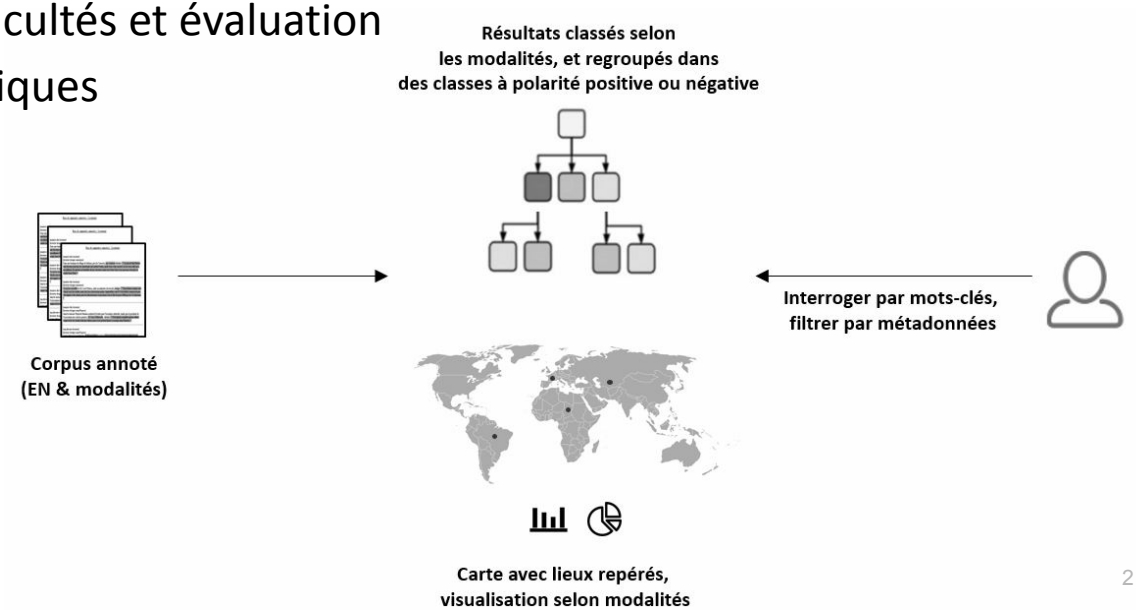
SAGEO, 5 mai 2021

Paris dans les récits de voyage d'écrivains arabes : repérage, analyse sémantique et cartographie de toponymes

Motasem Alrahabi (Sorbonne Université)
Carmen Brando (EHESS)

Plan

- Objectifs de l’étude
- Analyse spatiale de textes littéraires : les noms de lieu
- Corpus de travail
- REN pour la langue arabe: difficultés et évaluation
- Analyse des modalités linguistiques autour des toponymes
- Interprétation des résultats
- Cartographie
- Conclusion et perspectives



Analyse spatiale de textes littéraires: état de l'art

- Un lieu, un espace géographique avec une signification d'après un groupe de personnes (Tuan 1977), cela rejoint la notion d'espace vécu (Lefebvre 1991)
- En études littéraires, cartographier un récit de fiction permet de mesurer son spatialité et d'investiguer comment les auteurs, à travers de personnages, s'approprient d'un espace particulier, et sur plusieurs angles, d'après un courant, un auteur ou un groupe, à une époque ou plusieurs époques (Moretti 1999, Cooper et Gregory 2011, Julia Krôger 2021)
- Intérêt pour la manière dont les lieux sont désignés dans les textes (récits, romans) ainsi que pour leur catégorisation (toponymes, odonymes, lieux naturels, lieux de vie, lieux évoqués) et localisation dans un espace (géographique réel, fictionnel, ou abstrait)
- Besoin de cartographier les lieux de fiction revés (Piatti et al. 2009), le ressenti attaché à des lieux (Dominguès et al. 2019), ou encore les chronotopes de Bakhtine (1978) (Cartographies chronotopiques, Aron et al. 2017)

Méthodes mobilisées dans ces travaux :
linguistique outillée, **TAL, SIG**, analyse réseau

Corpus de travail

- Paris a une importance particulière dans l'imaginaire arabe et surtout à l'époque de la Renaissance culturelle du 19° siècle:
  - Déclin de l'Empire ottoman, campagne de Napoléon en Egypte et influence de la civilisation occidentale…
  - Renaissance littéraire, politique, culturelle et religieuse…
  - Réformateurs de toutes confessions: al-Tahtawi, al-Boustani, al-Yaziji, F. Marrache, J. Zaydan, al-Kawakibi…
- Le récit de voyage de 6 écrivains arabes entre 1830 et 1954 (environ: 265 614 mots)

Auteur	Titre	Date publication	Nombre de mots
رفاعة رافع الطهطاوي **Rifa'a al-Tahtawi**	تخليص الإبريز في تلخيص باريز ***L'Or de Paris***	1834	70231
أحمد فارس الشدياق **A. Fares Al-Shidyaq**	كشف المُخبَّا عن فنون أوروبا ***La découverte de ce qui est caché dans les arts de l'Europe***	1857	25870 *(sur 105743)*
أحمد زكي **Ahmad Zaki**	الدنيا في باريس ***L'Univers à Paris***	1900	50271
جرجي زيدان **Jurji Zaydan**	رحلة إلى أوروبا ***Voyage en Europe***	1912	16163 *(sur 26052)*
زكي مبارك **Zaki Mubarak**	ذكريات باريس ***Souvenirs de Paris***	1931	52690
مالك بن نبي **Malek Bin Nabi**	مذكرات شاهد للقرن ***Mémoires d'un témoin du siècle***	1965	50389

Corpus de référence et guide d’annotation

- Gold: premier tiers de chacun des six livres: environ 88 538 mots (lien).
- Annotation manuelle: 3 annotateurs avec l’outil Inception (Klie et al. 2018)
- Catégories retenues dans notre guide d’annotation (inspiré du modèle ESTER-2):
  - Lieux naturels et géographiques (Loc-Nature): نهر السين / la Seine...
  - Régions administratives (Loc-Admin): الحي اللاتيني / Quartier Latin...
  - Bâtiments et constructions (Loc-Building): دار الأوبرا / l’Opéra...
  - Chemins et axes de trafic (Loc-Path): بولفار سان ميشيل / boulevard Saint-Michel…
- Accord entre annotateurs 0.78 selon la mesure Kappa Fleiss
- Divergences: lieux historiques inconnus par certains annotateurs (*Chambre des pairs…*), confusion entre LOC-Admin et LOC-Nature (*Île-de-France…*), etc.

REN pour la langue arabe

- Difficultés:
  - Absence totale ou partielle des signes diacritiques (عمان: عمّان / عُمان)
  - Les lettres capitales n'existent pas.
  - L'agglutination de certaines particules, parfois combinées, aux noms propres (ليون / بليون)
  - Ambiguité: homonyme (ويشقها نهران أحدهما يقال له: نهر السين بفتح السين), etc.
- Quelques outils:
  - Stanza (Qi et al. 2020): librairie Python qui utilise le moteur CoreNLP de l'Université de Stanford.
  - Farasa (Abdelali et al. 2016): une suite d'outils pour le traitement de textes en arabe qui propose un module de REN basé sur un apprentissage semi-supervisé et des ressources multilingues de Wikipédia.
  - Madamira (Pasha et al. 2014) est également une suite d'outils pour le traitement de langue arabe : analyse morphosyntaxique, diacritisation, annotation des parties de discours, REN, etc.

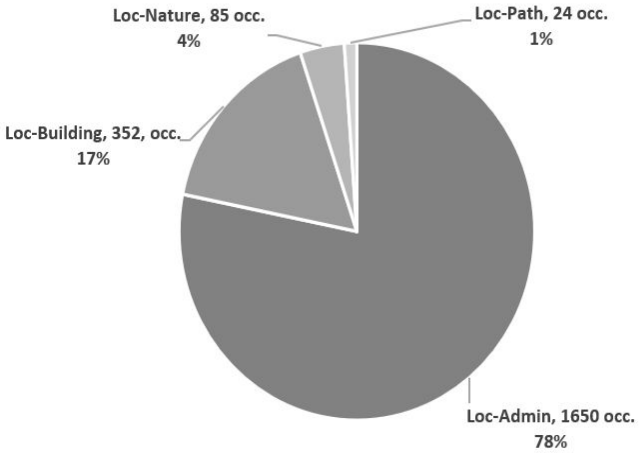
Évaluation de 3 outils de REN pour la langue arabe

- Échantillon: la partie annotée du premier livre dans la chronologie: *L'Or de Paris* (1834):
  - 149 phrases, 8741 tokens et 362 annotations manuelles
- Aucun des systèmes ne propose une granularité fine pour la catégorisation des lieux
  - garder uniquement le premier niveau d'annotation (Loc) en ignorant les spécifications Admin, Nature...
- Tokenisation BIO différente: ajuster manuellement l'alignement des trois sorties
- Modèles entraînés sur des textes modernes
- Stanza: meilleur performances en F-mesure
- Perspectives: élargir le gold et créer un nouveau modèle adapté à granularité fine.

	Mesure	Partielle	Exacte
Farasa	Précision	**0,54**	**0,42**
	Rappel	0,45	0,35
	F1	0,49	0,38
Stanza	Précision	0,51	0,41
	Rappel	0,72	0,58
	F1	**0,59**	**0,48**
Madamira	Précision	0,35	0,25
	Rappel	**0,55**	**0,39**
	F1	0,43	0,31

Création d’un gazetier: solution provisoire?

- Suite de chaîne de traitement (modalités et cartographie): besoin d’annotations précises
- Création d’un gazetier de toponymes français, à partir du gold, sans enrichissement.
- Prise en compte des formes agglutinées (proclitiques et enclitiques).
- Termes les plus fréquents parmi les 2111 EN reconnues:
  - *Paris, France, Marseille, la Seine, Lyon, Jardin des Plantes, Quartier Latin…*



Auteur	Loc-Admin	Loc-Building	Loc-Nature	Loc-Path
Rifa’a al-Tahtawi	390	159	39	0
Malek Bin Nabi	307	40	4	5
Zaki Mubarak	418	89	31	4
Ahmad Zaki	107	30	9	15
Jurji Zaydan	205	17	0	0
A. Fares Al-Shidyaq	223	17	2	0

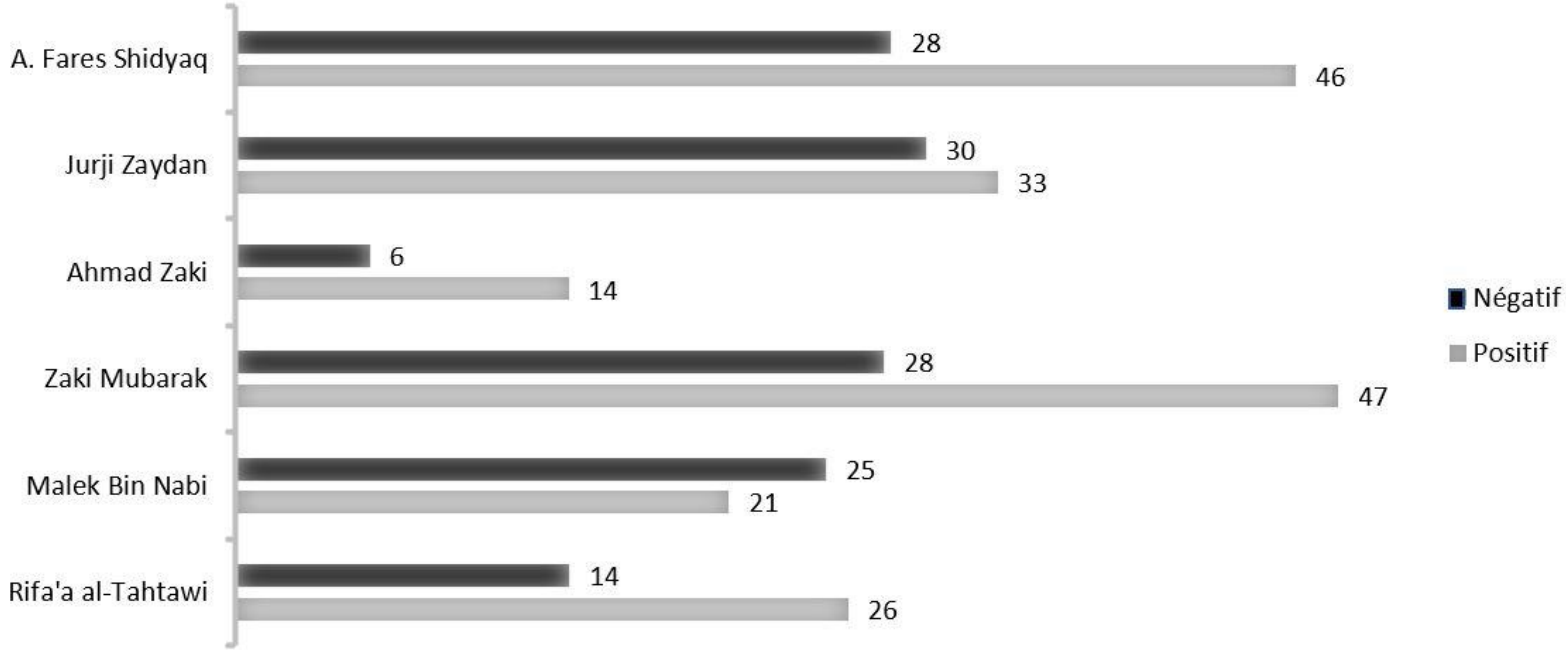
Évaluation de l’approche par dictionnaire

- Difficultés rencontrées:
  - Ambiguïté contextuelle
    - ليل → *Lille* ou nuit ; السين → *La Seine* ou ’s’ ; كان → *Cannes* ou l’auxiliaire être au passé…
  - Sur-composition: présence d’une EN dans deux catégories différentes (Moncla et Gaio 2015).
    - حديقة لوكسمبورغ / *le Jardin du Luxembourg*
- Évaluation sur un nouveau récit: *Est et Ouest* (شرق وغرب), M. Husayn Haykal (1888-1956)
  - 182 phrases (environ 6000 mots)
- 154 EN reconnues dans 81 phrases annotées
  - Précision 100% et Rappel 81%.
  - Les 36 EN non reconnues: lieux non répertoriés ou bien lieux répertoriés sous une transcription différente: نهر الصين au lieu de نهر السين et فنسين au lieu de فانسين…

Analyse des modalités linguistiques autour des EN de lieu

- Objectif: analyse des modalités dans le contexte des EN:
  - polarité positive: accord, appréciation, joie, soutien…(603 marqueurs au total, 13 catégories)
  - polarité négative: accusation, colère, tristesse, désaccord...(777 marqueurs 24 catégories)
- Une phase de prétraitement automatique est nécessaire :
  - Structuration au format TEI, ajout de métadonnées (auteur, date, éditeur…) et segmentation en phrases.
- Seules les phrases contenant à la fois EN et annotations sémantiques ont été gardées:
  - 692 phrases: 993 EN et 1311 modalités.
    - 785 sont « positives » (60%)
    - 526 sont « négatives » (40%).

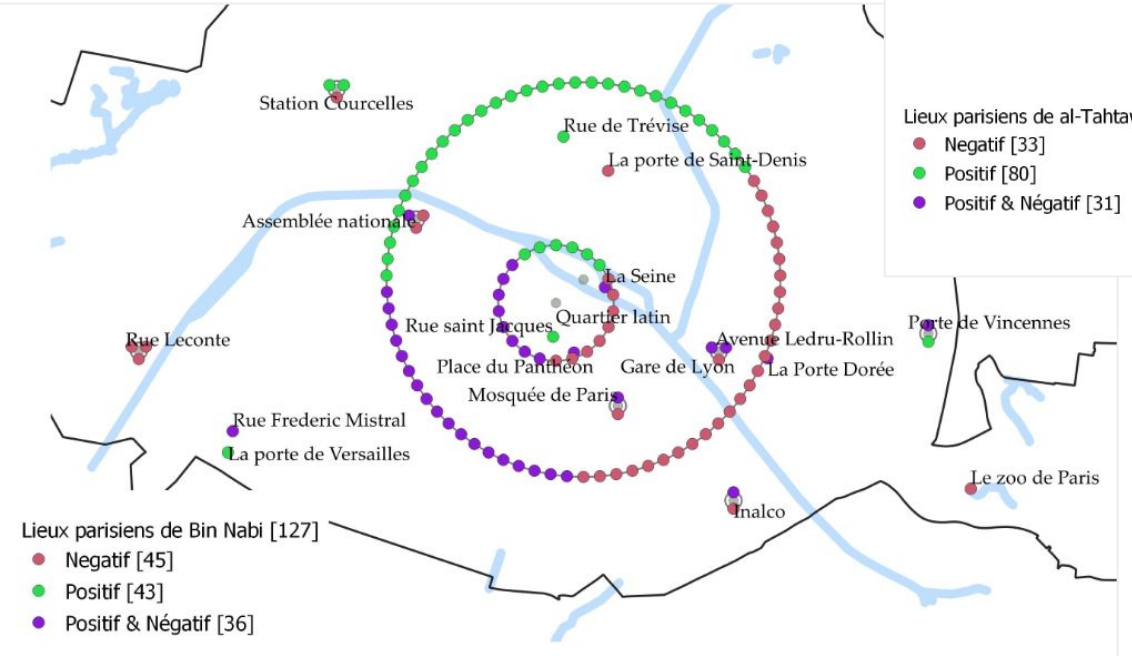
Analyse des modalités linguistiques autour des EN de lieu

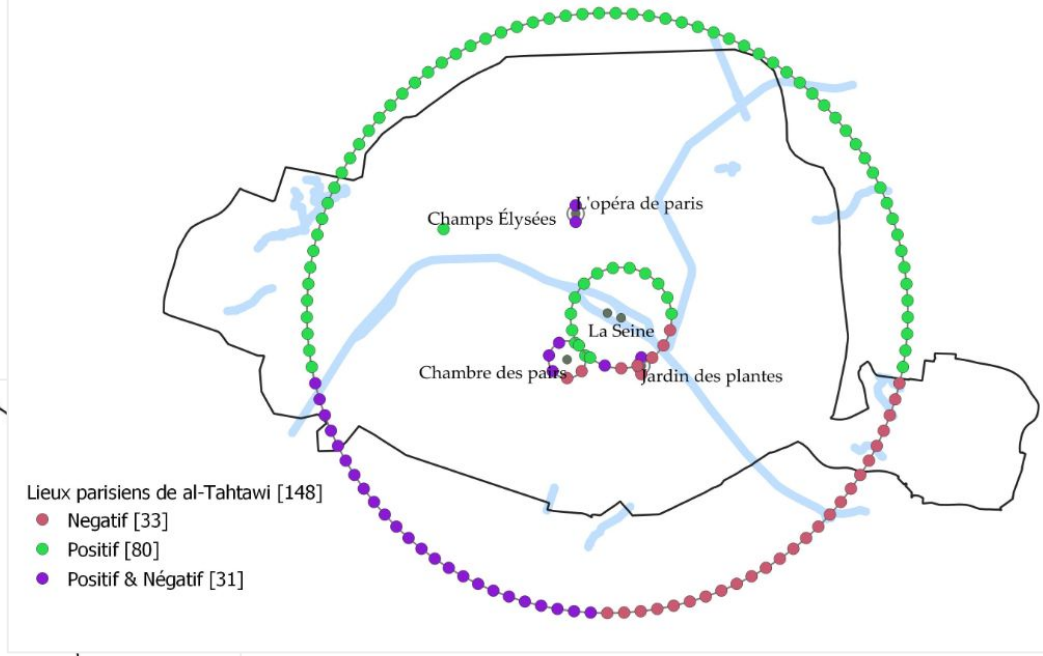


Interprétation des résultats

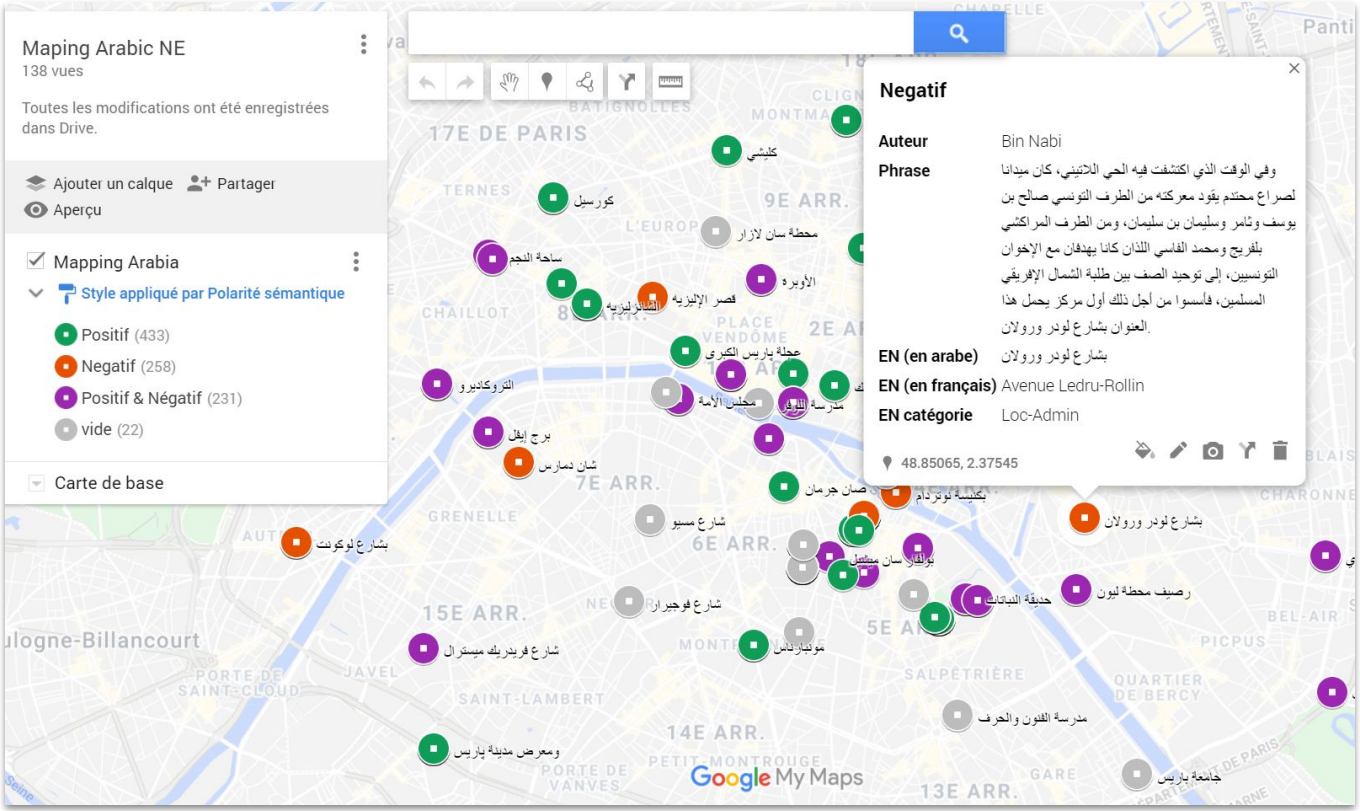
- Comparer les résultats obtenus avec des études critiques sur notre thème de recherche:
  - Khalil Al-Sheikh, *Paris in Modern Arabic literature*, 1998
- Deux visions opposées :
  - Vision francophile (ville de la science, de la connaissance et du développement…)
  - Vision anti-française: panarabisme, panislamisme (ville profane et des péchés, du découragement lié au retard perçu de la civilisation arabe…)
- Selon l'auteur, les 3 auteurs Tahtawi, Zaki et Mubarak (et Shidyaq?) entrent dans la première catégorie
  - Analyse automatique: Ils présentent presque deux fois plus de modalités positives que négatives.
- Pour l'auteur, Bin Nabi rentre dans la deuxième catégorie:
  - Analyse automatique: le seul auteur qui a rapporté plus de modalités négatives que positives
- Zaydan (pas dans le livre): sentiments presque aussi négatifs que positifs.

Cartographie thématique





Cartographie en ligne (dynamique)



Conclusion et perspectives

- Contribution:
  - État de l’art
  - Évaluation d’outils de REN pour la langue arabe
  - Analyse sémantique fine du contexte des toponymes
  - Cartographier les lieux selon leur polarité

- Perspectives:
  - Élargir le corpus et la couverture des EN
  - Annoter les comparaisons et les événements (rencontres, déplacements, déclarations…)
  - Relier chaque modalité à son objet (sa cible)
  - Générer des cartes émotionnelles selon le type d’EN

Merci pour votre écoute

Référence de l'article:

Motasem Alrahabi, Carmen Brando, Muhamed Alkhalil, Joseph Dichy "*Paris dans les récits de voyage d'écrivains arabes : repérage, analyse sémantique et cartographie de toponymes*", Revue Humanités numériques [en ligne], 3 |2021. A paraître.

Lien vers l'application:

https://obvil.huma-num.fr/ariane/humanistica/search