Université de Montréal

Text Mining a Noisy Corpus of Old Regime French Periodicals

A War Story

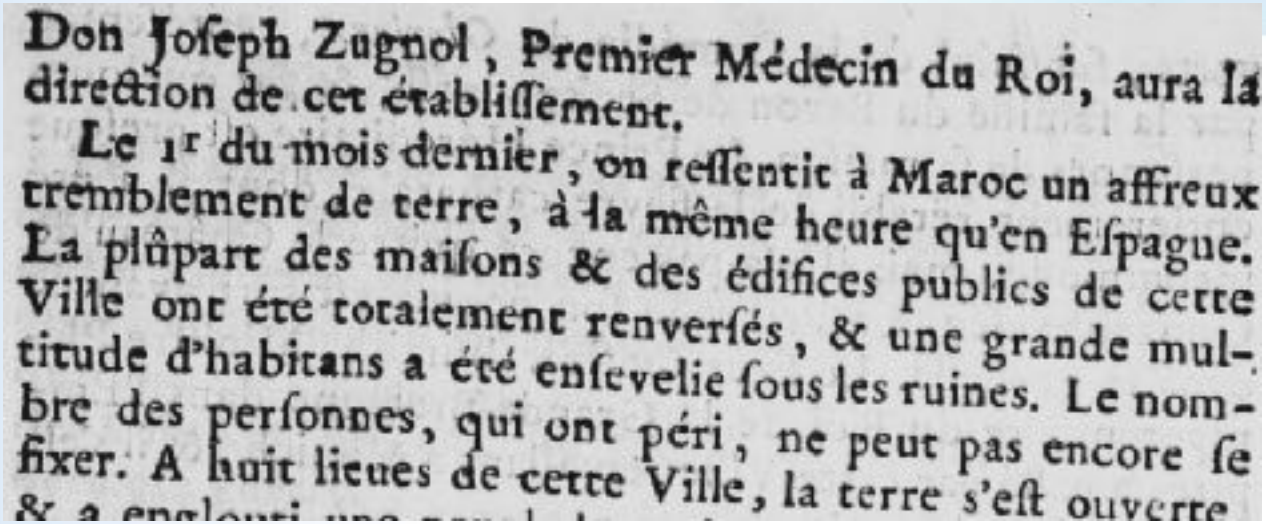
François Dominic Laramée
PhD Candidate in History, Université de Montréal
September 12th, 2017

THE BEST LAID PLANS OF MICE AND MEN…

« … no plan of operations extends with any certainty beyond the first contact with the main hostile force. »

— Helmuth von Moltke the Elder (1800-1891)

THE ENEMY



Don Joſeph Zugnol, Premier Médecin du Roi, aura la direction de cet établiſſement.

Le 1r du mois dernier, on reſſentit à Maroc un affreux tremblement de terre, à la même heure qu'en Eſpagne. La plûpart des maiſons & des édifices publics de cette Ville ont été totalement renverſés, & une grande multitude d'habitans a été enſevelie ſous les ruines. Le nombre des perſonnes, qui ont péri, ne peut pas encore ſe fixer. A huit lieues de cette Ville, la terre s'eſt ouverte

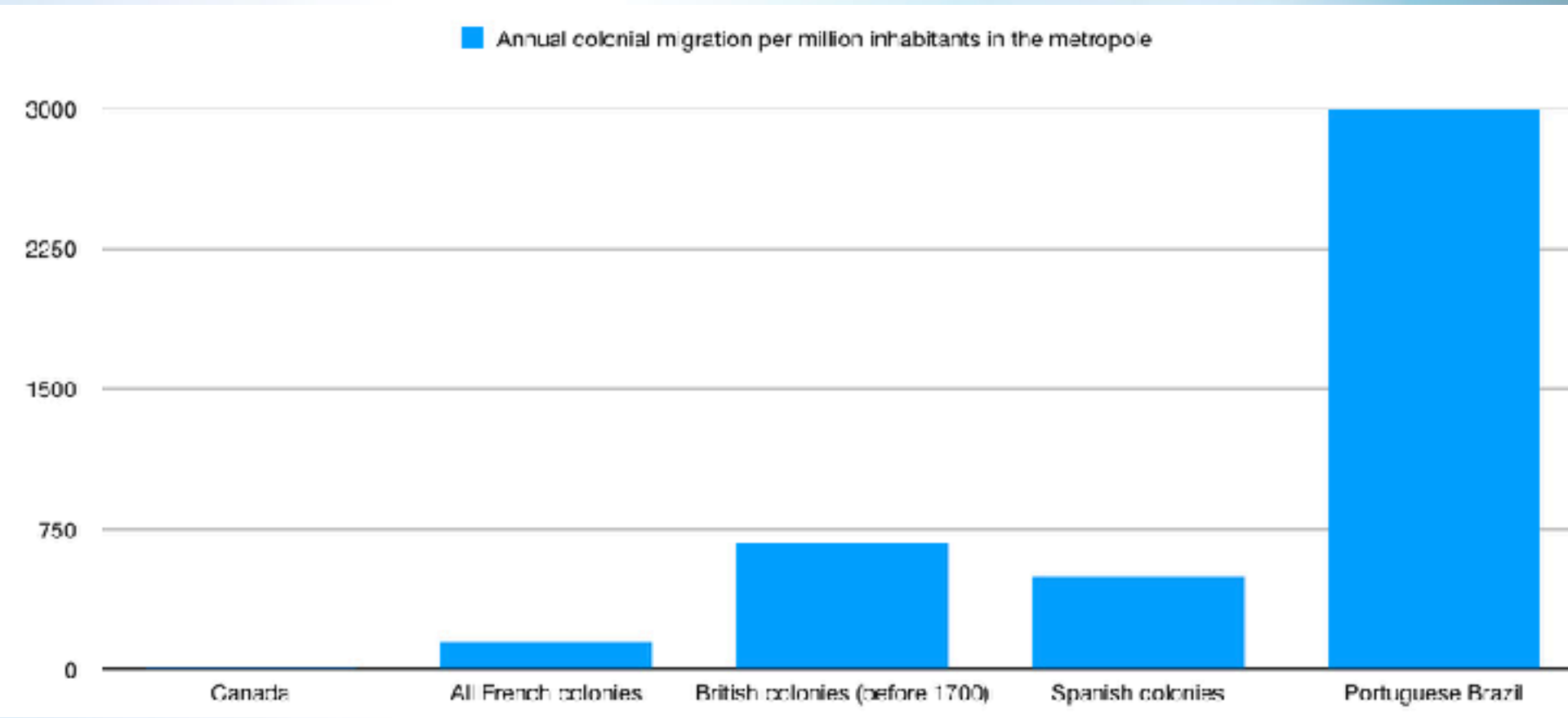
L-e Ir 'dùmois demih, On reRèntic à Maroc un affreux '- ViMe même beure qu'en Efpa<>ne, La plâpart des imifons & des-édificies publics de ctteVille ont été totalement renver * une grande mul,": titude d'habitans aéreenfevelie fous les ruines. Le nom- PfS encore sebredes personnes, qui ne peut pas encore [e ftrer.

PRESENTATION STRUCTURE

Five parts

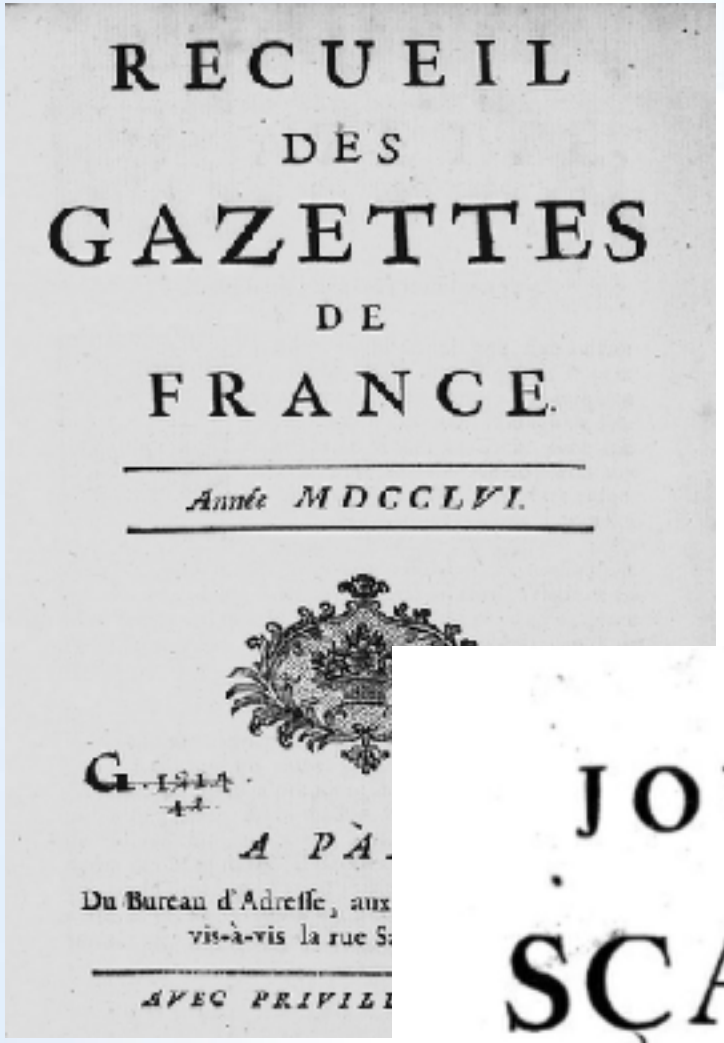
- Historical problem
- Corpus of Old Regime periodicals
- Methodology
- Gazette: a mostly digital analysis
- Other periodicals: a weakly digital analysis

COLONIAL MIGRATION DURING THE OLD REGIME



Data source: Landry (2006), « Les Français passés au Canada avant 1760 »

CORPUS

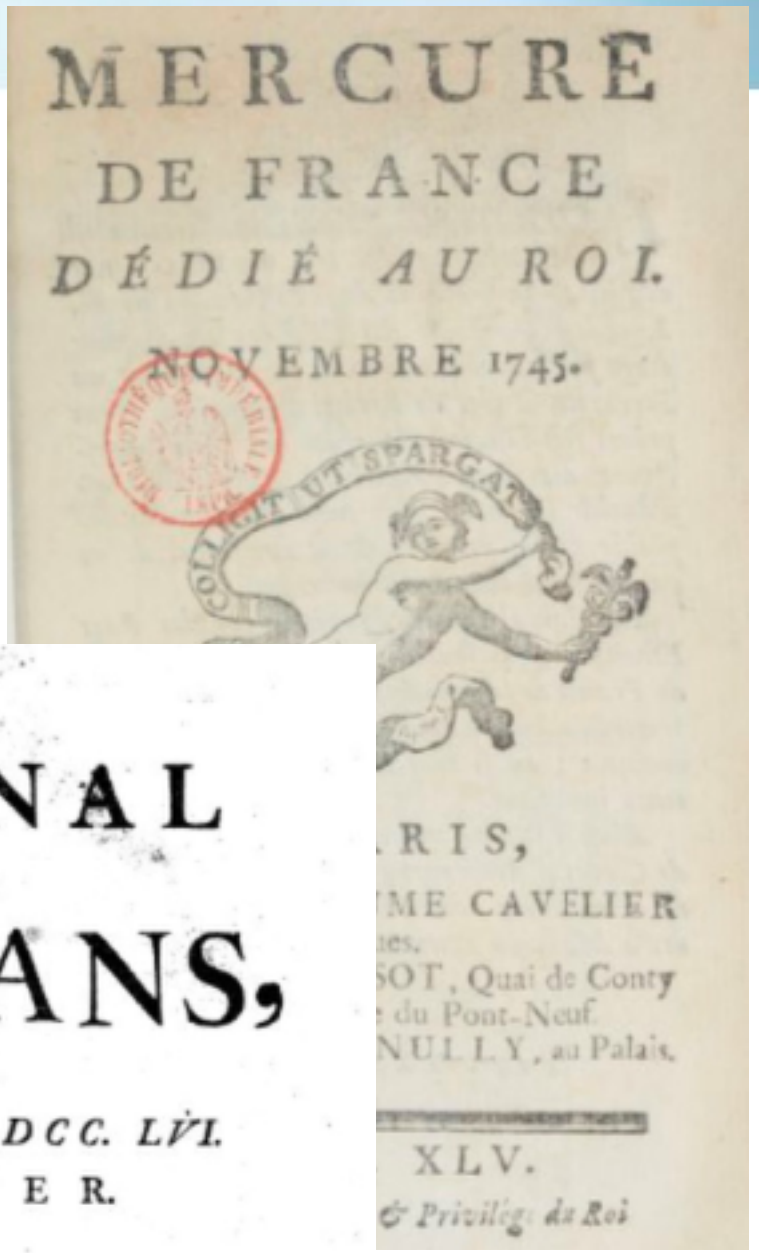

RECUEIL
DES
GAZETTES
DE
FRANCE.

Année MDCCLVI.

A PA

Du Bureau d'Adreſſe, aux
vis-à-vis la rue S

AVEC PRIVILE


MERCURE
DE FRANCE
DÉDIÉ AU ROI.

NOVEMBRE 1745.

COLLIGIT UT SPARGAT

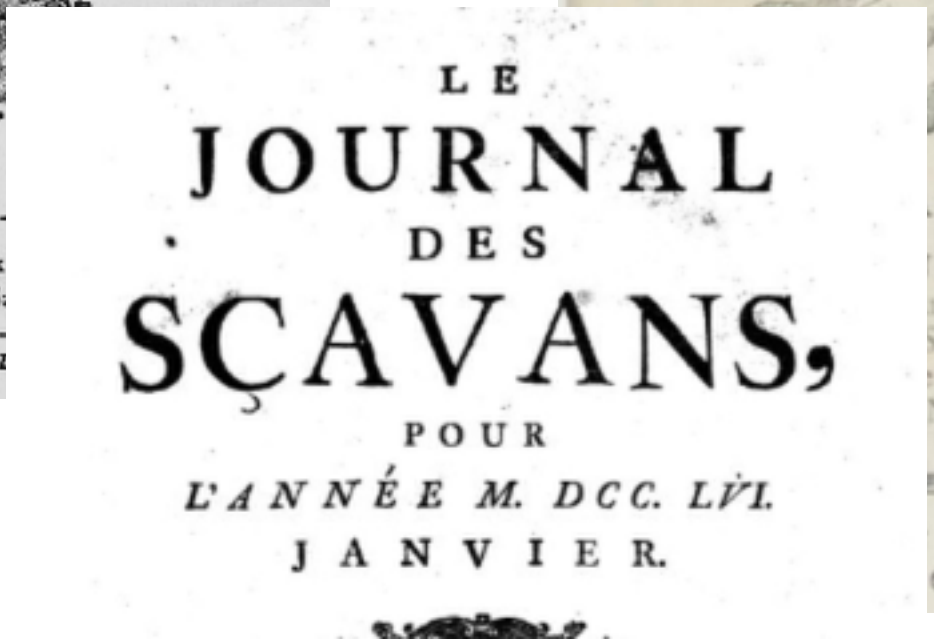
RIS,

ME CAVELIER

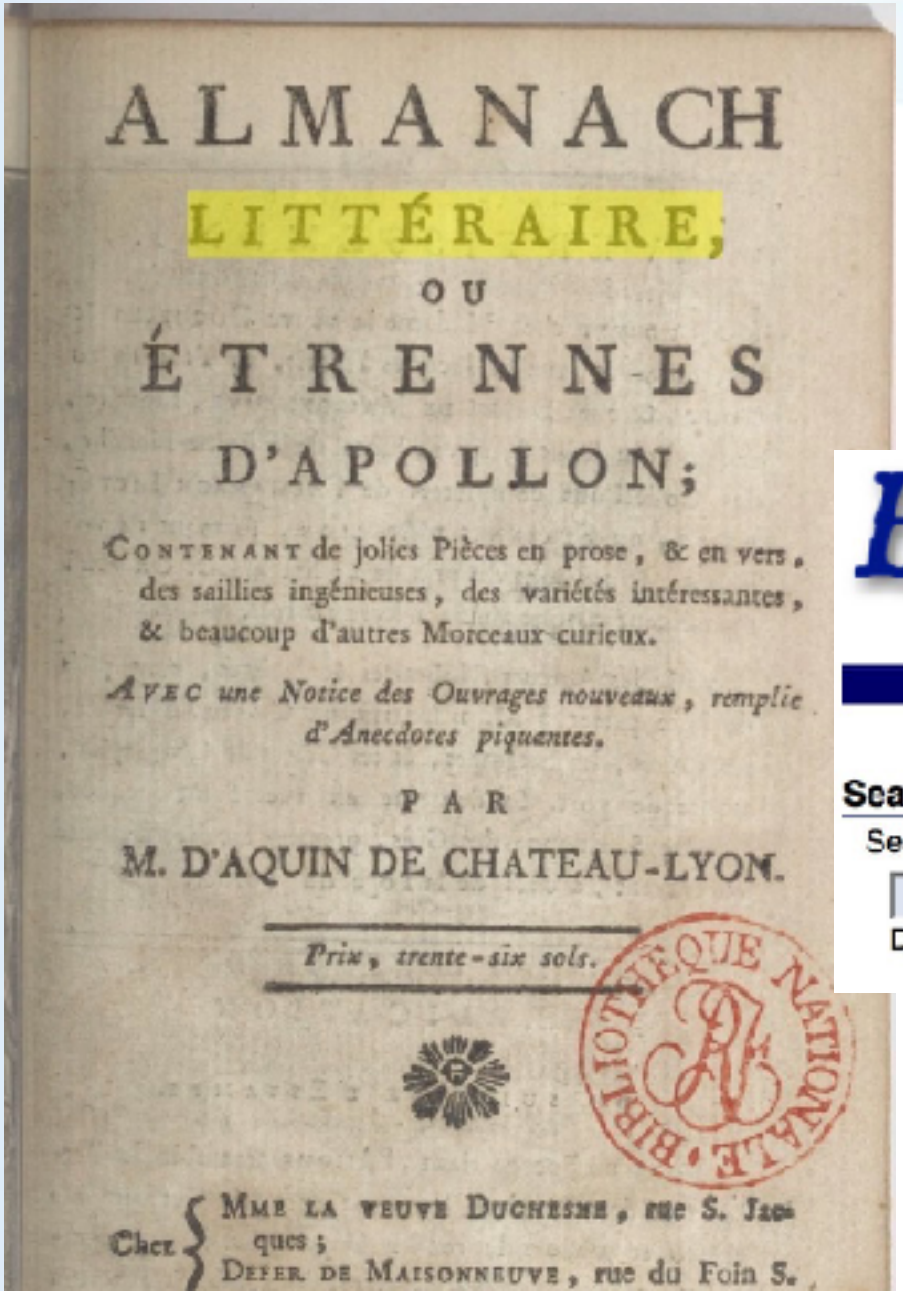
SOT, Quai de Conty
du Pont-Neuf.
NULLY, au Palais.

XLV.

& Privileg. du Roi


LE
JOURNAL
DES
SÇAVANS,
POUR
L'ANNÉE M. DCC. LVI.
JANVIER.

THE SILENCE OF POPULAR MEDIA



ALMANACH

LITTÉRAIRE,

OU

ÉTRENNES

D'APOLLON;

CONTENANT de jolies Pièces en prose, & en vers, des saillies ingénieuses, des variétés intéressantes, & beaucoup d'autres Morceaux curieux.

AVEC une Notice des Ouvrages nouveaux, remplie d'Anecdotes piquantes.

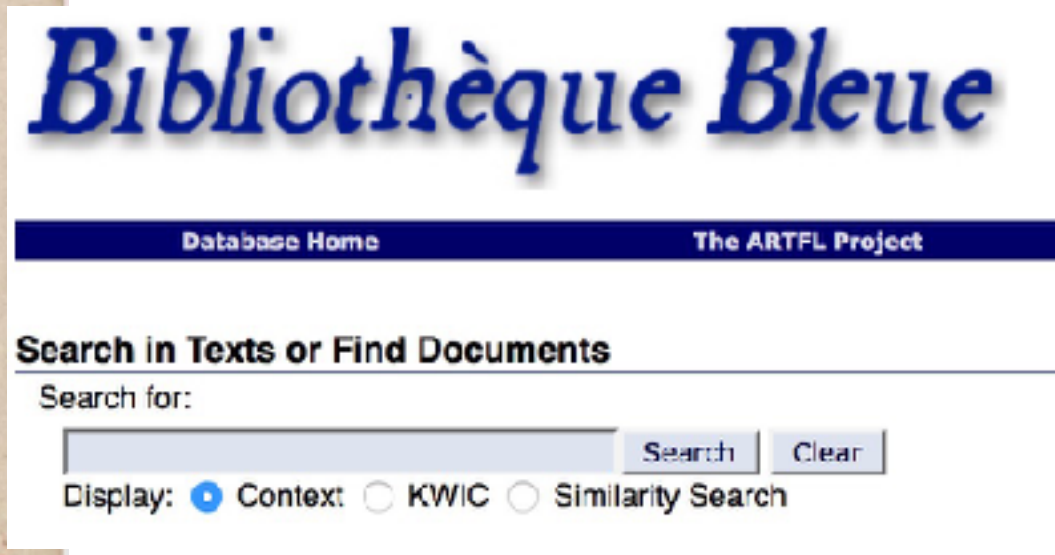
PAR

M. D'AQUIN DE CHATEAU-LYON.

Prix, trente-six sols.

BIBLIOTHÈQUE NATIONALE

Chez { MME LA VEUVE DUCHESNE, rue S. Jacques ; DEFER DE MAISONNEUVE, rue du Foin S.



CORPUS-SPECIFIC OCR ISSUES

- Problematic source documents
- High error variance
- Errors do not follow usual patterns

METHODOLOGY OVERVIEW

HEURISTIC PROJECT DESIGN

- Exploratory methods => Salient patterns
- Salient patterns => Small keyword set
- Small keyword set => Noise neutralization

KEYWORD-DRIVEN DATA RECONSTRUCTION

- Focused OCR correction
- Article subset selection
- Metadata extraction

LEVENSHTEIN'S ALGORITHM

Keyword	Candidate	Levenshtein Distance	Operations
Amérique	Amérique	0	—
Amérique	Amrique	1	Delete « é »
Amérique	Cmévrique	2	A —> C Insert « v »
Amérique	Musique	3	Delete « A » é —> u r —> s

LEVENSHTEIN APPLIED TO THE *GAZETTE*

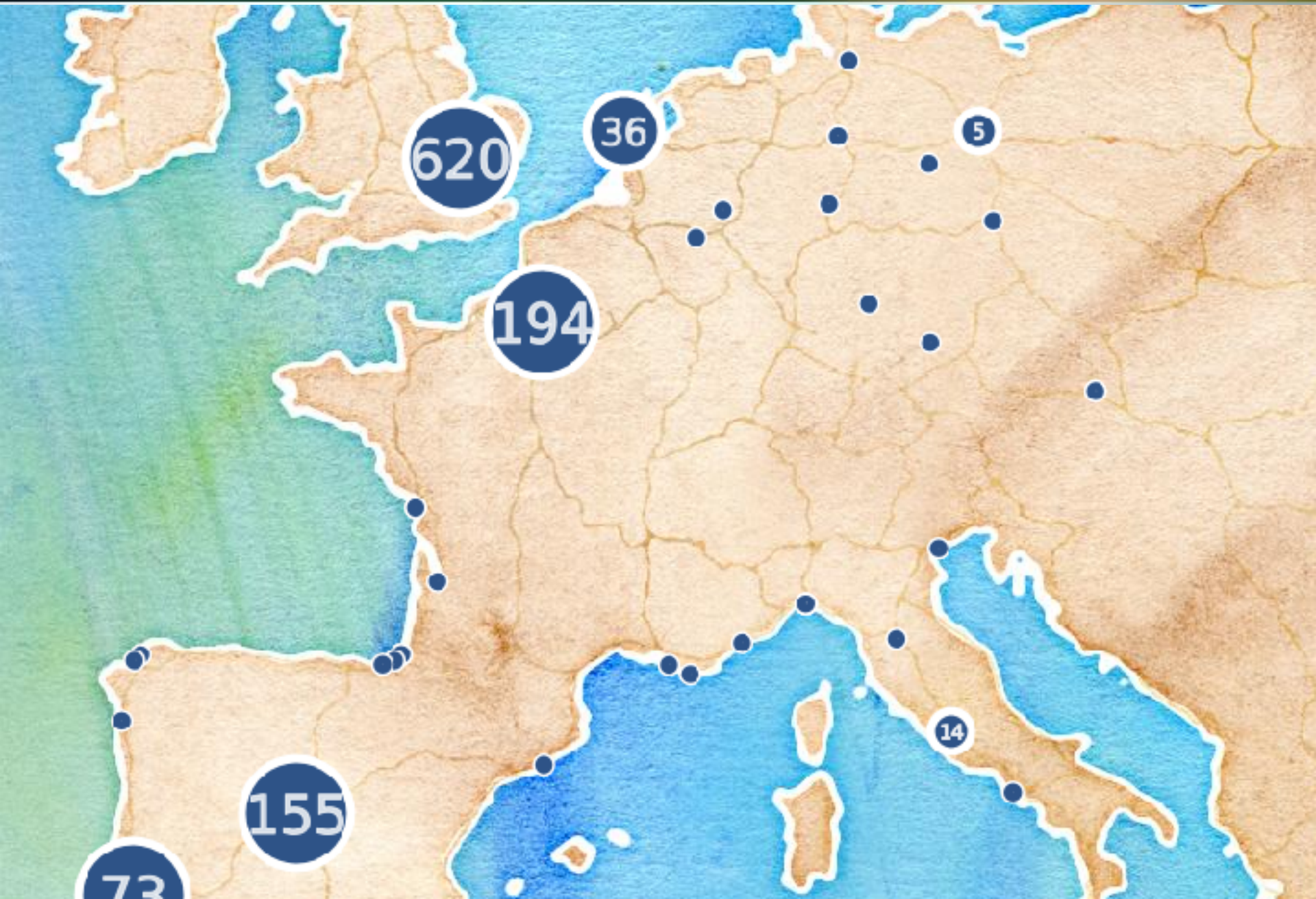
SOME OF THE RECOVERED INSTANCES

Keyword	Keyword tokens	Alternate types	Alternate type occurrences	Alternate types as % of all tokens	Examples
Amérique/ d'Amérique/ l'Amérique	485	36	128	20,9 %	l'amerique (59), d'amcrique (2), ramërique
Brésil	5	10	159	97,0 %	bresil (143), bretîl, brcfil
Canada/ Canadiens	139	3	3	2,1 %	canada*, en.canada
Colonie(s)	411	15	16	3,7 %	5lonie, coioniej
Domingue/ Saint-Domingue	88	11	13	12,9 %	jjomingue, saintdomingue
TOTAL	**1867**	**103**	**532**	**22,2%%**	**Increase: 28,5%**

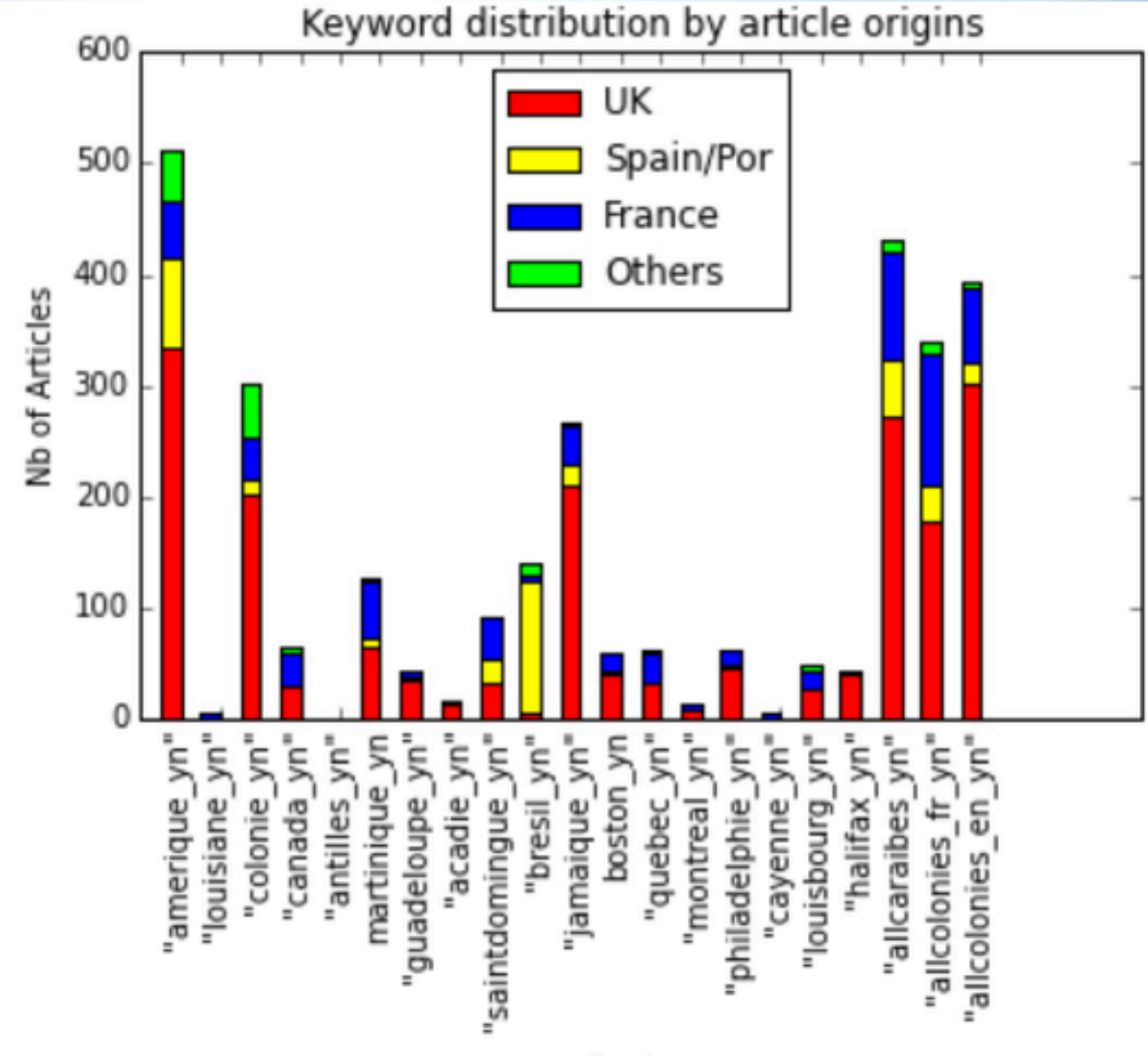
GAZETTE CORPUS METADATA FILE (sample)

id	"ville"	"annee_pub"	"amerique_yn"	"louisiane_yn"	"colonie_yn"	"canada_yn"
am1740_1	Londres	1740	1	0	0	0
am1740_10	Madrid	1740	1	0	0	0
am1740_11	Londres	1740	1	0	0	0
am1740_12	Londres	1740	1	0	0	0
am1740_13	Dresde	1740	1	0	0	0
am1740_14	Madrid	1740	1	0	0	0
am1740_15	Londres	1740	1	0	1	0
am1740_16	Londres	1740	1	0	0	0
am1740_17	Londres	1740	1	0	0	0
am1740_18	Madrid	1740	1	0	0	0
am1740_19	Londres	1740	0	0	1	0
am1740_2	Madrid	1740	1	0	1	0

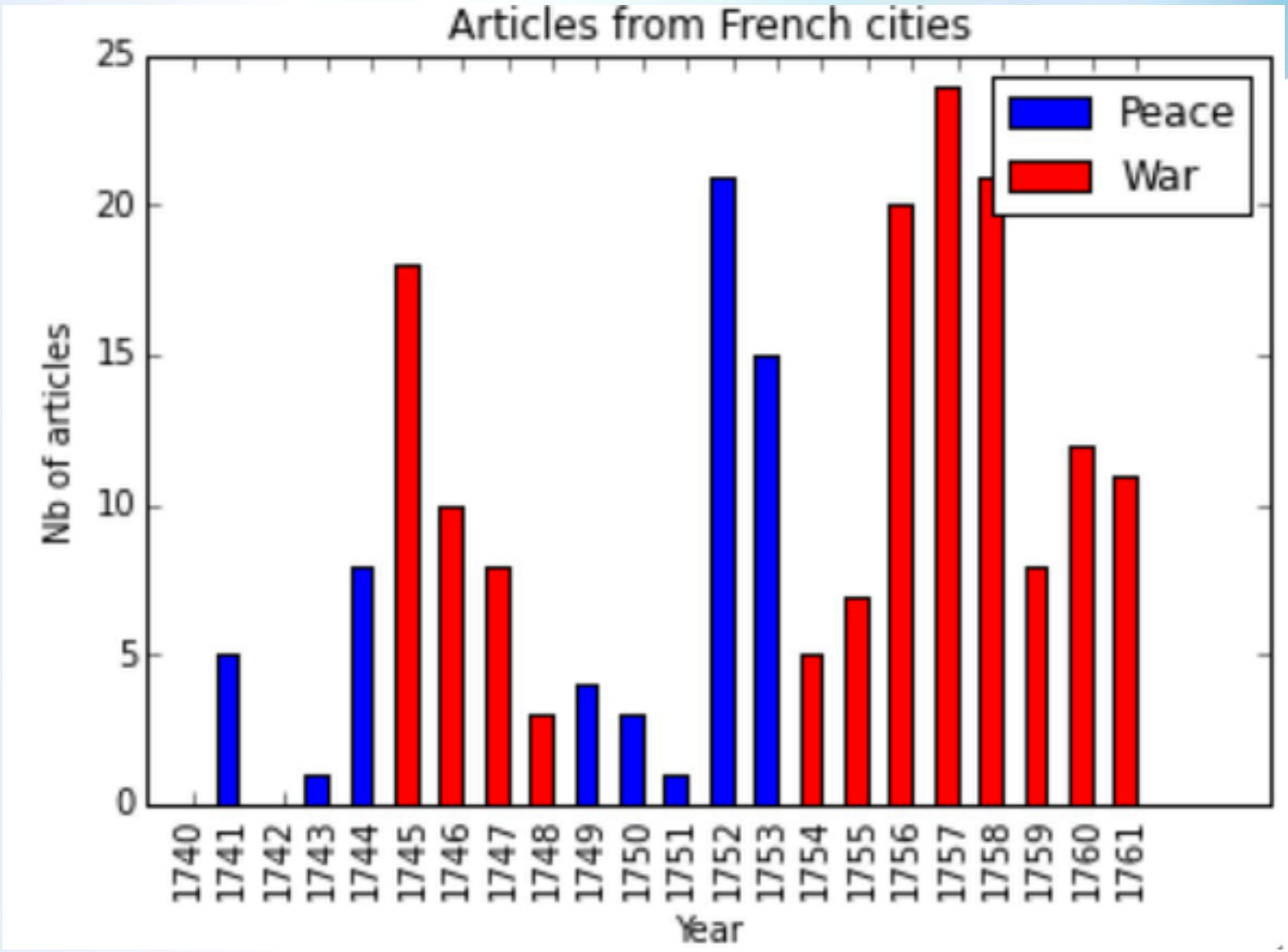
GAZETTE CORPUS — ARTICLE ORIGINS



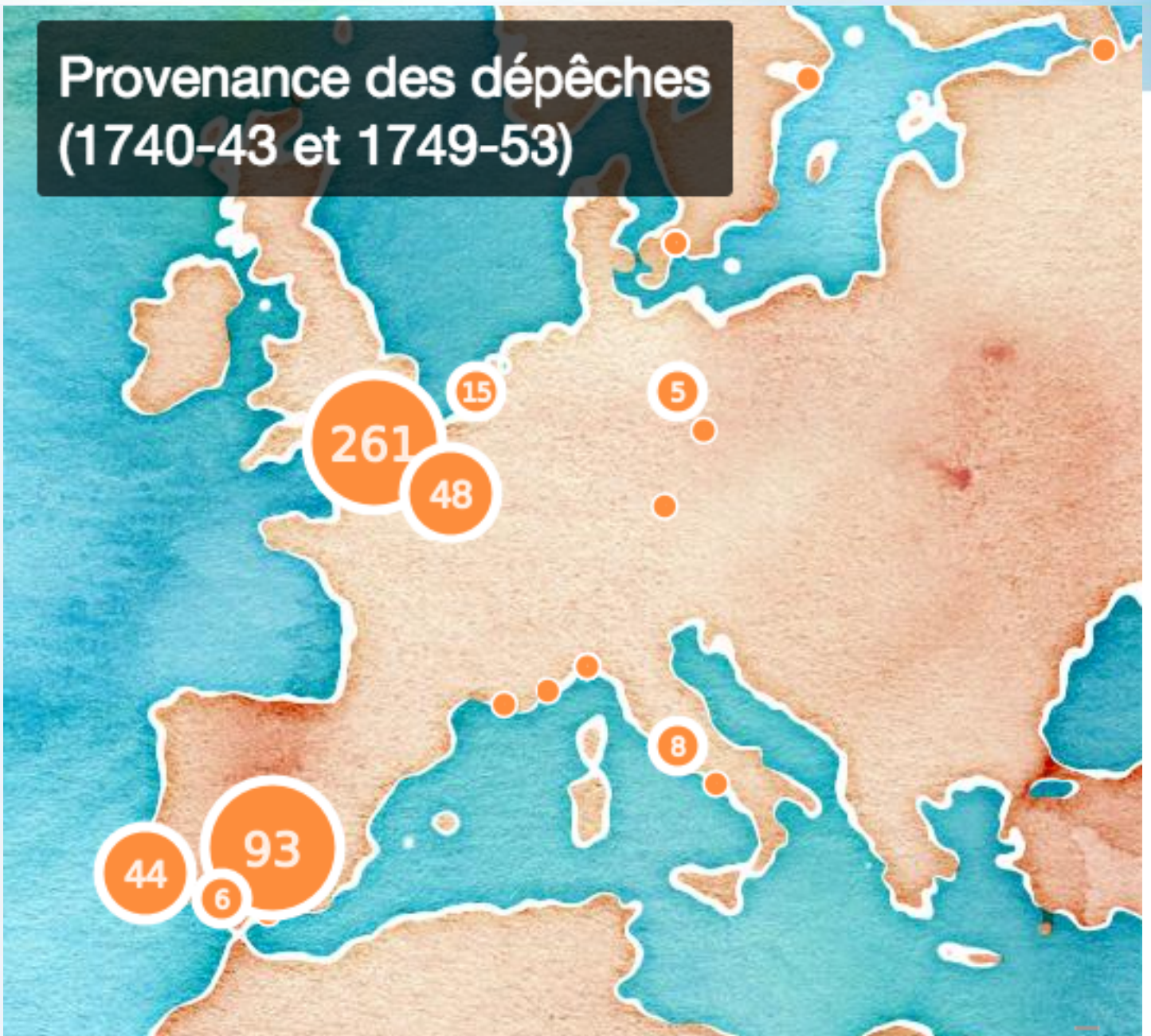
GAZETTE CORPUS — GEOGRAPHIC KEYWORD DISTRIBUTION



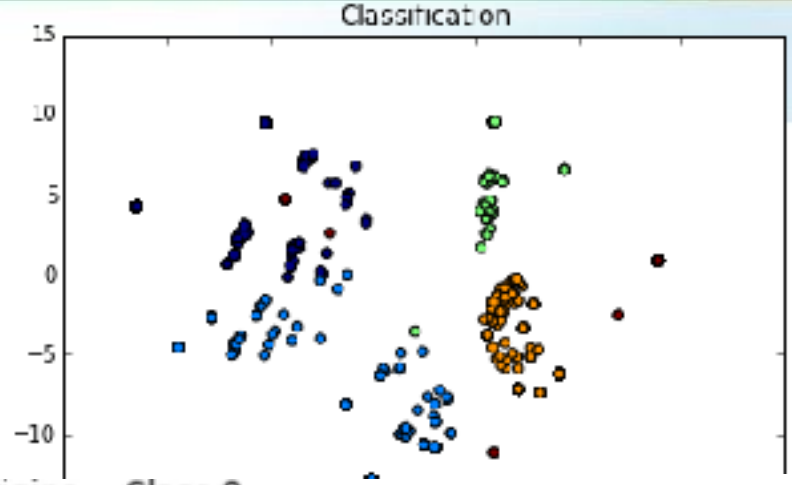
GAZETTE CORPUS — FRENCH ARTICLES AS TIME SERIES

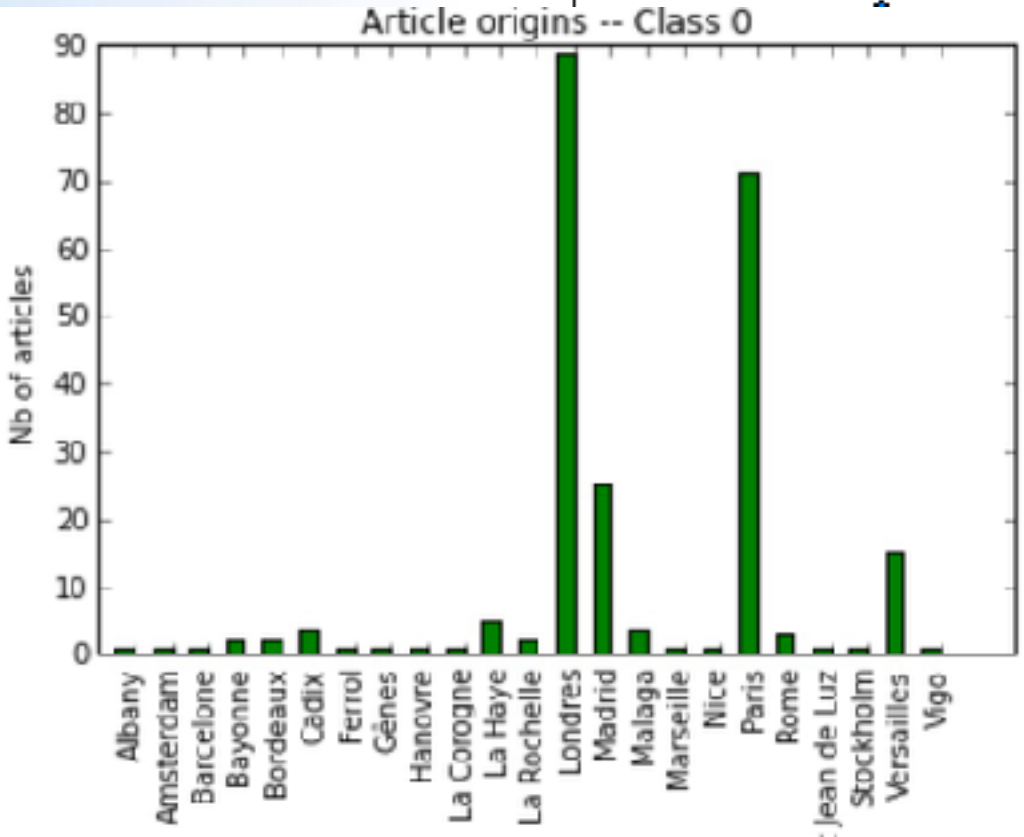


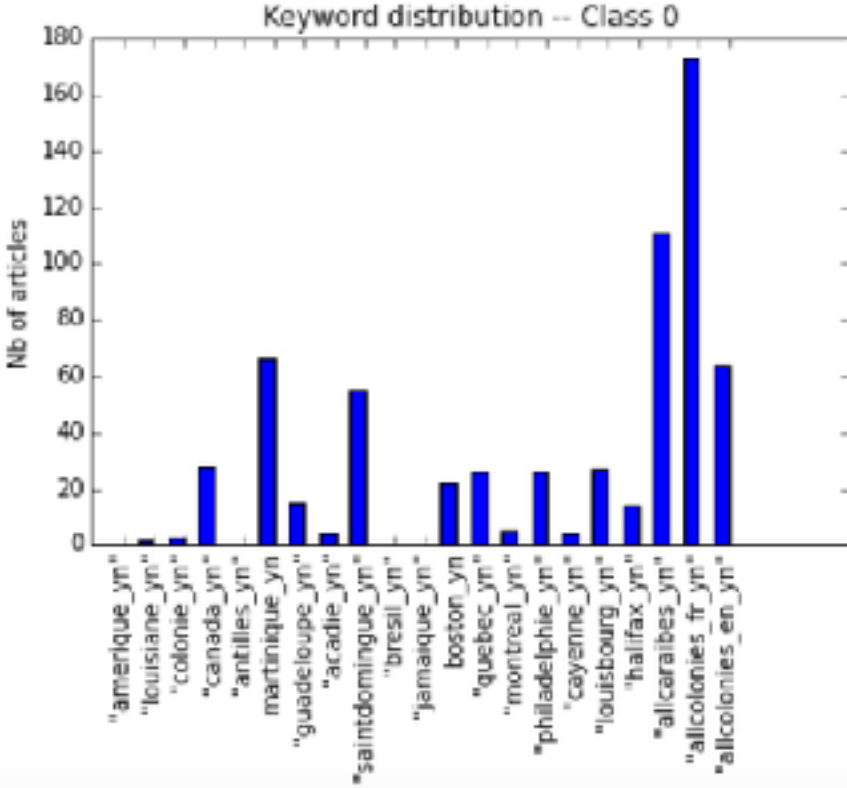
GAZETTE CORPUS — PEACETIME GEOGRAPHIC DISTRIBUTION



GAZETTE CORPUS — K-MEANS CLUSTERING







MERCURE and *JOURNAL DES SAVANTS*

WHEN METADATA FAILS TOO?

- Keyword distribution ==> Informed Close reading
- KWIC and Collocates as last remaining digital tools

THAT IS ENOUGH TO GET INTERESTING RESULTS

- Land of opportunity…
- … but also of great risk …
- … and irrevocably alien …
- … and less welcoming for the French than for others

CONCLUSIONS

THREE TIME SCALES, THREE IMAGINED AMERICAS

- Fast-moving current events: a dangerous, foreign land
- Slow-moving science: opportunities and risk
- Stable collective consciousness: imbalance of rewards

DIGITAL CONTRIBUTION TO THE ANALYSIS?

- More limited than expected
- Nothing groundbreaking
- But *sufficient to support a historical argument*

THANK YOU!

François Dominic Laramée
fdl@francoisdominiclaramee.com
www.francoisdominiclaramee.com
Twitter : @fdlaramee

Work funded by a doctoral grant from the Fonds de recherche du Québec - Société et Culture