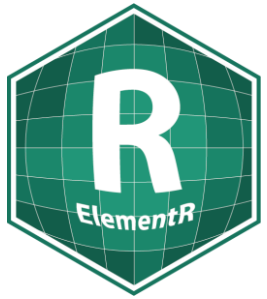


Faire de l'analyse textuelle sur R : pourquoi et comment ?

Table ronde

Groupe ElementR

16 janvier 2024

Les participant·es

Intervenant·es :

Camille Dabestani, Doctorante, UMR Géographie-cités

Mégane Fernandez, Doctorante, UMR Géographie-cités

Paul Gourdon, Ingénieur de recherche, UMR LATTS

Romain Leconte, Agrégé préparateur, UMR CMH

Anne-Cécile Ott, Post-doctorante, UMR CRIS

Etienne Toureille, Maître de conférences, UMR IDEES

Animation :
Léa Christophe, Doctorante, UMR Géographie-cités
Robin Cura, Maître de conférences, UMR Prodig

Déroulé de la table ronde

Tour d’horizon des projets des intervenant·es et de leurs résultats

Retour d’expériences sur les aspects techniques et méthodologiques

Echanges croisés sur différentes questions

Tour d’horizon des projets mobilisant de l’analyse textuelle

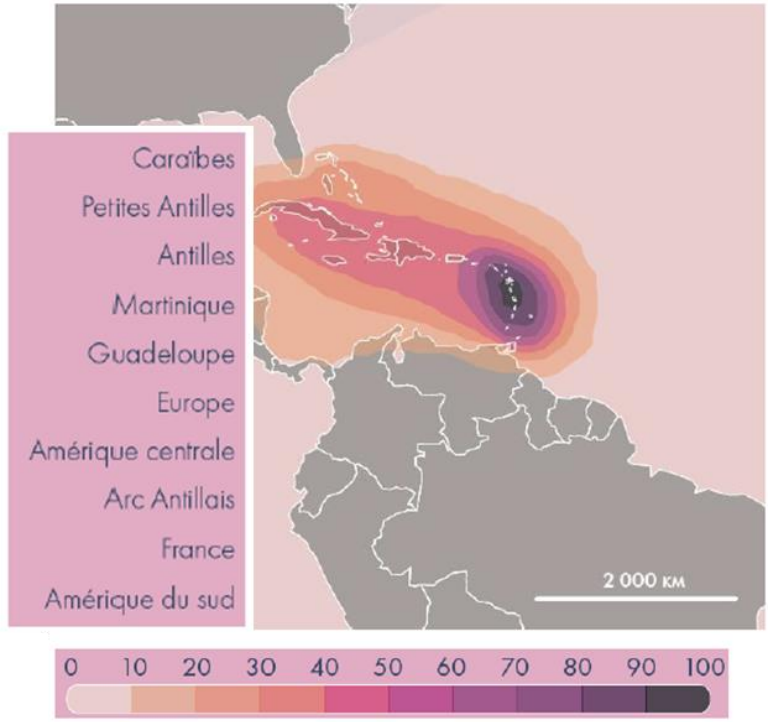
Etudier les supports de communication produits par les organisations régionales impliquées dans la Caraïbe

Camille Dabestani

Contexte global du projet présenté et question de recherche

Thèse en cours sur la construction des imaginaires macrorégionaux
→ focus sur la Caraïbe depuis les contextes guadeloupéen et martiniquais

- Pratiques et représentations socio-spatiales des macrorégions mobilisées (Antilles, Caraïbe, Amérique(s), etc.) et mises en relation dans les imaginaires (entre-deux, multi appartenance)
- Caraïbe comme construction, objet politique macrorégional flou et mouvant
- Catégories institutionnelles et politiques associées aux territoires français dans ce contexte (“régions ultrapériphériques”, “outremer”)



Dabestani & Marveaux. “Imaginer sa région du monde”. Poster présenté à la JIG 2022.

Trois corpus mobilisés :

ETUDIANT·ES : questionnaire cartographique numérique (dont cartes mentales), entretiens avec des étudiant·es à l’Université des Antilles

PRODUCTION DE SAVOIRS : champ académique, programmes scolaires, Wikipédia

INSTITUTIONS : discours produits par les organisations régionales impliquées dans la Caraïbe, entretiens “en appui”

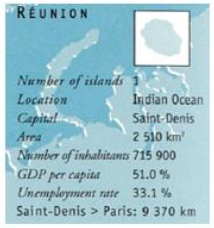
Etudier les supports de communication produits par les organisations régionales impliquées dans la Caraïbe (AEC, OECO, UE, Caricom) comme analyseur de production d’imaginaires « légitime » et dominant (Cussó & Gobin, 2008)

Corpus et analyses à effectuer

Corpus exploratoire :
6 publications officielles de la Commission européenne portant sur les RUP* de 1999 à 2020 de 2 à 50 pages selon les supports

Analyses à effectuer :

- Détecter et catégoriser les lieux, espaces et échelles
- Identifier les champs lexicaux et termes-clés employés



1999 and 2002


2010

2012 and 2017

*Régions ultrapériphériques de l'UE

Résultats obtenus

- Du territoire-Etat-UE au macrorégional

→ analyse lieux/échelles : changement d'échelle des entités spatiales dans le temps (avant 2002 : capitales/Etats vs. 2020 : organisations régionales/Etats voisins)

Table of residuals (chi² test) on the spatialities mobilised in the brochures (1999-2020)

	EU		States			Macroregions			OTs		Cities				
	EU	Admin. status (national or EU)	EU	Non EU	OTs	Organizations (non EU)	Europe	Other	OTs	Parts of OTs territories	Metropoles	Ots	EU	Non EU	Univ. / research centers
1999	3,93	-1,33	-1,33	-1,95	-0,87	-0,60	0,54	-1,37	-0,37	-2,16	2,42	2,46	5,33	-0,37	-1,67
2002	2,44	-1,20	-1,20	-1,74	-1,37	-0,54	-0,09	-1,11	-1,11	-1,93	2,99	3,75	2,65	-0,33	-0,83
2010	-1,01	1,22	-1,07	-1,74	0,25	-1,38	0,33	0,05	0,65	0,76	0,24	-0,21	0,11	-0,84	0,39
2012	-0,78	1,12	2,98	-1,52	0,38	-1,46	0,00	-1,54	0,47	1,30	-0,44	-1,21	-1,44	0,22	0,68
2017	0,02	-0,04	0,25	-0,75	0,04	0,32	-0,76	-1,73	0,12	1,47	-0,61	0,98	-1,29	0,07	0,01
2020	-1,31	-2,00	-2,00	8,76	0,34	4,62	0,53	7,01	-1,04	-3,23	-2,00	-3,32	-0,73	1,25	-0,10

- Du développement territorial à la coopération régionale

→ champs lexicaux (et formats) : des "territoires lointains et exotiques" (format atlas) au document de développement territorial

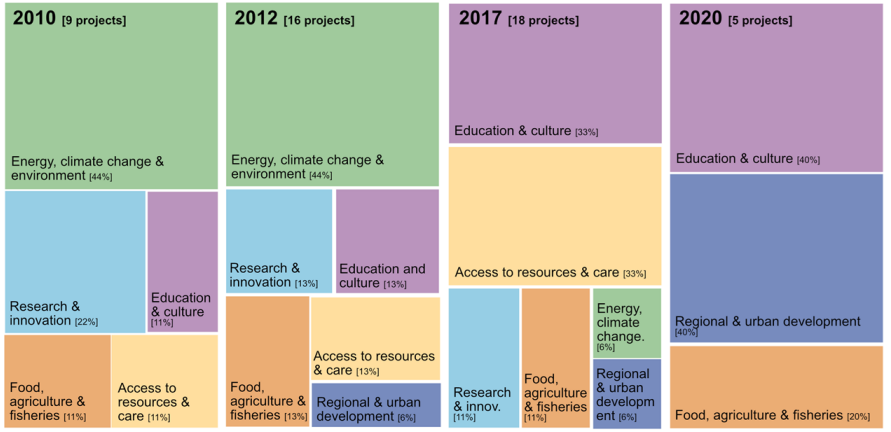
1999	2002	2010	2012	2017	2020
km	km	islands	project	project	caribbean
remote regions	ORS	km	regions	european	islands
Canary islands	GDP per capita	activities	islands	islands	project
Madeira	unemployment	guadeloupe	european	regions	cooperation
regions	EU	inhabitants	outermost	Martinique	TEECA
EU	Azores	madeira	research	regional	COCOA
azores	Madeira	martinique	euro	new	Martinique
area	regions	azores	EU	outermost	French
location	location	climate	development	euro	IHOI
specific	area	French Guiana	Réunion	development	IVY

remote/distance development projects (macro)regional coop.

- Fonctionnement par projet aux échelles macroréginales

→ champs lexicaux (et focus sur les thématiques des projets) : projets de développement territoriaux locaux (2010-17) aux projets macros (macrorégions et Etats voisins) de coopération (ex : économie de la connaissance)

Thematics of projects presented in booklets (1999-2020) [thematical groups are those of 2021, CE Website]



- Retrait de la parole institutionnelle et neutralisation

→ parole : multiplication des acteurs non-institutionnels

Etudier les terrains de recherche à l’échelle mondiale et pas dans une région donnée

Mégane Fernandez

Contexte global du projet présenté et question de recherche

- Thèse : “Les terrains de la recherche sur projet : dimensions structurelles, collectives et individuelles de la localisation des terrains de recherche

- Etat de l’art à la croisée de différents champs de recherche et de leurs méthodologies :
  - Approches critiques de la science
    Au sein des STS : épistémologies décoloniales et féministes, *ignorance studies* etc.
  - Géographie des sciences
  - Scientométrie spatiale
  - Humanités numériques

Etudier les terrains de recherche à l’échelle mondiale et pas dans une région donnée

Corpus et analyses à effectuer

Base de données utilisée : Cordis, Union européenne.
Données sur les projets de recherche et innovation financés par le programme de financement Horizon 2020 (2014-2020).
Informations sur les titres, résumés, participant.es, thèmes et domaines de recherche, montants du financement, type de financement etc.

Analyses à effectuer :

- Reconnaissance des entités nommées spatialisées
- Statistiques
- Analyses de réseau
- Cartographie ?

Résultats obtenus

- Travail toujours en cours
- Complété par des entretiens

Espaces les plus étudiés par les projets européens :

Macrorégion	Projets
Europe	516
Africa	78
EU	75
Earth	53
Mediterranean	49
Arctic	38
Middle East	27
Latin America	22
Asia	19
North Africa	19
South Africa	19

Pays	Projets
Italy	107
France	86
UK/United Kingdom	70
Germany	69
Spain	61
China	49
Egypt	35
Greece	34
Turkey	33
Belgium	29
Ireland	28

Différenciation des approches en fonction des espaces étudiés:

Europe

	n	%	val%
agricultural sciences	39	3.0	3.1
engineering and technology	63	4.8	5.0
humanities	302	23.2	24.0
medical and health sciences	96	7.4	7.6
natural sciences	270	20.7	21.4
social sciences	489	37.5	38.8
NA	45	3.5	NA

Mediterranean

	n	%	val%
agricultural sciences	6	4.9	5.0
engineering and technology	11	8.9	9.2
humanities	46	37.4	38.3
natural sciences	35	28.5	29.2
social sciences	22	17.9	18.3
NA	3	2.4	NA

Middle East

	n	%	val%
agricultural sciences	1	1.7	1.8
engineering and technology	1	1.7	1.8
humanities	31	53.4	55.4
medical and health sciences	1	1.7	1.8
natural sciences	4	6.9	7.1
social sciences	18	31.0	32.1
NA	2	3.4	NA

Africa

	n	%	val%
agricultural sciences	19	9.7	9.9
engineering and technology	16	8.2	8.3
humanities	34	17.4	17.7
medical and health sciences	21	10.8	10.9
natural sciences	44	22.6	22.9
social sciences	58	29.7	30.2
NA	3	1.5	NA

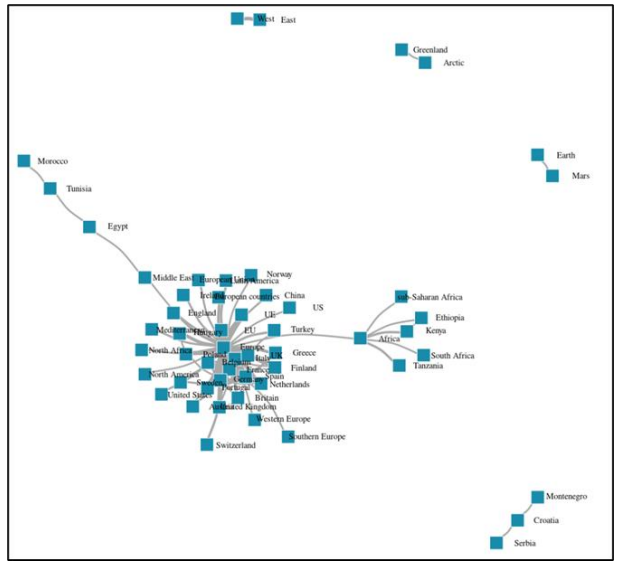
Arctic

	n	%	val%
agricultural sciences	1	1.2	1.3
engineering and technology	3	3.6	3.8
humanities	2	2.4	2.6
medical and health sciences	1	1.2	1.3
natural sciences	62	74.7	79.5
social sciences	9	10.8	11.5
NA	5	6.0	NA

Latin America

	n	%	val%
agricultural sciences	1	1.8	1.9
engineering and technology	2	3.6	3.7
humanities	6	10.9	11.1
medical and health sciences	14	25.5	25.9
natural sciences	6	10.9	11.1
social sciences	25	45.5	46.3
NA	1	1.8	NA

Espaces étudiés ensemble :



Analyser les politiques urbaines

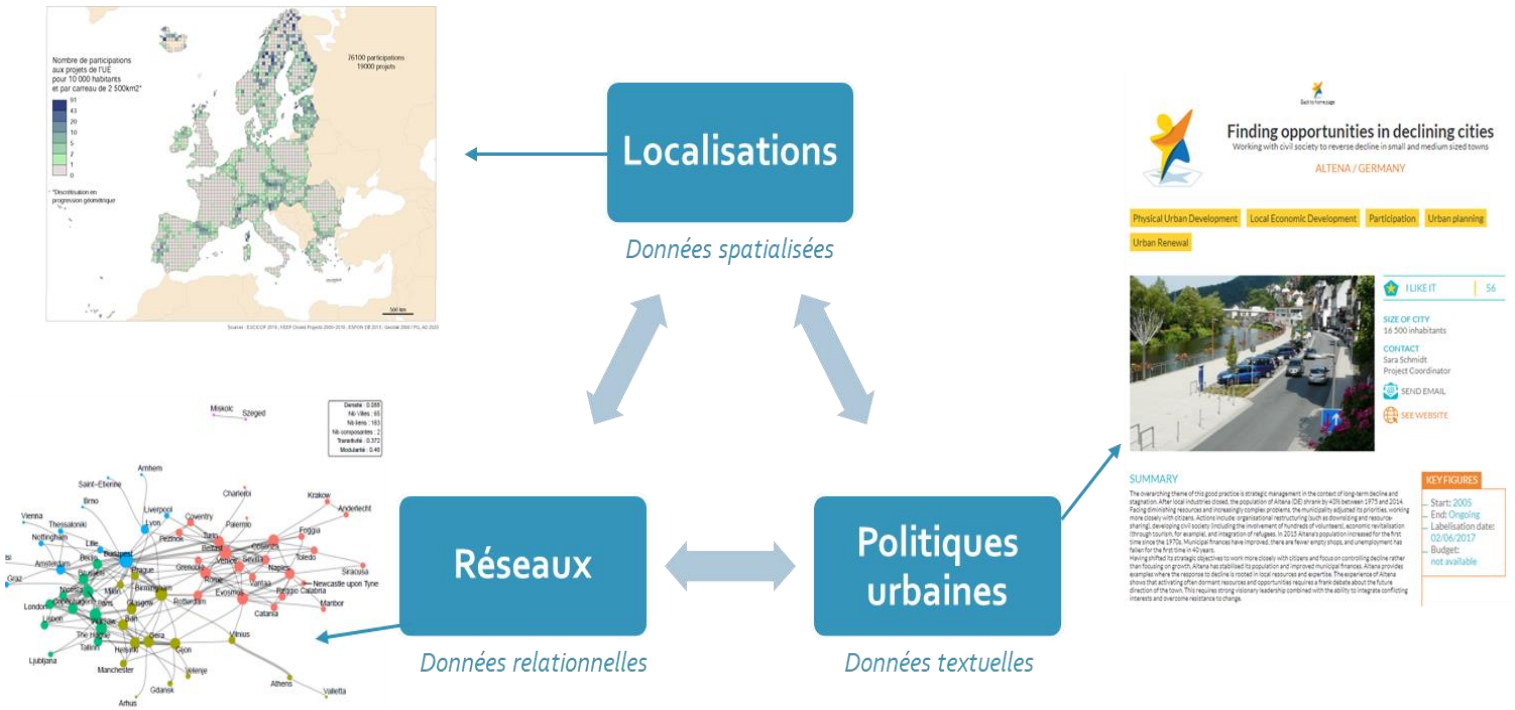
Paul Gourdon

Contexte global du projet présenté et question de recherche

- Une thèse sur la coopération politique entre villes européennes
- Une démarche exploratoire des données en 3 temps

Analyse textuelle des politiques urbaines :

a. **Saisir les discours des organisations transnationales sur les villes et les thèmes de la coopération**
b. **Voir comment les politiques locales sont transformées en “bonnes pratiques” (*best practices)***



Corpus et analyses à effectuer

Plusieurs corpus, notamment

- 87 rapports de projet URBACT, (2007-2020) = rapports préliminaires (baseline study) rédigés par les experts et exposant l'objet du projet de coopération
- 97 "bonnes pratiques" telles que labellisées par le programme en 2017

Analyses à effectuer :

1. **Analyses morpho-syntaxiques et nettoyage du corpus** : étiquetage des unités lexicales selon leur fonction morpho-syntaxique et transformation des textes en tableaux de données (une ligne par unité lexicale)

2. **Analyse lexicale** : fréquences d'utilisation des mots au sein des textes ou dans l'intégralité du corpus.

3. **Détection des mots-clés** : détection des locutions (groupes de mots) très fréquentes dans les textes, co-occurences de mots au sein d'unités textuelles (documents, paragraphes, phrases).

4. **Analyse des similarités et mots-spécifiques** : repérage des spécificités ou similarités lexicales entre les textes d'un corpus.

5. **Modélisation de thèmes** : construction automatique de thèmes (représentés par un ensemble de mots qui ont une forte co-occurrence au sein des documents) dans l'ensemble d'un corpus.

Résultats obtenus

- Une langue spécifique : proche des corpus d'anglais touristiques (- de verbes et + d'adjectifs)
- Un lexique du projet et du partenariat : primat de l'économie et de la "gouvernance", des textes extrêmement proches lexicalement
- Politiques urbaines présentées comme trans-sectorielles mais des oppositions nettes
- Un langage néolibéral : *politique de l'habilitation*

Un espace sémantique : mots-clés des projets URBACT

Faire la sociogenèse des manières enfantines de représenter le monde

Anne-Cécile Ott

Contexte global du projet présenté et question de recherche

- Thèse (soutenue en 2022) sur les **représentations enfantines du monde** :
  ➔ *Comment les enfants (se) représentent le monde ?*
  ➔ *Pourquoi ils le représentent d'une manière plutôt qu'une autre ? = Faire la sociogenèse des manières enfantines de représenter le monde*

- Enquête de terrain dans 4 écoles élémentaires parisiennes aux profils sociaux contrastés
  ➔ *248 enfants de CP, CE2, CM2*

- **Dispositif pluri-méthodologique**
- Observation
- Entretiens avec les enseignants
- Analyse de supports pédagogiques

PHASE	MÉTHODE	DÉROULEMENT
1	**Dessin commenté**	En classe entière Dans la classe ± 1h
2	***Brainstorming***	En classe entière Dans la classe ± 30-40 min
3	**Reconstitution d'un planisphère illustré**	Par groupes (5-7 enfants) Hors de la classe ± 30-40 min/gp
4	**Entretien**	Par groupes (2-4 enfants) Hors de la classe ± 20 min/gp

Corpus et analyses à effectuer

- Utilisation de l'analyse textuelle sur **2 corpus principaux** et 1 corpus complémentaire :
  - ➔ Commentaires des dessins
  - ➔ Réponses à la question 3 des entretiens (Q3) : "*Si vous deviez expliquer à quelqu'un ce que c'est le monde, vous diriez quoi ?*"
  - ➔ Mots et expressions du brainstorming en classe entière

Analyses à effectuer :

- Nettoyage des corpus et lemmatisation
- Tentative de POS tagging (*part of speech*)
- Bilans lexicaux
- Analyses lexicales de fréquences des termes, de co-occurrences
- Analyses des termes spécifiques à différents groupes d'enfants (selon l'âge, l'école, la classe, le genre, la classe sociale)

Résultats obtenus

- Dresser le **panorama du monde décrit par les enfants**

  ➔ Un monde polysémique : analyse de la dispersion lexicale et des termes utilisés
  ➔ Aide à la construction d'une typologie des représentations enfantines par analyses des domaines d'objets, échelles, significations associées au monde. *3 représentations idéaltypiques*
  ➔ Un monde entre convergence et différenciation : grande dispersion lexicale mais des récurrences
  ➔ Prépondérance du triptyque "monde-terre-planète(s)" : des représentations du monde à l'échelle globale et englobante ; une planétarisation des représentations du monde ?

- Des logiques de **différenciations sociales**. L'exemple de l'âge :

  ➔ Les enfants les plus jeunes ont été plus prolixes que leurs aînés
  ➔ Mais ils ont utilisé moins de termes distincts et d'hapax que les plus grands dans les commentaires des dessins. Plus de précisions et de concision chez les CM2
  ➔ Intéressant : c'est l'inverse pour les réponses à la Q3. Plus d'hapax chez les plus grands. Parfois dilution des propos chez les plus jeunes.

  ➢ Des manières différentes de représenter le monde selon les âges
  ➢ Des questions méthodologiques sur l'interprétation des analyses notamment

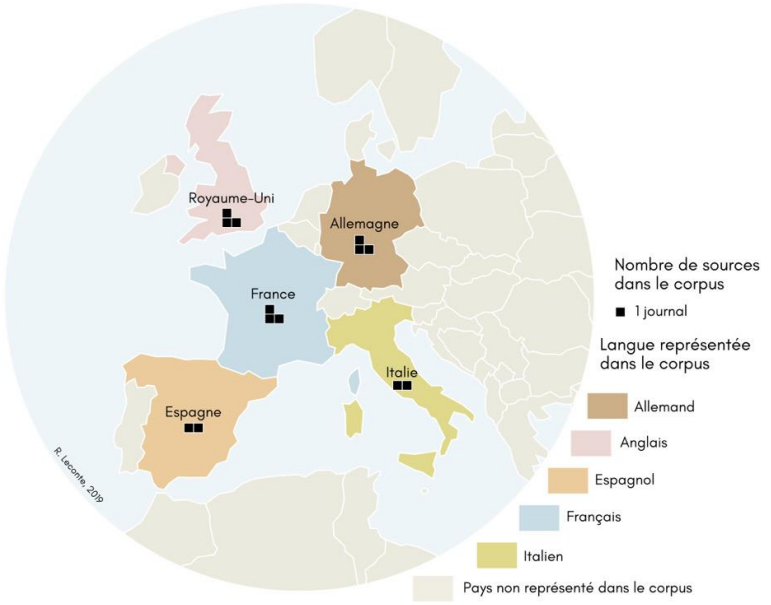
Analyser la médiatisation des frontières dans la presse européenne

Romain Leconte

Contexte global du projet présenté et question de recherche

- Thèse, 2022, Couvrir les frontières du Monde. Médiatisation des frontières et agenda géopolitique de la presse européenne (2013-2019)
  - ➔ Comment la presse couvre le Monde ?
  - ➔ Comment la presse couvre les frontières ?
  - ➔ = construction sociale des problèmes géo-politiques

- La presse quotidienne européenne
  - ➔ Matériau unique de la thèse
  - ➔ 13 quotidiens sur 5 ans
  - ➔ 5 langues



- Questions
  - ➔ Quelles frontières sont médiatisées ?
  - ➔ Les frontières occupent-elles une place de plus en plus importante dans la presse ?
  - ➔ Observe-t-on des basculements ou au contraire une stabilité du sens donné aux frontières ?
  - ➔ Quelles sont les fonctions du signal frontière dans le discours de presse ?
  - ➔ Peut-on observer des formes de convergence des agendas géopolitiques de la presse ouest-européenne dans la médiatisation des frontières ?

Corpus et analyses à effectuer

Un corpus général subdivisé en sous corpus
- accès aux journaux par leurs flux RSS
- textes courts

Tableau 5. Les quatre corpus de la thèse

Corpus	Contenu	Taille	Étiquetage	Chapitre
Complet	Titre + description	4,8 millions		2,2
Étranger	Titre + 1re phrase	728 646	201 pays du monde	2,3
Frontières	Titre + 1re phrase	12 787	« frontière »	3
Frontières + étranger	Titre + 1re phrase	5 666	« frontière » + 201 pays du monde	4

Analyses à effectuer :

1. **Nettoyage, mise en forme et préparation du corpus** : restructuration des données au format corpus ou tableau, nettoyage, dédoublonnage, pondérations, traduction automatisée vers l'anglais, lemmatisation

2. **Étiquetage** : étiquetage spatial (pays) et thématique (frontières, migrations)

3. **Analyse lexicales** : analyses de fréquence, spécificités lexicales, classifications (des documents, des sources et du temps)

4. **Analyse spatiale** : analyses spécifiques appliquées au vocabulaire spatial : flux (circulation de l'information) et réseaux (mise en ordre du Monde)

5. **Modélisation lexico-spatiale** : séquençage de l'agenda médiatique des frontières, convergence/divergence des sources, structures spatiales de la médiatisation du Monde et des frontières.

6. **Analyses quali** : mise en récit des événements

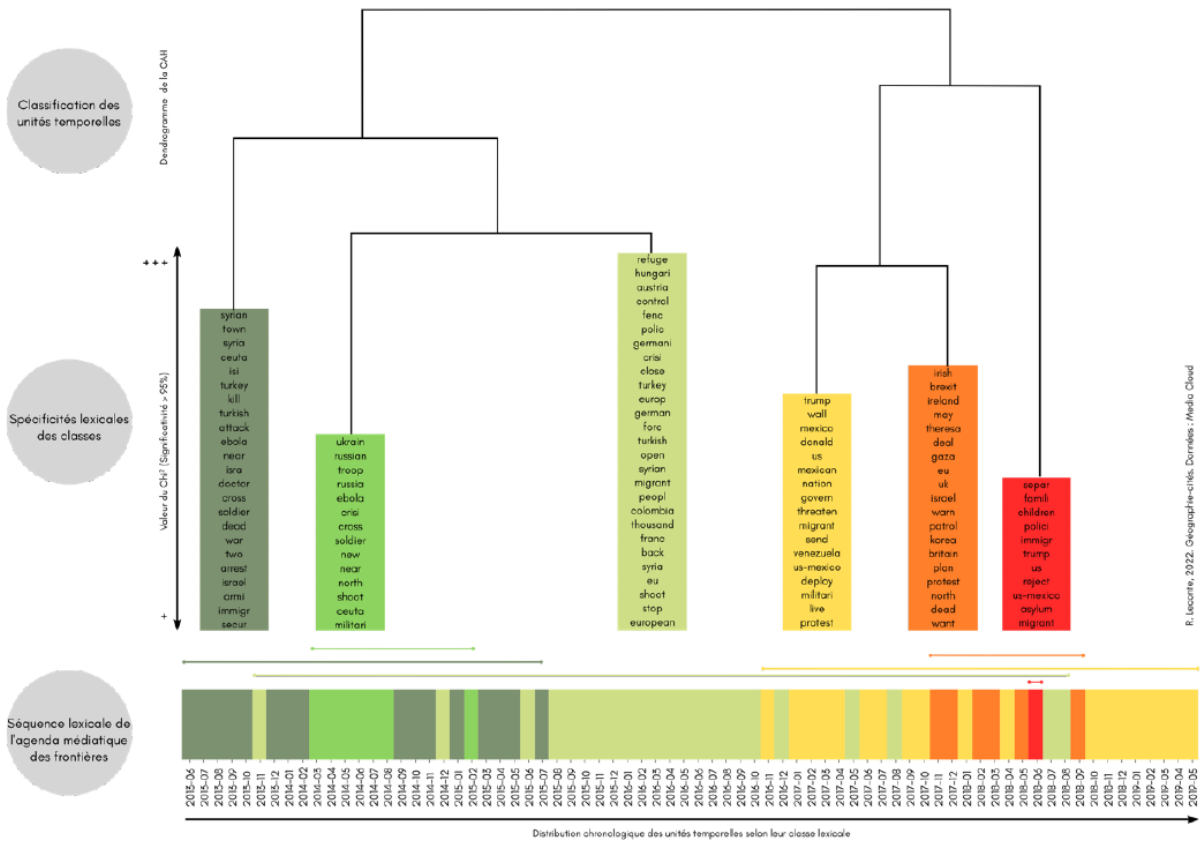
Résultats obtenus

- Sur la couverture du Monde :
  hiérarchie, proximité et colonialité

- Sur le discours sur les frontières :
  sécurité, conflits, migrations

- Sur la médiatisation des frontières :
  un tournant sémantique et spatial en 2015, autour de la question migratoire et sur l'espace européen

- Sur l'espace du corpus :
  2015, un moment de convergence des agendas médiatiques ouest-européens



Contribution de la statistique textuelle à l’analyse des représentations d’objets géographiques flous

Etienne Toureille

Contexte global du projet présenté et question de recherche

En réalité trois “projets”

- ***Thèse de doctorat***: analyse des représentations sociales des limites de l’Europe chez les étudiants turcs (3 ans → 6 ans)
  - Analyse de questions ouvertes (techniques de mots associés): “
- ***post-doc 1- ODYCCEUS*** : analyse de l’agenda-géomédiatique de ladite “crise migratoire” de 2015
  - Corpus de presse quotidienne nationale (textes courts - titres flux RSS et textes)
- ***post-doc 2 ANR IMAGEUN*** : analyser la forme socio-spatiale de grandes régions (macro-régions)

Bilan: Plus qu’un “projet”, l’analyse de texte est au coeur d’une méthodologie qui interroge les discours sur le monde à travers différents langages (textuel, cartographique, etc.)

Proposer une analyse systématique des représentations d’objets géographiques flous

- L’ “Europe” comme région / notion | population: les étrangers à travers leurs représentations médiatiques

=> analyse exploratoire de matériaux textuels bruts (questionnaires / données de presse).
=> détection d’entités (identification d’objets spatialement objectivés, recours à des onthologies)

Corpus et analyses à effectuer

Une formation en analyse quantitative de données, dont statistique textuelle

- ***Master géoprisme, formations INED*** (analyse de données).
- ***Usage préalables de logiciels hors R*** (SPAD, Iramutec, Rthemis - via RCmdr).

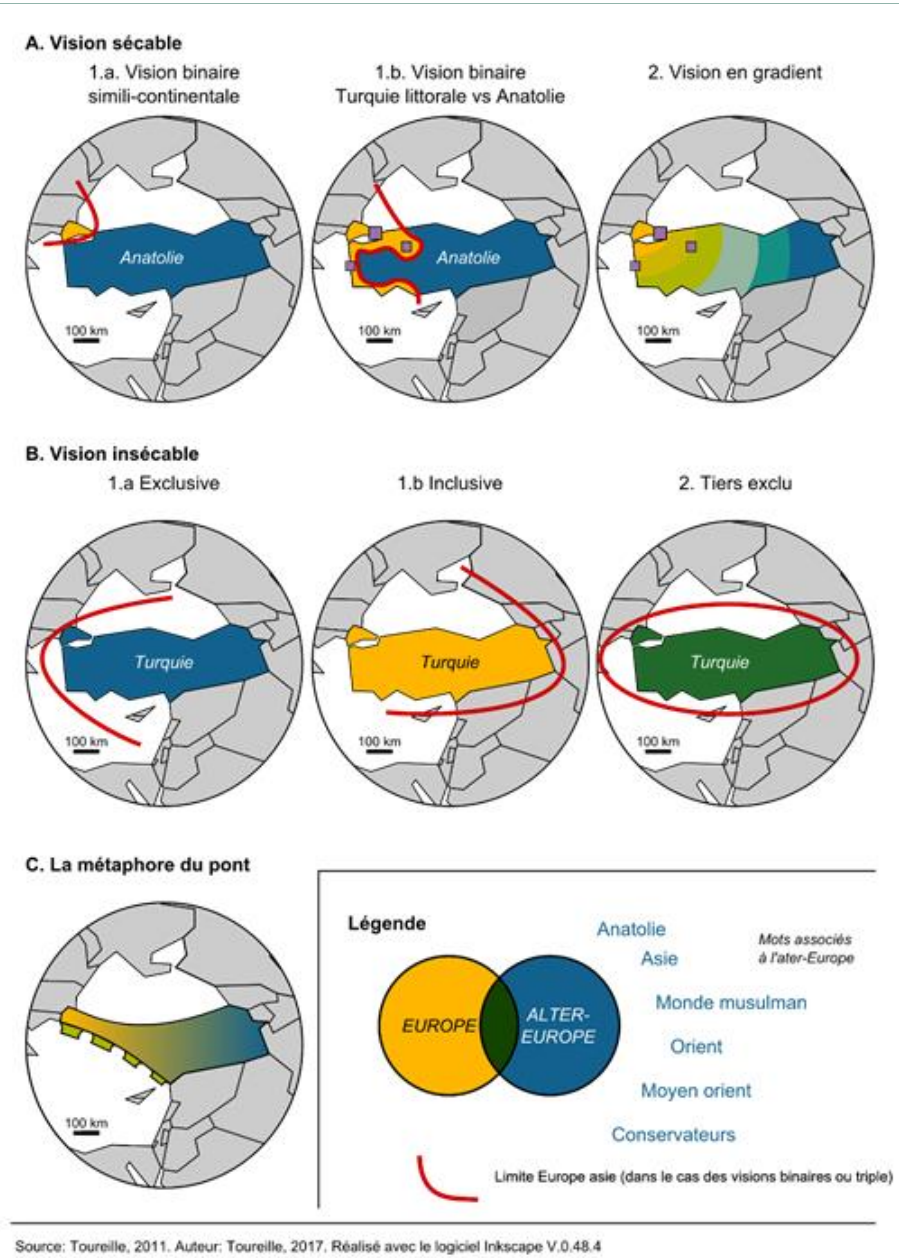
→ R reste un outil, passage progressif compte tenu de l'obsolescence des autres outils (choix de Quanteda)

→ Choix de R compte tenu d'habitudes personnelles et de l'environnement de travail (géo. / stat.).

Objectifs / méthodologies et outils R utilisés: transposition de méthodes dans R

- ***Analyse exploratoire de questionnaires d'enquête (statistique textuelle - Lebart et Salem, 1992)***
  - Analyse des spécificités lexicales – chi2 via Quanteda
  - Analyses factorielles (AFC/ ACM) – packages généralistes ade4 / Factominer sur le TLA
  - Classifications (CAH sur TLA , CDH de Reinert) – Rainette (J. Barnier)
- ***Analyse de la saillance de sujets d'intérêt identifiés par des mots clefs (migrants)***
  - Etiquetage via des dictionnaires (listes de pays) – Quanteda
  - Calculs de saillance
  - Confrontation des sous-ensembles avec le vocabulaire utilisé.
- ***Analyses interlinguistiques et croisement de corpus différents*** (usage / construction d'onthologies)

Résultats obtenus 1



Source: Toureille, 2011. Auteur: Toureille, 2017. Réalisé avec le logiciel Inkscape V.0.48.4

Tableau 5. – Vocabulaire spécifique des étudiants en fonction du type de tracé appliqué à la Turquie.

Type de tracé	Nb de réponses	En % de répondants	Nb de mots distincts	% de mots distincts	Nb. moyen de mots par réponse	Mots spécifiques*	Exemples
Anatolie	36	10 %	89	63 %	3,9	Modern, fashion, contempo-rary, liberty	negociations, modern, crisis, germany, France (étudiant n° 55)
Bosphore	173	20 %	295	40 %	4,3	Us, christianity, euro, italy, history	euro/pope/ christianity/ fashion/ freedom (étudiant n° 194)
Complètement intégrée	69	19 %	170	55 %	4,6	Development, education, power, language, wealth, colonialism	Self/reliance/ education/ money/ language/ development (étudiante n° 334)
Complètement exclue	86	24 %	218	60 %	4,2	Racism, colo-nial, football, selfishness, the**, exploitation	Exploitation/ racism/crisis/ benefit/ insincerity (étudiant n° 459)

*Seules les formes significatives sont représentées ici.

** Forme non prise en compte dans l'analyse (mot vide), n = 364 étudiants.

Résultats obtenus 2 (limites et perspectives)

Limites

- La statistique textuelle (comme l'analyse de données) prend du temps, des éléments non-automatisables (construction de corpus, sélection des formes, compréhension du matériau).
- Un coût de formation en entrée, notamment pour l'acquisition des outils avancés (NLP et méthodes supervisées, par exemple) ou des conceptualisations extra disciplinaires (linguistique, informatique).
- Certaines méthodes sont plus accessible dans d'autres langages (Python, cf. NLP et plus généralement outils de linguistique: méthodes d'apprentissage supervisées).

Si c'était à refaire

- Découverte tardive du caractère accessible des méthodes recourant à des modèles de langue (étiquetage morphosyntaxique).
- Créer des collectifs pour échanger sur ces méthodes et se rassurer, penser interdisciplinaire (même si ce n'est pas facile, cf. linguistes).

Perspectives

- Arrêter d'être dépendant de R (passage à Python sur certains outils - cf. Spacy plutôt que Spacyr)
- Formation aux méthodes recourant au NLP (vectorisation de texte, méthodes d'apprentissage pour les opérations déductives nécessitant une supervision).

Retour d’expériences sur les aspects techniques et méthodologiques

	Types d'analyses	Packages utilisés	Environnement informatique	Ressources d'apprentissage	Pistes abandonnées
Camille	Lexiques, mots-clés	quanteda, udpipe	RStudio en local (sous Windows)	Tutos quanteda, bibliographie, ressources locales	
Mégane	POS tagging, reconnaissance des entités nommées	Spacyr, tidytext, quanteda, stringr	RStudio, en local et sur huma-num	Documentation des packages, forums, cheat sheets	irec pour recodage, udpipe
Paul	lexiques, mots-clés, thèmes, similarités, réseau de concept	spacyr, tidytext, udpipe	RStudio, en local et sur huma-num / + Cortext-Manager	vignette des packages, Stack Overflow	analyses temporelles
Romain	Lexiques, mots-clés, analyse de réseau	Quanteda, rainette,translateR	RStudio sur Huma-num	Doc packages, ressources locales (Géo-cités, projet)	Word embedding, NLP
Anne-Cécile	lexiques, mots-clés, similarités/différences	Rtemis	RStudio en local (sous Windows)	Carnet hypothèse pour Rtemis, ressources locales (=Paul)	POS tagging, analyse des correspondances
Etienne	Exploratoire/inductive; Déductive /étiquetage	Quanteda, Tidytext, Spacyr, Rainette, spacyr	Rstudio en local et sur R-Huma num	Formation initiale, INED, tuto (exemple Quanteda), manuels.	

Echanges croisés sur différentes questions