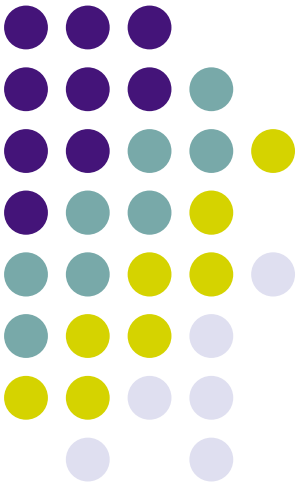
Les mots d'Obama et les maux de McCain



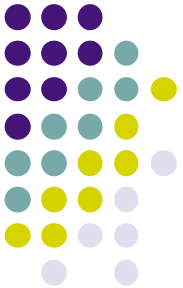
J. Savoy
Institut d'informatique
Université de Neuchâtel

Linguistique computationnelle & sciences politiques

Plan

- **Remarques méthodologiques**
- Autour du mot (fréquence)
- Comparer les corpus
- Applications au corpus électoral
  - Sur-emploi & sous-emploi
  - Evolution

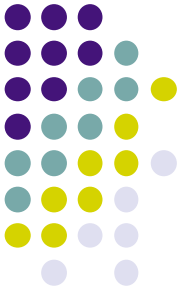
Remarques préliminaires



Politique et informatique ?

- Quelles sont les *mots* qui caractérisent le mieux un texte ? (moteur de recherche ?)
  Les mots ne sont pas le simple fruit du hasard.
  En comparaison avec au moins un autre document.
- Applications
  - Représentation compacte d'un corpus (site web, blog, ensemble de blogs)
  - Etablir des cartographies de la blogosphère (distance)
  - Recherche d'opinion / sentiment sur le web (journaux) (modèle de langue)

Remarques préliminaires



Et pourquoi la politique ?

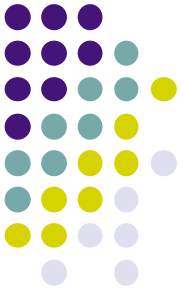
- Nombreux documents à disposition (sur le web)
- Qualité des documents (orthographe)
- Facile à comprendre (large public)
- Touche l'actualité
- Plusieurs études / corpus existent
  - Comparer, mesurer, évaluer les outils en reprenant des corpus existants
- Et en plusieurs langues / cultures / pays
- Limité au lexique ?
  caractères, bigrammes, trigrammes, syntagmes,

Remarques préliminaires

Tout document est-il bon à prendre ?

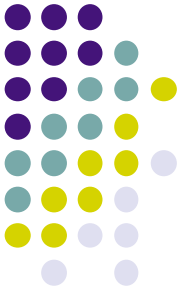
- Les œuvres littéraires d'un même auteur s'avèrent, au niveau de l'analyse lexicale, relativement distantes si l'on change :
  - de composition (roman, théâtre, essai)
  - de genre (prose, vers)
  - époque (début et fin de carrière)
- Différents documents / discours en politique
  - discours du Trône (Canada)
  - discours inaugural (Québec)
  - discours d'investiture (France)
  - discours électoral (convaincre, encourager, compétence)

Prétraitement



- Nécessaire
  - Organiser le corpus (metadata)
  - Unifier l'encodage des caractères
  - Signes de ponctuation (‘‘ ” → " ou “→ ') et accents
- Normaliser les dénominations
  - “US”, “U.S.”, “USA”, “United States”, …
- Normaliser l'écriture (majuscules / minuscules)
  - “Nous”, “nous”, “US”, “United States”, …
    mais "who" et "WHO" (World Health Organization)
- Puis découpage en mots / phrases
  - “La chatte mange la souris.”
    une phrase de 5 formes ou 4 vocables

Découpage en phrase / mots



- Outil rapide mais pas toujours évident …
  - McDonald's
  - can't (I'll, you're, O'Reilly)
  - U.S. President (Washington, D.C.)
  - C|net (Micro$oft)
- Problèmes divers "-"
  - the aluminium-export ban
  - a text-based medium
  - a final "take-it-or-leave-it" offer
  - the 45-year old
  - the New York-New Haven railroad
- Via un outil d'analyse syntaxique (D. Labbé, Grenoble) pour déterminer le bon lemme ("est" → verbe, nom)

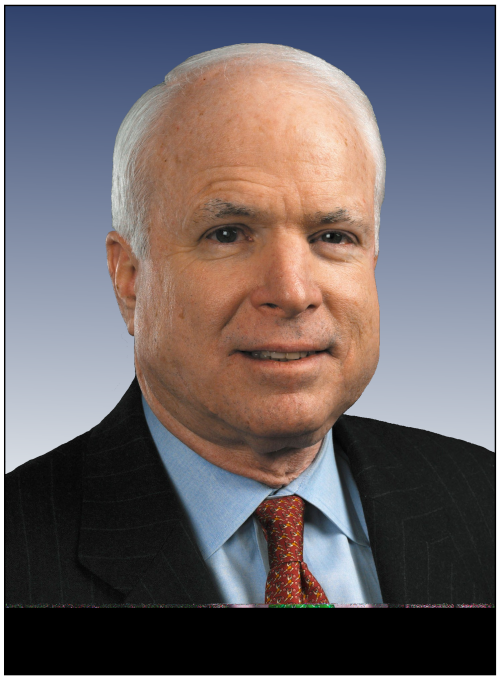
Plan

- Remarques méthodologiques
- **Autour du mot (fréquence ?)**
- Comparer les corpus
- Applications au corpus électoral
  - Sur-emploi & sous-emploi
  - Evolution

Corpus électoral US

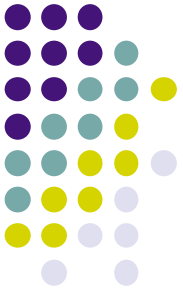
Comment, au niveau lexical, les discours les départagent ?

Democrats



Republicans

Les corpus



- Discours de B. Obama et J. McCain (seulement eux)

  B. Obama :150 discours depuis 10 février 2007
  420 410 formes, 9014 vocables (7792 pour 2008)

  http://www.barackobama.com/

  J. McCain : 94 discours, depuis 25 avril 2007
  206 899 formes, 9401 vocables (7663 pour 2008)

  http://www.johnmccain.com/

- Différence entre le discours électoral et le discours comme membre / chef du gouvernement
- Discours électoraux français
  (élection présidentielle, mai 2007)
  limité à Nicolas Sarkozy (11) et Ségolène Royal (17)

Les mots les plus fréquents

Trivial …

Prendre les plus fréquents !

Prendre les lemmes, (e.g., “be” pour “is” “are” “was”)

Est-ce utile ?

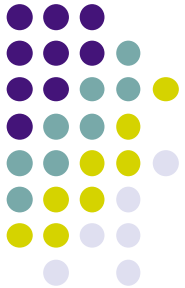
(Permutation)

	McCain		Obama	
Rang	**Lemme**	**Freq.**	**Lemme**	**Freq.**
1	the	5,05%	the	4,47%
2	and	3,97%	be	4,01%
3	to	3,51%	and	3,64%
4	be	3,33%	to	3,17%
5	of	3,00%	*that*	2,49%
6	in	1,97%	of	2,39%
7	a	1,87%	*we*	2,05%
8	*I*	1,52%	a	1,97%
9	*that*	1,48%	in	1,83%
10	*we*	1,42%	*I*	1,46%

Les mots les plus fréquents

PS		PDC		PRD		UDC	
tf	vocable	*tf*	vocable	*tf*	vocable	*tf*	vocable
237	*nous*	643	*nous*	178	*être*	864	*suisse*
198	*politique*	347	*suisse*	176	*suisse*	456	*pas*
192	*doit*	261	*pas*	166	*doit*	445	*politique*
190	*pas*	245	*être*	143	*politique*	384	*ne*
178	*ne*	230	notre	138	*nous*	323	*être*
150	*être*	222	*ne*	108	sécurité	321	état
133	*suisse*	177	*politique*	108	*ne*	320	AI
132	culture	174	PDC	91	*pas*	295	droit
106	culturelle	156	*doit*	90	doivent	286	UDC
104	sociale	144	formation	88	armée	248	étranger

Les mots les plus fréquents



	Suisse	France	S. Royal	N. Sarkozy
1	suisse	je	je	je
2	*nous*	*pas*	nous	pas
3	*pas*	*ne*	vous	ne
4	politique	*nous*	pas	france
5	être	france	france	nous
6	*ne*	vous	ne	veux
7	doit	veux	veux	si
8	droit	si	j	parce
9	notre	parce	notre	être
10	doivent	être	faire	français

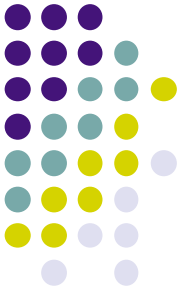
Interprétation

«Ce qui me dérange dans le discours électoral, c'est la haute fréquence des "je".»

«Pas les "nous".»

«Non, non. C'est le "je" que je ne supporte mal.»

Les mots les plus fréquents



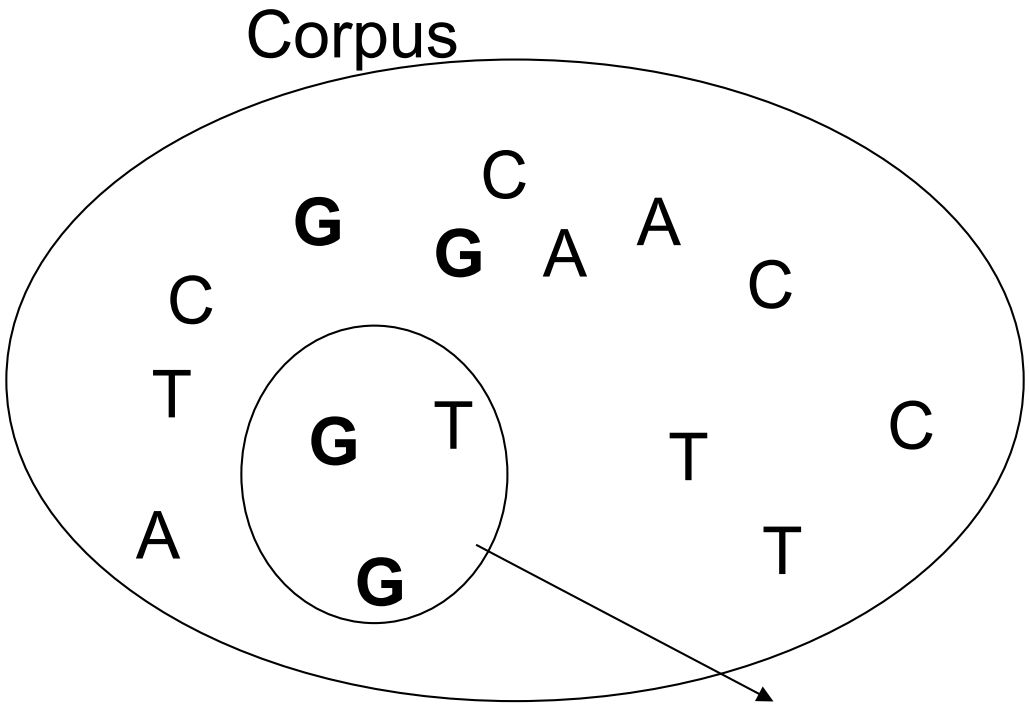
- Tenir compte des mots les plus fréquents ?
  - Peu d'intérêt
  - Pas de grandes découvertes
  - Comparaison difficile
  - Chaque auteur est pris indépendamment de l'autre
- Définir les mots propres à un auteur en fonction de l'ensemble du corpus (ou de l'autre / des autres auteur-s)
- On souhaite avoir les suremplois et les sous-emplois de chaque auteur

Plan

- Remarques méthodologiques
- Autour du mot (fréquence)
- **Comparer les corpus**
- Applications au corpus électoral
  - Sur-emploi & sous-emploi
  - Evolution

Mesure d'association

Idée : Modèle probabiliste d'apparition des mots dans un sous-corpus comparé à un autre



Taille
Corpus : 15
Sous-corpus : 3

On s'intéresse à "**G**"

Mesure d'association

Exemple : Le mot = "Bush" et les discours de McCain'08

Ce mot apparaît 26 fois. Comme McCain'08 représente 22% du total, on devrait trouver 22% de 424 soit 94 fois

	McCain'08	**US-**	**Total**
"Bush"	26	398	424
non "Bush"	154 339	474 331	628 670
Total	154 365 (22%)	474 729 (78%)	629 094 (100%)

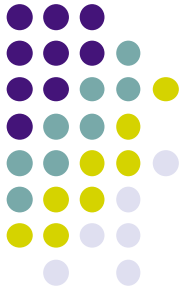
Plan

- Remarques méthodologiques
- Autour du mot (fréquence)
- Comparer les corpus
- **Applications au corpus électoral**
  - Sur-emploi & sous-emploi
  - Evolution

Corpus électoral US

	Sur-emploi	Sous-emploi
McCain	*government*, *Obama*, honor, freedom, power, public, …	because, why, McCain, *Bush*, street, working, …
2007	property, freedom, Islamic, construe, Reagan, enemy, …	because, school, jobs, McCain, children, working, …
2008	Obama, government, Canada, federal, small, judicial, …	why, because, McCain, college, *Bush*, …
Obama	*because*, *why*, *McCain*, college, *Bush*, street, …	government, Obama, honor, freedom, intend, …
2007	bullet, page, Joshua, Chicago, kids, poverty, …	senator, economic, tax, John, trade, government, …
2008	McCain, John, *Bush*, *jobs*, Washington, …	government, Obama, Congress, public, law, …

Corpus électoral US (Obama)

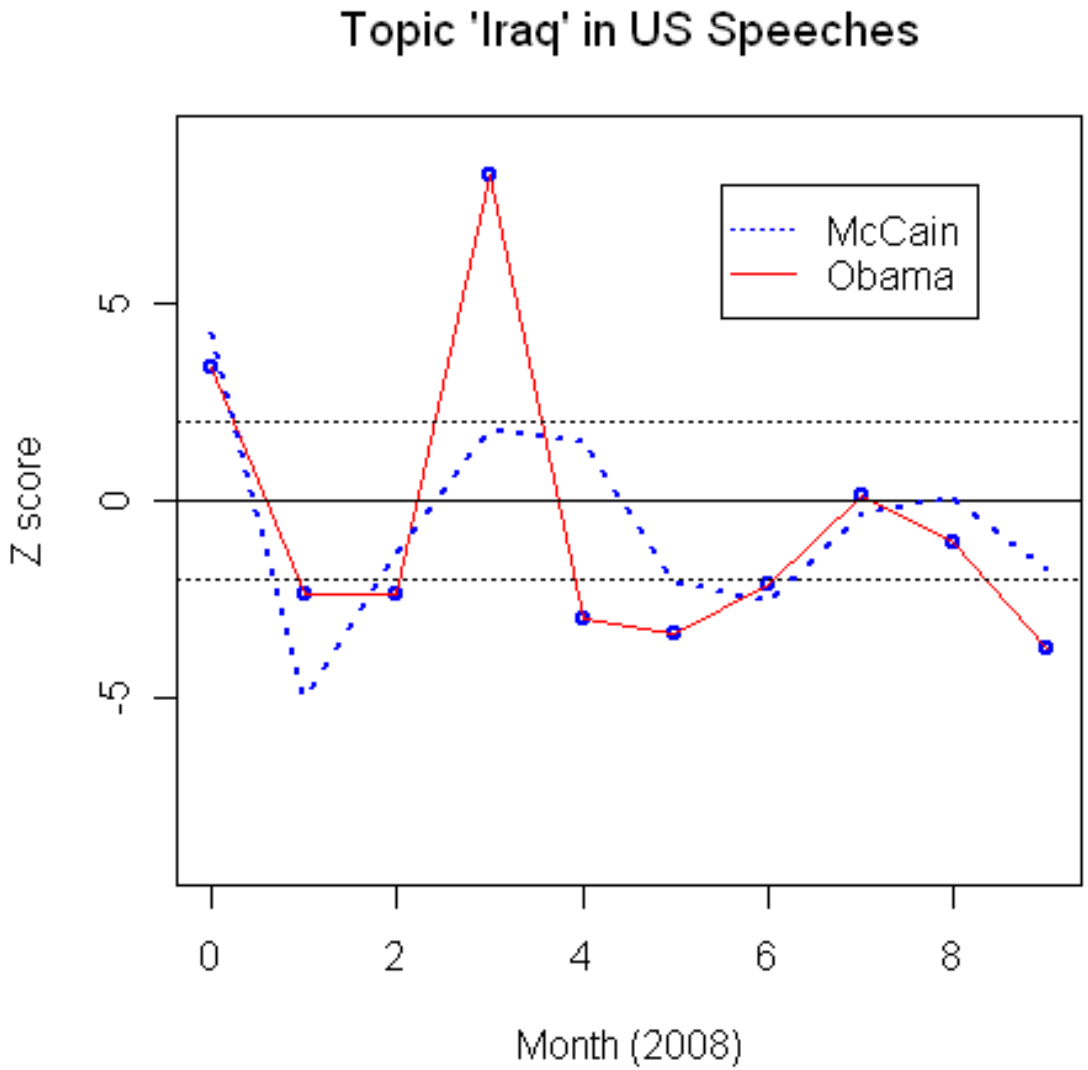


2008	**Sur-emploi**	**Sous-emploi**
May	hemisphere, Cuba, Latin, freedom, …	Iraq, kids, nuclear, market, …
Jun.	Israel, patriotism, Jewish, cities, …	politics, war, veteran, people, …
Jul.	Berlin, women, cyber, Marshall, …	politics, change, tell, story, …
Aug.	Joe Biden, oil, energy, renewable, …	war, white, school, law, …
Sep.	financial, school, courses, McCain,…	war, Iraq, oil, energy, women, …

Corpus électoral US

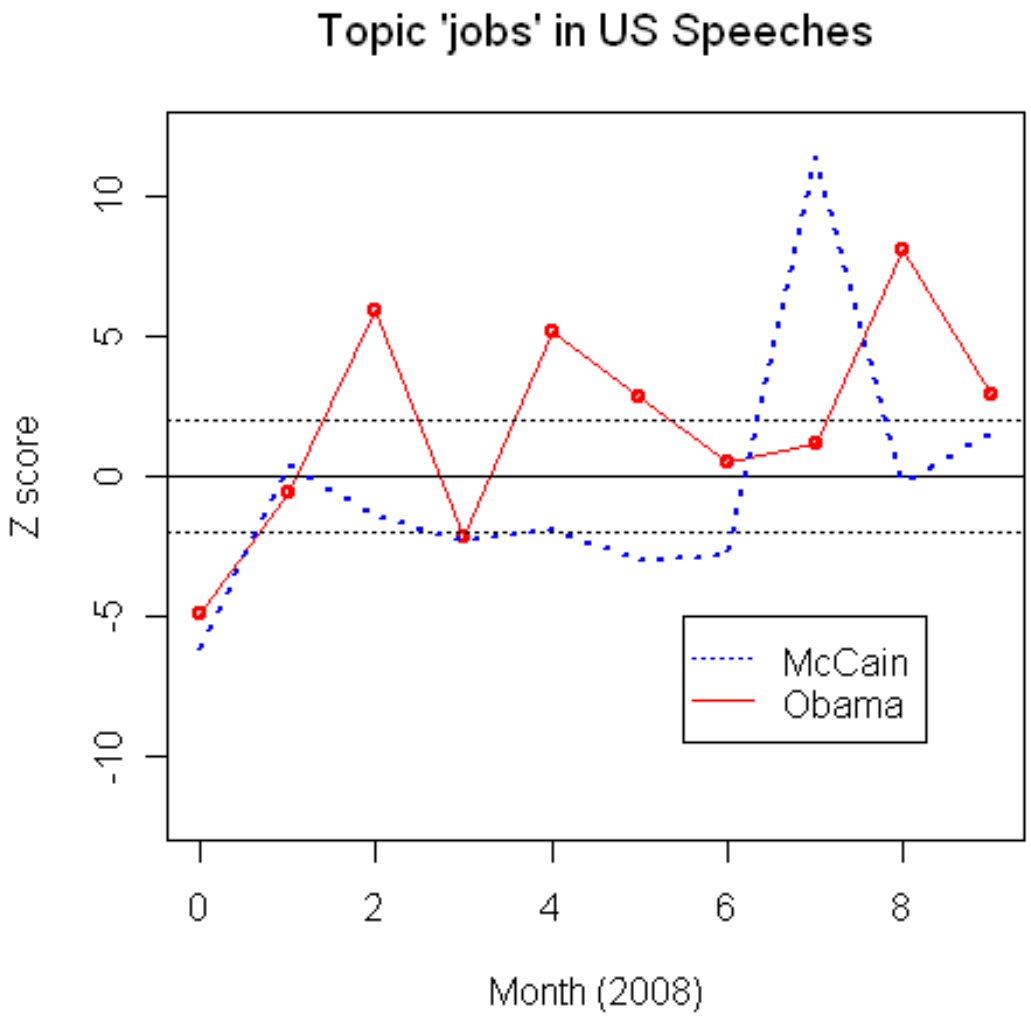
Thème “Iraq”

Mois par mois



Corpus électoral US

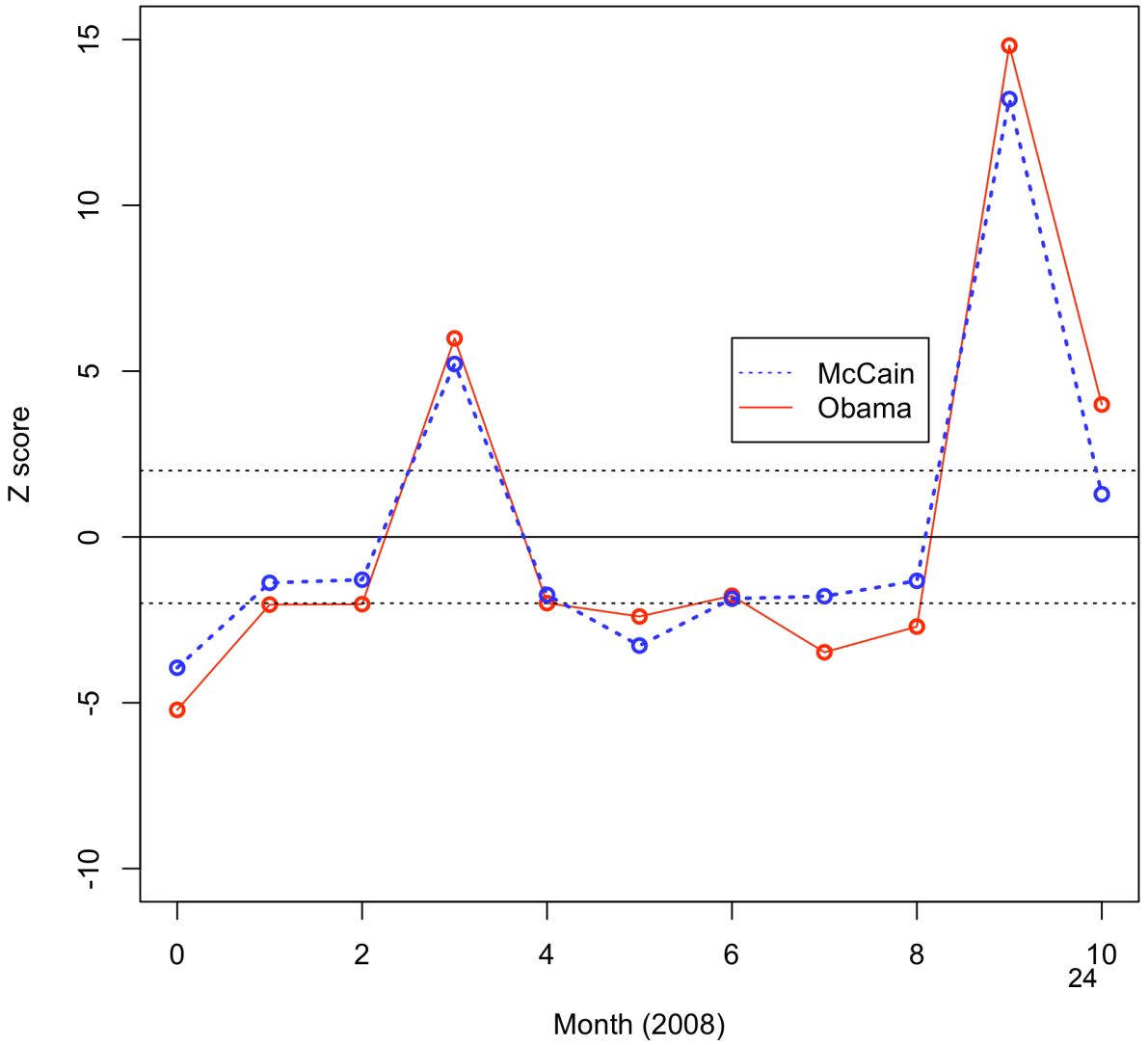
Suivie d'un thème
"jobs", mois par mois



Corpus électoral US

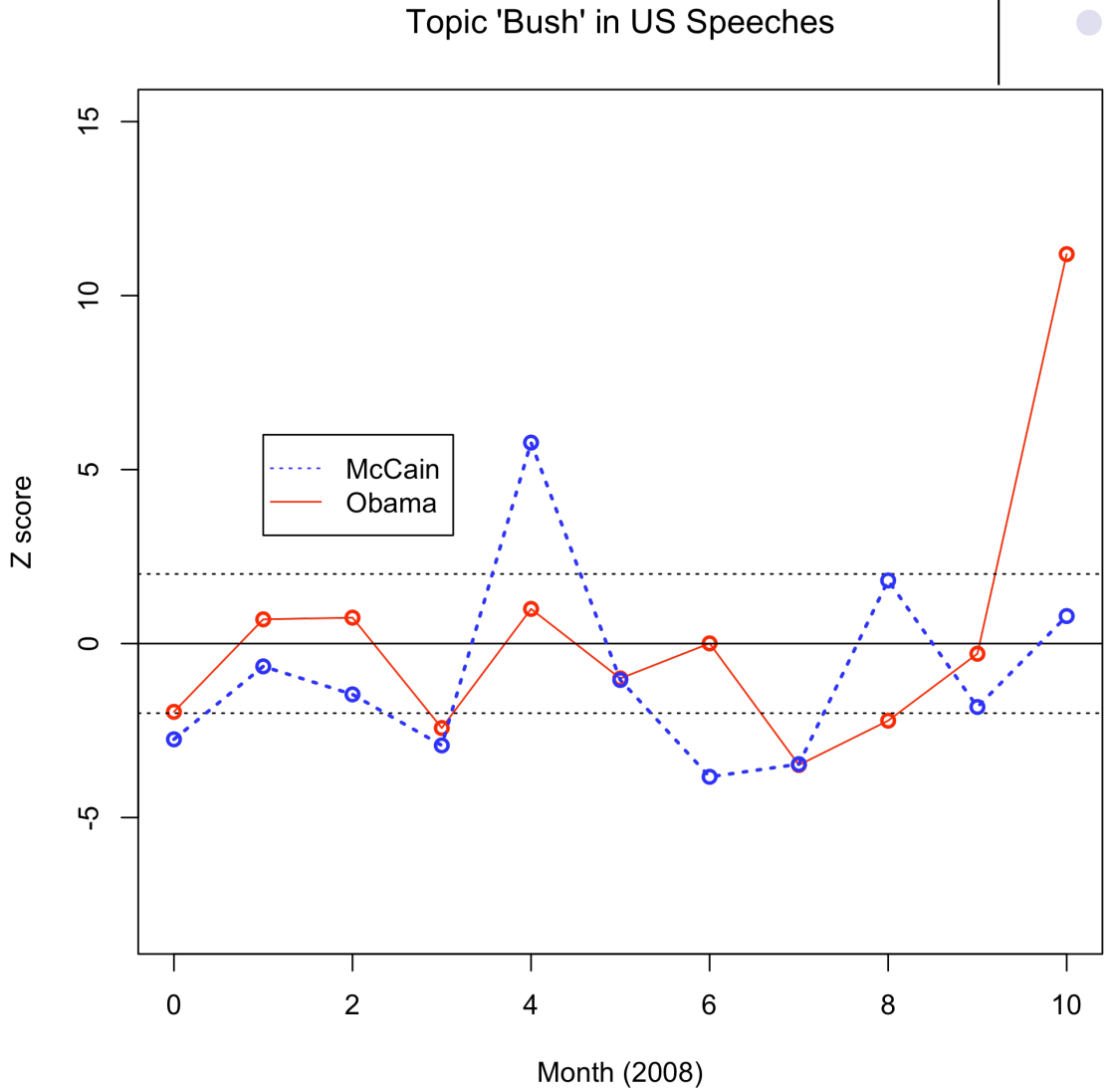
Suivie d'un thème "financial", mois par mois.

Les deux candidats se suivent à la trace…



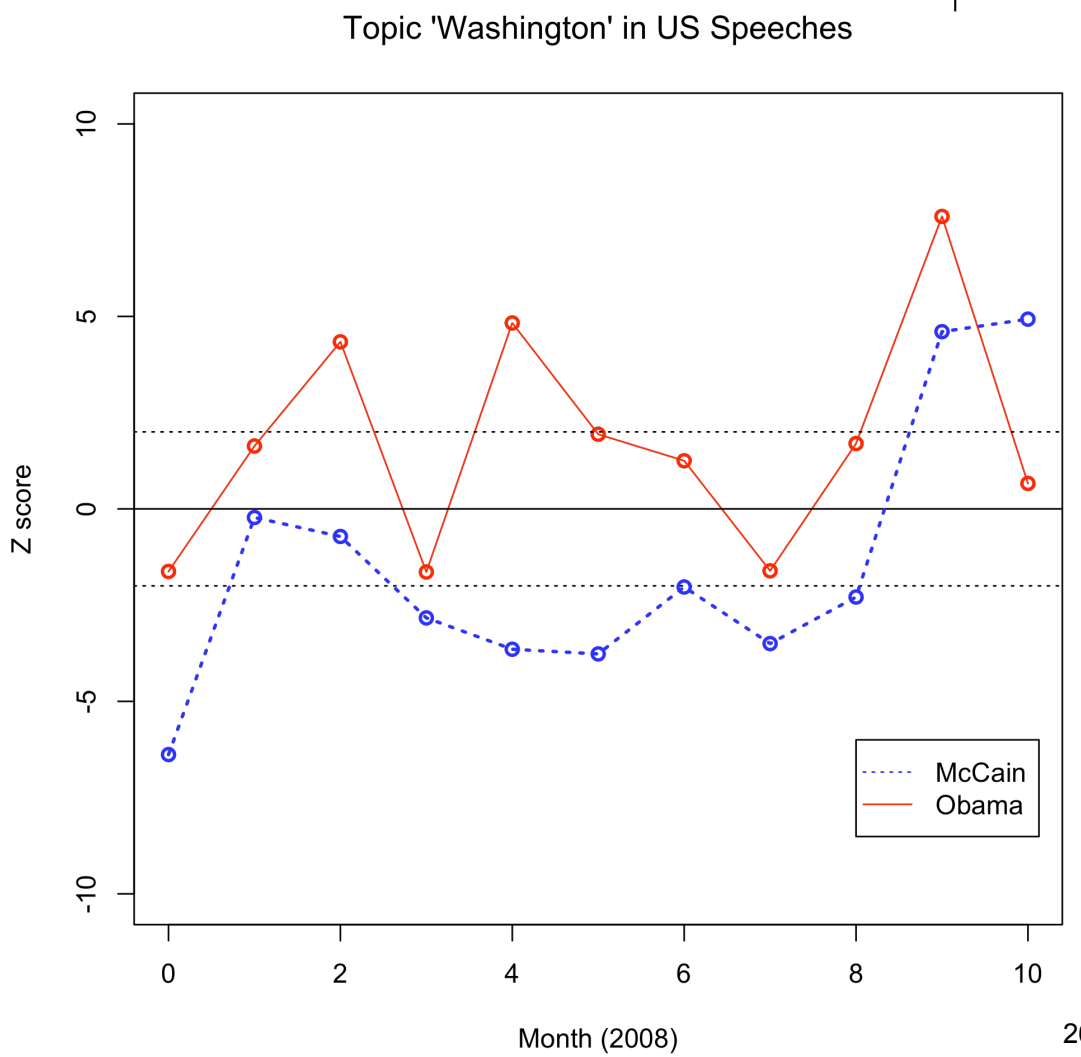
Corpus électoral US

Suivie d'un thème "Bush", mois par mois



Corpus électoral US

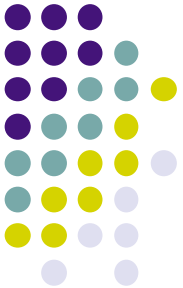
Suivie d'un thème "Washington", mois par mois



Autour d’un mot

	Obama 2008
6	Washington we can
6	failure politician Washington
5	Washington player expect
5	status quo Washington
5	know happen Washington
5	dime Washington lobbyist
5	broken system Washington
4	Washington twenty six
4	Washington think long
4	Washington game Washington
4	they back Washington
4	politician Washington think
4	George Bush Washington

Corpus électoral US

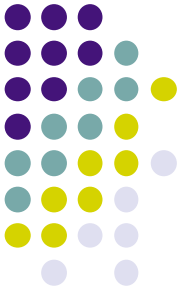


- Prendre en compte les bigrammes (séquences de deux mots)
  - Mais en éliminant les mots fréquents
    sinon on retrouve “in the ”, “of the ”, …
- Tenir compte des parties du discours (POS) ?
  - Contraintes imposées sur les bigrammes admissibles
  - NN – NN (e.g., “carbon pollution”)
  - JJ – NN (e.g., “middle class”)

Corpus électoral US

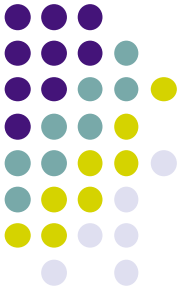
Z	McCain 2008	Z	Obama 2008
28.5	Senator Obama	20.0	Senator McCain
8.4	small business	17.2	John McCain
8.1	government spending	13.9	Wall Street
6.7	tax increase	11.9	middle class
6.6	bad economy	11.4	tax cut
6.3	higher tax	11.0	Main Street
6.2	business tax	9.6	tax break
6.2	flex fuel	9.1	Insurance company
6.1	law enforcement	8.5	George Bush
5.9	more job	8.4	more year
5.9	energy security	7.9	oil company
5.6	great country	7.6	rescue plan
5.6	tax rate	7.5	21st century

Corpus électoral US



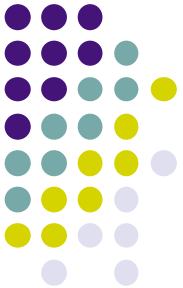
freq.	McCain 2008	freq.	Obama 2008
50	President I will	69	President United States
28	I elected President	67	President I will
25	you thank you	57	United States America
22	thank you thank	42	I running President
21	I believe we	40	we can afford
21	health care system	38	million new jobs
20	dependence foreign oil	35	we can choose
18	small business owner	34	we will make
17	I thank you	34	I President we
16	thank you I	33	President we will
16	I will work	33	I will make
15	I will make	32	will make sure
12	our country I	26	change we need

Corpus électoral US



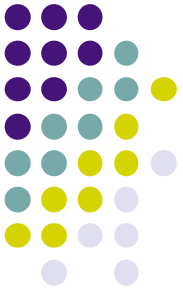
Z	McCain 2008	Z	Obama 2008
5.0	hybrid flex fuel	8.2	State of America
4.6	nuclear power plant	5.6	common sense regulation
4.6	cost of energy	5.5	last eight year
4.5	strong have courage	5.4	middle class family
4.5	stronger better country	5.2	capital gain tax
4.5	selfishness in Washington	4.8	source of energy
4.5	mess of corruption	4.6	world class education
4.4	percent of America	4.6	month in Iraq
4.3	manufacture of hybrid	4.4	time for change
4.3	excess of wall	4.4	job of tomorrow

Et pour la Suisse ?



PS	PDC	PRD	UDC
état	nous	PRD	AI
II	PDC	radical	UDC
culture	demandons	mission	neutralité
culturelle	énergie	armée	gauche
artiste	internet	défense	naturalisation
encouragement	enfant	sécurité	rente
art	notre	militaire	état
autogestion	énergétique	*easy*	nationalité
CO2	PDC	imposition	milliard
pro	formation	*tax*	étranger

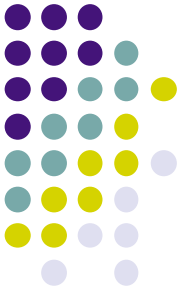
Et Obama Président ?



Sur-emploi du Président

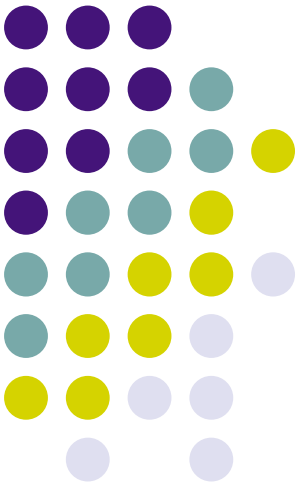
budget	thank
Chrysler	Turkey
department	secretary
recovery plan	recovery act
new foundation	economic recovery
American recovery	new investment
reinvestment act	mutual interest
auto loan	mutual respect
higher education	
health care reform	kind of energy
clean energy economy	long term deficit

Conclusion



- Termes caractéristiques d'un corpus
  mot isolé, séquence bigramme & trigrmme
- Le choix du lexique …
  - Le "nous", "je", mais pas le "she"
  - L'Irak vs. les "jobs" (Washington, Bush)
- Allez plus loin …
  - Suivi et veille technologique (Web)
  - Sémantique
  - Mesure de distance entre les discours politiques
  - Les mots "gagnants" ?

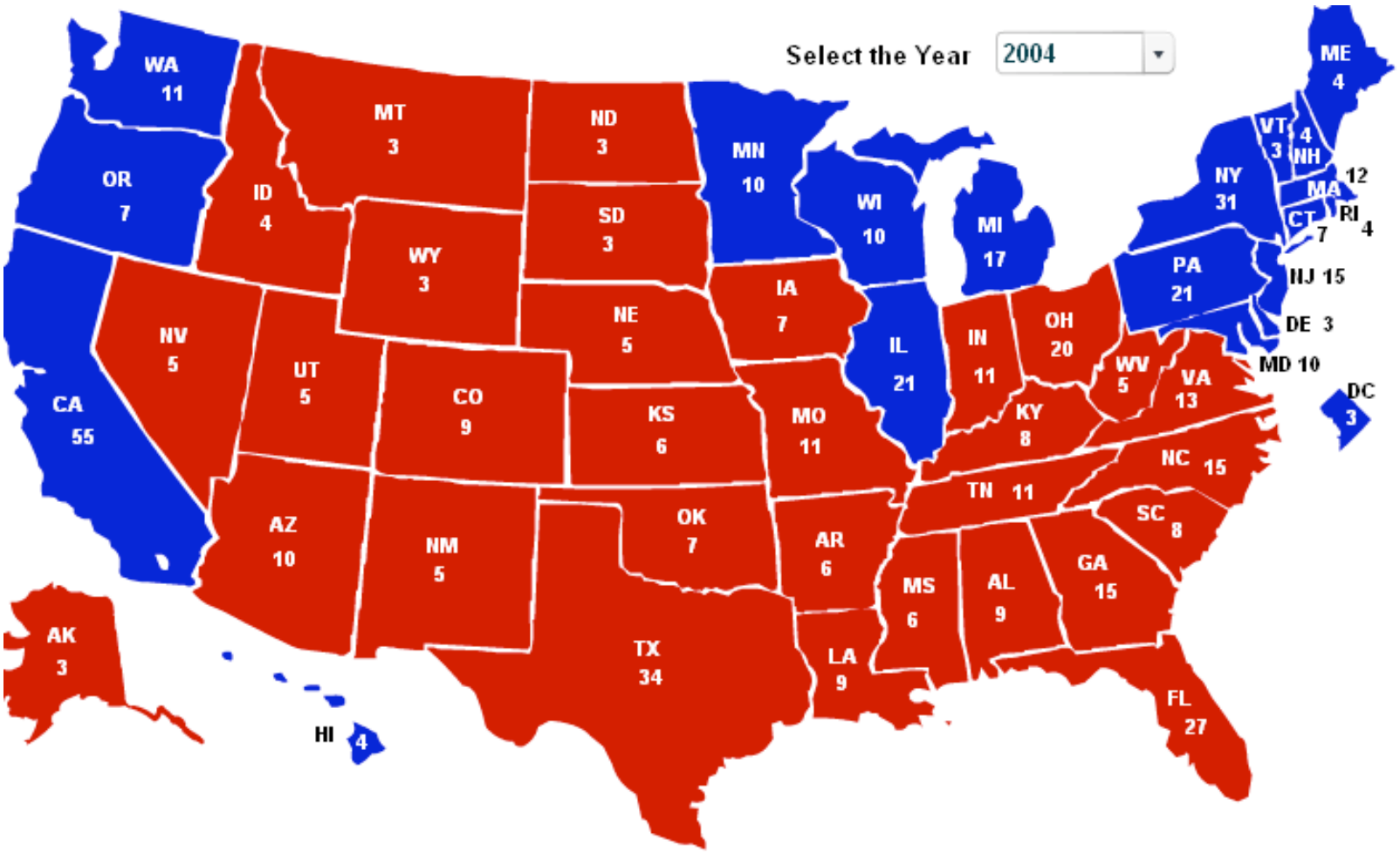
Les mots d'Obama et les maux de McCain



J. Savoy
Institut d'informatique
Université de Neuchâtel

Linguistique computationnelle & sciences politiques

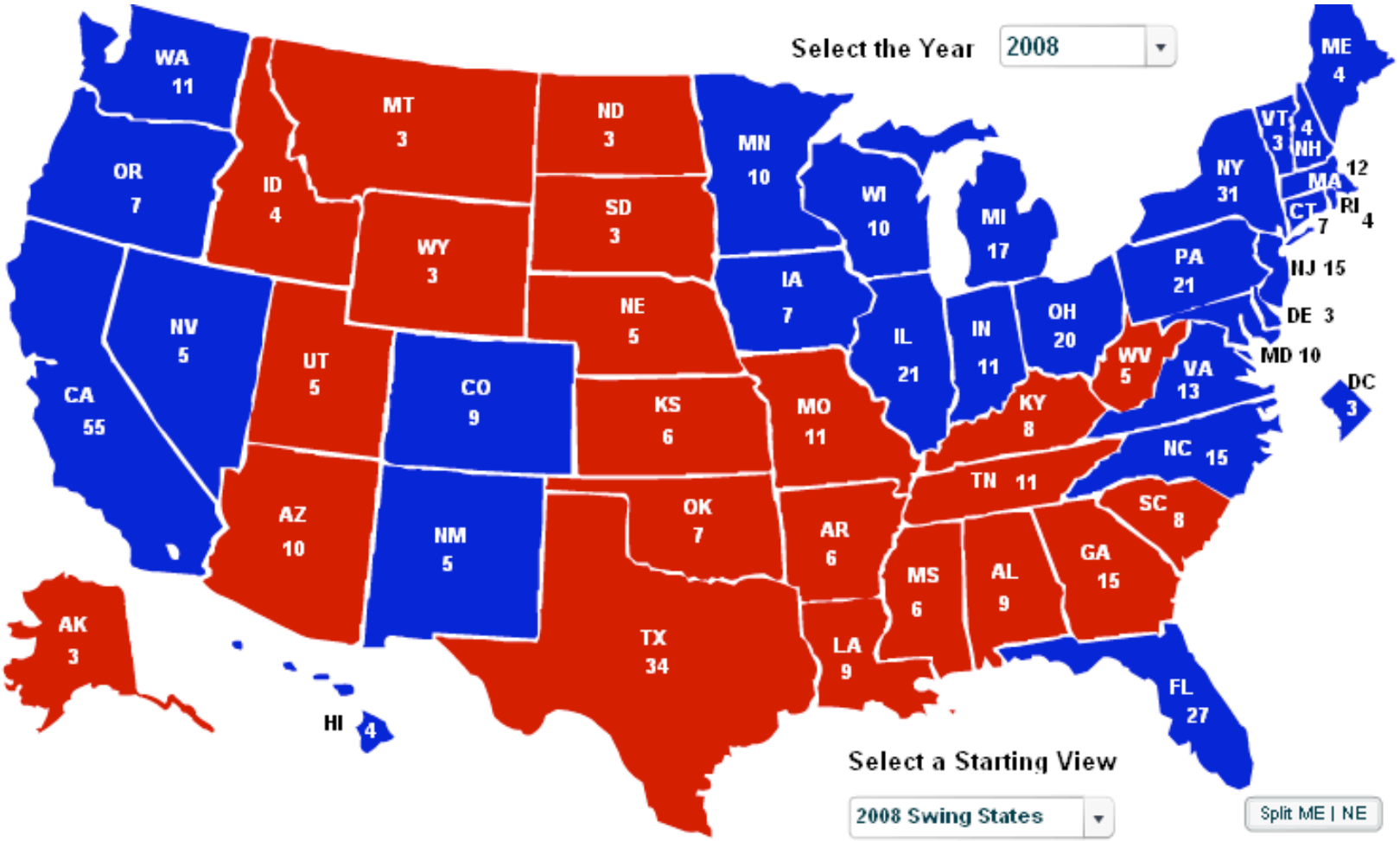
Grands électeurs 2004

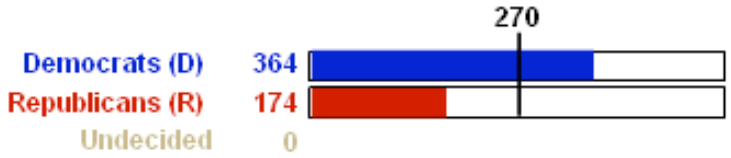


Republicans

Democrats

Grands électeurs 2008

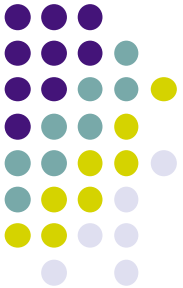




Democrats

Republicans

Bigrammes et *n*-grammes



L'importance de l'ordre des mots

Un vieillard en or avec une montre en deuil
Une reine de peine avec un home d'Angleterre
Et des travailleurs de la paix avec des gardiens de la mer
Un hussard de la farce avec un dindon de la mort
Un serpent à café avec un moulin à lunettes
Un chasseur de corde avec un danseur de têtes
Un maréchal d'écume avec une pipe en retraite
Un chiard en habit noir avec un gentleman au maillot
Un compositeur de potence avec un gibier de musique
…
Cortère de J. Prévert