Construction of an encyclopedic dictionary of tourism from scientific corpus

Lichao Zhu, Fabrice Issac, Touria Aït El Mekki, Olivier Hu

Construction of an encyclopedic dictionary of tourism from scientific corpus

Lichao Zhu[1], Fabrice Issac[2], Touria Ait El Mekki[3], Olivier Hû[3]

[1]Université Paris Cité – lichao.zhu@gmail.com

[2]Université Sorbonne Paris Nord – fabrice.issac@gmail.com

[3]Université d'Angers – [touria.aitelmekki, olivier.hu]@univ-angers.fr

Abstract

In this article, we present our method in order to create a nomenclature while respecting the choice of the experts, not only to identify the terms but also to be able to account for their origins and their specificities. Starting by compiling a corpus of scientific articles from various electronic tourism journals, we labelled the corpus with certain compound words and proper nouns were identified in it. We then implemented different strategies to identify candidates using syntactic patterns. A particular attention was paid to the presentation of syntactic and statistical results for experts in the field of tourism (interface and terminology sheet)

Keywords: textometry, tourism, scientific corpus, linguistic modeling, supervised learning

Résumé

Nous nous présentons dans cet article notre méthodologie pour la création de la nomenclature en respectant la volonté des experts, de non seulement identifier les termes mais aussi d'être capable de rendre compte de leurs origines et de leurs spécificités. Nous avons commencé par constituer un corpus d'articles scientifiques issus de différentes revues électroniques du tourisme. Ce corpus a été étiqueté, certains mots composés et noms propres y ont été identifiés. Nous avons ensuite mis en œuvre différentes stratégies pour identifier des candidats en utilisant des patrons syntaxiques. Nous avons porté une attention toute particulière à la présentation des résultats syntactico-statistiques pour les experts du domaine du tourisme (interface et fiche terminologique).

Mots clés : textométrie, tourisme, corpus scientifique, modélisation linguistique, apprentissage supervisé

1. Introduction

La recherche se définit en discipline mais aussi, et peut être de plus en plus, en domaine ou en champ. La lexicométrie, le TAL ou les humanités numériques sont des exemples de ces domaines qui mettent en œuvre plusieurs disciplines. Se crée alors un « langage commun » formé de termes issus des différentes disciplines et de termes forgés pour le domaine ou le champ lui-même. Ces termes sont issus d'un consensus : ils sont le produit d'une littérature scientifique et d'un temps de maturation importants.

Le domaine du tourisme en France est une évidence économique et sociétale. Il est également un champ de recherche transdisciplinaire (économie, géographie, histoire, sociologie…), mais qui ne disposait pas ou peu de structures dédiées à la recherche. Démarré en 2019, le Groupement d'intérêt scientifique (GIS) *Études touristiques* entend participer à combler ce manque et propose de regrouper les chercheurs pour lesquels le tourisme est la thématique

centrale. Dans cette optique la direction du GIS a initié un travail épistémologique avec, entre autres, l'élaboration d'un ouvrage de référence de type dictionnairique et/ou encyclopédique.

Pour répondre à cette problématique nous avons constitué un corpus d'articles scientifiques issus de différentes revues en ligne du tourisme. Ce corpus a ensuite été étiqueté, certains mots composés et noms propres y ont été identifiés. Cette ressource nous permet de mettre en œuvre dans un premier temps différentes stratégies classiques pour identifier des candidats (tf-idf, fréquence, spécificité, information mutuelle…) en utilisant des patrons syntaxiques. Nous suivons en cela une démarche proche de celle proposée avec l'outil *Termostat* (Drouin 2003).

2. Contexte

Dans le cadre de la création de cette encyclopédie du Tourisme, la direction du GIS a tout d'abord structuré sa démarche avec la création d'un comité éditorial en parallèle du comité scientifique déjà existant. Ce comité éditorial est composé de différents experts du GIS, volontaires et co-optés dans un souci de pluralité des disciplines représentées sans pour autant prétendre à l'exhaustivité.

Ce comité éditorial a en charge : 1. l'appel à participation auprès des chercheurs du réseau du GIS afin que les auteurs potentiels proposent des termes-candidats ; 2. la sélection des entrées du dictionnaire encyclopédique parmi les termes-candidats proposés par les membres du GIS ; 3. la définition des règles éditoriales auxquelles les auteurs devront se soumettre : taille des articles, usages des illustrations, gestion des citations… ; 4. le suivi éditorial et les relations avec les auteurs : relances, modifications, validations.

Pour éviter la posture d'expert et le piège d'une analyse critique restreinte aux disciplines des membres du comité éditorial, la direction du GIS a décidé d'initier un travail d'analyse de la production scientifique du domaine. Il a donc été entamé la construction d'un corpus représentatif de la production scientifique en Tourisme dont l'exploitation doit répondre à deux objectifs : extraire automatiquement des unités lexicales qui pourront être proposées au comité éditorial ; enrichir la compréhension et l'appréhension par le comité éditorial d'un candidat-terme via la mise en évidence de son usage au sein du corpus. La volonté des experts est de non seulement identifier des termes mais aussi d'être capable de rendre compte de leurs origines et de leurs spécificités notamment disciplinaires.

Pour ce travail, nous nous intéressons donc à l'extraction de termes spécialisés en tourisme issus d'articles scientifiques. L'idée est de proposer des termes à valeur thématique (*thematic valuable words*) à partir de méthodes qui allient des approches statistiques et modélisation lexicographique et syntaxique. Nous utilisons ensuite les annotations du comité éditorial pour améliorer, dans une approche supervisée, la précision des propositions de termes. Par rapport au schéma de traitement classique, nous ne recourons pas à une référence déjà existante (Nazarenko et al. 2009) mais nous nous appuyons sur notre corpus et les avis de nos experts.

2.1. Lexicographie et dictionnairique

Même si nous considérons que le travail réalisé ici comporte un volet encyclopédique, les outils et méthodologies à mettre en œuvre se rapportent principalement à la lexicographie. Le volet dictionnairique (au sens de Pruvost 2005) ne doit cependant pas être oublié et sera traité dans un second temps. Les différents aspects à prendre en compte pour la réalisation concrète du dictionnaire encyclopédique, notamment sur la question de l'usage ne doivent cependant pas être complètement occultés.

Nous nous concentrons ici sur deux tâches : construire une nomenclature et proposer aux experts des outils permettant d'en établir leur sens. Les apports de l'informatique à la lexicographie et à la dictionnairique sont bien connus. On trouvera par exemple dans Béjoint 2007, Jacquet-Pfau 2005 et Bertin 2009 un certain nombre de points militant pour l'utilisation à tous les niveaux de l'outil informatique (construction, représentation, exploitation). Parmi tous les points abordés, le corpus tient une place centrale. Concernant son utilisation retenons qu'il permet : d'établir une nomenclature représentative ; d'avoir une information fidèle à l'usage ; de travailler en synchronie mais aussi en diachronie ; et d'exhiber des unités lexicographiques complexes.

3. Nomenclature

L'élaboration de la nomenclature est le premier pas vers la construction des dictionnaires. Il s'agit de définir le lexique sinon exhaustif du domaine du moins représentatif. Nous ne nous trouvons pas ici dans le cas d'un dictionnaire général et la tâche s'en trouve simplifiée. Cependant cette notion de représentativité n'en reste pas moins posée et c'est le corpus de par son volume et son origine qui dans un premier temps sera utilisé pour construire la nomenclature. Cette étape pose un grand nombre de questions pour tous les types de dictionnaires. Il s'agit finalement de définir quelles informations vont être présentées. C'est une tâche importante au cours de laquelle le lexicographe doit se poser un grand nombre de questions (De Villers 2011). Il faut non seulement établir une liste de candidats mais aussi : (1) Placer les termes sur un axe spatial (y a-t-il des variations sémantiques selon la région) et/ou temporel (le sens a-t-il changé au cours du temps). (2) Garder/rejeter le candidat selon son niveau de spécialisation (L'homme and Polguère 2008). (3) Décider de critères d'insertion des emprunts.

Ce questionnement présuppose la construction d'une première liste de candidats. Celle-ci peut être construire en (i) faisant appel à l'introspection des experts (ii) en utilisant une nomenclature existante (iii) en utilisant des outils statistiques permettant d'extraire les termes saillants.

D'un point de vue statistique la fréquence et la répartition ne peuvent être les seuls critères. Les experts le sont dans leur domaine, dans le champ du tourisme, mais ne sont pas des lexicographes. A nous de proposer des candidats, d'expliquer pourquoi statistiquement ils ont une importance et de laisser au groupe d'experts le choix sur : Le terme monolexical ou polylexical, ses variantes flexionnelles, pourquoi il a été sélectionné (quelle mesure) et sa répartition au sein du corpus de manière absolue et en diachronie.

4. Élaboration d'un corpus d'articles scientifiques

Le corpus est issu de trois revues scientifiques[1], identifiées par le comité scientifique du GIS comme représentatifs du champ, de par leur large couverture pluridisciplinaire et leur ancienneté. La récupération s'est faite par aspiration des pages du site, ce qui a permis d'ajouter tout un ensemble de méta données (auteur, le cas échéant traducteur, bibliographie…). L'ensemble du corpus est structuré en XML en utilisant une DTD spécifique. Un tel corpus est a priori représentatif de la recherche en tourisme pas du tourisme et de ses acteurs. Le vocabulaire et les usages professionnels en sont par définition exclus.

1 *Mondes du tourisme* : https://journals.openedition.org/tourisme/, *Teoros* : https://journals.openedition.org/teoros/ et *ViaTourism* : https://journals.openedition.org/viatourisme/

L'ensemble de ces informations permet d'envisager un traitement des éléments lexicaux sous différents axes. Il est ainsi possible, et pertinent pour notre étude, d'observer l'évolution du lexique en diachronie. Les textes d'article extraits sont tokénisés et étiquetés à l'aide de l'outil *Corpindex*[2]. Cet étiquetage se fait par application de dictionnaires généraux et une levée d'ambiguïté à partir de règles ce qui rend le corpus exploitable au niveau des parties du discours. Au total nous disposons d'une ressource composée d'articles réparties sur 20 ans comportant plus de 4 millions de mots. Si ce corpus n'est pas la ressource *définitive* par rapport au domaine du tourisme nous estimons qu'il le représente de manière satisfaisante mais c'est encore une fois les experts qui nous diront s'il faut en reconsidérer les contours.

5. Méthodologie

Nous indiquons ici les différentes mesures, ou stratégies, mises en œuvre afin de constituer des listes pertinentes de candidats. On distingue les traitements sur des unités monolexicales (p. ex. fréquence), polylexicales (p. ex. information mutuelle) et transversaux (répartition). L'ensemble des traitements ne présente pas de caractéristiques particulières et est classique dans le cadre d'outils terminologiques. Notre objectif est de construire une fiche descriptive, la plus complète possible, qui sera soumise aux experts.

5.1. Mesure de dispersion

Le corpus est trié chronologiquement. De ce fait, une recherche de l'ensemble des occurrences d'un terme permet à la fois de mesurer sa répartition (cette mesure est calculée à l'aide d'un calcul du coefficient de variation obtenu par le rapport entre l'écart-type et la moyenne) et son évolution dans le temps. Il est à noté que cette mesure considère le corpus comme étant monolithique. Plus cette mesure est faible plus le terme est présent sur l'ensemble du corpus. À titre d'illustration, le mot de, qui est extrêmement bien réparti, a une mesure de 0,03, le terme *chaîne hôtelière* à une mesure égale à 1,8 qui indique une répartition relativement uniforme à l'inverse de touriste postmoderne avec une mesure de 3,47.

Nous nous sommes intéressé aux scores TF-IDF (Salton and Buckley, 1988) des noms dans l'ensemble des documents. Ainsi, nous avons retenu le meilleur score ($maxTFIDF_{t_k}$) et la moyenne des scores ($\overline{TFIDF}_{t_k}$) pour chaque nom monolexical. Ces mesures que nous proposons aux experts aident à faire émerger des noms (unités monolexicales) à l'aune des fréquences en mesurant l'importance d'une occurrence dans un document et son importance dans l'ensemble des documents. Le croisement des deux suppute qu'un mot a de la valeur thématique. Ainsi, parmi les 20 noms qui ont obtenu les meilleurs scores, nous remarquons des mots tels que *surf, nautism, weather, tabasco* et *hospitalité*. Il est à noter que cette mesure est à prendre avec précaution. Comme l'indique Claveau (2012) il existe des mesures, comme Okapi, qui semblent plus efficaces. Notre indicateur est donc une baseline qu'il faudra ajuster.

5.2. Unités polylexicales

La recherche de collocations est motivée par le fait qu'elles sont un bon indicateur du genre des textes (Spina and Tanganelli 2012), la notion de genre étant ici associée aux articles scientifiques liés au champ du tourisme. Par ailleurs, ces collocations permettront d'identifier des unités monolexicales *productive* et par conséquent potentiellement intéressante pour la nomenclature. Nous utilisons, arbitrairement, deux mesures afin de proposer aux experts un

2 https://github.com/ProfessorFabrice/Corpindex

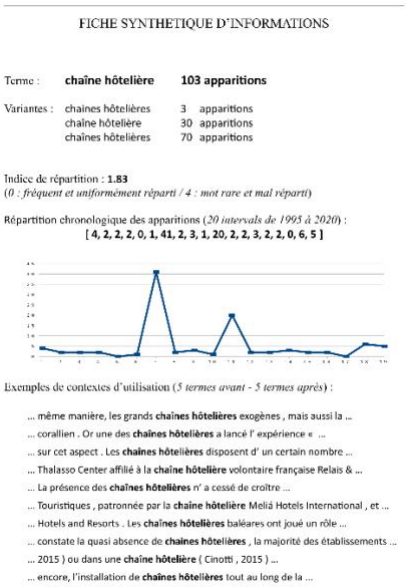
ensemble d'unités polylexicales : le z-test et l'information mutuelle. Nous envisageons, si cela s'avère nécessaire, d'utiliser d'autres mesures comme la mesure de gravité lexicale (Daudaravičius and Marcinkevičienė 2004). L'intérêt de disposer de deux mesures est le fait que les résultats obtenus présentent un certain nombre de différences (sur les 200 premières unités trouvées par les deux méthodes, moins d'un quart sont identiques).

Le z-test propose une mesure capable de rendre compte du fait que les mots qui composent une collocation apparaissent plus fréquemment ensemble que ne le voudrait le hasard. La méthode probabiliste basée sur l'information mutuelle, quant à elle, calcule la quantité d'informations partagée par deux mots. Dans les deux cas nous filtrons a priori les couples à partir de patrons syntaxiques. L'utilisation des résultats permet en outre d'identifier soit des prédicats soit des arguments pour des entrées données :

- *(pratiques, opérateur, aménagement, activité, expériences) touristiques*
- *espace (naturel, public, naturels, publics, rural, urbain)*

6. Présentation des livrables

A l'issue des différents traitements réalisés nous proposons deux types de résultats aux experts. Le premier est une fiche synthétique construite pour certaines unités (voir figure 1). Le deuxième est une interface de type concordancier qui en outre permet de retourner facilement au texte complet (voir figure 2). Ces deux ressources doivent à la fois refléter avec exactitude la réalité du corpus et être accessible pour un public non familiarisé avec ce type de résultats.


FICHE SYNTHETIQUE D'INFORMATIONS

Terme : **chaîne hôtelière** **103 apparitions**

Variantes : chaines hôtelières 3 apparitions
chaîne hôtelière 30 apparitions
chaînes hôtelières 70 apparitions

Indice de répartition : **1.83**
(0 : fréquent et uniformément réparti / 4 : mot rare et mal réparti)

Répartition chronologique des apparitions (*20 intervals de 1995 à 2020*) :
[4, 2, 2, 2, 0, 1, 41, 2, 3, 1, 20, 2, 2, 3, 2, 2, 0, 6, 5]

Exemples de contextes d'utilisation (*5 termes avant - 5 termes après*) :

... même manière, les grands **chaînes hôtelières** exogènes , mais aussi la ...
... corallien . Or une des **chaînes hôtelières** a lancé l' expérience « ...
... sur cet aspect . Les **chaînes hôtelières** disposent d' un certain nombre ...
... Thalasso Center affilié à la **chaîne hôtelière** volontaire française Relais & ...
... La présence des **chaînes hôtelières** n' a cessé de croître ...
... Touristiques , patronnée par la **chaîne hôtelière** Meliá Hotels International , et ...
... Hotels and Resorts . Les **chaînes hôtelières** baléares ont joué un rôle ...
... constate la quasi absence de **chaînes hôtelières** , la majorité des établissements ...
... 2015) ou dans une **chaîne hôtelière** (Cinotti , 2015) ...
... encore, l'installation de **chaînes hôtelières** tout au long de la ...

Figure 1 - Exemple de fiche synthétique d'information

7. Conclusion, discussion, perspectives

Le travail présenté ici est le premier volet d'un projet visant à co-construire un dictionnaire encyclopédique capable de couvrir le champ du tourisme : la nomenclature. C'est une problématique qui nécessite évidemment les compétences d'experts du domaine mais également en lexicographie et, pour ce volet particulier du projet, de TAL. La méthodologie mise en œuvre consiste à développer des outils heuristiques de découverte de candidats au sein d'un corpus constitué d'articles scientifiques. Dans un deuxième temps les candidats sont présentés aux experts (fiche descriptive, outil d'exploration du corpus) et les raisons de l'acceptation ou du refus sont explicités. Par exemple les mots *voyage* et *station* malgré des statistiques les classant parmi les termes les plus pertinents, n'intéressent pas les experts qui les jugent trop imprécis. Ils préfèrent les termes *station verte* ou *station intégré*e malgré un classement bien plus faible. De la même manière les termes *station balnéaire* également bien placés ne seront pas choisis car considérés comme une instance de *station intégrée*.

Une fois les termes sélectionnés la démarche est donc double. D'une part un processus itératif pour améliorer nos outils d'extraction. Comme nous l'avons évoqué nous confronterons les termes ainsi triés et ceux annotés et retenus par les experts afin de proposer une stratégie d'apprentissage supervisé intégrée dans notre processus de découverte de nouveaux candidats.

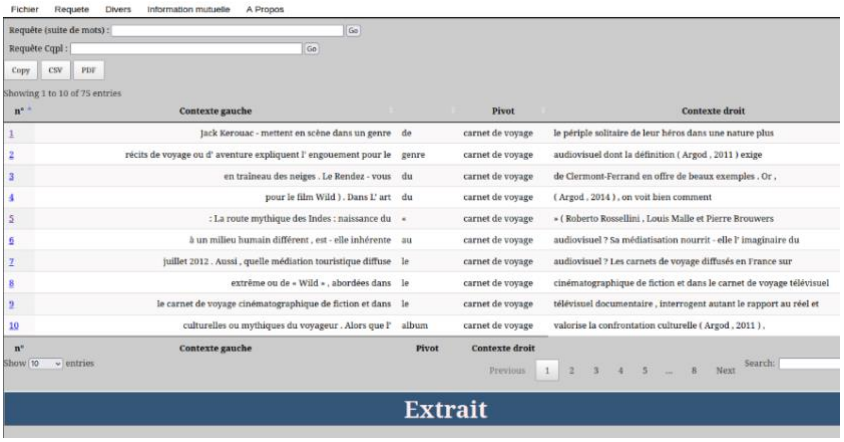


Figure 2 - Capture d'écran de l'outil d'exploration du corpus

D'autre part, un travail de formation et d'information en direction des experts dans le cadre de leur travail épistémologique sera nécessaire. Ils vont devoir formaliser leur démarche, accompagné en cela par les outils que nous mettons à leur disposition afin de les aider à appréhender les usages terminologiques de la communauté des chercheurs en Tourisme.

Références

Henri Béjoint. (2007). Informatique et lexicographie de corpus : les nouveaux dictionnaires. *Revue française de linguistique appliquée*, 12(1), pp. 7-23.

Annie Bertin. (2009). Grammaire et dictionnaire : le parti pris des mots. *Linx. Revue des linguistes de l'université Paris X Nanterre*, 61, pp. 71-85.

Vincent Claveau. (2012). Vectorisation, okapi et calcul de similarité pour le tal : pour oublier enfin le tf-idf (vectorization, okapi and computing similarity for nlp : Say goodbye to tf-idf) [in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pp. 85-98.

Vidas Daudaravičius and Rūta Marcinkevičienė. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2), pp. 321-348.

Marie-Éva De Villers. (2011). *Profession lexicographe*. Les Presses de l'Université de Montréal.

Patrick Drouin. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), pp. 99-115,

Christine Jacquet-Pfau. (2005). Pour un nouveau dictionnaire informatisé. *Ela. Études de linguistique appliquée*, 1, pp. 51-71.

Marie-Claude L'homme and Alain Polguère. (2008). Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*, pp. 191-206.

Adeline Nazarenko, Haifa Zargayouna, Olivier Hamon, and Jonathan Van Puymbrouck. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Revue TAL, 50, varia*, pp. 257–281, *URL https://hal.archives-ouvertes.fr/hal-00516698*.

Jean Pruvost. (2005). Quelques concepts lexicographiques opératoires à promouvoir au seuil du xxie siècle. *Éla. Études de linguistique appliquée*, 137(1), pp. 7-37.

Gerard Salton and Christopher Buckley. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5), pp. 513-523.

Stefania Spina and Elena Tanganelli. (2012). Collocations as an index for distinguishing text genres. *Corpus*, 11, 01.