

## Approches quantitatives des corpus textuels

André Salem  
Systèmes linguistiques, énonciation et discursivité  
(SYLED - EA 2290)  
Université de la Sorbonne nouvelle - Paris 3

André Salem Barcelona 2006

## L'équipe CLA<sup>2</sup>T - Syled

- CLA<sup>2</sup>T-Syled, Université de la Sorbonne nouvelle - P3
- SYLED : Système linguistiques énonciation et discursivité
- CLA<sup>2</sup>T : Centre de lexicométrie & d'analyse automatique des textes
- Revue : *Lexicometrica* <http://www.cavi.univ-paris3.fr/lexicometrica/>
- Logiciel : Lexico3  
<http://www.cavi.univ-paris3.fr/ilpqa/ilpqa/tal/LexicoWWW/>
- Principales publications
  - Pratique des Segments répétés* (1987)
  - Statistique textuelle* (1994) avec L. Lebart
  - Exploring Textual Data* (1998) avec L. Lebart & L. Berry
  - Les linguistiques de corpus* (2000) avec B. Habert & A. Nazarenko
  - Explorations textométriques* (2005) avec S. Fleury

André Salem Barcelona 2006

## Plan

- Approches quantitatives du texte (rappels)
- Les unités de décompte
- Topographie textuelle
- Résonance textuelle

André Salem Barcelona 2006

## Point de vue :

Vite, la Constitution de l'Europe !

par V. Giscard d'Estaing

*Le Monde* du 10 juillet 2004

*.../ Le projet de Constitution (...) vient d'être adopté à l'unanimité par le Conseil européen. (...) Au total c'est le projet de la Convention qui a été adopté, avec quelques retouches. Notre projet n'a pas été détricoté ! Sur les 14 740 mots que comprend le nouveau texte dans sa partie constitutionnelle, 14 000 mots proviennent de notre projet soit 95 %. 680 mots seulement ont été modifiés.*

*La plupart de ces retouches se situent en retrait de notre texte. On est donc mal placé pour critiquer notre manque d'audace ! /.../*

André Salem Barcelona 2006

## Des traditions très éloignées

- Philologie
- Linguistique
- Analyse du discours
- Analyse de contenu
- Intelligence artificielle
- Linguistique computationnelle
- Statistique textuelle / lexicométrie

André Salem Barcelona 2006

## Des préoccupations différentes

- 1970 Etudes littéraires (Racine - Corneille)
- 1980 Textes politiques (Partis, Syndicats)
- 1990 Réponses à des questions ouvertes dans les enquêtes socioéconomiques
- 2000 Extraction de ressources textuelles à partir de corpus de textes  
Corpus de documents sur le *web*.

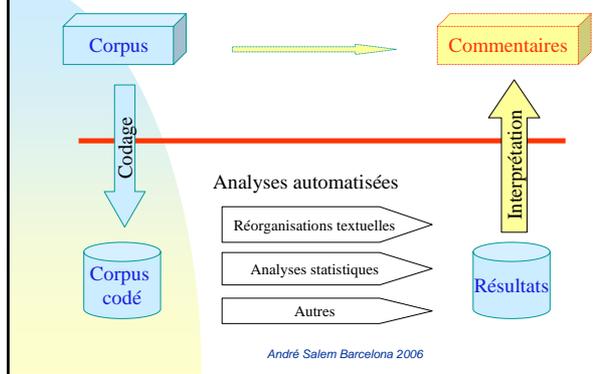
André Salem Barcelona 2006

### Des besoins différents

- Scanner un texte
- Corriger les fautes d'orthographe, de syntaxe
- Résumer un texte
- Traduire un texte
  
- Constituer des typologies de textes
- Explorer à l'aide des méthodes textométriques

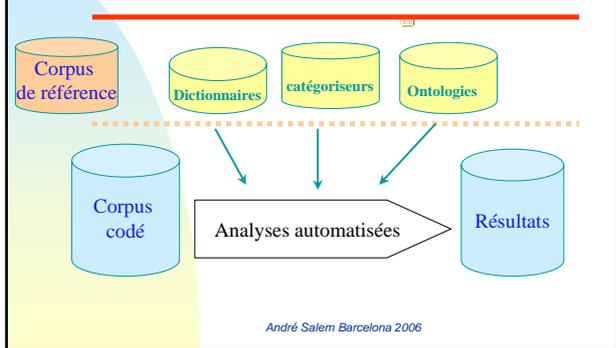
André Salem Barcelona 2006

### Analyses textuelles sur corpus



André Salem Barcelona 2006

### Ressources textuelles



André Salem Barcelona 2006

### Déclaration des droits de l'homme et du citoyen

26 août 1789

Les représentants du peuple français, constitués en Assemblée nationale, considérant que l'ignorance, l'oubli ou le mépris des droits de l'homme sont les seules causes des malheurs publics et de la corruption des gouvernements, ont résolu d'exposer, dans une déclaration solennelle, les droits naturels, inaliénables et sacrés de l'homme, afin que cette déclaration, constamment présente à tous les membres du corps social, leur rappelle sans cesse leurs droits et leurs devoirs ; afin que les actes du pouvoir législatif et ceux du pouvoir exécutif, pouvant être à chaque instant comparés avec .....

André Salem Barcelona 2006

### « Synthèse » produite par CORDIAL

Ce texte est assez court, ce qui rend délicate une analyse de contenu.  
 Le domaine "anatomie" est le plus saillant du texte  
 Le domaine "administration" est également un domaine d'importance.  
 Le domaine "géométrie" est le troisième domaine saillant.  
 Dans la thématique de ce texte, la collectivité, par opposition à l'univers et à l'humain, occupe une place capitale.  
 D'une façon plus précise, l'analyse des thèmes généraux de ce texte indique une prédominance des thèmes suivants : "Le droit" et "La volonté".

André Salem Barcelona 2006



André Salem Barcelona 2006

## La Bible (1) : Hapax

- **Petit Robert**
- **Hapax legomenon** :
  - «Chose dite une seule fois »
- **Mot, forme, emploi, dont on ne peut relever qu'un exemple à une époque donnée**
- **Habitude des commentateurs de signaler des emplois uniques dans le texte biblique**

André Salem Barcelona 2006

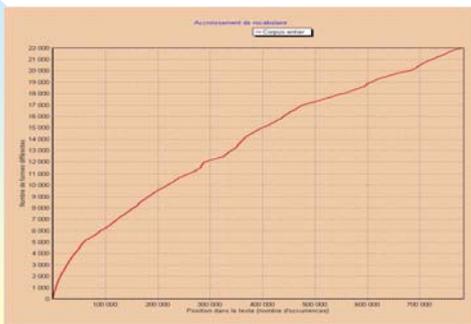
## La Bible (2) : Hapax

Nombre des occurrences 783 408  
 Nombre des formes 22 021  
 Fréquence maximale 32 669  
 Nombre des hapax 7 688

- |                         |                        |
|-------------------------|------------------------|
| ■ <b>Mots « rares »</b> | ■ <b>Mais aussi</b>    |
| • 1 87 Abaddon          | • 1 4514 abaissant     |
| • 1 88 Abagtha          | • 1 4520 abaissent     |
| • 1 89 Abana            | • 1 4525 abandon       |
| • 1 90 Abandonnés       | • 1 4541 abandonnerez  |
| • 1 91 Abandonnez       | • 1 4542 abandonneront |
| • 1 97 Abdeel           | • 1 4546 abandonnons   |
| • 1 100 Abdiel          | • 1 4549 abattait      |
| • 1 104 Abets           | • 1 4551 abattent      |
| • 1 107 Abiasaph        | • 1 4552 abattes       |
| • 1 109 Abib            | • 1 4555 abattis       |
| • 1 116 Abigal          | • 1 4559 abattrais     |
| • 1 117 Abihail         | • 1 4560 abattras      |

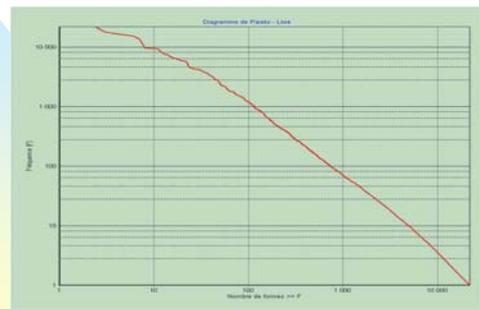
André Salem Barcelona 2006

## La Bible (3) : Accroissement du vocabulaire



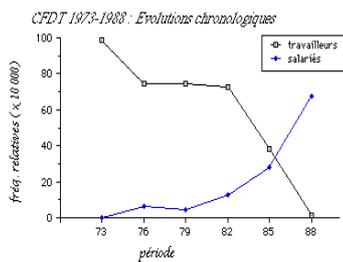
André Salem Barcelona 2006

## La Bible (4) : Diagramme de Pareto



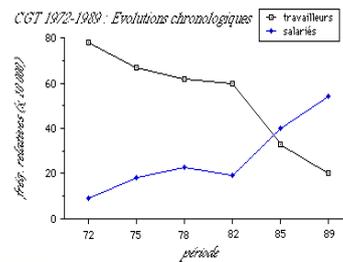
André Salem Barcelona 2006

## Les formes *travailleurs* et *saliés* à la CFDT



André Salem Barcelona 2006

## Les formes *travailleurs* et *saliés* à la CGT



André Salem Barcelona 2006

### Segments répétés (1)

*période 1*

tous les travailleurs	22
pour les travailleurs	13
intérêts des travailleurs	6
aspirations des travailleurs	6
ensemble des travailleurs	19
catégories de travailleurs	6
expression des travailleurs	2
permettre aux travailleurs	6

*période 2*

tous les salariés	6
pour les salariés	8
intérêts des salariés	3
aspirations des salariés	3
ensemble des salariés	4
catégories de salariés	6
expression des salariés	3
permettre aux salariés	2

André Salem Barcelona 2006

### Segments répétés (2)

sur les masses populaires et avant tout sur la classe ouvrière	3
CFDT1 CGT1 CGT1	
la défense des intérêts matériels et moraux des travailleurs	3
CFTC FO1 FO2	
le retour aux quarante heures sans diminution de salaire	3
CFDT1 CGT3 CGT3	
la propriété sociale des moyens de production et d'échange	3
CFDT1 CFDT2 CFDT2	
la séparation et l'équilibre des pouvoirs issus du suffrage universel	3
CGT1 CGT1 CGT2	

André Salem Barcelona 2006

### Segments répétés (3)

- [Congrès CGT 1978] § L'avenir socialiste de la France
- § le régime capitaliste fait apparaître de manière aiguë et massive ses tares et ses absurdités, son incapacité profonde /.../
- § Dès sa fondation, la CGT s'est assigné **pour but** de transformer la société capitaliste en mettant un terme à l'exploitation capitaliste /.../
- § **Pour la CGT, le socialisme est indissociable de la liberté/.../**
- /.../
- § Les réflexions exposées dans ce document constituent **une** base commune aux organisations de la CGT pour la poursuite de la réflexion et de la discussion tant avec les travailleurs qu'avec les forces intéressées à ce but et pour laquelle la CGT demeure disponible.

André Salem Barcelona 2006

### Procédures d'analyse lexicométrique (cas général)

#### Le tableau lexical entier (TLE)

Num	Forme	P1	P2	P3	P4	P5	P6	P7	P8
0	de	886	875	853	753	746	757	669	591
1	les	641	688	569	549	534	687	555	526
2	la	648	550	597	581	482	505	486	449
3	et	449	524	480	530	463	467	461	399
4	le	348	376	398	374	356	321	327	265
5	à	382	384	350	361	319	308	266	262
6	que	349	390	351	298	317	287	287	217
7	qui	222	276	310	261	271	268	267	204
8	des	262	262	240	201	245	243	274	217
9	il	300	233	285	199	248	229	203	128
10	l	221	260	250	236	210	201	206	187
11	pour	214	249	252	220	194	181	183	172
12	en	171	199	153	192	167	169	147	111
13	qu	184	189	225	164	184	139	115	98
14	d	170	158	170	151	162	161	152	150
15	nous	180	250	167	132	155	100	182	104

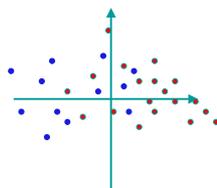
André Salem Barcelona 2006

### Analyses typologiques

(cas général)



Classification automatique



Analyse factorielle

André Salem Barcelona 2006

### Les unités de décompte

- Ngrammes (lib, ber, rté)
- Racines (liberté(s), libre, libérer)
- Lemme (liberté, libérés)
- Forme graphique (liberté)
- Segments (liberté de la presse)
- Cooccurrences (liberté + égalité)

André Salem Barcelona 2006

## Les types généralisés (Tgen)

■ sous-ensemble d'occurrences du corpus



■ Exemples de Tgen(s)

- ◆ les occurrences d'un segment répétés
- ◆ les cooccurrences de deux formes à l'intérieur de phrases.
- ◆ un ensemble de formes présentant un lien au plan sémantique
- ◆ le résultat d'un surlignage sélectif par un humain
- ◆ une classe de fréquence

André Salem Barcelona 2006

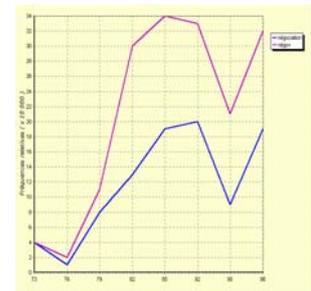
## CFDT 1973-1998

Ventilations de la forme :

***négociation***

et du TGen

***négo+***

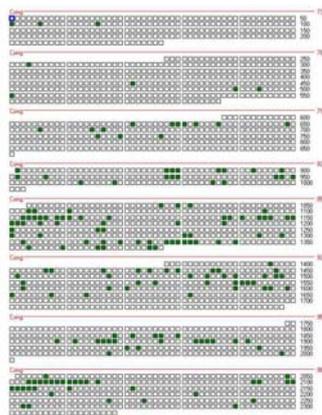


André Salem Barcelona 2006

## CFDT 1973-1998

Vers une topographie textuelle

TGen : ***négo+***



André Salem Barcelona 2006

## CFDT congrès de 1998, §§ 2051-2052

§ dans les **négociations** d'entreprise et de branche, dans les fonctions publiques et les entreprises publiques, la CFDT lie de manière dynamique et diversifiée les salaires et l'emploi.  
 § le choix de l'emploi par la RTT fait de la compensation salariale un des éléments de la **négociation**, sans a priori dans un sens ou dans l'autre. dans ces **négociations**, les équipes syndicales prennent en compte le volume d'emplois créés, l'ampleur de la RTT, la participation de l'entreprise et le niveau des salaires et de ses éléments accessoires (intéressement, participation, actionariat...).

André Salem Barcelona 2006

## Formes spécifiques

Question ouverte

Termes spécifiques pour les jeunes diplômés

Terme	F	f	spec
financières	173	34	+04
problème financiers	16	6	+03
couple	95	19	+03
responsabilités	21	7	+03
raisons financières	92	20	+03
situation économique	11	4	+02

André Salem Barcelona 2006

## Corpus de lettres de réclamation

*segment*

*cooccurents*

ma facture  
 une facture  
 cette facture  
 votre facture

recevoir  
 recevoir  
 conteste[r]  
 contester  
 accuser réception  
 retourner

André Salem Barcelona 2006

## La résonance textuelle

**Explorations de corpus à partir de préoccupations issues de l'analyse de discours**

André Salem  
Systèmes linguistiques, énonciation et discursivité (SYLED - EA 2290)  
Université de la Sorbonne nouvelle - Paris 3

André Salem Barcelona 2006

### Topographie bi-textuelle (1) M. Zimina

volet français

volet anglais

Termes	Frq Tot	Frq	Sp.
fonctionnaires	43	43	100
des fonctionnaires	14	14	21
des fonctionnaires	14	14	31
de loyauté	36	14	22
loyauté	42	14	21
de loyauté politique	22	10	17
loyauté politique	24	10	16
de fonctionnaires	7	7	16
obligation de loyauté politique	15	8	14
obligation de loyauté	21	8	12

Termes	Frq Tot	Frq	Sp.
servants	50	31	55
civil servants	46	29	52
civil	304	41	40
loyalty	42	14	20
duty	100	15	15
political loyalty	25	10	16
of political	29	10	15
duty of	45	11	15
officers	38	10	14
duty of political loyalty	23	9	14
service	110	14	14
civil service	58	11	13

### Topographie bi-textuelle (2)

volet français

volet anglais

Termes	Frq Tot	Frq	Sp.
fonction	38	10	19
fonction	16	7	16
fonction	19	6	13
service police	9	4	11
police	216	9	10
service police	3	3	9

André Salem

### Débat Mitterrand / J. Chirac (présidentielles de 1988)

- ÉLIE VANNIER
- *Nous reviendrons au problème du chômage tout à l'heure...*
- JACQUES CHIRAC
- ... Je voudrais simplement dire un mot sur la présentation que fait M. Mitterrand du chômage.
- Non, monsieur Mitterrand, nous n'avons pas tous échoué de la même façon...
- FRANÇOIS MITTERRAND
- ... C'était pire avec vous!... Vous avez doublé...
- JACQUES CHIRAC
- ... Vous me permettez de parler de l'actualité...
- FRANÇOIS MITTERRAND
- Ah oui... c'est cela!... Vous voulez éviter le passé lorsqu'il est lourd!...
- JACQUES CHIRAC
- J'assume toutes mes responsabilités, monsieur Mitterrand...

André Salem Barcelona 2006

### Introduction à la résonance textuelle (1)

sélection des paragraphes contenant un terme

termes d'induction

Locuteur A

alignement

Locuteur B

André Salem Barcelona 2006

### Introduction à la résonance textuelle (2)

résonance dans des textes alignés en paragraphes

termes d'induction

Locuteur A

alignement

Locuteur B

réponse

André Salem Barcelona 2006

## Introduction à la résonance textuelle (4)

**application** : présidentielles françaises de 1988  
(deux tours de parole consécutifs)

### F. Mitterrand

il faut d'abord distinguer, c'est un problème qui a été vraiment exagéré et compliqué à plaisir. il y a plusieurs catégories de personnes visées par le débat actuel. il y a d'abord ceux qui ne sont pas des **immigrés**, qui sont les enfants d'**immigrés** et qui sont nés sur notre sol. ceux-là ont vocation. ils sont français, sauf s'ils en décident autrement à l'âge de dix-huit ans. il y a, ensuite, les naturalisés; ce sont les **immigrés** qui désirent devenir français. là, l'administration étudie leur cas et il aboutit à reconnaître le droit à la naturalisation, selon son propre rythme. je n'insiste pas. et puis il y a les **immigrés** /.../\_/

### J. Chirac

je voudrais répondre, moi, très clairement en m'appuyant sur mon bilan dans cette affaire; parce que c'est très gentil de faire des promesses, mais enfin, encore faut-il qu'elles soient rendues crédibles par un bilan, s'agissant de l'**immigration** tout court, il faut la stopper, parce que nous n'avons plus les moyens de donner du travail à des étrangers, aussi, naturellement, en supposant quelques souplesses naturellement, mais il faut la stopper. s'agissant de **immigration** clandestine, il faut évidemment lutter contre cette **immigration** avec beaucoup d'énergie et reconduire les.../

André Salem Barcelona 2006

## Les séries textuelles chronologiques 4

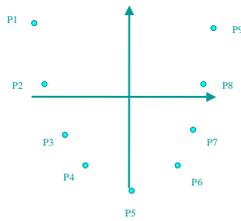
- Quelles sont les formes qui sont à la base de l'évolution ?
- Quelles sont celles qui apparaissent (ou disparaissent) lors de chaque période ?
- Peut-on mettre en évidence certains segments de texte particulièrement caractéristiques de l'évolution lexicale du corpus ?

André Salem Barcelona 2006

## Les séries textuelles chronologiques 2

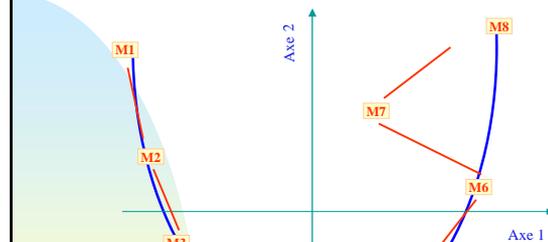
Corpus

P1	P2	P3	P4	P5	P6	P7	P8	P9
----	----	----	----	----	----	----	----	----



André Salem Barcelona 2006

## Les séries textuelles chronologiques 3



André Salem Barcelona 2006

## Les séries textuelles chronologiques (applications)

- Etudes de presse
- Discours politique
- Etudes longitudinales
  - domaine médical
  - communication d'entreprise
  - réponses de clientèle
- Veille technologique
- Trajectoires professionnelles

André Salem Barcelona 2006

## Conclusions

- constituer des ensembles d'unités sur la définition desquelles le chercheur peut agir plus aisément le temps d'une expérience
- cartographier l'extension spatiale de ces unités à travers un découpage du corpus dont le grain peut être réglé à volonté
- utiliser les données de structure, d'alignement, etc. entre les différents éléments de corpus parallèles

André Salem Barcelona 2006

