ILCEA

Revue de l'Institut des langues et cultures
d'Europe, Amérique, Afrique, Asie et Australie

39 | 2020

**Les humanités numériques dans une perspective
internationale : opportunités, défis, outils et méthodes**

Le Trésor du lexique patrimonial galicien et portugais : un portail pour l'étude du lexique dialectal

*O Tesouro do léxico patrimonial galego e português: um portal para o
estudo do léxico dialetal*

*Treasury of the Galician and Portuguese Heritage Lexicon: a Website for
the Study of Dialectal Lexicon*

Marta Negro Romero et Xandra Santos Palmou

Édition électronique

URL : <http://journals.openedition.org/ilcea/9933>

DOI : 10.4000/ilcea.9933

ISSN : 2101-0609

Éditeur

UGA Éditions/Université Grenoble Alpes

Édition imprimée

ISBN : 978-2-37747-174-4

ISSN : 1639-6073

Référence électronique

Marta Negro Romero et Xandra Santos Palmou, « Le Trésor du lexique patrimonial galicien et portugais : un portail pour l'étude du lexique dialectal », *ILCEA* [En ligne], 39 | 2020, mis en ligne le 03 mars 2020, consulté le 09 octobre 2020. URL : <http://journals.openedition.org/ilcea/9933> ; DOI : <https://doi.org/10.4000/ilcea.9933>

Ce document a été généré automatiquement le 9 octobre 2020.

© ILCEA

Le Trésor du lexique patrimonial galicien et portugais : un portail pour l'étude du lexique dialectal¹

O Tesouro do léxico patrimonial galego e portugués: um portal para o estudo do léxico dialetal

Treasury of the Galician and Portuguese Heritage Lexicon: a Website for the Study of Dialectal Lexicon

Marta Negro Romero et Xandra Santos Palmou

1. Introduction

- ¹ Le Trésor du lexique patrimonial galicien et portugais (TLPGP)² est un projet international auquel participent des universités et des équipes de recherche de Galice, du Portugal et du Brésil, sous la coordination de la Professeure Rosario Álvarez Blanco, membre de l'Institut de la langue galicienne de l'Université de Saint-Jacques-de-Compostelle³. Il en est ressorti un corpus numérique dont l'accès est libre et qui met à la disposition de la communauté scientifique et du reste de ses usagers le lexique dialectal, géographiquement référencé, issu de ces trois territoires. L'objectif de cette initiative est de conserver et de diffuser un grand nombre de documents lexicaux, pour la plupart inconnus ou d'accès restreint, liés à la culture traditionnelle, notamment aux traditions et aux savoirs qui ont disparu ou sont sur le point de disparaître à cause de divers changements sociaux et culturels. Il ne s'agit donc pas de réaliser de nouveaux travaux sur le terrain mais plutôt d'en faire connaître les études existantes, en vue d'en traiter et d'en divulguer le contenu.
- ² Au cours de ces dernières décennies, suite au développement des corpus linguistiques et à la prolifération des bases de données terminologiques, l'accès à ces dernières est devenu d'autant plus facile qu'elles intègrent un outil de recherche⁴. Les progrès de l'informatique l'ayant rendu possible, ce sont aujourd'hui des outils qui sont non

seulement utilisés par les chercheurs mais également par le grand public. C'est dans ce sens que le *TLPGP* s'inscrit, cherchant à collecter, numériser, traiter et mettre gratuitement à la disposition de quiconque une grande quantité de données lexicales. Un tel trésor se distingue toutefois du reste de notre corpus et de nos bases lexicales par deux aspects : d'une part, il intègre, sur une seule plate-forme de consultation, les informations lexicographiques contenues dans les textes compilés avec une représentation cartographique des données générées de façon automatique ; d'autre part, le *TLPGP* se situe dans une perspective internationale, les champs linguistiques abordés étant le galicien, le portugais européen et le brésilien, trois langues qui partagent la même origine : la branche linguistique galaïco-portugaise.

- 3 Ces deux caractéristiques lui confèrent sans nul doute un caractère novateur dans son domaine, encore qu'elles aient supposé deux des plus grands défis à relever au niveau de la conception et du développement. Ainsi, l'incorporation d'images et de cartes générées à chaque recherche implique non seulement une innovation technologique pour y parvenir mais également la mise en place d'une interface à même de faciliter leur visualisation simultanée, ce à quoi s'ajoute la différence géographique des territoires étudiés ; en effet, si le *continuum* linguistique observé entre la Galice et le Portugal va de pair avec leur proximité géographique, il n'en est pas de même dans le cas du Brésil, un pays dont l'étendue est largement supérieure.
- 4 Quoi qu'il en soit, comme nous l'avons déjà souligné, la sélection de variétés linguistiques issues de trois territoires différents se justifie par le fait qu'elles ont la même origine historique : le sous-groupe linguistique galaïco-portugais. C'est pourquoi la plupart des linguistes des deux langues partagent le même point de vue en ce qui concerne leurs traditions. Cependant, quoique le galaïco-portugais ait été généré sur le territoire de l'actuelle Galice et dans le nord du Portugal, au nord du Douro, l'histoire politique a séparé ces deux espaces suite à l'indépendance du Portugal, de sorte que son unité linguistique s'est rompue, donnant lieu à deux branches. Alors que le portugais s'est répandu sur les autres continents et a connu une grande évolution (Dubert García, 1995)⁵, le galicien a subi un processus de « castillanisation » sous le règne de la Castille. Ainsi donc, l'union linguistique et historique entre les trois zones géographiques justifie l'enjeu d'élaborer un projet lexical de cette ampleur, en dépit de la complexité de son élaboration. Il permet en effet d'obtenir des renseignements révélateurs sur l'étude des évolutions diachroniques, des divergences et des similitudes entre ces communautés idiomatiques.

2. Typologie des sources

- 5 Les sources lexicales qui composent ce corpus numérique se caractérisent par leur hétérogénéité typologique (Álvarez, Álvarez Pérez, Saramago & Sousa, 2009 ; Álvarez Pérez & Sousa, 2013). Nous pouvons en établir trois sortes :
- 6 **a) Monographies ethnolinguistiques.** Une partie très importante des documents sont des travaux académiques de diverses natures, aussi bien des thèses de doctorat et des mémoires de master que des productions académiques réalisées dans le cadre de cours de doctorat et dans les dernières années de licence, examinées par des professeurs d'université et soutenues devant un jury. Ces travaux sont donc assez hétérogènes en termes de longueur, de profondeur de recherche, de qualité, de mode de présentation de l'information et d'analyse. Certaines monographies ont été rédigées, alors que

d'autres comprennent un glossaire, lesquelles peuvent être organisées alphabétiquement ou combiner les processus d'élaboration d'un dictionnaire idéologique avec ceux d'un dictionnaire alphabétique.

- 7 **b) Bases de données d'atlas linguistiques et d'enquêtes dialectales.** À l'instar du cas précédent, ces bases de données contiennent non seulement le matériel qui sera cartographié mais aussi d'autres types d'informations (dessins, photographies, proverbes...) qui ne peuvent pas, dans de nombreux cas, être présentées dans la publication imprimée mais qui pourront désormais être intégrées au *TLPGP*. À cet égard, il convient de mentionner, en raison de sa pertinence pour le territoire étudié, l'*Inquérito Linguístico por correspondência* (*Enquête linguistique par correspondance*), entamé par Manuel de Paiva Boléo en 1942 et conservé à la Faculté des lettres de l'Université de Coimbra.
- 8 **c) Autre matériel avec lexique dialectal.** Pour ce qui est du galicien et du portugais européen, il existe plusieurs ouvrages qui étudient la culture locale d'un point de vue anthropologique ou ethnolinguistique, à l'instar des travaux pionniers de la dialectologie galicienne, de la première moitié du ^{xx}e siècle, menés par plusieurs philologues allemands tels que Ebeling, Krüger et Schneider, ainsi que des travaux des chercheurs du Séminaire d'Études Galiciennes (Xaquín Lorenzo, Fermín Bouza Brey, Antón Fraguas, etc.). Il existe également de nombreux ouvrages qui rassemblent des lexiques dialectaux publiés par des auteurs moins connus dans des publications locales ou régionales ou sur des sites web qui ne sont généralement pas pris en considération dans les bibliographies scientifiques.
- 9 Actuellement, le corpus du *TLPGP* se compose de 209 travaux, dont 42 appartiennent au Brésil, 97 à la Galice et 68 au Portugal, auxquels il faut en ajouter 2 autres, à la fois liés à la Galice et au Portugal. Il convient toutefois de tenir compte du fait que la base de données continue à être mise à jour avec l'incorporation de nouveaux ouvrages, de sorte que ces chiffres s'accroîtront à l'avenir⁶.

3. Conception de la base de données

- 10 Le principal défi dans la conception de la base de données a été de collecter et de tirer profit de la grande variété d'informations issues de différents types de sources sans pour autant renoncer à leur harmonisation et à leur mise en commun. Dans ce sens, en s'étant adaptée à ces diverses sources, elle s'avère être un outil suffisamment souple pour surmonter une telle difficulté et répondre aux besoins de ses usagers. Ayant agglutiné et relié les données entre elles, elle permet à ces derniers de les comparer et de mener toute sorte d'études.
- 11 Pour ce faire, la base de données a été configurée à partir de dix-neuf champs (tableau 1), sur lesquels nous ne nous attarderons pas outre mesure, le nom de chacun d'entre eux nous paraissant suffisamment clair. Les quatorze premiers, marqués en gris dans le tableau, visent à enregistrer tous les renseignements fournis par l'œuvre originale, tandis que les six suivants relèvent de la responsabilité exclusive de l'équipe du *TLPGP*. De ces six projets, deux sont hachurés en rose, car ce sont ceux sur lesquels nous travaillons dans la phase de développement actuelle du projet, ce que nous expliquerons plus en détail ci-après.

Saisie des données (chaque ouvrage)														Traitement des données					
														Chaque ouvrage				Dans la base de données	
1	2	3	4	5	7	8	10	11	11 b	12	13	15	20	9	16	17	6	18	19
variante	phonétique	classe ou catégorie grammaticale	définition	exemples et dictons	page	section	référence bibliographique	code géographique	tableau des lieux	images	terme de référence	références aux textes	introduction d'un terme	classification sémantique	lemme de chaque langue	classe ou catégorie grammaticale	commentaires du manipulateur	synonyme de référence	lemme commun

Tableau 1 - La structure de la base de données.

- 12 Quoique les champs gris soient conçus pour collecter les informations contenues dans les textes du corpus, il n'est pas obligatoire de tous les remplir. Pour introduire une œuvre dans la base, il faut seulement compléter les champs *variante* (1) – enregistrement de la variante telle qu'elle apparaît dans l'ouvrage – et *code géographique* (11) – lequel relie ladite variante avec le ou les endroits où l'œuvre en question est documentée (municipalités pour la Galice et le Portugal ; méso-régions pour le Brésil) –, ce qui est indispensable pour la production automatique ultérieure de la carte. Les douze autres champs sont facultatifs et ne sont remplis que s'il existe des informations à ce sujet, toujours dans le respect de la littéralité du texte.
- 13 Par ailleurs, comme l'explique González Seoane (2015 : 54), les textes non lexicographiques ne peuvent pas être directement transférés dans la base de données, étant donné qu'ils nécessitent un degré d'intervention plus élevé. En l'occurrence, les formulaires correspondant au champ 1 (variante) sont directement extraits des textes (les légendes de la carte, dans le cas des atlas) et illustrés dans le champ 4 (définition) à travers un ou plusieurs extraits du même texte, lesquels permettent de comprendre le sens du mot et son contexte d'usage (dans le cas des atlas, c'est le titre de la carte qui apparaît généralement dans le champ 4).

4. Traitement des données compilées

- 14 Comme nous venons de l'expliquer, le traitement du matériel commence par l'introduction d'un certain nombre de données dans la base de données, suivant les instructions du manuel. Une fois cette phase terminée, la lemmatisation, la catégorisation grammaticale et l'affectation d'un classificateur sémantique sont effectuées, tel qu'on peut le voir dans l'exemple suivant.

Image 1 - Extrait de l'ouvrage Pérez Froiz (1972).

1 Variante	2 Trans- cription	4 Défini- tion	7 Page	8 Section	9 Classi- ficateur séman- tique	10 Référer- ce	11 Code géogra- phique	16 Lemme	17 Caté- gorie gramma- ticale
picado de troita	picado de tr[o]ta	Pecoso	87	El cuerpo humano	7.1	PérezF 1972	36059	picado da troita	loc a
pulgar	pul[ɲ]ar	Pulgar	87	El cuerpo humano	7.1	PérezF 1972	36059	polgar	sm
queixo	[kej]xo	Mentón	87	El cuerpo humano	7.1	PérezF 1972	36059	queixo	sm
roxo	r[o]xo	Pelirrojo	87	El cuerpo humano	7.1	PérezF 1972	36059	roxo	a, s

Tableau 2 - Introduction des données de l'ouvrage Pérez Froiz (1972) dans la base de données.

- 15 Dans le processus de lemmatisation⁷, qui suit des critères préétablis, un lemme standardisé (champ 16) est attribué à la variante (champ 1), coïncidant avec le lemme attribué par l'auteur de l'œuvre. Le lexique rassemblé dans les textes étant d'une grande variété formelle, l'attribution du lemme permet l'unification des variantes. C'est ainsi que sont regroupées sous un même lemme les variantes phoniques (sous le lemme *bailar* « danser », nous trouvons *baillar*, *balhar*, *beilar* et *beillar*), les flexionnelles (le lemme *cantar* « chanter » inclut l'infinitif *cantar* et d'autres formes du paradigme comme *cantai* ; *cadeira* « hanche » regroupe le singulier *cadeira* et le pluriel *cadeiras* ; et *neno* « enfant » contient le féminin *nena* et le masculin *neno*) et les orthographiques (*edra* et *hedra* se trouvent sous le lemme *hedra* « lierre »). Cependant, les formes au sens identique mais dont la racine est différente (*testa* / *fronte* « front ») sont situées sous des lemmes différents, ainsi que les variantes à la même racine présentant des

dissemblances morphologiques, que ce soit parce que l'une vient de l'autre suite à un processus morphologique synchronique (*mexerico* / *mexeriqueiro* « personne qui se plaint »), ou qu'elles diffèrent morphologiquement les unes des autres, selon le moment de l'histoire de la langue (*apalleirar* / *empalleirar* « empilage de paille ou d'herbe séchée »).

- 16 Le lemme choisi coïncidera, autant que possible, avec la forme standard, portugaise ou galicienne, selon le cas. À cet égard, il est à noter que les index des lemmes galicien et portugais sont indépendants, c'est-à-dire qu'ils apparaissent séparément dans les recherches (*miñoca* et *minhoca* « ver de terre » ou *temón* et *temão* « gouvernail »), encore qu'un travail de connexion de ces deux catégories de lemmes soit en cours d'élaboration en ce moment. Dans le cas de la Galice, le *Vocabulario Ortográfico da Lingua Galega* (Vocabulaire orthographique de la langue galicienne) et le *Dicionario da Real Academia Galega* (Dictionnaire de l'Académie royale de Galice) sont pris comme références ; pour ce qui est du portugais européen et brésilien, ce sont les dictionnaires *Porto Editora*, *Aurélio*, *Priberam*, *Aulete* et *Houaiss*. Lorsque le mot n'est enregistré dans aucun de ces ouvrages de référence au niveau normatif, d'autres critères sont mis en œuvre :
- 17 Étymologique. Le mot *devanado* est trouvé dans l'une des œuvres répertoriées ; comme l'étymon est *DEPANARE*, nous avons décidé d'établir comme lemme *debanado*, en cohérence avec le résultat du -P- latin intervocalique.
 - Cohérence avec la famille de mots. L'une des variantes apparaissant dans l'un des documents enregistrés est *miniñeiro* ; comme *meniño* « enfant » est le terme primitif dont il dérive et qu'il s'agit de la forme standard, nous avons établi comme lemme *meniñeiro* « personne qui aime les enfants ».
 - Proximité des langues voisines, principalement le portugais dans le cas du galicien et vice-versa. Nous avons trouvé, dans un livre en galicien, la forme *trobexar* « tonner ». Ni ce verbe ni aucune de ses variantes phoniques ne sont normatifs en galicien, tandis qu'en portugais, il existe la forme standard *trovejar*. C'est pourquoi nous avons choisi de lemmatiser en galicien avec *trobexar* (l'orthographe avec s'expliquant par l'existence des formes galiciennes *trebón* et *torbón*).
- 18 Par ailleurs, si l'origine étymologique de la variante est inconnue et que nous ignorons l'existence d'autres formes de la même famille de mots ou équivalents dans les langues voisines, nous choisissons de conserver la forme telle qu'elle a été enregistrée. Les variantes problématiques sont ponctuées d'un point d'interrogation dans le champ 6 (commentaires du manipulateur), ce qui nous permettra de les localiser plus en avant, car il est très probable que le matériel que nous avons compilé dans le *TLPGP* puisse nous aider à trouver la famille de mots « perdue ».
- 19 L'attribution des lemmes s'accompagne de l'établissement d'une catégorie grammaticale normalisée, selon une série d'étiquettes préétablies. Dans les travaux qui composent le *TLPGP*, les auteurs n'attribuent pas toujours une catégorie à la forme lexicale collectée ; néanmoins, lorsqu'ils le font, les critères suivis sont très divers en termes de notes utilisées et de quantité d'informations. À cet égard, il est à souligner que la catégorie que nous établissons est liée au lemme (champ 16) et non à la variante (champ 1), qu'elle tient compte du fonctionnement des mots dans la variété standard et qu'elle ne coïncide pas nécessairement avec la catégorie définie par l'auteur. Par exemple, dans l'ouvrage *GonzálezGM 2002*, l'auteur assigne la catégorie *n.f.* au mot *labor* « travail », alors qu'en galicien standard, il est masculin, ce pour quoi nous plaçons *n.m.* dans le champ 17.

- 20 Face aux nombreux corpus linguistiques désormais proposés en ligne et autres trésors de la langue, une nouveauté présentée par le *TLPGP* est, d'une part, l'organisation onomasiologique de l'ensemble des données incluses. Ainsi, dans un premier temps, un schéma conceptuel a été élaboré, lequel structure tous les éléments lexicaux introduits, si bien que, lors de l'incorporation de chaque document, celui-ci est étiqueté au moyen d'un classificateur sémantique (champ 9) prenant en compte un maximum de trois niveaux. Il ne fait donc aucun doute que l'organisation onomasiologique a été un autre défi important à relever, mais elle s'avère être d'une grande aide au moment d'effectuer des recherches, grâce à un filtre qui affine notre sélection de données. D'autre part, la classification sémantique évite d'éventuels problèmes de désambiguïsation. L'accès à une liste alphabétique est également possible, de sorte que les deux possibilités sont mises en œuvre.
- 21 Après avoir atteint un volume important de variantes et de lemmes⁸, nous avons conçu, pour les membres de l'équipe de travail, une application à usage interne qui facilite le traitement des données. Si la variante est enregistrée dans la base de données du *TLPGP*, cette dernière propose elle-même un lemme, un classificateur sémantique et un classificateur grammatical, après que nous nous sommes situés dans la colonne correspondante. Si nous sommes d'accord avec la proposition, il suffit de cliquer dessus pour qu'elle soit enregistrée dans la base de données. Dans l'exemple suivant du travail *PereiraG 1983*, l'application propose, pour la variante portugaise *abobreira*, le lemme *aboboreira* (image 2), le classificateur sémantique 3.2.9 « Potager et fruits » (image 3) et la catégorie grammaticale *n.f.* (image 4). Comme nous avons accepté toutes les propositions faites, le résultat final est celui de l'image 5.

[← Voltar á índice](#)

Obra: PereiraG 1983

+ Nova entrada		✓ Gardar cambios		✗ Cancelar cambios		Filtros destacados: Sen filtrar									
id	1	16	suxestións 16	CS	suxestións CS	17	suxestións 17	2	4	3	5				
15	abobreira		undefined									Planta que produz as abóboras. Noutros puntos do país chamáman abóbora tanto á planta como ao fruto da mesma.			

Image 2 - Proposition de lemme.

[← Voltar á índice](#)

Obra: PereiraG 1983

+ Nova entrada		✓ Gardar cambios		✗ Cancelar cambios		Filtros destacados: Sen filtrar									
id	1	16	suxestións 16	CS	suxestións CS	17	suxestións 17	2	4	3	5				
15	abobreira				undefined							Planta que produz as abóboras. Noutros puntos do país chamáman abóbora tanto á planta como ao fruto da mesma.			
					3.2.9 Horta e froita										

Image 3 - Proposition de classificateur sémantique.

[← Voltar á índice](#)

Obra: PereiraG 1983

+ Nova entrada		✓ Gardar cambios		✗ Cancelar cambios		Filtros destacados: Sen filtrar									
id	1	16	suxestións 16	CS	suxestións CS	17	suxestións 17	2	4	3	5				
15	abobreira						undefined					Planta que produz as abóboras. Noutros puntos do país chamáman abóbora tanto á planta como ao fruto da mesma.			
							sf								

Image 4 - Proposition de catégorie grammaticale.

[← Voltar á índice](#)Obra: **PereiraG 1983**

Filtros destacados: Sen filtrar									
id	1	16	suxestións 16	CS	suxestións CS	17	suxestións 17	2	4
15	abobreira	abobreira	aboboreira	3.2.9	3.2.9	sf	sf		Planta que produz as abóboras. Noutros puntos do país chamáman abóbora tanto á planta como ao fruto da mesma.

Image 5 - Enregistrement traité de l'œuvre *PereiraG* (1983).

- 22 Enfin, nous tenons à souligner que la participation au projet *Patrimoine lexical de la Gallaecia* (*Patrimonio léxico da Gallaecia*⁹), ayant fait suite au *TLPGP*, a permis de concevoir et de mettre en place une application informatique (image 6) qui facilite la mise en relation des géosynonymes. Une telle application permet d'attribuer un hyperonyme aux différents lemmes ou variantes liés à un concept, tant du domaine galicien que du portugais européen et brésilien. Les recherches pour l'établissement de l'hyperonyme peuvent être ajustées en sélectionnant un champ sémantique spécifique et en utilisant également une définition. Elle permet, en outre, de relier les ensembles de lemmes galiciens et portugais, puisqu'ils apparaissent séparément dans l'application de recherche du *TLPGP*, tel que nous l'avons mentionné ci-avant.

Vinculación de xeosinónimos

Tesouro de léxico patrimonial

Variante	Lema	Concepto	Idioma
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Definición	Obra	Clasificador semántico	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
Lema vínculo	<input type="text"/>	Concepto	<input type="text"/>
<input type="button" value="Limpar selección"/>	<input type="button" value="Limpar todo"/>	<input type="button" value="Aplicar aos seleccionados"/>	<input type="button" value="Aplicar aos seleccionados"/>
<input type="button" value="Buscar"/>			
0 Rexistros			
Variante	Lema	Clase	...
Clasificador semántico			
Lema vínculo			
Concepto			

Image 6 - Application pour la liaison de géosynonymes.

- 23 Quant à l'image 7, elle présente un exemple de connexion où, sous le concept *nogueira* « noyer », les divers noms existant en galicien et en portugais sont regroupés pour désigner cet arbre : *concheira*, *concheiro*, *noceira*, *noceiro*, *nozeira*, *nogal*, *nogueira* et *nogueiro*. En outre, dans la colonne « lema vínculo » (« lemme lien »), la forme *noceira* est placée pour relier les formes galicienne et portugaise *noceira* et *nozeira*, divergeant seulement au niveau de l'orthographe. Elles apparaissent ainsi représentées sous le même symbole et la même couleur et non sous deux dénominations différentes dans la représentation cartographique automatiquement générée à partir de cette application.

Tesouro de léxico patrimonial

Variante <input type="text"/>	Lema <input type="text"/>	Concepto <input type="text" value="nogueira"/>	Idioma gl pt eu
Definición <input type="text"/>		Obra <input type="text"/>	Clasificador semántico Horta e froita

Variante	Lema	Clase	...	Clasificador semántico	Lema vínculo	Concepto
anoeiro	noceiro	sm	3.2.9	Horta e froita		nogueira
anozéra	nozeira	sf	3.2.9	Horta e froita	noceira	nogueira
concheira	concheira	sf	3.2.9	Horta e froita		nogueira
concheiro	concheiro	sm	3.2.9	Horta e froita		nogueira
nezeiré	nozeira	sf	3.2.9	Horta e froita	noceira	nogueira
noceira	noceira	sf	3.2.9	Horta e froita	noceira	nogueira
noceira	noceira	sf	3.2.9	Horta e froita	noceira	nogueira
noceira	noceira	sf	3.2.9	Horta e froita	noceira	nogueira
noceira	noceira	sf	3.2.9	Horta e froita	noceira	nogueira
noceira	noceira	sf	3.2.9	Horta e froita	noceira	nogueira
nogal	nogal	sm	3.2.9	Horta e froita		nogueira
nogueira	nogueira	sf	3.2.9	Horta e froita		nogueira
nogueira	nogueira	sf	3.2.9	Horta e froita		nogueira
nogueira	nogueira	sf	3.2.9	Horta e froita		nogueira
nogueira	nogueira	sf	3.2.9	Horta e froita		nogueira
nogueira	nogueira	sf	3.2.9	Horta e froita		nogueira
nogueira	nogueira	sf	3.2.9	Horta e froita		nogueira

Image 7 - Notion *nogueira*.

- 24 En ce qui concerne la représentation cartographique, il convient de remarquer que la répartition des dénominations existantes d'un concept a été enregistrée sur une seule carte de l'Eurorégion Galice-Nord du Portugal (image 8), laquelle suppose une autre avancée récente puisqu'il n'avait pas été possible, jusqu'à présent, de représenter de manière unifiée le territoire galicien et portugais dans l'application de consultation du *TLPGP*. Cette application cartographique, bientôt disponible, donnera non seulement lieu à des recherches par concept et par champ sémantique mais elle offrira aussi la distribution des lemmes existants. Ses usagers pourront également désélectionner les lemmes qu'ils ne veulent pas voir représentés sur le territoire en fonction de leurs intérêts.

Image 8 - Carte de l'Eurorégion Galice-Nord du Portugal.

5. Fonctionnement du portail

- 25 La plateforme de consultation des données lexicales contenues dans le *TLPGP* est en accès libre sur Internet depuis 2014. Sur la page d'accueil, son utilisateur dispose d'informations générales sur le projet et ses objectifs, les équipes de travail, les travaux composant le corpus – qui continue à s'enrichir –, ainsi qu'un inventaire des travaux en lexique dialectal de la Galice, du Portugal et du Brésil. Des chiffres actualisés sont également fournis pour le nombre d'ouvrages, d'enregistrements, de lemmes et de variantes saisis dans la base de données. Tous ces renseignements peuvent être consultés en trois langues : galicien, portugais et anglais. Sur cette même page, il est dorénavant possible d'effectuer des recherches de deux types – simples et avancées –, en plus de pouvoir choisir de les réaliser par *lemme* ou *variante dialectale*.
- 26 En ce qui concerne la recherche simple, il n'apparaît qu'une seule case vide, où il suffit d'introduire l'unité lexicale à étudier. De plus, au fur et à mesure que le mot est écrit, l'application propose, à travers un menu déroulant, différents lemmes ou variantes par ordre alphabétique contenant les caractères introduits. À cet égard, il est à noter, comme on peut le constater sur l'image 10, que les unités et expressions pluriverbales ont leur propre entrée dans la base de données, du fait qu'elles ont été dûment étiquetées à cette fin.

Image 9 - Recherche simple.

Image 10 - Liste déroulante avec les propositions de recherche.

- 27 Une fois que la forme lexicale a été écrite, l'outil renvoie les résultats sur un écran, où l'information est organisée en trois parties, comme on peut le voir sur l'image suivante, en plus de conserver l'espace destiné à introduire le lemme ou la variante dans la zone supérieure (en blanc).

Image 11 - Présentation de l'information.

- 28 L'image présente une colonne à gauche (hachurée en bleu), où se trouvent les filtres correspondant aux localisations, aux champs sémantiques, aux lemmes et aux variantes enregistrées. L'utilisateur est alors en mesure d'affiner sa recherche en les activant ou en les désactivant, pouvant notamment décider de ne sélectionner que les territoires lusophones et le champ sémantique 1.2 Météorologie.
- 29 Dans la deuxième zone, située dans la partie centrale de l'écran (hachurée en gris), la plateforme offre les données correspondant à la forme consultée, accompagnées d'une indication du travail de leur collecte, du lemme et de la catégorie qui y est associée, du domaine linguistique dont relève chacune (galicien, portugais européen ou brésilien) et de la classification sémantique correspondante. De plus, une icône indique s'il y a une image associée, laquelle pourra être ouverte en cliquant dessus. Par ailleurs, lorsque le curseur est placé sur l'un des exemples, l'endroit où le mot a été enregistré s'éclaire sur la ou les cartes.
- 30 Enfin, dans la troisième zone (hachurée en rouge), la cartographie¹⁰ apparaît automatiquement, reliée aux enregistrements de la partie supérieure. Si le curseur est situé sur les zones en bleu, c'est-à-dire les communes ou les mésorégions où le lemme ou la variante saisie dans la zone de recherche a été recueillie, l'enregistrement correspondant s'allume.
- 31 En revanche, si la recherche avancée est activée, les trois cases qui apparaissent permettent de sélectionner le lieu, le champ sémantique et la catégorie du mot. Lorsque l'utilisateur aura cliqué sur chacun d'entre elles, un menu déroulant s'ouvrira avec les options existantes pour chaque champ, l'option de marquer ceux qui l'intéressent étant dès lors possible. Il pourra, par exemple, établir qu'il souhaite obtenir les lemmes existant au Portugal liés au temps et classés comme expressions verbales. L'application émettra alors une liste avec les lemmes ou variantes classées par ordre alphabétique qui répondent à ces critères dans tous les ouvrages contenus dans la base de données (image 13).

Image 12 - Recherche avancée.

Image 13 - Cases déroulantes avec des options de recherche.

Image 14 - Résultats de la recherche avancée avec les critères « Portugal », « Temps », « Expression verbale ».

- 32 Nous considérons que les deux possibilités de recherche, simple et avancée, s'avèrent très intuitives pour un usager moyen et qu'elles n'impliquent aucune difficulté pour les personnes habituées à l'utilisation des outils lexicographiques en ligne, comme les corpus textuels ou les dictionnaires.

6. Lignes d'utilisation du *TLPGP*

- 33 De tout ce qui a été dit jusqu'à présent sur le *TLPGP*, on peut en déduire que les lignes d'action qu'il nous offre sont multiples et qu'il constitue un outil très utile pour divers domaines d'étude.
- 34 Il s'agit en effet d'un outil fondamental pour l'élaboration de dictionnaires, de corpus d'unités pluriverbales et de vocabulaires thématiques, en plus d'aider à la mise en place de la norme, car il permet notamment de :

- 35 a) Connaître des mots qui ne figurent pas dans les dictionnaires et les ouvrages de référence lexicographiques et/ou qui ne font pas partie des compilations standard (par exemple : *abaxada* « urticaire », *investres* « membranes fœtales », *arreleixar* « sonner le glas », etc.).
- 36 b) Avoir accès aux nouvelles significations des formes déjà introduites dans les dictionnaires (par exemple : *acolleita* « infection », *alcanzar* « haleter », *morriñento* « avare », etc.).
- 37 c) Effectuer des recherches directes d'unités complexes, qui permettent d'en obtenir une meilleure visibilité lorsqu'elles apparaissent sur un pied d'égalité avec les unités mono-lexicales (par exemple : *non dar chío* « ne dire mot »).
- 38 d) Faire des recherches au moyen de champs sémantiques ayant recours à une nouvelle façon de visualiser les données grâce à la *recherche avancée*. Comme indiqué, les recherches peuvent être effectuées au moyen d'un ou plusieurs champs, aussi bien dans les trois territoires du projet (Galice, Portugal et Brésil) que dans un seul d'entre eux.
- 39 Le portail pouvant offrir une vue simultanée des variantes existant en Galice, au Portugal et au Brésil, il permet de mener des travaux de nature contrastive ou axés sur l'analyse des processus de variation et de changement au sein de l'ensemble linguistique galicien, portugais et brésilien. Il peut aussi révéler l'existence de « discontinuités » dans la répartition territoriale actuelle de certaines solutions et, par là même, mettre en lumière la naissance de nouveaux processus, ou leur propagation, leur changement et leur décadence s'y étant produits. Par ailleurs, le fait que le *TLPGP* soit constitué d'ouvrages d'époques très diverses, y compris l'actuelle, permet de savoir si certains processus de changement se poursuivent dans le temps ou s'ils se sont perdus.
- 40 De plus, grâce à l'application à même d'établir des hyperliens à partir des données du *TLPGP*, il est possible d'effectuer des recherches onomasiologiques, c'est-à-dire d'analyser tous les signifiants existants pour un concept. Par exemple, pour la notion « partie saillante des os qui composent l'articulation de la jambe avec le pied »¹¹, plusieurs dénominations apparaissent dans l'application : *nocello*, *tornocelo*, *novelo*, *mondongo*, *garrão*, *calcañar*, *mocotó*..., dont certaines ne sont pas enregistrées dans les dictionnaires. Une autre possibilité offerte par le *TLPGP* est d'analyser la motivation se trouvant à l'origine du nom attribué à un concept et les mécanismes utilisés dans la création des différentes dénominations, qui ont tendance à être principalement de nature métaphorique.
- 41 Enfin, le portail fournit à la communauté scientifique de nombreux renseignements sur le patrimoine matériel et immatériel traditionnel de la Galice, du Portugal et du Brésil, lesquels peuvent être utilisés pour une recherche historique et ethnographique. Dans un nombre important d'enregistrements, leur définition apporte ainsi des précisions sur les aspects matériels (outils, usages...) et les us et coutumes y étant associés, accompagnées, dans certains cas, de dessins ou de photographies.

7. Conclusions

- 42 Nous avons expliqué, dans ce travail, les principales caractéristiques et possibilités d'exploitation du *TLPGP*. Nous avons également mis en avant les difficultés et les défis auxquels nous avons fait face. En guise de conclusion, nous tenons à souligner les

aspects qui font de cette plateforme un projet unique. En premier lieu, elle représente une avancée par rapport aux corpus et aux banques terminologiques traditionnels en termes d'informations et de présentation, puisqu'outre la numérisation, le traitement et l'étiquetage des données lexicales, cette plateforme comprend des cartes générées automatiquement par rapport à chaque recherche. Il s'agit par conséquent d'un outil qui contribue aux études lexicographiques et dialectales les plus diverses, tout en mettant à disposition, sur Internet et en libre accès, une grande quantité de données qui étaient restées, pour la plupart, inédites jusqu'à présent.

- 43 Il incorpore également d'autres renseignements contenus dans les œuvres sources, telles que des images ou des proverbes. C'est pour toutes ces raisons que nous sommes à même de présenter cette plateforme comme une ressource hybride, combinant les avantages d'un dictionnaire basé sur des corpus textuels aux possibilités d'une ressource dialectale et permettant de sauvegarder les données ethnolinguistiques que transmettent les textes originaux.

BIBLIOGRAPHIE

ÁLVAREZ PÉREZ Xosé Afonso & SOUSA Xulio (2013), « The Tesouro do léxico patrimonial galego e portugués. A Galician and Portuguese word bank », *Dialectologia. Especial issue*, 4, 5-24.

ÁLVAREZ Rosario, ÁLVAREZ PÉREZ Xosé Afonso, SARAMAGO João & SOUSA Xulio (2009), « Presentación de un corpus digital de léxico tradicional : Tesouro do léxico patrimonial galego e portugués », *Fonetica și Dialectologie*, 28, 5-19.

ÁLVAREZ Rosario [coord.] (2014), *Tesouro do léxico patrimonial galego e portugués (TLPGP)*, Saint-Jacques-de-Compostelle : Instituto da Lingua Galega, en ligne sur <<http://ilg.usc.gal/Tesouro/>>.

ÁLVAREZ DE LA GRANJA María & NEGRO ROMERO Marta (2015), « O processo de lematização no Tesouro do léxico patrimonial galego e português », F. C. Altino et al. (dir.), *Anais do III CIDS: Congresso Internacional de Dialectologia e Sociolinguística*, Londrina (Brésil) : Universidade Estadual de Londrina, 848-862.

BENITO Alejandro et al. (2016), « Aspatio-temporal visual analysis tool for historical dictionaries », F. J. García-Peñalvo (dir.), *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'16)*, ACM, New York : USA, 985-990.

BENITO Alejandro et al. (2018), « Creating Meaningful Narratives in Collections of Historical Lexical Data », *Journal for Geographic Information Science, GI_Forum 2018*, 6(2), 50-57.

BERNHARD Delphine (2014), « Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian », *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, Reykjavík : Iceland, 23-29.

CASTRO Ivo (2013), « Formação da língua portuguesa », P. Raposo et al. (dir.) *Gramática do português*, Lisbonne : Gulbenkian, 1, 7-16.

CORBELLA DÍAZ Dolores et al. [dir.] (2018), *Historia del léxico español y Humanidades digitales*, Berlin : Peter Lang.

DUBERT GARCÍA Francisco (2015), « Del verbo gallego medieval al contemporáneo: entre el contacto y la autonomía lingüísticos », *50th International Congress on Medieval Studies*, Kalamazoo, Michigan : Western Michigan University.

FERREIRA Aurélio Buarque de Holanda (2009), *Novo Dicionário Aurélio da Língua Portuguesa*, Curitiba : Editora Positivo.

GONZÁLEZ GONZÁLEZ Margarita (2002), *El gallego hablado de Cangas de Morrazo* (thèse de doctorat), Université d'Oviedo, Oviedo.

GONZÁLEZ SEOANE Ernesto (2015), « Aspectos metodolóxicos do deseño e elaboración do *Tesouro do léxico patrimonial galego e portugués* (TLPGP) », O. Castro & M. Liñeira (dir.), *Trama e urda. Contribucións multidisciplinares desde os estudos galegos*, Saint-Jacques-de-Compostelle : Consello da Cultura Galega, 45-60.

HOUAISS Antônio (2001), *Dicionário Houaiss da Língua Portuguesa*, Rio de Janeiro : Objetiva.

INFOPÉDIA (2003), *Dicionário da Língua Portuguesa da Porto Editora*, en ligne sur <<http://www.infopedia.pt/lingua-portuguesa/>>.

LEXIKON EDITORA DIGITAL (2007), *iDicionário Aulete da língua portuguesa*, en ligne sur <<http://aulete.com.br>>.

NEGRO ROMERO Marta (2015), « O *Tesouro do léxico patrimonial galego e portugués* no marco do espazo lingüístico galego-portugués (PE e PB) », O. Castro & M. Liñeira (dir.), *Trama e urda. Contribucións multidisciplinares desde os estudos galegos*, Saint-Jacques-de-Compostelle : Consello da Cultura Galega, 77-98.

MUTTER Christina & WIATR Aleksander (2018), « The Virtual Research Environment of VerbaAlpina and its Lexicographic Function », J. Čibej, V. Gorjanc, I. Kosem & S. Krek (dir.), *Proceedings of the XVIII EURALEX International Congress*, Ljubljana : University Press, Faculty of Arts Ljubljana, 775-785.

OBERHOLZER Susanne & KUNZMANN Markus (2017), « Geolinguistic Documentation of Multilingual Areas – VerbaAlpina and the Challenges of Digital Humanities », I. Buchstaller & B. Siebenhaar (dir.), *Language Variation - European Perspectives VI. Selected papers from the 8th International Conference on Language Variation in Europe (ICLaVE 8)*, Leipzig : Benjamins, 199-213.

PEREIRA A. Gomes (1983), *O século XIX na linguagem popular dos concelhos de Alijó, Mesão Frio, Mondim de Basto, Murça, Ribeira de Pena, Sabrosa, Santa Marta de Penaguião e Vila Real*, Vila Real : Comissão Regional de Turismo da Serra do Marão.

PÉREZ FROIZ María Inés (1972), *El léxico de Santa María de Oirós* (mémoire de licence), Universidad de Santiago de Compostela, Saint-Jacques-de-Compostelle.

POLÍTICA IN (2008), *Dicionário Priberam da Língua Portuguesa*, en ligne sur <<http://www.priberam.pt/DLPO/>>.

REAL ACADEMIA GALEGA (2012), *Dicionario da Real Academia Galega*, en ligne sur <<http://www.realacademiagalega.org/dicionario>>.

SANTAMARINA Antón & GONZÁLEZ GONZÁLEZ Manuel [dir.] (2004), *Vocabulario ortografico da lingua galega*, La Corogne : Real Academia Galega / Instituto da Lingua Galega, <<http://www.realacademiagalega.org/recursos-volg>>.

VANDERSCHUEREN Clara & MENDES Amália (2015), « Panorama de los corpus y textos del portugués europeo contemporáneo », M. Iliescu & E. Roegiest (dir.), *Manuel des antologies, corpus et textes romans*, Berlin, Boston : Walter De Gruyter, 58-80.

VAN KEYMEULEN Jacques & DE TIER Véronique (2010), « Pilot Project: a Dictionary of the Dutch Dialects », A. Dykstra & T. Schoonheim (dir.), *Proceedings of the XIV Euralex International Congress*, Leeuwarden, Pays-Bas : Fryske Akademy, 754-763.

NOTES

1. Cet article fait partie du projet *Patrimonio léxico da Gallaecia* (en français, *Patrimoine lexical de la Gallaecia*), qui s'inscrit dans *Patrimonio cultural da Eurorrexión Galicia-Norte de Portugal: valoración e innovación* (0358_ GEOAPARD_1_E, en français, *Patrimoine culturel de l'Eurorégion Galice-Nord du Portugal : évaluation et innovation*), projet transfrontalier du programme opérationnel EP-INTERREG V A Espagne-Portugal (POCTEP) 2014-2020.
2. L'accès au TLPGP est libre et gratuit sur Internet depuis 2014 : <<http://ilg.usc.es/Tesouro/en/>>.
3. La liste complète des membres de chaque équipe peut être consultée sur le lien suivant : <<http://ilg.usc.gal/Tesouro/en/equipo.html>>.
4. Afin d'approfondir des questions liées aux dernières contributions des humanités numériques en matière de lexicographie diachronique et de corpus linguistiques dans le domaine portugais et hispanique, les travaux de Van Keymeulen & De Tier (2010), Bernhard (2014), Vanderschueren & Mendes (2015), Benito *et al.* (2016), Oberholzer & Kunzmann (2017), Corbella Díaz *et al.* (2018), Mutter & Wiatr (2018) et Benito *et al.* (2018) sont disponibles.
5. Le chercheur portugais Ivo Castro l'explique comme un processus remontant au XI^e siècle, puisque c'est à cette époque-là que vit le jour le Comté *Portucalense*. Ce fait historique a conduit à une plus grande séparation de ce comté du reste du royaume de Galice, dans le nord de la Péninsule ibérique. Ainsi, le territoire situé au nord du Minho (aujourd'hui, la Galice) appartenait au royaume de León, tandis que celui situé au sud de ce fleuve allait devenir indépendant sous le nom de Portugal. Cela a accéléré le processus de différenciation entre les deux comtés et l'histoire extérieure du galicien et du portugais, c'est-à-dire que leur histoire sociale et politique a cessé d'être commune, les choix linguistiques ayant également pris des chemins différents. Il s'est ensuivi, suite à l'intégration du royaume de León à la couronne de Castille en 1230, que la langue galicienne est devenue partie intégrante de la zone d'influence du castillan, d'où elle importait le lexique et certaines caractéristiques grammaticales (Castro, 2013 : 9 ; *apud* Dubert García, 2015).
6. Pour plus d'informations sur les sources, voir : <<http://ilg.usc.es/Tesouro/en/corpus.html>>.
7. Une explication détaillée de ce processus apparaît dans Álvarez de la Granja & Negro Romero (2014).
8. Au moment d'écrire ces lignes, il existe 105 249 variantes et 66 436 lemmes dans la base de données.
9. Ce projet est coordonné depuis l'Institut de langue galicienne de l'Université de Saint-Jacques de Compostelle par la Professeure Rosario Álvarez Blanco, avec la participation du Centre de Linguistique de l'Université de Porto et le Centre d'Études en lettres de l'Université de Trás-os-Montes et de l'Alto Douro. En même temps, il fait partie du projet transfrontalier *Patrimoine Culturel de l'Eurorégion Galice-Nord du Portugal : valorisation et innovation* (GEOARPAD, pilote 4), lequel appartient au programme opérationnel EP-INTERREG V A Espagne-Portugal (POCTEP), dont l'objectif fondamental est la connaissance et la valorisation du patrimoine culturel de l'Eurorégion Galice - Nord du Portugal.

10. Il est à noter que, dans le cas de la Galice, la carte montre non seulement les frontières géographiques officielles de cette Communauté autonome, mais aussi les zones frontalières dans lesquelles est parlé le galicien.

11. « Cheville » en français.

RÉSUMÉS

Le projet international *Tesouro do léxico patrimonial galego e português* (TLPGP, en français, *Trésor du lexique patrimonial galicien et portugais*) intègre un matériel lexicographique dialectal de la Galice, du Portugal et du Brésil, référencé géographiquement dans une seule banque de données, dont l'accès est libre via Internet. Le but de cette initiative est d'aider à conserver et à mettre à la disposition de la communauté scientifique de tels documents, dont beaucoup sont inconnus ou très difficiles à consulter, provenant de sources diverses et dans un état de conservation hétérogène.

Dans notre article, après avoir brièvement expliqué la typologie du matériel qui compose le TLPGP, nous nous pencherons sur certains aspects liés à son traitement, l'information lexicale y étant lemmatisée, classée sémantiquement et grammaticalement et référencée géographiquement, ce qui permet diverses sortes de recherches et la production automatique de la cartographie de chaque mot. En outre, nous montrerons le fonctionnement du portail, en accès libre depuis 2014, et les différents types de consultation qui peuvent y être effectués, en fonction des intérêts de chaque usager. Finalement, nous exposerons quelques-unes des utilisations possibles de ce corpus dialectal, notamment aux niveaux lexicographique et dialectologique.

O projeto internacional *Tesouro do léxico patrimonial galego e português* (TLPGP) integra material lexicográfico dialectal da Galiza, Portugal e Brasil, referenciado geograficamente numa única base de dados e com acesso gratuito através da Internet.

O objetivo desta iniciativa é ajudar a preservar e disponibilizar para a comunidade científica esses documentos, em grande número desconhecidos ou pouco acessíveis para ser consultados, de várias fontes e em estado de conservação heterogêneo.

No nosso artigo explicaremos brevemente a tipologia dos materiais que compõem o TLPGP e abordaremos alguns aspectos relacionados com o seu processamento, posto que a informação léxica é lematizada, classificada semântica e gramaticalmente e referenciada geograficamente, pelo que permite fazer pesquisas de distinto tipo e gerar de forma automática a cartografia de cada uma das palavras. Além disso, mostraremos como funciona o portal, com acesso gratuito desde o ano 2014, e os diferentes tipos de consulta que podem ser realizados, dependendo dos interesses de cada usuário. Finalmente, apresentaremos alguns dos possíveis usos deste corpus dialectal, particularmente nos níveis lexicográfico e dialetológico.

The international project *Tesouro do léxico patrimonial galego e português* (TLPGP—*Treasury of the Galician and Portuguese heritage lexicon*)—integrates dialectal lexicographic material from Galicia, Portugal and Brazil, geographically referenced in a single database and with free access over the Internet.

The aim of this initiative is to help preserve documents from various sources and in a heterogeneous state of conservation and make such documents, which are largely unknown or very difficult to consult, available to the scientific community.

The current article briefly explains the typology of the materials that make up the TLPGP and discuss some aspects related to its processing, since the lexical information is lemmatised, semantically and grammatically classified and geographically referenced, enabling different types of research and the automatic generating of the cartography of each of the words. In addition, the article shows how the site, with free access since 2014, works and the different types of consultation that can be carried out, depending on the interests of each user. Finally, it presents some of the possible uses of this dialectal corpus, particularly at the lexicographic and dialectological levels.

INDEX

Mots-clés : corpus numérique, lexique dialectal, portugais brésilien, portugais européen, galicien

Palavras-chave : corpora digitais, léxico dialetal, português brasileiro, português europeu, galego

Keywords : digital corpora, dialectal vocabulary, Brazilian Portuguese, European Portuguese, Galician

AUTEURS

MARTA NEGRO ROMERO

Université de Saint-Jacques-de-Compostelle

XANDRA SANTOS PALMOU

Université des Antilles – pôle de la Martinique