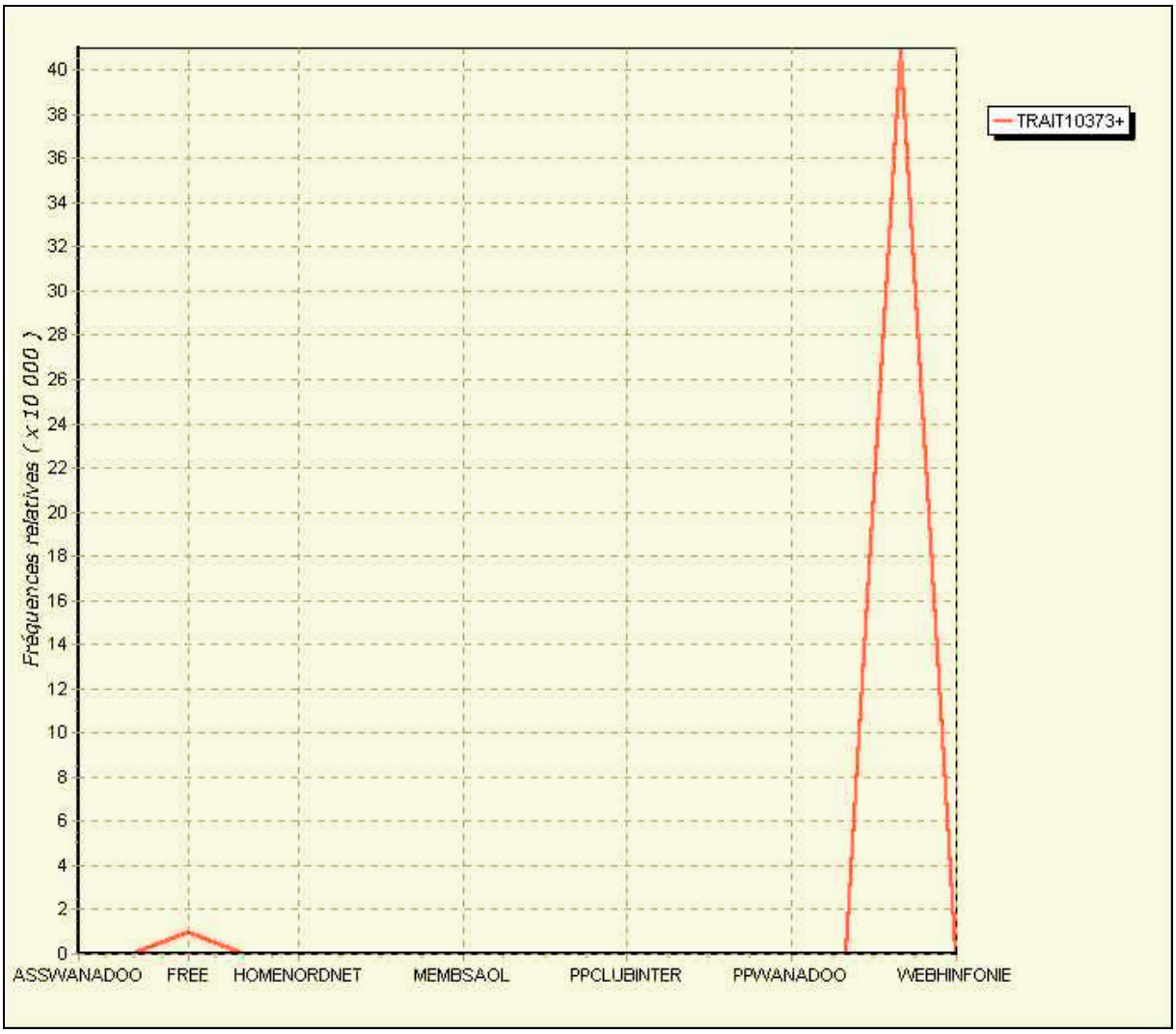
Le pic correspond à l'hébergeur SITEVOILA.

4.1.1.3.2.6 Examen du trait FRAME (occurrence > 10)

FRAME#TYPE#NORESIZE#VALUE#NORESIZE#TYPE#FRAMEBORDER#VALUE#NO#TYPE#MARGINWIDTH#VALUE#0#TYPE#SCROLLING#VALUE#NO#TYPE#NAME#VALUE#MENU#TYPE#MARGINHEIGHT#VALUE#0#TYPE#SRC#VALUE#MENUHTML#	37
FRAME#TYPE#SRC#VALUE#BLANKHTM#TYPE#NAME#VALUE#F4#	28
FRAME#TYPE#MARGINHEIGHT#VALUE#0#TYPE#NORESIZE#VALUE#NORESIZE#TYPE#SRC#VALUE#HTTPSITEVOILAFRPUBLICDIVERSTOPHTML#TYPE#SCROLLING#VALUE#NO#TYPE#MARGINWIDTH#VALUE#0#TYPE#NAME#VALUE#TOP#	20
FRAME#TYPE#NORESIZE#VALUE#NORESIZE#TYPE#FRAMEBORDER#VALUE#NO#TYPE#MARGINWIDTH#VALUE#0#TYPE#SCROLLING#VALUE#NO#TYPE#NAME#VALUE#TOP#TYPE#MARGINHEIGHT#VALUE#0#TYPE#SRC#VALUE#TOPHTML#	20
FRAME#TYPE#SRC#VALUE#SOMMAIREHTM#TYPE#SCROLLING#VALUE#AUTO#TYPE#TARGET#VALUE#PRINCIPAL#TYPE#NAME#VALUE#SOMMAIRE#	19
FRAME#TYPE#NORESIZE#VALUE#NORESIZE#TYPE#FRAMEBORDER#VALUE#NO#TYPE#MARGINWIDTH#VALUE#0#TYPE#SCROLLING#VALUE#NO#TYPE#NAME#VALUE#TOPLEFT#TYPE#MARGINHEIGHT#VALUE#0#TYPE#SRC#VALUE#TOPLEFTHTML#	17
FRAME#TYPE#NAME#VALUE#MENU#TYPE#SRC#VALUE#MENUHTML#	16
FRAME#TYPE#SRC#VALUE#MENUHTML#TYPE#NAME#VALUE#MENU#	15
FRAME#TYPE#SRC#VALUE#MENUHTM#TYPE#NAME#VALUE#MENU#	15
FRAME#TYPE#FRAMEBORDER#VALUE#NO#TYPE#MARGINWIDTH#VALUE#0#TYPE#NAME#VALUE#MAIN#TYPE#MARGINHEIGHT#VALUE#0#TYPE#SRC#VALUE#MAIN1HTML#	15
FRAME#TYPE#SRC#VALUE#SOMMAIREHTM#TYPE#TARGET#VALUE#PRINCIPAL#TYPE#NAME#VALUE#SOMMAIRE#	14
FRAME#TYPE#SRC#VALUE#MENUHTM#	13
FRAME#TYPE#NORESIZE#VALUE#NORESIZE#TYPE#MARGINWIDTH#VALUE#0#TYPE#SCROLLING#VALUE#NO#TYPE#NAME#VALUE#TOPLEFT#TYPE#MARGINHEIGHT#VALUE#0#TYPE#SRC#VALUE#TOPLEFTHTML#	12
FRAME#TYPE#SRC#VALUE#ACCUEILHTM#TYPE#NAME#VALUE#PRINCIPAL#	11
FRAME#TYPE#MARGINWIDTH#VALUE#0#TYPE#SCROLLING#VALUE#NO#TYPE#NAME#VALUE#VIDE#TYPE#MARGINHEIGHT#VALUE#0#TYPE#SRC#VALUE#VIDEHTML#	10
FRAME#TYPE#NORESIZE#VALUE#NORESIZE#TYPE#MARGINWIDTH#VALUE#0#TYPE#SCROLLING#VALUE#NO#TYPE#NAME#VALUE#TOP#TYPE#MARGINHEIGHT#VALUE#0#TYPE#SRC#VALUE#TOPHTML#	10

Ventilation du trait FRAME (les 5 premiers)

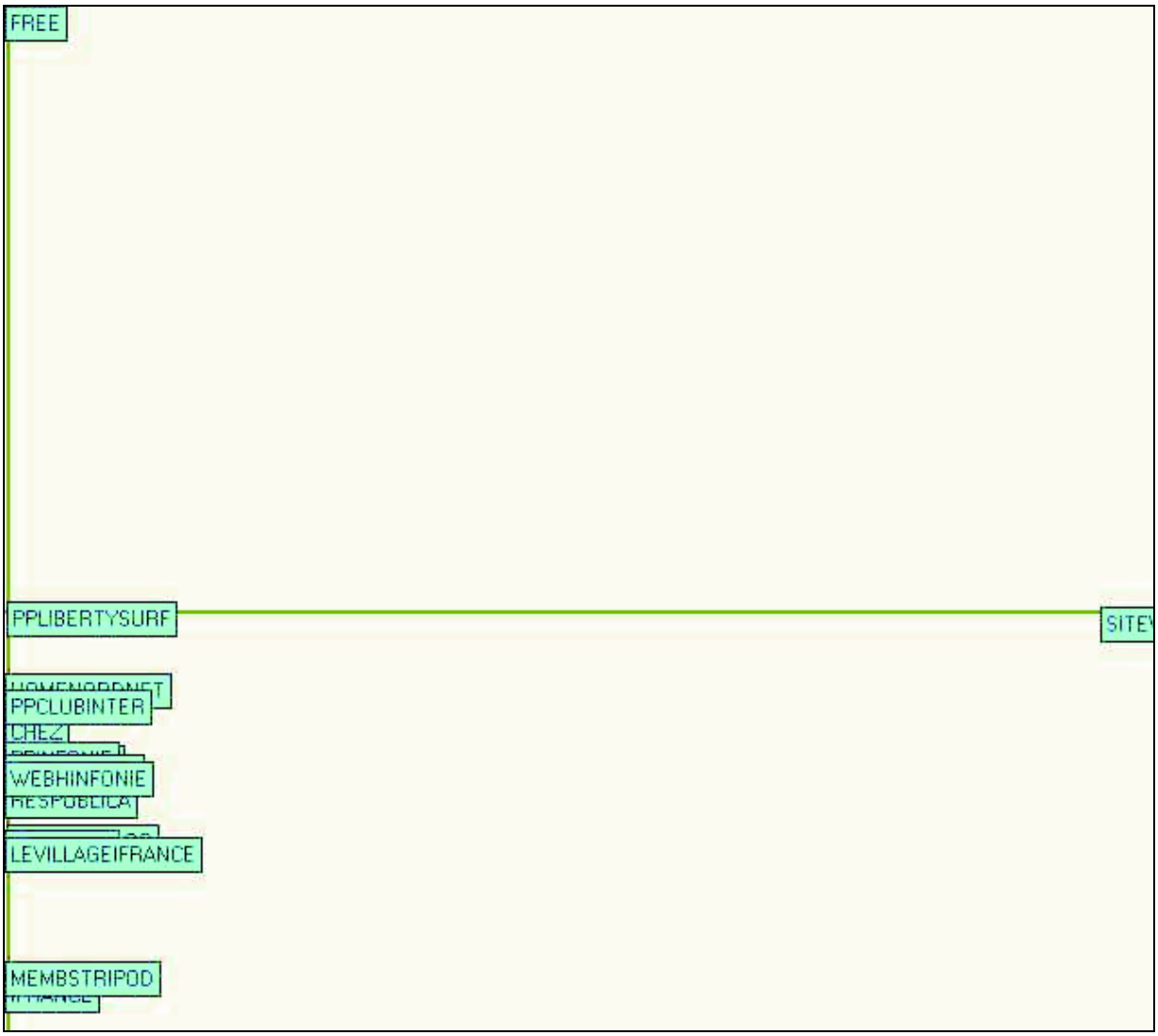
Le trait n'est pas présent sur toutes les familles d'hébergeurs (l'ordre correspond à celui du tableau des "Principales caractéristiques de la partition"). On retrouve cette disparité sur la graphique de ventilation du trait :



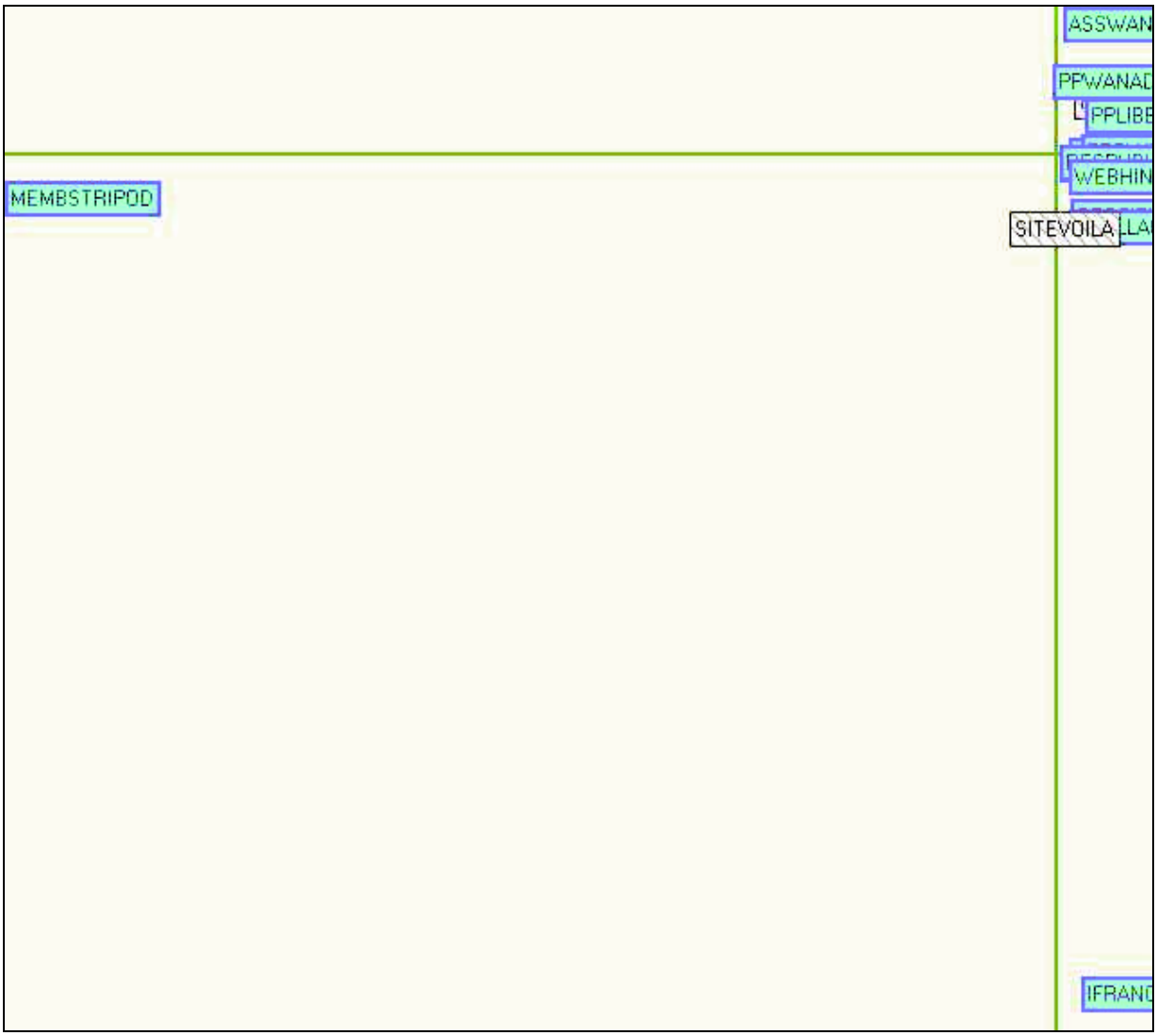
Le pic correspond de nouveau à l'hébergeur SITEVOILA.

4.1.1.3.2.7 AFC

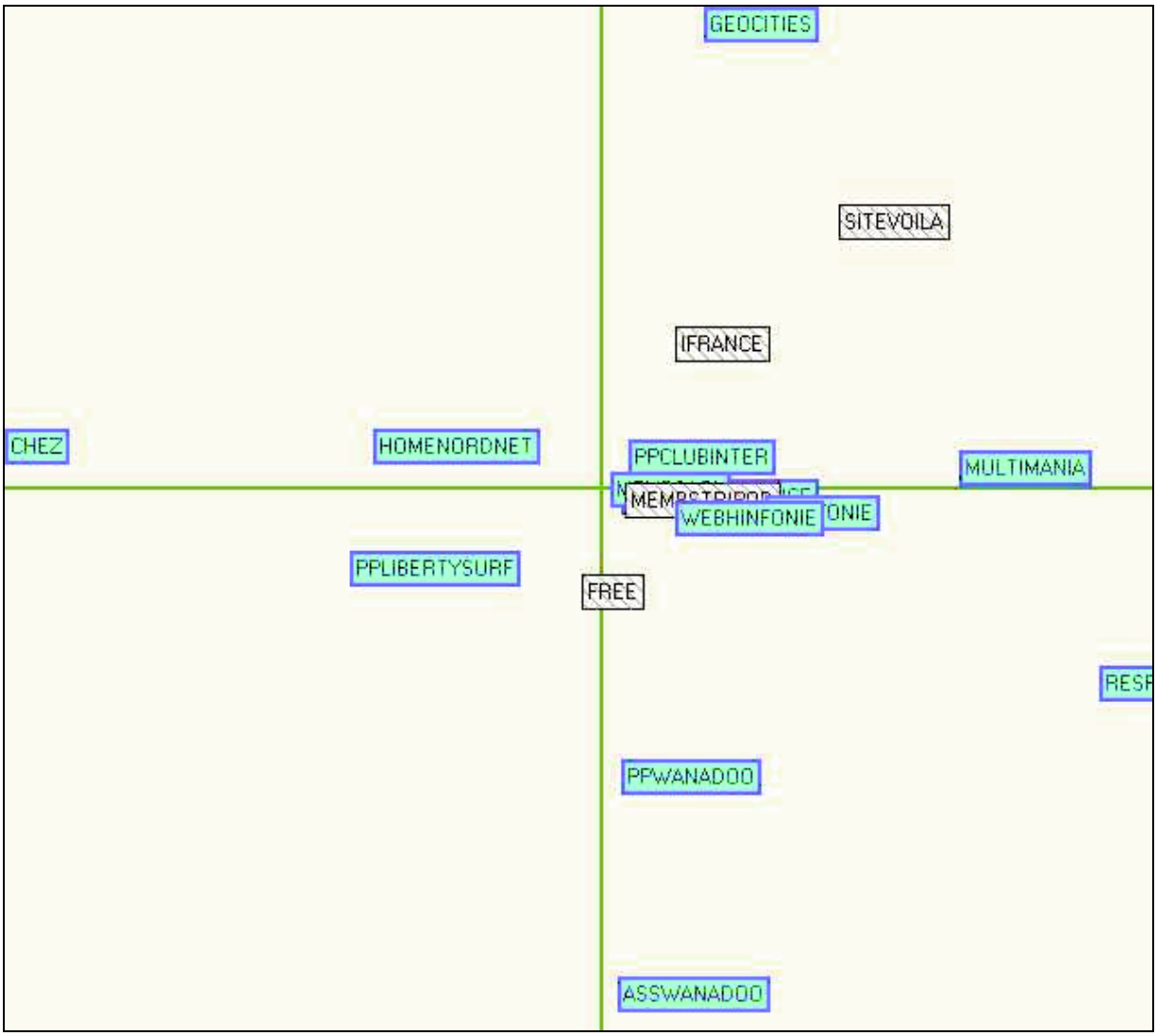
4.1.1.3.2.7.1 N° 1



4.1.1.3.2.7.2 N° 2


ASSWAN
PPWANAD
PPLIBE
WEBHIN
MEMBSTRIPOD
SITEVOILA
IFRAND

4.1.1.3.2.7.3 N° 3


GEOCITIES
SITEVOILA
IFRANCE
CHEZ
HOMENORDNET
PPCLUBINTER
MULTIMANIA
WEBHINFONIE
PPLIBERTYSURF
FREE
PPWANADOO
ASSWANADOO

4.1.1.3.3 Balises FONT, FRAME, TABLE, BODY, STYLE

4.1.1.3.3.1 Etat du corpus

```
FIVEFEATURES3-HEB-CLEAN-CP100000
2523145 2523145 164260 1707405 31307 166855 10766 5000000 3 3 132948 0 0

*** Résultat de la segmentation du fichier: FIVEFEATURES3-HEB-CLEAN-
CP100000.TXT ***
Délimiteurs .,:;!?/_-\"'()[]{}§$£

      nombre des occurrences :     1707405
      nombre des formes      :       31307
      frequence maximale     :      166855
      nombre des hapax       :       10766
      nombre des clés(type)  :           3
      nombre des clés(ctnu)  :      132948

*** Fin de la segmentation du fichier:      FIVEFEATURES3-HEB-CLEAN-
CP100000.TXT ***
```

4.1.1.3.3.2 Principales caractéristiques de la partition : HEBERGEUR

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
ASSWANADOO	4101	292	98	304	TRAIT493
CHEZ	254253	5937	2302	56773	TRAIT27
FREE	330119	5390	1785	71406	TRAIT872
GEOCITIES	119182	5421	2081	7520	TRAIT27
HOMENORDNET	17878	836	333	4400	TRAIT27
IFRANCE	57673	2301	823	3328	TRAIT14840
MEMB-AOL	37355	1976	771	3001	TRAIT27
MEMB-TRIPOD	22212	1307	509	1702	TRAIT5
MULTIMANIA	387137	10305	3558	28541	TRAIT47
PP-CLUBINTER	99055	3430	1305	8762	TRAIT27
PP-INFONIE	2021	148	54	857	TRAIT141
PP-LIBERTYSURF	16899	1026	420	4203	TRAIT27
PP-WANADOO	335033	10495	3881	28720	TRAIT27
RESPUBLICA	564	7	4	553	TRAIT10
SITEVOILA	4619	139	46	804	TRAIT69027
VILLAG-IFRANCE	15248	800	314	1829	TRAIT47
WH-INFONIE	4056	347	150	628	TRAIT47

Valeur des traits du tableau précédents :

TRAIT493	FONT#TYPE#SIZE#VALUE#2#TYPE#COLOR#VALUE#1C6DDD#
TRAIT27	FONT#TYPE#SIZE#VALUE#1#
TRAIT872	FONT#TYPE#FACE#VALUE#TIMESNEWROMAN#TYPE#SIZE#VALUE#2#TYPE#COLOR#VALUE#000000#
TRAIT14840	FONT#TYPE#SIZE#VALUE#2#TYPE#COLOR#VALUE#FFFFFF#TYPE#FACE#VALUE#VERDANAARIALHELVETICA#
TRAIT5	FONT#TYPE#SIZE#VALUE#2#TYPE#FACE#VALUE#ARIAL#
TRAIT47	FONT#TYPE#SIZE#VALUE#2#
TRAIT141	FONT#TYPE#FACE#VALUE#ARIAL#
TRAIT10	BODY
TRAIT69027	FONT#TYPE#FACE#VALUE#GENEVAARIAL#TYPE#SIZE#VALUE#1#TYPE#COLOR#VALUE#ALGO#

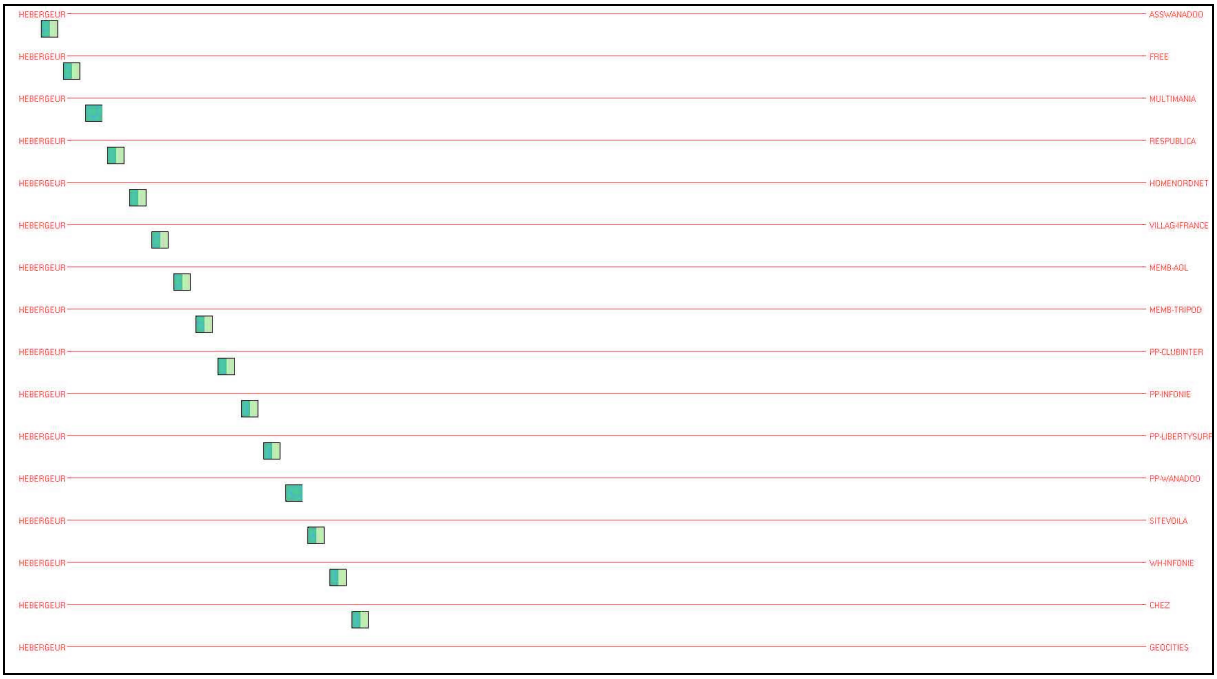
4.1.1.3.3.3 Examen du trait STYLE (occurrence > 100)

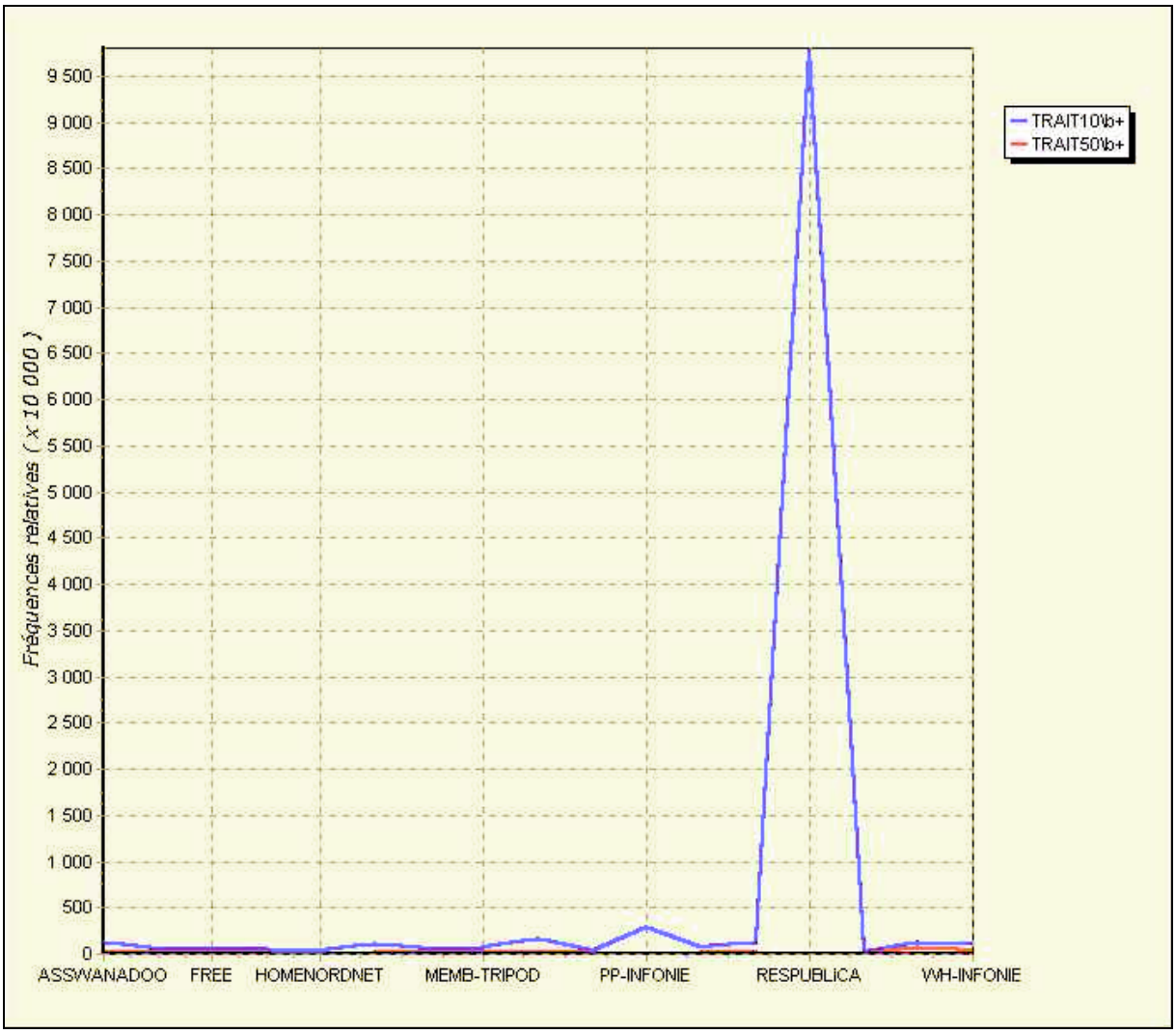
STYLE#TYPE#TYPE#VALUE#TEXTCSS#	3574	
STYLE	**3079**	**TRAIT50**
STYLE#TYPE#FPROLLOVERSTYLE#VALUE#FPROLLOVERSTYLE#	299	
STYLE#TYPE#ID#VALUE#NOFSTYLESHEET#	17	
STYLE#TYPE#KORESQUE#VALUE#KORESQUE#	10	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#NAME#VALUE#CUSTOM#	4	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#TITLE#VALUE#ANIMATION#	4	
STYLE#TYPE#TYPE#VALUE#SPOOKYCSS#	3	
STYLE#TYPE#TITLE#VALUE#NONE#TYPE#TYPE#VALUE#TEXTCSS#	3	
STYLE#TYPE#ID#VALUE#EZ2#	2	
STYLE#TYPE#TYPE#VALUE#TEXTJAVASCRIPT#	2	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#SRC#VALUE#STYLECSS#	2	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#SRC#VALUE#STYLEGJMCSS#	1	
STYLE#TYPE#ID#VALUE#DEMOINDEX121119996270STYLES#	1	
STYLE#TYPE#CONTENT#VALUE#PETERLYNCHINVESTISSEMENTBOURSEANALYSEFONDAMENTALEFINANCERECOMMANDATIONSBOURSIERES#	1	
STYLE#TYPE#ID#VALUE#MENU#	1	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#ID#VALUE#JOUSTSTYLES#	1	
STYLE#TYPE#TITLE#VALUE#VE#TYPE#TYPE#VALUE#TEXTCSS#	1	
STYLE#TYPE#HREF#VALUE#OLIVIEREVERAERE#TYPE#REL#VALUE#OWNS#	1	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#SRC#VALUE#MKCSS#	1	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#NAME#VALUE#MIDNIGHTBLACK#	1	
STYLE#TYPE#TYPE#VALUE#TEXTCSS#TYPE#SRC#VALUE#STYLEMENUCSS#	1	
STYLE#TYPE#TEXT#VALUE#TEXTCSS#	1	
STYLE#TYPE#SRC#VALUE#HTTPPERSOWANADOOFRALAINLAEMLESTYLESCSS#TYPE#TYPE#VALUE#TEXTCSS#	1	
STYLE#TYPE#ID#VALUE#UPESE26595STYLES#	1	
STYLE#TYPE#TYPE#VALUE#TXTCSS#	1	
STYLE#TYPE#ID#VALUE#A712STYLES#	1	
STYLE#TYPE#TYPE#VALUE#	1	
STYLE#TYPE#ID#VALUE#CLASSEMENTCLASSIC125556STYLES#	1	
STYLE#TYPE#HREF#VALUE#HREF#	1	
STYLE#TYPE#ID#VALUE#SOMMAIREDESPAROLES8756STYLES#	1	

4.1.1.3.3.4 Coprésence des traits BODY et STYLE

La "faiblesse" du trait BODY n'est semble-t-il pas compensée par le trait STYLE :

STYLE	**3079**	**TRAIT50**
BODY	**17252**	**TRAIT10**

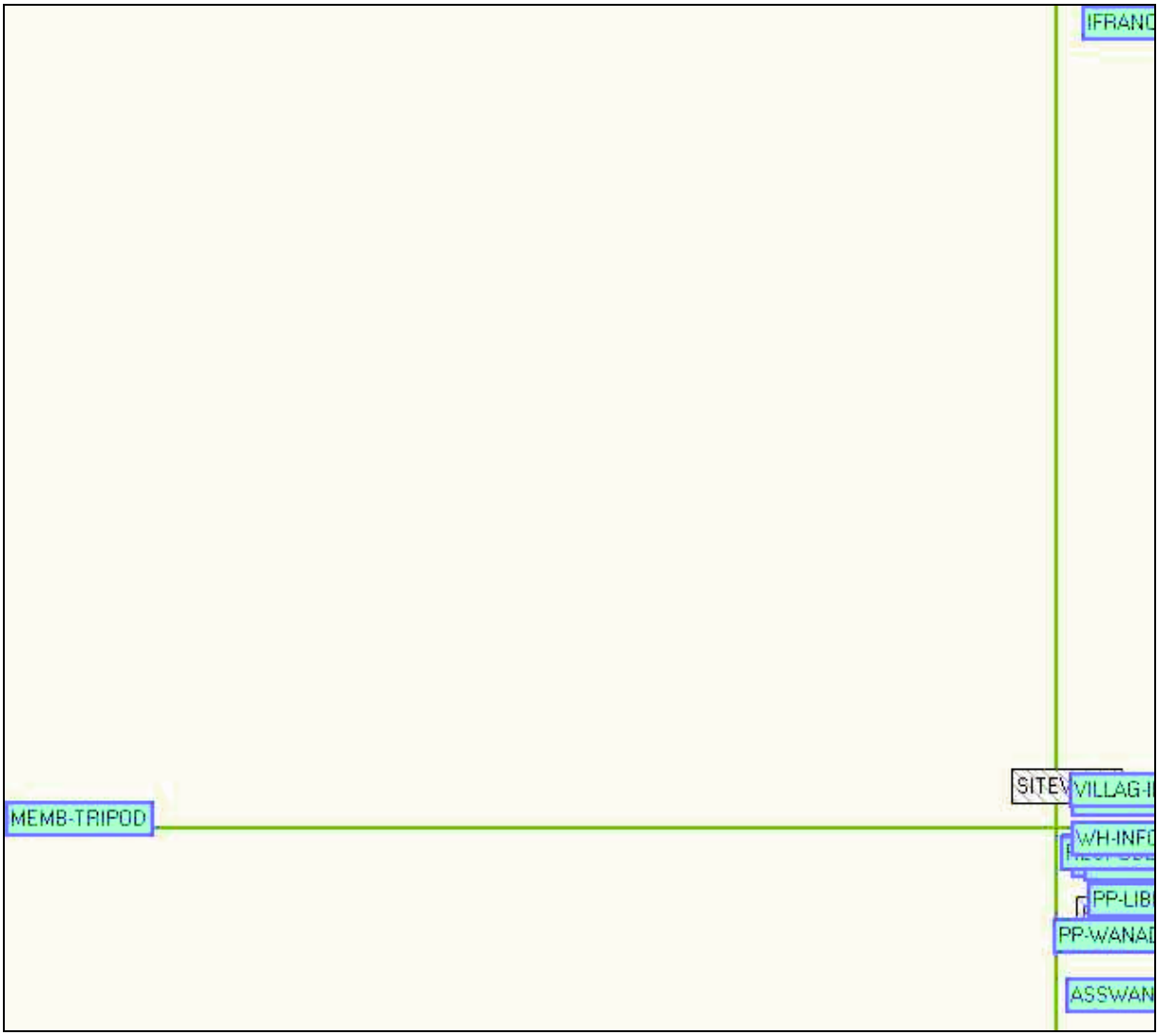



TRAIT10\b+
TRAIT50\b+
Fréquences relatives (x 10 000)
9 500
9 000
8 500
8 000
7 500
7 000
6 500
6 000
5 500
5 000
4 500
4 000
3 500
3 000
2 500
2 000
1 500
1 000
500
0
ASSWANADOO
FREE
HOMENORDNET
MEMB-TRIPOD
PP-INFONIE
RESPUBLICA
VVH-INFONIE

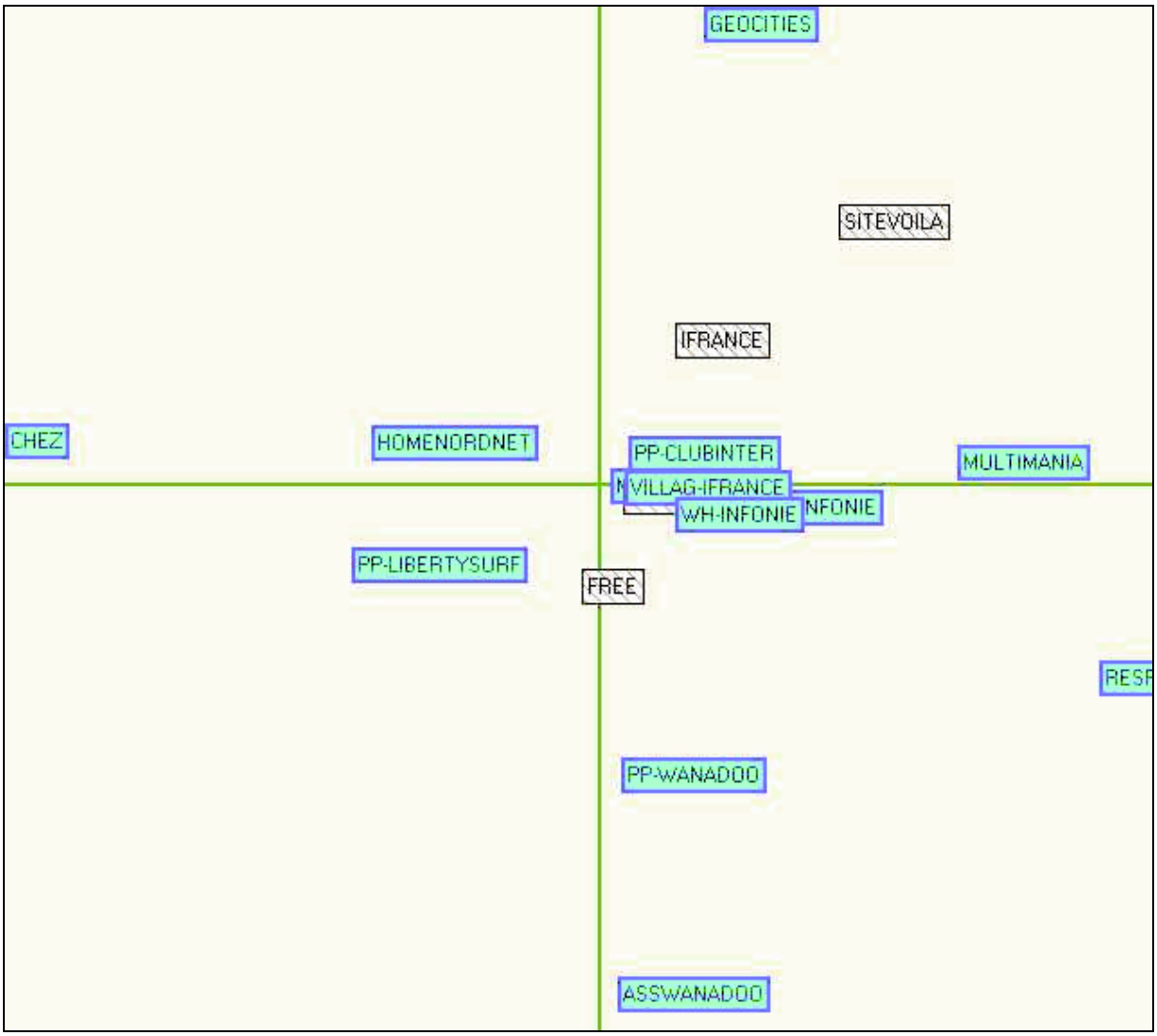
4.1.1.3.3.5 AFC

4.1.1.3.3.5.1 N° 1

4.1.1.3.3.5.2 N° 2


IFRANC
SITEV
VILLAG-I
MEMB-TRIPOD
WH-INFO
PP-LIB
PP-WANAD
ASSWAN

4.1.1.3.3.5.3 N° 3


GEOCITIES
SITEVOILA
IFRANCE
CHEZ
HOMENORDNET
PP-CLUBINTER
MULTIMANIA
VILLAG-IFRANCE
WH-INFONIE
NFONIE
PP-LIBERTYSURF
FREE
RESF
PP-WANADOO
ASSWANADOO

4.1.2 Chaîne Lynx

4.1.2.1 Corpus textuel

A définir

4.1.2.1.1 Traitements avec Lexico

Analyse avec Lexico : voir ce qu'il est possible de faire

4.1.2.1.1.1 Etat du corpus lu sous Lexico

Corp100000LynxLexClean2

```
36468265 36468265 883441 23142754 732857 732259 393167 5000000 18 3 150564
0 0

*** Résultat de la segmentation du fichier: Corp100000LynxLexClean2.TXT ***
Délimiteurs .,:;!?/_-\"'()[]{}§$

      nombre des occurrences :    23142754
      nombre des formes      :      732857
      frequence maximale     :      732259
      nombre des hapax       :      393167
      nombre des clés(type)  :           3
      nombre des clés(ctnu)  :      150564

*** Fin de la segmentation du fichier:      Corp100000LynxLexClean2.TXT ***
```

Principales caractéristiques de la partition : HEBERGEUR

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
ASSOCWANADOO	71361	12385	6831	3155	de
CELEBRITIES	45	41	37	2	et
CHEZ	3039294	167580	81706	90783	de
CLUBINTERNET	1363382	105443	55120	54108	de
FREE	3130767	164423	86793	86124	de
GEOCITIES	2014537	160540	87570	35136	de
IFRANCE	993558	78910	31521	34598	de
LEVILLAGEIFRANCE	332517	31139	11425	12354	de
LEVILLAGEORG	745	445	350	25	de
LIBERTYSURF	235113	41341	26961	4457	de
MEMBERSAOL	888380	81489	41446	24303	de
MEMBRESTRIPOD	407388	76087	52837	10843	de
MULTIMANIACOM	5291130	279735	148090	180547	de
MULTIMANIAFR	443	276	212	17	et
NORDNET	340866	40707	21821	10768	de
NOUNOUVEVETTE	16	16	16	1	les
PERSOINFONIE	33050	7629	4505	963	de
PERSORESPUBLICLA	918	622	554	19	a
PERSOWANADOO	4832510	225223	109914	177787	de
RESPUBLICA	7767	228	31	675	root
SITEVOILA	95709	13485	7426	3892	de
WEBHOMEINFONIE	63258	15220	9521	2398	de

Vocabulaire fréquent (extrait)

Formes (ordre lexicométrique)	Fréquence
de	732259
a	422020
la	356170
et	304657
le	283397
les	227077
l	220802
des	219700
d	205388
en	187026
un	173258
est	172049
1	159120
du	150573
une	125817
*	120317
pour	117084
sur	113920
que	109088
dans	97508
2	91507
qui	91317
vous	90384
par	84635
pas	82647
s	82524
au	78936
il	76421
ou	73836
the	72271
n	71956
Le	70566
plus	63821
NomDeFichier	63524
ne	62803
3	60878
avec	60747
ce	60569
Les	58888

Formes (ordre lexicométrique)	Fréquence
Les	58888
La	57050
France	50148
0	49342
qu	49023
on	48947
se	48897
sont	46003
4	43743
A	43362
L	42877
site	42841
of	41238
to	41184
5	39866
I	39780
je	39004
page	38803
and	38775
Marie	38509
Jean	35637
Il	35525
son	35454
c	34802
+	34420
nous	32256
mais	31988
x	31586
Naissance	31262
C	31181
y	30320
aux	29636
10	28898
tout	28549
etre	27881
votre	27512
6	27441
2000	26763
cette	26502

Formes (ordre lexicométrique)	Fréquence
cette	26502
D	26326
me	25650
root	25538
in	25508
12	25129
Date	24777
bien	24433
octets	24209
si	24196
Sexe	24185
Lieu	24176
comme	24018
meme	23801
faire	23531
elle	23433
fait	23257
Je	23194
Retour	23058
Pour	22985
you	22843
m	22716
11	22554
8	22347
Deces	22198
The	22138
7	21625
for	21619
sa	21563
tres	21343
peut	21086
01	20519
Pierre	20457
Un	19361
lui	19232
J	19169
j	19148
ont	19032
ses	18919

Formes (ordre lexicométrique)	Fréquence
ses	18919
30	18840
F	18782
DE	18642
ai	18378
deux	18278
t	18227
N	18196
20	18143
tous	17988
leur	17965
En	17858
Vous	17670
UrL	17500
aussi	17261
is	17099
02	16920
Saint	16707
9	16643
15	16572
autres	16569
sans	16519
vers	16508
13	16380
apres	16173
ete	16111
ces	16065
1x	15870
21	15650
avant	15419
18	15355
16	15353
mon	15309
principale	15288
Si	15285
Page	15112
ils	15082
E	15057
here	15039

4.1.2.2 Matrice de mots

La chaîne Lynx a produit un fichier qui doit être utilisé pour générer une/des matrice(s) (de mots) : cf travail fait sur corpus 15000