De nouvelles perspectives dans l'analyse du langage

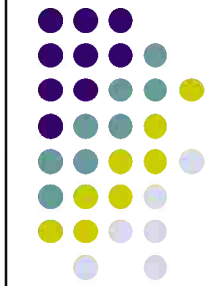
Prof. Jacques Savoy

chaire de linguistique computationnelle

Université de Neuchâtel

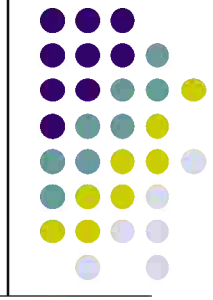
L'écrivain: automate de Jaquet-Droz (1774-1776)

Sommaire



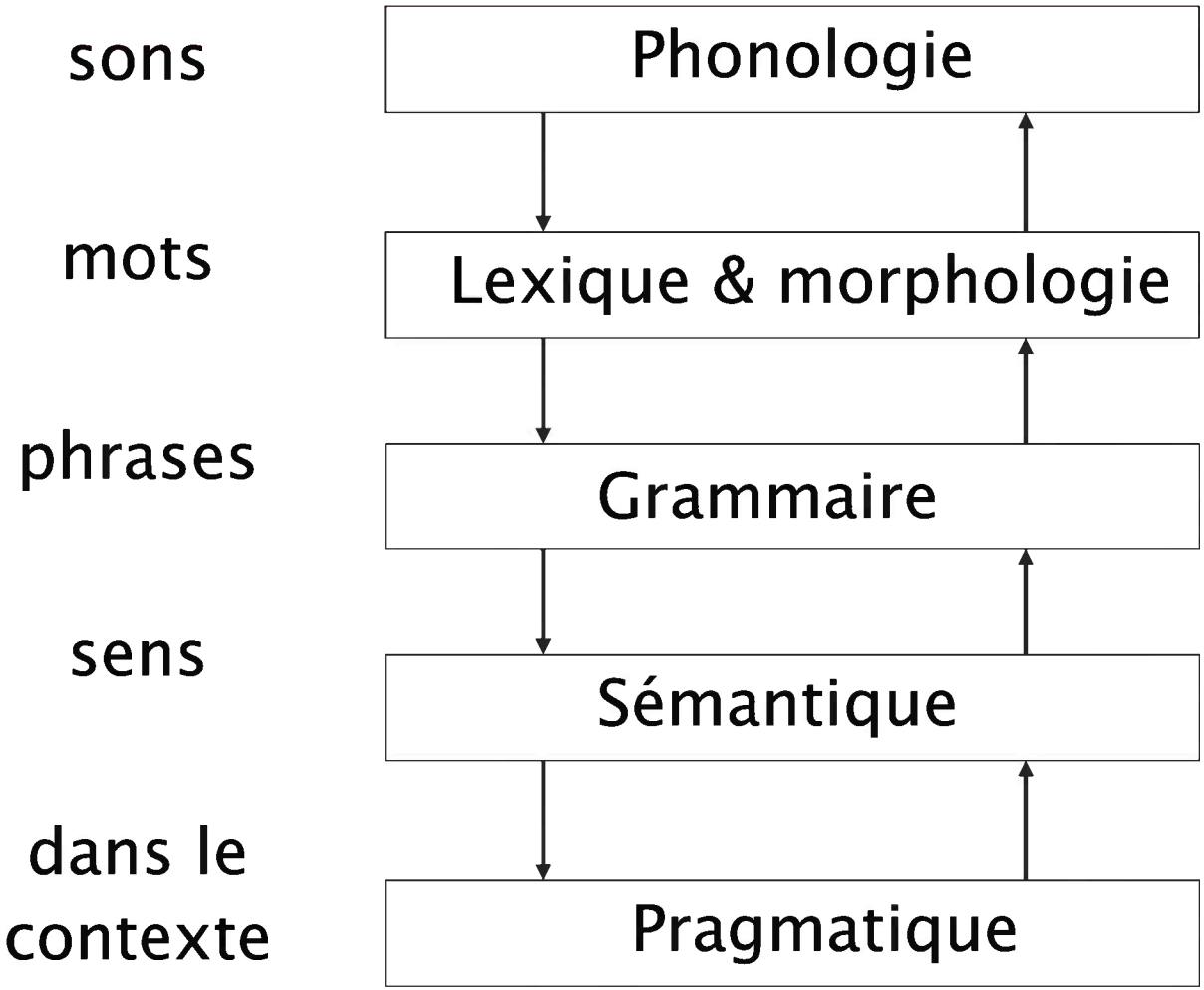
- **Linguistique et informatique**
- L'hortogaffe!
- Loi de Zipf
- Profilage et Cie
- Déterminer l'auteur d'un écrit ou d'un roman
- 2016 : Détection des menteurs (fake news)

Langage, Langue

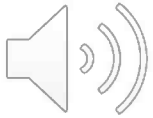


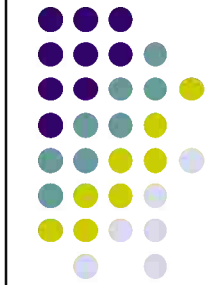
- Langage : système de communication le plus efficace mais il demeure complexe.
- Langue : mode d'expression propre à une communauté.
- Langue naturelle et artificielle.
- Propre à l'homme ?
  - et chez les animaux ?
  - et les plantes ?
- Formes
  - Seulement écrit et oral ?
- Linguistique et informatique ?

Domaines de la linguistique



Phonétique





Au Père Spicace,

Un grand malheur est arrivé à l'abbaye.

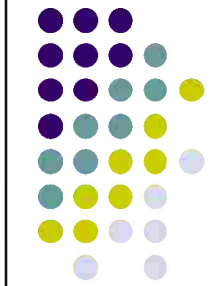
Pendant que l'abbé Nédictine donnait les dernières grâces, l'abbé Quille perdit l'équilibre dans l'escalier et tomba inanimé dans les bras du Père Iscope.

Un seul restait joyeux : le père Fide.

Quant à l'abbé Tise, il n'y comprenait rien. Il aurait bien voulu que le saint Plet l'aide à comprendre.

A la fin, le Père Nod et le Père Collateur servirent à boire et chacun pût se remettre de ses émotions.

Domaines de la linguistique



Phonétique : Cendrillon a un soulier en verre ou vair?

Lexique : Je suis le président.

je être/suivre le président

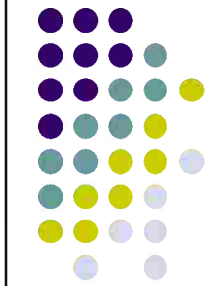
Syntaxe : sujet(je), verbe(suis), det(le), nom(président).

Sémantique : Etre ou suivre ?

Pragmatique : Qui est le suiveur ? Qui est le président ?

Exemple 2 : "On vit des évènements graves."

Synomymie / Polysémie



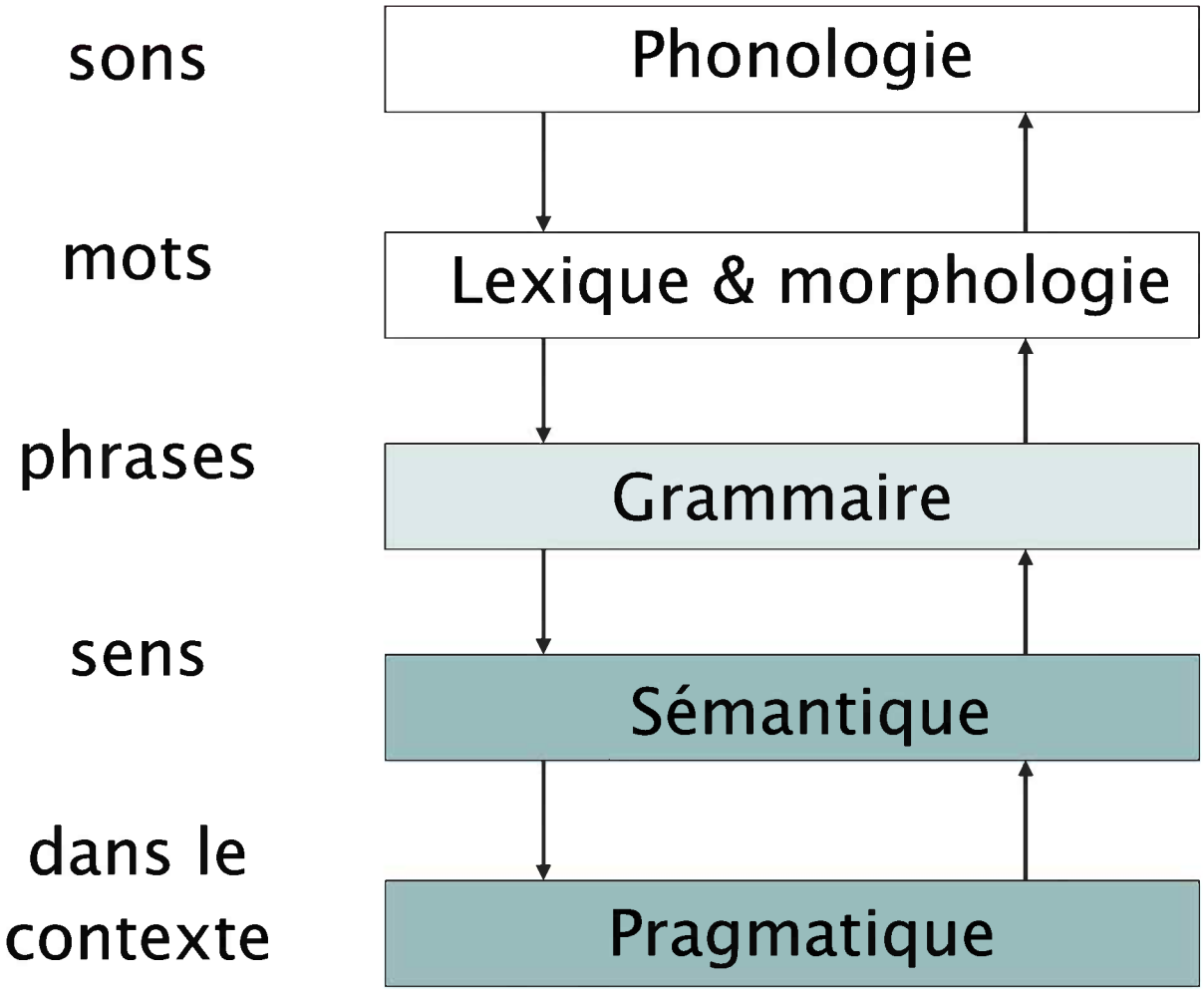
Références

Mr Major est arrivé en France aujourd'hui. Le premier ministre rencontrera le Président demain. Le leader du parti conservateur partira ensuite pour Moscou où il rencontrera Mr Gorbachev. Mme Major rejoindra son mari en Russie où ce fils d'un artiste de cirque est relativement inconnu.

Idiomes

Marcher sur des œufs (*to skate on a thin ice*).
Quand les poules auront des dents (*pigs might fly*).

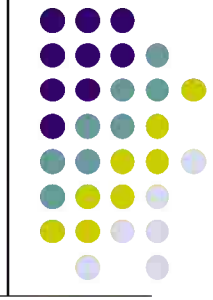
Domaines de la linguistique



Famille de langues

- Origine des langues
- Diversité des langues
- Famille de langues

Diversification des langues



5 000 à 6 000 langues vivantes

- 2 197 en Asie
- 2 092 en Afrique
- 1 310 en Océanie
- 1 002 en Amérique
- 230 en Europe.

Seulement 600 sont écrites.

80 % de la population mondiale parle 75 langues différentes.

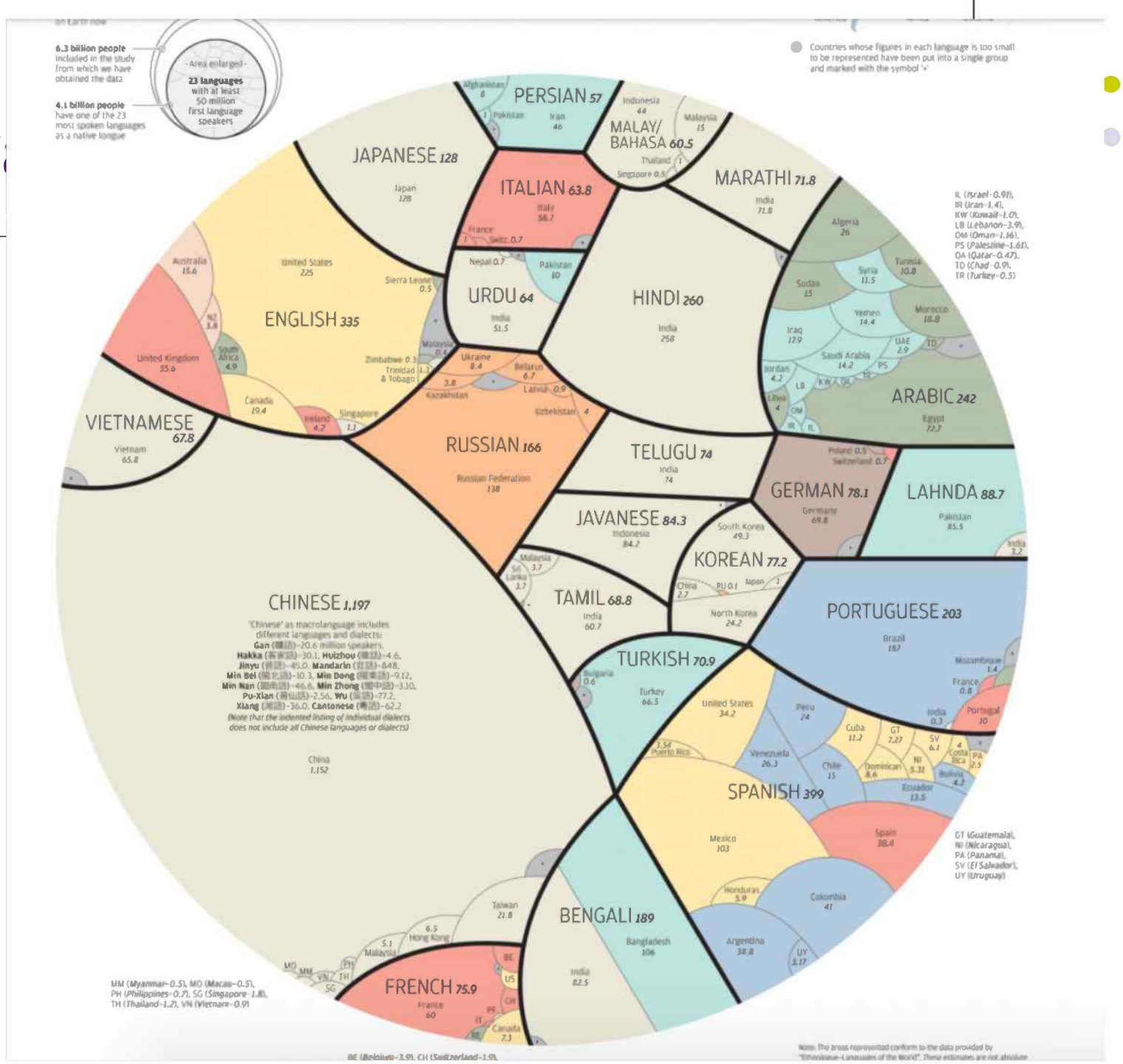
40 % de la population mondiale parle 8 langues différentes.

75 langues sont parlées par plus de 10 M personnes.

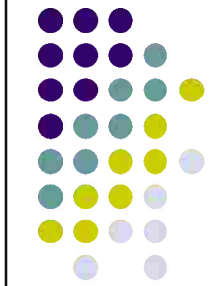
20 langues sont parlées par plus de 50 M personnes.

8 langues sont parlées par plus de 100 M personnes.

Motiv

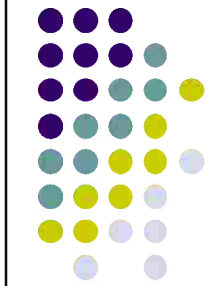

6.3 billion people included in the study from which we have obtained the data
4.1 billion people have one of the 23 most spoken languages as a native tongue
-Area enlarged-
23 languages with at least 50 million first language speakers
Countries whose figures in each language is too small to be represented have been put into a single group and marked with the symbol '+'
CHINESE 1,197
'Chinese' as macrolanguage includes different languages and dialects:
Gan (贛語)-20.6 million speakers.
Hakka (客家話)-30.1, Huizhou (徽語)-4.6,
Jinyu (晉語)-45.0, Mandarin (官話)-848,
Min Bei (閩北語)-10.3, Min Dong (閩東語)-9.12,
Min Nan (閩南語)-46.6, Min Zhong (閩中語)-3.10,
Pu-Xian (莆仙語)-2.56, Wu (吳語)-77.2,
Xiang (湘語)-36.0, Cantonese (粵語)-62.2
(Note that the indented listing of individual dialects does not include all Chinese languages or dialects)
China 1,152
SPANISH 399
ENGLISH 335
HINDI 260
ARABIC 242
PORTUGUESE 203
BENGALI 189
RUSSIAN 166
JAPANESE 128
PERSIAN 57
MALAY/ BAHASA 60.5
ITALIAN 63.8
MARATHI 71.8
URDU 64
TELUGU 74
GERMAN 78.1
LAHNDA 88.7
JAVANESE 84.3
KOREAN 77.2
TAMIL 68.8
TURKISH 70.9
VIETNAMESE 67.8
FRENCH 75.9
IL (Israel-0.9), IR (Iran-1.4), KW (Kuwait-1.0), LB (Lebanon-3.9), OM (Oman-1.16), PS (Palestine-1.61), QA (Qatar-0.47), TD (Chad-0.9), TR (Turkey-0.5)
GT (Guatemala), NI (Nicaragua), PA (Panama), SV (El Salvador), UY (Uruguay)
MM (Myanmar-0.5), MO (Macau-0.5), PH (Philippines-0.7), SG (Singapore-1.8), TH (Thailand-1.2), VN (Vietnam-0.9)
BE (Belgium-3.9), CH (Switzerland-1.9),

Identification d'une langue



1. Strč prst skrz krk
2. Mitä sinä teet?
3. Mam swoją książkę
4. Nem fáj a fogad?
5. Er du ikke en riktig nordmann?
6. Добре дошли в България!
7. 정보검색시스템
8. 我不是中国人
9. Fortuna caeca est

Evolution linguistique (main)



allemand	Hand
anglais	hand
danois	hånd
roumain	mine
italien	mano
espagnol	mano
polonais	ręka
russe	рука
irlandais	lámh
japonais	手

Famille indo-européenne

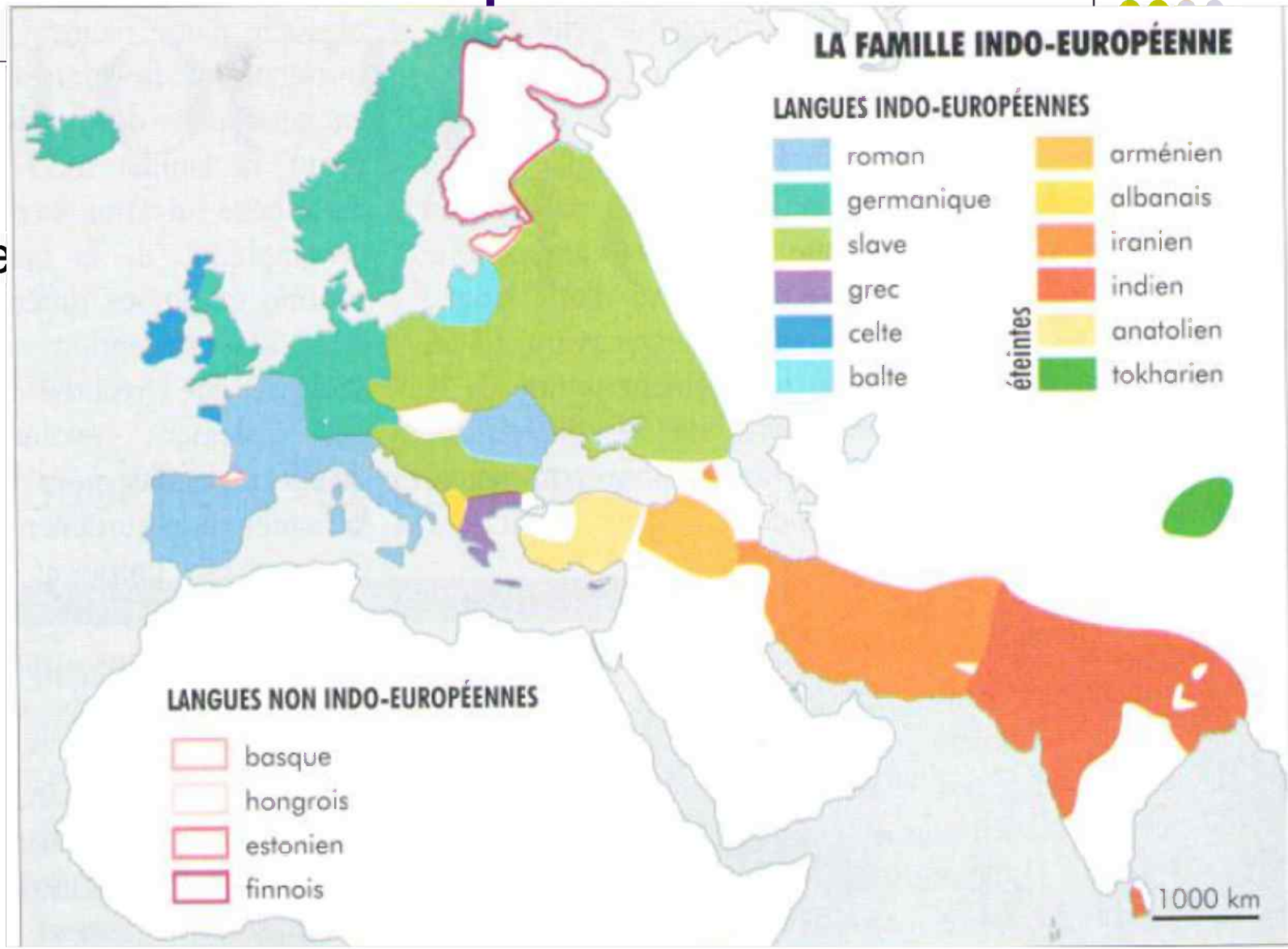
Chaque membre partage de similarités

Je / moi

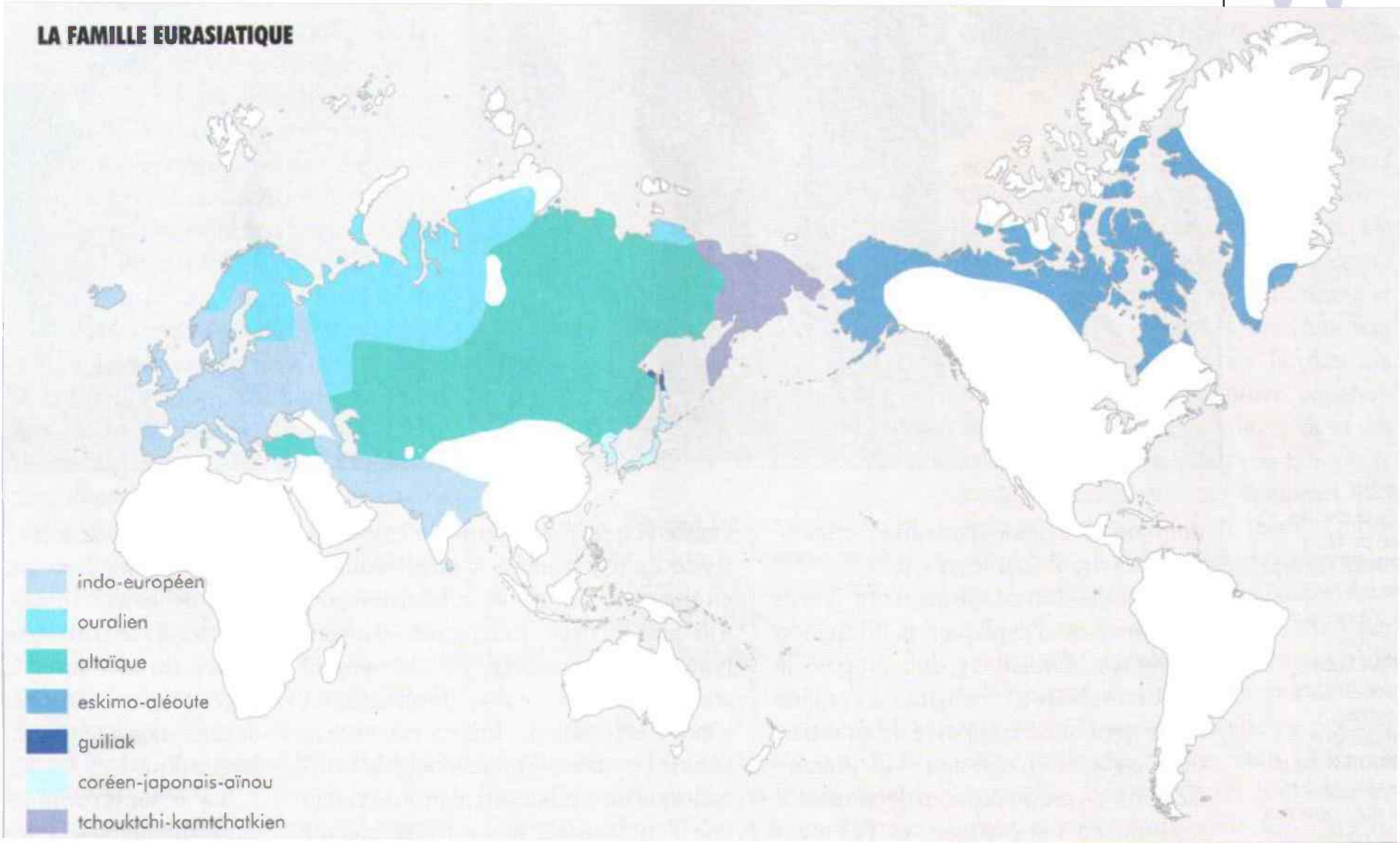
Tu / toi

Ich / mich

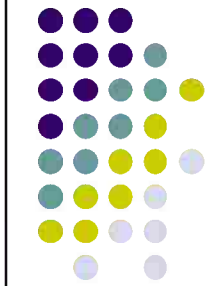
du / dich



Famille eurasiatique



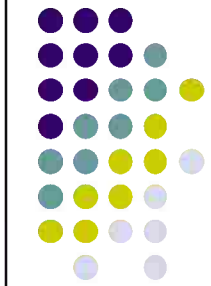
Le français



- Langue latine (espagnol, italien, portugais, …).
- "La plus germanique des langues romandes"
  Henriette Walter

- Proto-français
- *842 : Le Serment de Strasbourg*
- Français
- 1539 : Ordonnance de Villers-Cotterêts (François I^{er})
- 1635 : Académie française (Richelieu)
- 1694 : premier dictionnaire de l'Académie

Le français ne vit pas isolé

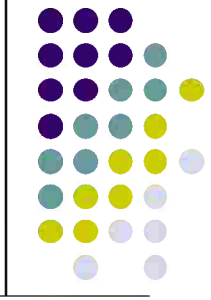


Emprunt de mots d'autres langues

- Gaulois : crème, benne, bouc, galet, luge, …
- Allemand : maçon, espion, hareng, bonnet, gigot, …
- Italien : ducat, florin, piastre, alarme, sentinelle, bille, …
- Anglais : interview, cash, rail, pedigree, sport, …
- Russe : zibeline, vodka, tsar, …
- Hongrois : sabre, goulasch
- Japonais : karaté, judo, sushi, …

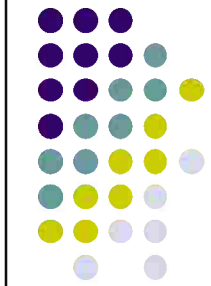
- Et "trop" et "guère" viennent de ?

Sommaire



- Linguistique et informatique
- **L'hortogaffe**!
- Loi de Zipf
- Profilage et Cie
- Déterminer l'auteur d'un écrit ou d'un roman
- 2016 : Détection des menteurs (fake news)

L'orthographe



- La communication est facilitée par une bonne orthographe.
- Signe de sérieux (compétence).
- Complexité de la relation entre la phonétique et la graphie dans certaines langues (dont le français ou l'anglais).

- Intérêt relativement récent (Révolution française).

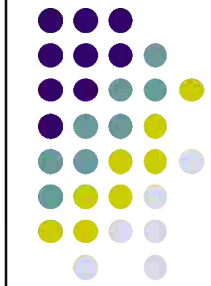
- Favorisé en France.
- Mais en perdition ?

L'orthographe XVIIIe

Madame de Sévigné (XVIIIe)

"Monsieur vous me permettres de souhaitter la paix car ie trouve uec vostre permission quune heure de Conuersation vaut mieux que cinquante lettres, quand vous seres icy etque iauray lhonneur devous voir ievous ferey demeurer dacort quela querre est vne fort sotte chose …"

Sur les réseaux sociaux



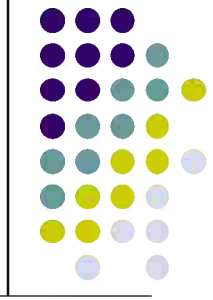
Toi t'est un qui n'a rien vue

Toi t'est un qui n'a rien vue contre le FCC. Je te rappel que contre le fcc sur 11 joueur sur le terrain il y avait que 4 joueurs licencié. Sinon c'était des essais pour voir le niveau des nouveaux arrivées.

FC Bosna

fc bosna j aisspaire que momo revienne sest le seule qui peux nous faire monté comme entreneure

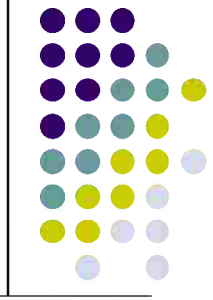
Est-ce si simple ?



- Des règles simples
- Correspond à l'étymologie
- Invariable dans le temps

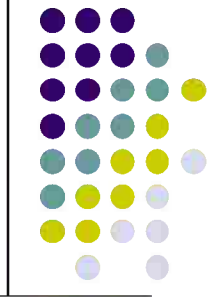
- En perdition avec les réseaux sociaux ?
- "Avec leurs acronymes, smileys et émoticons, les jeunes ne connaissent plus l'orthographe!"

Variations



- Variantes possibles :
  - yogurt, yoghurt, yoghourt.
  - gnôle, gnole, gniole, gnaule, niaule.
- Avec des noms propres
  - Oscar vs. Oskar.
  - Taylor vs. Tailor.
  - Katherine vs. Katherin vs. Catherine vs. …
  - Shaksper vs. Shakspere vs. Shakspeare vs. Shakespeare, …

Variations temporelles



- Orthographe correcte se modifie :

  arcs-boutants (1694)

  arc-boutants (1718)

  arcs-boutants (1740)

  arc-boutants (1762)

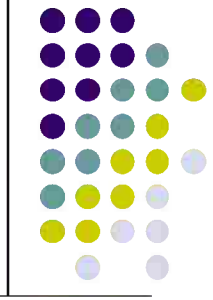
  arcs-boutants (1835)

- L'influence des grands écrivains

  Chinois & Anglois → Anglais (Voltaire)

F. de Closets: Zero faute. L'orthographe, une passion française. Mille et Une Nuits, Paris, 2009.

Etymologie

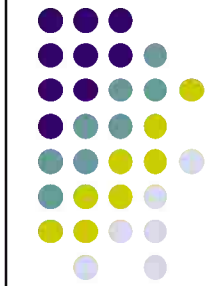


- Relation pas évidente…
  - hauteur → altus
  - pomme → poma
  - huile → olea (et les autres langues latines)
- Accents et prononciation (pas toujours)
  - pôle et police → pas de différence
  - pêcheurs et pécheurs → pas de différence
- Accent indique une lettre perdue
  - âne → asinus
  - théâtre → theatrum
- Arbitraire (a**p**ercevoir et a**pp**araître)

Réforme de l'orthographe

Ancienne	Nouvelle	Ancienne	Nouvelle
pique-nique	piquenique	a priori	apriori
porte-clé	porteclé	statu quo	statuquo
auto-stop	autostop	hot-dog	hotdog
mille-feuille	millefeuille	cow-boy	cowboy
ping-pong	pingpong	week-end	weekend
québecois	québécois	crémerie	crèmerie
événement	évènement	diesel	diésel
edelweiss	édelweiss	revolver	révolver
veto	véto		
chariot	charriot	combatif	combattif
eczéma	exéma	nénuphar	nénufar
relais	relai		

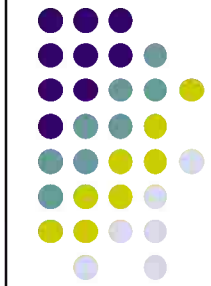
Et la grammaire



Accord du participe passé

sans auxiliaire	les chats **lavés** mangent.
avoir (c.o.d.)	Vous avez **envoyé** une lettre. Je l'ai bien **reçue.**
avec infinitif	L'histoire que j'ai **entendu** *raconter*.
	La cantatrice que j'ai **entendue** *chanter*.
	(les deux sont correctes, dès 1990)
avec en (invar.)	J'ai **cueilli** des fraises et j'*en* ai **mangé**.
	Ses ordres, s'il *en* a **donnés** ne me sont pas **parvenus**.
être (sujet)	Les chats sont **lavés.**
"se" est c.o.d.	Les chats se sont **lavés.**
"se" est c.o.ind.	Les chats se sont **lavé** les pattes.
que faire ?	Les chats de Marie que Jean a **vus** (**vue**) sont **lavés.**
	Jean a vu les chats (ou Marie)?

Linguistique computationnelle



Trouver et analyser des régularités dans la langue.

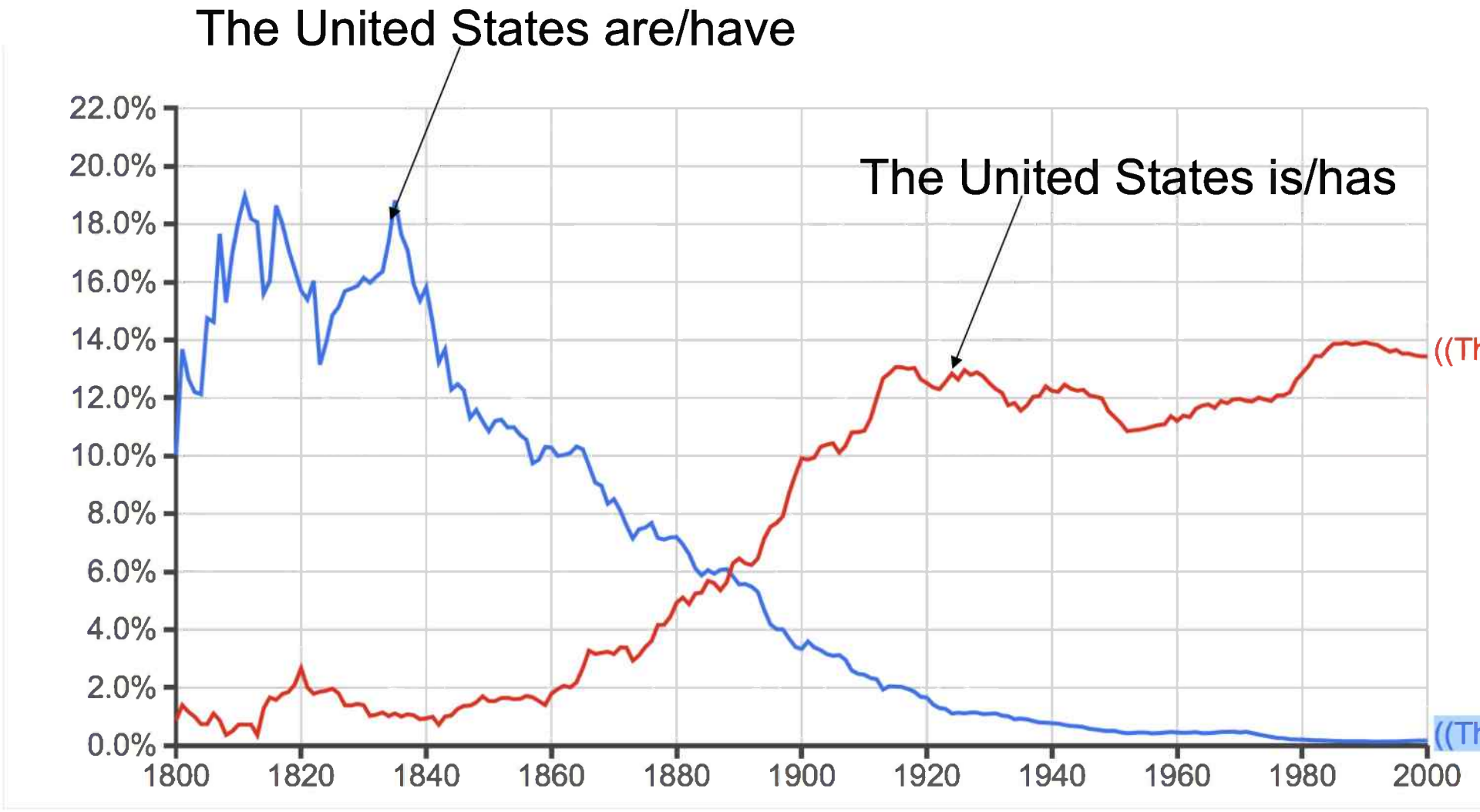
Aspects plus quantitatifs.

Disposer d'un corpus pour l'analyse

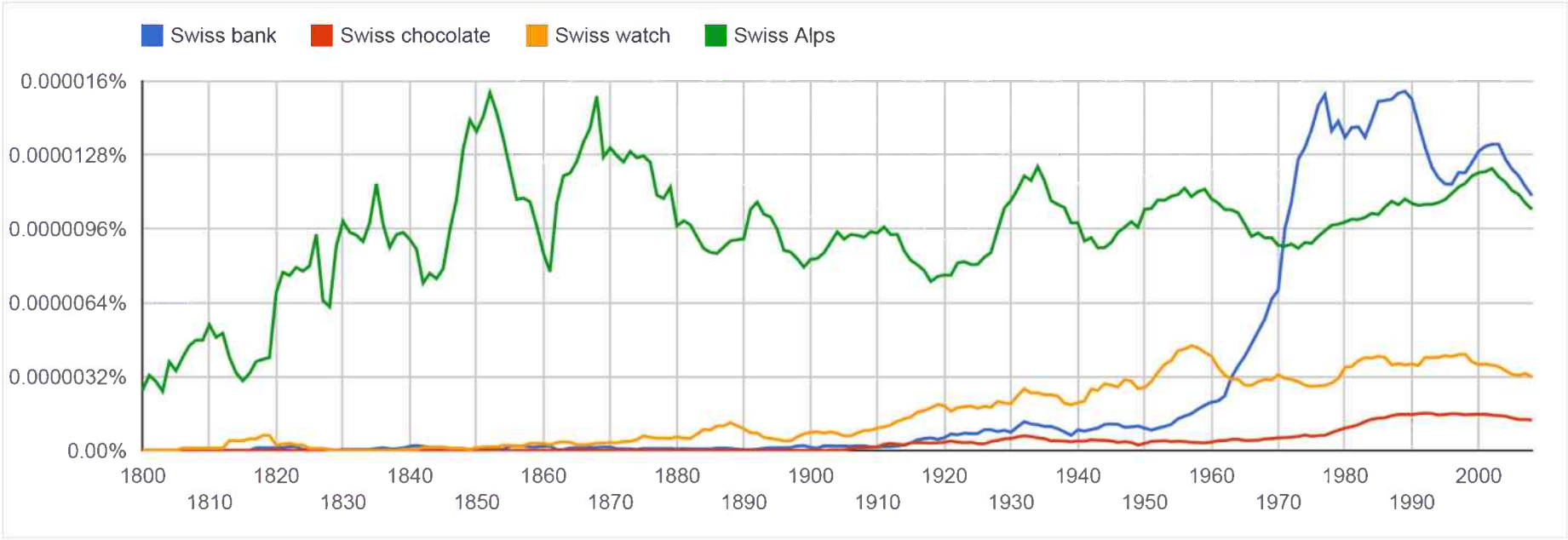
- Journaux
- Débats parlementaires
- Réseaux sociaux

- Google N-Gram

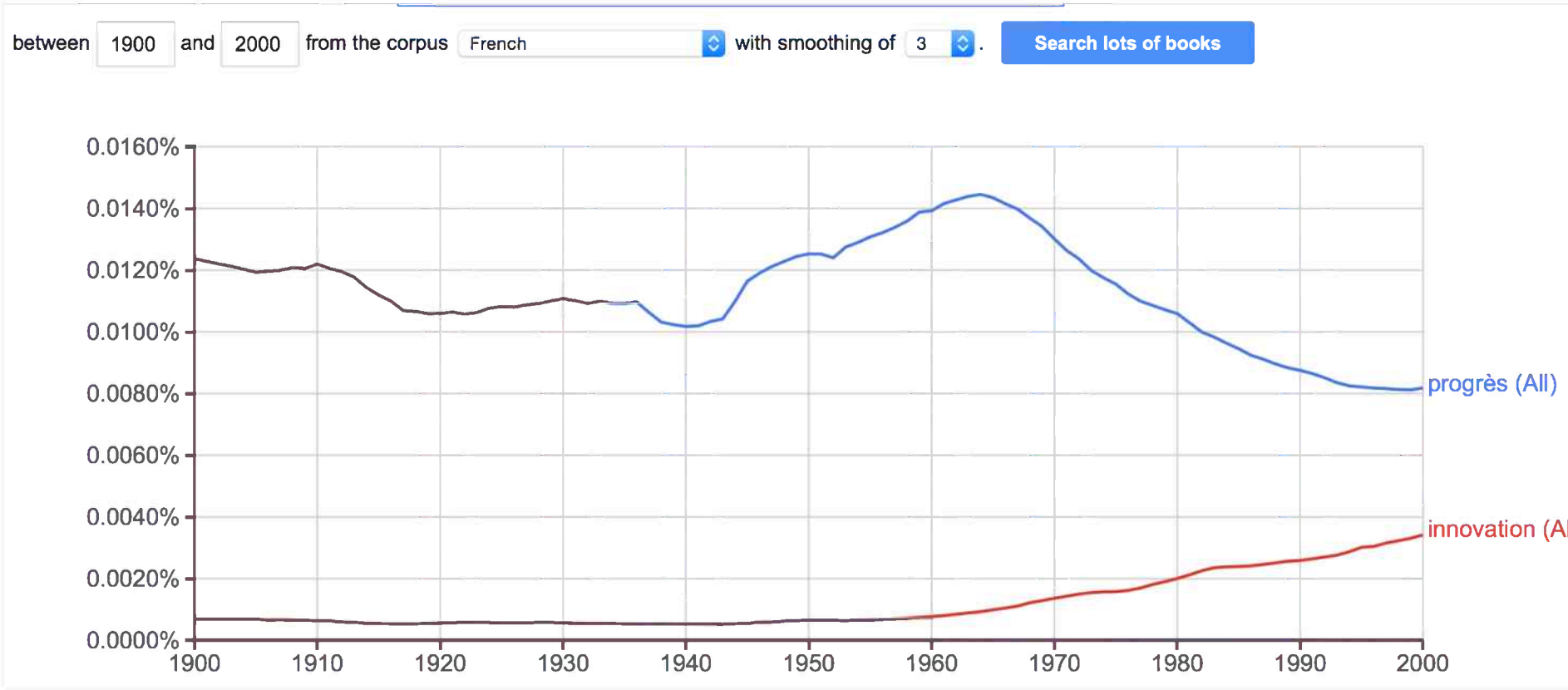
Google N-Gram Viewer



Google N-Gram Viewer



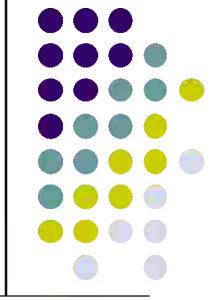
Idée de progrès



Et Internet

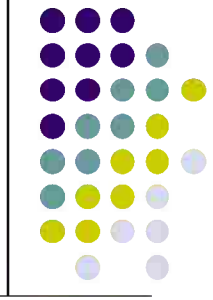

Google Books Ngram Viewer
Graph these comma-separated phrases: internet
case-insensitive
between 1800 and 1900 from the corpus French with smoothing of 3.
Search lots of books
0.00000120%
0.00000110%
0.00000100%
0.00000090%
0.00000080%
0.00000070%
0.00000060%
0.00000050%
0.00000040%
0.00000030%
0.00000020%
0.00000010%
0.00000000%
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
internet
(click on line/label for focus)

Sommaire



- Linguistique et informatique
- L'hortogaffe!
- **Loi de Zipf** (une langue étrangère, c'est très simple!)
- Profilage et Cie
- Déterminer l'auteur d'un écrit ou d'un roman
- 2016 : Détection des menteurs (fake news)

La loi de Zipf



Classons les mots d'après leur fréquence d'occurrence,
du plus fréquent au moins fréquent.

Quel mot est le plus fréquent en français ?

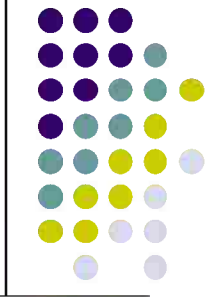
Une illustration du principe du moindre effort.
Ces mots sont si fréquents qu'ils passent inaperçus.

Rang	Mot	Fréquence rel.	cumulée
1	de	5,48 %	5,48 %
2	la	3,08 %	8,56 %
3	l	2,35 %	10,91 %
4	le	2,34 %	13,25 %
5	à	1,98 %	15,23 %
6	les	1,90 %	17,13 %
7	et	1,72 %	18,85 %
8	des	1,69 %	20,54 %
9	d	1,59 %	22,14 %
10	en	1,38 %	23,52 %
11	du	1,27 %	24,79 %
12	a	1,19 %	25,98 %
13	un	1,14 %	27,12 %
14	une	0,97 %	28,09 %
15	est	0,81 %	28,90 %
16	dans	0,77 %	**29,67** %

Rang	Mot	Fréquence rel.	cumulée
1	der	3,46 %	3,46 %
2	die	3,44 %	6,90 %
3	und	2,13 %	9,02 %
4	in	1,78 %	10,80 %
5	den	1,29 %	11,93 %
6	von	0,95 %	12,88 %
7	das	0,95 %	13,84 %
8	mit	0,84 %	14,68 %
9	im	0,81 %	15,49 %
10	zu	0,79 %	16,28 %
11	für	0,76 %	17,05 %
12	des	0,70 %	17,75 %
13	auf	0,69 %	18,43 %
14	sich	0,65 %	19,09 %
15	dem	0,62 %	19,70 %
16	ein	0,60 %	20,30 %

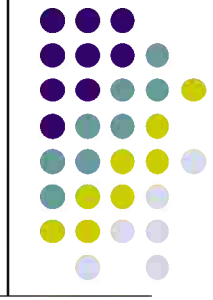
Rang	Mot	Fréquence rel.	cumulée
1	il	6,0 %	6,0 %
2	di	3,18 %	9,18 %
3	e	2,81 %	12,03 %
4	essere	2,80 %	14,82 %
5	che	2,48 %	17,31 %
6	a	1,90 %	19,21 %
7	avere	1,87 %	21,08 %
8	un	1,69 %	22,77 %
9	del	1,64 %	24,41 %
10	non	1,61 %	26,02 %
11	in	1,29 %	27,31 %
12	al	1,09 %	28,40 %
13	per	1,05 %	29,45 %
14	si	1,04 %	**30,49 %**
15	una	1,04 %	31,53 %
16	con	0,84 %	32,37 %

Sommaire



- Linguistique et informatique
- L'hortogaffe!
- Loi de Zipf
- **Profilage et Cie**
- Déterminer l'auteur d'un écrit ou d'un roman
- 2016 : Détection des menteurs (fake news)

Analyse des écrits



Peut-on distinguer un écrit rédigé par un homme de celui écrit par une femme ?

Par exemple, écrivez le récit de votre dernier voyage.

Peut-être cela n'est pas vraiement possible…

Examinons quelques cas extraits des blogs (votre journal personel sur le Web).

Exemple : Ecrit par ?

Un homme ou une femme ?

et pourquoi ?

Hier nous avons eu notre deuxième compétition de jazz. Dieu merci on n'a pas été évalué... Nous étions si mauvais. Comme, j'avais tellement honte, je ne voulais même plus parler à personne après. Je me sentais tellement en colère, et je voulais pleurer, mais ... bon c'est passé.

Exemple : Ecrit par ?

Un homme ou une femme ?

et pourquoi ?

Hier *nous* avons eu notre deuxième compétition de jazz. Dieu merci on *n'a pas* été évalué... *Nous* étions si *mauvais*. Comme, *j'*avais tellement honte, *je ne* voulais même plus parler à personne après. *Je me* sentais tellement en *colère*, et je voulais *pleurer*, mais ... bon c'est passé.

Exemple : Ecrit par ?

Un homme ou une femme ?

et pourquoi ?

Mon super patron avait accepté de me laisser une semaine de «travail». Il m'a finalement rendu mon rapport après huit jours fous! Maintenant, je n'ai que le reste de la semaine et une semaine entière après mes vacances pour finir ce fichue travail.

Exemple : Ecrit par ?

Un homme ou une femme ?

et pourquoi ?

Mon super *patron* avait accepté de me laisser une *semaine* de «*travail*». Il m'a finalement rendu mon *rapport* après huit *jours* fous! Maintenant, je n'ai que le *reste* de la *semaine* et une *semaine* entière après mes *vacances* pour finir ce fichue *travail*.

Genre (Schler *et al.*, 2005)

Termes	Hommes	Femme
travail	68.1 ± 0.6	56.5 ± 0.5
argent	43.6 ± 0.4	37.1 ± 0.4
sports	31.2 ± 0.4	20.4 ± 0.2
tv	21.1 ± 0.3	15.9 ± 0.2
sexe	32.4 ± 0.4	43.2 ± 0.5
famille	27.5 ± 0.3	40.6 ± 0.4
nourriture	23.9 ± 0.3	30.4 ± 0.3
amis	20.5 ± 0.2	25.9 ± 0.3
dormir	18.4 ± 0.2	23.5 ± 0.2
pos-émotions	248.2 ± 1.9	265.1 ± 2
nég-émotions	159.5 ± 1.3	178 ± 1.4

Fréquence (dans les blogs) pour 10 000 mots (et erreur standard)

Genre

Femmes	Hommes
Pronoms	Articles
Négations	Adjectifs
Emotions	Noms
Mots cognitifs	Prépositions
Mots sociaux	Mots longs
Verbes	Nombres
et, dans, pour, avec	Gros mots
	Technologies

Différences possibles, mais le rapport signal / bruit est faible !

Plagiat ?


SARKOZY SORT UN LIVRE
LE PARI DE LA SINCÉRITÉ
... CE N'EST PAS MOI QUI L'AI ÉCRIT !

Plagiat ?

Un vrai problème…

●●○○○ Swisscom 08:33 87%

News

Suisse Il y a 10h

Uni: les «ghostwriters» ont toujours autant la cote

Un nombre croissant d'universitaires en Suisse engagent une tierce personne pour qu'elle rédige leurs travaux scientifiques. Consciente du problème, l'Uni de Saint-Gall vient de déposer plainte.

Plagiat ?

J.F. Kennedy, 1961

“Ask not what your country can do for you, ask rather what you can do for your country”.

W. Harding, 1923

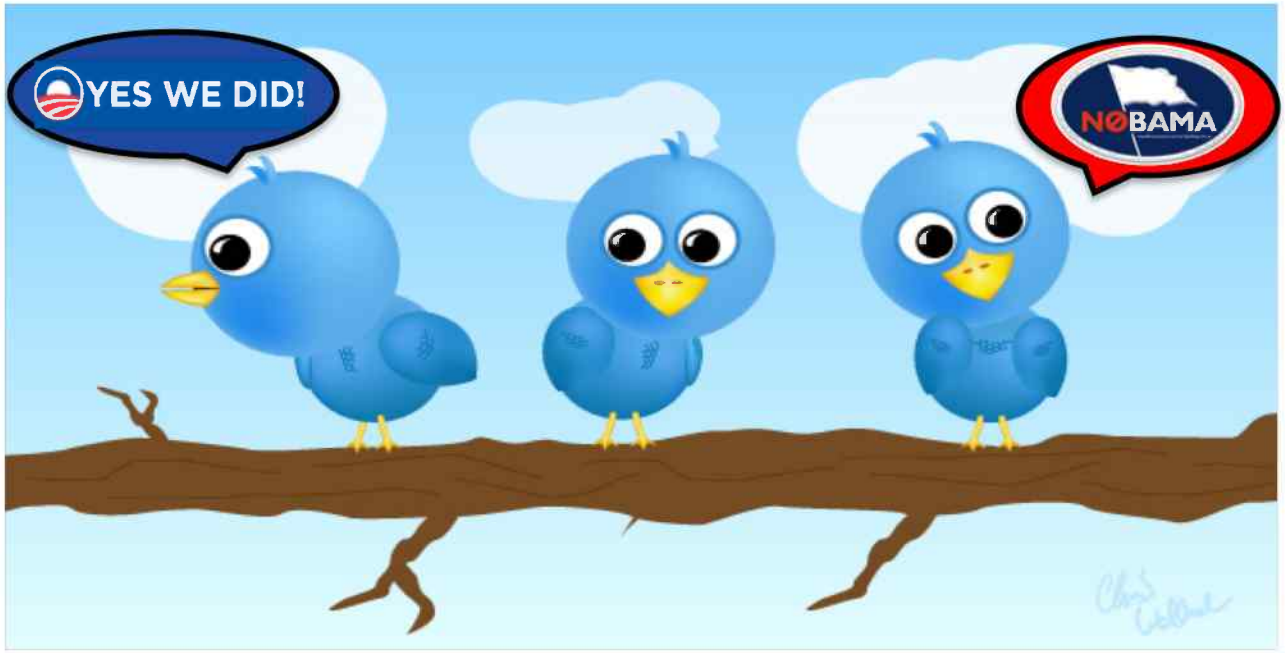
“We need to be thinking not so much of what the country can do for us but what we can do for our country”.

P. St. John, 1930s (headmaster of Choate)

“It’s not what Choate does for you, but what you can do for Choate”.

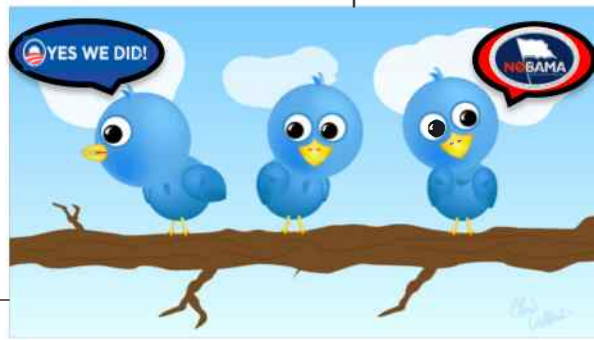
J. Humes: Confessions of a White House Ghostwriter: Five Presidents and Other Political Adventures. Regnery Publ., 1997.

Les réseaux sociaux



Sylwester, K., & Purver, M. 2015. *Twitter Language Use Reflects Psychological Differences between Democrats and Republicans*. PLoS

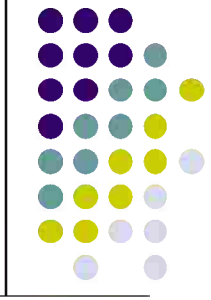
Social Networks



	Republicans	*Democrats*
I, me (je)		+
We, our (nous)	+	
jurons		+
émotions positives		+
émotions négatives	=	=
religion	+	
sentiments		+

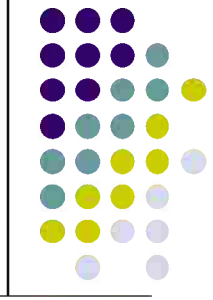
Sylwester, K., & Purver, M. 2015. *Twitter Language Use Reflects Psychological Differences between Democrats and Republicans*. PLoS

Sommaire



- Linguistique et informatique
- L'hortogaffe!
- Loi de Zipf
- Profilage et Cie
- **Déterminer l'auteur d'un écrit ou d'un roman**
- 2016 : Détection des menteurs (fake news)

Attribution d'auteur



- Etant donné un ensemble de textes dont les auteurs sont certains (disons que l'on a des textes de quatre auteurs distincts), on dispose d'un nouveau texte (un roman, une pièce de théâtre) dont l'auteur est inconnu (ou douteux), pouvez vous me donnez le véritable nom de ce nouveau texte (extrait ou non des 4 auteurs que l'on dispose).

- Lettre ou courriel menaçant
- Revendication d'un attentat
- Testament
- Œuvre littéraire (le cas Elena Ferrante)

Qui est l'auteur ?

Comme auteurs possibles : John F. Kennedy, B. Obama, A. Lincoln. Attribuez chaque texte à son auteur.

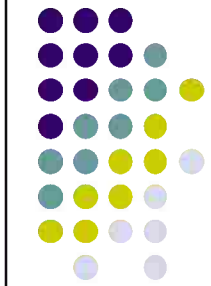
Texte 1: "Il y a quatre-vingt sept ans, nos pères donnèrent naissance sur ce continent à une nouvelle nation conçue dans la liberté et vouée à la thèse selon laquelle tous les hommes sont créés égaux."

Texte 2: "Yes, we can."

Texte 3: "Mes chers compatriotes, ne demandez pas ce que votre pays peut faire pour vous, demandez ce que vous pouvez faire pour votre pays."

Texte 4: "Ich bin ein Berliner."

A la poursuite d'Elena…



Problème classique en attribution d'auteur :
Etant donné un texte, qui l'a écrit ?

Avec Elena Ferrante, un intérêt mondial.

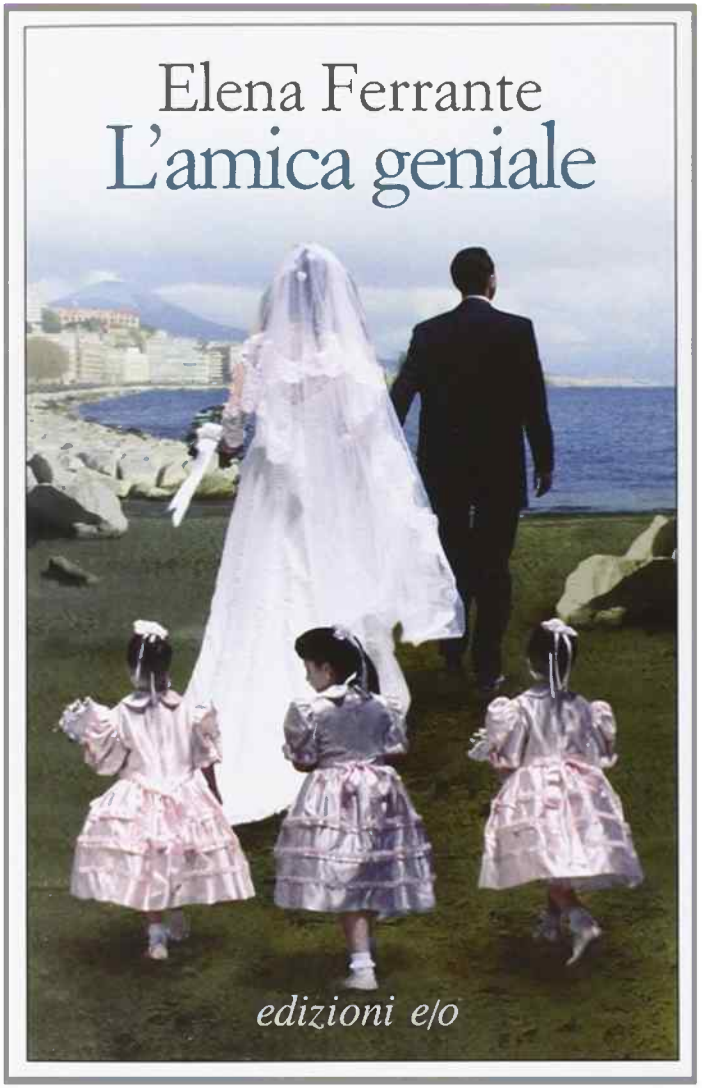
1992: Le premier roman: *L'amore molesto* (film en 2007)
2014 : *L'Amica geniale*

A la poursuite d'Elena…

Problème classique en attribution d'au
Etant donné un texte, qui l'a écrit ?

Avec Elena Ferrante, un intérêt mond

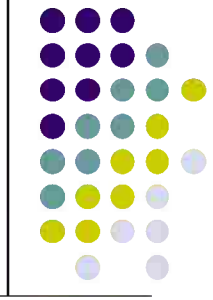
1992: Le premier roman: *L'amore mo*
2014 : *L'Amica geniale*



L'affaire Molière-Corneille

- Jean Baptiste Poquelin (1622-1673)
- 1645-1659 (14 ans).
  années difficiles, production faible.
- 1659-1673 (14 ans) production abondante, comédien, directeur du théâtre du Roi.
- 1658 Corneille & Molière à Rouen
- *Psyché* (1671), pas de doute, les deux auteurs ont écrits ensemble.
- Le droit d'auteur a un sens différent.
- Pierre Louys (octobre 1919) sur le style et la versification.

Attribution d'auteur



- Notre problème : retrouver le véritable auteur d'un roman ou pièce de théâtre.
- Chaque auteur a-t-il /elle un style personnel ?
- Pourquoi utiliser un nom de plume ?
- Dans quel contexte ?

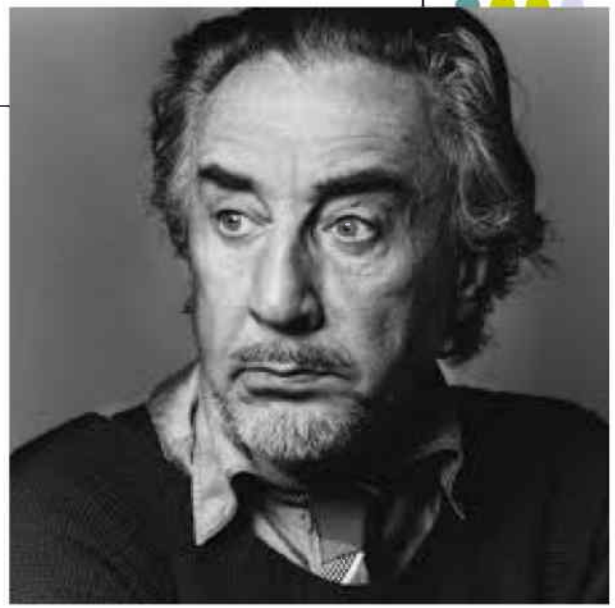
- Quelques cas en littérature (Ferrante n'est pas unique).
- L'affaire Molière-Corneille
- Shakespeare
- et plus récemment…

Shakespeare

- Est-ce que Shakespeare a bien écrit toutes les pièces ?
  - Plusieurs auteurs ont été nommés Bacon, Marlowe ou Giovani Florio.
  - “Shakespeare” est peut-être un nom-de-plume pour un groupe d'écrivains ?
- Pièces écrite par plus d'un auteur
  - *Two Noble Kinsmen* – Shakespeare & Fletcher
  - *Edward III* – Shakespeare? & Kyd?
  - *Titus Andronicus* – Shakespeare & Peele?
  - *Henry VIII* – Shakespeare & Fletcher?
  - *Timon of Athens* - Shakespeare & Fletcher?

Craig H., Kinney A.F.: Shakespeare, Computers, and the Mystery of Authorship. Cambridge University Press, 2009.

Gary – Ajar en France



- Romain Gary (1918-1980)
  Romancier français (Roman Kacew).
- Emile Ajar apparaît en 1973 avec
  le roman *Gros-Câlin*
  “un nouvel style” rapporte la presse.
- 1974: débute la recherche pour trouver qui est “Ajar”.
- 1975: Ajar reçoit le prix Goncourt (*La Vie devant soi*).
- 1975: le cousin de Gary est-il “Ajar” ?
- 1980: R. Gary écrit qu'il est le véritable E. Ajar.

Le cas Galbraith–Rowling

- Depuis 2007, on dit que J.K. Rowling (Harry Porter) désire écrire un roman policier.
- *The Cuckoo's Calling* est publié le 4 avril 2013 sous le nom de Robert Galbraith.
- Qui est ce R. Galbraith?
- Le 13 juillet 2013, *The Sunday Times* indique que J.K. Rowling est le véritable auteur de ce roman.

Style et ordinateur

Lemma	Carofiglio	De Luca	Ferrante	Starnone
il	4,18 %	**5,86 %**	**4,33 %**	4,55 %
di	2,82 %	2,80 %	2,55 %	2,56 %
e	2,43 %	2,31 %	2,13 %	2,14 %
essere	**2,65 %**	2,11 %	**2,07 %**	2,06 %
che	2,15 %	1,59 %	2,36 %	2,10 %
a	1,44 %	1,87 %	1,92 %	1,65 %
avere	1,73 %	1,11 %	1,54 %	1,70 %
un	1,52 %	1,60 %	1,12 %	1,25 %
del	1,25 %	1,31 %	1,10 %	1,21 %
non	1,52 %	1,47 %	1,42 %	1,23 %

Distance de Labbé

Texte A
la bocca del lupo.

la: 1
bocca: 1
del: 1
lupo: 1

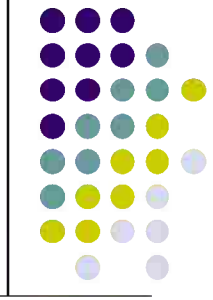
Texte B
la bocca del leone.

la: 1
bocca: 1
del: 1

leone: 1

$$Dist(A,B) = (|1-1| + |1-1| + |1-1| + |1-0| + |0-1|) / (2 \cdot 4) = 2 / 8 = 0,25$$

Distance de Labbé



- Chaque roman correspond à un texte.
- Prendre tous les mots ayant une fréquence > 2.
- Les romans singées E. Ferrante.

Id	Année	Titre
46	1992	L'amore molesto
47	2002	I giorni dell abbandono
48	2006	La figlia oscura
49	2011	*L'amica geniale*
50	2012	*Storia del nuovo cognome*
51	2013	*Storia di chi fugge e di chi resta*
52	2014	*Storia della bambina perduta*

Distance de Labbé (lemmes)

Rang	Distance	Id	Titre	Id	Titre	
1	**0.111**	51	Ferrante	52	Ferrante	
2	**0.121**	50	Ferrante	51	Ferrante	
3	**0.128**	49	Ferrante	50	Ferrante	
4	**0.134**	50	Ferrante	52	Ferrante	
5	**0.142**	145	Veronesi	147	Veronesi	
6	**0.146**	42	Faletti	44	Faletti	
7	**0.150**	43	Faletti	44	Faletti	
8	**0.154**	41	Faletti	42	Faletti	
9	**0.157**	42	Faletti	43	Faletti	
10	**0.161**	38	De Silva	39	De Silva	
11	**0.161**	49	Ferrante	51	Ferrante	
…						
33	**0.193**	52	Ferrante	132	Starnone	**<-**
38	**0.195**	51	Ferrante	131	Starnone	**<-**
41	**0.196**	51	Ferrante	132	Starnone	**<-**
…						
84	**0.216**	25	Carofiglio	147	Veronesi	**Erreur**

Mais un modèle a ses limites …

Le classement FIFA (octobre 2017)

Rang	Pays
1	Brésil
2	Allemagne
3	Argentine
4	Suisse
5	Pologne
6	Portugal
7	Chili
8	Colombie
9	Belgique
10	France

Et les prochains ?

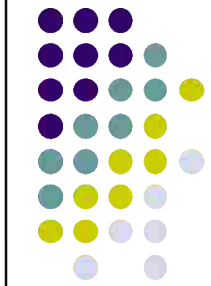
Evidence #2

- Un texte = tous les romans écrits par un même auteur.
- Ignorez les mots apparaissant une ou deux fois.
- Appliquer la distance de Labbé (min: 0; max: 1).

Conclusion

1. Dist (Ferrante, Starnone) = 0,177
2. Dist (Picollo, Veronesi) = 0,220
3. Dist (Nesi, Veronesi) = 0,226
4. Dist (De Silva, Veronesi) = 0,227

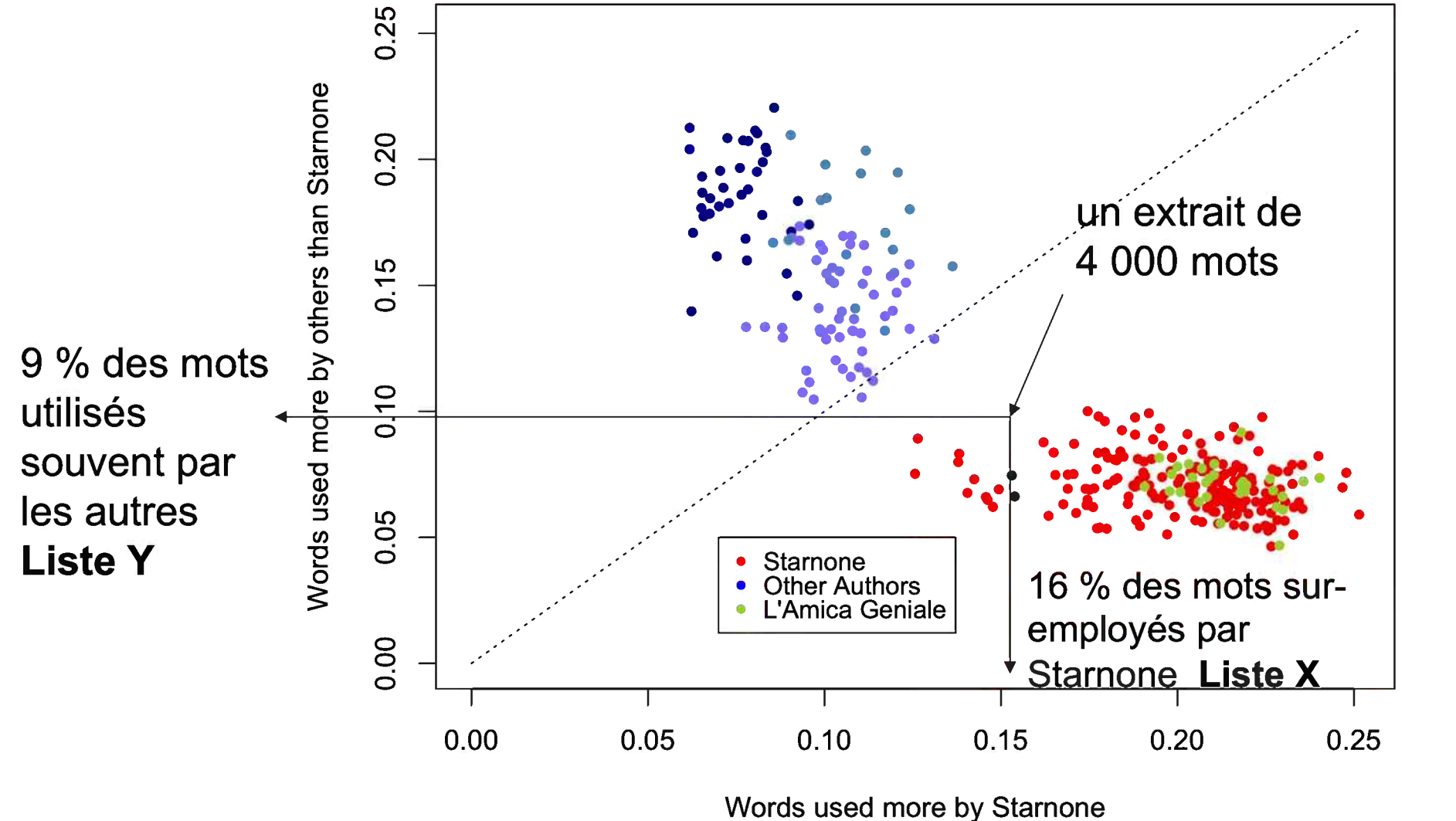
Evidence #3: Le test Zeta



- Etablir une liste des mots employés souvent par Starnone mais peu par les autres auteurs. C'est la liste X.
  (par exemple, *volta, er, sua, solo, poi, senza, della, quando*, …)
- Et une liste des mots utilisés fréquemment par les autres auteurs (sans considérer les romans de Ferrante). C'est la liste Y.
  (par exemple, *volto, muro, strana, nonstante, propria, ognuno*, …)

- Prenons un passage de 4 000 mots d'un roman (*L'amica geniale* de Ferrante). Combien de mots appartiennent à la liste X et à la liste Y ?

Le test Zeta

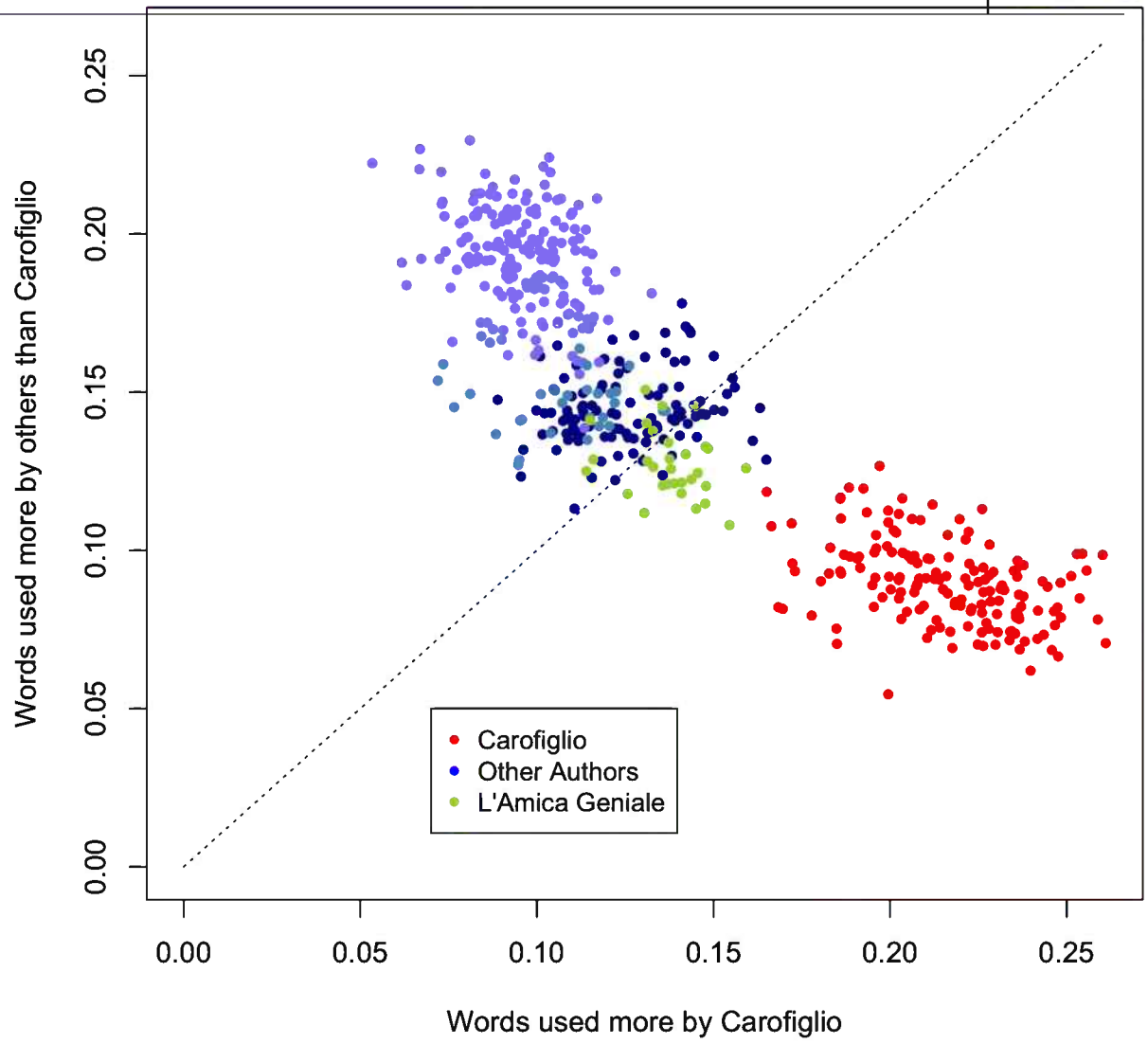

Clustering by authors
Words used more by others than Starnone
0.00
0.05
0.10
0.15
0.20
0.25
Words used more by Starnone
Starnone
Other Authors
L'Amica Geniale
un extrait de
4 000 mots
9 % des mots
utilisés
souvent par
les autres
Liste Y
16 % des mots sur-
employés par
Starnone Liste X

Le test Zeta

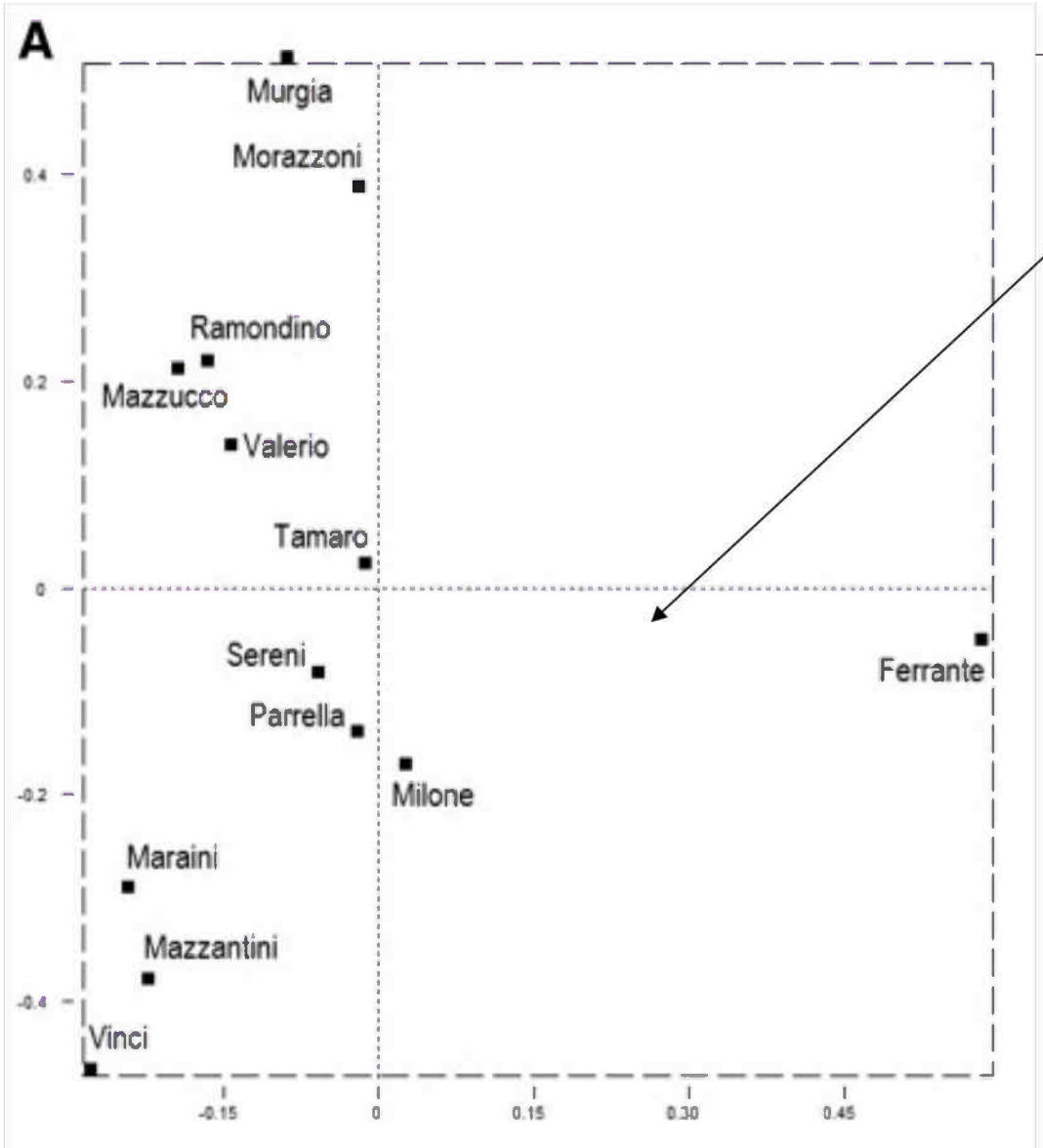
Contre-exemple

Carofiglio

Style différent de Ferrante

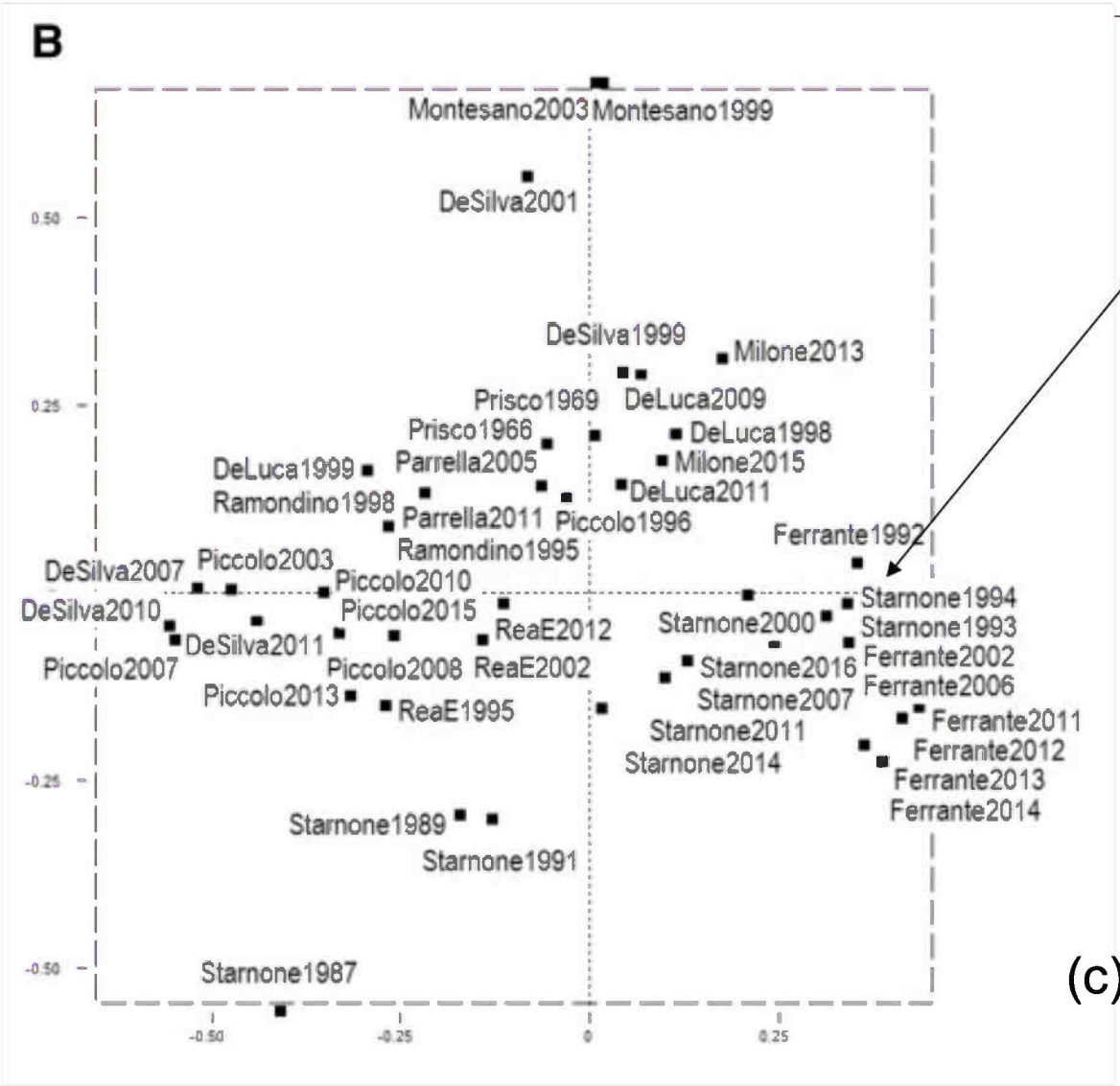


Les auteures



Large distance entre *Ferrante* et les autres auteures

Les auteur-e-s de Campania



Distance très faible entre *Ferrante* et Starnone

(c) A. Tuzzi, 2018

Nous découvrons que …

- Evidence #1

  Roman par roman

  Storia Della Bambina Perduta proche de *Lacci*
- Evidence #2

  Par auteur: Ferrante très proche de Starnone
- Evidence #3 (Zeta)

  L'amica geniale a un vocabulaire

  proche de celui de Starnone
- Evidence #4 (Homme et Campania)

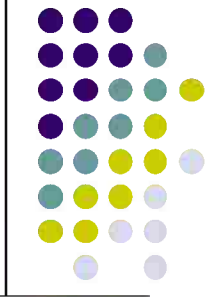
  Toujours Starnone!

Domenico Starnone

Analyse des mots

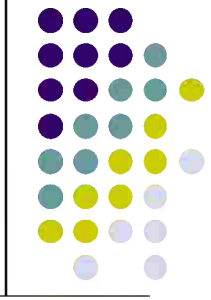
Mots	Corpus	Ferrante (6,5%)	Starnone (6,4%)
padre	9 815	833 (8,5%)	1 170 (11,9%)
madre	8 246	1 104 (13,4%)	762 (9,2%)
perciò	1 263	222 (17,6%)	254 (20,1%)
persino	1 351	266 (19,7%)	205 (15,2%)
temere	1 345	274 (20,4%)	207 (15,4%)
tono	2 135	421 (19,7%)	286 (13,4%)
strunz	85	18 (21,2%)	63 (74,1%)

Vérification



- Est ce que cette conclusion est falsifiable?
- Eliminons Starnone de notre corpus…
- Appliquons nos modèles pour recherche le véritable auteur.
- 12 noms apparaissent au moins une fois au premier rang :
  Balzano, Mazzucci, Milone, Tamaro, Veronesi,
  Brizzi, Carofiglio, Giordano, Fois, Murgia, Raimo, Sereni

Sommaire



- Linguistique et informatique
- L'hortogaffe!
- Loi de Zipf
- Profilage et Cie
- Déterminer l'auteur d'un écrit ou d'un roman
- **2016 : Détection des menteurs (fake news)**

2016 : Election aux Etats-Unis

Un candidat hors norme apparaît: D. Trump.

Il gagne les primaires républicaines.

Et gagne l'élection le 8 novembre 2016

Donald Trump – Parti républicain

Colistier : Mike Pence

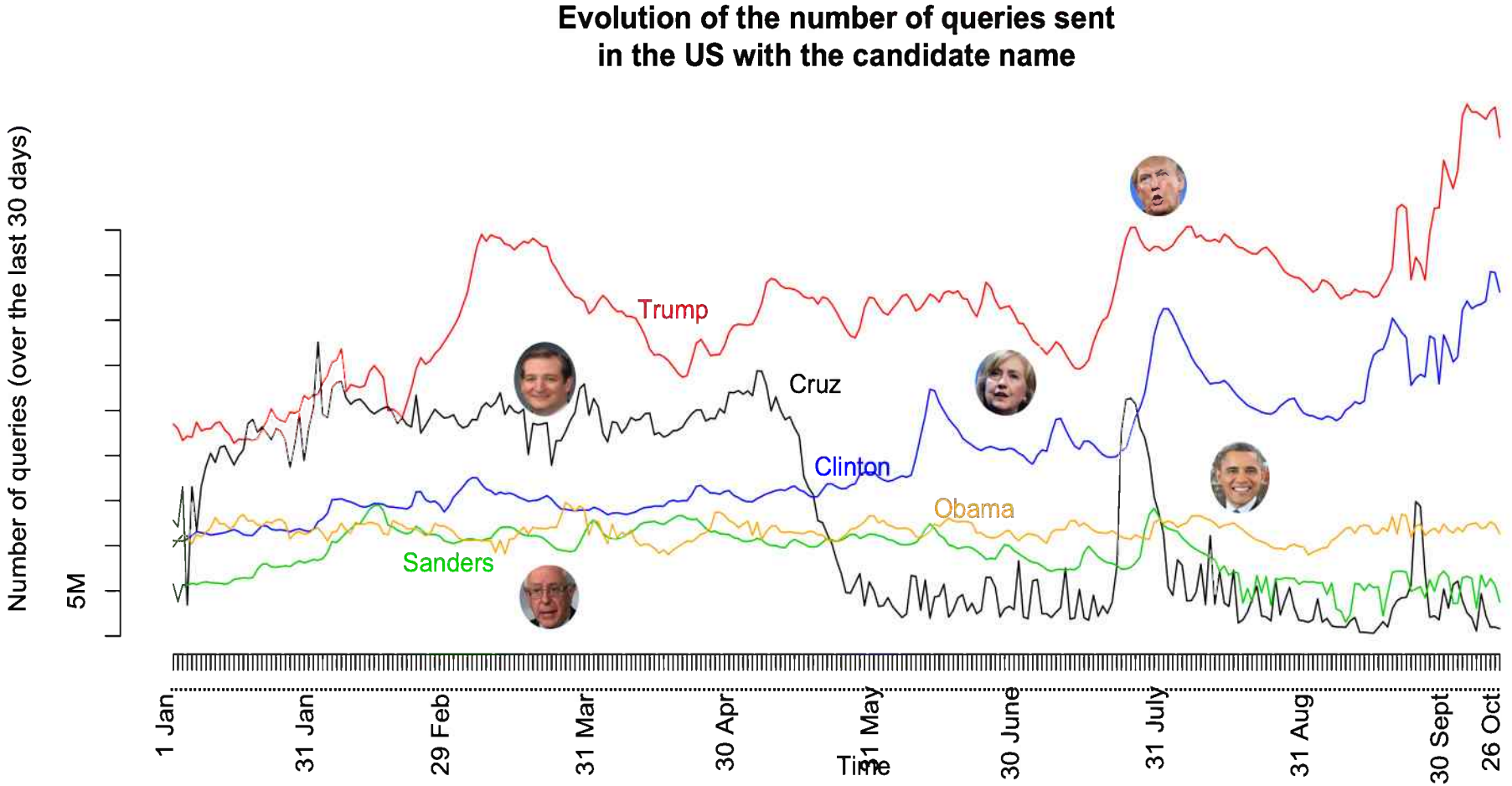
Voix	62 984 828
	46,1 %
Grands électeurs	304

Hillary Clinton – Parti démocrate

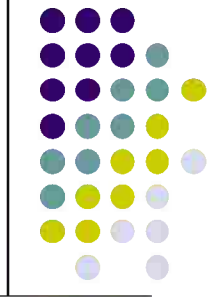
Colistier : Tim Kaine

Voix	65 853 514
	48,2 %
Grands électeurs	227

2016 : Election aux Etats-Unis



Discours électoral



- Comment gagner une élection ?
- Communication démontrant les qualités du candidat.
- Homme fort (un dur à cuire et pas une lavette).
- Connaissant les problèmes et les solutions.
- Paraître honnête avec de l'empathie (mais juste un peu).

- Parler direct et franc: Vers la simplicité!
- Par un lexique choisi, un rythme dans les phrases, …

- Mots courts, phrases courtes, simplissime, à la tweet

Mots très fréquents

Les lemmes les plus fréquents dans les primaires.

Le Je/moi et le Nous.

	Bush	Cruz	Trump	Clinton	Sanders
1e	be	be	be	be	be
2e	the	the	**I**	the	the
3e	to	and	the	to	**I**
4e	**we**	to	**we**	**I**	to
5e	that	I	and	and	and

Savoy J. (2018). Analysis of the style and the rhetoric of the 2016 US presidential primaries. *Digital Scholarship in the Humanities*, 33(1), 143-159 .

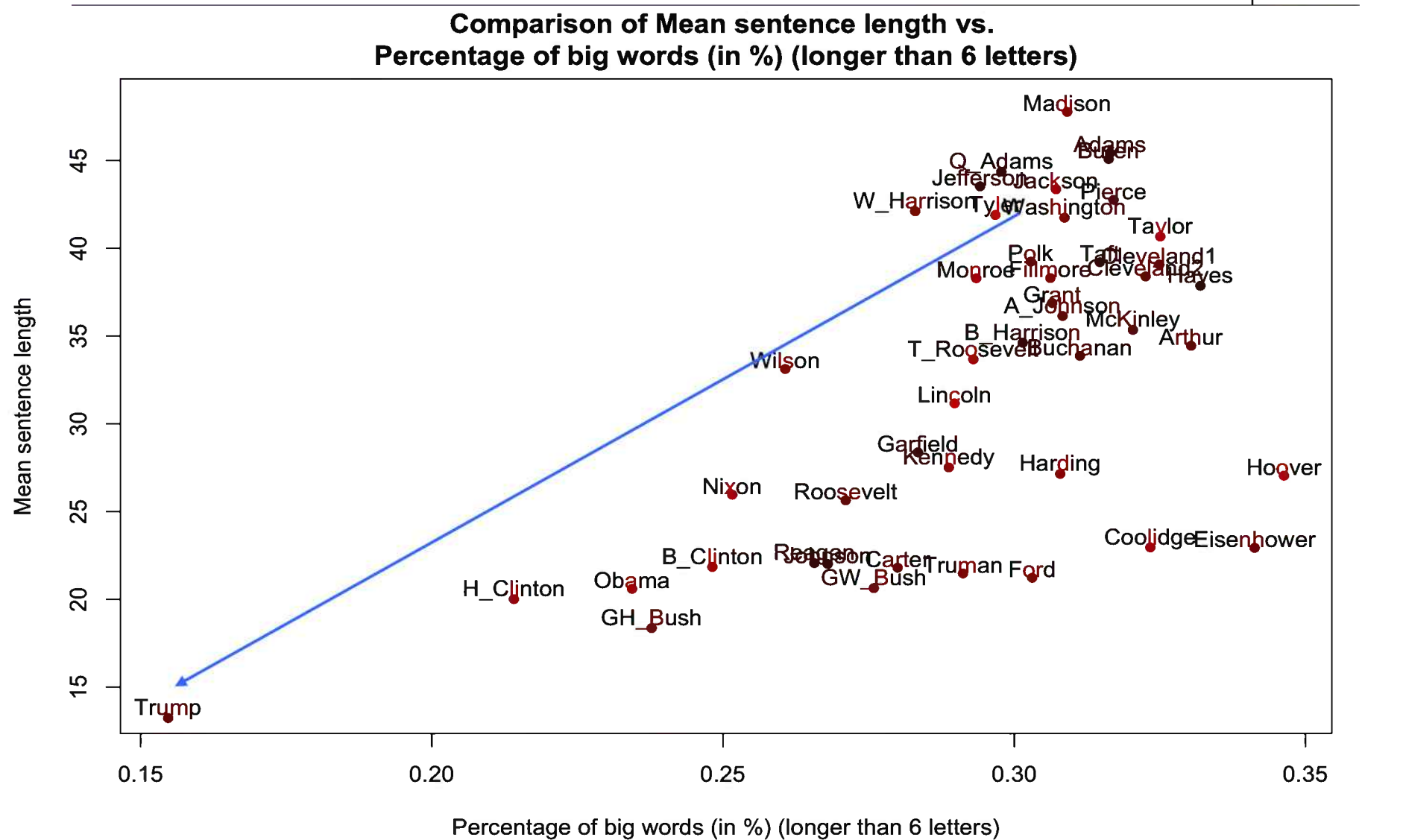
Simplicité

- Pourcentage de mots complexes (Big Words, BW)
- Hypothèse : les mots longs sont plus complexes.

$$BW = \frac{\mid \#longueur(mots) > 5 \mid}{\mid \#de\ mots \mid}$$

"One finding of cognitive science is that words have the most powerful effect on our minds when they are simple. The technical term is basic level. Basic-level words tend to be short. … Basic-level words are easily remembered; those messages will be best recalled that use basic-level language." (Lakoff & Wehling, 2012, p. 41).

Simplicité


Comparison of Mean sentence length vs.
Percentage of big words (in %) (longer than 6 letters)
Mean sentence length
Percentage of big words (in %) (longer than 6 letters)
15
20
25
30
35
40
45
0.15
0.20
0.25
0.30
0.35
Madison
Adams
Buren
Q_Adams
Jefferson
Jackson
Pierce
W_Harrison
Tyler
Washington
Taylor
Polk
Monroe
Fillmore
Cleveland1
Cleveland2
Hayes
Grant
A_Johnson
McKinley
B_Harrison
Arthur
T_Roosevelt
Buchanan
Wilson
Lincoln
Garfield
Kennedy
Harding
Hoover
Nixon
Roosevelt
Coolidge
Eisenhower
B_Clinton
Reagan
Johnson
Carter
Truman
Ford
GW_Bush
H_Clinton
Obama
GH_Bush
Trump

Mesures stylistiques

Transcrits des débats des primaires.

LMP : Longueur moyenne des phrases.

BW : % de mots complexes.

LD: densité des mots porteurs de sens.

	Bush	Cruz	Trump	Clinton	Sanders
LMP	17,3	19,4	**13,7**	20,5	19,7
BW	24,4 %	26.4 %	**18,3** %	24,1%	26,4%
LD	40,7 %	44,6 %	**36,6 %**	40,4 %	43,6 %

Savoy J. (2018). Analysis of the style and the rhetoric of the 2016 US presidential primaries. *Digital Scholarship in the Humanities*, 33(1), 143-159.

Phrases caractéristiques

D. Trump, Speech, Sept 13th, 2016

"American cars will travel the roads, American planes will soar in the skies, and American ships will patrol the seas."

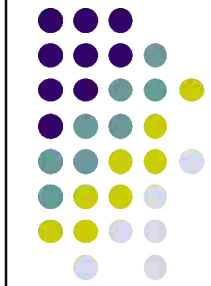
D. Trump, TV debate, March 10th, 2016

"They don't like seeing bad trade deals, they don't like seeing higher taxes, they don't like seeing a loss of their jobs where our jobs have just been devastated."

H. Clinton, Interview on CNN, February 24th, 2016

"So what people talk to me about is how I'm going to get incomes up, how I'm going to make sure the Affordable Care Act works, and get prescription drug costs down, and make college affordable and relieve student debt."

Mensonge et Fake news

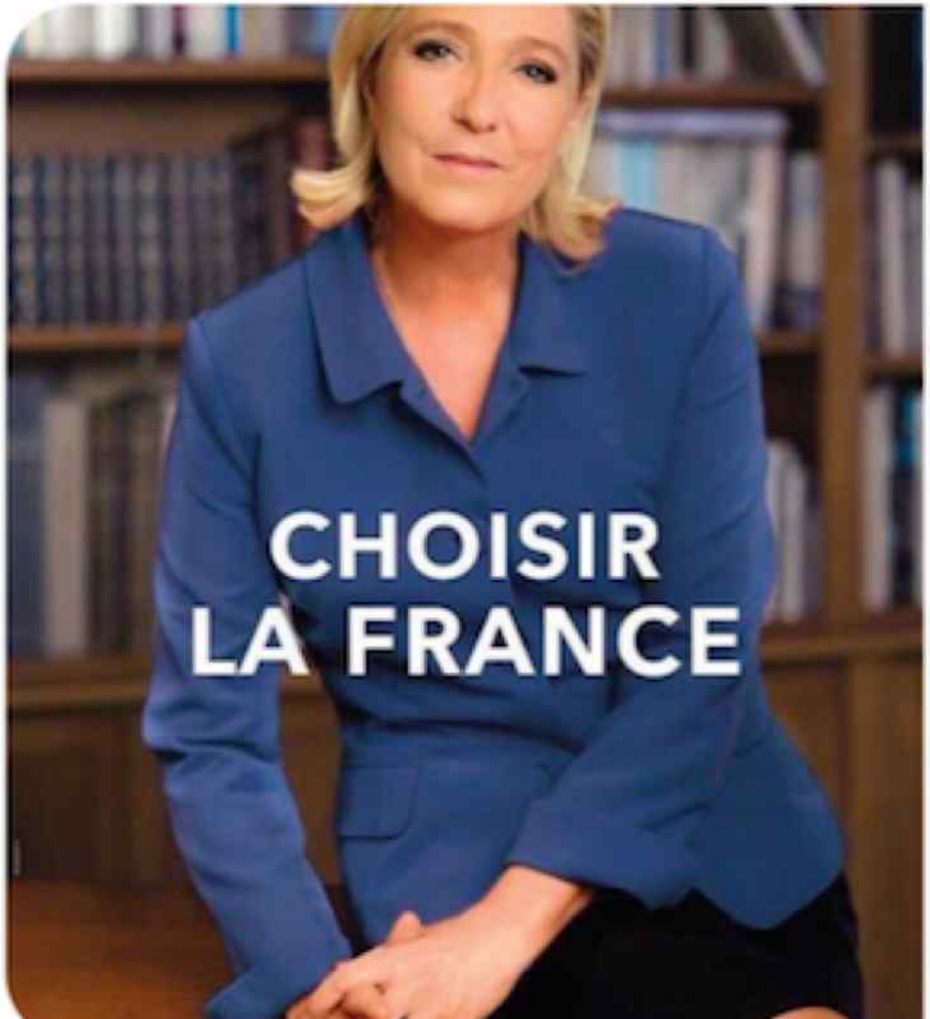


- Enoncé délibéré d'un fait que l'on sait contraire à la vérité.
- Nouvelles trompeuses, information fallacieuse, infox.
- *Fake news* : nouvelle qui est intentionnellement fausse, mais vérifiable, pour induire en erreur le lecteur (ou le consommateur).

- Propagande ?

- Et la satire, moquerie, canular ?

Photos / Images



Photos / I

- Exemples

@realDonaldTrump Poor Jeb. I could've sworn I saw him outside Trump Tower the other day! pic.twitter.com/EMeIEdsinX

4:42 PM - 22 Jan 2016

RETWEETS 46 LIKES 42

Dans un document ?

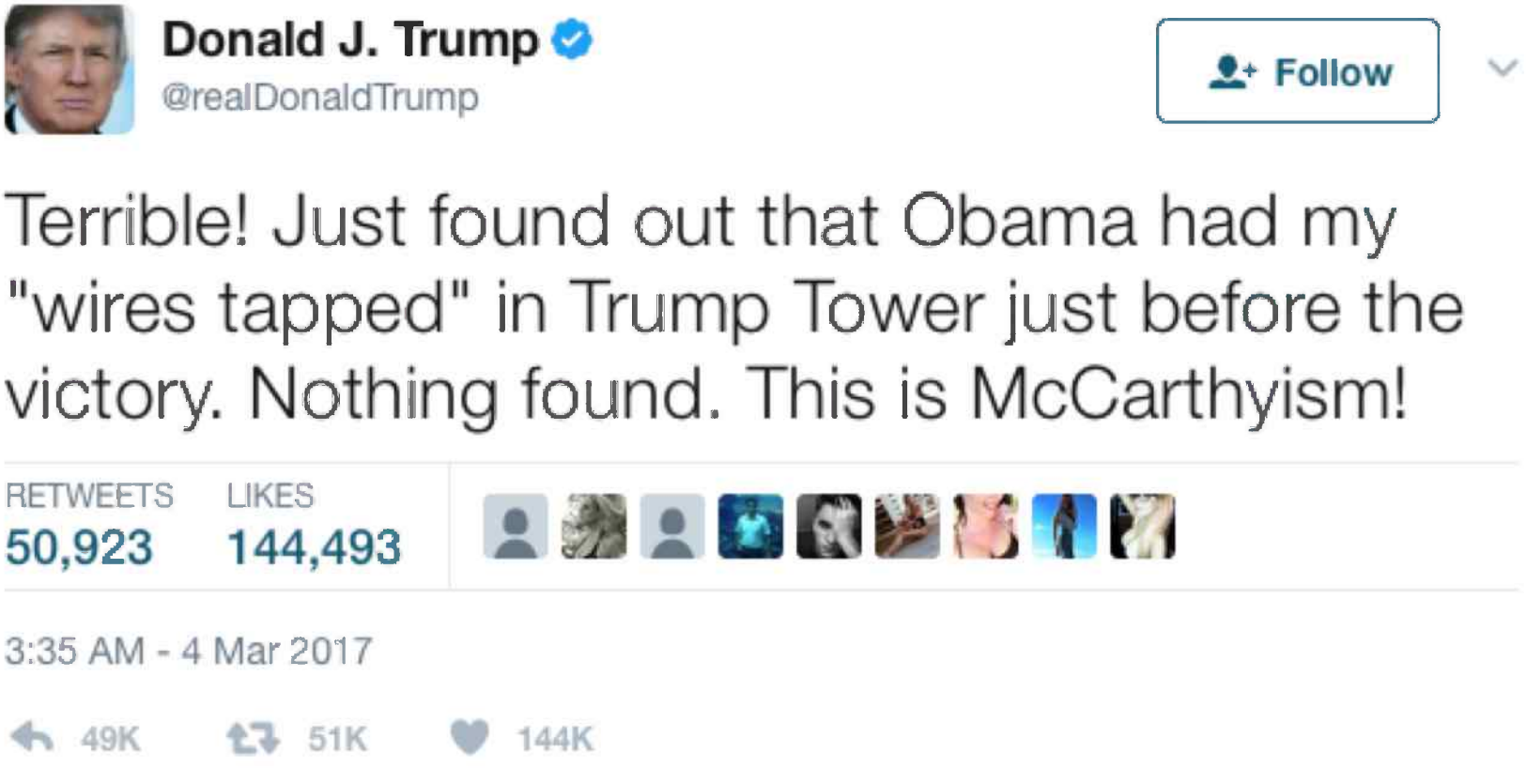
Les faux ont toujours existés.

Donation de Constantin : L'empereur Constantin I^{e} (4^{e} siècle) transfert l'autorité sur Rome et la partie occidentale de l'empire romain au pape.

En 1439, Lorenzo Valla conclut que ce document est une contrefaçon (présence de termes anachroniques indiquant un texte écrit au 8^{e} siècle).

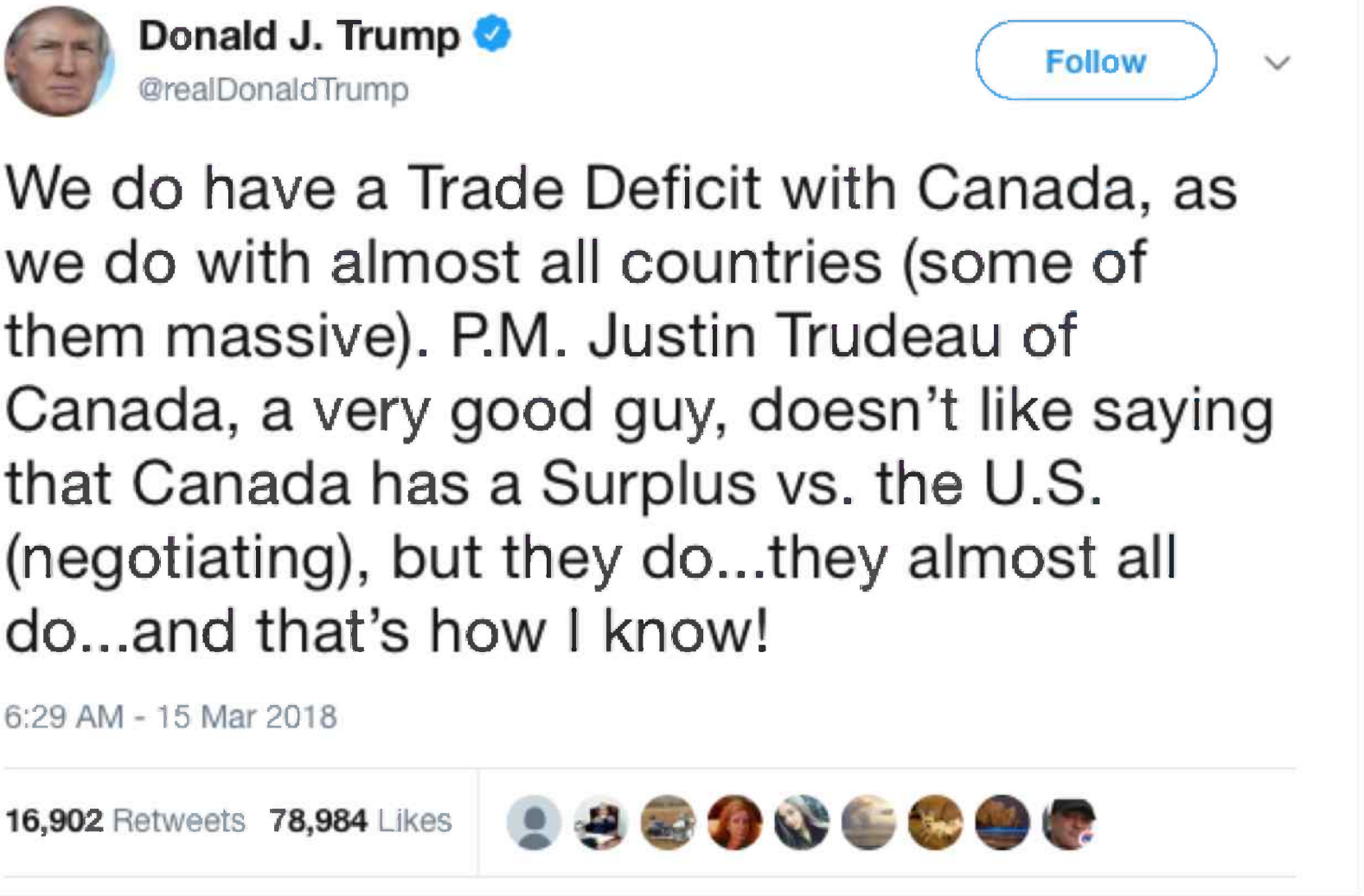
Evidences externes : analyse papier, encre, écriture manuscrite, contexte historique, …

Sur Internet



Shu K., Silva, A., Wang, S., Tang, J., Liu, H. (2017). *Fake News Detection on Social Media: A Data mining Perspective*. Proceedings SIGKDD Explanatory Newsletter, 19(1), 22-36.

Fact-checking

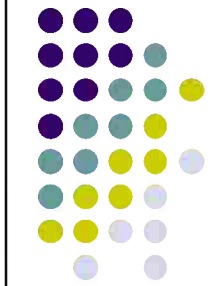


Mensonge ?

- Si on analyse une histoire, le menteur a tendance à :
  - écrire une histoire plus courte;
  - utiliser des mots courts et phrases courtes;
  - moins de détails, moins de nombres (moins précis);
  - moins de conjonctions (mais, sauf), moins de références au temps, lieu, mouvement;
  - plus d'émotions;
  - plus de mots cognitifs (penser, croire, …);
  - plus de certitude, plus de références aux autres;

Pennebaker, J.W. (2011). The Secret Life of Pronouns. What our Words say about us. New York: Bloomsbury Press.

Vrai ou Faux ? Expérience 1$



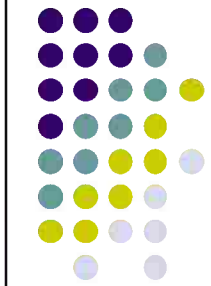
“No, I didn't take your dollar.”

“I don't believe in stealing. I did it once a long time ago; I was … younger.”

“It really offends me that you would accuse me of something like that.”

“Why would I? I would never even think to look in the book to look for a dollar.”

Vrai ou Faux ?

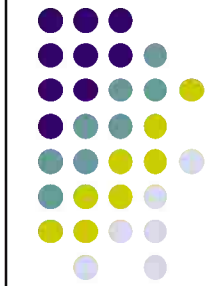


“Mistakes were quite probably made by the administrations in which I served.” Henry Kissinger

“I'm going to say this again: I did not have sexual relations with that woman, Miss Lewinsky.” Bill Clinton

“I want to state clearly and without qualification: I did not take steroids, human growth hormone, and any other banned substances.” Un cycliste américain

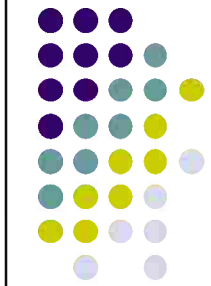
Le cas Stephen Glass



- Journaliste au *New Republic Magazine* (NY).
- Déc. 1995 à mai 1998, 41 longs articles et sur différents thèmes dans un style émouvant et avec des citations colorées.
- Une étoile montante du journal.
- Mais six articles sont complètement faux et 21 autres largement imaginaires (14 authentiques).

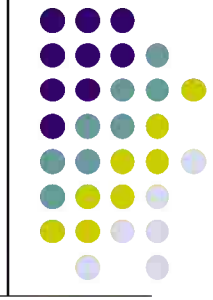
- Détection automatique : taux de succès 60 à 65%.

Raisons : contexte sociétal



1. Préférence pour les nouvelles brèves;
2. Alphabétisation fonctionnelle (littératie) (45 % en France, Italie, US (OCDE));
3. Se forger une opinion via ses amis (biais de représentativité);
4. Biais de confirmation (répétition implique la véracité).
5. Les communautés jouent le rôle de relais, d'amplificateur. La répétition du message est important.

(Re)Diffusion des mensonges



- Comment obtenir une large diffusion ?

  Nombre de personnes sur tweeter (2,6 milliards);

  Recherche des nouvelles via ces réseaux (62 % aux US).

- Les réseaux sociaux :
  - gratuits;
  - rapides;
  - faciles à utiliser;
  - faciles à propager;
  - pas d'intermédiaire (contrôle, vérification);

Exemples de communautés

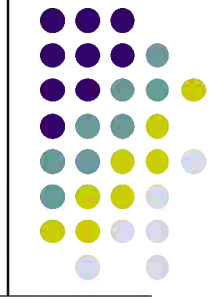
Chemtrails	Théorie du complot
Vaccin et autisme	Extra-terrestre
Energie libre	Reptiliens

Pizzagate (30 oct. 2016, contre H. Clinton (4 déc. 2016))

Quattrociocchi, W. (2016). *Fake Online News Spreads Through Social Echo Chambers.* Scientific American. November 2016.

Bessi, A., *et al.* (2016), *Users polarization on Facebook and Youtube* in PLOS ONE, 11 (ISSN 1932-6203)

Diffusion (Retweet) ?



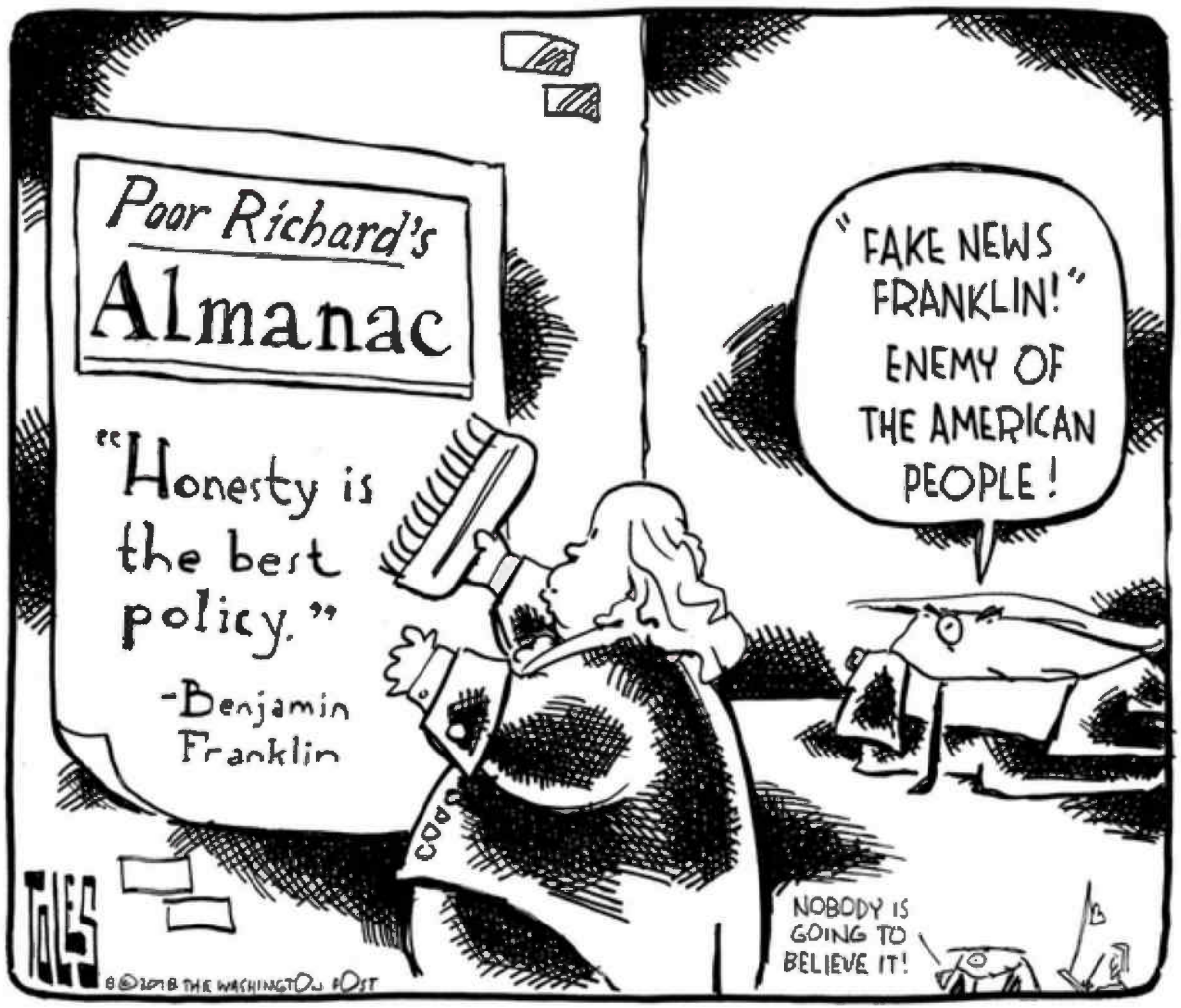
Pour un tweet, les chances de retweet augmentent si :

1. Contenu surprenant
2. Emotionnel négatif (colère)
3. Conforme à notre a priori

Exemple : contenu scandaleux, négatif avec une image.
Amplification par la communauté (chambre d'écho).

Del Vicario, M. *et al*. (2016). *The spreading of misinformation online*. PNAS, 113(3), 554-559.

En conclusion



(Tom Toles)

Exercice: Créer votre infox

Prenons une affaire,

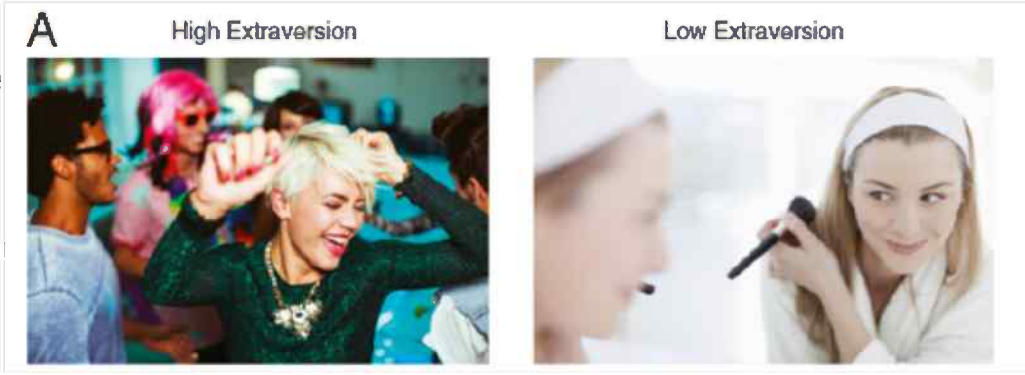
Ajoutez une photo (ou vidéo),

Et provoquez la colère (sur un a priori)

Pierre Maudet, …

Conclusion

- Les langues évoluent
- Et l'informatiquer
  - Facilité de diffusion (nouvelle forme d'expression (tweets))
  - Moteur de recherche (accès à l'information)
  - Correction de l'orthographe
  - Aide à la traduction
  - Qui est le véritable auteur ?
  - Profilage (publicité)
  - Détection de mensonge
- Illectronisme



Conclusion

Merci de votre attention.

Questions ?