Mots. Les langages du politique

102 | 2013

Les discours sur l'enseignement supérieur et la recherche

Pays et nation dans un corpus de presse en 2007 et 2009. Étude en linguistique de corpus en vue de l'élaboration de dictionnaires spécialisés

Radia Hannachi et Delphine Seguin-Giuliani

Édition électronique

URL : <http://journals.openedition.org/mots/21405>

DOI : 10.4000/mots.21405

ISSN : 1960-6001

Éditeur

ENS Éditions

Édition imprimée

Date de publication : 2 septembre 2013

ISBN : 978-2-84788-414-2

ISSN : 0243-6450

Référence électronique

Radia Hannachi et Delphine Seguin-Giuliani, « *Pays et nation* dans un corpus de presse en 2007 et 2009. Étude en linguistique de corpus en vue de l'élaboration de dictionnaires spécialisés », *Mots. Les langages du politique* [En ligne], 102 | 2013, mis en ligne le 02 septembre 2015, consulté le 06 mai 2019. URL : <http://journals.openedition.org/mots/21405> ; DOI : 10.4000/mots.21405

Pays et nation dans un corpus de presse en 2007 et 2009. Étude en linguistique de corpus en vue de l'élaboration de dictionnaires spécialisés

Les dictionnaires entretiennent diverses relations avec la notion de corpus. Leur rôle, qui consiste à décrire une langue, semble leur imposer de fait un processus méthodologique qui ne pourrait que difficilement s'écarter des extraits du langage authentique¹ qu'ils doivent expliciter.

Dès le 18^e siècle, en 1755, un dictionnaire fut élaboré par Samuel Johnson à partir de fiches de travail comportant des citations authentiques (Williams, 2006, p. 152). Cet ouvrage a ouvert la voie à de nombreux autres dictionnaires tirés d'une étude de *corpus*². Par la suite, différents dictionnaires feront usage de corpus, que ceux-ci reposent sur des textes littéraires en latin, en langue vernaculaire ou bien sur des extraits de langage non littéraire.

Cependant, il faut attendre la deuxième moitié du 20^e siècle pour qu'un mode de conception radicalement différent du dictionnaire voie le jour, du fait du développement exponentiel des technologies d'information et de communication. Les travaux relatifs au français fondamental et l'élaboration du dictionnaire qui s'est ensuivi (Gougenheim, 1964) entrent dans ce type de projet innovant, de même que le projet de réalisation du *Trésor de la langue française*. Un autre projet, cherchant à élaborer un dictionnaire complet de la langue anglaise, prendra également place en Grande-Bretagne dans les années quatre-vingt.

1. Nous entendons par *langage authentique* un ensemble de faits de langue attestés et qui n'ont pas été créés en vue d'une exploitation linguistique. Notre corpus est ainsi constitué de textes complets tous issus de journaux diffusés ou publiés. Il suit en cela les recommandations données par John M. Sinclair (1991) sur la façon dont un corpus doit être composé, et nous impose de garder en considération le fait que les résultats de nos recherches valent pour le genre médiatique considéré.
2. Le mot *corpus*, ici, n'est pas entendu au sens qu'on lui donne dans le domaine de la linguistique de corpus dans la mesure où il ne s'agit pas de données informatisées. Toutefois, pour l'élaboration des définitions, le recours systématique à des fiches de travail organisées et répertoriant des citations était une innovation dans le cadre du travail de Samuel Johnson.

HCTI (Héritage et constructions dans le texte et l'image, EA 4249), Université de Bretagne-Sud / LiCoRN (groupe de recherche en lexicographie, corpus et ressources numériques)
radia.hannachi@univ-ubs.fr
LiCoRN
giulianidelphine@gmail.com

À cette époque, le projet Cobuild a débuté avec l'objectif de créer un corpus d'anglais contemporain pour la composition d'un dictionnaire et d'une grammaire entièrement nouveaux. La jeune discipline de la linguistique de corpus et le projet Cobuild se sont entre-nourris dans cet environnement lexicographique innovant (Sinclair, 1987). La nouveauté résidait dans le fait de recourir uniquement à un très grand corpus d'anglais contemporain³ pour réaliser un dictionnaire. Patrick Hanks, un des linguistes de corpus ayant participé à la réalisation de ce projet, a préconisé l'emploi de prototypes lexicographiques pour réaliser des dictionnaires basés sur corpus (Hanks, 1994, 2000, 2006). C'est de ce type de prototypes qu'il va être question ici. Nous proposons en effet dans cette étude de voir en quoi les prototypes lexicographiques peuvent permettre une appréhension plus fine de la polysémie et, par exemple, aider à réaliser des dictionnaires pour apprenants de haut niveau.

Dans un premier temps, nous préciserons la nature du corpus que nous utilisons, ainsi que les méthodes d'analyse qui lui sont appliquées. Nous expliquerons ensuite ce que sont les prototypes lexicographiques et en quoi ils peuvent être utiles dans la réalisation de dictionnaires spécialisés. Enfin, nous proposerons une étude de cas avec la réalisation de ces prototypes pour les mots *nation* et *pays*.

Méthodologie

Le corpus IntUne

IntUne est un projet intégré du PCRD6 (6^e Programme cadre de recherche et développement) de l'Union européenne⁴ qui a impliqué trente-deux partenaires dans quinze pays différents. L'objectif était d'étudier les évolutions concernant la façon dont la citoyenneté européenne se construit. Il s'agit d'un projet complexe et pluridisciplinaire réunissant quatre groupes principaux de recherche : un groupe média, un groupe élites, un groupe expert et un groupe masse.

Le groupe média se composait de quatre équipes nationales. Il avait pour finalité l'élaboration de deux corpus médiatiques multilingues et leur analyse selon une méthodologie issue des principes de la linguistique de corpus (Williams, Bayley, 2012). Le corpus français a été assemblé à l'université de Bretagne-Sud sous la responsabilité de Geoffrey Williams. L'accès à ce corpus est restreint, mais en tant que participante au projet IntUne et à la réalisation du corpus, nous pouvons présenter une étude qui lui est rattachée (Giuliani, 2009 ; Williams *et al.*, 2012).

3. Ce corpus se composait de plus de 7 millions de mots, traités de façon uniquement automatisée (Sinclair, 1987).

4. <http://www.intune.it/> (consulté le 3 avril 2013).

Ce corpus, par sa taille (environ 40 millions de mots pour sa partie française) et son format informatique, suit les recommandations formulées par Sinclair en la matière : « *Corpus has to have the material in electronic form [...] The only guidance I would give is that a corpus should be as large as possible.* »⁵ (Sinclair, 1991, p. 14-18)

Méthode d'analyse

Une fois le corpus construit et opérationnel, il s'agit de l'interroger. Nous utilisons *Xaira*⁶, en raison de ses capacités à fonctionner sans difficultés majeures sur des corpus de taille importante, ainsi que de son aptitude à gérer des structures complexes de texte. Il permet également un fonctionnement de type *corpus-driven* et l'identification de collocations, ce qui correspond à notre approche du corpus qui se veut inductive (Tognini-Bonelli, 2001).

Selon l'approche *corpus-driven*, le linguiste ne cherche pas à imposer des modèles déjà existants, ou à faire correspondre à son intuition les données qui sont extraites au cours de l'analyse du corpus. Il est plutôt question de partir de l'observation des données, et non d'une hypothèse *a priori*. Dans ce type de recherche, si le mot étudié est *pays*, il sera ainsi procédé au repérage de toutes ses occurrences dans le corpus, mais non à celui de structures dont l'hypothèse aurait été faite préalablement, pour voir si elles sont effectivement présentes ou non.

Firth résumait avec cette formule la meilleure façon d'appréhender le sens d'un mot : « *You shall know a word from the company it keeps* »⁷ (Firth, 1957, p. 11). Cette phrase implique que c'est par le contexte du mot⁸ que l'on parviendra à mieux le comprendre.

Xaira est un concordancier qui extrait des lignes de contextes au sein desquels les mots clefs choisis apparaissent, et qui calcule des collocations. L'étude de corpus postule que si des concepts sont reliés dans la réalité au sein d'un type de discours, ces relations pourront être atteintes à travers l'analyse d'un corpus : « *In so far as the central concepts relate to each other, similar relational patterns emerge in their surface constructs, words* »⁹ (Williams, 1998, p. 151).

5. « Le matériau d'un corpus doit se présenter sous forme numérique [...] Le seul conseil que je donnerais est qu'un corpus doit être aussi étendu que possible. »
6. Le concordancier *Xaira* (Xml Aware Indexing and Retrieval Architecture) peut se télécharger librement à l'adresse suivante : <http://www.oucs.ox.ac.uk/rts/xaira/> (consulté le 3 avril 2013).
7. « Vous connaîtrez un mot en fonction de sa compagnie habituelle. »
8. Dans notre étude, le contexte se compose le plus souvent de trois mots à gauche et trois mots à droite du mot clef dans le calcul des collocations. Lorsque nous menons une étude de type qualitatif, celui-ci est généralement d'une dizaine de mots à gauche du mot clé étudié et d'une dizaine de mots à droite. Nous l'avons étendu quand la compréhension de l'occurrence était gênée par un manque d'information.
9. « De la même façon que des concepts centraux sont reliés les uns aux autres, des structures de relations similaires émergent dans leurs constructions de surface : les mots ».

L'étude et le calcul des collocations¹⁰ peuvent alors être utilisés pour montrer ce qui est typique ou ce qui ne l'est pas au sein d'un corpus fondé sur un type de discours (Williams, 2002, p. 45). L'analyse menée dans cet article repose donc sur ce type de calcul. La recherche de collocations permet de faire émerger ce qui est typique pour le mot clef étudié, dans un type de discours (ici, le discours médiatique). À partir du moment où l'on dispose de ces connaissances sur les sens que le discours médiatique donne à des concepts particulièrement complexes et utiles dans différents domaines, il devient possible d'implémenter les ressources lexicographiques afin qu'elles représentent davantage la réalité discursive à laquelle des étudiants spécialisés sont confrontés.

Dans cet article, nous allons réaliser différents prototypes lexicographiques pour les mots *pays* et *nation*. Ils se fonderont sur la matière lexicographique déjà existante et sur les collocations de ces mots dans le corpus IntUne.

Dictionnaires et prototypes lexicographiques

La pratique didactique courante tend à envisager les dictionnaires papier comme des ouvrages d'appoint. Le développement de dictionnaires électroniques (TLFi, *Petit Robert* électronique) ne semble pas avoir permis un accroissement sensible de l'utilisation de ces ressources en classe de didactique (Verlinde *et al.*, 2005). Il apparaît, entre autres facteurs, que cette faible utilisation naît d'une inadéquation entre les besoins des utilisateurs et les informations présentées dans les dictionnaires (Williams 2012 ; Verlinde *et al.*). Le fait de présenter les données du dictionnaire non plus sous la forme d'un article conçu comme un tout (avec citations littéraires, connaissances encyclopédiques, explication du sens du mot, etc.) mais en donnant davantage d'autonomie aux différentes informations qui constituent habituellement un article, voire en ajoutant certaines informations comme les fréquences de certains usages (Verlinde *et al.*) ou en utilisant des phrases explicatives formées sur le modèle de la Full Sentence Explication (Williams, 2012), pourrait réduire ces difficultés. L'utilisation de prototypes lexicographique peut permettre d'aller dans le sens de telles modifications et de donner des informations plus pertinentes aux étudiants de langue étrangère lors de cours de FLE (français langue étrangère) ou de FOS (français sur objectifs spécifiques), dans la mesure où ils peuvent, par exemple, rendre possible la comparaison fine de traductions données pour différents mots (Williams, 2012, p. 196).

La théorie du prototype en sémantique a connu un grand succès et son potentiel dans le domaine de la linguistique est bien documenté (Kleiber,

10. Ce terme pose de nombreux problèmes de définition. John Rupert Firth parlait de la « compagnie habituelle » des mots pour décrire ce phénomène. Dans notre étude, les collocations sont des regroupements au niveau syntagmatique statistiquement significatifs de mots avec un mot clef ou nœud étudié.

1990 ; Taylor, 1995). Cette théorie prend ses racines dans la pensée de Wittgenstein. Celui-ci a élaboré le modèle de la ressemblance de famille en montrant que la catégorie *game*¹¹ se compose non pas de membres qui partagent tous un ensemble de conditions nécessaires et suffisantes, mais de membres qui partagent au moins un attribut avec un autre membre de la catégorie. Cet attribut commun donne une ressemblance de famille à l'ensemble et permet de dénommer les différentes formes de jeux sous un nom unique, quand bien même certaines formes n'ont aucun point en commun (Taylor, 1995, p. 39).

Eleanor Rosch a par la suite développé le modèle des prototypes. Le prototype est alors la meilleure instance d'une catégorie et pour appartenir à la catégorie, un membre doit partager au moins un attribut avec ce prototype (Kleiber, 1990, p. 48). Rosch s'est plus tard inspirée de la ressemblance de famille, le prototype devenant non plus l'objet représentant le mieux la catégorie, mais une « entité abstraite construite sur la base des propriétés typiques de la catégorie » (*ibid.*, p. 63).

Cette théorie du prototype a connu de nombreux développements et applications dans le domaine de la linguistique. Parmi ces applications, nous basons notre travail sur la méthodologie développée par Patrick Hanks¹² (1994, 2000, 2006). Elle se fonde sur ce que Georges Kleiber appelle la « version étendue » (Kleiber, 1990, p. 150) de la théorie du prototype. Cette « version étendue » s'appuie sur la ressemblance de famille, selon laquelle il n'y a pas dans la catégorie un centre et une périphérie. Hanks pose les prototypes comme se construisant à partir d'un ensemble de composantes, de sens potentiels (il parle de « meaning potential ») qui sont présents dans le prototype et s'activent dans les usages du mot étudié. Ces sens potentiels rencontrent les attributs présents dans le modèle de la ressemblance de famille : aucun n'est nécessaire ou central dans la catégorie. Il formule ainsi sa pensée : « *Words only have meanings when they are put into context. In isolation, they have meaning potentials, which are composed of any number of rather fuzzy semantic components, some or all of which are activated when the word is used.* » (Hanks, 2006, p. 2)¹³

Partant de ce postulat, il pose que le sens d'un mot est l'ensemble des « *word's meaning potential that are activated when it is used in its more normal contexts* »¹⁴ (*ibid.*, p. 3), ces contextes les plus normaux pouvant se mesurer à l'aide de grands corpus et de calculs statistiques de collocations. Le prototype

11. L'exemple de Wittgenstein fonctionne très bien avec les mots *Spiel* et *game*, un peu moins bien avec le français *jeu*.

12. Patrick Hanks est un lexicographe britannique qui a été directeur de la première édition du Cobuild, et éditeur en chef des éditions Collins.

13. « Les mots n'ont de signification que mis en contexte. Isolément, ils ont des sens potentiels constitués d'un certain nombre de composantes sémantiques plutôt floues ; l'ensemble ou certains de ces sens potentiels sont activés quand le mot est utilisé. »

14. « les sens potentiels d'un mot qui sont activés quand il est utilisé dans ses contextes les plus courants ».

n'a alors plus beaucoup de points communs avec celui de la « version standard » de la théorie. Il cherche avant tout à permettre l'élaboration de définitions opérantes se basant sur des faits de langue attestés pour des mots complexes.

Hanks parle ainsi de *prototype lexicographique* et non plus de *prototype sémantique*. Nous pourrions résumer la nuance entre ces deux notions de la façon suivante : d'une part, le prototype lexicographique se base essentiellement sur la théorie de la ressemblance de famille élaborée par Wittgenstein, il rassemble différents sens qui sont associés au travers d'un continuum ; d'autre part, le prototype sémantique est généralement fondé sur la théorie roschienne selon laquelle les prototypes se composent d'un sens central et d'une périphérie. Il serait possible, dans notre étude, d'isoler une signification prototypique, centrale pour les mots *nation* et *pays* dans un sous-corpus ou un contexte donné. Mais aucun sens isolé, donné comme central, ne pourrait nous permettre de cerner avec un maximum d'efficacité la polysémie des mots ici étudiés. C'est pourquoi nous fonderons notre étude sur l'élaboration de prototypes lexicographiques tels qu'ils sont développés par Hanks et Williams (Williams, 2012 ; Williams *et al.*, 2012).

Nous allons maintenant tenter d'élaborer ces prototypes lexicographiques pour les mots *pays* et *nation*¹⁵ en nous appuyant sur la matière lexicographique déjà existante et sur le corpus. Nous essaierons ainsi de voir dans quels contextes ces mots peuvent être employés le plus fréquemment et quels sens ils peuvent recouvrir.

Nation et pays, quelles différences ?

Prototypes basés sur la matière lexicographique

Le mot *pays*

Cette troisième partie s'appuie sur les recherches menées au cours d'une thèse financée par le projet IntUne (Giuliani, 2009 ; Williams *et al.*, 2012).

Dans un premier temps, nous nous intéresserons ici au mot *pays* et aux sens qu'en donnent plusieurs dictionnaires français.

1. Territoire d'une nation ; *nation*, État. 1b. Ensemble des habitants d'une *nation*.
 2. Région, contrée envisagée du point de vue physique, climatique, économique, etc.
 3. Lieu, région d'origine.
 4. Village, localité.
- (Larousse, 2011)¹⁶

15. Notre étude ne porte pas sur le mot *État* qui est beaucoup moins présent dans le corpus IntUne, ce qui rend son étude moins fiable.

16. Nous soulignons.

1. Territoire habité par une collectivité et constituant une réalité géographique dénommée ; *nation*.
2. Les gens, les habitants du pays.
- 3a. Le pays de quelqu'un, patrie à laquelle on appartient par sa naissance.
- 3b. Le pays de quelque chose, terre d'élection, milieu particulièrement favorable à, riche en.
4. Région géographique plus ou moins nettement limitée, considérée surtout dans son aspect physique.
5. Petite ville, village.
(*Petit Robert*, 2008)
- A. Division territoriale habitée par une collectivité, et constituant une entité géographique et humaine. Dans une acception large, *nation*, État. Dans une acception restreinte, partie plus ou moins étendue d'une *nation* : province, région, canton.
- B. *Nation*, région à laquelle on appartient par la naissance. Dans une acception large, *nation* d'origine, patrie. Dans une acception restreinte, province natale, village natal.
- C. Par métonymie : ensemble des habitants, des citoyens d'une *nation*.
- D. Au figuré : domaine fabuleux, irréel, généralement peuplé d'êtres et de choses imaginaires, rêvés.
(*Trésor de la langue française informatisé*)

Le dictionnaire *Larousse* nous fournit cinq sens pour le mot *pays* : il peut être le territoire d'une nation ou la totalité des habitants de cette nation, il peut s'agir d'une région considérée du point de vue de son climat, de son économie, etc. Il peut s'agir d'un lieu de naissance, d'une unité administrative : un village ou une petite ville.

Le deuxième dictionnaire, *Le Petit Robert*, livre des sens proches mais formulés différemment. On repère encore cinq sens principaux. *Pays* désigne un territoire habité par une communauté, et par métonymie, c'est aussi cette communauté. Cela se rapproche de la seconde partie du sens 1 donné par le dictionnaire Larousse ; cependant, il n'est pas question ici du mot *nation*. Il peut s'agir d'un lieu de naissance, d'un lieu particulièrement favorable à quelque chose, d'une région géographique, d'une petite ville ou d'un village.

Le *TLFi* reprend ces différents sens en les organisant différemment. Nous ne trouvons plus cinq sens numérotés, mais quatre. Le sens métonymique de « communauté » n'est plus présenté en première ou deuxième position comme dans le *Larousse* ou le *Petit Robert* mais en troisième position. Il est néanmoins précisé dès le premier paragraphe que ce mot représente une « entité géographique et humaine »¹⁷. Différentes possibilités d'interprétation sont données au lecteur dans les deux premiers paragraphes avec la mention « dans une acception large [...] [ou] restreinte ».

17. Nous soulignons.

Enfin, il enrichit la description de ce mot en précisant l'usage fréquent du mot *pays* dans le sens métaphorique de « domaine fabuleux ».

Dans ces trois définitions, on retrouve le sens de « territoire pourvu de frontières », d'« entité géographique », de « lieu où se trouve une communauté », d'« unité administrative », de « village » et de « lieu de naissance ».

À l'aide de ces informations, un premier prototype peut être élaboré pour le mot « pays ».

Un pays est :

- une étendue de terre physique,
- une étendue de terre avec un gouvernement et des frontières officielles, un lieu administratif,
- une étendue où se trouvent des habitants,
- une nation,
- un lieu de naissance, d'origine,
- une étendue qui peut se considérer du point de vue de son économie,
- une petite ville, un village¹⁸.

Ce prototype mis au point, nous allons maintenant nous intéresser au mot *nation*.

Matière lexicographique sur le mot *nation*

Les ressources lexicographiques utilisées plus haut donnent les définitions suivantes :

1. Grande communauté humaine, le plus souvent installée sur un même territoire et qui possède une unité historique, linguistique, culturelle, économique plus ou moins forte.
2. Communauté politique distincte des individus qui la composent, et titulaire de la souveraineté.
(*Larousse*, 2011)
1. Groupe d'hommes auxquels on suppose une origine commune.
2. Groupe humain, généralement assez vaste, qui se caractérise par la conscience de son unité (historique, culturelle) et la volonté de vivre en commun.
3. Groupe humain constituant une communauté politique, établie sur un territoire défini ou un ensemble de territoires définis, et personnifiée par une autorité souveraine.
4. Ensemble des individus qui composent ce groupe.
(*Petit Robert*, 2008)
- A. Groupe d'hommes dont les membres sont unis par une origine réelle ou supposée commune et qui sont organisés primitivement sur un territoire.
- B1. Groupe humain, généralement assez vaste, dont les membres sont liés par des affinités tenant à un ensemble d'éléments communs ethniques, sociaux (langue, religion, etc.) et subjectifs (traditions historiques, culturelles, etc.) dont la cohérence repose sur une aspiration à former ou à maintenir une communauté.

18. Pour une explication plus étendue de ce prototype et de celui de nation, voir Williams *et al.*, 2012.

2. Groupe humain stable, établi sur un territoire défini constituant une unité économique caractérisée par une auto-conscience ethnique (marquée par l'idée de la communauté d'origine et de destinée historique), une langue et une culture communes, formant une communauté politique personnifiée par une autorité souveraine et correspondant à un stade évolué du mode et des rapports de production.
 3. Par métonymie : territoire occupé par ce groupe humain.
 - C1. Ensemble des personnes qui composent ce groupe (cf. B2)
 - D. Désigne une collectivité par opposition à un individu, une classe, un groupe. Collectivité nationale.
- (*Trésor de la langue française informatisé*)

De nombreux sens potentiels communs émergent de ces quatre définitions. Un prototype lexicographique pour le mot nation pourrait être le suivant.

Une nation est :

- un pays considéré comme un groupe humain ;
- un groupe humain d'une origine commune supposée ;
- un groupe humain généralement assez vaste ;
- un groupe humain qui se caractérise par la conscience de son unité (historique, culturelle) ;
- un groupe humain qui se caractérise par une langue commune ;
- un groupe humain qui se caractérise par une religion commune ;
- un groupe humain qui a la volonté de vivre en commun ;
- un groupe humain constituant une communauté politique ;
- un groupe humain établi sur un territoire défini ;
- un groupe humain établi sur un ensemble de territoires définis ;
- un groupe humain personnifié par une autorité souveraine ;
- une communauté politique distincte des individus qui la composent.

Nous allons maintenant confronter ces données au corpus IntUne afin de voir si dans un corpus de discours médiatique, elles sont confirmées, infirmées ou modifiées par les découvertes réalisées.

Nation et Pays dans le corpus IntUne

Pays

Le mot *pays* a une très forte présence dans le corpus IntUne. On dénombre 15 448 occurrences en 2007 et à 13 319 reprises en 2009. Il est mentionné dans chaque article de notre corpus et représente environ 0,07 % des mots du corpus. L'analyse du mot *pays* dans le corpus se veut exploratoire, inductive. Elle va ainsi reposer sur l'étude des collocations de ce mot afin de mettre en relief les différentes thématiques qui lui sont associées.

Pour calculer les collocations, nous utilisons le logiciel Xaira et prenons une fenêtre de trois mots à gauche et trois mots à droite du mot clef étudié. Nous retirons ensuite tous les mots apparaissant à moins de trois reprises et

reclassons les collocats restants en fonction de l'importance de leur z-score¹⁹ (voir tableau 1 en annexe).

Une grande majorité des collocations reste identique entre les deux années de collecte du corpus, ce qui met en évidence une certaine stabilité sémantique du mot *pays*. La dimension économique du concept est indubitablement présente à travers les collocats *émergents*, *industrialisés* ou *développés* par exemple.

D'autre part, la notion d'étendue géographique est marquée par des mots comme *dans*, où et dans une certaine mesure, par les collocats *nordiques*, *occidentaux*. Cependant, ce dernier type de collocat, bien que se construisant sur une appellation géographique, peut correspondre à des critères largement culturels ou économiques.

Les collocations qui rendent compte d'une étendue géographique montrent un mot fréquent qui est *du*. Cette préposition apparaît dans la colligation²⁰ *du pays*, utilisée pour situer un événement dans une perspective géographique. Les journaux parlent essentiellement de faits de guerres dans ces occurrences. Ainsi, *du pays* est utilisé massivement pour l'Afghanistan, l'Irak, le Soudan. Quand *du pays* est utilisé pour désigner une étendue géographique de façon plus précise, il apparaît dans la vaste majorité des cas que le pays en question subit le terrorisme ou la guerre. Le fait que les journaux préfèrent généralement traiter de problèmes peut expliquer ce phénomène. Néanmoins, *Les Échos*, qui accordent moins d'importance à ce type de faits, sont aussi concernés par ce comportement, puisque lorsqu'il est question de situer quelque chose à l'aide de la colligation, les informations sont le plus souvent mauvaises. Par ailleurs, en 2009, ce type d'usage apparaît toujours avec la même importance.

Au Darfour, région de l'ouest du	pays	, en proie à la guerre civile	fi170307 ²¹
èvement de 2 français et de 3 afghans dans le sud ouest du	pays	.	su060407
fanatiques hindous avaient détruit une mosquée, au nord du	pays	. Des émeutes inter-religieuses s'en ét	ec160409

19. Le z-score est un type de calcul statistique qui permet d'obtenir les collocations les plus significatives d'un point de vue sémantique en écartant les mots présents de façon très courante dans le corpus.

20. Le terme de *colligation* est proposé par J. R. Firth dans les années cinquante. En faisant la synthèse des travaux récents sur la question, Dominique Legallois (Legallois 2008) en propose la définition suivante : « La colligation se définit par la relation régulière observée dans les énoncés, entre une unité lexicale particulière et une catégorie grammaticale. [...] La colligation est donc un phénomène de comportement grammatical préférentiel, de la même façon que la colligation est un phénomène de comportement lexical préférentiel. »

21. Ce type d'abréviation signifie ici que l'occurrence est extraite du journal *Le Figaro* (soit les deux premières lettres du fichier « fi ») du 17 mars 2007. Dans les autres abréviations, nous avons choisi « ec » pour *Les Échos*, « li » pour *Libération*, « su » pour *Sud Ouest* et « ou » pour *Ouest-France*.

Une entité politique est également présente, notamment avec le mot *euro-péen* qui concerne les pays membres de l'Union européenne. Une façon intéressante de dépeindre un espace politique dans la presse française est alors l'usage de la formule *notre pays*. On dénombre 607 occurrences de cette construction dans le corpus, et le journal qui l'utilise le plus est *Le Figaro* avec 208 occurrences, puis *Libération* avec 123 occurrences, tandis que *Les Échos* n'apparaissent qu'en troisième position. Cela peut s'expliquer par le fait que le quotidien *Les Échos*, spécialisé dans l'information économique et financière, accorde moins de place que les deux autres aux commentaires strictement politiques. La formule *notre pays* est beaucoup moins présente dans le corpus 2009, ce qui s'explique par le fait que celui-ci n'a pas été collecté lors d'une échéance électorale jugée importante pour la France. Lorsque la formule est utilisée, les collocats signalent l'existence de différents problèmes :

	Notre pays	est en train de « décrocher sous anesthésie »	fi240307
	Notre pays	est dans une situation difficile et effectivement menacé	fi370307
problèmes se pose à l'ensemble de	notre pays	. Faut-il pour autant les ignorer ? Je m'y refuse. Sans	li050209

À partir du moment où des problèmes sont exposés, des solutions doivent être suggérées, cela peut arriver mais avec une importance moindre en termes de fréquence :

instaurer pour pouvoir lancer les réformes utiles à	notre pays	.	fi130307
---	------------	---	----------

Toujours dans cette thématique politique, *pays* peut aussi être utilisé pour désigner une entité qui appartient à un groupe caractérisé par une réglementation politique commune. Le groupe le plus présent est l'Union européenne. En 2007, tous les pays qui en font partie connaissent une bonne croissance économique à l'exception de la France. En 2009, le contexte des élections législatives européennes apparaît dans ce cadre et la crise économique fait ressortir de lourds problèmes d'entente entre les pays composant l'Union :

croissance dans les principaux	pays	de l'Union en 2007 à l'exception de la France	ec190207
de présenter la liste dans l'ensemble des	pays	de l'Union européenne. Sous ce même drapeau,	li2110209
FMI décoche [...] une flèche en direction des	pays	de l'Union européenne dont le manque de coordination	ec230409

La dimension culturelle est présente à travers différents collocats. Les pays peuvent par exemple être regroupés selon différents points communs, par exemple ethnique et culturel (*anglo-saxons, arabes*), religieux et économique (les pays du *Golfe* sont majoritairement à la fois de population musulmane et exportateurs de pétrole). Un autre type de collocats rend également compte de cette thématique culturelle, les pays étant cette fois compris en tant qu'entités infranationales. Le mot *basque* est par exemple une partie de la construction *pays Basque*, qui désigne une nation non reconnue et sans existence officielle. Le mot *Loire* est surtout utilisé dans l'expression *pays de Loire*²², qui désigne le terroir connu pour ses châteaux de style Renaissance ou ses vins.

L'unité administrative est présente à travers différents collocats : *pays de Lorient* réfère à une étendue dépendant d'une ville, et qui regroupe plusieurs communes de façon institutionnelle et officiellement reconnue. Le collocat *Galles* permet également de référer à une unité administrative reconnue, mais fonctionnant cette fois au Royaume-Uni. Enfin, le lieu d'origine est présent à travers l'adjectif possessif *notre* et les collocats *natal* ou *origine*.

Ces collocations font émerger différentes thématiques, différents sens pour le mot *pays*. Ceux-ci, comme on l'a vu, peuvent en outre se relier les uns aux autres. En conséquence, après ce panorama des collocations, le prototype lexicographique préalablement construit peut évoluer de la façon suivante :

Un pays est :

- une étendue économique,
- une étendue de terre physique,
- un espace politique doté d'un gouvernement,
- une unité administrative plus ou moins grande,
- un lieu d'origine,
- une entité complexe constituant une réalité culturelle.

Nation

On dénombre 807 occurrences de ce mot dans le corpus 2007, et il apparaît à 339 reprises en 2009. Il représente environ 0,004 % du total des mots du corpus dans le premier cas, contre 0,002 % dans le second. Les deux journaux qui l'emploient le plus sont les deux quotidiens à ligne éditoriale politique, *Libération* et *Le Figaro*. Voir ses collocats dans le tableau 2 en annexe.

Ces collocats font intervenir différentes thématiques. Une première thématique forte, présente dans les deux corpus IntUne, renvoie à la dimension symbolique du mot *nation*. L'entité semble alors très propice à la métaphore, elle est presque humaine. Elle est aussi associée à des idées positives et fortes : la nation intègre, on l'exalte. Cette dimension symbolique se retrouve en 2009 :

22. Il est étonnant de constater qu'ici, ce n'est pas la région administrative « Pays de la Loire » qui est majoritairement citée dans les journaux du corpus IntUne, mais un terroir, le « Pays de Loire ».

Irak : l'agonie d'une	nation	, confirmant notamment	li190207
L'arrière petit-fils du père de la	nation	Indienne, Tushar Gandhi	lio50309

Le contexte politique est présent de diverses façons. Des candidats à l'élection présidentielle peuvent souhaiter employer ce concept en refusant d'en laisser le monopole à leurs adversaires. La nation est aussi associée à la République, toutes deux correspondant à des principes essentiels et fondamentaux de la France. En 2007, cette entité politique est très souvent représentée par la personne de son président. En 2009, le contexte politique apparaît via des collocations comme *indépendante* ou *république*. La nation est alors une entité qui bénéficie d'une reconnaissance par d'autres ensembles politiques et qui en tant que telle peut agir (honorer quelqu'un en le citant à l'ordre de la Nation, par exemple). Ce contexte d'une reconnaissance obtenue ou demandée auprès d'autres ensembles est également présent en 2007, mais ne se manifeste pas de façon aussi importante dans les collocations.

président de la République, c'est l'homme de la	nation	, ce n'est pas l'homme d'un clan, d'un parti	li170407
membre des comptes économiques de la	nation	dresse un bilan financier terrifiant de la décentr	ou090207
donner par étapes à la Corse un statut de	nation	indépendante en Europe. Officiellement formé	fio20209

La nation est aussi un concept très fortement associé à l'idée d'unité. L'unité rejoint dans les deux corpus IntUne la thématique politique, puisque dans la vaste majorité des cas, quand cette dimension est évoquée, c'est par des hommes politiques. Les différents articles définis et indéfinis (*la, une*) renforcent cette idée puisqu'ils sont actualisés au singulier, de même que des mots comme *unité*, *unie* ou *entière* qui montrent que cette entité est fédératrice et porteuse d'une forte unité.

réaliser le rêve d'une	nation	unie, chacun doit pouvoir participer au destin	li120307
------------------------	--------	--	----------

Cette idée d'une unité de la nation rejoint le fait qu'il s'agit d'une communauté humaine dotée d'origines culturelles et historiques supposées communes, une *ethnie*, une *patrie*. Dans ces cas, les deux substantifs sont associés au mot *nation* au sein d'énumérations : ces différents concepts sont ainsi présentés comme proches ou équivalents. En 2007, le contexte culturel montre avant tout la nation française et note, de ce point de vue, son rayonnement international, notamment à travers le musée du Louvre. Le contexte historique est

présent à la même période pour expliquer l'importance du concept de nation dans la campagne présidentielle ou remémorer d'anciens conflits.

posée, des personnes décédées à une ethnie, une	nation	, une race ou une religion déterminée.	su230407
qui vient rappeler que le Louvre est au service de la	nation	tout entière », la lettre poursuit : « ce projet	fio50207

Le corpus 2009 fait apparaître des thématiques qui n'étaient pas présentes dans les collocations de 2007. Il s'agit essentiellement de contextes sportifs, économiques et territoriaux. Le contexte sportif apparaît déjà dans le corpus 2007, mais sans que cela soit mis en relief par les collocations. Il donne une image plutôt positive des nations concernées. On parle alors d'une *grande nation* dans un sport quelconque. Par ailleurs, le syntagme *Tournoi des Six Nations* est présent dans de nombreuses occurrences.

avait décroché 5 médailles à Bordeaux en 2006 (2 ^e)	nation	derrière la Grande-Bretagne), quels sont les	ou290307
La France, grande	nation	du football, avec ses 2,5 millions de licenciés	lio60309

L'apparition du contexte économique en 2009 se manifeste via différents collocats. Le collocat *économique* apparaît uniquement dans le cadre de prévisions qui sont faites par un organisme, la « Commission économique de la nation ». Le collocat *sociale* apparaît également dans ce contexte économique. Enfin, le collocat *Live* se rapporte au nom d'une société musicale, *Live nation*, particulièrement active à l'époque de la collecte de ce corpus.

ives et les dangers. Pour la cohésion sociale de la	nation	, il est peut-être encore temps de changer	ou140209
--	--------	---	----------

Le contexte territorial apparaît aussi dans les collocations en 2009, montrant que lorsqu'on parle d'une nation, il est question d'une relation entre celle-ci et un espace physique. Les occurrences du corpus relatent essentiellement, dans ce domaine, des faits d'agression. Ce sont le plus souvent les mots du président iranien qui sont repris et qui pointent les problématiques territoriales entre pays occidentaux et pays arabes.

militaire pour priver de terres une	nation	entière sous le prétexte de la souffrance juive »	ec210409
Israël, qui a privé « de terres une	nation	entière sous le prétexte de la souffrance juive	su210409

Ces collocations font émerger différents sens pour le mot *nation*. En conséquence, le prototype lexicographique préalablement construit peut évoluer de la façon suivante pour décrire ce qu'est *nation* dans notre corpus :

Une nation est :

- un groupe humain personnifié ;
- un groupe humain qui se caractérise par la conscience de son unité (historique, culturelle, économique) ;
- un groupe humain qui se caractérise par une pratique sportive commune ;
- un groupe humain qui a la volonté de vivre en commun ;
- un groupe humain constituant une communauté politique distincte des individus qui la composent ;
- un groupe humain établi sur un territoire défini.

L'élaboration des prototypes lexicographiques de *pays* et *nation* met en lumière un certain nombre de caractéristiques lors de leur activation dans les discours médiatiques français.

D'une part, si l'on met en comparaison le prototype lexicographique établi à l'aide de l'analyse du corpus avec celui qui est fondé sur la matière lexicographique existante, on s'aperçoit de certaines différences. Le mot *pays* n'est ainsi que très rarement – voire jamais – mentionné comme étant une petite ville ou un petit village dans le discours médiatique courant. Par ailleurs, le sens d'« espace économique », qui n'est pratiquement pas mentionné dans les dictionnaires, semble important pour *pays*. Le mot *nation* présente le même comportement en ce que le sens de « groupe humain qui se caractérise par une pratique sportive commune », non mentionné dans les dictionnaires, est bien présent dans le corpus IntUne (Tournoi des Six Nations, rang occupé parmi les nations participant aux Jeux olympiques, etc.).

D'autre part, les différences entre ces deux notions sont mises en relief lors de l'analyse de leur présence dans le corpus. Ainsi, les prototypes lexicographiques construits à l'aide de l'analyse du corpus IntUne montrent que ces deux mots sont utilisés dans des contextes bien précis : plutôt économique pour *pays*, plutôt culturel ou sportif pour *nation*.

L'élaboration de prototypes lexicographiques à l'aide de l'analyse inductive et systématique d'un vaste corpus semble donc pertinente pour implémenter des dictionnaires et les rendre plus représentatifs du sens de mots complexes dans des domaines précis. Cette étude n'est qu'une ébauche concernant des mots représentant des concepts complexes, et il serait intéressant de

confronter dans la suite des recherches cette analyse à une analyse sur d'autres corpus, constitués dans d'autres pays, et avec d'autres types de textes.

Références

- FIRTH John Rupert, 1957, *Papers in Linguistics 1934-1951*, Oxford, Oxford University Press.
- GIULIANI Delphine, 2009, « Représentation et construction d'une identité européenne à travers la presse française. Une étude en linguistique de corpus au sein d'un projet de l'Union européenne », thèse de doctorat dirigée par G. Williams, soutenue à l'Université de Bretagne-Sud, en ligne [<http://tel.archives-ouvertes.fr/tel-00634154/>], consultée le 3 avril 2013.
- GOUGENHEIM Georges et al., 1964, *L'élaboration du français fondamental. Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.
- HANKS Patrick, 1994, « Linguistic norms and pragmatic exploitations, or why lexicographers need prototype theory, and vice versa », *Papers in Computational Lexicography, Complex '94*, Budapest, Hungarian Academy of Sciences.
- 2000, « Do word meanings exist ? » *Senseval : Evaluating Word Sense Disambiguation Programmes*, A. Kilgarriff, M. Palmer éd., *Computers and the humanities*, n° 34 (1-2), p. 205-215.
- 2006, « Metaphoricity is a Gradable », *Corpora in Cognitive Linguistics*, vol. I, *Metaphor and Metonymy*, Berlin - New York, Mouton de Gruyter.
- KLEIBER Georges, 1990, *La sémantique du prototype*, Paris, PUF.
- LEGALLOIS Dominique, 2008, « Peut-on mesurer la naturalité des énoncés ? », *Mémoires de la Société de linguistique de Paris*, n° 16, 2008, Paris.
- SINCLAIR John, 1987, *Looking Up*, Oxford, Collins ELT.
- 1991, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.
- TAYLOR John R., 1995, *Linguistic categorization. Prototypes in linguistic theory*, Oxford, Oxford University Press.
- TOGNINI-BONELLI Elena, 2001, *Corpus Linguistics at Work*, Amsterdam, John Benjamins.
- VERLINDE Serge, SELVA Thierry, BINON Jean, 2005, « Dictionnaires électroniques et environnement d'apprentissage du lexique », *Revue française de linguistique appliquée*, 2005 (2), vol. X, p. 19-30.
- WILLIAMS Gelfrey, 1998, « Collocational networks. Interlocking patterns of lexis in a corpus of plant biology research articles », *International Journal of Corpus Linguistics*, n° 3 (1), p. 151-171.
- 2002, « In search of representativity in specialised corpora », in *International Journal of Corpus Linguistics*, n° 7 (1), 43-63.
- 2006, « La linguistique de corpus, une affaire prépositionnelle », *revue Texto*, en ligne [<http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf>], consultée le 3 avril 2013.
- 2012, « Art for dictionaries sake. Comparing cultural outlooks through dictionaries

and corpora », *English Dictionaries as Cultural Mines*, R. Facchinetti éd., Cambridge Scholars Publishing, Newcastle upon Tyne, p. 171-200.

WILLIAMS Gelfrey, PIAZZA Roberta, GIULIANI Delphine, 2012, « Nation and supnation. A tale of three Europes », P. Bayley, W. Geoffrey éd., *European Identity. What the Media Say*, Oxford, Oxford University Press.

WILLIAMS Gelfrey, BAYLEY Paul, 2012, *European Identity. What the Media Say*, Oxford, Oxford University Press.

Annexe

2007			2009		
Collocations	Fréquence	z-score	Collocations	Fréquence	z-score
Émergents	368	244,9	Émergents	340	246,1
Basque	249	148,4	Basque	248	155,6
Galles	127	148,1	Loire	196	155,5
Loire	219	146,6	Développés	154	153,4
Industrialisés	81	130,9	Galles	153	149,5
Dans	3 393	113,4	Industrialisés	58	112,6
Européens	335	99,4	Dans	2 764	100,2
Développés	99	97,1	Européens	318	97,8
Pauvres	155	83,4	Riches	157	90,9
Notre	638	82,1	Pauvres	152	90,7
Nordiques	38	71,1	Baltes	43	83,4
Baltes	38	70	Occidentaux	86	67,8
Scandinaves	44	69,2	Les	3 421	58,3
Du	2 915	68	Du	2 329	55,8
Riches	111	61,4	Arabes	72	54,9
Arabes	90	60,9	Nordiques	23	54,5
Lorient	254	57,9	Autres	452	51
Anglo-saxons	49	56,2	Anast	11	49,9
Occidentaux	73	55,6	Notre	376	49,3
Civilisé	25	54,7	Natal	27	48,9
Exportateurs	38	53	Africains	54	48,7
Où	523	51,9	Anglo-saxons	35	43,6
Les	3 427	50,6	Scandinaves	19	42,3
Africains	65	49,1	Exportateurs	25	41,9
Autres	475	48,4	Auge	12	41,3
Natal	29	47	Où	407	39,2
OCDE	47	46,7	Membres	143	39,2
Voisins	86	45,8	Golfe	47	38
Origine	168	44,7	Dogon	11	36
Bigouden	15	43,7	Europe	227	35,3

Tableau 1. Les trente premières collocations du mot pays dans le corpus IntUne

2007			2009		
Collocations	Fréquence	z-score	Collocations	Fréquence	z-score
Pupilles	3	50,7	Live	22	161,6
La	645	49,2	Ticketmaster	5	131,3
Exaltation	4	45,5	La	232	25,8
Symboles	9	42,4	Fondamentaux	4	24,1
Intégrante	4	35,5	Entière	5	22,6
Ethnie	3	30,6	Terres	4	17,7
Conceptions	4	29,2	Indépendante	3	16,1
Agonie	3	25,5	Concerts	3	12,6
Fondent	3	23,9	Reconnaissance	3	12,3
République	21	21,9	Père	6	11
Musulmane	4	21	Une	58	9,6
Discours	17	20,9	Seule	6	9,5
Unie	3	19,8	Football	4	9,4
Patrie	4	19,2	Sauver	3	9,3
Islam	6	18,1	Américaine	5	8,8
Nationalisme	3	17,7	Toute	8	8,6
Menacée	4	16,8	De	172	8,4
Symbole	6	15,3	Petite	5	8,3
Entière	5	15,2	Économique	7	8,1
Une	139	15	Menace	3	7,8
Monopole	4	14,9	Classement	3	7,7
Race	3	13,3	Supérieur	3	6,8
Fermée	3	13,1	Grande	6	6,5
Identité	9	12,8	Ordre	4	6,4
Vives	3	12,8	Veulent	3	6,4
Grande	18	12,7	Sous	7	5,8
Turque	3	12,4	Reste	6	5,6
Unité	6	12,3	Sociale	3	5,5
Idée	12	12	République	3	4,9
Ouverte	6	10,9	Notre	6	4,5

Tableau 2. Collocations du mot nation dans le corpus IntUne