ED Cultures et sociétés UPEC

Approches numériques des données textuelles.

- http//:textopol.org
- P. Fiala, J.-M. Leblanc.
- Traitements textométriques des discours sociopolitiques.
- CEDITEC: Discours, information, communication.

Textométrie en questions
Perspective typologique et prospective.

- **Présentation d'outils d'analyse textuelle standard sous forme interactive, comparative et appliquée.**
- **Discussion des possibilités offertes par ces instruments dans les recherches en SHS (histoire, socio, AD, humanités) ou dans les formations professionnalisantes (SIC, professorat, traduction).**
- **Présentation de résultats récents en textométrie et des projets en cours de l'équipe Textopol, Textobserveur, site de recherche, formations.**

Démarche :
A quoi ça sert et comment ça marche?

- De la lecture de résultats à la maîtrise et à la comparaison des outils.
  - Résultats de divers traitements numériques de données textuelles.
  - Questions auxquelles répondent ces résultats.
  - Questions soulevées par ces traitements.
  - Repères pour les approches quantitatives (lexicométrie, 1960-1990; textométrie, 1990-2011) et qualitatives en analyse du discours .

- Prise en main rapide d'outils et de corpus test

Les décomptes fréquentiels médiatiques, version réduite de l'analyse des discours politiques.

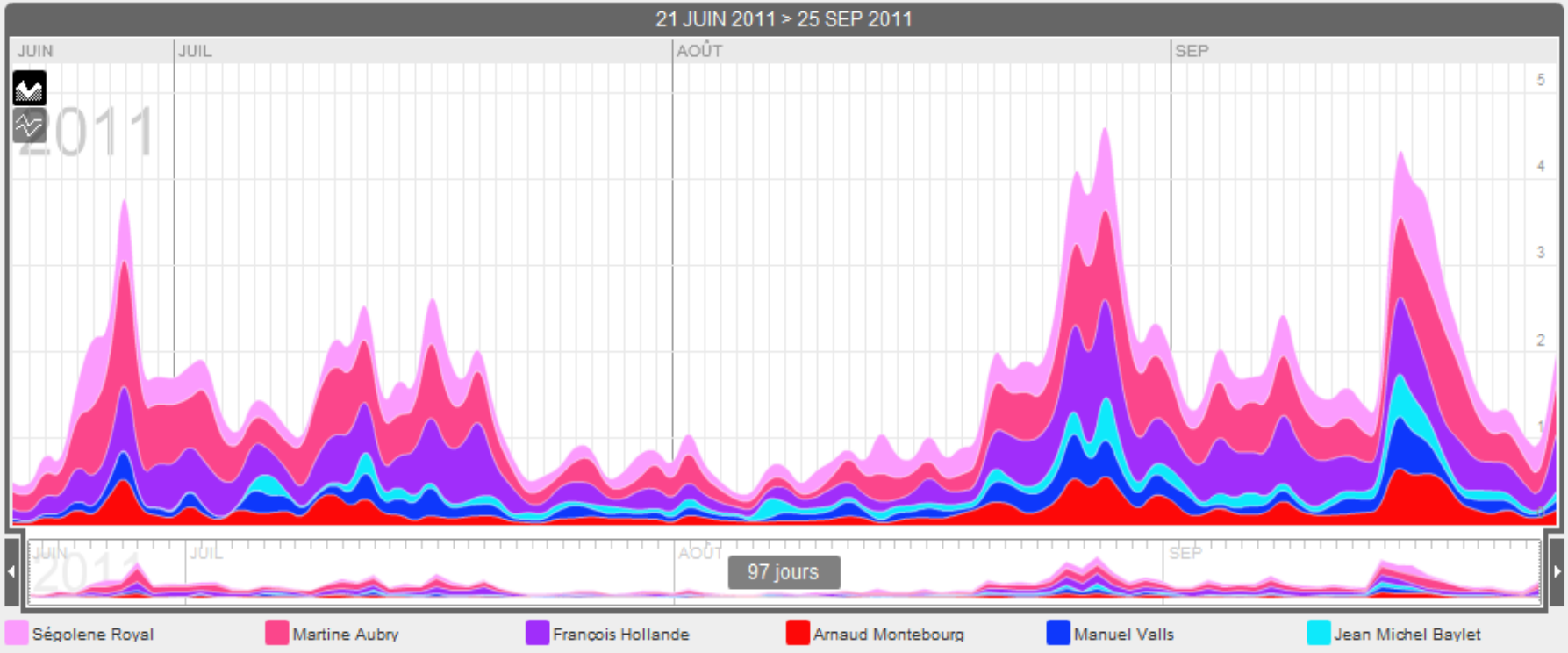
- **Développement rapide depuis 2005 dans le commentaire politique et sur le WEB 2.**

- **Standardisation des objets et des méthodes**
  - **Blogométrie et buzz.**
  - **Mesures des « querelles sémantiques ».**

- **Risques :**
  - **Données partielles ou mal connues, peu documentées.**
  - **Interprétation sauvage des listes.**
  - **Généralisations abusives.**
  - **Instrumentalisation par les officines (prétendus Instituts) d'opinion publique.**

Des nuages sur le WEB: Buzz à la une.
La Primaire du PS depuis le 21 juin
http://fr.linkfluence.net/

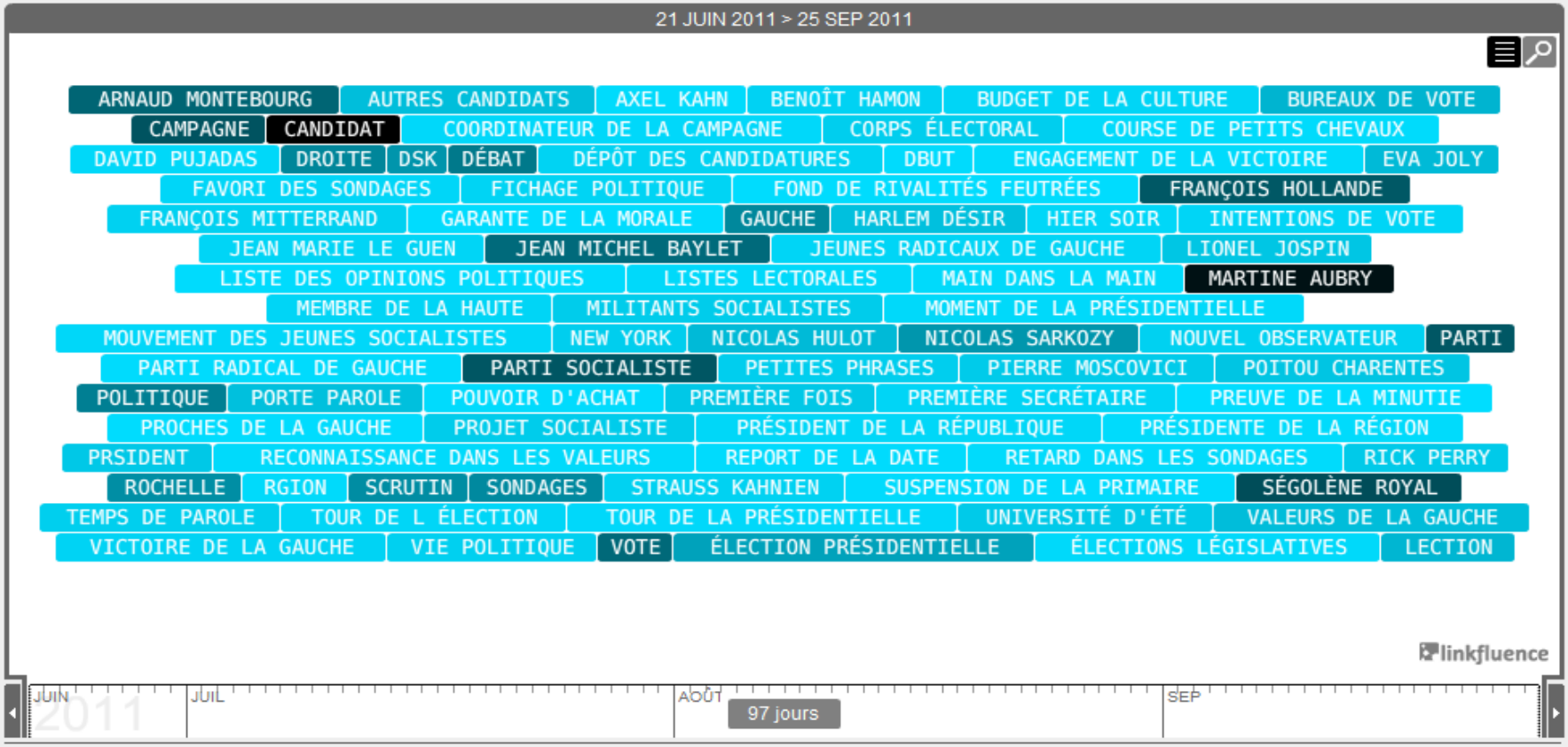
La présence des candidats dans les conversations sur le Net



La Primaire du PS depuis le 21 juin

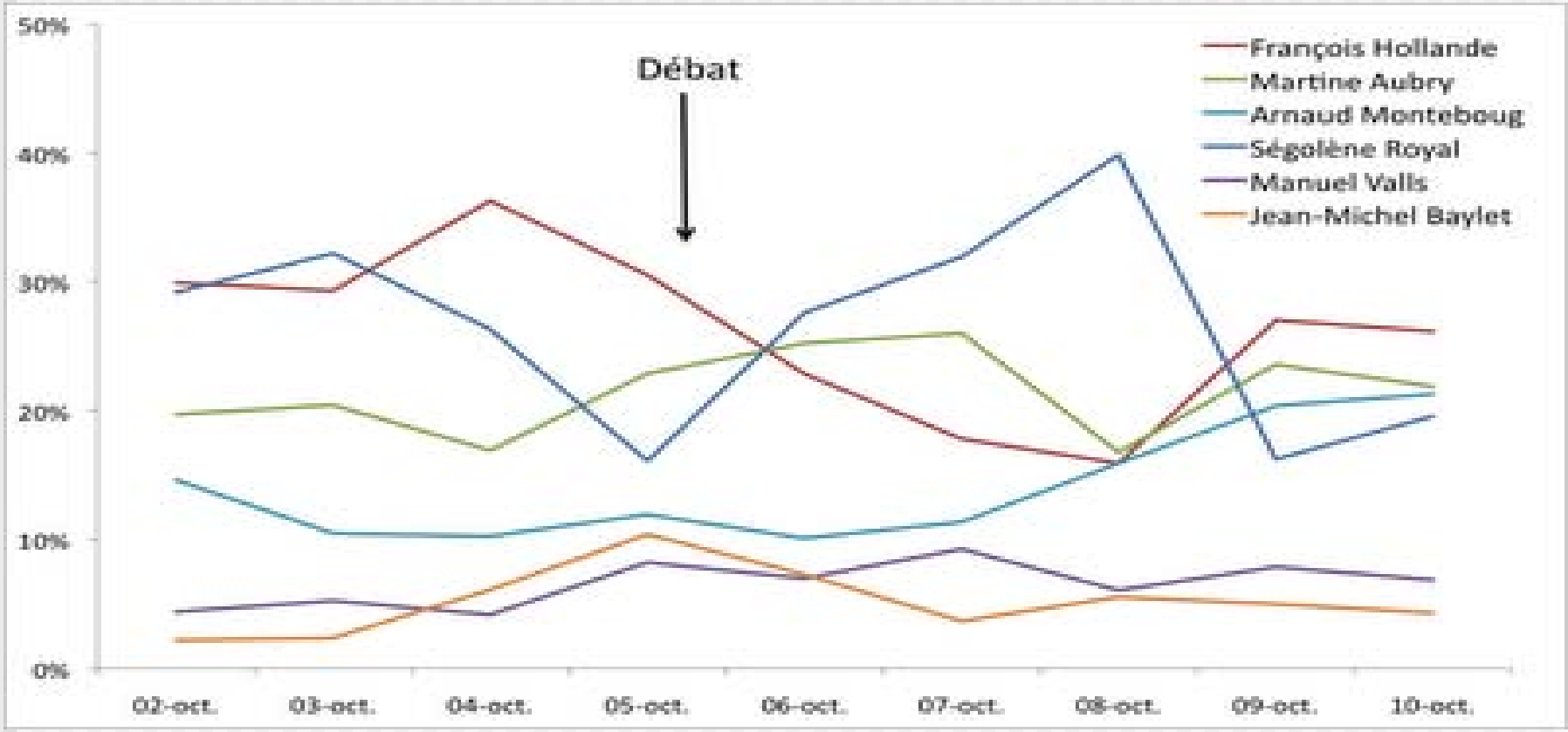
http://fr.linkfluence.net/

Les mots associés à la primaire



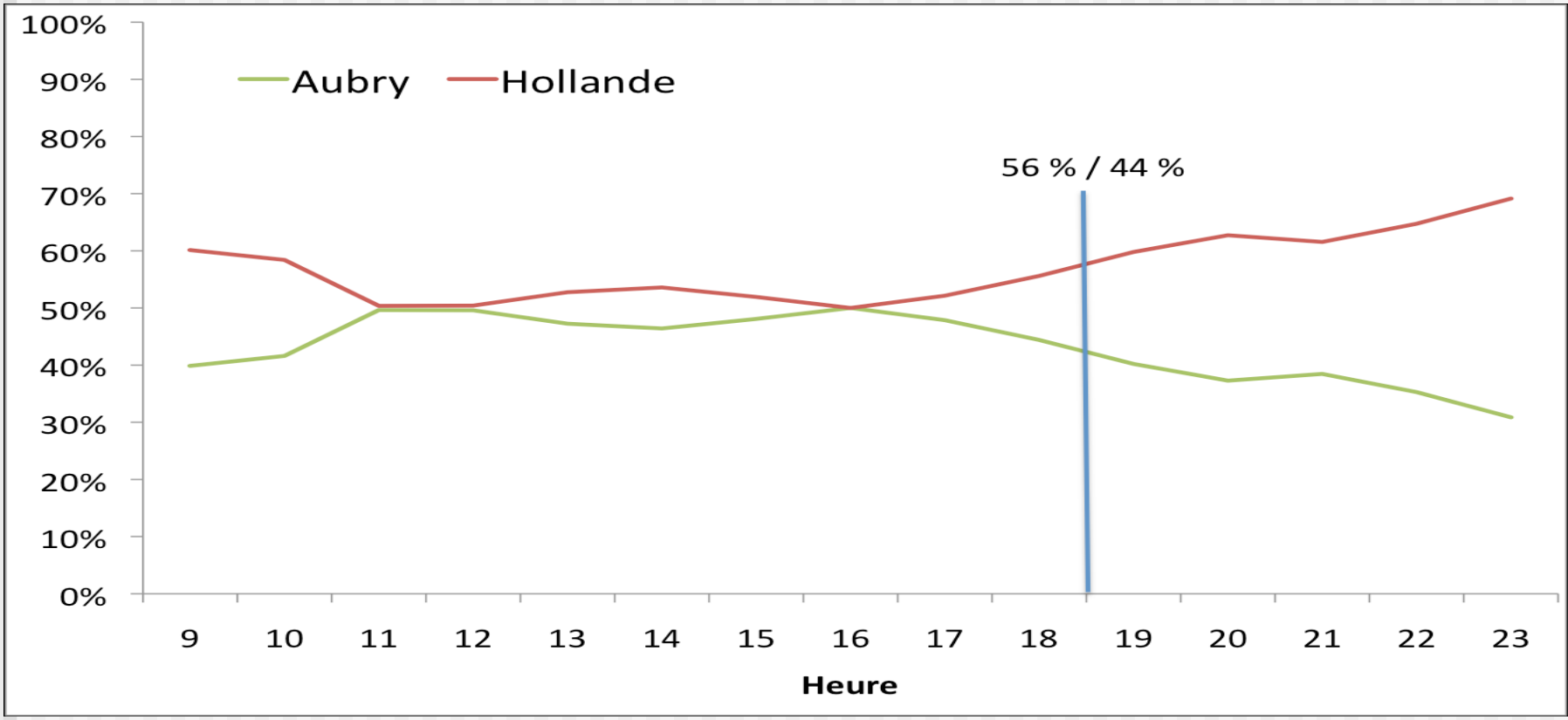
La Primaire du PS
L'outil Buzz

http://blog.veronis.fr/2011/10/primaire-le-buzz-de-la-journee-de-vote.html



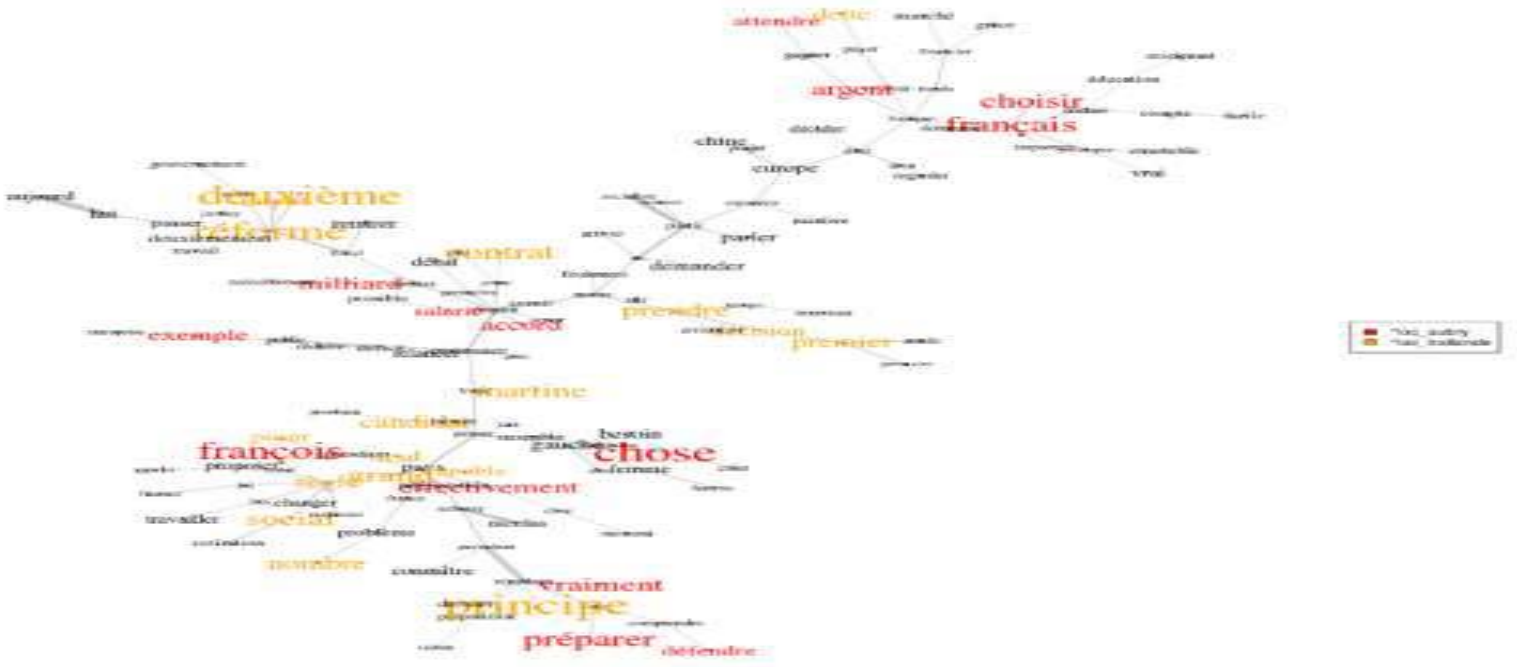
La Primaire du PS
Buzz du deuxième tour

http://blog.veronis.fr/2011/10/primaire-le-buzz-de-la-journee-de-vote.html



La Primaire du PS
Les mots des candidats

http://pascal-marchand.fr/spip.php?article50



Aborder la mesure avec précaution dans la recherche SHS

- **Garder ses distances face à la standardisation simplificatrice des objets et des méthodes.**
- **Maîtriser et comparer les outils; développer de maquettes et des procédures autonomes originales.**
- **Partager et transmettre les acquis théoriques, méthodologiques, les données sur des sites autonomes. Exemples:**
  - Texto: revue.texto.net ; Veronis: http://blog.veronis.fr/
- **Suivre les débats critiques.**
  Exemple: *Les mots sont importants*
  - *contact@lmsi.net*

Questions auxquelles un traitement numérique apporte des réponses documentées

- Dans **texte**, chez un auteur, une source:
  - Etudier les attestations, la forme, la distribution d' un mot graphique, d'une liste de mots.
  - Etablir et classer les contextes d'un mot dans un corpus: **concordances, KWIC.**
  - Etablir le vocabulaire, le lexique, l'index exhaustif ou non, avec un classement lexicographique ou fréquentiel : **Tableau lexical.**
  - Etablir la phaséologie (mots composés, idiotismes) classés alphabétiquement, par longueur ou fréquence: **Segments répétés (N-gram), inventaires distributionnels.**

Questions auxquelles un traitement plus riche apporte des décomptes documentés

- Dans un texte ou un ensemble de textes
  - Etiqueter, lemmatiser, classer et compter les mots par catégories grammaticales (?)
  - Identifier des thématiques d'un texte (?!)
  - Résumer automatiquement un texte (?!!)
  - Comparer deux textes (traductions) lignes à lignes: Alignement
  - Evaluer sur le plan lexical, stylistique, sémantique, les propriétés d'un texte par comparaison à un texte de référence(?!!)

Questions auxquelles un traitement statistique évolué apporte des constats probabilistes

- Dans un **corpus** divisé en parties (chronologique, par sources, auteurs, genres, chapitres, etc.)
- Mesurer et comparer la distribution d'un mot, d'un ensemble de mots, **sa spécificité** chronologique, par auteur, etc.
- Etablir le vocabulaire **représentatif**, spécifique de chaque partie.
- Etablir la liste des **énoncés** les plus représentatifs statistiquement.

Autres questions auxquelles un traitement statistique approfondi apporte des résultats

- **Dans un corpus divisé en parties (chronologique, par sources, auteurs, genres, chapitres, etc.) mesurer et comparer:**
  - **La richesse du vocabulaire**
  - **L'accroissement lexical**
  - **L'évolution du vocabulaire**
  - **La répartition statistique et linguistique des fréquences, THF, HF, MF, BF, Hapax.**
  - **Les associations lexicales : les cooccurrences**

Autres questions auxquelles un traitement statistique probabiliste apporte des résultats

- **Dans un corpus divisé en parties (chronologique, par sources, auteurs, genres, chapitres, etc.) , mesurer et comparer:**
  - **La proximité statistique des parties: AFC**
  - **Les corrélations entre distribution**
  - **L'évolution du vocabulaire**
  - **Les associations lexicales : les cooccurrences**
  - **Voisinages et identification des textes**

Des disciplines et des domaines proches et interdépendants

- **Lexicologie et lexicographie : de l'analyse systématique du vocabulaire à la mise en forme des dictionnaires : l'exemple du TLF.**
- **Linguistique de corpus, TAL (traitement automatique des langues), et ADT (analyse des données textuelles).**
- **AD (analyse du/des discours), TAD-TND (traitement automatique, numérisé des discours).**
- **Lexicométrie-stylométrie-textométrie-logométrie?**
- **Philologie numérique; qualité des étiquetages, des filtres**

Coup d'œil bibliographique

- Une ressource : textopol.free.fr/Documents/Bibliographie/

- **Quelques références récentes**
  - Lemercier & Zalc: *Méthodes quantitatives en Histoire*, La découverte, 2008.
  - Dictionnaires méthodologiques et conceptuels
    - Charaudeau-Maingueneau,
    - Siblot (et autres),
  - Revues:*Langages, Mots, Semen, Lexicometrica*
  - Une collection en cours: Champion

Jalons: Fondements philologiques de la statistique lexicale

- La création manuelle d'index (listes des mots référencés d'une œuvre, d'un corpus) et de concordances (contextes des mots) a des origines grammaticale, religieuse, philosophique, littéraire.
- Objectifs: attestation exhaustive, analyse de la forme, et sens « vrai » des mots dans leur contexte.
- Premières concordances: Homère (3ème siècle ACN); Bible (13ème), Dante, Shakespeare, Leibnitz, Descartes.
- L'index et la concordance, tâches simples sur le plan informatique; essentielles dans tout programme d'analyse lexicale ou textuelle.

Jalons: Les « banques » de données textuelles: du tas instable ou indifférencié…

- **Web**

  Données indifférenciées, non documentées, non structurées, accès libre, moteurs multiples .

- **Google labs**

  550 MM (anglais)

  45 MM (français: 16e-21siècles)

  Données indifférenciées, non documentées, non structurées, accès libre, moteur faibles .

… à des organisations (hyper)structurées

- **Gallica (BNF)**
  - **Données historiques documentées, accès limité.**
- **Athena, ABU**
  - **Données philosophiques et littéraires, libres, faiblement structurées**
- **Frantext: 90M**
  - **Données philosophiques et littéraires, documentées et structurées, accès, payant.**
- **Prospéro**
  - **Structuration évolutive, balisage permanent.**
- **Brown, BNC,ANC: Tradition anglo-saxonne, échantillonage, représentativité sociolinguistique**

Bases de données et moteurs: principes initiaux de la statistique lexicale.

- **Bases numériques de données textuelles et moteurs de recherche : du tas de mots (Google) aux banques structurées (dates, auteurs, genres, titres, etc.)**
- **L'index et la concordance, essentiels dans tout programme d'analyse textuelle, tâches simples sur le plan informatique, permettent une présentation sous formes de listes ou de tableaux du stock lexical et de ses propriétés distributionnelles.**

Quelques fondement de l'analyse des discours sociopolitiques (1965-1985)

- **Fondements et méthodes multiples : histoire culturelle, philosophique, sciences politiques, psychosociologie, anthropologie, sociologie, analyse du discours**
- **Objectiver et finaliser les savoirs : politique, économique, social, et discursif.**
- **Ancrages lexicologiques, énonciatifs, textuels**
- **Principes rhétoriques, techniques argumentatives pour la persuasion et la réfutation.**
- **Importance des types de situation et des genres de discours**

Tradition française d'analyse du discours sociopolitique (ADSP), Déconstruction critique

- **L'approche lexicologique et discursive française : les moments révolutionnaires et les usages langagiers 1789-1794, 1848, 1870, 1920, 1968. Une tradition critique, « de gauche ».**
- ***Un site actuel: Les mots sont importants: contact@lmsi.net***

- **La tradition lexicographique et conceptuelle allemande: Les *Thesaurus* philologiques;**

  Brunner, Gonze, Koselleck : *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland, 1972- 1997*

Jalon: trois courants herméneutiques en ADSP ancrés dans les sciences du langage(1960-2000)

- De la sémiotique (Greimas) à la pragmatique argumentative.
- De la syntaxe harrissienne (M. Pêcheux: AAD) aux approches énonciatives
- Lexicométrie: de la statistique lexicale aux typologies discursives

Une première étape: un laboratoire de lexicométrie (Saint Cloud, 1965-2000) La statistique lexicale au pied nu

- **Le corpus syndical; des tracts en mai**
- **Le dictionnaire des fréquences syndicales**
- **De *travailleur* à *salarié***

La réduction lexicométrique (Maurice Tournier 1980)

1. De la curiosité à la question de recherche
2. De la question de recherche aux hypothèses de recherche
3. Des hypothèses au recueil de données
4. Des données à la constitution du (des) corpus
5. Le choix des traitements informatisés adaptés aux hypothèses
6. Le formatage du corpus lexicométrique
7. Le traitement automatisé

Approches quantitatives des discours sociopolitiques

1. **La lexicométrie: une méthodologie probabiliste appliquée à des grands corpus textuels (1965-2006): programmes politiques, résolutions de congrès, textes littéraires.**
2. **Chaînes de logiciels de documentation, de tri et de comparaison de vastes données alpha-numériques: textes et décomptes statistiques.**
3. **Méthode de traitement et d'analyse automatisée des discours de masse**

Principes et méthodologie de base

- **Des hypothèses sociopolitiques historiques, argumentatives, stylistiques, énonciatives, sémantiques, portant sur les genres, les scénographies, les ethos, les positions idéologiques, politiques, etc.**
- **Un recueil échantillonné de données discursives calibrées**
- **Un corpus multipartitionné**

Procédures standard : documentaires, statistiques, probabilistes

- Segmentation, indexation, distribution des formes
- Concordances, segments répétés, inventaires distributionnels
- Calcul et comparaison de la richesse lexicale des parties
- Calcul des proximités entre les parties, par AFC, ou classification hiérarchique
- Calcul des surreprésentations et sous-représentations lexicales : spécificités lexicales
- Recherche des cooccurrences lexicales

Norme endogène d'interprétation et procédure de vérification

- **Vérification des hypothèses par interprétations des comparaisons internes et retour aux données textuelles**
- **Vérification expérimentale des mesures par modification des seuils et établissement de corrélation**
- **Montage de contre-épreuves**
- **Mise en visibilité : établissement de topographies et tableaux standard**

Dispositifs, modularité et développements lexicométriques.

- Des bases de textes accessibles
- Des outils informatisés robustes interfacés, articulés entre eux
- Une démarche cumulative
- Une démarche comparative
- Le retour au texte par l'hypertextualité généralisée
- Etiquetage morphosyntaxique, sémantique des corpus

TEXTOPOL Espace en ligne de Formation et recherche Analyse linguistique et quantitative des corpus sociopolitiques

- **Un espace didactique de formation à la recherche**
- **Une base de textes sociopolitiques accessibles**
- **Des outils informatisés robustes interfacés, articulés entre eux**
- **Des outils expérimentaux**
- **Forum, archive**

Un site d’expérimentation et de formation
http//:textopol.free.fr
Une démarche empirique

- Appliquée
- Cumulative
- Opératoire
- Comparative

Six outils robustes quantitatifs évolués en accès facile

- Lexico3
- Hyperbase
- Weblex (TXM)
- Cordial
- Tropes
- Alceste
- Sphinx
- Développement de Textobserveur

Outils de l'expérimentation textométrique

- **Des instruments lexicométriques de base:**
  - Lexico 3 (André Salem, Paris-3).
  - Hyperbase (Etienne Brunet, Nice)

- **Accès à d'autres outils lexicométriques:**
  -Alceste: analyseur sémantique par cooccurrence;
  -Weblex (TXM): constructeur de réseaux lexico-sémantique.

- **Bases de textes et de corpus**

Comparer les outils.

- **Lexico3: une expérimentation souple**
  - Une seule base, plusieurs balises
  - Expérimentation locale rapide
  - Démarche endogène
  - Topographie des textes
- **Hyperbase: les gros corpus stables**
  - Une base par type de balise
  - Expérimentation globale
  - Démarche exogène possible
  - Mesure de la proximité intertextuelle
  - Référenciation aisée, interface développé.
- **Des fonctions analogues**
  - Fonctions documentaires similaires
  - Fonctions hypertextuelles analogues
  - Modes de calcul relativement comparables

Types d'outils de l'analyse textuelle

- **Des instruments lexicométriques de base:**
  - Hyperbase, (Etienne Brunet, Nice);
  - Lexico 3, (André Salem, Paris-3).
  - Weblex: constructeur de réseaux lexico-sémantique développe en plate-forme TXM (Heiden)

- **Outils linguistico-documentaires:**
  -Cordial: étiqueteur morphosyntaxique;
  -Tree-Tagger: analyseur syntaxique
  -Alceste: analyseur sémantique par cooccurrence;
  -Tropes : analyseur grammatical du contenu des énoncés
  -Sphinx: base d'analyse de contenu

Autres outils lexicométriques

- **Weblex: cooccurrenceur récursif**
- **Cordial: étiqueteur statistique**
- **Alceste: cooccurrenceur sémantique**
- **Métromètre: versificateur**

Des outils linguistiques et statistiques évolués pour des recherches sémantiques : Tropes et Alceste

Tropes : fondement linguistique, avec étiquetage morphologique et ontologie sémantique adaptable

Alceste: fondement statistique et linguistique distributionnel,

Cordial, outil robuste propriétaire à base normative.

www.synapse-fr.com

- Correcteur orthographique évolué, interfacé,
- Etiqueteur morphologique (taux de résussite: 90%)
- Analyseur syntaxique
- Etiqueteur sémantique (ontologique)
- Décompteur
- Comparateur (corpus de comparaison par genre)
- Evaluateur, résumeur, diagnostics stylistique et sémantique.

Tropes, outil linguistique d'analyse de contenu

- Aspirateur de sites
- Etiqueteur morphologique
- Analyseur syntaxique
- Etiqueteur sémantique (ontologique)
- Décompteur

Alceste, outil statistique de construction de classes sémantiques sur une base cooccurrentielle

- **Raciniseur morphologique**
- **Construction et optimisation d'unité de contexte cooccurrencielle**
- **Etablissement de classes d'énoncés cooccurrenciels par classement hiérarchique.**
- **Outils textométriques de tri, de navigation, et de visualisation.**

Un protocole d'observation sur des corpus partitionnés.

Distributions statistiques (Tableau Lexical Complet)

- 1. Propriétés zipféennes du TLC
- 2. Sur le vocabulaire : Index global et par partie
- 3. Proximités des parties du corpus
- 4. Les cooccurrences et lexicogrammes

Distributions linguistiques (Tableau Lexical Réduit)

- 1. Les outils documentaires
- 2. Catégories énonciatives
- 3. Catégories grammaticales

Distributions statistiques (Tableau Lexical Complet)

- 1. Propriétés zipféennes du TLC
- Les relations rangs-fréquences
- Les relations formes-occurrences : taille et vocabulaire.
- Les hapax
- Les hautes fréquences
- Les moyennes fréquences
- La richesse de vocabulaire
- 2. Sur le vocabulaire : Index global et par partie
- Voc partagé
- Voc original
- Voc spécifique
- Voc banal
- 3. Proximités des parties du corpus
- Analyse Factorielle des Correspondances
- Classification Hiérarchiques Ascendantes
- 4. Les cooccurrences et Les lexicogrammes: attirances probabilisées du vocabulaires

Distributions linguistiques (Tableau Lexical Réduit)

1. Les outils documentaires :
index, concordance, segments répétés, listes de mots

2. Catégories énonciatives
- Pronoms et déterminants du discours
- Modalités : assertives, interrogatives, exclamatives, jussives
- Négations
- Les auxiliaires modaux : Vouloir, pouvoir, devoir, falloir, avoir à
- Les Déterminants

3. Catégories grammaticales
- Noms propres
- Substantifs et adjectifs
- Verbes
- Adverbes et connecteurs
- Interjections, phatiques

Les vœux politiques des présidents de la 5e République (PF-JML, TEXTOPOL, 2003)

- **Un genre de discours calibré: un rituel discursif**
- **Présidents de droite, président de gauche**
- **Lexico3, hyperbase, Alceste**

La gauche plurielle (PF-JML, CEDITEC 2000)

- 6 semestres du gouvernements Jospin
- Des genres de discours distincts
- Ministre politiques et ministres techniques
- Des Femmes et des Hommes

Premiers ministres sous la 5e République (1981-2005) (AFC, Marchand, 2007)

- Un circuit politico-thématique sinueux
- Un axe temporel

Dirigeants de gauche et la droite sous la 3e République (AFC, Damon Mayaffre)

- Deux gauches
- Deux droites

La parole des présidents de la 5e République (Mayaffre, 2005)

- Démarche extensive
- Présidents de droite, président de gauche
- Chirac et jospin
- Hyperbase

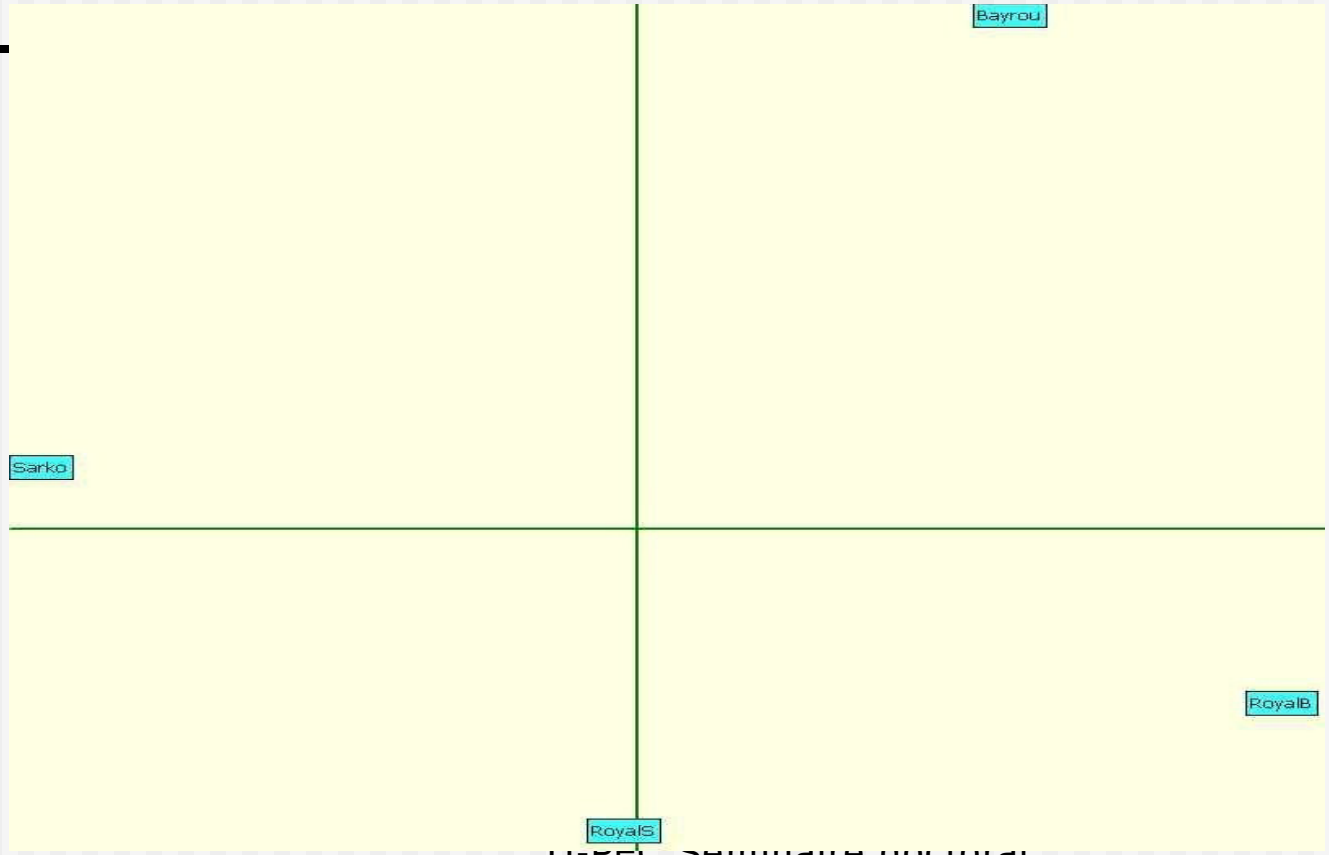
Ségolène en campagne 2007 Une application lexicométrique prototypique (M. Tournier, 2007)

- **La valse à quatre: Royal-Bayrou; Royal-Sarkozy**

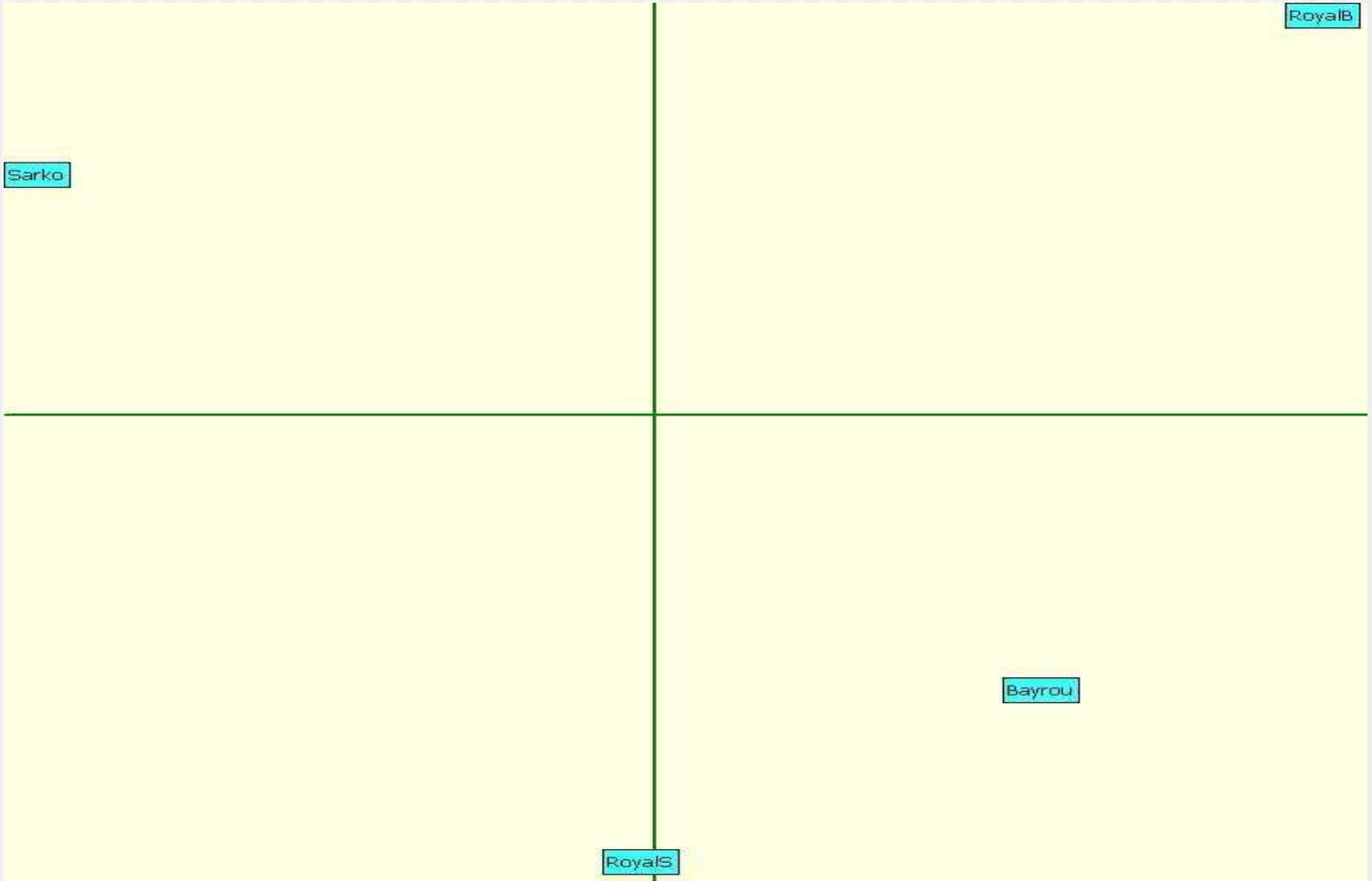
- L'ensemble du corpus comprend (journalistes exclus) 40455 occurrences et 4356 formes textuelles, réparties de la façon suivante :

Partie	*Nb occurrences*	*Nb formes*	*Nb hapax*
Bayrou	7462	1425	791
RoyalB	9316	1746	976
RoyalS	11427	2029	1070
Sarko	12250	2133	1147

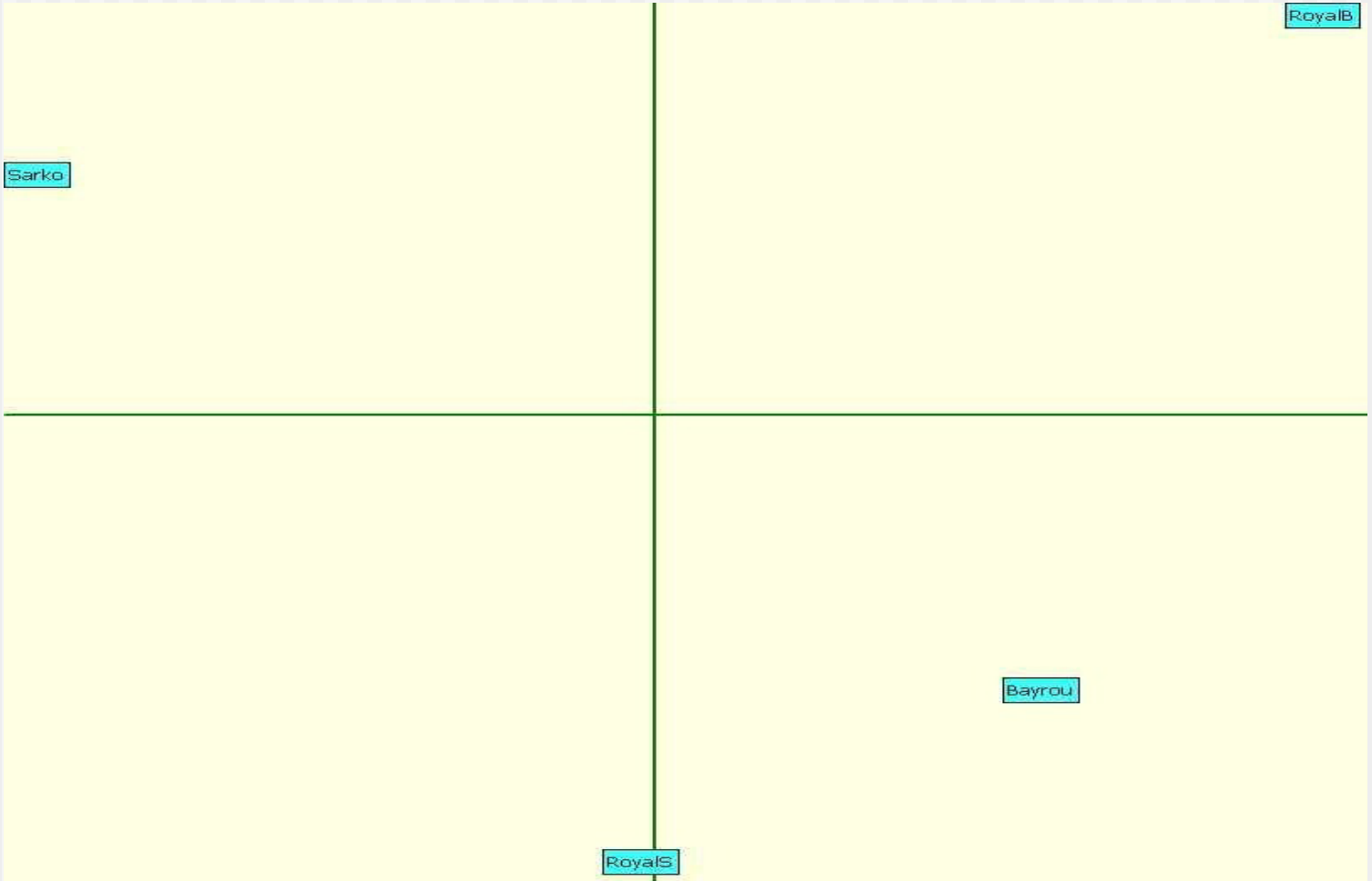
Ségolène en campagne 2007 (M. Tournier, 2007)



Ségolène en campagne 2007 (M. Tournier, 2007) AFC Axe 1 et 3

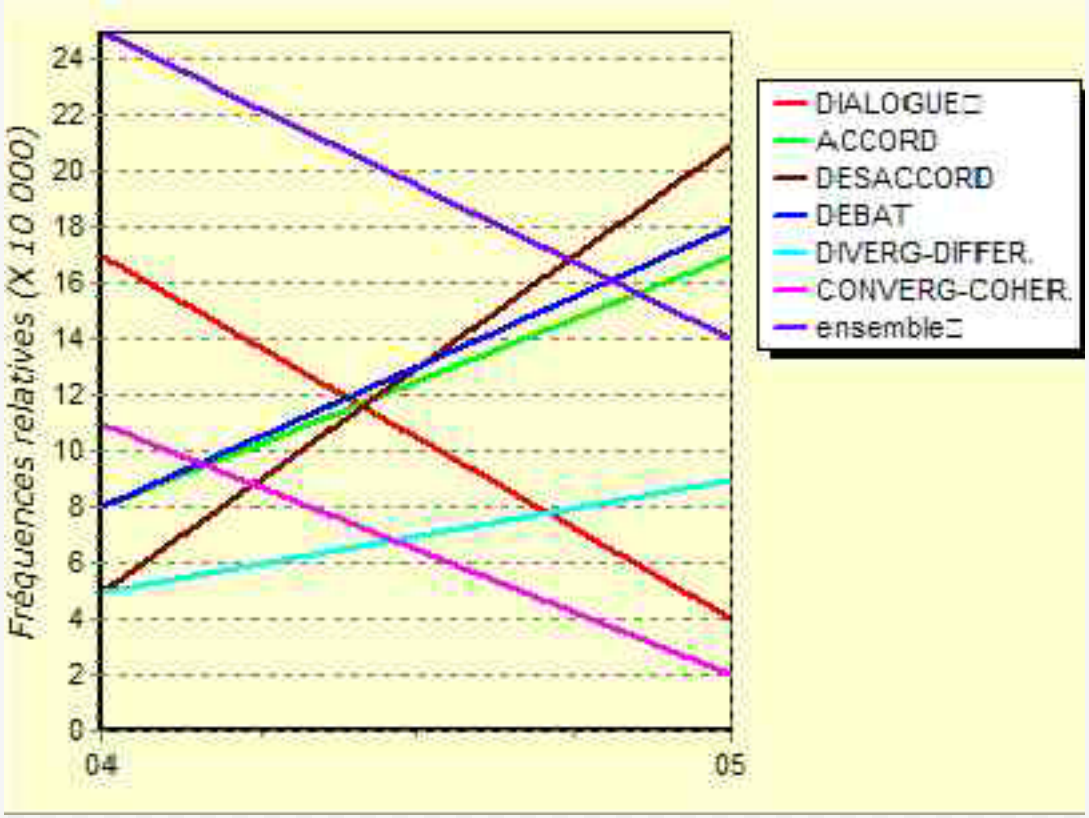


Ségolène en campagne 2007 (M. Tournier, 2007) AFC Axe 1 et 3



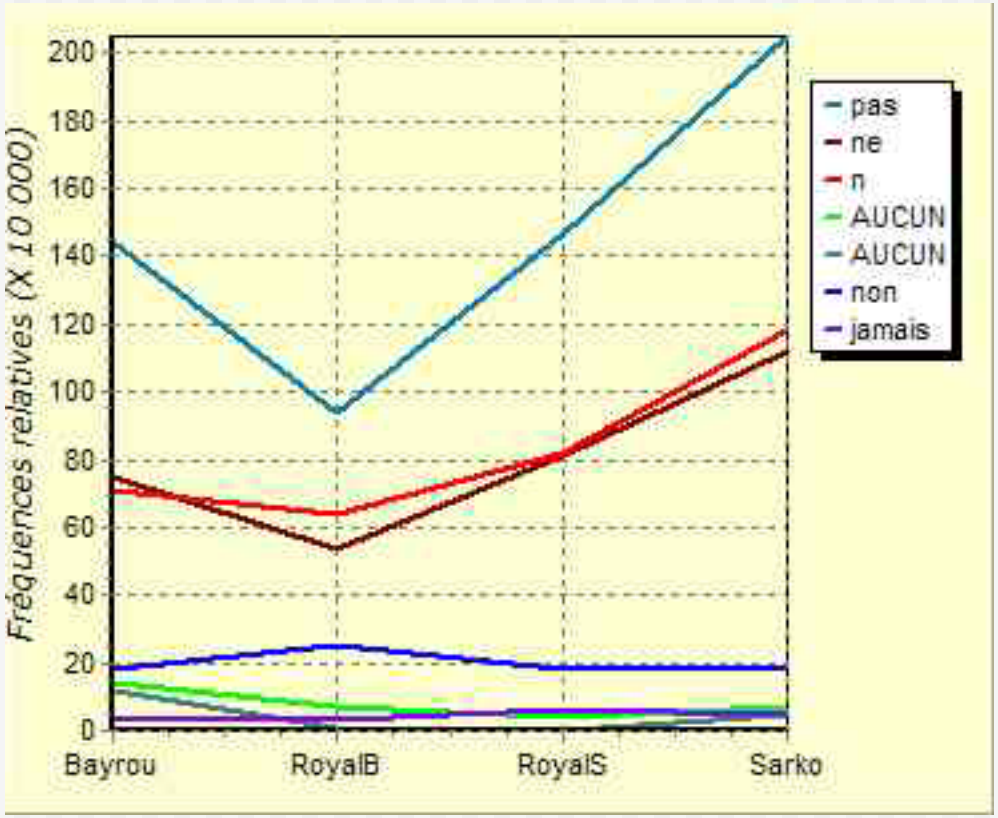
Ségolène en campagne 2007 (M. Tournier, 2007) AFC Axe 1 et 3

Ségolène en campagne 2007 (M. Tournier, 2007)
Duo Royal-Bayrou

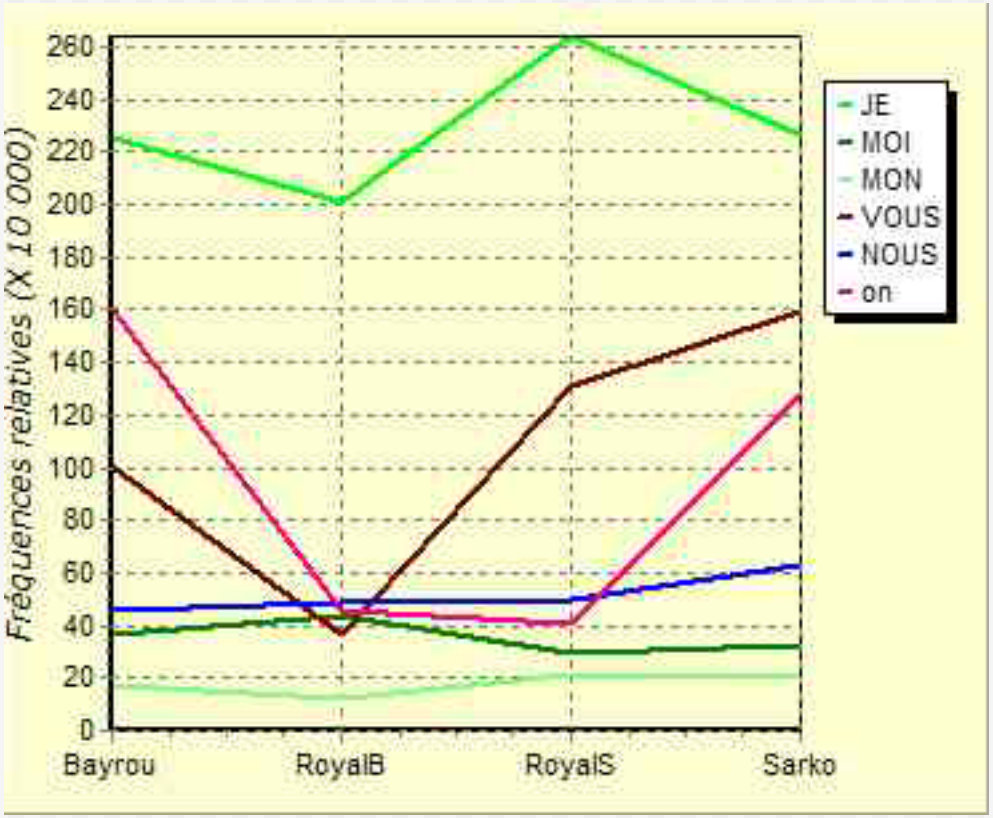


Ségolène en campagne 2007 (M. Tournier, 2007)
Les négations

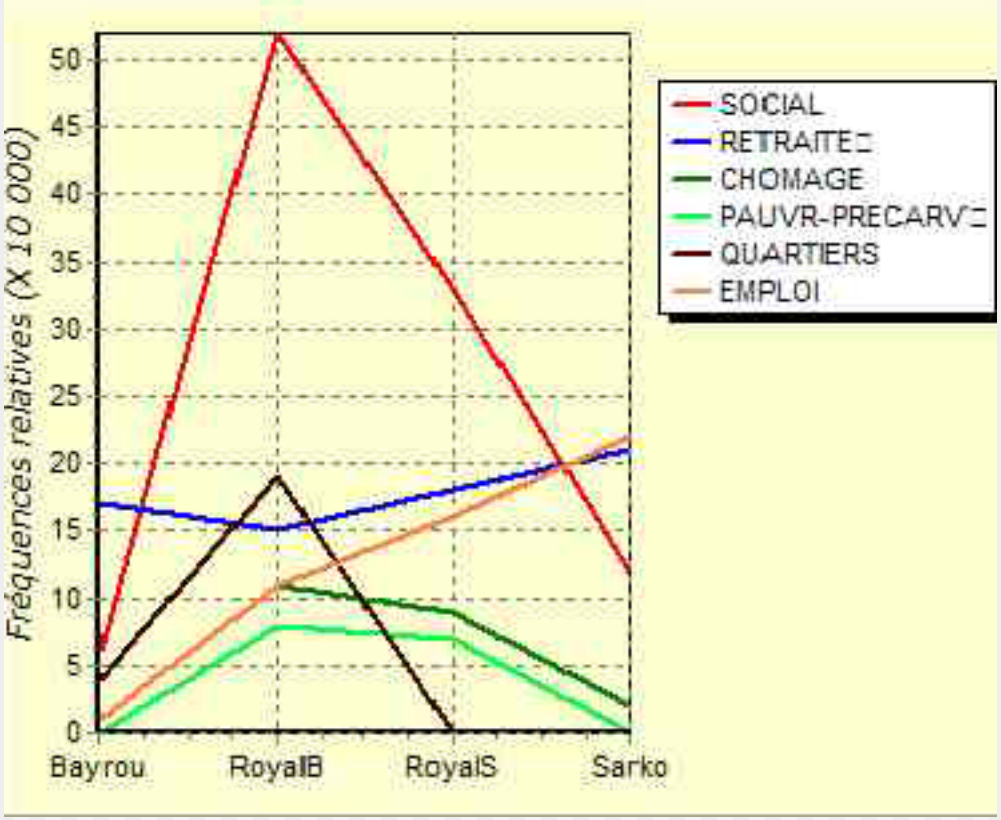
—



Ségolène en campagne 2007 (M. Tournier, 2007)
Moi-je

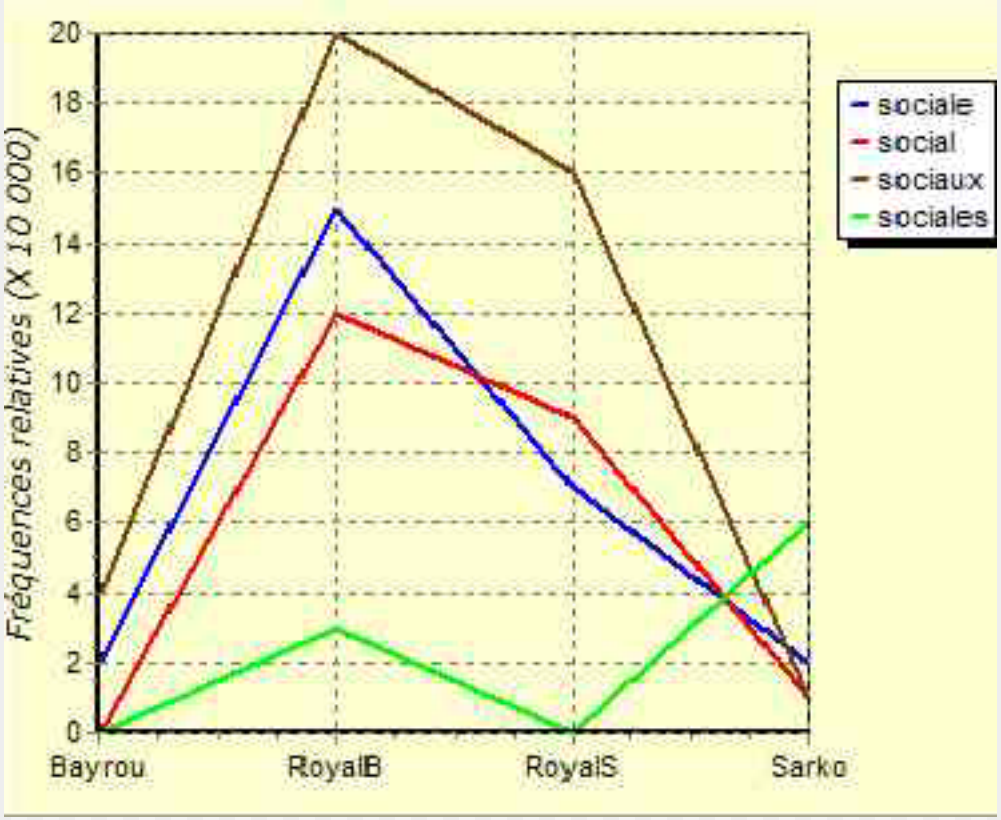


Ségolène en campagne 2007 (M. Tournier, 2007)
Et le social?

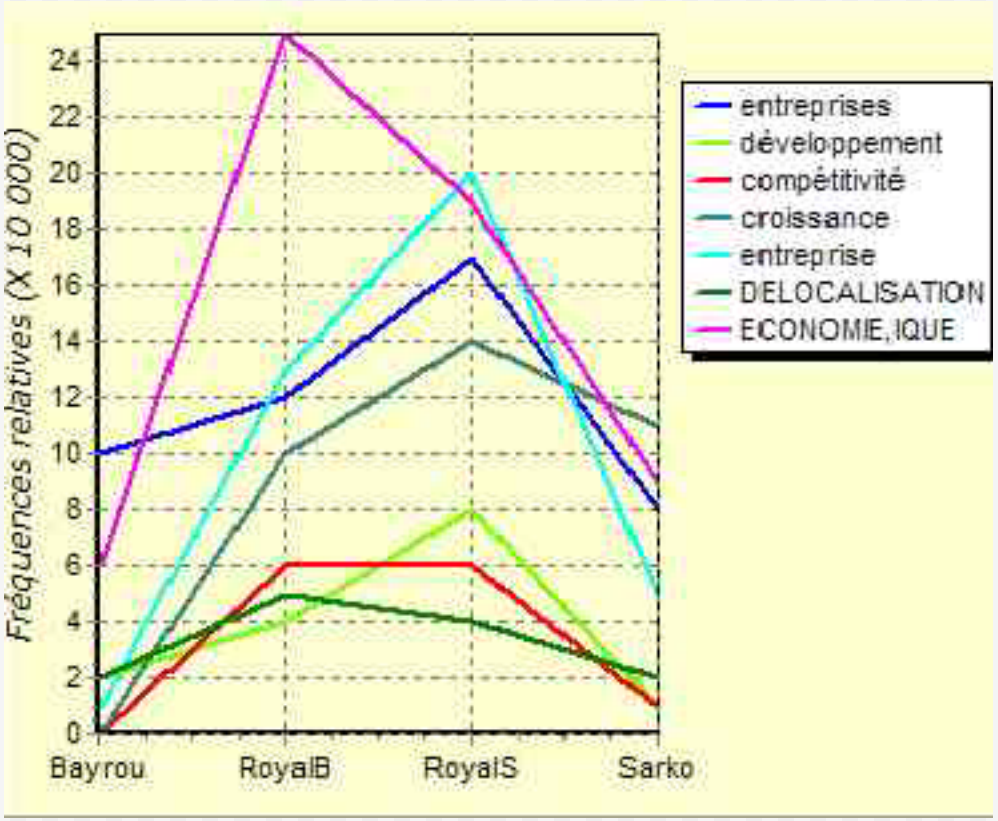


Textométrie et ADP
http://textopol.free.fr/

Ségolène en campagne 2007 (M. Tournier, 2007)
Et le social?

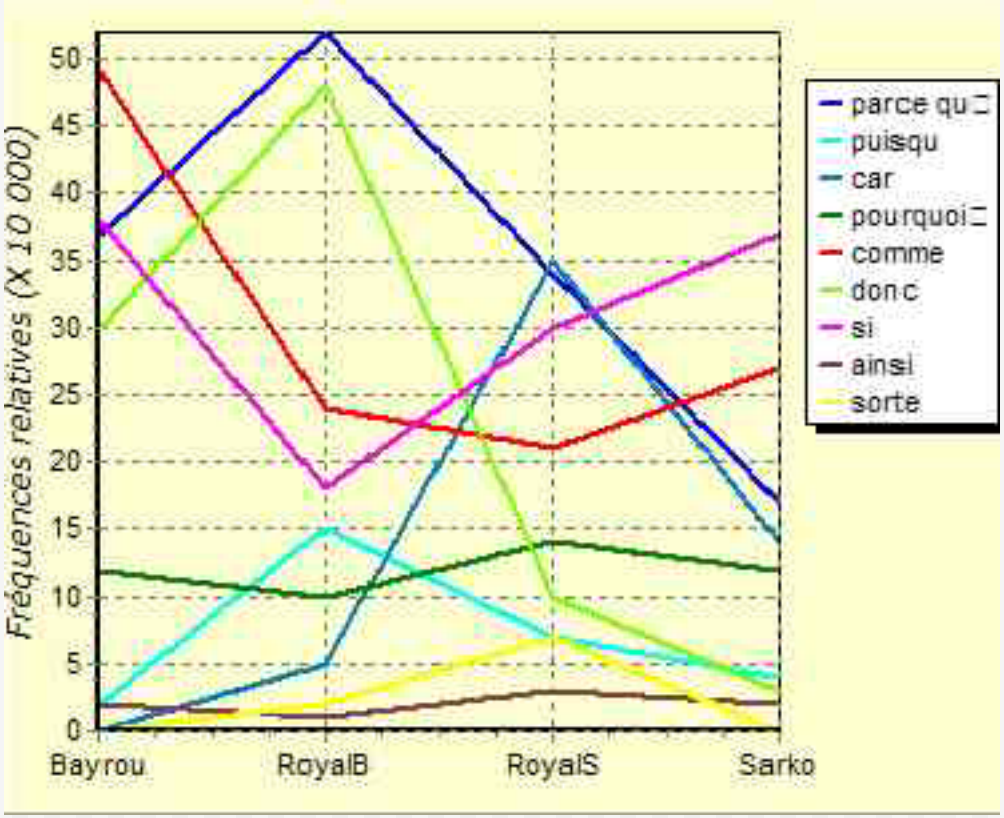


Ségolène en campagne 2007 (M. Tournier, 2007) Et l'économie?



Ségolène en campagne 2007 (M. Tournier, 2007) Connecteurs?

—



Campagne présidentielle 2007 Ségolène en campagne 2006 (P. Marchand, 2007)

- **Les débats à l'américaine au PS**

Campagne présidentielle Les vœux 2007 (PF-JML, TEXTOPOL, 2007)

- Le croisement d'un genre rituel et d'un hypergenre : le discours électoral
- Candidats vs non candidat
- Présidentiables vs non présidentiables

Campagne présidentielle 2007 Ecologie de gauche ou de droite? (P. Marchand, 2007)

- Les déclarations des candidats chez Monsieur Hulot

Les mots de l'humanitaire: La parole des Prix Nobel de la Paix: questions et hypothèses

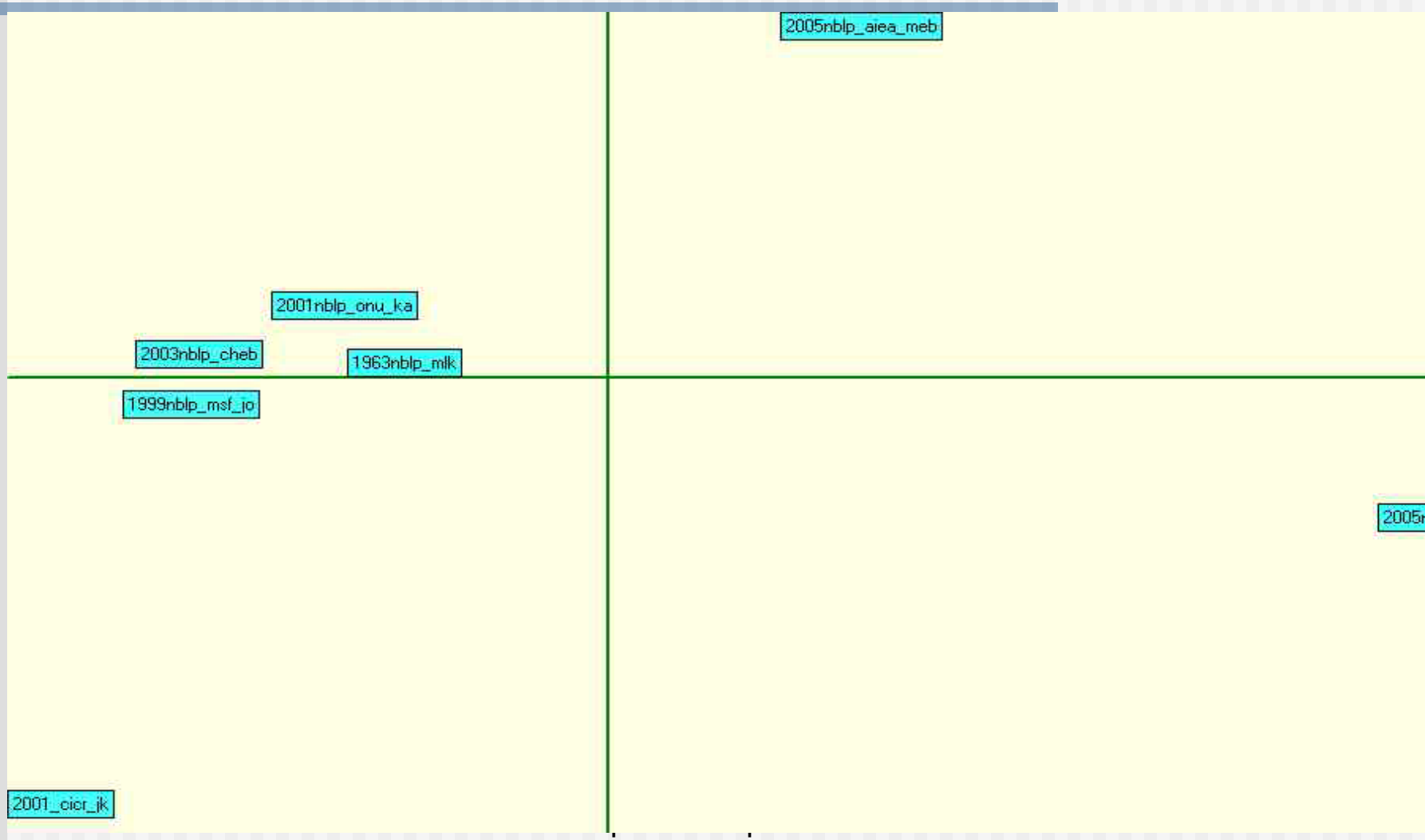
- Quelle parole politique?
- Quel éthos du Nobel de la Paix?
- Quel rapport à l'humanitaire?
- Quels rapports entre la paix, l'humanitaire et la politique?
- La Position des acteurs et de leurs institutions?

Corpus 7PNBL, un échantillonnage plausible

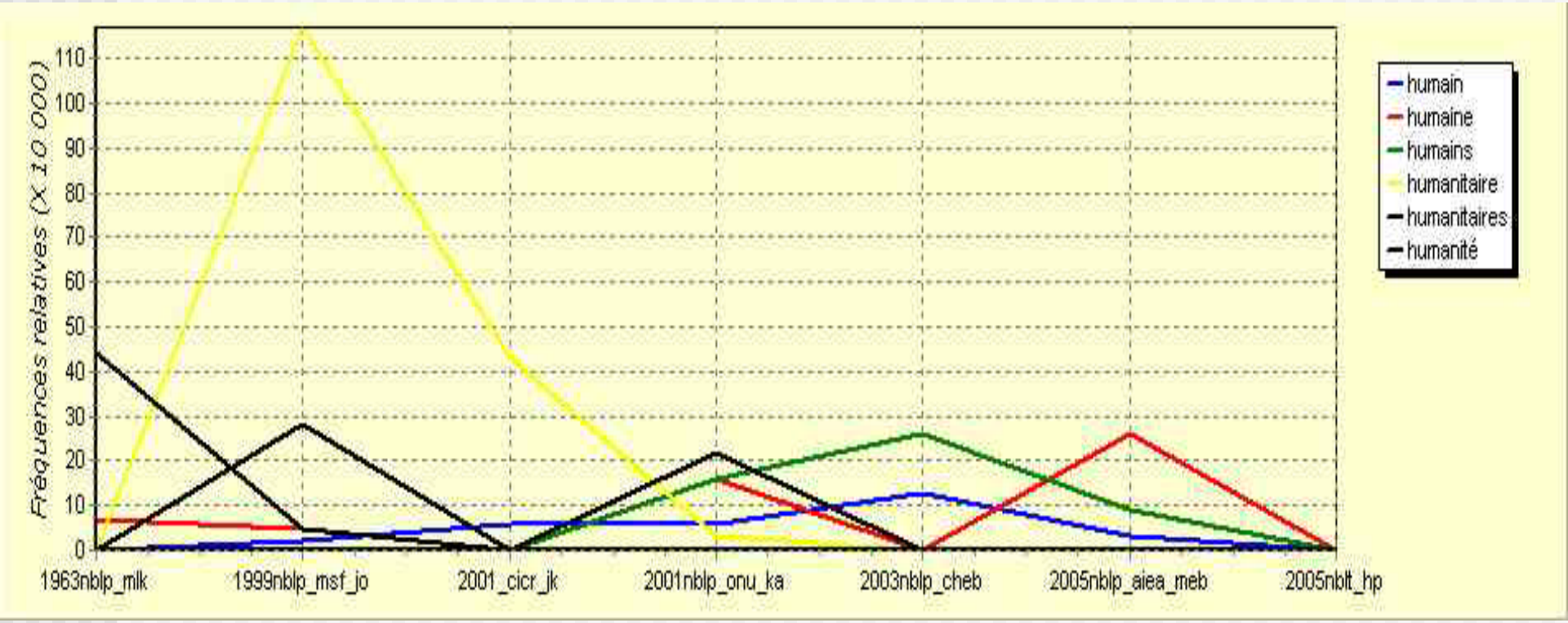
locuteur	Institution	date	Longueur	vocabulaire
Martin Luther King		1963	1344	555
James Orbinski	MSF	1999	3477	1070
Jakob Kellenberg	CICR	2001	1609	552
Kofi Annam	ONU	2001	3050	938
Chirin Ebadi		2003	741	339
Mohamed Elbaradei	AIEA	2005	3011	1113
Harold Pinter		2005	5591	1816

Analyse factorielle des correspondances - loc

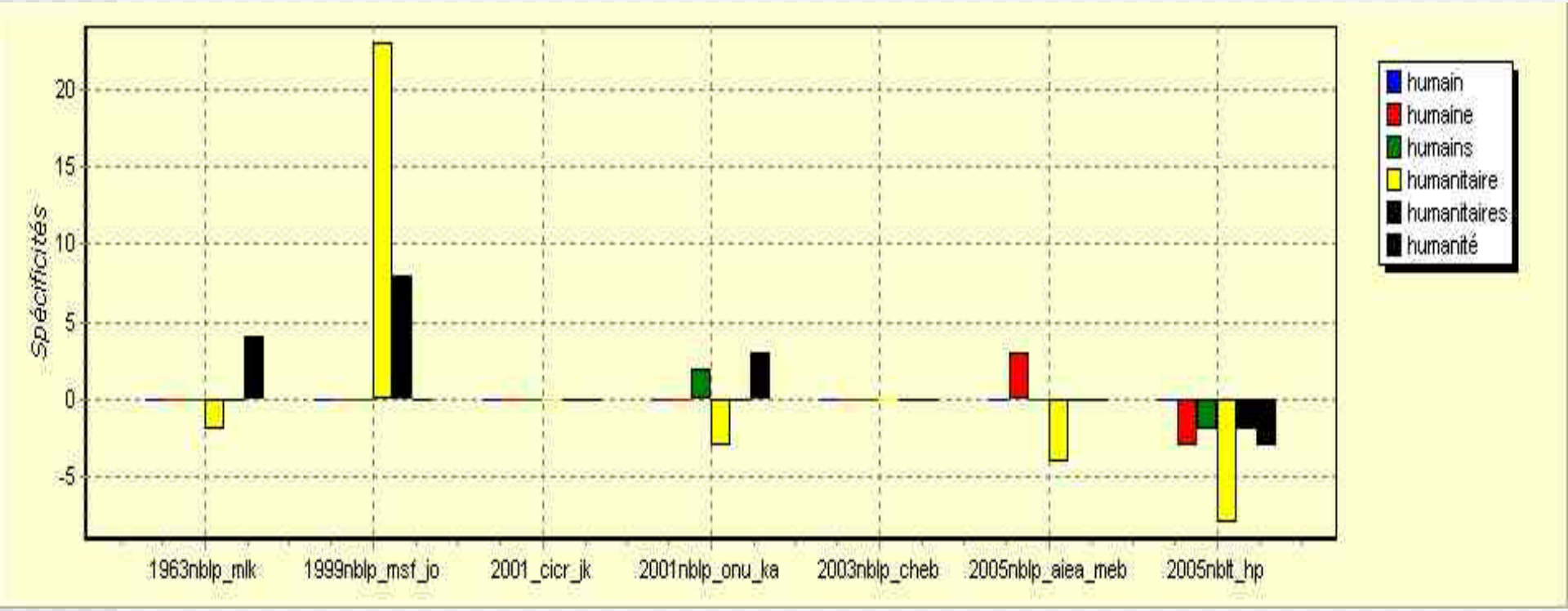
Corpus : 7nblDate : lundi 6 novembre 2006 - 17:18Partition = locNombre de parties :



Les termes de l'humanitaire dans 7NBL



Les termes de l'humanitaire dans 7NBL



Les termes de l'humanitaire dans 7NBL

Concordances de : humanitaires (tri avant)

dition nécessaire pour mener des actions **humanitaires** . msf s ' est construit en réfutant
européenne mobilisable pour des actions **humanitaires** . nous demandons aux ne s opérations
militaires qualifiées " d ' **humanitaires** " . nous affirmons avec force le pri , ni
une purification ethnique avec des **humanitaires** . les humanitaires ne peuvent faire ni la
cun des acteurs . le temps et l ' espace **humanitaires** ne sont pas ceux des politiques . bilités
politiques et non des impératifs **humanitaires** . l ' acte humanitaire est le plus ethnique
avec des humanitaires . les **humanitaires** ne peuvent faire ni la guerre ni la paix
interventions qualifiées de " militaro**humanitaires** " . l ' action humanitaire a pour
droits des victimes et des organisatios **humanitaires** . il définit aussi le devoir des etats à
histoire a montré que les préoccupations **humanitaires** , issues de la société civile

Les mots de la terreur:

- Parole transversale dans la dichotomie gauche droite

Conclusion: La textométrie: un chemin original dans l'ADSP française

- Appliquée
- Cumulative
- Opératoire
- Comparative
- Exhaustive
- Expérimentale
- Evolutive