Trotter, David/Bozzi, Andrea/Fairon, Cédrick (éd.) (2016). *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux*. Nancy, ATILF :
http://www.atilf.fr/cilpr2013/actes/section-16.html.

Réseau Tramontana (2012-2013). Méthodologie de la collecte dynamique de matériaux culturels auprès des habitants et modes de restitution des résultats

Le présent article est écrit à deux mains. Fabrice Bernissan donnera (1) une présentation de la teneur du projet Réseau Tramontana qui lie sept structures situées en France, Italie et Portugal. Ce projet de recherche fait la part belle à la linguistique même si cette discipline ne représente pas la partie la plus importante de nos travaux. Le même auteur présentera (2) deux faits de langue relevés lors d'enquêtes menées dans les vallées haut-pyrénéennes, alors que Giovanni Agresti proposera (3), à partir de trois cas de figure tirés d'un corpus de collectes réalisées dans le massif du Gran Sasso d'Italia (Abruzzes), une démarche de sociolinguistique pour que le travail de terrain se transforme, aussi, en un levier de développement local.

1. Le projet *Réseau Tramontana*

Les savoirs traditionnels de l'Europe constituent des ponts entre les personnes, les peuples et leur environnement. Ces savoirs sont parfois délaissés et/ou malmenés sous l'effet de la globalisation. L'objectif général de notre projet est de sauvegarder la mémoire du paysage et de la communauté et d'utiliser les matériaux collectés comme outils d'éducation, de valorisation du sujet et de développement des territoires.

L'axe majeur de *Réseau Tramontana* (désormais Tramontana) consiste, après l'élaboration de grilles de questionnements ouverts utiles lors de nos collectes[1], à mettre en place une action concrète d'enquêtes de terrain dans un souci de respect de la diversité et de l'intégrité des territoires, de sauvegarde de la mémoire et de promotion de l'identité et des patrimoines immatériels. Pour parvenir à notre objectif nous avons mis en place, par la coopération, un protocole méthodologique d'enquêtes de terrain. Le consensus fait sur cette question, nous avons ensuite réalisé 790 enquêtes de terrain qui regardent diverses disciplines : la linguistique, l'anthropologie, l'ethnologie, la sociologie et la musicologie. Ces 790 enquêtes ont été numérisées puis ont fait l'objet d'un archivage selon la méthodologie retenue (fiches chronothématiques et, moins souvent, transcription diplomatique).

L'ensemble des matériaux collectés sont (ou seront à terme) mis en commun pour leur étude et leur diffusion. Les axes de travail qui ont été retenus au fil de nos rencontres sont les suivants :

[1] Nous proposons en annexe un exemple de grille en rapport avec l'enquête linguistique.

1. Usages des cloches
2. Des chèvres aux vaches : les animaux et l'élevage
3. Des saints et des pèlerinages
4. De Noël à Pâques : rites d'hiver
5. Légendes d'origine des lieux, des choses et des animaux
6. Chants et travail
7. Mines, barrages et usines
8. Départ et retour au pays
9. Perceptions de l'évolution de la société rurale et de l'environnement
10. Pratique, perception, représentation des langues

La diffusion des résultats obtenus occupe une place centrale du projet Tramontana. Nous prévoyons à terme de mettre en ligne 200 extraits d'enquêtes. Nous faisons un effort particulier autour de la communication puisque nous avons organisé trois forums (en Gascogne, dans les Abruzzes et au Portugal) et terminerons ce projet par un séminaire de bilan en Toscane. Enfin le volet communication sera complété par des publications par voie de presse, par des articles dans des revues spécialisées, la production de CD et l'édition d'un fascicule méthodologique.

Nous observons, au point où nous en sommes, que les contenus linguistiques de nos enquêtes s'articulent autour de quatre axes.

1. Le recueil d'éléments théoriques et méthodologiques portant sur la toponymie narrative ;
2. Le recueil de faits de langue en raison de particularités dialectales saillantes : notamment à Pietracamela (massif du Gran Sasso d'Italia) et dans les vallées pyrénéennes (Extrême de Salles, Azun, Pays Toy) ;
3. Le recueil de matériaux sociolinguistiques : l'auto-conscience et la perception linguistiques (la nature, le rôle et « l'importance » de la variété linguistique locale, le « lien à la langue » du sujet interviewé, la transmission) ;
4. Le recueil, chez l'habitant, de textes écrits (manuscrits, références de publications confidentielles et autres écrits mineurs) de préférence rédigés dans la variété linguistique locale ou portant sur l'histoire et la culture locales.

Les partenaires du projet Tramontana sont :

En France : l'association Eth Ostau comengés, qui réalise des enquêtes ethnologiques depuis 2007 ainsi que des documentaires vidéo en gascon, principalement dans le département de la Haute-Garonne et de l'Ariège ; l'association Numériculture-Gascogne, qui mène des enquêtes linguistiques et ethnologiques dans le département du Gers et des Pyrénées-Atlantiques depuis 2007, et l'association Nosauts de Bigòrra, qui conduit depuis 2001 des enquêtes linguistiques et ethnographiques et produit des documentaires DVD dans les Hautes-Pyrénées.

En Italie : l'association Bambun, qui réalise des enquêtes anthropologiques et publie dans les Abruzzes ; l'association La Leggera, qui mène des enquêtes de terrain (musique, danse) et anime des ateliers de restitution et d'exploitation artistique en Toscane. Et l'association Langues d'Europe et de la Méditerranée-Italia, qui organise des colloques de linguistique et publie (région des Abruzzes et Italie centro-méridionale).

Enfin, au Portugal : l'association Nodar-Binaural, qui mène des enquêtes sonores et audiovisuelles de terrain et procède à des restitutions publiques (district de Viseu principalement).

Les financements, obtenus après parfois de longues et délicates démarches, nous permettent de mener ce projet sur une période de vingt mois. Les principaux partenaires sont l'Union européenne à hauteur de 200.000 € (50 % du budget), les États par le biais des Ministères de la culture, certaines Régions et collectivités locales, et enfin, les Parcs nationaux et régionaux.

2. Deux faits de langue observés dans le gascon des Hautes-Pyrénées

Nous avons été extrêmement surpris de rencontrer une aussi forte variété dialectale lors des enquêtes menées dans la vallée du Gave (Hautes-Pyrénées). Nous rejoignons le propos de Martin Glessgen qui, s'exprimant à propos de la différenciation des langues romanes, écrit : « les montagnes ne sont pas des limites en tant que telles mais plutôt des zones de fracture naturelles, autour desquelles se produisent, dans des processus séculaires, des phénomènes de séparation linguistique »[2]. Nous pensons que nous pouvons appliquer ce principe à l'échelle microdialectale. En effet les zones qui nous intéressent concernent toutes des entités proches et de faible taille à la fois géographique et démographique : nos points d'enquêtes sont disséminés dans les vallées d'Argelès, d'Azun, de Batsurguère, du Castelloubon et le Pays Toy (vallée de Barège et haute vallée du Gave).

Un point de phonologie à particulièrement retenu notre attention. On remarque, en effet, que les habitants des trois communes situées dans l'Extrême de Salles (au nord de la vallée d'Argelès) font évoluer la finale des mots [e] > [ej]. Ainsi nous relevons :

Argelès	Extrême de Salles Commune de Gez	Français
[bu'de]	[bu'dej]	'beurre'
[le]	[lej]	'laid'
[pes]	[pej]	'pied'
[karti'e]	[karti'ej]	'quartier'
[kasta'Qe]	[kasta'Qej]	'châtaignier'
[bulan'je]	[bulan'jej]	'boulanger'
[kusi'ne]	[kusi'nej]	'cuisinier'

[2] Glessgen 2012, 349.

Argelès	Extrème de Salles Commune de Gez	Français
[au'je]	[au'jej]	'berger'
[ba'ke]	[ba'kej]	'vacher'
[pa'pe]	[pa'pej]	'papier'
[pu'me]	[pu'mej]	'pommier'
[da're]	[da'rej]	'dernier'
[prezo'ne]	[prezo'nej]	'prisonnier'
[pati'sje]	[pati'sjej]	'pâtissier'
['je]	[jej]	'hier'
[a'3es]	[a'3ejs]	'affaires'
[pla'ze]	[pla'zej]	'plaisir'
[fre'zje]	[frezj'ej]	'fraisier'
[gruze'je]	[gruze'jej]	'groseillier'
[sa'be]	[sa'bej]	'savoir'
[bu'le]	[bu'lej]	'vouloir'
[pu'de]	[pu'dej]	'pouvoir'
[jes]	[jejs]	Gez
[arje'les]	[arje'ljes]	Argelès

Les matériaux relevés donnent à voir que le phénomène touche tous les types de mots du lexique, y compris les mots affixés. De nouvelles enquêtes devront nous permettre de vérifier que les conjugaisons suivent le même traitement.

Un autre point linguistique nous intéresse fortement : il s'agit de la variation lexicale rencontrée dans ces petites vallées distantes d'à peine quelques kilomètres. Le tableau ci-dessous montre la grande variété lexicale que comportent ces petites entités.

Argelès /Castel-loubon	Azun	Pays Toy	Français
[man'dCrCs]	[tu'matCs]	[try'fes]	'pommes de terre'
[bu'kaj]	[ky'buG]	[lun'tDrC]	'ouverture dans une grange'
[eska'bDlC]	[ka'jDrC]	[la ska'bDlC]	'chaise'
[bas'tu]	[tCt'Gu]	[tCt'Gu]	'bâton'
[ares'tet]	[ba'liC]	[ares'tet]	'râteau'
['driQ]	['dret]	[Gi'Qau]	'peu'
['kesC] / [ka'mizC]	['kesC]	['kesC]	'chemise'
[la'bDs]	[la'bDs]	[Ga'bDts]	'alors'
[brespa'ja]	[ber'na]	[bre'na]	'collation l'après-midi'
[se'C]	[ky'bat]	[ky'bat]	'seau'
[je'lCs] / [tam-bu'rDu]	[je'lCs]	[tambu'rDu]	'tombereau'

Nous observons lors de nos enquêtes le fort attachement des habitants à leur propre forme lexicale. Les locuteurs emploient ces formes à défaut de tout autre. Néanmoins ils connaissent, pour la plupart, les termes employés par leurs voisins. Cette connaissance microdialectale provient vraisemblablement de la fréquentation traditionnelle des mêmes marchés, à Argelès, et des mêmes foires. Par ailleurs des mariages entre valléens ont donné lieu à des déplacements et donc à la pénétration des formes employées localement. La différenciation lexicale donne lieu à de (gentilles) moqueries de part et d'autre de la vallée des Gaves, et dans tous les cas à une différenciation quant à l'appartenance des locuteurs : l'appartenance irréductible à une vallée et à un groupe humain déterminé.

On le voit, le terrain n'en finit pas de délivrer des matériaux qui attendent une analyse plus poussée. Gageons que les autres territoires que couvre le projet Tramontana seront aussi riches du point de vue de la linguistique.

3. De la collecte de matériaux culturels à la linguistique du développement social

Même s'il ne s'agit pas d'un projet linguistique *stricto sensu*, dans toutes ses phases Tramontana tourne autour de la parole-mémoire. Elle est d'abord collectée moyennant la caméra vidéo dans le cadre d'entretiens individuels. Par la suite, elle est régulièrement réinjectée dans l'espace collectif par le biais de laboratoires, *focus group*,

présentations publiques et publications diverses. Elle est donc susceptible d'être analysée non seulement quant au contenu, à ce qu'elle raconte, ou à la forme (linguistique, paralinguistique ou épi-linguistique) qui l'actualise, mais également par rapport à la dialectique permanente qu'elle établit entre le sujet et le territoire, l'identité et le maillage social qui l'enveloppent et qui l'habitent.

Dans la perspective du Projet, il est à notre sens nécessaire, voire urgent que la prise en compte de cette dialectique dépasse le niveau descriptif de la sociolinguistique classique pour atteindre celle que nous avons ailleurs appelée la « linguistique du développement social » (Agresti en préparation, désormais LDS) : une linguistique d'intervention[3] à même d'explorer et d'exploiter les patrimoines linguistico-mémoriels des communautés *pour mieux les faire vivre*. Cette urgence est due aux changements socio-économiques parfois abrupts, à la baisse démographique qui frappe à quelques exceptions près toutes les montagnes d'Europe ainsi qu'à la fragilisation des liens intergénérationnels et à la pulvérisation culturelle-mémorielle qui s'ensuivent.

Synthèse oblige, nous nous bornerons à présenter ici trois cas de figure paradigmatiques qui montrent comment l'analyse linguistique peut gagner en rendement social et même théorique grâce au dialogue constant avec le terrain et d'autres disciplines à la fois distinctes et complémentaires (notamment : l'anthropologie, l'histoire, l'ethnographie, toutes présentes dans Tramontana) : 1-2) la prise en compte de la portée narrative de l'interprétation et de la création toponymiques; 3) la mise en question de la notion de ' jargon de métier ' par la socialisation et l'incorporation de ce même jargon de la part de l'ensemble de la communauté linguistique dont il relève.

Précisons nos cadre et corpus. Nous avons enquêté dans huit communes des Abruzzes (situées autour du massif du Gran Sasso d'Italia, le sommet des Apennins) et dans une commune plus méridionale, San Marco dei Cavoti (Campania). Au total, cet espace compte 22648 habitants sur une superficie de 457,66 km^2. Nous avons interviewé, entre août 2012 et septembre 2013 et toutes communes confondues, 102 témoins d'un âge moyen de 84 ans. Cela fait environ 130 heures d'entretiens filmés en plus d'une vingtaine d'heures et de centaines de photos concernant d'autres moments de la vie et individuelle et sociale des territoires étudiés.

3.1. La toponymie narrative

La toponymie est un carrefour scientifique et disciplinaire. Non seulement le linguiste ou le philologue, mais également l'historien, le géographe, le juriste peuvent s'y pencher et l'éclaircir. Qui plus est, sous le praxème ' toponyme ' se cachent des objets sociohistoriques et linguistiques fort divers. De surcroît, elle se situe au cœur de la dialectique de sujet, langue, espace et est donc censée rapprocher, voire souder ces trois éléments. Au vu de cette formidable complexité, il est plusieurs manières d'aborder l'analyse toponymique : non seulement en synchronie et en diachronie,

[3] Cf. le colloque International de la SHESL *Linguistique d'intervention. Des usages sociopolitiques des savoirs sur le langage et les langues*, Paris 26-28 janvier 2012.

mais également en croisant les deux coupes temporelles via la prise en compte de l'imaginaire individuel et collectif justifiant ou remotivant le sens interne au toponyme lui-même. Ce qui nous intéresse plus particulièrement dans la perspective de la LDS, c'est la dimension proprement narrative, discursive, de la toponymie populaire : la mémoire des lieux, racontée, spécialement à l'oral, pourrait être, sous certaines conditions, à la fois acte de lien intracommunautaire et intergénérationnel ainsi qu'acte de rapprochement et d'appréhension, de connaissance du territoire.

Nous avons ailleurs approfondi les tenants et aboutissants de cette notion de « toponymie narrative » (Agresti, en préparation), publié des approches d'analyse du discours appliquées à l'analyse toponymique (Agresti 2012) et illustré quelques cas de figure lors d'un colloque international. Nous reprenons ici quelques propos débattus à cette occasion :

> un toponyme ne donne normalement que des informations très générales ou alors très ponctuelles [...] sur l'histoire ou la nature du lieu. Pourtant, ces lieux sont le cadre de nos vies, et donc, éventuellement, de nos *récits de vie*. Il est donc une seconde greffe linguistique de l'humain sur le territoire : le *sens* que tel ou tel lieu – et partant de tel ou tel toponyme – a pour tel ou tel sujet. Chez ce dernier, l'amour pour un quartier, une ville, une région etc. se dit par son évocation verbale ; il est donc inscrit en quelque sorte dans les toponymes désignant ces cadres de vie. À ce niveau, l'imaginaire individuel se croise sans cesse avec l'imaginaire collectif, le présent dialectise et souvent déforme les toponymes, qui peuvent devenir opaques voire être remplacés. […] Nous retenons l'idée que tout toponyme peut être un résonateur de la mémoire, et donc un vecteur de l'imaginaire. [Au sujet des Atlas de toponymie narrative] le niveau discursif ferait [...] ' parler ' le territoire par la voix de ses habitants se racontant à partir de repères toponymiques. Il en résulterait un tissu à la fois narratif et topologique-relationnel qui pourrait, à l'avenir, donner sa contribution à ce que la richesse culturelle, mémorielle de ce territoire ne soit pas dispersée, avec tout ce qui peut s'ensuivre au niveau du bien-être (ou alors du malaise) social[4].

Parmi les témoignages collectés au cours des enquêtes Tramontana il y en a deux qui ont particulièrement attiré notre attention en ce qu'ils permettent non seulement de préciser cette notion de toponymie narrative, mais également de revenir sur l'analyse toponymique tout court.

3.1.1. La création toponymique

Le premier concerne l'origine des toponymes dits ' populaires ' – ou plus exactement leur acte de baptême – dont normalement les spécialistes prétendent que l'on ignore toujours les auteurs. Or, dans un entretien collecté le 13 janvier 2013 à Arsita (village de la province de Teramo, à 470 m. d'altitude), le témoin GB, plutôt jeune en fait par rapport à la moyenne (il est né en 1945) nous a raconté comment il a jadis baptisé une fontaine à l'entrée du village, la *Fontana degli innamorati* ("Fontaine des amoureux"). Voici un passage de son récit :

---

[4] Agresti et Pallini sous presse. L'analyse du discours ne s'est que récemment, et encore assez marginalement, penchée sur l'analyse toponymique. Pour une liste essentielle des publications en domaine francophone et italophone, cf. nos *Références* en fin d'article.

> lì si andava a lavare i panni [...] anche mia madre andava col cestino su in testa [...] e il bambino in braccio, però lei andava sempre in buon'ora la mattina, per non aspettare che si affolava, capito ? [...] E allora… si andava a fare le abbeverate e lì si scambiavano pure le *parole*. Per esempio mia sorella ha conosciuto un teramano, [...] mio cognato, andando a prendere l'acqua lì, a lavare pure certi panni di casa [...]. Lui è passato con un cavallo, che stava a fare la trebbiatura in montagna a Colle Mesole [...] vede 'sta signorina e hanno cominciato a scambiare parole, non so che avrebbe detto : ‹ dove abiti ›, ‹ come ti chiami ›, fatto sta che è ritornato poi di giorno altre volte eh, non è passato molto tempo che si è fidanzato con mia sorella, per dire… Come pure tante altre : andavano ad attingere l'acqua lì, e si incontravano – a piedi allora, si camminava – andavano ad attingere l'acqua e si fermavano. Passavano i giovanotti, magari, non l'avevano vista mai e lì s'incontravano e parlavano. Tutto qui. Era un incontro dove poter dire, scambiarsi qualche parola. […] Andavano con la conca a prendere l'acqua [...] beh, tante magari facevano il tragitto dietro la curva lassù, lo buttava l'acqua e ritornavano un'altra volta, un altro viaggio [*rires*]… per incontrare un'altra volta il giovanotto… […] e io ho dato il nome degli innamorati, la ‹ fontana degli innamorati › […] beh perlomeno io l'ho battezzata così, mo, gli altri non lo so, io l'ho trasmesso, se gli altri lo dicono pure non lo so.

Dans ce passage on remarquera la valeur et fonction sociale de la parole qui établit un rapport presque synonymique entre la relation amoureuse, la donne topologique et l'échange langagier (« lì si scambiavano pure le parole »), analogie fréquente dans la culture populaire où souvent parler à deux équivaut à faire l'amour. Ce qui nous intéresse le plus est cependant ici l'acte de baptême du lieu par un sujet connu, historique, qui est tout à fait conscient des raisons qui l'ont poussé à baptiser la fontaine ainsi que de son rôle de relais de connaissance. Il s'observe en effet transmettre (« io l'ho trasmesso ») ce toponyme qui est à la fois subjectif et ancré dans le maillage social, dans la culture locale et traditionnelle pour peu que l'on considère tout un répertoire culturel, littéraire célébrant les mille fontaines des amoureux de notre héritage (au moins) euro-méditerranéen.

3.1.2. L'interprétation narrative toponymique

Le second témoignage ayant trait à la toponymie narrative concerne l'interprétation re-motivante de toponymes à l'origine obscure, comme dans le cas de *Nerito*, village situé à 835 m. d'altitude à l'ouest de Fano Adriano (dans la province de Teramo). Nous avons enregistré une interprétation invraisemblable au point de vue étymologique mais, pour cette raison même, intéressante parce qu'illustrant un volet fondamental de l'histoire locale : la culture de la transhumance. Notre témoin a été BR, un ancien berger qui, né dans une cabane (!) à Sacrofano, au nord de Rome, en janvier 1926, a pendant toute sa vie active parcouru à pied les sept étapes qui liaient traditionnellement Fano Adriano à la Capitale (ce qui fait qu'aujourd'hui les gens de Fano ont un accent *romanesco*). D'après BR l'origine du toponyme *Nerito* serait ramenable à la culture pastorale en ce qu'il cristalliserait une expression des bergers lorsqu'il leur arrivait de se disputer à cause de la répartition des pâturages. Chaque village avait en effet un droit de pâturage, qui correspondait à une parcelle de territoire délimitée par des enceintes, par des *reti* ("grillages"). Ainsi, à force de répéter qu'il ne fallait pas violer ces parcelles, la désignation du lieu aurait abouti à *Nerito* (« *non* si passano

le *reti* », “on ne dépasse pas les grillages” > *non-reti* > [nëÈritë]). Cette étymologie est de toute évidence totalement fantaisiste. *Nerito* n'a pas d'étymologie sûre : chez Giammarco (1990, 265) elle résulterait d'une base prélatine hydronymique *NER + suffixe phytonymique -IT-. Mais enfin : notre témoin avait déjà réfléchi à l'origine du toponyme en question et était déjà parvenu à une réponse, découlant de son vécu et de sa propre expérience du territoire. La culture de la transhumance lui avait permis d'interpréter, donc de mettre en narration, un toponyme qui, justement parce qu'étymologiquement obscur, se rendait disponible pour de nouvelles lectures.

3.2. Jargons de métier et langue locale

Le troisième cas de figure concerne la nature du rapport liant le jargon de métier à la langue locale dont il est pétri. Il est question ici du jargon des cardeurs de Pietracamela, appelé ‘ trignano ’ par ses habitants. Dans ce village de la province de Teramo perché à 1005 mètres dans le massif du Gran Sasso d'Italia, jusqu'à la moitié des années 30 du XX^{e} siècle l'économie locale reposait sur l'élevage mais aussi sur le cardage de la laine, qui était pratiqué exclusivement par les hommes. Les cardeurs travaillaient et se déplaçaient toujours à deux, quittaient Pietracamela en automne et rentraient au printemps après avoir parcouru parfois des centaines de kilomètres vers le Nord de la Péninsule (Toscane et Marches surtout, mais aussi Modène, Bologne, l'Ombrie et la Romagne). Au fil du temps et au cours de ces longs périples, les cardeurs avaient développé un jargon leur permettant de crypter leurs conversations lorsqu'ils se trouvaient chez leurs clients qui les hébergeaient parfois pendant quelques jours. Voici juste quelques exemples de ce jargon :

[lamaˈnɛfra] “ le chat”
[laˈtʃikkəfəˈmənda] “la table appareillée”
[lapəˈʃallə] “la femme, la patronne”
[lamarˈkuttʃa] “la bouche”

Or, si ce jargon est assez bien documenté, notamment par le dialectologue abruzzais Giammarco (1964, 228-233), nous avons pu constater pendant nos entretiens Tramontana que, cinquante ans après, ce patrimoine lexical (Giammarco avait collecté en 1964 98 entrées) s'est certes pulvérisé mais, finalement, se maintient encore et même de manière plutôt surprenante. En effet, parmi les sept témoins que nous avons interviewé il n'y avait qu'un ancien cardeur (BF, né en 1922), sans doute le dernier[5], et, pourtant, tous nos interlocuteurs – y compris les femmes – ont montré une certaine connaissance du *trignano*. À partir de ce constat, nous avons cherché non seulement à élargir le répertoire de Giammarco, mais également à comprendre les raisons de cette

[5] Avec RI, plus jeune, qui reprit ce métier en des temps assez récents mais qui n'habite plus Pietracamela.

maintenance : les cardeurs, une fois rentrés au printemps, se plaisaient à raconter au village leurs expériences vécues à travers l'Italie centre-septentrionale, ce qui propageait aisément leur parole dans le maillage social de Pietracamela – qui, par ailleurs, avait déjà la conscience d'être une sorte d'îlot linguistique ou culturel à cause d'une variété dialectale bien singulière par rapport aux villages voisins[6].

En résumant, le jargon *trignano*, deuf fois cryptique à l'extérieur de la communauté (parce que jargon de métier et jargon construit à partir de l'‘ étrange ’ dialecte local), était tout à fait diffus ou à tout le moins familier en son sein. Ce simple fait a vraisemblablement contribué d'une part à accentuer la singularisation linguistico-culturelle de Pietracamela, et de l'autre à nuancer le statut du jargon des cardeurs en tant que code lexicalement distinct de la ‘ langue commune ’. En effet, à notre avis une importante continuité s'est établie entre ces deux codes, notamment en raison des références à la culture locale présente dans le *trignano*. Ainsi, l'expression

[kattəˈlǝɲəsanjuˈwɔnnə]

littéralement « que dit San Giovanni? », qui n'est pas attestée chez Giammarco et que nous avons collectée chez deux de nos témoins, signifie en fait « quelle heure est-il? », car à Pietracamela l'horloge du village était situé – et l'est encore – derrière l'église de San Giovanni. L'élaboration de la stratégie cryptique propre aux jargons de métier puise dans un univers local marqué par une connivence sociale et des références partagées. Ce jargon est donc une véritable langue identitaire, non seulement des cardeurs, mais plus largement des habitants de Pietracamela. Caractère identitaire qui s'est maintenu jusqu'à nos jours et qui pourrait par là être l'un des leviers de projets de développement local par sa capacité de rassembler les gens, de moins en moins nombreux hélas, soucieux de faire vivre leur village et leur culture malgré le redoutable processus de dépeuplement qui n'a pas l'air de ralentir.

C'est dans cette perspective que tâche de s'engager la LDS. C'est également l'horizon du projet Tramontana.

Université de Teramo — Giovanni AGRESTI

Université de Paris-Sorbonne, EA 4080
(Linguistique et lexicographie latines et romanes) — Fabrice BERNISSAN

[6] Nous sommes en train d'étudier la langue de Pietracamela, recherche qui fera l'objet prochainement d'une publication spécifique.

Références bibliographiques

Agresti, Giovanni, en préparation. *Actualité des racines. Pour une linguistique du développement social, Rom*e, Aracne. Sortie prévue : printemps 2013.

Agresti, Giovanni, 2012. *Toponymes en discours. Trois recherches en Méditerranée*, Rome, Aracne.

Agresti, Giovanni / Pallini, Silvia, en préparation. « Vers une toponymie narrative : récits autobiographiques et ancrages géographiques dans deux villages de la Haute Vallée du Vomano (Italie) ». *Colloque international Défis de la toponymie synchronique. Structures, contextes et usages (Rennes, 22-23 mars 2012)*. Actes en préparation par les soins de Bethina Schnabel-Le Corre et Jonas Löfström.

Bernissan, Fabrice, 2012. « Le comput des locuteurs de l'occitan », *Revue de linguistique romane* 76, 467-512.

Bernissan, Fabrice, 2013. *Toponymie gasconne entre Adour et Arros. Contribution à la lexicographie, à l'ethnologie et à la philologie occitanes*. Thèse de doctorat présentée à l'Université Paris 4 – La Sorbonne, Sarrebruck, Presses académiques francophones.

Betemps, Alexis, 2002. « Toponymie rurale et mémoire narrative (Vallée d'Aoste) », *Rives nord-méditerranéennes* 11, 15-31.

Bouvier, Jean-Claude (ed.), 1997. « Nommer l'espace », *Le Monde alpin et rhodanien* 2-4.

Bouvier, Jean-Claude / Guillon, Jean-Marie (ed.), 2001. *La toponymie urbaine : significations et enjeux*, Paris, L'Harmattan.

Boyer, Henri / Paveau, Marie-Anne (ed.), 2008. « Toponymes. Instruments et enjeux », *Mots, Les langages du politiques* 86.

Bruno, Giuliana, 2002. *Atlante delle emozioni*, Milano, Bruno Mondadori.

Céfaï, Daniel (ed.), 2003. *L'enquête de terrain*, Paris, La Découverte Mauss.

Dalbera, Jean-Philippe, 2004. « Du toponyme à la toponymie », in Ranucci, Jean-Claude / Dalbera, Jean-Philippe (ed.), « Toponymie de l'espace alpin : regards croisés », Corpus, *Les Cahiers* 2, 5-20.

Fabre, Paul, 1997. « Ce que la toponymie peut apporter à la... toponymie », in : Bouvier, Jean-Claude (ed.), « Nommer l'espace », *Le Monde alpin et rhodanien* 2-4, 13-20.

Genre, Arturo, 1986. «I nomi, i luoghi e la memoria», *Quaderni della Valle Stura* 4, Demonte, Comunità Montana Valle Stura, 3-10.

Giammarco, Ernesto, 1964. «I gerghi di mestiere in Abruzzo», *Abruzzo* II, 2, 219-239.

Giammarco, Ernesto, 1990. *Toponomastica Abruzzese e Molisana*, Roma, Edizioni dell'Ateneo.

Glessgen, Martin, 2012 [2007]. *Linguistique romane. Domaines et méthodes en linguistique française et romane*, Paris, Armand Colin.

Kristol, Max, 2002. « Motivation et remotivation des noms de lieux : réflexions sur la nature linguistique du nom propre », *Rives nord-méditerranéennes* 11, 105-120.

Lafont, Robert, 2007 [1994]. *Il y a quelqu'un. La Parole et le Corps*, Limoges, Lambert-Lucas.

Lecolle, Michelle / Paveau, Marie-Anne / Reboul-Touré, Sandrine (ed.), 2009. « Le nom propre en discours », *Les Carnets du CEDISCOR* 11. <http://cediscor.revues.org/729>.

Marrapodi, Giorgio, 2007. « Tassonomia dei sistemi toponimici popolari: individualità del TN e ricorsività lessicale », in : Finco, Franco (ed.), *Atti del secondo convegno di toponomastica friulana*, Udine, Società Filologica Friulana, 259-278.

Mohia, Nadia, 2008. *L'expérience de terrain*, Paris, La Découverte.

Nora, Pierre (ed.), 1984-1992. *Les lieux de mémoire*, Paris, Gallimard, tome 2, vol. 3, 283-315.

Pelen, Jean-Noël (ed.), 2002. *Rives méditerranéennes* 11 (« Récit et toponymie »).

Persi, Peris (ed), 2010. *Emotion & Territories / Emotional geographies. Actes du V^e Colloque international des biens culturels, Fano (Pesaro-Urbino), 4, 5 et 6 septembre 2009*, Dipartimento di Psicologia e del Territorio, Università degli Studi di Urbino « Carlo Bo », Istituto di geografia.

Ranucci, Jean-Claude, 2004. « Micro-toponymie des Alpes-Maritimes : strates motivationnelles », in : Ranucci, Jean-Claude / Dalbera, Jean-Philippe (ed.). « Toponymie de l'espace alpin : regards croisés », Corpus, *Les Cahiers* 2, 203-224.

Rivoira, Matteo, 2009. « L'Atlante Toponomastico del Piemonte Montano (ATPM) : principes, méthodes et résultats », *Géolinguistique* 11, 29-49.

Sanchez, Barbara, 2002. « Récits de la rue et de la ville: Aix-en-Provence », *Rives nord-méditerranéenne*s 11, 91-103.

Vineis, Edoardo (ed.), 1981. *La toponomastica come fonte di conoscenza storica e linguistica*, Pisa, Giardini.

Zamboni, Alberto, 1994. «I nomi di luogo», in : Serianni, Luca / Trifone, Pietro (ed.), *Storia della lingua italiana, II. Scritto e parlato*, Torino, Einaudi, 859-878.

Les étapes dans le travail rédactionnel du DÉRom (*Dictionnaire Étymologique Roman*)[1]

Introduction

Mon article a trois objectifs stratégiques : dans l'immédiat, faire découvrir la méthode de travail du DÉRom aux romanistes intéressés. Dans le *Livre Bleu* du projet, document qui réunit les règles internes de fonctionnement du projet, on présente un tableau synthétique des « Étapes de rédaction » ; les pages qui suivent représentent une introduction à ce document essentiel pour tout rédacteur du DÉRom.

À moyen terme : former la base d'un module d'enseignement pour l'École d'été franco-allemande en étymologie romane qui aura lieu en juillet 2014 à l'ATILF et que nous espérons toujours aussi réussie que la première édition qui s'est déroulée en 2010 à Nancy[2].

Et, dans une perspective à long terme, donner une idée aux futures générations de lexicologues et d'étymologistes de la manière de travailler au début du XXI^{e} siècle.

Rendre claires les étapes de rédaction est essentiel dans le contexte d'une entreprise lexicologique qui mobilise une importante communauté de romanistes : plus d'une cinquantaine de linguistes originaires de douze pays européens et qui représentent tous les domaines linguistiques de la Romania. Devant la diversité des collaborateurs, une préoccupation constante du projet est de donner des règles de rédaction qui assurent une méthodologie de travail harmonieuse et une cohérence interne du dictionnaire.

La rédaction d'un article du DÉRom est un processus complètement informatisé qui commence par : (1) des travaux préparatoires ; (2) continue par la rédaction de la bibliographie ; (3) la rédaction des matériaux ; (4) la rédaction du commentaire ; (5) la révision par domaines géographiques ; (6) la révision générale ; (7) la révision finale.[3]

En ce qui suit, je vais détailler chacune de ces étapes en les illustrant par des exemples concrets de la pratique du DÉRom.

---

1 Je remercie Madame Béatrice Stumpf, ATILF CNRS, d'avoir eu la gentillesse de relire notre article.

2 École d'été franco-allemande en étymologie romane (ATILF, Nancy, 26-30 juillet 2010).

3 Dans ce sens, cet article est complémentaire à Delorme 2011.

1. Travaux préparatoires

1.1. Article modèle : */'kad-e-/

Avant de commencer le travail de rédaction d'un article du DÉRom, le rédacteur doit imprimer l'article-modèle */'kad-e-/, qui constitue l'article échantillon publié au début du DÉRom en 2008 par Eva Buchi. Les rédacteurs doivent lire cet article avec attention et ils doivent le garder à côté d'eux comme un modèle pendant tout le processus rédactionnel.

En lisant l'article modèle, le rédacteur peut déjà s'apercevoir de la structure d'un article du DÉRom : il apparaît en premier l'étymon protoroman, sa classe grammaticale et la définition du signifié de l'étymon protoroman reconstruit :

*/'kad-e-/ v.intr. « être entraîné à terre en perdant son équilibre ou son assiette » (Buchi 2008–2013 *in* DÉRom s.v. */'kad-e-/)

Vient ensuite le corps de l'article qui peut comporter une ou plusieurs subdivisons contenant chacune les issues romanes. Par exemple, l'article */'kad-e-/ comporte deux subdivisons : I. Flexion en */'-e-/ et II. Flexion en */-'e-/.

Par la suite, nous avons dans l'ordre les divisions : « Commentaire » du rédacteur, « Bibliographie », « Signatures » et « Date de mise en ligne de cet article ».

1.2. Choix d'un article dans la nomenclature du DÉRom

Après avoir lu l'article */'kad-e-/, les rédacteurs sélectionnent un article dans la nomenclature DÉRom.

L'objectif du DÉRom est de reconstruire le lexique héréditaire de l'ancêtre des langues romanes, le protoroman, raison pour laquelle la nomenclature du DÉRom est constituée à partir de la liste de mots panromaniques proposée en 1962 par Fischer,ILR2 et sur la base du REW. C'est pourquoi dans la nomenclature du DÉRom on peut lire dans la première colonne le numéro et le lemme du REW qui représente un étymon latin, tandis que dans la deuxième colonne figurent les lemmes que le DÉRom propose pour les entrées correspondantes du REW, comme nous pouvons l'observer dans l'extrait suivant :

Lemme REW	Lemme(s) DÉRom	Rédacteur(s)	État
4. *abante*	*/a'bante/	Michela Russo	D
92. *acer*, 2. **acrus*	*/'akr-u/	Christoph Groß	
98. *acētum*	*/a'ket-u/[1]	Jérémie Delorme	RD
ø	*/a'ket-u/[2]	Jérémie Delorme	RD

Lemme REW	Lemme(s) DÉRom	Rédacteur(s)	État
136. *ad*	*/a/	Michela Russo	D
172. *adjūtāre*	*/ad'iut-a-/	Victor Celac	D
240. *aer*, *-re*	*/a'ere/	Xavier Gouvert	RG
276. *ager*	*/'agr-u/	Julia Alletsgruber	

Dès cette première étape, le rédacteur peut se familiariser avec quelques principes de base du DÉRom concernant la représentation écrite des lemmes :

- ils sont présentés en notation phonologique, afin de signaler leur caractère reconstruit à partir d'unités lexicales orales
- ils comportent un astérisque pour dire que les étymons proposés ne sont pas attestés mais qu'ils ont été trouvés par la méthode de la grammaire comparée-reconstruction
- ils comportent un accent s'ils constituent des unités qui peuvent porter l'accent
- des traits d'union séparent les différents morphèmes lexicaux et grammaticaux

Il faudra tenir compte que dans ce tableau les lemmes ne sont donnés qu'à titre provisoire et qu'ils pourront être réajustés en fonctions des données romanes et de leur analyse.

1.3. L'éditeur de texte

L'éditeur XML choisi pour saisir les articles du projet DÉRom est oXygen. Les directeurs du projet fournissent une licence pour tout rédacteur du projet désirant saisir lui-même ses articles.

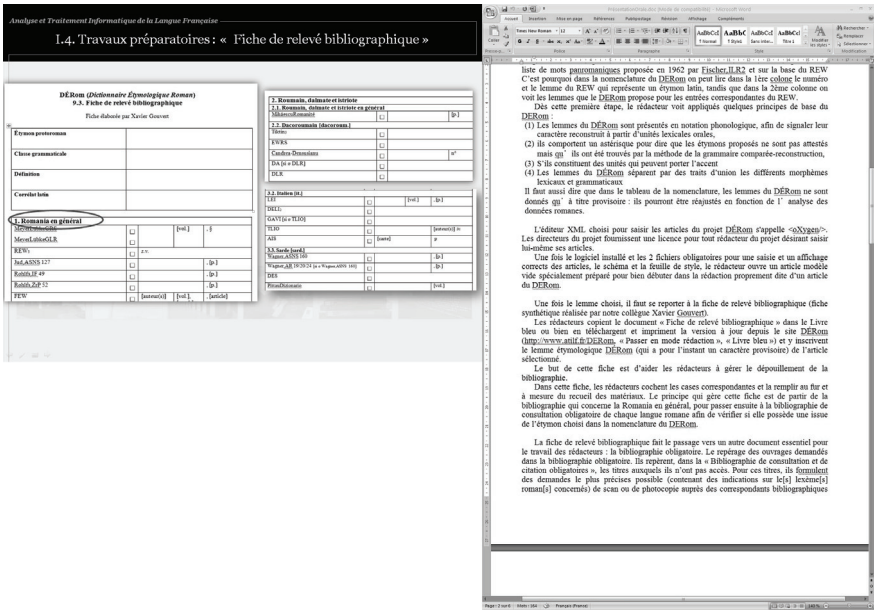
Quand le logiciel et les fichiers obligatoires pour une saisie et un affichage corrects des articles sont installés conformément aux indications techniques du chargé informatique du projet, Gilles Souvay, le rédacteur peut ouvrir un article modèle vide spécialement préparé pour bien débuter dans la rédaction proprement dite d'un article du DÉRom.

1.4. La fiche de relevé bibliographique

Une fois le lemme choisi, le rédacteur doit se reporter à la fiche de relevé bibliographique, fiche synthétique qui recense toutes les sources bibliographiques de consultation obligatoire et qui sont organisées dans un ordre préalablement établi qui respecte strictement l'organisation d'un article du DÉRom.

Les rédacteurs copient le document « Fiche de relevé bibliographique » dans le Livre bleu ou bien en téléchargent et impriment la version à jour depuis le site DÉRom (http://www.atilf.fr/DERom, « Passer en mode rédaction », « Livre bleu ») et y inscrivent le lemme étymologique DÉRom de l'article sélectionné, lemme qui a pour l'instant un caractère provisoire.

Le but de cette fiche dont nous avons ci-dessous un aperçu est d'aider les rédacteurs à gérer le dépouillement de la bibliographie.



Dans cette fiche, les rédacteurs cochent les cases correspondantes et la remplissent au fur et à mesure qu'ils recueillent les matériaux. Le principe qui gère ce document est de partir de la bibliographie qui concerne la Romania en général, pour passer ensuite à la bibliographie de consultation obligatoire de chaque langue romane afin de vérifier si elle possède une issue de l'étymon choisi dans la nomenclature du DÉRom.

1.5. La bibliographie obligatoire

La fiche de relevé bibliographique fait le passage vers un autre document essentiel pour le travail des rédacteurs : la bibliographie obligatoire et le repérage des ouvrages demandés dans la bibliographie obligatoire.

Les rédacteurs repèrent dans la « Bibliographie de consultation et de citation obligatoires » les titres auxquels ils n'ont pas accès. Pour ces sources, ils formulent des demandes précises contenant des indications sur le lexème roman concerné de scan ou de photocopie auprès des correspondants bibliographiques cités dans la troisième colonne du document (« Scan ou photocopie disponible auprès de […] »). Comme dans cet exemple, Pascale Baudinot pour le PatRom :

PatRom	Cano González (Ana María)/Germain (Jean)/Kremer (Dieter) (éd.), 2004–. *Dictionnaire historique de l'anthroponymie romane. Patronymica Romanica* (PatRom), Tübingen, Niemeyer.	Pascale Baudinot [AURICULA, BRACCHIUM, CAPUT, CAUDA, COLLUM, CORNŪ, CORPUS, CŪLUS, DĒNS, DIGITUS, FRŌNS, HOMŌ, LINGUA, MANUS, NĀSUS, OCULUS, PANTEX, POLLEX, TESTA, VENTER] [À citer par auteur(s), volume/tome, colonne(s) et article]

Les correspondants bibliographiques dont les adresses mail figurent sur la liste des membres du projet répondent très rapidement aux demandes de scan ou de photocopie des rédacteurs.

2. La rédaction de la bibliographie : Romania en général

Au fur et à mesure que les documents bibliographiques sont identifiés, ils doivent être imprimés : ces matériaux constitueront ensuite le dossier de l'article rédigé, dossier qui deviendra avec le temps lui-même un outil de travail et une référence à laquelle le rédacteur reviendra chaque fois qu'un problème surviendra au cours de la rédaction et de la révision de son article.

Par exemple, pour l'article */ti'tion-e/, la fiche de relevé bibliographique pour la Romania en général a été complétée de la manière suivante :

1. Romania en général				
MeyerLübkeGRS MeyerLübkeGLR	✓		[vol.] 1	, § 118-119, 135, 306-307, 350, 404-405, 450, 454, 509
REW_3		*s.v. tītio, -ōne*		
Jud,ASNS 127	Ø			, [p.]
Rohlfs,IF 49	Ø			, [p.]
Rohlfs,ZrP 52	Ø			, [p.]
FEW		[auteur(s)] Müller [année] 1966 *in* FEW	[vol.], 13/1, [p.col.] 356a-359b	, [article] tītio
LausbergSprachwissenschaft LausbergLingüística LausbergLinguistica LausbergLinguística	□ ✓ □ □		[vol.] 1	, § 179-182, 231-235, 253, 273, 304, 405, 452-455

Une fois les sources réunies, les rédacteurs s'approprient le contenu des textes ainsi réunis par la lecture, le surlignage, éventuellement par la discussion avec un

collègue. Dans cette phase initiale, il s'agit de se forger une vue d'ensemble et surtout pas de résoudre des questions de détail.

Une fois la bibliographie générale constituée, les rédacteurs rédigent la section « Bibliographie » de l'article dans le fichier XML déjà créé selon les indications fournies dans la partie consacrée aux travaux préparatoires et en prenant modèle sur l'article */'kad-e-/.

Ce qui peut surprendre, c'est que dans le document XML prédéfini pour la saisie des articles, la bibliographie ne correspond pas au début de l'article, mais elle est placée après les matériaux des issues romanes et après le commentaire. Pourtant, il est important que cette étape soit accomplie en premier car elle permet d'avoir une vision générale sur les phénomènes linguistiques qu'on va ensuite retrouver dans les différents langues romanes.

3. La rédaction des matériaux

Par la suite, les rédacteurs dépouillent les sections pertinentes des titres cités sous « 2.1. Généralités » et « 2.2. Dacoroumain » de la « Bibliographie de consultation et de citation obligatoires ». Au fur et à mesure de l'avancement des travaux, ils remplissent la « Fiche de relevé bibliographique ». Ainsi, pour l'article */ti'tion-e/, la fiche de relevé bibliographique pour le Dacoroumain a été complétée de la manière suivante :

2.2. Dacoroumain [dacoroum.]				
Tiktin$_3$				
EWRS				
Candrea-Densusianu	Ø			n°
DA [si ø DLR]	□			
DLR				
Graur,BL 5	Ø			, [p.]
Cioranescu Ciorănescu	□			n° 8443
Frăţilă,MedRom 19	Ø			, [p.]
MDA				
DELR	□			
ALR SN		[carte] 1214		p 182, 250, 346, 520, 848

Quand la « Fiche de relevé bibliographique » a été complétée et les références retrouvées photocopiées, le rédacteur dispose du dossier nécessaire pour rédiger la section « Matériaux » de l'article. Les matériaux sont à saisir sous les cognats corres-

pondants qui se présentent toujours dans le même ordre, conformément aux indications fournies sous le point (2.) du document « Normes rédactionnelles » du *Livre bleu* du DÉRom et en prenant modèle sur l'article */'kad-e-/.

Les rédacteurs commencent par rédiger les matériaux pour le roumain et ses dialectes, ensuite ils procèdent de même pour les cognats istroroumain, méglénoroumain, aroumain, dalmate et istriote, par la suite, pour l'italien, sarde, frioulan, ladin et romanche, ensuite pour le français, francoprovençal, occitan et gascon, et finalement pour le catalan, espagnol, asturien, galicien et portugais.

Ainsi, pour l'article */ti'tion-e/ la saisie des matériaux pour le dacoroumain à partir de la fiche de relevée bibliographique et selon la documentation trouvée a donné le résultat suivant :

*/ti'tion-e/ > dacoroum. *tăciune* s.m. « incandescent, tison » (dp. 1620, Tiktin3 ; EWRS ; Cioranescu n° 8443 ; DLR ; MDA ; SalaPhonétique 166, 225 ; ALR SN 1214 p 182, 250, 346, 520, 848). (Jactel/Buchi 2012–2013 *in* DÉRom s.v. */ti'tion-e/)

Si une source consultée comporte la référence d'une publication intéressante pour l'article qu'ils sont en train de traiter au niveau plus ou moins panroman, les rédacteurs doivent se la procurer et la mettre de côté pour la rédaction du commentaire.

Le processus est terminé quand la « Fiche de relevé bibliographique » est entièrement remplie et la partie « Matériaux » de l'article est rédigée.

Si les rédacteurs ont besoin de citer une source non encore référencée par la bibliographie générale il faut proposer un sigle. Ainsi, pour créer le sigle d'une monographie par exemple, il faut juxtaposer le nom de l'auteur (ou du premier auteur s'il y en a plusieurs) et un élément (de préférence le mot sémantiquement central) du titre. Par exemple :

AppelChrestomathie = Appel (Carl), 1930[6] [1895[1]]. *Provenzalische Chrestomathie mit Abriss der Formenlehre und Glossar*, Leipzig, Reisland.

Les rédacteurs envoient à Pascale Baudinot leur proposition de sigle, ainsi que les références précises des sigles créés en respectant les indications données dans le document « DÉRom Règles de siglaison ». Elle intègre le nouveau sigle dans la bibliographie DÉRom en ligne, ce qui va permettre à la fois de l'identifier lors du contrôle en ligne de l'article rédigé, mais aussi de la retrouver et de l'afficher lors d'une recherche sur la bibliographie utilisée dans le DÉRom.

4. La rédaction du commentaire

Une fois les matériaux rédigés, les rédacteurs se penchent sur le « Commentaire » de l'article en prenant comme modèle l'article */'kad-e-/.

Même si le « Commentaire » est la partie la plus libre dans la rédaction d'un article du DÉRom, les rédacteurs sont tenus à respecter trois divisions constitutives qui se retrouvent dans tous les articles et qui se matérialisent sous forme de paragraphes.

Le premier paragraphe est toujours consacré aux branches romanes qui présentent des cognats conduisant à reconstruire l'étymon protoroman, comme c'est le cas ici pour */ti'tion-e/ :

> À l'exception du dalmate, branches romanes présentent des cognats conduisant à reconstruire protorom. */ti'tion-e/ s.m. « morceau de bois incandescent, tison ; maladie des céréales d'origine cryptogamique qui les convertit en poussière noirâtre, charbon ». (Jactel/Buchi 2012–2013 *in* DÉRom s.v. */ti'tion-e/)

Le deuxième paragraphe est consacré à des courtes explications concernant les subdivisions de l'article en fonction des sens qu'ils présentent, comme c'est le cas pour l'article */ti'tion-e/ qui présente deux subdivisions correspondant aux deux sens 'tison' et 'maladie des céréales', ou en fonction des changements morphologiques comme le changement des classes flexionnelles du genre pour les substantifs, comme c'est le cas de l'article */'pɔnt-e/, ou de la déclinaison pour les verbes */'kad-e-/, par exemple.

Le troisième paragraphe est toujours consacré au corrélat du latin écrit : le rédacteur doit obligatoirement le donner s'il existe ; s'il n'existe pas, il est explicitement formulé en mentionnant toujours depuis quand il est attesté en latin écrit, comme nous pouvons le voir pour le sens II. [« charbon (maladie des céréales »)] de l'article */ti'tion-e/ :

Le corrélat du latin écrit de I., *titio*, *-onis* s.m. « tison », est connu depuis Varron (* 116 – † 27, OLD ; « mot populaire d'après Lactance «, Ernout/Meillet4)[10]. Le latin écrit de l'Antiquité ne connaît pas, en revanche, de corrélat de II. (Jactel/Buchi 2012–2013 *in* DÉRom s.v. */ti'tion-e/)

5. La phase de contrôle et de révision

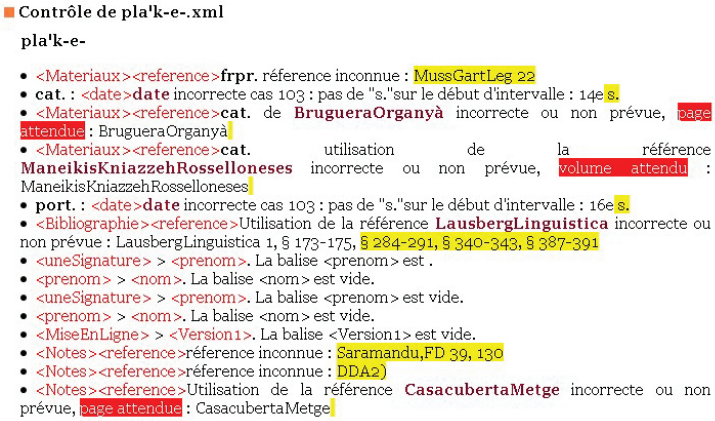
5.1. Le contrôle en ligne

La phase de rédaction proprement dite accomplie, l'article doit être soumis au contrôle en ligne : http://www.atilf.fr/DERom, en activant « Passer en mode rédaction », ensuite « Articles XML », en choisissant la fonction « Contrôler un fichier XML ». Cette étape permet aux rédacteurs de corriger les éventuelles erreurs détectées par le système informatique. Ils procèdent ensuite à la visualisation en ligne de l'article à partir de l'URL :

http://www.atilf.fr/DERom, « Passer en mode rédaction », « Articles XML », « Visualisation d'un fichier XML »

Par la suite, ils impriment l'article et le relisent attentivement, en portant une attention particulière aux problèmes de présentation matérielle (italiques, espaces parasites, ponctuation à la fin des notes etc.), puis ils corrigent les erreurs éventuellement détectées lors de cette relecture. Ensuite, ils procèdent de nouveau à la visualisation en ligne de leur article.

Pour donner un exemple, je procède au contrôle en ligne d'un des articles que je suis en train de rédiger et qui n'est donc pas encore en ligne, plus précisément l'article */pla'k-e-/, ce processus va donner le résultat suivant :



Contrôle de pla'k-e-.xml

pla'k-e-

- <Materiaux><reference>**frpr.** référence inconnue : MussGartLeg 22
- **cat.** : <date>**date** incorrecte cas 103 : pas de "s."sur le début d'intervalle : 14e s.
- <Materiaux><reference>**cat.** de **BrugueraOrganyà** incorrecte ou non prévue, page attendue : BrugueraOrganyà
- <Materiaux><reference>**cat.** utilisation de la référence **ManeikisKniazzehRosselloneses** incorrecte ou non prévue, volume attendu : ManeikisKniazzehRosselloneses
- **port.** : <date>**date** incorrecte cas 103 : pas de "s."sur le début d'intervalle : 16e s.
- <Bibliographie><reference>Utilisation de la référence **LausbergLinguistica** incorrecte ou non prévue : LausbergLinguistica 1, § 173-175, § 284-291, § 340-343, § 387-391
- <uneSignature> > <prenom>. La balise <prenom> est .
- <prenom> > <nom>. La balise <nom> est vide.
- <uneSignature> > <prenom>. La balise <prenom> est vide.
- <prenom> > <nom>. La balise <nom> est vide.
- <MiseEnLigne> > <Version1>. La balise <Version1> est vide.
- <Notes><reference>réference inconnue : Saramandu,FD 39, 130
- <Notes><reference>réference inconnue : DDA2)
- <Notes><reference>Utilisation de la référence **CasacubertaMetge** incorrecte ou non prévue, page attendue : CasacubertaMetge

Nous pouvons constater que 14 erreurs ont été faites parmi lesquelles le sigle MussGartLeg 22 pour l'édition de Mussafia et Gartner des légendes des 10 apôtres, par exemple, n'est pas reconnu parce que le rédacteur n'a pas signalé à la responsable de la bibliographie, Pascale Baudinot, la citation dans l'article de cette nouvelle source pour DÉRom. Il faudra donc lui transmettre le sigle que nous proposons pour cette source, en l'occurrence le sigle du DEAF, et son développement détaillé, selon les normes requises pour la rédaction de la bibliographie que nous avons pu voir ci-dessus.

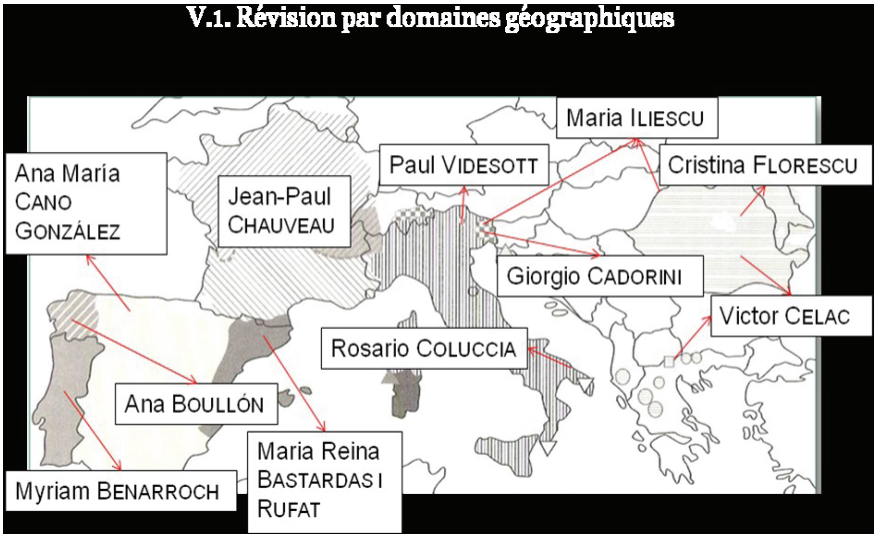
D'autres types d'erreurs peuvent apparaître : des erreurs de balisage, comme par exemple pour la date « 14^{e} s. » il s'agit d'une erreur de balisage au niveau du 's.' de 'siècle' ; ou encore des oublis : pour la référence BrugueraOrganyà, par exemple, le système informatique nous attire l'attention sur le fait que la référence à la page citée doit être précisée lorsqu'on cite cette source.

Ce système de contrôle en ligne du fichier XML permet de corriger et d'harmoniser la saisie d'un article qui sera prêt pour l'étape suivante lorsque plus aucune anomalie n'est détectée par le système informatique.

5.2. Les révisions par domaines géographiques

Quand le rédacteur a fini de corriger les dernières erreurs de balisage, l'article peut être converti en version Word et PDF et envoyé ensuite aux réviseurs par domaine géographiques via une adresse mail collective qui réunit tous les réviseurs de cette étape.

Ainsi, pour l'article */ti'tion-e/, les réviseurs qui ont signé l'article ont été pour *Romania du Sud-Est* : Victor Celac et Cristina Florescu ; pour *Italoromania* : Giorgio Cadorini, Rosario Coluccia et Paul Videsott. Pour la *Galloromania* : Jean-Paul Chauveau ; pour *Ibéroromania* : Maria Reina Bastardas i Rufat, Ana Boullón et Ana María Cano González. Pour avoir une idée plus précise de la complexité et de l'importance de cette étape de rédaction, nous pouvons visualiser la répartition des réviseurs sur la carte de la Romania :



Les réviseurs par domaines géographiques vérifient donc, chacun pour le domaine géographique dont il a la responsabilité, l'exactitude des données et la justesse des analyses proposées et envoient leurs propositions de modifications aux rédacteurs.

Suite aux propositions de modifications formulées par les réviseurs par domaines géographiques, complétées, si nécessaire, par d'autres échanges, le rédacteur élabore une deuxième version de leur article. Ils répètent le processus d'échange avec les réviseurs par domaines géographiques jusqu'à ce que tous les doutes soient levés.

Au fur et à mesure, ils tiennent à jour la rubrique « Signatures » de l'article, en tenant compte de l'apport concret des différents réviseurs à l'article traité, indépendamment de la place qu'occupent ces derniers dans l'organigramme du projet.

Les étapes de révisions des articles constituent un va et vient continu entre les rédacteurs et les réviseurs comme on peut se rendre compte en lisant le tableau de bord dressé dans la publication de Jérémie Delorme (Delorme 2011).

Une fois la révision par domaine géographique accomplie, les rédacteurs envoient la nouvelle version de l'article à Jean-Pierre Chambon et, en version Word (.doc) et en version PDF (.pdf), à Günter Holtus pour la révision générale.

Ils révisent l'article du point de vue de la grammaire comparée-reconstruction et de la synthèse romane et envoient leurs propositions de modifications aux rédacteurs.

Suite aux propositions de modifications formulées par Jean-Pierre Chambon et Günter Holtus, ils élaborent une troisième version de leur article. Ils répètent le processus d'échange avec eux jusqu'à ce que tous les doutes soient levés.

A ce moment, les rédacteurs renvoient la troisième version de leur article, en version Word (.doc) et PDF (.pdf), aux réviseurs par domaines géographiques (à l'adresse collective deromrevision@atilf.fr).

Cette étape de révision suppose un feedback très prompt et actif des deux côtés. Quelque fois, une information qui peut sembler très intéressante à un réviseur d'un certain domaine géographique ne peut être spécifiée que dans une note de l'article puisqu'elle est plutôt d'intérêt idioroman et donc n'est pas essentielle pour la reconstruction de l'étymon protoroman. C'est, par exemple, de la note concernant le masculin *pod* ('pont') du dacoroumain : il s'agit d'une donnée linguistique qui ne concerne plus la phase protoromane de la langue car il s'agit d'un emprunt du slave et concerne donc une phase plus tardive de la langue, il s'agit donc d'un développement idioroman. Le rédacteur n'a pas rejeté ce type d'information puisqu'il l'a donnée dans une note, car il a considéré qu'elle peut être utile au lecteur qui consulte le dictionnaire et qui s'intéresse particulièrement au dacoroumain. Pourtant, le rédacteur a choisi de donner cette information dans une note, car il ne doit pas perdre de vue l'objectif final d'un article du DÉRom qui est la reconstruction protoromane de l'étymon traité et non pas ses développements idioromans :

*/'pɔnt-e/ > dacoroum. *punte* s.f. « passerelle réservée aux piétons » (dp. 1649, DRH B, 34, 124 ; Tiktin3 ; EWRS ; Candrea-Densusianu n° 1474 ; Cioranescu n° 6971 ; DLR ; MDA ; ALRR – M pl. 55 ; NALR – O pl. 47 ; ALRR – MD 539*[1])[2], […]

2. […] – Dans le sens de « pont », cette issue a été évincée par dacoroum. *pod* s.n. (< protosl. **podъ* s.m. « sol », IEEDSlavic ; dp. 1563/1583, DLR ; MDA). (Andronache 2008–2013 *in* DÉRom s.v. */'pɔnt-e/)

5.3. Les révisions par domaines géographiques

Une fois le processus de révision accompli, les rédacteurs envoient leur article, en version <oXygen/> (.xml), à travers l'adresse collective deromdirection@atilf.fr, à Éva Buchi et Wolfgang Schweickard. Les directeurs du DÉRom révisent l'article, notamment du point de vue de la politique scientifique et de la cohérence générale du projet, et envoient leurs propositions de modifications aux rédacteurs. L'échange se répète et, suite aux propositions de modifications formulées par Éva Buchi et Wolfgang Schweickard, ils élaborent une cinquième version de leur article et l'envoient, en version <oXygen/> (.xml), à travers l'adresse collective deromdirection@atilf.fr, à Éva Buchi et Wolfgang Schweickard.

Par la suite, Éva Buchi publie l'article sur le site Internet du DÉRom (http://www.atilf.fr/DERom). L'article peut dorénavant être consulté en libre accès, mais le rédac-

teur peut continuer d'améliorer son article en fonction des nouvelles informations retrouvées, ce qui fait que nous avons le plus souvent deux dates de mise en ligne d'un article : celle de la publication initiale et celle de la dernière modification qui a été apportée. Entre les deux, plusieurs versions de l'article ont pu voir le jour.

6. Conclusion

La rédaction d'un article du DÉRom constitue un processus de longue maturation pendant lequel le rédacteur donne au moins cinq versions de l'article suite au travail sur les matériaux des divers cognats romans et aux échanges avec les réviseurs par domaines géographiques, les réviseurs généraux, à nouveau les réviseurs par domaines géographiques et les réviseurs finaux. La rédaction reste un processus ouvert aux améliorations tant que les articles sont publiés en ligne, car le DÉRom n'est pas seulement un dictionnaire d'une grande ouverture, mais aussi un dictionnaire ouvert.

ATILF CNRS - Université de Lorraine Marta ANDRONACHE

Références bibliographiques

DEAF = Baldinger (Kurt) *et al*., 1974–. *Dictionnaire Étymologique de l'Ancien Français*, /Tübingen / Paris / Niemeyer / Klincksieck.

Delorme, Jérémie, 2011. « Généalogie d'un article étymologique : le cas de l'étymon protoroman * /βi'n-aki-a/ dans le *Dictionnaire Etymologique Roman* », *Bulletin de la Société de linguistique de Paris*, 106/1, 305-341.

DÉRom = Buchi, Éva/Schweickard, Wolfgang (dir.), 2008-. *Dictionnaire Étymologique Roman (DÉRom)*, Nancy, ATILF, site Internet (http://www.atilf.fr/DERom).

Fischer,ILR2 = Fischer, Iancu, 1969. « III. Lexicul. 1. Fondul panromanic », in : Rosetti, Alexandru (éd.), *Istoria limbii române*, Bucarest, Editura Academiei Republicii Socialiste România, volume 2, 110-116.

REW_3 = Meyer-Lübke (Wilhelm), 1930-1935[3] [1911-1920[1]]. *Romanisches Etymologisches Wörterbuch*, Heidelberg, Winter.

Mussafia, Adolf/Gartner, Theodor, 1895. *Altfranzösische Prosalegenden aus der Hs. der Pariser Nationalbibliothek Fr. 818*, I. Teil [seul paru], Wien – Leipzig (Braumüller).

PatRom = Cano González (Ana María)/Germain (Jean)/Kremer (Dieter) (éd.), 2004–. *Dictionnaire historique de l'anthroponymie romane. Patronymica Romanica* (PatRom), Tübingen, Niemeyer.

Le corpus *Per-Fide* et ses applications : une étude de cas sur les expressions idiomatiques

1. Introduction

L'étude dont il est question ici vise précisément à souligner l'impact significatif que la linguistique de corpus peut avoir sur la recherche et l'enseignement dans les domaines touchant à l'étude du langage et à la traduction (Williams, 2005). Nous entendons ici proposer une réflexion sur l'utilité d'un corpus multilingue (tel que le corpus *Per-Fide*) dans le cadre du travail de traduction, sur la base d'un exemple concret, celui des expressions idiomatiques, dans une perspective contrastive français-portugais.

2. Le corpus multilingue Per-Fide : nouvel outil de traduction contextuelle en ligne

Les corpus multilingues parallèles sont des outils précieux pour l'analyse contrastive, mais ils sont bien souvent limités au niveau des langues mises en regard et/ou des types de textes considérés. Le projet *Per-Fide* (Araújo *et al.*, 2010) dont nous voulons rendre compte dans le cadre de cette étude trouve précisément son origine dans la nécessité de construire, avec des textes appartenant à des domaines plus diversifiés (littéraire, technico-scientifique, journalistique, juridique et religieux), des corpus alignés à la phrase, qui mettent en parallèle des textes originaux en portugais dans ses différentes variantes (portugais européen, portugais brésilien, portugais africain) et leurs traductions dans six autres langues (espagnol, russe, français, italien, allemand et anglais), ainsi que des textes originaux dans ces six langues étrangères et leurs traductions vers le portugais.

2.1. Mode de recherche monolingue

Le corpus *Per-Fide* peut être consulté librement à partir d'un concordancier (http://www.per-fide.ilch.uminho.pt/query), qui permet de procéder à différents types de recherche. Dans cette interface, l'utilisateur peut entrer des critères de recherche non seulement pour la langue source mais aussi pour la langue cible ou bien pour les deux à la fois afin de visualiser rapidement l'élément recherché dans différents corpus simultanément :

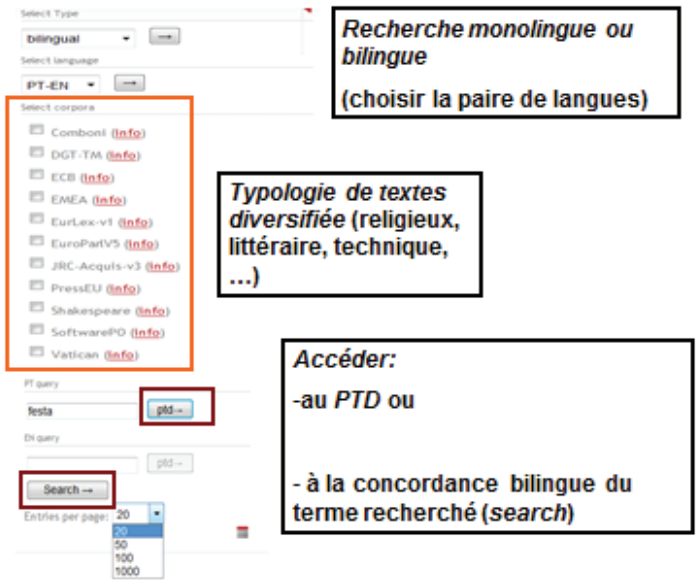


Figure 1 : Interface de recherche du corpus *Per-Fide*

Pour obtenir la concordance monolingue d'un mot donné, il suffit de taper le mot en question. Cette concordance affiche ligne par ligne chaque occurrence d'une forme vedette au sein de son contexte d'origine :

Figure 2 : Concordance monolingue de *chat*

Il semblerait que le mot *chat* soit, dans le corpus *Europarl*, un excellent catalyseur d'expressions idiomatiques : *appeler un chat un chat / acheter chat en poche.*

2.2. Mode de recherche bilingue (simple ou double)

Une recherche bilingue devrait nous permettre de voir si l'homologue portugais de *chat* est aussi producteur d'expressions figées. La figure qui suit montre les résultats d'une requête qui permet de visualiser rapidement l'élément recherché dans différents contextes, mais aussi ses traductions en portugais, qui peuvent différer suivant les cas :

Figure 3 : Concordance bilingue de *chat*

Si l'on compare les inventaires de structures qui correspondent en portugais à l'expression idiomatique *appeler un chat un chat* du français, on constate que la redondance sémantique du français (*un chat un chat*) se perd en portugais et il semblerait que l'on n'ait plus affaire, dans cette langue, à des *chats* mais à des *bœufs* (*bois*) ou à des *choses* (*coisas*) : *chamar os bois pelos nomes / designar as coisas pelos próprios nomes*. En demandant au bi-concordancier de repérer tous les couples de phrases dans lesquels *gato* apparaît du côté portugais et *chat* du côté français, le traducteur ou le linguiste peut déterminer, assez rapidement, les cas de correspondance directe entre ces deux termes :

Figure 4 : Concordance bilingue double de *chat-gato*

Cette recherche bilingue double montre que l'expression portugaise *comprar gato por lebre* est réellement l'équivalent de l'expression *acheter chat en poche* en français.

2.3. Recherche d'expressions multi-mots

Le corpus *Per-Fide* devient un outil encore plus précieux pour ceux qui souhaitent non seulement traduire des mots simples, mais également des mots plus complexes que l'on peut trouver dans différents contextes, ce qui est très souvent le cas dans la pratique. Si l'on tient à «contrer le trop-plein d'information, à confirmer une intuition personnelle ou encore vérifier si les équivalents proposés par les dictionnaires bilingues sont vraiment utilisés» (Langlois, 1996), on peut combiner plusieurs mots dans la syntaxe d'interrogation (simple ou double), et demander que le mot-vedette (ici, *taureau*) soit précédé ou suivi par un mot donné (par exemple *prendre/par les cornes*) :

Figure 5 : Concordance autour d'une expression idiomatique

Ce corpus permet donc de traduire des combinaisons de mots, des tournures idiomatiques et imagées dans leur contexte d'utilisation, en s'appuyant sur une base de données de traductions existantes, régulièrement mise à jour. L'utilisateur peut ainsi comparer les alternatives de traduction qui s'affichent et choisir celle(s) qui convien(nen)t le mieux au contexte de sa traduction. Il est intéressant de voir que l'expression *prendre le taureau par les cornes* (dont le sens propre – *se lancer dans l'arène et faire face aux coups de cornes de l'animal* – est reconnaissable sous le sens figuré – *faire face aux difficultés* – mais son usage en diffère) ne donne pas nécessairement lieu à une traduction littérale puisque l'on trouve, à côté de l'équivalent le plus direct *pegar/agarrar o touro pelos cornos*, d'autres formes de traduction (*tomar as rédeas da situação* / *enfrentar corajosa e rapidamente as dificuldades* / *encarar o problema de frente* / *meter mãos à obra* / *dar o braço a torcer*, …) qui restituent, chacune à leur manière, le sens (à savoir *s'attaquer aux difficultés de front*) de cette expression.

2.4. Mode de recherche par les dictionnaires probabilistes de traduction

À travers l'interface de recherche du corpus *Per-Fide* l'utilisateur peut également activer le dictionnaire probabiliste de traduction (PTD) d'un mot en cliquant (non pas sur *search* mais) sur *ptd* (cf. *supra*, fig. 1). Voyez ci-dessous un exemple de recherche simple (le mot *chat*) par PTD :

Figure 6 : Dictionnaire probabiliste de traduction de *chat*

L'affichage en temps réel d'un PTD lorsque l'utilisateur passe la souris sur n'importe quel mot d'une concordance est un véritable atout de ce corpus. Ce PTD est capable d'identifier (grâce aux méthodes statistiques) les correspondances traductionnelles de l'expression recherchée. Dans le sous-corpus *Europarl*, *gato* est le meilleur candidat à la traduction du mot *chat* mais il est intéressant de voir que, dans le réseau des autres équivalents de ce mot, on trouve également, bien que dans une moindre mesure, les mots *coisas* et *nomes* (qui figurent, on le rappelle, dans les expressions idiomatiques *chamar os bois pelos nomes* / *chamar as coisas pelos próprios nomes* signalées plus haut). Plutôt que d'avoir à parcourir de longues listes de concordances pour examiner les équivalences traductionnelles d'une unité, l'utilisateur peut donc très rapidement activer le PTD d'un terme pour en déterminer, en un seul clic, les correspondants les plus directs. Cette approche du lexique est quantitative, mais ne doit, en aucun cas, négliger le retour à la concordance, qui est la seule à pouvoir offrir

une représentation fidèle et claire du voisinage d'une forme vedette dans une langue et des différentes manières de la rendre dans une autre langue. Pour assurer ce retour à la concordance, il suffit de cliquer sur la flèche qui figure à droite de chaque candidat à la traduction (cf. *supra*, fig. 6) et on a aussitôt accès à la concordance double *chat-gato*, *chat-coisas*, *chat-nomes*, etc., ce qui nous permet donc de visualiser ces termes en contexte et de voir qu'ils s'inscrivent, pour la plupart, dans des groupes de mots reliés par de fortes relations sémantiques et syntaxiques. Pour simplifier notre appréhension, le PTD opère donc le découpage initial du texte en unités, mais il est clair que ces unités ne sont pas seulement des mots isolés mais, également des unités complexes, plus larges, qu'il est nécessaire de retrouver et d'expliciter par le contexte. Nous verrons, notamment avec la méthode des cooccurrences du programme *CooCS*, comment le repérage de ces formes complexes peut se faire automatiquement.

Il faut préciser, par ailleurs, que ces PTDs fonctionnent selon un principe de circularité dans la mesure où ils affichent non seulement les correspondances traductionnelles du terme recherché ,mais également les traductions de chacune de ces correspondances, ce qui permet de revenir à la langue de départ de notre requête initiale :

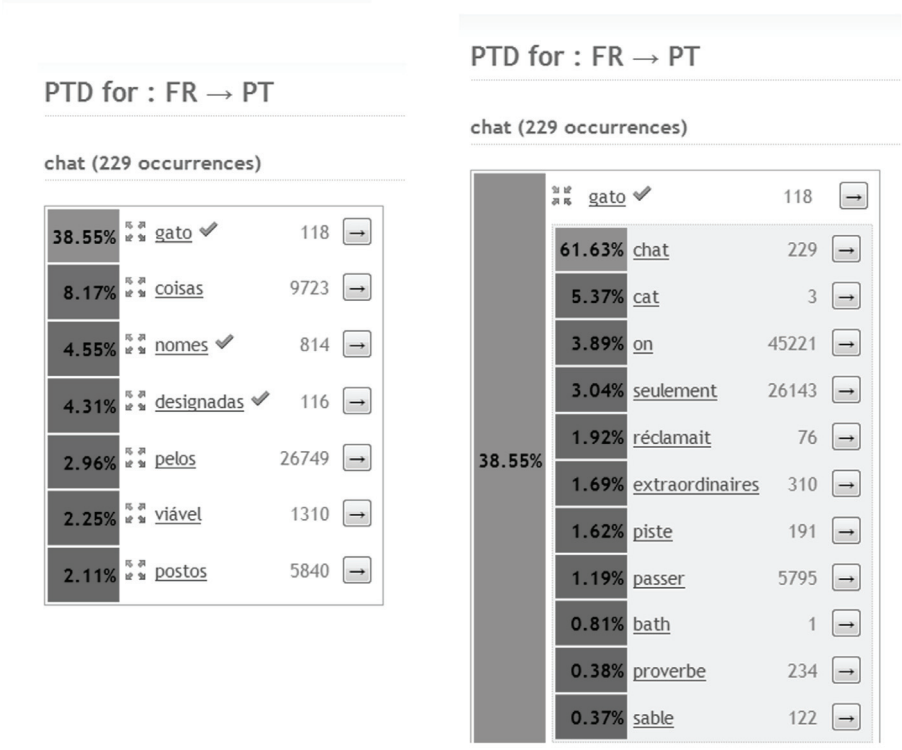


Figure 7 : Circularité des dictionnaires probabilistes de traduction

En effet, en cliquant sur les flèches centrifuges qui se trouvent à gauche de chaque alternative de traduction, l'utilisateur gère le PTD de chacune de ces alternatives, ce qui lui permet, par exemple, de conclure, de manière très simple et rapide, que le pourcentage de correspondance de *chat-gato* est nettement plus élevé du PT au FR que l'inverse (→ 38,55% contre 61,63%). On remarquera, en effet, que le PTD est traversé par un gradient chromatique qui mesure le degré de probabilité d'un terme d'être traduit par un autre. Le rouge pur s'impose lorsque le degré de correspondance entre deux termes tend vers les 0%, le vert pur lorsque ce degré se rapproche, au contraire, des 100% et le jaune pur s'assume comme une couleur intermédiaire qui marque un degré de correspondance de l'ordre des 50%. Il est à noter que plus ce degré de correspondance baisse, plus l'on aura affaire à des stratégies de traduction indirecte. En effet, comme on l'a signalé plus haut, d'autres termes (*coisas, bois*) entrent en concurrence avec le terme *gato* dans la traduction des expressions idiomatiques françaises comportant le nom de cet animal et donc la correspondance *chat-gato* se situe nécessairement dans un pourcentage inférieur à 50 % (d'où la couleur orange des 38,55%).

Il est intéressant de remarquer qu'un même terme peut donner lieu à des PTDs différents en fonction du corpus dans lequel on effectue la requête. En effet, si l'on calcule, par exemple, le PTD de *chat* dans un corpus plus technique lié à l'informatique, on tombe alors (non plus sur le terme *chat* mais) sur celui de *conversa* qui nous fait entrer, de plein fouet, dans le monde des messages SMS, des *chats*, des courriers électroniques très prisés par les jeunes.

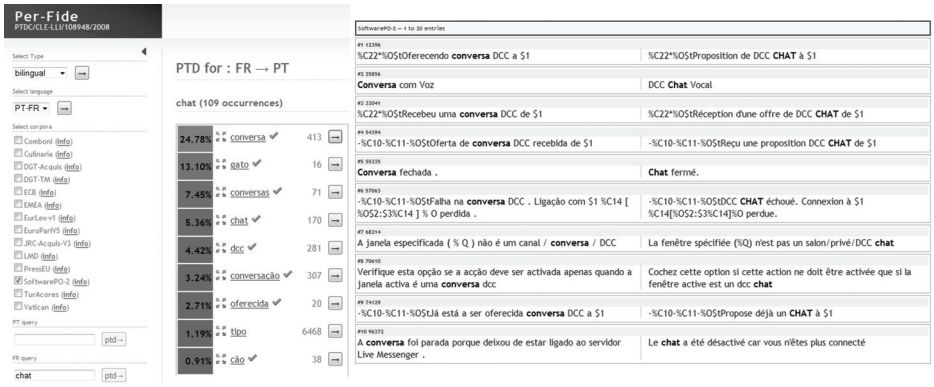


Figure 8 : Technicité des corpus (chat = conversa ≠ gato)

C'est justement parce que nous avons conscience de l'impact direct du degré de technicité des corpus sur les résultats obtenus que nous avons cherché à diversifier au maximum la typologie de textes qui intègre le corpus *Per-Fide*.

3. Apports de la textométrie pour l'analyse de corpus

Outre la compilation de corpus, le projet *Per-Fide* vise également la création de ressources (e.g. PTDs et mémoires de traduction) qui peuvent être gratuitement téléchargeables sur http://per-fide.ilch.uminho.pt/site.pl/resources.pt.

Les mémoires de traduction, en particulier, peuvent être combinées avec des méthodes lexicométriques pour faciliter le repérage des attractions ou répulsions entre des formes et leur analyse ultérieure en tant qu('anti)cooccurrences par le biais d'outils de mesure adaptés comme, par exemple, les logiciels *Lexico3* et *CooCS*[1] réalisés dans le cadre du laboratoire SYLED de l'Université Paris 3 – Sorbonne Nouvelle.

Pour cette analyse, nous continuons à travailler sur la mémoire de traduction Europarl. Après une première étape de dépouillement léxicométrique avec *Lexico3*, on obtient un inventaire numérique hors-contexte qui permet d'identifier déjà un lexique de base et des champs sémantiques caractéristiques du corpus. Parmi les formes les plus employées dans *Europarl*, on constate, bien évidemment, la prédominance de formes fonctionnelles à faible contenu sémantique (articles, prépositions, etc.) mais également la prééminence statistique de substantifs renvoyant à des animaux ou des parties du corps.

3.1. Étude des spécificités lexicales

En déterminant par la statistique les attractions lexicales remarquables autour d'une forme-pôle, la méthode des cooccurrences spécifiques produit par *CooCS* fournit des listes de formes associées deux à deux et hiérarchisées par un indice (de spécificité) signalant le degré de probabilité de la 'cooccurrence lexicale'[2]. Le tableau ci-après présente les résultats de cette statistique de concomitances lexicales avant ou après la forme *tête* au cours de ses 2505 apparitions dans *Europarl*:

---

[1] *Lexico3* est disponible en téléchargement sur www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW. *CooCS* est téléchargeable sur le site www.cavi.univ-paris3.fr/ilpga/individus/martinez/accueil.htm.

[2] Le signe positif ou négatif indique une attraction ou une répulsion entre mots. Le nombre (de 2 à 50) indique le degré de surprise par rapport à ce qui est prévu par la statistique. Un indice de 3 représente 1/100 de chances pour que la rencontre entre deux formes ait lieu, 4 représente 1/1000, etc. Ex.: le mot CHAT attire le mot APPELONS avec un indice de +40 ce qui signale une attraction qui avait une chance sur 1000000000000000000000000000000000000000 de se produire.

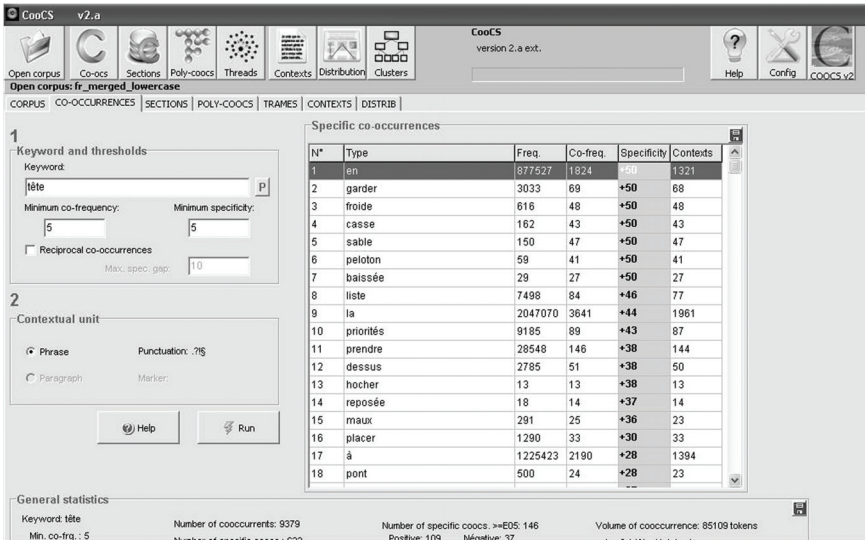


Figure 9 : Cooccurrents positifs du pôle *tête*

On calculera ainsi que, dans ses 2505 phrases d'occurrence, le pôle tête apparaît plus souvent en co-fréquence avec le nom peloton qu'avec le nom maux (co-fréq. 41 contre 25). La forme peloton se distingue par un indice de spécificité maximal (+50) qui rend compte de ses surprenantes coïncidences avec le pôle tête (41 sur 59, soit deux tiers de ses occurrences contre 25 sur 291 pour *maux*). Comme on peut le constater, en dressant un inventaire hiérarchisé des associations répétées autour d'un pôle donné, la méthode cooccurrentielle définit ce que l'on peut appeler sa valence lexicale, c'est-à-dire sa capacité à attirer d'autres formes de manière récurrente en contexte (Martinez, 2003).

3.2. Au-delà de la cooccurrence binaire : les poly-cooccurrences

En examinant les systèmes d'attraction lexicale au-delà du lien de cooccurrence binaire, on peut recenser des attractions multiples et simultanées qui s'opèrent autour d'une forme pivot et d'identifier, par conséquent, des expressions semi-figées ou idiomatiques dont le sens ne peut être déduit à partir des mots qui les constituent (Martinez, 2012). On trouve ci-dessous un extrait du graphe qui reflète une partie du complexe réseau cooccurrentiel élaboré à partir du pôle *tête* :

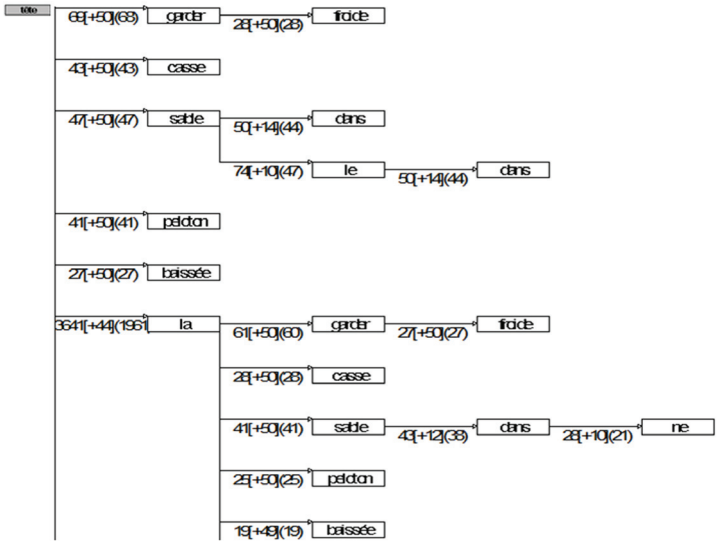


Figure 10 : Réseau de cooccurrence du pôle *tête*

En explorant, avec ce niveau de précision, l'environnement contextuel du pôle *tête*, on découvre à ce dernier un certain nombre de cooccurrents précédemment mis en avant par la méthode de cooccurrence initiale. Articulés autour des verbes (*casser, garder, placer, figurer,…*), qui restent les cooccurrents les plus spécifiques de *tête*, apparaissent des prédicats nominaux (*sable, peloton, liste,* etc.) et adjectivaux (*froide*, *haute*, *baissée,* etc.). Les cartouches d'information qui jalonnent ce graphe nous renseignent sur la réalité statistique de ces attractions multiples :

xx [yy] (zz) = co-fréquence [indice de spécificité] (contextes de rencontre)

L'arbre donne donc pour chaque nœud le nombre exact de rencontres et le nombre de contextes où ces coïncidences ont lieu. Considérons le chemin de cooccurrence *tête + sable*. Le diagnostic de la dernière cooccurrence avec la forme sable, 47[+50](47), nous informe sur l'existence de 47 contextes de réalisation pour cette poly-cooccurrence. En poursuivant la lecture à droite du chemin, on voit que celui-ci se prolonge en deux réalisations plus spécifiques intégrant les formes *le* et *dans* pour lesquels on trouvera respectivement 47 et 44 réalisations en contexte. La concordance bilingue qui suit restitue une partie des contextes de réalisation de cette poly-cooccurrence :

FR	PT
(1a) Nous ne pouvons nous *enfoncer la tête dans le sable* et nier que c'est déjà le cas.	(1b) Não podemos *enterrar a cabeça na areia* e negar a realidade do que está a acontecer.
(2a) Il ne sert à rien de *s'enfouir la tête dans le sable* et d' affirmer sans cesse le contraire selon la devise […]	(2b) Não vale a pena *enterrar a cabeça na areia* e afirmar continuamente o contrário, adoptando o lema […]
(3a) Certains *mettent leur tête dans le sable* comme des autruches et pensent que tout ira bien.	(3b) Alguns *enfiam a cabeça na areia* e pensam que tudo acabará bem.
(4a) Vous pouvez continuer à *vous cacher la tête dans le sable* si bon vous semble.	(4b) Podem continuar a *enterrar a cabeça na areia*, se preferirem.
(5a) L'autruche, alors qu'elle est capable de *fourrer sa tête dans le sable* dans certaines situations, est aussi l'animal qui court le plus vite sur le sol.	(5b) A avestruz , embora possa em determinado momento *enfiar a cabeça na areia*, é também o corredor mais veloz em terra.

Tableau 1 : Concordance bilingue de la poly-cooccurrence verbe + *la tête dans le sable* (*Europarl*/*Per-Fide*)

Il est intéressant de voir, en effet, que l'association *tête dans le sable* apparaît bel et bien dans le sous-corpus *Europarl* du corpus *Per-Fide* et on ne peut s'empêcher de constater les variations du verbe qui affectent cette poly-cooccurrence. En effet, on trouve aussi bien le verbe *s'enfoncer* que les verbes *se cacher, s'enfouir, s'enterrer, fourrer, mettre*, … *la tête dans le sable*. Certains auteurs parlent de classes d'équivalences (Gross, 1988) ; de possibilité de paradigme (Gross, 1996) ; de variantes lexicales (García, 2011). C'est cette même variation verbale que l'on observe sur la séquence *tête* + *baissée* (cf. *supra*, figure 10) :

FR	PT
(6a) Je demande donc à la Commission si elle veut *se jeter tête baissée* dans ce problème?	(6b) Pergunto aqui à Comissão: quer ela *meter-se às cegas* neste problema?
(7a) Au lieu de soutenir Israël en ce sens, d'importants États membres de l'UE *se précipitent tête baissée* sur l'État juif.	(7b) Em vez de apoiar Israel nesta questão, proeminentes Estados-Membros da UE *caem irreflectidamente* sobre o Estado judaico.
(8a) Il est regrettable que la Commission ait préféré *foncer tête baissée*, en s'en tenant à une proposition idéologique qui ne tient pas compte des opinions de la population.	(8b) É lamentável que a Comissão tenha optado , em vez disso , por *avançar de rompante* com a sua proposta ideológica que não presta atenção às opiniões dos cidadãos .
(9a) La Russie *plonge tête baissée* dans le piège et soutient indirectement et sans le savoir les forces en Géorgie qui pourraient vouloir utiliser le conflit pour satisfaire leurs propres fins politiques.	(9b) A Rússia *está a cair exactamente nesta mesma armadilha* e apoia mesmo, de modo indirecto e involuntário, as forças que, na Geórgia, poderão querer usar o conflito para obtenção dos seus próprios fins políticos.

FR	PT
(10a) J'espère donc que le Conseil entendra la voix de la sagesse et que les ministres ne *poursuivront* pas *tête baissée* le scénario de l'unification, qui serait sans issue.	(10b) Espero, portanto, que o Conselho oiça a voz da sensatez e que os Ministros não prossigam, *cabisbaixos*, o cenário da unificação, que não teria saída .

Tableau 2 : Concordance bilingue de verbe + *tête baissée* (*Europarl*/*Per-Fide*)

Dans le contexte plus immédiat à gauche de *tête baissée*, on trouve également une grande panoplie de verbes : *se jeter, se précipiter, foncer, plonger, poursuivre… tête baissée.* Outre cette variation d'ordre purement lexical, on ne peut s'empêcher de souligner la très grande diversité de moyens linguistiques mis en œuvre pour rendre, en portugais, cette expression idiomatique : *meter-se às cegas, avançar de rompante, cair (irreflectidamente), cabisbaixos*, … C'est, au contraire, une traduction littérale que l'on trouve en portugais (*enterrar a cabeça na areia)* pour l'expression française *s'enfoncer la tête dans le sable* mentionnée auparavant. Il semblerait que la traduction portugaise de l'expression idiomatique *jeter de l'huile sur le feu* subisse, en revanche, d'immenses variations dans notre corpus :

FR	PT
(11a) Les institutions européennes auraient dû y penser en 1991-1992, lorsqu'elles ont reconnu, contrairement aux règles en vigueur en matière de protection des minorités, la Slovénie d'abord et la Croatie ensuite, et ainsi *jeté de l'huile sur le feu* et déclenché la catastrophe des Balkans.	(11b) As instituições da UE deveriam ter pensado neste aspecto em 1991 1992 quando, contra as regras em vigor relativas à protecção das minorias, se reconheceu primeiro a Eslovénia e mais tarde a Croácia, *tendo* assim *posto achas na fogueira* que, mais tarde, se transformou na catástrofe dos Balcãs.
(12a) Une opération militaire de grande ampleur dans le nord de l'Iraq ne ferait qu'*ajouter de l'huile sur le feu* dans la région et pourrait également causer de sérieux problèmes en Turquie.	(12b) Uma intervenção militar em grande escala no norte do Iraque seria apenas *lançar mais achas para a fogueira* naquele país, além de que poderia igualmente causar graves problemas à Turquia.
(13a) Par ailleurs, il faut également faire comprendre au Président palestinien, M. Arafat, que son attitude ne favorise pas non plus le processus de paix. Il ne doit pas *mettre de l'huile sur le feu* en brandissant la menace d'une déclaration d'indépendance unilatérale de l'Etat palestinien.	(13b) De resto, deve ser dito que a atitude do presidente da Palestina, Arafat, não ajuda em nada o processo de paz, não devendo este *deitar mais achas na fogueira* ao ameaçar com a proclamação unilateral da independência da Palestina.
(14a) En présentant les villes et villages arabes comme un «cinquième front », les dirigeants des partis de droite *ont versé de l'huile sur le feu*.	(14b) Apresentando as cidades e as aldeias árabes como uma "quinta frente", os dirigentes dos partidos de direita *atiraram lenha para a fogueira*.

Tableau 3 : Concordance bilingue de verbe + *de l'huile sur le feu* (*Europarl*/*Per-Fide*)

En effet, alors qu'en français, l'expression ne manifeste, une fois encore, qu'une simple variation du verbe (*jeter/mettre/verser de l'huile sur le feu*), les alternatives de traduction proposées, en portugais, pour cette expression française opèrent une variation aussi bien au niveau du verbe qu'au niveau du nom et même de la préposition : *pôr/atirar/lançar/deitar* (verbes) *achas/lenha* (noms) *na/para a* (prépositions) *fogueira*. Une recherche sur corpus présente donc l'intérêt d'offrir une représentation fidèle des différentes mutations morphosyntaxiques et sémantiques qu'une expression idiomatique peut subir en contexte. Évidemment, cette variation est rarement répertoriée dans les dictionnaires. Par exemple, si l'on considère l'expression idiomatique dont il vient d'être question, on s'aperçoit qu'elle apparaît sous une forme bien plus pétrifiée dans les dictionnaires portugais *Infopédia* (http://www.infopedia.pt/) et *Priberam* (http://www.priberam.pt/dlpo/) puisque elle y est, respectivement, fixée comme suit : *deitar achas/lenha na fogueira ; deitar lenha na fogueira*. Mais un survol de notre corpus montre que les traductions de cette expression peuvent diverger considérablement de cette version dictionnarisée. En effet, en contexte, une expression de ce type peut ne mettre en avant que son sens propre (*lançar petróleo/gasolina na fogueira*) (cf. (15a)-(16b)) ou donner, au contraire, à son sens figuré une grande liberté d'expression (*evitar inflamar o debate, atear o fogo de um sentimento, alimentar o próprio monstro,* ….) (cf. (17a)-(19b)) :

FR	PT
(15a) Saddam *jette* en effet constamment *de l'huile sur le feu* - qui n'a rien de sacré - du conflit israélo-palestinien.	(15b) Com efeito, Saddam Hussein lança constantemente *petróleo para a fogueira* profana do conflito israelo-palestiniano.
(16a) Plusieurs pays de la région peuvent y prendre part directement car ils *jettent* sciemment de *l'huile sur le feu* terroriste palestinien et libanais.	(16b) Esse facto pode preocupar diversos países da região, porquanto *lançam*, a torto e a direito, *gasolina na fogueir*a terrorista da Palestina e do Líbano.
(17a) Nous devons tous éviter de jeter de l'huile sur le feu en soulevant des questions qui pourraient être jugées menaçantes dans nombre de capitales nationales.	(17b) Todos nós devemos *evitar inflamar o debate* levantando questões que possam ser interpretadas como uma ameaça em muitas capitais nacionais.
(18a) […], des déclarations qui contribuent à *jeter de l'huile sur le feu* d'un sentiment antisémite qui se répand et qui est déjà trop violent en Europe.	(18b) […] declarações essas que contribuem para *atear o fogo* de um sentimento anti-semita que grassa e é já muito violento na Europa.
(19a) Si nous n'évitons pas les réponses extrémistes, nous pourrions bien *jeter de l'huile sur le feu* que nous essayons d'éteindre.	(19b) Se não evitarmos reacções extremistas, podemos acabar por vir a *alimentar o próprio monstro* que pretendemos destruir.

Tableau 4 : Concordance bilingue de *jeter de l'huile sur le feu* (*Europarl*/*Per-Fide*)

Ces expressions caractérisées par le figement constituent un phénomène extrêmement complexe du point de vue formel et sémantique (Lamiroy, 2008 ; Mejri, 2008) qui présente un intérêt certain sur le plan contrastif, puisque, comme nous venons d'en rendre compte, la plupart d'entre elles échappent à une traduction directe lors de la recherche d'équivalents (figés ou non) dans une autre langue.

4. Considérations finales

Il est donc important de motiver l'intégration des corpus et des outils d'analyse de corpus dans l'enseignement de la traduction (outre celui de la linguistique et des langues) car le recours à ce type de ressources améliore la qualité des traductions par rapport à la seule utilisation de dictionnaires, notamment au niveau du choix des termes et de l'utilisation d'expressions idiomatiques (Frérot, 2010). Comme le note, à juste titre, Kraif (2006), les outils d'alignement et de concordance bilingue sont encore largement sous-exploités. Familiariser les utilisateurs – traducteurs, étudiants ou linguistes – avec les principaux concepts-clés du domaine et avec les techniques utilisées pour exploiter des corpus (Araújo *et al.,* 2011) se révèle donc essentiel. Nous avons tenu à montrer que le corpus multilingue *Per-Fide* met gratuitement à disposition une interface qui permet différents modes de recherche (monolingue ou bilingue simple et double) ; dictionnaire probabiliste de traduction (PTD) dans des langues et des domaines de spécialité variés. Les corpus parallèles, qui y sont gratuitement consultables en ligne, permettent d'étudier deux langues, en les confrontant et les éclairant réciproquement. Les PTDs sont tout aussi précieux, par exemple pour la traduction statistique automatique, mais aussi pour assister le travail des lexicographes et des terminologues en réalisant une partie du travail préliminaire de reconnaissance. Les textes sont, en effet, des objets complexes qui se construisent sur des relations d'attraction lexicale que les PTDs ne parviennent qu'à suggérer. Pour mettre à jour ces phénomènes de cooccurrence binaire ou multiple (notamment les expressions figées idiosyncrasiques et imprévisibles d'une langue donnée), il est possible, comme on l'a vu, d'utiliser toutes ces ressources en complémentarité avec d'autres outils pour mener des études plus textométriques.

Université du Minho — Sílvia ARAÚJO
Université du Minho — Ana CORREIA
Institut de Linguistique Théorique et Computationnelle — William MARTINEZ

Références bibliographiques

Araújo, Sílvia / Almeida, José-João / Dias, Idalete / Simões, Alberto, 2010. «Apresentação do projecto Per-Fide: Paralelizando o Português com seis outras línguas», *Linguamática* 2, 71-74.

Araújo, Sílvia / Oliveira, Ana / Dias, Idalete, 2011. «Como pesquisar em corpora», *Workshop présenté aux I Journées Internationales Corpora & Traduction*, 22 et 23 novembre 2011, Université du Minho, Braga - Portugal http://perfide.ilch.uminho.pt/site.pl/workshop.pt.

Frérot, Cécile, 2010. «Outils d'aide à la traduction : pour une intégration des corpus et des outils d'analyse de corpus dans l'enseignement de la traduction et la formation des traducteurs», *Les Cahiers du GEPE*, 2. Outils de traduction - outils du traducteur?

Gross, Maurice, 1988. «Les limites de la phrase figée», *Langages*, 90, 7-22.

Gross, Gaston, 1996. *Les expressions figées en français : noms composés et autres locutions*, Paris, Ophrys.

García-Page, Mario, 2011. «Collocations complexes (application à l'espagnol)», *Lingvisticae Investigationes*, 34, 68-111.

Kraif, Olivier, 2006. «Qu'attendre de l'alignement de corpus multilingues?», *Traduire. 4e Journée de la traduction professionnelle* 210, 17-37.

Lamiroy, Béatrice, 2008. «Les expressions figées: à la recherche d'une définition», in : Blumenthal, Peter & Mejri, Salah (ed.), *Les séquences figées : entre langue et discours.* Zeitschrift für Französische Sprache und Literatur, Beihefte 36, 85-98.

Langlois, Lucie, 1996. *Bitexte, bi-concordance et collocation*, Thèse pour l'obtention de la Maîtrise en Traduction, Université d'Ottawa, Canada.

Martinez, William, 2003. *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de Doctorat en Sciences du Langage, Université de la Sorbonne Nouvelle - Paris 3, France.

Martinez, William, 2012. «Au-delà de la cooccurrence binaire : poly-cooccurrences et trames de cooccurrence», *Corpus* 11.

Mejri, Salah, 2008. «Figement et traduction: problématique générale», *Meta* : *journal des traducteurs* 53, 244-252.

Williams, Geoffrey (dir.), 2005. *La Linguistique de Corpus*, Rennes, Presses Universitaires de Rennes.

Presentazione del *Tesoro dei Lessici degli Antichi Volgari Italiani* (*TLAVI*)

Mi sono dedicato negli ultimi anni alla realizzazione di un thesaurus lessicale di impianto onomasiologico di volgarismi ricavati da repertori lessicali (liste di parole, glossari, vocabolari) di area italoromanza risalenti all'epoca medievale: un repertorio che ho denominato *Tesoro dei Lessici degli Antichi Volgari Italiani* (*TLAVI*). Del TLAVI ho realizzato, oltre a una versione cartacea, anche una versione elettronica (oggetto di questa breve presentazione), costituita da un sito Internet di presentazione del progetto (all'indirizzo: www.tlavit.it) con, al suo interno, un software per la ricerca nel corpus lessicale.

1. Le fonti del TLAVI

Per la messa in opera del TLAVI ho fatto una selezione preliminare di repertori antichi (editi in anni recenti o meno recenti) da cui estrarre il materiale lessicale utile. Ho scelto il 1500 come limite cronologico per la selezione per il motivo che è a partire dalla fine del Quattrocento che il livello di 'dialettalità', cioè di specificità geolinguistica, dei repertori lessicali e degli altri documenti provenienti dalle varie regioni italiane si riduce drasticamente per l'azione livellatrice e uniformatrice esercitata dal modello toscano.

Collegandosi al sito Internet del progetto compare una pagina di presentazione e, sulla sinistra, il menu con i collegamenti alle altre pagine del sito. Il primo collegamento, *Le fonti del TLAVI*, permette di accedere alla pagina in cui si trova l'elenco completo dei lessici finora sottoposti a spoglio per la realizzazione del TLAVI[1]:

(1) Glossario latino-volgare della Biblioteca Universitaria di Padova (GBP) (Arcangeli 1997)[2];
(2) Glossario francese-veneto (GFV) (Baldelli 1988 [1962]);
(3) Glossario italiano-arabico (GIA) (Teza 1893);
(4) Glossario latino-bergamasco (GLB) (Lorck 1893, 95-163);
(5) Glossario latino-eugubino (GLE) (Navarro Salazar 1985);
(6) Glossarietto latino-padovano (GLP) (Arcangeli 1992, 202-204);

---

[1] Le denominazioni dei repertori, solo nel bisogno opportunamente semplificate, sono dedotte dai titoli delle relative edizioni.

[2] Rispetto all'elenco che compare nella pagina del sito, aggiungo qui il riferimento bibliografico (si rimanda pertanto alla bibliografia per il recupero delle altre informazioni).

(7) Glossario latino-reatino (GLR) (Baldelli 1971 [1953]);
(8) Glossario latino-sabino (GLS) (Vignuzzi 1984);
(9) Glossario ligure al *Tresor* di Brunetto Latini (GLT) (Vitale Brovarone 2008);
(10) Glossario latino-trentino (GLTZ) (Zingerle 1900);
(11) Glossario latino-veneto (GLV) (Gualdo 1997);
(12) Glossario latino-velletrano (GLVG) (Giuliani 2010[2]);
(13) Glossario latino-volgare della Biblioteca comunale di Perugia (GLVP) (Gambacorta 2007);
(14) Glossario latino-volgare quattrocentesco (GLVQ) (Vignali 2001);
(15) Glossario provenzale-italiano (GPI) (Castellani 1980 [1958]);
(16) Glossarietto tedesco-italiano (GTI) (Scarpa 1991);
(17) Lessico latino-bergamasco (LLB) (Contini 1934);
(18) Lemmario settentrionale di Carpentras (LSC) (Colotti 1999 e 2000-2001);
(19) Vocaboli volgari da un glossario latino di Bartolomeo Sachella (VBS) (Marinoni 1962);
(20) *Vocabula* di Domenico d'Arezzo (VDA) (Pignatelli 1998);
(21) Volgarismi del *Declarus* di Senisio (VDS) (Marinoni 1955);
(22) *Vocabula* di Goro d'Arezzo (VGA) (Pignatelli 1995);
(23) Vocabolarietto milanese-fiorentino (VMFmil) (Folena 1952);
(24) Vocabolarietto milanese-fiorentino (VMFfio) (Folena 1952);
(25) *Vallilium* di Nicola Valla (VNV) (Gulino 2000).

Per ognuno dei repertori do, quando disponibili, le seguenti informazioni:

– il nome dell'editore[3] e il rinvio bibliografico all'edizione[4];
– la segnatura del codice che lo contiene, oltreché una breve descrizione del codice;
– il nome dell'autore e/o del copista[5];
– l'altezza cronologica di compilazione[6];
– il tipo di volgare presente e alcuni suoi tratti salienti[7].

Le schede descrittive dei repertori si basano precipuamente sulle informazioni ricavabili dalle relative edizioni. Ho cercato poi di integrare tali informazioni (sia

[3] Un editore, avendo solo repertori lessicali editi c'è sempre.

[4] Dei lessici raccolti sono pochi quelli che hanno conosciuto più edizioni nel tempo; piuttosto, edizioni integrali sono state precedute da edizioni parziali: ciò che è accaduto con *GLB*, una porzione del quale era stata pubblicata una ventina d'anni prima da Giusto Grion (1870): in tali circostanze, ho preso in considerazione l'ultima (che si presuppone essere filologicamente più attendibile della precedente) edizione in ordine di tempo (esemplare a proposito è *VMF*, in precedenza dato alle stampe – "malamente", come dichiarato in modo schietto dal successivo editore Folena 1952 – da Fanfani 1874-1875).

[5] Quando non siano rimasti nell'anonimato.

[6] L'assegnazione di un repertorio a una certa area geografica o a una certa altezza cronologica può raggiungere gradi anche assai elevati di approssimazione. Inoltre, in mancanza di indicazioni specifiche limitatamente alla data di compilazione dell'opera, ho assunto come data valida quella del manoscritto che lo conserva: a meno che non risulti evidente in qualche modo – in genere l'indicatore è costituito dal dato linguistico – che il manoscritto è di molto posteriore al glossario.

[7] Molto sinteticamente ho dato conto di quei tratti considerati decisivi per l'attribuzione del repertorio a una più o meno ampia casella dello scacchiere geolinguistico italoromanzo.

quelle relative ai codici che quelle relative ai testi ivi contenuti) con altre attinte da studi successivi (quando ce ne sono).

Riporto di seguito un esempio di scheda descrittiva, nel caso specifico relativa a *GLE*:

> 5) Glossario latino-eugubino (GLE)
> Editore: Maria Teresa Navarro Salazar (Navarro Salazar 1985).
> Codice: A,4,5 della Biblioteca del Real Seminario de San Carlos (Saragozza). Il presente codice miscellaneo consta di 133 fogli, di cui ben 73 riservati alla trattazione di aspetti morfosintattici della lingua latina e ai lessici latini (tra cui si trova anche il nostro glossario, da 61r a 86r), investiti del compito di procurare ai discenti le informazioni lessicali di cui avessero abbisognato. Lo spazio rimanente comprende contenuti vari: scritti a carattere religioso (sacramenti, un alleluia, un'orazione funebre, una predica), modelli di lettere e documenti professionali ad uso e consumo dei tirocinanti nelle professioni notarile e mercantile, quattro liriche petrarchesche, l'introduzione del Lucidario, etc.
> Autore: Ugovino Angeli. Fu *magister* e console del quartiere di S. Martino a Gubbio.
> Datazione: la composizione di GLE risale alla prima metà del Trecento, e probabilmente al suo secondo venticinquennio (periodo di tempo durante il quale Ugovino ricoprì la carica di console a S. Martino). Il codice è invece più tardo: il termine *a quo* è individuato da Navarro Salazar in alcuni anni prima del 1364 (anno in cui, stando a Briquet, si interrompe la fabbricazione della carta che ha come filigrana una colonna semplice con capitello dorico) e quello *ad quem* nel 1418 (raccomandandosi alla tavola cronologica a c. 70v che riporta il calendario di quest'anno).
> Volgare: la curatrice dell'edizione, per quanto attiene alla sostanza linguistica degli *interpretamenta*, assegna il volgare rappresentato al tipo 'umbro settentrionale' (riprendendo la dicitura di Ugolini). All'assegnazione a tale distretto geolinguistico congiurano alcuni tratti: fra di essi quello, *in absentia*, della mancata palatalizzazione di [a] (con l'eccezione di *fieto* [236] "fiato"), fenomeno che oggi interessa un'area che copre parte dell'Umbria settentrionale e della Toscana orientale: nell'eugubino medievale non si danno invece tracce di tale fenomeno. Un altro tratto è quello, *in praesentia*, del dittongamento non metafonetico (tale perché si realizza a prescindere dal tipo di vocale finale) di [ɛ], a esemplificare il quale si possono convocare le seguenti forme: *ceriesca* (784) "ciliegia", *cielo* (180), *diece* (118), *pieco* (619), *piei* (444), *pieie* "piedi" (274), *schiera* (411), *siero* (929), etc.: questo tipo di dittongamento «è simile a quello che troviamo a Perugia lungo tutto il Trecento» (Navarro Salazar 1985, 67).

Le fonti selezionate presentano molte differenze sul piano contenutistico, strutturale e linguistico. Più specificamente:

- 15 di essi rientrano nella tipologia delle compilazioni bilingui latino-volgari (con il latino sempre 'lingua lemmatica', senza eccezioni): si tratta di *GBP*, *GLB*, *GLE*, *GLP*, *GLR*, *GLS*, *GLTZ*, *GLV*, *GLVG*, *GLVP*, *GLVQ*, *LLB*, *LSC*, *VDA*, *VGA*;
- 5, sempre del tipo bilingue, dispongono del volgare per tradurre una lingua coeva non italoromanza, ovvero il francese (*GFV*, *GLT*), il provenzale (*GPI*), il tedesco (*GTI*) e l'arabo (*GIA*) (solo in quest'ultimo caso con un rapporto inverso, cioè con la lingua allotria non a lemma bensì in funzione di glossa);
- 1 (*VMFmil/fio*), bilingue, oppone due idiomi italoromanzi;
- 2 (*VBS* e *VDS*) sono vocabolari monolingui della lingua latina da cui uno stesso editore (Marinoni) ha prelevato e ordinato in forma di elenco alfabetico le voci volgari;

– 1 (*VNV*) è un vocabolario bilingue con il siciliano (e il fiorentino) in esponente e il latino lingua delle definizioni.

Per scendere nello specifico del tipo di volgare riscontrabile in ciascun repertorio, nella metà dei lessici è presente un volgare dell'area settentrionale: ora di più facile attribuzione a un'area più o meno ristretta, come quello di *GFV* (veneto occidentale), *GLB* (bergamasco), *GLP* (veneto con tratti più specificamente padovani), *GLTZ* (trentino-veronese), *LLB* (bergamasco-bresciano) e *VMFmil* (lombardo-milanese); ora più 'coineggiante' e quindi non passibile di essere racchiuso entro i confini di un'area circoscritta, come quello di *GBP* (lombardo-veneto con qualche screziatura ligure), *GLT* (ligure orientale-genovese), *GLV* (coinè padana, con macchie sia lombarde sia venete), *GLVP* (dell'area genericamente emiliana), *GLVQ* (idem), *LSC* (genericamente settentrionale, con venetismi) e *VBS* (genericamente lombardo, con ammiccamenti al milanese). Volgare 'di transizione' fra Italia settentrionale e Toscana, a motivo della convivenza in esso di elementi settentrionali e toscani, può essere considerato quello di *GPI*. Scendendo lungo la penisola, i primi a farsi incontro sono gli aretini *VGA* e *VDA*; a seguire *VMFfio* e, non lontano, proveniente da Gubbio, *GLE*; rispetto a quest'ultimo, *GTI* sembrerebbe dover essere collocato un po' più a est, in area marchigiana. Nel territorio laziale, poi, ci si imbatte in *GLR*, *GLS* e *GLVG*: dal primo all'ultimo però i tratti di più manifesta vernacolarità digradano in una coinè di tipo romano. Al tipo linguistico napoletano (ma le presenze toscaneggianti e latineggianti sono tante) va ascritto *GIA*. La Sicilia, infine, conta due repertori: *VDS* e *VNV* (quest'ultimo però, linguisticamente parlando, in 'comproprietà' con la Toscana).

Per quanto riguarda l'altezza cronologica di compilazione, i più antichi sono *GFV* e *GPI*, entrambi fatti risalire dagli addetti ai lavori all'inizio del Trecento. Sempre trecenteschi sono, in ordine di tempo, *GLE* (1325-1350), *VDS* (1348), *VGA* (metà del secolo) e *GLT* (di cui non pare determinabile una data, neanche per approssimazione, all'interno di questo centennio). A cavallo fra il secolo XIV e il secolo XV si posizionano invece *VDA* (1365-1413/4) e *GTI*. Tutti gli altri repertori risultano di fattura quattrocentesca: dell'inizio del secolo *GIA* (ma potrebbe risalire anche alla fine del secolo precedente), *GLTZ* (1400-1425), *GLVP*, *LLB* (1429, o poco prima); della metà del secolo circa *GBP* (1435-1460, o forse anche prima), *GLP* e *VBS* (1440-1447); in una data compresa fra l'inizio e la metà dello stesso secolo *GLV*. Al secondo cinquantennio vanno ascritti *VMF* (1485), *GLVG* (1486), *GLR* (dopo il 1491), *GLS* (prima del 1497) e *LSC*. Per *GLB* e *GLVQ* si parla genericamente di Quattrocento. Chiude il secolo, o meglio apre quello nuovo, *VNV* (1500).

Per metà delle compilazioni, infine, non si hanno indicazioni di sorta sull'identità sull'autore: questa situazione deriva, è ammissibile ipotizzare, dal particolare status di molte di queste operette, dalla funzione meramente gregaria, di ausilio allo studio del latino, e quindi non concepite perché potessero vivere di vita propria al di fuori degli scopi meramente pratici (primariamente scolastici) per cui furono state pensate.

2. La struttura del TLAVI

Attraverso il collegamento *Consulta il TLAVI* disponibile nel menu della pagina iniziale del sito, è possibile accedere alla sezione che conserva il software per la ricerca. Una volta entrati in questa sezione, è immediatamente visibile un secondo menu (posto in alto, centralmente) attraverso cui si può accedere alle tre sezioni *Categorie*, *Lemmi* e *Fonti*.

Nella prima sezione è disponibile la prima delle due modalità di ricerca fondamentali: quella all'interno delle categorie e sottocategorie onomasiologiche. Qui l'utente può prendere piena visione dell'albero onomasiologico; innanzitutto, compare l'elenco delle macrocategorie:

> Il mondo vegetale
> Il mondo animale
> I cibi e la loro preparazione, l'alimentazione
> L'abbigliamento
> Gli ornamenti
> La pulizia personale e la cura del corpo
> L'abitazione e altri edifici e costruzioni
> Gli oggetti dell'arredamento
> […]

A fianco delle macrocategorie (e, all'interno di ciascuna di esse, delle sottocategorie), si notano tre icone:

- la prima, costituita da una freccia rivolta verso il basso, serve per inoltrarsi all'interno della categoria in questione, e quindi all'interno delle diverse sottocategorie 'figlie', e per aprire quindi l'albero onomasiologico interno;
- la seconda, costituita da una lente di ingrandimento con una linea verticale al suo interno, 'lancia' la *Ricerca verticale*, cioè la ricerca (alfabetica) dei lemmi di una categoria e quindi anche all'interno delle sue categorie figlie (restringendo, a mano a mano che si 'scende' all'interno della categoria, l'ambito della ricerca);
- la terza, costituita da una lente di ingrandimento con una linea orizzontale al suo interno, permette la *Ricerca orizzontale*, cioè la ricerca solo a quel particolare livello onomasiologico in cui ci si trova (e non anche all'interno delle eventuali sottocategorie annidate).

Facciamo un esempio. Cliccando sulla freccia della terza categoria in ordine di comparsa (*I cibi e la loro preparazione, l'alimentazione*), l'utente vedrà comparire l'elenco delle 4 sottocategorie di primo grado (il grado è indicato da dei pallini):

- Cibi e bevande
- La preparazione e la cottura dei cibi
- La consumazione dei cibi
- Processi chimici relativi a cibi e bevande

A questo punto è possibile inoltrarsi all'interno della prima sottocategoria (*Cibi e bevande*) e, cliccando sulla freccia, visionare le sottocategorie[8]:

- Paste e pani
- Minestre e zuppe
- Carne e pesce
- Latte, uova e formaggi
- Frutta e verdura
- Dolci
- Ingredienti, condimenti, sostanze per conservare i cibi
- Cibo per animali
- Bevande
- Caratteristiche di cibi e bevande

In alternativa, senza aprire la tendina delle sottocategorie, si possono cercare direttamente tutti i lemmi della categoria *Cibi e bevande*; come ho detto,

– con la *Ricerca verticale* è possibile cercare i lemmi (e, sotto ciascun lemma, le forme) di quella categoria e anche delle sottocategorie che in essa sono annidate, senza distinzioni[9]:
  aceto s. m.
  acqua s. f.
  acquarello s. m.
  […]
  agliata s. f.
  agresto s. m.
  animella s. f.
  […]

– con la Ricerca orizzontale si possono cercare solo i lemmi (e le forme) di quel particolare livello onomasiologico (senza quindi mostrare anche quelli delle sottocategorie annidate):

  cibo s. m.
  compagna s. f.
  companaggio s. m.
  companatico s. m.
  conserva s. f.
  grascia s. f.
  guazzetto s. m.
  guazzino s. m.
  vitto s. m.
  vivanda s. f.

[8] La presenza della freccia indica che esistono ulteriori sottocategorie. Quando la freccia è assente, la sottocategoria in questione è l'ultima in grado.

[9] Il risultato della ricerca compare nella stessa pagina, sotto l'elenco delle categorie. Può anche essere visualizzata cliccando su *Lemmi* del menu in alto.

Se si desidera proseguire e visualizzare solo i lemmi della sottocategoria di 2° grado *Paste e pani*, si dovrà cliccare sulla relativa icona di ricerca:

[…]

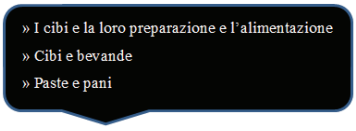
buccellato s. m.
casoncello s. m.
[…]
cialdone s. m.
collura s. f.
crescenza s. f.
[…]

"Aprendo" il lemma *casoncello*, per esempio, si otterrà il seguente risultato:

casoncello s. m.

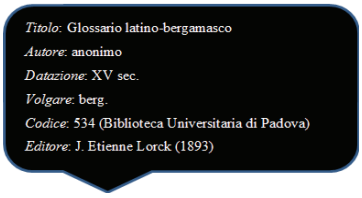
1. 'tipo di pasta' [Paste e pani]
casonçel (GLB)
casonzel (GLB)
cosonçelo (GLTZ)

Se si sposta il puntatore sopra il nome della categoria *Paste e pani*, posto tra parentesi quadre, si apre un piccolo riquadro che mostra tutto l'albero onomasiologico relativo allo specifico lemma:



1. 'tipo di pasta' [Paste e pani]

Allo stesso modo, passando con il puntatore sopra la sigla di un glossario e poi cliccandoci sopra si apre un riquadro che dà le principali informazioni relative a quel glossario:



casonçel (GLB)

Tornando al menu posto in alto, il secondo collegamento è rappresentato da *Lemmi*: esso rimanda all'elenco alfabetico di tutti i lemmi. Ad esempio, la ricerca del lemma *madre* nello spazio bianco del campo di ricerca restituisce questo risultato:

madre s. f.

1. ‘madre’ [La parentela]
mader (GLB)
madre (GBP, LSC, VDA)
mare (LSC)
matre (GLVG, GTI, LSC, VNV)

2. ‘madre, suora’ [I religiosi e i loro uffici e benefici]
madre (GBP)

3. ‘utero’ [Ossa e altre parti interne del corpo umano]
mader (GLB)

Sono quindi attestate, per il lemma *madre*, tre accezioni diverse: la prima con quattro forme (*mader*, *madre*, *mare*, *matre*), la seconda e la terza con una (*madre* e *mader* rispettivamente). Come nell’esempio precedente, anche qui è possibile aprire i relativi riquadri della categoria onomasiologica e del glossario.

Attenzione, però: se si apre il lemma *madre* non partendo dalla ricerca lemmatica (*Lemmi*) ma entrando, attraverso la ricerca per categorie (*Categorie*), dalla relativa categoria, l’informazione non pertinente a quella specifica categoria compare con un effetto di ‘scolorimento’ con lo scopo di mettere in risalto solo l’informazione pertinente (vd. fig. 1).

Ancora, l’informazione grammaticale viene data solo quando non c’è corrispondenza di ‘status’ grammaticale della forma con il lemma sotto il quale è rubricata (l’esempio è qui dato da *cani*):

cane s. m.

1. ‘cane’ [Animali quadrupedi]
can (GLVQ)
cane (GBP, GIA, GLVQ, GTI, LSC, VDA, VNV)
cani pl. (GLVQ, VNV)
chane (VMFfio)

Le icone che affiancano lo spazio bianco destinato alla ricerca di un lemma (ma anche di una delle forme attestate) servono per selezionare la *Ricerca semplice*, per adoperare le funzioni di *Ricerca avanzata* (che permettono di cercare un certo lemma o una certa forma – o una loro parte – all’interno di uno specifico glossario o di una specifica categoria) e per *Cancellare la ricerca*.

Infine, la terza sezione raggiungibile dal menù, *Le fonti*, offre l’elenco completo dei repertori-fonte del TLAVI:

Glossario latino-volgare della Biblioteca Universitaria di Padova (GBP)
Glossarietto francese-veneto (GFV)
Glossario italiano-arabico (GIA)
Glossario latino-bergamasco (GLB)
Glossario latino-eugubino (GLE)
Glossarietto latino-padovano (GLP)
Glossario latino-reatino (GLR)
[…]

Cliccando, ad esempio, su GLE, si apre inferiormente una tendina con una sintetica scheda descrittiva, che ho già detto comparire quando, all’interno di un articolo lemmatico, si clicca sulla sigla di un glossario (in fig. 2, è aperta la scheda di GLV).

3. Lo spoglio delle fonti del TLAVI

Vorrei ora spendere alcune parole sulle modalità di spoglio dei repertori-fonte e di acquisizione dei dati lessicali, cioè sul 'trattamento' del materiale lessicale raccolto. Prima di tutto, come forse è prevedibile, non ho potuto garantire a tutte le voci una collocazione all'interno della struttura onomasiologica. In sede di spoglio delle fonti, ho in genere escluso una voce in tre casi:

(1) quando non sono riuscito a cogliere, con un sufficiente grado di certezza, il suo valore semantico;
(2) perché il suo significato è semanticamente troppo generico e quindi non incasellabile all'interno della (di una?) struttura onomasiologica;
(3) perché non ho ritagliato, all'interno della struttura onomasiologica, una casella idonea ad accoglierla.

1) Bisogna per prima cosa rammentare che i testi collezionati per lo spoglio sono repertori lessicali, in cui in genere il volgare riveste la funzione di idioma 'glossante' (del latino, perlopiù): i testi sono dunque rappresentati in specie da liste di parole in cui l'elemento d'appoggio e di 'confronto' per l'interpretazione semantica di una voce volgare è dato dal solo elemento lessicale ad essa giustapposto (che si trovi a lemma o, comunque solo in pochi casi, in posizione di elemento glossante). Prendiamo due esempi dal glossario latino-velletrano (*GLVG*), edito da Giuliani (2010[2]):

(168) Hec colus la rocha
(184) Hic jems la vernata
(225) Hec falz vinaria lo roncio

In casi consimili, le forme latine a lemma rappresentato l'unico strumento di verifica nella decodificazione delle forme volgari: non di rado sono anche imprescindibili, come nel primo esempio, in cui soltanto la forma latina *colus* può instradarci verso la corretta assegnazione del significato "conocchia' alla forma *rocha*, e risolvere l'impasse conseguente alla sua situazione di omonimia con *rocca* "fortificazione; cima; etc.".

In altri casi, oltre alla lingua 'altra' soccorrono anche 'espansioni' con funzione di determinante (nei due esempi che seguono, tratti da *GLS*, è questione rispettivamente di un complemento di specificazione e di una relativa):

(3) Hec gemma -me, l'ochio dela vite
(165) Hoc subtegmen -nis, la trama che se tesse

Limitatamente ai repertori-fonte che hanno un impianto metodico, c'è anche un altro importante indicatore per la determinazione del valore semantico di una forma: l'inquadramento all'interno di una, piuttosto che di un'altra, area semantica (alcune volte 'dichiarata', altre no). Si prenda il seguente esempio:

[De tonstrina et eius armis]
[422] Hoc verriculum -li la scopecta

Se non ci si trovasse nella sezione indicata tra parentesi quadre ('L'attività e gli strumenti del barbiere'), non avrei potuto registrare – come invece ho fatto – la forma *scopecta* entro la casella relativa agli strumenti di lavoro del barbiere, ma forse avrei dovuto registrarla nella casella degli strumenti manuali di vario uso (che ho creato per dare accoglienza a forme relative a oggetti per i quali non c'è un contesto che ne circoscriva lo specifico campo d'afferenza).

2) Ci sono forme che restano fuori dal TLAVI perché semanticamente troppo generiche o, comunque, aventi una gamma di significati troppo ampia e diversificata[10] perché possa risultare fattibile una loro collocazione all'interno dello schema onomasiologico (il caso limite – e più emblematico – è rappresentato dalla forma *cosa*), almeno così come esso è stato articolato. Mi riferisco praticamente a forme quali *copertura* e *imbucà* 'imboccare' (*VBS*), *sicuru* 'sicuro' (*VDS*) e *debito* 'dovuto, giusto' (*LSC*), *parte* e *libero* (*GLVG*), *ombrosa* (*GPI*) e *remane* 'rimane' (*GLVQ*), e così via.

3) Ancora, restano fuori quelle forme per le quali non ho 'aperto' uno scomparto onomasiologico: scomparto in genere non aperto a motivo della scarsa consistenza numerica del gruppo di forme pertinenti a uno specifico 'centro di interesse'. Per esempio, avrei potuto aprire una macrocategoria riservata alle 'sostanze chimiche', se non fosse che è molto esigua la colonia dei rappresentanti lessicali di questo settore: *antemonio* 'antimonio' (*GIA*), *argiento vivo* 'mercurio' (*GIA*), *sal alchali* 'sale alcalino' (*GIA*), *salanitro* 'salnitro' (*VNV*), *sal armoniacho* 'sale ammoniaco' (*GIA*), *sal comone* 'sale comune' (*GIA*), *sal gema* 'salgemma' (*GIA*), *sal nitro* 'sale nitro' (*GIA*), e pochi altri.

In conclusione, voglio precisare che il TLAVI, nella sua versione informatica, è da considerare un repertorio *in progress*: è infatti prevista, a intervalli più o meno regolari, la 'riapertura' del corpus del TLAVI al fine di ampliarlo con nuovi lessici-fonte e quindi nuove voci (lemmi e forme) volgari; uniche restrizioni – oltre a quella, ovvia, della presenza di corpi lessicali volgari da estrarre – saranno le stesse in vigore per i lessici-fonte già accolti: il non superamento del limite cronologico del 1500 e l'attendibilità filologica delle edizioni di riferimento.

Sapienza Università di Roma Alessandro ARESTI

[10] Il che capita spesso, come si è detto, con i repertori-fonte alfabetici privi di indicatori contestuali utili ad enucleare, di una certa forma, una determinata accezione o un determinato spettro di accezioni 'vicine'.

Bibliografia

Arcangeli, Massimo, 1992. «La tradizione dei glossari latino-volgari (con un glossarietto inedito)», *Contributi di filologia dell'Italia mediana* 6, 193-209.

Arcangeli, Massimo, 1997. *Il glossario quattrocentesco latino-volgare della Biblioteca Universitaria di Padova (ms. 1329)*, Firenze, Accademia della Crusca.

Aresti, Alessandro, 2010. «Un *Glossario dei glossari* degli antichi volgari italiani. Preliminari, risultati, prospettive», *Bollettino dell'Atlante Lessicale degli Antichi Volgari Italiani* 1, 9-32.

Aresti, Alessandro, 2012. «La variazione linguistica e lessicale nella tradizione dei glossari medievali in volgare», in: Bianchi Patricia / De Blasi Nicola / De Caprio Chiara / Montuori Francesco (ed.), *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali*, Atti dell'XI congresso SILFI (Napoli, 5-7 ottobre 2010), Firenze, Cesati, vol. 2, 39-48.

Aresti, Alessandro, 2013. «Tesoro dei Lessici degli Antichi Volgari Italiani (TLAVI)», *Zeitschrift für romanische Philologie* 129 (4).

Baldelli, Ignazio, 1971 [1953]. «Glossario latino-reatino del Cantalicio», in: Id., *Medioevo volgare da Montecassino all'Umbria*, Bari, Adriatica, 195-238; pubblicato in precedenza in: *Atti dell'Accademia toscana di Scienze e Lettere "La Colombaria"*, 17, 1953, 367-406.

Baldelli, Ignazio, 1988 [1962]. «Un glossarietto francese-veneto del Trecento», in: Id., *Conti, glosse e riscritture dal secolo XI al secolo XX*, Napoli, Morano, 1988, 159-168; pubblicato in precedenza in: *Atti dell'VIII congresso internazionale di studi romanzi*, Firenze, 1960, vol. 2, 757-763.

Castellani, Arrigo, 1980 [1958]. «Le glossaire provençal-italien de la Laurentienne (ms. Plut. 41, 42)», in: Id., *Saggi di linguistica e filologia italiana e romanza (1946-1976)*, Roma, Salerno Editrice, 1980, vol. 3, 90-133; pubblicato in precedenza in: *Lebendiges Mittelalter, Festgabe für Wolfgang Stammler*, Friburgo (Svizzera), Universitätsverlag, 1958, 1-43.

Colotti, Maria Teresa, 1999. «La storia della lingua italiana attraverso i glossari: prodromi all'edizione del lemmario settentrionale di Carpentras (seconda metà del XV sec.)», *La nuova ricerca. Pubblicazione annuale del Dipartimento di Linguistica, Filologia e Letteratura moderna dell'Università degli Studi di Bari* 8, 123-64.

Colotti, Maria Teresa, 2000-2001. «L'edizione del lemmario settentrionale di Carpentras (seconda metà del XV sec.): lettere b–c», *La nuova ricerca. Pubblicazione annuale del Dipartimento di Linguistica, Filologia e Letteratura moderna dell'Università degli Studi di Bari* 9-10, 201-38.

Contini, Gianfranco, 1934. «Reliquie volgari dalla scuola bergamasca dell'Umanesimo», *L'Italia dialettale* 10, 223-40.

Fanfani, Pietro, 1874-1875. «Vocabolarietto milanese-fiorentino», *Il Borghini* 1, 311-314, 343-346, 361-363, 370-374.

Folena, Gianfranco, 1952. «Vocaboli e sonetti milanesi di Benedetto Dei», *Studi di filologia italiana* 10, 83-144.

Gambacorta, Carla, 2007. «Un glossario latino-volgare (Biblioteca comunale Augusta di Perugia, ms. B 56)», *Contributi di Filologia dell'Italia Mediana* 21, 79-134.

Giuliani, Valentina, 2010[2]. *Il glossario inedito di Domenico Gallinella (Velletri 1486)*, Roma, Aracne, (prima ediz.: 2007).

Grion, Giusto, 1870. «Il pozzo di S. Patrizio», *Il Propugnatore* 3, 80-8.

Gualdo, Riccardo, 1997. «Dal papa allo «strazarolo»: un inedito glossario latino-veneto (1450)», *Studi linguistici italiani* 23 (2 della 3ª serie), 180-218.

Gulino, Giuseppe, 2000. *Il* Vallilium *di Nicola Valla*, Aachen, Shaker Verlag (a cura di Mario de Matteis).

Lorck, Jean Etienne, 1893. *Altbergamaskische Sprachdenkmäler (IX.-XV. Jahrhundert)*, Halle, Max Niemeyer Verlag.

Marinoni, Augusto (ed.), 1955. *Dal "Declarus" di A. Senisio: i vocaboli siciliani*, Palermo, Pubblicazioni del Centro di Studi filologici e linguistici siciliani.

Marinoni, Augusto, 1962. «Vocaboli volgari da un glossario latino di Bartolomeo Sachella», in: AA. VV., *Saggi e ricerche in memoria di Ettore Li Gotti*, Palermo, Centro di studi filologici e linguistici siciliani, vol. 2, 226-59.

Navarro Salazar, Maria Teresa, 1985. «Un glossario latino-eugubino del Trecento», *Studi di lessicografia italiana* 7, 21-155.

Pignatelli, Cinzia, 1995. «Vocabula Magistri Gori de Aretio», *Annali Aretini* 3, 273-339.

Pignatelli, Cinzia, 1998. «Vocabula Magistri Dominici de Aretio», *Annali Aretini* 6, 36-166.

Scarpa, Emanuela, 1991. «Uno sconosciuto glossarietto italo-tedesco», *Studi di filologia italiana* 49, 59-74.

Teza, Emilio, 1893. «Un piccolo glossario italiano e arabico del Quattrocento», *Rendiconti della Reale Accademia dei Lincei. Classe di scienze morali, storiche e filologiche*, serie V, 2, 77-88.

Vignali, Luigi, 2001. «Un glossario latino-volgare quattrocentesco e il *Vocabolarium breve* di Gasparino Barzizza», in: Bongrani, Paolo / Dardi, Andrea / Fanfani, Massimo / Tesi, Riccardo (ed.), 2001, *Studi di storia della lingua italiana offerti a Ghino Ghinassi*, Firenze, Le Lettere, 3-87.

Vignuzzi, Ugo, 1984. *Il «Glossario latino-sabino» di Ser Iacopo Ursello da Roccantica*, Perugia, Università Italiana per Stranieri.

Vitale Brovarone, Alessandro, 2008. «Un glossario ligure al *Tresor* di Brunetto Latini (Paris, Bibliothèque Nationale de France, fr. 1113)», *Bollettino dell'Atlante Lessicale degli Antichi Volgari Italiani* 1, 53-69.

Zingerle (von), Wolfram, 1900. «Eine wälschtirolische Handschrift (Um das Jahr 1400)», *Zeitschrift für romanische Philologie* 24, 388-394.

Un'ontologia per il DiTMAO (*Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*)

1. Premessa

Il Progetto DiTMAO pone, fra i propri obiettivi principali, quello di rendere accessibili sul Web, in formato digitale, la terminologia contenuta nei testi di argomento medico-farmaceutico in antico occitano, organizzata in forma di dizionario[1]. Come capita di frequente nella lessicografia storica, anche in questo progetto si è affrontato il problema della modalità con cui realizzare la corrispondenza fra i valori concettuali che le parole, appartenenti ai campi semantici considerati, esprimono nei testi del corpus e quelli equivalenti nella terminologia medico-farmaceutica attuale, ove, naturalmente, tale persistenza esista[2]. Il lavoro lessicografico che è stato avviato consente di indagare non solo fenomeni linguistici, ma anche, tramite essi, di descrivere un universo concettuale sincronico, ovvero un dominio di conoscenze relativo alla medicina e alla farmacopea occitana di epoca medievale. Esso, infatti, se analizzato solo sulla base di fenomeni storici e culturali, risulterebbe incompleto, o impreciso.

Se la redazione delle voci, inoltre, viene realizzata tenendo presente anche la prospettiva diacronica che ha inevitabilmente inciso sull'evoluzione dei valori concettuali di termini che sono sopravvissuti fino ai nostri giorni, si pongono tutte le condizioni favorevoli:

---

[1] Si veda la comunicazione di Corradini («La realizzazione del Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan (DiTMAO): problemi di organizzazione della conoscenza medico-farmaceutica attestata dai manoscritti in occitano antico»), presentata in questa stessa sessione, dove sono descritti in dettaglio gli scopi del progetto e i testi sui quali lo spoglio lessicale viene condotto. Il progetto è coordinato da Guido Mensching, Università di Gottinga, finanziato dalla DFG e sviluppato in collaborazione con l'Università di Pisa (Corradini) e l'Università di Colonia (Bos).

[2] Tale fenomeno si verifica in particolare riguardo a testi antichi di argomento tecnico per i quali i dizionari generali di una lingua, che non abbiano utilizzato spogli specifici, forniscono informazioni lacunose o non puntuali. Osservazioni a questo proposito, concernenti termini tecnici del *Corpus Hippocraticum* e relative definizioni presenti negli strumenti lessicografici disponibili per il greco antico, si leggono in Bozzi 1982 (in particolare nelle note introduttive alle pp. 1-3). In questo genere di opere, inoltre, si può verificare il fenomeno della specializzazione semantica di parole di uso comune le quali, in contesti particolari, o sull'esempio di modelli greci o latini ove ciò accade molto di frequente, si caricano di accezioni tecniche specifiche.

- per avere un quadro dettagliato del settore specifico sul quale la ricerca viene condotta;
- per mettere a disposizione di molte comunità di filologi e linguisti una mole di dati organizzata in forma ottimale ai fini della consultazione sia nella tradizionale veste di un dizionario cartaceo, sia mediante sistemi informatici di ultima generazione semanticamente orientati.

Per raggiungere gli obiettivi indicati, le voci devono essere redatte seguendo un duplice sistema di classificazione logico-semantica: quello aderente al periodo medievale nel quale esse furono impiegate e quello coerente con l'uso che esse hanno assunto attualmente.

A tale scopo e considerando soprattutto che il risultato finale della ricerca dovrà essere consultabile sul Web, oltre che, come detto, su supporto cartaceo, si è ritenuto opportuno adottare appropriate tecnologie del cosiddetto Web semantico e una serie di applicazioni basate su strumenti software per il trattamento di testi, immagini e lessici, sviluppati all'ILC-CNR di Pisa[3].

In questo contributo si prende in esame l'approccio metodologico tipico dei modelli concettuali utilizzati per le ontologie di dominio grazie al quale è stato possibile dare una descrizione formale esplicita del particolare dominio di interesse, cioè la medicina e la farmacopea medievale in antico occitano[4].

La riflessione preliminare si è basata sull'analisi delle informazioni e dei materiali disponibili, alcuni dei quali in formato cartaceo, altri in formato elettronico.

Sono stati utilizzati:

- i manoscritti, ovvero le fonti primarie, editi o in corso di edizione[5];
- i termini medici in antico occitano scritti in alfabeto ebraico, con relative corrispondenze in ebraico e in arabo[6];
- indici lemmatizzati di testi editi disponibili in formato elettronico.

I materiali, dunque, sono eterogeni, redatti in lingue e alfabeti diversi ed è indispensabile che il sistema da adottare consenta, come già detto, di stabilire un rete di relazioni semantiche diacroniche fra il significato posseduto dai termini in epoca medievale e quello assunto nell'età contemporanea.

---

[3] L'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa (‹www.ilc.cnr.it›) ha una pluridecennale esperienza nel trattamento dei testi mediante sistemi informatici; fra questi, è in fase molto avanzata di sviluppo un'applicazione Web di filologia computazionale per la produzione di edizioni elettroniche di testi e per la comparazione fra testi antichi e loro antiche traduzioni. Si veda, per esempio, il sito del progetto ERC Ideas Advanced Grant 249431, "Greek into Arabic. Philosophical Concepts and Linguistic Bridges" (‹www.greekintoarabic.eu›), che ha lo scopo di contribuire, mediante confronto semiautomatico fra alcuni capitoli delle *Enneadi* di Plotino e la traduzione di essi in lingua araba, nota come *Teologia di Aristotele*, all'incremento del *Glossarium Graeco-Arabicum* redatto presso la Ruhr-Universität Bochum.

[4] In questo ambiente il termine "ontologia" significa "specificazione esplicita di una concetto". Si veda Gruber (1993, 199-220). Cfr. anche Staab / Studer (ed.) (2009).

[5] Si veda Corradini (1997).

[6] Si veda Bos / Mensching / Hussein / Savelsberg (2011).

2. Le motivazioni alla base della scelta ontologica per l'annotazione semantica strutturata

Nella prassi redazionale di tipo tradizionale è sempre esistita la possibilità di utilizzare parole chiave in grado di classificare tematicamente le definizioni, come, per esempio, i tecnicismi, gli arcaismi, gli usi metaforici, i sinonimi. Tuttavia, nell'ambito di una più precisa organizzazione semantica dei dati contenuti in un vocabolario riferito ad uno specifico dominio di conoscenza e, per di più, relativo ad un'epoca remota, il fatto di inserire ogni entrata di una eventuale lista di parole-chiave in uno schema precostituito, ove compaiano voci di livello superiore, voci ad esse soggiacenti e relazioni fra voci e sotto-voci, assume senza dubbio notevoli valori aggiunti[7].

Ciò, infatti:

- consente ai redattori di dare una struttura semantica omogenea ed esplicita alle definizioni della terminologia per la composizione del dizionario cartaceo;
- rende più produttive le fasi di consultazione, studio e ricerca da parte, soprattutto, degli specialisti che accederanno alla versione Web del dizionario e all'archivio dei testi;
- rende lo schema concettuale indipendente dalla lingua nel quale vengono descritte classi, sottoclassi e relazioni[8];
- soprattutto, rende lo schema concettuale indipendente dalla lingua del testo studiato. Ciò vale in primo luogo per i testi multilingui (per esempio, le opere che contengono parti in latino, lingue romanze, greco, ebraico, arabo), come nel caso del progetto DiTMAO, sul quale tale ipotesi di organizzazione ontologica dei dati del dominio si sta sperimentando con successo.

---

[7] Lo schema logico e concettuale potrebbe essere definito 'tassonomia', adottando un termine più comunemente usato nel settore degli studi umanistici rispetto a quello di 'ontologia' che, nato nell'ambito degli studi logici e filosofici, è stato oggi riqualificato dagli ingegneri dell'informazione che ad esso hanno attribuito una connotazione tecnologica e strumentale. La differenza fra le due espressioni consiste soprattutto nel fatto che l'ontologia consente di dare una rappresentazione esplicita di uno specifico dominio di conoscenza, mettendo in relazione fra loro le componenti concettuali le quali non sono quasi mai isolate le une rispetto alle altre, e alle quali è possibile applicare una indefinibile quantità di attributi. Il potenziale descrittivo delle ontologie, pertanto, è molto potente soprattutto nel caso in cui lo si voglia applicare a un dominio specifico e delimitato come, appunto, la medicina e la farmacopea di epoca medievale. L'utilizzo delle ontologie, come elemento descrittivo del lessico, si è molto diffuso anche in conseguenza dell'affermazione della teoria del Lessico Generativo (Pustejovsky, 1995). Un'interessante applicazione di questo metodo al lessico tecnico presente nei lavori di Ferdinand de Saussure si legge in Ruimy / Piccini / Giovannetti (2012, 1043-1056).

[8] Resta inteso, comunque, che le etichette concettuali attribuite a ciascun elemento dello schema devono essere espresse in forma linguistica e, nella fattispecie, l'idioma scelto dal gruppo di ricerca è il francese. Ciò significa che per interrogare il lessico e ritrovare tutti i termini che condividono valori concettuali o partecipano delle medesime relazioni fra concetti diversi, è necessario operare la selezione su liste predisposte in francese. Il vantaggio dell'approccio ontologico consiste anche nel fatto che tali etichette, in numero non rilevante, possono essere facilmente tradotte per un accesso multilingue ai dati del dizionario, sempre che il termine usato nella traduzione (per es., l'inglese) ricopra il concetto di quello francese originario.

La conoscenza dello specifico dominio che i testi, sulla base dei quali viene redatto il dizionario, veicolano, viene pertanto rappresentata in forma ontologicamente strutturata, esplicita. La valutazione puntuale della terminologia medico-farmaceutica medievale in antico occitano è un caso-studio particolarmente interessante poiché necessita di strumenti di analisi più fini di quelli offerti da semplici indici di parole-forma o di lemmi presenti nelle fonti, corredati di eventuali concordanze.

Oltre ad essi, appare oggi sempre più funzionale interrogare la base dei dati terminologica o gli stessi testi utilizzando, come chiave di accesso, un concetto o un tema generico (es. "unguento"), oppure una relazione fra due concetti (es. "unguento" per "ferita"), oppure ancora una relazione fra più elementi (es. "unguento" per "ferita" in una determinata parte del corpo, come la "testa")[9].

I risultati che si ottengono con programmi di tipo tradizionale al fine di produrre vocabolari terminologici di specifici settori, inoltre, potrebbero non essere esaustivi perché è accertato, per esempio:

- che una delle fonti (o chi effettua una ricerca) denoti uno stesso tema con parole diverse da quelle utilizzate da un'altra fonte;
- oppure che in fase di recupero delle informazioni ('information retrieval') venga usata una chiave di accesso (es. "unguento", "ferita", "testa") diversa da quella, semanticamente identica, che è attestata. Ciò provoca una evidente impossibilità di recuperare le informazioni che invece sono presenti, sia pure in una veste diversa.

Il superamento di questi limiti e la certezza di ottenere risultati esaustivi si ottiene mediante un sistema di 'data modeling' grazie al quale l'utente viene assistito nell'organizzazione e nella scrittura dello schema concettuale tipico del dominio di sua particolare competenza ed interesse[10]. La progettazione di questo innovativo sistema è stata effettuata anche per il progetto di edizione elettronica di un corpus di manoscritti di F. de Saussure[11].

Prima di entrare nel dettaglio dello schema ontologico con cui è stata organizzato il dominio della terminologia medico-farmaceutica dell'antico occitano, desideriamo fornire un esempio concreto che contribuisca subito a chiarire almeno la più evidente delle ragioni che hanno spinto ad adottare questo impianto metodologico e a realizzare le conseguenti componenti del sistema informatico.

Si tratta del caso in cui termini differenti sussumono il medesimo valore semantico inserito nello schema: per es., il concetto di "unguento", in antico occitano *oignement* (e varianti *oinhement*, *oinnement*, *onhement*, *hongemen*) e *onguent*, ma anche

[9] Le etichette o denominazioni dei concetti, per comodità espositiva, vengono qui espresse in italiano, mentre, come detto, nel sistema esse compaiono in francese.

[10] In rete si trovano una grande quantità di ottimi siti accademici ove ricavare informazioni, anche di tipo divulgativo, su questi aspetti. Fra i sistemi di scrittura ('editor') più noti e gratuiti per la creazione di ontologie si indica qui *Protégé* (‹http://protege.stanford.edu/overview›).

[11] Si veda il contributo a questo convegno di Pesini / Del Grosso / Bozzi: «Ferdinand de Saussure e la linguistica romanza. Un'applicazione web per l'edizione elettronica dei manoscritti».

dura confeccio e *apostolico*, che sono tipi particolari di unguento. Il problema sarebbe parzialmente risolvibile anche con altri metodi, come, per esempio, la preparazione di tabelle di corrispondenza con cui la macchina sarebbe in grado di equiparare almeno le varianti (grafiche, fonetiche, morfologiche) di una medesima parola. Le caratteristiche linguistiche del corpus, tuttavia, sono tali da indurre la procedura a generare errori: l'utilizzo meccanico delle corrispondenze da parte di un programma di calcolatore potrebbe, infatti, far considerare varianti di uno stesso termine forme che in realtà appartengono a parole diverse.

Con l'attribuzione del valore concettuale "unguento" si superano tutte queste difficoltà e i due lemmi *oignement* e *onguent*, con tutti gli altri eventuali sinonimi oltre che naturalmente tutte le varianti, sono unificati su base semantica e concettuale. L'insieme dei contesti nei quali i lemmi ricorrono si possono ottenere selezionando la voce "unguento" nello schema ontologico predisposto.

Nelle fasi iniziali del progetto si era valutata la possibilità di risolvere il problema della equivalenza semantica fra termini differenti in una stessa lingua, e/o appartenenti a lingue diverse, e/o appartenenti a tranche diacroniche molto distanti, mediante l'utilizzo di un modello, noto nel settore del trattamento automatico del linguaggio col nome di WordNet (*EuroWordNet*, per le lingue europee)[12]. Esso si basa sulla descrizione del valore semantico grazie alla concatenazione di liste di sinonimi ('synset'): in pratica, una parola viene definita, come capita spesso anche nei vocabolari cartacei redatti con metodi descrittivi tradizionali, mediante altre parole che denotano il medesimo concetto o si riferiscono al medesimo oggetto della realtà. Abbiamo, tuttavia, constatato che le versioni del sistema sono state principalmente adottate per lingue parlate contemporanee come, per esempio, l'italiano (*ItalWordNet*)[13], anche allo scopo di avere a disposizione uno strumento di controllo per i programmi di traduzione automatica. L'impianto metodologico centrato sulla sinonimia, che è alla base di *WordNet*, di fronte alla necessità dei redattori di DiTMAO di tenere sotto controllo tecnicismi del lessico medico-farmaceutico e botanico, correlati anche ad aspetti di semantica diacronica (tutti elementi che è impensabile trattare con l'utilizzo di 'synset'), si è dimostrato non pertinente alle finalità.

---

[12] Si vedano il sito ufficiale di *WordNet* (‹www.illc.uva.nl/EuroWordNet›) e Rodrìguez / Climent / Vossen et alii (1998, 117-152).

[13] A questo proposito si veda Marinelli / Roventini / Enea (2004, 465-468); il manuale d'uso di *ItalWordNet* è accessibile sul sito dell'ILC-CNR (‹www.ilc.cnr.it›). Solo di recente, nell'ambito di un progetto che l'ILC sta realizzando in collaborazione con il *Perseus Project* (Tufts University, Boston), si sta effettuando un utilizzo di *WordNet* in lessicografia greca classica (*AncientGreekWordNet*) con lo scopo, però, di favorire lo sviluppo di nuove modalità di interrogazione, semanticamente orientate, di archivi testuali, a fini principalmente didattici. Al momento attuale non esiste ancora una pubblicazione scientifica dedicata all'argomento.

3. Lo schema ontologico

Il primo passo per la definizione di classi, proprietà e regole dello schema ontologico è consistito nell'analisi dei materiali e nella successiva verifica dell'esistenza di standard, riconosciuti a livello internazionale, idonei allo scopo. Alcune prove preliminari di popolazione[14] dello schema ontologico, eseguite al fine di verificare la validità e l'applicabilità di tali standard a questo specifico progetto, hanno evidenziato come esse non fossero sufficienti a rappresentare il dominio di conoscenza relativo a tale ambito nei materiali disponibili[15]. È stato necessario, quindi, estendere lo schema ontologico, pur mantenendo la compatibilità con gli standard di partenza. Si è trattato di un processo iterativo che, grazie a successive fasi di popolamento dello schema di volta in volta affinato, ha consentito di progettarne uno pienamente rispondente ai requisiti. Senza scendere in eccessivi dettagli non necessari in questa occasione, descriviamo qui i punti principali dell'attività che si è svolta su due fronti diversi, ma correlabili fra loro.

Sono stati predisposti due schemi.

Il primo, descritto con linguaggio UML[16], è in grado di rappresentare le classi principali e le relazioni per quanto riguarda gli aspetti grammaticali (grafico-fonetici e morfologici), che sono spiegate in dettaglio nel contributo di Maria Sofia Corradini citato sopra alla nota n. 1. Grazie a tale linguaggio, è possibile catalogare lemmi, sottolemmi, sinonimi e varianti (morfologiche, grafico fonetiche, ecc.), indicando l'alfabeto, la categoria grammaticale, il numero, il significato, la lingua, il nome scientifico, l'eventuale corrispettivo in un'altra lingua antica, il periodo nel quale ogni singola voce (lemma, sottolemma e/o variante) era in uso e in quali documenti è attestata.

Lo schema è stato popolato con i reali lemmi estratti dalle fonti che, entrando nella classificazione ontologica, ne hanno assunto le relative proprietà. Questo processo è avvenuto grazie ad una specifica interfaccia grafica, semplice ed intuitiva.

---

[14] 'Popolare/popolazione' sono termini tecnici adoperati dagli informatici che si occupano di sistemi ontologici e si riferiscono alla "attività di attribuzione di istanze ai singoli elementi dello schema".

[15] Fra gli standard esistenti, sono stati considerati quelli più adatti al tema di questo progetto: si tratta di FRBR-oo (*Functional Requirements for Bibliographic Records-object oriented*) adottato dalla 'International Federation of Library Associations' (IFLA), per il quale si veda ‹http://archive.ifla.org/VII/s13/wgfrbr/FRBR-CRMdialogue_wg.htm›. Esso incorpora elementi dello schema concettuale e terminologico presente nel CIDOC-CRM (*Conceptual Reference Model*), validato come standard ISO-21127 nel 2006, per il quale si veda ‹www.cidoc-crm.org›. Abbiamo, inoltre, preso in considerazione il *Lexical Markup Framework* (LMF), ISO-24613 validato come standard nel 2008, per il quale si veda ‹www.lexical-markupframework.org›.

[16] UML ('Unified Modeling Language'). Nello schema i nomi delle classi e delle relazioni, per convenzione, sono scritti in inglese. Viene utilizzato il francese come lingua del progetto per quanto riguarda la denominazione dei campi presenti nella scheda di immissione dei dati nell'ontologia e per le fasi di consultazione. Si veda anche la nota n. 8.

Essa ha l'aspetto di una scheda per l'immissione di metadati bibliografici. Con questo sistema sono stati inseriti fino ad oggi i dati relativi a più di 1500 lemmi.

Il secondo schema è invece in grado di rappresentare le classi, sottoclassi e relazioni per quanto riguarda gli aspetti semantico-concettuali. Esso è correlato a quello precedente mediante la superclasse 'OldOccitanicMedicalTerminology' e consente di attribuire ai lemmi un valore concettuale (fig. 1).

Anche 'OldOccitanicMedicalTerminology' racchiude classi e relazioni: esse descrivono la classificazione medievale dei termini medici e botanici a partire da tre sottoclassi principali: 'AnimalWorld' "mondo animale", 'VegetalWorld' "mondo vegetale", 'MineralWorld' "mondo minerale".

Il legame tra la 'OldOccitanicMedicalTerminology' e la 'ModernMedicalTerminology' rende esplicita la capacità del sistema di esprimere la relazione diacronica fra i termini.

Si riportano di seguito alcuni esempi di termini relativi al globo oculare attestati nella *Notomia* di Anric de Mondavilla, testo in lingua d'oc che accoglie numerosi latinismi di ambito anatomico. Tali voci mostrano corrispondenze o discrepanze di differente tipologia rispetto alle voci equivalenti nella lingua moderna, secondo un processo parallelo a ciò che avviene nella gran parte dell'area romanza come, per esempio:

- *cornea* "cornea" che, pur nel mutamento di forma, mantiene nella lingua moderna (*cournèio*) il medesimo significato posseduto nel medioevo;
- *uvea* "iride", che è stato sostituito, nella denominazione della medesima entità, da *iris*;
- *aranea*, termine oggi scomparso assieme all'entità a cui si riferiva, che è stata concettualmente assimilata alla retina.

4. Dalla classificazione della terminologia ai testi digitalizzati

Il processo di classificazione della terminologia medico-farmaceutica, secondo i due schemi predisposti dai redattori con l'ausilio del personale informatico che collabora al progetto, non è isolato dai testi nei quali la terminologia è documentata. Per questa ragione, le informazioni inserite nello schema ontologico saranno utilizzate per effettuare una lemmatizzazione semi-automatica su ulteriori testi resi disponibili in formato elettronico. Inoltre, il componente di filologia computazionale che gestisce tutto l'archivio, sarà in grado di associare alle trascrizioni delle opere del corpus anche le corrispondenti immagini digitalizzate delle fonti originali. Un componente aggiuntivo consentirà, infine, di utilizzare la classificazione dello schema ontologico anche per consentire ai redattori di introdurre annotazioni su testi e immagini (o porzioni di immagini).

Pisa, ILC-CNR — Andrea BOZZI
Università di Bologna — Damiana Luzzi

Bibliografia

Bos, Gerrit / Mensching, Guido / Hussein, Martina / Savelsberg, Frank, 2011. *Medical Synonym Lists from Medieval Provence: Shem Tov Ben Isaac of Tortosa, Sefer ha-Shimmush*, Book 29. Part 1: *Edition and Commentary of List 1* (Hebrew-Arabic-Romance/Latin), Leiden, Brill.

Bozzi, Andrea, 1982. *Note di lessicografia ippocratica. Il trattato sulle arie, le acque, i luoghi*, Roma, Edizioni dell'Ateneo.

Corradini, Maria Sofia, 1997. *Ricettari medico-farmaceutici medievali nella Francia meridionale*, Firenze, Olschki.

Gruber, Thomas Robert, 1993. «Translation Approach to Portable Ontology Specifications», *Knowledge Acquisition*, 5 (2), 199-220.

Marinelli, Rita / Roventini, Adriana / Enea, Alessandro, 2004. «Building a Maritime Domain Lexicon: Few Considerations on the Database Structure and the Semantic Coding», *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Paris, The European Language Resources Association, II, 465-468.

Pustejovsky, James (1995). *The Generativ Lexicon*, MIT Press, Cambridge, MA.

Rodrìguez, Horacio / Climent, Salvador / Vossen, Piek *et al*., 1998. «The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology», *Computers and the Humanities* 32 (2-3), 117-152.

Ruimy, Nilda / Piccini, Silvia / Giovannetti, Emiliano, 2012. «Les Outils Informatiques au Service de la Terminologie Saussurienne», pubblicato in linea negli atti del Congrès Mondial de Linguistique Française (Lyon, 2012), consultabile liberamente all'indirizzo www.ilc.cnr.it/viewpage.php/sez=ricerca/id=917/vers=ita, 1043-1056.

Staab, Steffen / Studer, Rudi (ed.), 2009. *Handbook on Ontologies*, Berlin – Heidelberg, Springer.

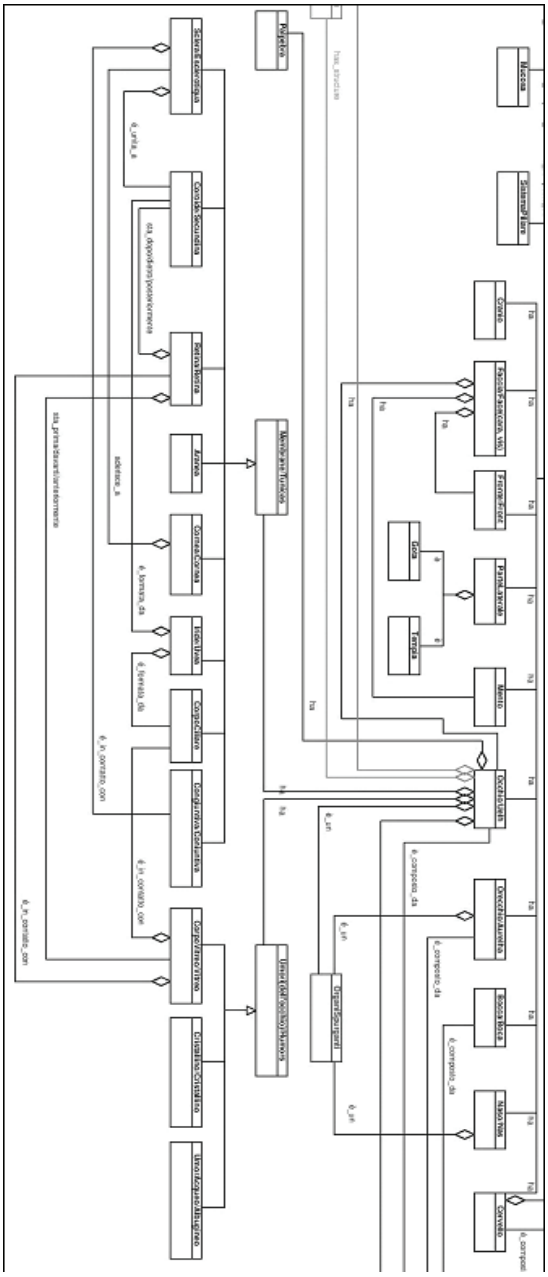


Fig. 1. Schema, descritto in UML, degli aspetti semantico-concettuali relativi al globo oculare

Categorías semánticas para describir estructuras argumentales en un ámbito de especialidad[1]

Los estudios sobre tipologías nominales de sustantivos no abundan en las lenguas de especialidad. Algo que resulta sorprendente si tenemos en cuenta el poder descriptivo que una tipología nominal puede tener en un ámbito de especialidad, donde existe una mayor restricción del significado de las unidades lingüísticas y de sus combinaciones léxicas. En este artículo exponemos, por un parte, los primeros pasos de una tipología nominal en el ámbito de especialidad del medio ambiente y en el subdominio de los desastres naturales. Por otro lado, planteamos aquí una hipótesis según la cuál si los términos que aparecen en los argumentos de los verbos de un ámbito de especialidad se clasifican y describen según los roles semánticos que suelen ocupar –basándonos en un estudio de corpus– es posible predecir los verbos asociados con una estructura argumental dada. Se trata de una idea que cobra especial interés en el ámbito de la traducción, ya que estas predicciones tienen un alcance interlingüístico.

El objetivo de este artículo es plantear las bases necesarias para sistematizar la descripción de la estructura argumental, requisito previo para verificar la hipótesis señalada. En concreto, exponemos un protocolo para establecer las categorías semánticas de los conceptos del subdominio de los desastres naturales. Utilizamos la estructura conceptual de EcoLexicon (ecolexicon.ugr.es), un tesauro visual que representa los conceptos especializados del medio ambiente en redes semánticas. Para llevar a cabo la clasificación de conceptos en categorías, nos basamos en las relaciones semánticas de EcoLexicon. En este artículo tomamos como punto de partida los términos en español pero dado que la equivalencia entre términos está establecida en EcoLexicon, la clasificación podría ser fácilmente extrapolable a otras lenguas.

1. La estructura argumental como predictor léxico interlingüístico

Los estudios sobre clasificación de sustantivos en categorías abundan. Algunos adoptan una perspectiva estrictamente lingüística, como las clases de objetos de Gross[2] (1994), las clases léxicas de Bosque (1999) o la tipología nominal de Flaux et

[1] Esta investigación ha sido realizada en el marco del proyecto de investigación RECORD: Representación del Conocimiento en Redes Dinámicas [Knowledge Representation in Dynamic Networks, FFI2011-22397], financiado por el Ministerio de Ciencia e Innovación de España.

[2] Es cierto que las clases de objetos tienen como perspectiva una aplicación computacional.

Van Velde (2000). Otros, desde una perspectiva computacional, han dado lugar a ontologías tales como WordNet Miller 1990, FrameNet (Baker *et al* 1998), VerbNet (Kipper 2005) o ADESSE[3] (García-Miguel *et al* 2010).

Ninguno de estos estudios, sin embargo, es directamente transferible a la clasificación de conceptos en un área de especialidad. En ese sentido, el grupo LexiCon está trabajando en un sistema de clasificación que permita estructurar la base de conocimientos EcoLexiCon en categorías semánticas nominales. EcoLexicon representa de forma visual el conocimiento especializado en el ámbito de las Ciencias Ambientales. Hasta ahora, la organización de los conceptos de esta base de datos se ha basado en roles semánticos (AGENTE, PROCESO, PACIENTE, RESULTADO), pero dicha clasificación resulta insuficiente, ya que no los estructura según sus rasgos semánticos.

El objetivo último de la clasificación de los conceptos de EcoLexiCon en categorías semánticas es doble. Por una parte, nos permitirá mejorar la información fraseológica de la base de conocimiento. Hasta ahora, los recursos lexicográficos especializados han prestando poca atención a la combinatoria de los términos. Resulta sin embargo sorprendente si tenemos en cuenta que cada término tiene unas preferencias léxicas que varían de una lengua a otra. Por ejemplo, los verbos que se combinan con un término en una lengua no pueden traducirse utilizando equivalencias de la lengua general, puesto que en los lenguajes de especialidad, cada idioma cuenta con reglas se combinatoria léxica propias que a menudo se basan en la semántica de sus argumentos. Desde esta perspectiva, una tipología de las clases semánticas de los distintos ámbitos de las Ciencias Ambientales nos permitiría alcanzar un mayor poder descriptivo de la fraseología propia de cada ámbito. La segunda utilidad de esta clasificación sería la posibilidad de predecir la traducción multilingüe de verbos basándonos en las estructuras actanciales. Esta idea se basa en un hipótesis de que los verbos equivalentes en distintas lenguas *próximas* comparten una misma estructura actancial (Buendía Castro 2013). En el siguiente ejemplo puede comprobarse cómo, a pesar de que los verbos *spew, éjecter* y *expulsar* no siempre son equivalentes directos en la lengua general, sí funcionan como equivalentes dentro del subdominio de la volcanología dado que, como puede observarse, estos verbos comparten una misma estructura actancial. Es decir, sus argumentos tienen el mismo rol semántico y función sintáctica. Además, los términos que actúan como argumentos (*volcán, lava*) pertenecen a una misma clase conceptual (ACCIDENTE GEOGRÁFICO, MATERIAL GEOLÓGICO).

Campo: Ciencias Ambientales Ámbito: Volcanología Marco: RELEASE frame			
Rol semántico	agente	verbos	TEMA
Sintaxis	Sujeto		Complemento Directo
Clase semántica	ACCIDENTE GEOGRÁFICO		MATERIAL GEOLÓGICO
Lexicalizaciones	*The volcano*	spewed	lava and ashes
	El volcán	expulsó	lava y cenizas
	Le volcan	a éjecté	de la lave et des cendres

Tabla 1. Estructura argumental de los verbos *spew, expulsar* y *éjecter*

Nuestra hipótesis es que si conseguimos clasificar todos los conceptos de la base de conocimiento EcoLexiCon en clases conceptuales, seremos capaces de establecer un mecanismo semiautomático que permita averiguar la traducción de un verbo dado en contexto dentro de un ámbito de especialidad. Esto es lógico si tenemos en cuenta que cada verbo selecciona ciertas categorías en cada uno de sus argumentos y que, a su vez, cada estructura argumental lleva asociada unos verbos concretos. Por ejemplo, dentro del subdominio de la volcanología, a partir del enunciado (1) se deduce la estructura argumental (2) y las posibles traducciones de ese verbo en francés e inglés (3 y 4 respectivamente). Como se observa en la tabla 1, la estructura argumental de los verbos *spew, expulsar* y *éjecter* coincide en cuanto a roles semánticos (AGENTE, TEMA) y categorías semánticas (ACCIDENTE GEOGRÁFICO, MATERIAL GEOLÓGICO). Esto nos permite decir que es muy probable que los verbos que comparten una estructura argumental dada sean equivalentes. Si esta hipótesis es cierta, una vez que hayamos estudiado las estructuras argumentales de los verbos de cada subdominio será posible determinar la traducción de un verbo de un subdominio de especialidad a partir de su estructura argumental.

1. El volcán sigue expulsando lava y cenizas.
2. [AGENTE/S/ACCIDENTE GEOGRÁFICO/] V [TEMA/COD/MATERIAL GEOLÓGICO]
3. Verbos en francés correspondientes a esta estructura: *éjecter, rejeter, cracher, éjecter*
4. Verbos en inglés correspondientes a esta estructura *expel, eject, spit, erupt*

2. Hacia una tipología de las clases semánticas del Medio Ambiente

Dado que cada ámbito de especialidad tiene unos patrones lingüísticos propios, distintos a la lengua general, el establecimiento de categorías semánticas de un campo de especialidad debe hacerse atendiendo a la idiosincrasia de cada ámbito, en este caso el Medio Ambiente. Presentamos aquí la metodología que seguimos para establecer estas categorías semánticas. El primer paso consiste en la constitución de un corpus sobre el subdominio que estudiamos, por ejemplo el de la volcanología, seguido de su análisis utilizando herramientas semiautomáticas como SketchEngine[3] o AntConc[4] y, por último, la representación de los resultados en la base de datos EcoLexiCon. Después, explotamos los resultados siguiendo varios procedimientos que explicamos en los apartados siguientes.

2.1. Constitución y análisis del corpus

El primer paso de nuestro estudio es la constitución corpus comparable en inglés, francés y español con el fin de establecer equivalencias interlingüísticas. Para ello existen dos métodos.

El primer método consiste en la recopilación de artículos científicos, de divulgación y periodísticos de un sub-ámbito de especialidad dado, por ejemplo la volcanología, que pertenece al ámbito más general de la Sismología, los desastres naturales metereológicos (Meteorología) o los movimientos de ladera de cadenas montañosas (Geología). Extraemos los artículos de revistas especializadas a través de bibliotecas electrónicas y los convertimos a formato txt. Al ser ámbitos tan restringidos, un corpus pequeño de unas doscientas mil palabras es a menudo suficiente para que sea representativo. En este caso, hemos constituido corpus sobre desastre naturales meteorológicos.

Otra opción complementaria para constituir un corpus especializado es la herramienta automática WebBootCat, integrada en SketchEngine. Esta herramienta permite buscar en la web de manera automática textos en los que aparezcan distintas combinaciones de número variable de un conjunto de palabras clave. Por ejemplo, para obtener un corpus de los desastres naturales causados por fenómenos meteorológicos hemos hecho una búsqueda a partir de esta lista de palabras clave: *tifón, huracán, tornado, tsunami, precipitación, vientos, huracán, terremoto, tormenta tropical, fenómeno, intensidad, destrucción, daños, costa, desastres*. Hemos combinado de manera automática estas palabras en grupos de tres palabras para buscar textos que las contengan. Los tipos de texto que hemos obtenidos son artículos divulgativos o artículos de prensa sobre fenómenos meteorológicos.

Al aunar estos dos procedimientos, hemos obtenido un corpus de 300.000 palabras, que resulta representativo si tenemos en cuenta lo restringido que es este campo.

---

[3] ‹http://www.sketchengine.co.uk/›

[4] ‹http://www.antlab.sci.waseda.ac.jp/software.html›

2.2. Categorías léxicas: del término al verbo y del verbo al término

El verbo selecciona los sustantivos con los que se combina, por ejemplo el verbo *solucionar* requiere en su COD sustantivos como *problema, situación*. Pero también sucede lo contrario, y es que los sustantivos, sobre todo en las lenguas de especialidad, también imponen restricciones léxicas al verbo con el que se combinan. El corpus se procesa con el programa SketchEngine (Kilgarriff *et al* 2004). La función "Word List" nos permite acceder a una lista de las palabras clave más frecuentes del corpus, de donde extraemos los términos más recurrentes. Por ejemplo, en el corpus de desastres naturales obtenemos términos como *huracán, tsunami*, *tormenta tropical*. Observamos en la figura 1 cómo el término *huracán* aparece en el corpus con una serie de verbos prototípicos tales como *golpear, arrasar, destruir, azotar.*

densas Myanmar. En 2008, un intenso	huracán golpeó	la costa de Myanmar (antes llamada Birmania
casi imposible, tanto, quizá, como que un	huracán arrase	una costa. Uno puede tomar medidas y precaverse
El poderoso viento y la lluvia de un	huracán golpean	un edificio Florida . Los huracanes son
Louis-Georges Arsenault. "Normalmente, un	huracán destruye	las frágiles viviendas de los más pobres
las huellas de la hecatombe; impetuosos	huracanes derribaron	altas torres, y no faltan, más bien sobran
Pero lo peor no queda ahí, pues ciclones y	huracanes arrasarán	todo lo que encuentren a su paso. Esta
mal tiempo también a causado desástres: el	huracán azotó	la región, arrancó techos y rompió árboles
cinco años, el 29 de agosto del 2005, ese	huracán provocó	devastadoras consecuencias en Nueva Orleans
sin hogar. Las fuertes precipitaciones del	huracán causaron	abundantes inundaciones en varias partes
de última hora. El huracán Katrina	golpeó	ayer la costa de Florida causando importates

Figura 1. Concordancia de [huracán + V]

Estudiamos los verbos asociados con cada término gracias a la función "Word Sketch", que permite una visualización más rápida y directa que la lista de concordancias. Tal y como observamos en la tabla 2, obtenemos una lista de los verbos que se combinan de manera más frecuente con cada uno de estos términos en posición de sujeto y objeto, dentro de estructuras como "<Term> V N", donde <Term> representa cualquiera de los términos sobre los que se consulta el "Word Sketch". Estas dos estructuras indican que el AGENTE puede aparecer en primera o segunda posición argumental.

	"<Term> V N"	"N V <Term>"
huracán	*tocar, pasar, afectar, producir, azotar, atravesar*	*formar, ser, causar, acercar, impulsar, afrontar, originar*
tsunami	*viajar, golpear, devastar, provocar, alcanzar*	*generar, provocar, llegar, producir, causar*
tormenta tropical	*tocar, dañar, producir, causar, arrojar, originar, ocasionar*	*desarrollar, pasar, originar, producir*

Tabla 2. Verbos asociados con los términos *huracán, tsunami, tormenta tropical*

Observamos que estos términos seleccionan verbos que expresan las relaciones causa-efecto (*causar, producir, originar, ocasionar*), verbos de movimiento (*pasar, tocar, alcanzar, atravesar*), verbos de destrucción (*azotar, dañar*) y verbos de existencia (*originar, formar, ser, producir*).

A su vez, esta información sirve, de manera inversa, para estudiar los sustantivos que aparecen en las posiciones clave de cada uno de estos verbos. Esto nos da información sobre las categorías semánticas con las que se combina cada verbo. Desde esta perspectiva, es la propia combinatoria del verbo, dentro de cada lenguaje de especialidad, lo que nos guía en el establecimiento de las categorías léxicas. Así, las categorías son más significativas y alcanzan un mayor poder predictivo que si las estableciéramos independientemente del corpus.

Una vez que obtenemos la lista de verbos, estudiamos la estructura argumental de cada uno de ellos prestando especial atención a los sustantivos que aparecen como argumentos. Estos pueden agruparse en grandes categorías según la similitud de sus rasgos. Por ejemplo, es evidente que *cometa* y *meteorito* pertenecen a una misma categoría diferente a la de conceptos como *partícula* o *huracán*. El siguiente paso consiste en extraer las categorías léxicas que se combinan con cada verbo. Así, el verbo *golpear* aparece en nuestro corpus combinado con los siguientes tipos de sustantivo en las posiciones de Agente y Paciente:

Categoría léxica	N del Agente
CUERPO CELESTE	cometa, meteorito, asteroide,
PARTÍCULA ELEMENTAL	partícula, neutrinos, electrones, fotones
RADIACIÓN	sol, rayo de luz, radiación
DESASTRE NATURAL	huracán, tornado, tifón, tsunami, tormenta tropical, terremoto, inundación

Tabla 4. Algunos agentes del corpus sobre desastres naturales meteorológicos

Categoría léxica	N del Paciente
LUGAR NATURAL	suelo, tierra, bosque, costa, playa
LUGAR ARTIFICIAL	edificio, colegio, casa
LUGAR POLÍTICO	región, continente, país, ciudad

Tabla 5. Algunos pacientes del corpus sobre desastres naturales meteorológicos

Estas listas de los sustantivos más comunes en los argumentos de cada verbo nos llevan a constituir una primera clasificación de las categorías semánticas como CUERPO CELESTE o PARTÍCULA ELEMENTAL. Se trata sin embargo de una primera clasificación que es necesario verificar mediante otros procedimientos que

nos permitan tener criterios lingüísticos más fiables. Para ello, utilizamos tests distribucionales basados en las relaciones conceptuales propias de cada categoría como veremos en el siguiente apartado.

2.3. Tests distribucionales basados en relaciones conceptuales

Con el fin de asentar cada categoría y sus miembros sobre criterios sólidos, establecemos para cada categoría una serie de tests distribucionales basados en las relaciones semánticas que cada concepto activa dentro del propio corpus y en la base de conocimiento EcoLexicon. Si tomamos el ejemplo de la categoría DESASTRE NATURAL en EcoLexiCon, vemos que está caracterizada por las proposiciones conceptuales siguientes:

- Un DESASTRE NATURAL causa PÉRDIDADES HUMANAS/MATERIALES/ECONÓMICAS
- Un DESASTRE NATURAL daña el ENTORNO
- Un DESASTRE NATURAL ocurre en un periodo de tiempo corto
- Un DESASTRE NATURAL ocurre de manera violenta

Los sustantivos *huracán, tornado, tifón, tsunami, tormenta tropical, terremoto, inundación* cumplen la veracidad de estas proposiciones. Todos ellos causan pérdidas humanas, materiales y/o económicas, dañan el entorno y ocurren rápido y de forma violenta.

- Un *huracán/tornado/tifón/tsunami/tormenta tropical/terremoto/inundación* causa PÉRDIDADES HUMANAS/MATERIALES/ECONÓMICAS.
- Un *huracán/tornado/tifón/tsunami/tormenta tropical/terremoto/inundación* daña el MEDIO AMBIENTE.
- Un *huracán/tornado/tifón/tsunami/tormenta tropical/terremoto/inundación* ocurre en un periodo de tiempo corto.
- Un *huracán/tornado/tifón/tsunami/tormenta tropical/terremoto/inundación* ocurre de manera violenta.

Sin embargo, este procedimiento no soluciona enteramente el problema de la constitución de las clases semánticas y su justificación. Por una parte, no podemos tomar como único punto de referencia la base de datos EcoLexicon puesto que no todos los términos que encontramos en los argumentos de los verbos están representados en ella. Esto no es de extrañar, puesto que la base de datos EcoLexiCon no reúne aún la totalidad de los términos relacionados con el medio ambiente. Por otra parte, algunos sustantivos que aparecen en los argumentos de los verbos, como *edificio* o *país*, no pertenecen directamente a este ámbito y por lo tanto no están representados en EcoLexiCon. Para salvar este escollo, planteamos tests distribucionales basados en las propiedades sintáctico-semánticas de los sustantivos en cuestión. Nos inspiramos en trabajos sobre clases léxicas que aplican estos procedimientos como los de Gross (1994) o Flaux y Van Velde (2000).

Este es el caso de la clase LUGAR ARTIFICIAL. Estos sustantivos se caracterizan por necesitar de la acción humana como agente de su creación, por eso son compatibles con esta estructura:

- LUGAR ARTIFICIAL fue construido por...
- El *edificio / colegio / casa* fue construido por...

Los referentes de estos sustantivos se caracterizan por tener un modelo ideal previo a su existencia que constituye su máxima expresión. Es decir, que para que un ente llegue a la categoría de *edificio* debe cumplir unas características, entre ellas la de tener una finalidad propia. Desde el punto de vista lingüístico esto se manifiesta porque el sustantivo es compatible con estructuras que indican su funcionalidad:

- LUGAR ARTIFICIAL sirve para V
- El *edificio / colegio / casa* ha sido construido para trabajar / estudiar / vivir

Estos sustantivos suelen tener un poseedor, por lo que pueden insertarse en el sintagma nominal:

- El LUGAR ARTIFICIAL de N (posesor)
- El *edificio / colegio / casa* de Juan

Además, en ciertos casos, pueden actuar por metonimia como sujetos de verbos que requieren un agente humano:

- El LUGAR ARTIFICIAL ha pensado / decidido / ordenado
- El *colegio / establecimiento / juzgado ha pensado / decidido / ordenado*

Estos N aparecen también con verbos de movimiento:

- Ir a / venir de LUGAR ARTIFICIAL
- Voy al / vengo del *colegio / establecimiento / juzgado*

Este tipo de tests son eficaces para descartar la inclusión de ciertos elementos dudosos en el grupo. Por ejemplo, los sustantivos *tienda de campaña* o *paseo marítimo* no verifican la totalidad de las proposiciones anteriores y por lo tanto no forman parte de la clase léxica LUGAR ARTIFICIAL:

- *La tienda de campaña fue construida por
- La tienda de campaña de Juan
- ?La tienda de campaña ha decidido que
- Voy a la / vengo de la tienda de campaña

- El paseo marítimo fue construido por
- *El paseo marítimo de Juan
- *El paseo marítimo ha decidido que
- Voy al / vengo del paseo marítimo

No obstante, es importante señalar que las fronteras entre clases léxicas no están tan claramente delimitadas como pueda parecer en un principio. Los límites entre una clase y otra son difusos. Es el caso de los sustantivos *tienda de campaña* y *paseo marítimo*, que se encuentran en la periferia de la clase LUGAR ARTIFICIAL. Además, hay que tener en cuenta que un sustantivo puede pertenecer a dos clases léxicas diferentes a la vez puesto que el significado es polifacético. Es el contexto lo que activa un aspecto semántico u otro de cada sustantivo. Por ejemplo, el sustantivo *playa* puede funcionar como LUGAR NATURAL o FORMACIÓN GEOLÓGICA en los ejemplos siguientes:

- Lisa ha tomado el sol en la playa.
- La playa se formó en una erupción volcánica.

Tampoco el procedimiento mediante el que se establecen los tests distribucionales es infalible puesto que estos no dejan de estar basados en cierta subjetividad del lingüista que los "crea" a partir de su propia intuición o conocimientos enciclopédicos. Con el fin de obtener una categorización lo más exacta posible y, sobre todo, de obtener unos resultados adaptados a la categorización que el propio discurso especializado establece de manera interna, combinamos este procedimiento con otro mediante el cual la categorización se extrae del propio corpus.

2.4. Categorías semánticas basadas en patrones de conocimiento

Los patrones de conocimiento (*knowledge patterns*) constituyen según algunos autores (Condamines 2002; Barrière y Abago 2006; Cimiano y Staab 2006) uno de los métodos más fiables para establecer relaciones semánticas entre conceptos. Nos basamos en la hipótesis de que los términos pertenecientes a una clase semántica atienden a los mismos patrones de relaciones conceptuales. Por ejemplo, dentro del campo de la sismología, los términos *erupción* y *terremoto* se comportan de manera similar dentro de la estructura causal "La erupción / el terremoto provoca N", como en el enunciado "La erupción / el terremoto provoca daños / destrozos". De esta manera, si localizamos todos los términos lexicalizados en una relación conceptual dada, podremos decir que pertenecen a la misma clase léxica. En el ejemplo anterior, los sustantivos *daños* y *destrozos* comportan los mismos rasgos semánticos y distribucionales y por lo tanto pueden agruparse dentro de una misma clase léxica.

El primer paso de este procedimiento es por lo tanto la búsqueda de los patrones de conocimiento dentro del corpus en los que aparece cada término. En primer lugar, estudiamos las relaciones conceptuales que el hiperónimo del término en cuestión mantiene con otros conceptos dentro la base de datos EcoLexiCon. Por ejemplo, una de las relaciones conceptuales del término *huracán* es la de hiperonimia (is_a) con la categoría léxica EVENTO EXTREMO (Huracán *is_a* extreme event). A continuación, observamos la representación en EcoLexiCon que las relaciones conceptuales de la categoría EVENTO EXTREMO mantiene con otros conceptos de la ontología son las siguientes:

- AN EXTREME EVENT causes HUMAN/ECONOMIC/MATERIAL LOOSES
- AN EXTREME EVENT affects THE ENVIRONMENT
- AN EXTREME EVENT occurs in a SHORT PERIOD OF TIME

De ahí, deducimos que estas relaciones conceptuales se lexicalizan en el corpus a través de enunciados como estos:

- Un huracán causa pérdidas humanas/económicas/ materiales.
- Un huracán afecta al medio ambiente.
- Un huracán arrasa de manera rápida.

El objetivo de este procedimiento no es otro que detectar cómo se lexicalizan en el corpus esas relaciones, con el fin de agrupar todos los términos que atienden a los mismos patrones de conocimiento bajo una misma categoría léxica. Por ejemplo, dentro de la meteorología, todos aquellos daños que provoca un *huracán*, como por ejemplo *muertes, pérdidas, destrucción* pertenecerán a una misma categoría, en este caso la de PÉRDIDAS HUMANAS/ECONÓMICAS/MATERIALES. Para averiguar cuáles son los sustantivos que forman parte de esta categoría, realizamos una búsqueda en el corpus del patrono [An EXTREME EVENT causes HUMAN/ECONOMIC/MATERIAL LOOSES]. Para realizar la búsqueda, debemos estudiar primero cómo se lexicaliza la causa dentro del corpus que estamos estudiante. El verbo de causalidad por antonomasia es *causar.* Este patrono se representa en SketchEngine de la siguiente manera: [lemma="tifón"] []{1,2} [lemma="causar" [] {1,2} [tag="N.*"]. De esta manera, obtendremos una concordancia (figura 2) en la que observamos cuáles son los N que resultan de la acción de *tifón*. Podemos hacer lo mismo con otros sustantivos de comportamiento similar como *huracán*.

[lemma="tifón"] []{1,2} [lemma="causar"] []{1,2} [tag="N.*"]

lluvias en la isla de Hokkaido. En China el	tifón Talim ha causado importantes daños	y un centenar de víctimas mortales.
trae entre manos... "Parece que el	tifón no causará ningún efecto	durante este fin de semana, ¡gracias a
Morakot' llegó el domingo por la tarde , el	tifón ha causado la muerte	de tres personas y un desaparecido en las
Así se derrumbaba el hotel Chin Shuai. El	tifón Morakot ha causado en Taiwán	las peores inundaciones de los últimos
paso por la isla filipina de Luzón El	tifón Ketsana ha causado un centenar	de muertos y centenares de miles de desplazados
decretase la alerta roja y advirtiese de que el	tifón podria causar severos daños	en varias áreas costeras de la zona.
fuerte vendaval y las lluvias que arrastra el	tifón Shanshan ha causado hoy la muerte	a ocho personas y heridas a más de doscientas

[lemma="huracán"] []{1,2} [lemma="causar"] []{1,2} [tag="N.*"]

Sandy no pasó directamente sobre Haití, el	huracán ha causado estragos	en la empobrecida isla caribeña. Varios
antes del paso de 'Sandy'. Anteriormente, el	huracán había causado un muerto en Jamaica	. El ciclón avanza en dirección norte-noreste
los datos reflejados en su página web. El	huracán ha causado daños por un importe	de unos 50.000 millones de dólares (39.000
ha cobrado al menos 39 vidas. Asimismo el	huracán ha causado grandes daños	materiales. La elevación del nivel de las
Nueva York y Nueva Yersey. En Nueva York, el	huracán ha causado muchas víctimas	, la mayoría de ellas fallecidas por árboles
afectado muy gravemente las inundaciones. El	huracán ha causado inundaciones	en cinco departamentos del sur de Haití
afectado muy gravemente las inundaciones. El	huracán ha causado inundaciones en cinco departamentos	del sur de Haití y en otras regiones como
, pero incluso tras amainar su fuerza el	huracán ha causado gran destrucción	en la ciudad más grande del mundo. Inundaciones
las redes sociales. Hasta el momento el	huracán ha causado pérdidas por más	de mil 100 millones de dólares en la zona
todo el domingo para Nueva Inglaterra". El	huracán ha causado ya las cancelaciones	de numerosos vuelos y la suspensión de
tropical 'Jova', que hasta hace poco era un	huracán, ha causado la muerte	de siete personas en la costa del Pacífico

Figura 2. Corcordancia de "tifón causa N" y "huracán causa N"

También es posible realizar una búsqueda más amplia en el corpus para observar los patrones de causalidad. Mediante la búsqueda: [tag="N.*"] [lemma="causar" []{1,2} [tag="N.*"] obtenemos una concordancia como la que se muestra en la figura 3.

el tercio norte de Filipinas, donde sus	lluvias torrenciales causaron las peores inundaciones
perdieron la vida ahogados en riadas y	corrimientos de tierra causados por los aguaceros
Las aguas de la inundación se llevaron	vehículos y causaron destrozos en edificios

Figura 3. Concordancia de "N causa N"

Además de esto, es necesario ampliar las estructuras de causalidad a otros verbos como *causar, provocar, originar, propiciar, favorecer, activar.* El resultado de estas búsquedas es que, a partir de estos patronos, observamos cuáles son los sustantivos que tienen una misma distribución y obtenemos así una visión más precisa de los miembros conforman cada categoría como observamos en la siguiente tabla.

DESASTRE NATURAL	huracán, lluvias torrenciales, corrimientos de tierra, inundaciones, aguaceros
PÉRDIDAS HUMANAS/ECONÓMICAS/MATERIALES	estragos, muertos, inundaciones, destrucción, pérdidas, destrozos

Tabla 6. Miembros de dos categorías conceptuales

En un proceso circular de retroalimentación, una vez que las categorías semánticas están más claramente definidas y sabemos cuáles son los sustantivos que las integran, podemos avanzar en el conocimiento del corpus para obtener nuevas categorías y patronos. Esto quiere decir que, si sabemos cuáles son los miembros de la categoría PÉRDIDAS HUMANAS/ECONÓMICAS/MATERIALES podemos estudiar de manera rápida qué tipo de pérdidas causa cada tipo de desastre natural mediante una búsqueda como esta:

```
*DUAL
=cause_of/effect_of
1: [tag="N.*"] [] {0,5} [lemma="causar"|"provocar"|"ocasionar"] []{0,5} 2: [tag="N.*"]
1: [tag="N.*"] [] {0,5} [lemma="causar"|"provocar"|"ocasionar"] []{0,5}
2[lemma="daño.*"|"víctima.*"|"muerte.*"|"muerto.*"
|"estrago.*"|"destrucción.*"|"inundaciones.*"|"desplazados.*"]
```

Tabla 8. Búsqueda de N que causan "PÉRDIDAS HUMANAS/ECONÓMICAS/MATERIALES"

Esta búsqueda nos da acceso, de manera automática, a una lista de las causas de cada tipo de desastre natural como vemos en la tabla 6 donde se muestran los N más

frecuentes de la estructura <Term> causa N. Esto nos permite obtener una mayor precisión que la descrita en la tabla 6 sobre los miembros que constituyen la categoría PÉRDIDAS HUMANAS/ECONÓMICAS/MATERIALES. En concreto, esta búsqueda añade a esta categoría los términos: *movimiento de terreno, tragedia, erosión, inundación, daño, alud, lodo, desbordamiento, daños.*

PÉRDIDAS HUMANAS/ECONÓMICAS/MATERIALES		
tsunami	*tormenta tropical*	*inundación*
movimiento de terreno	alud	desbordamiento
tragedia	lodo	daños
erosión	inundación	
inundación	muerte	
daño		
agua		
muerte		
viento		

Tabla 9. Resultados de <Term> causa N

Este tipo de información es útil desde al menos dos puntos de vista. Por una parte, nos da una información muy útil para profundizar en la semántica de cada término a través de su combinatoria léxica. Por ejemplo, observamos que los N que aparecen en el corpus como causas de *tsunami* son más frecuentes que los de *tormenta tropical* y estos, a su vez, más frecuentes que los de *inundación*. Esto indica que el espectro semántico de estas palabras es distinto y nos conduce a una descripción del significado de cada término más preciso:

Un *tsunami* suele actuar sobre el *terreno*, provocando *daños* materiales como *erosión, inundación, muerte* y daños no materiales como *tragedia* y *muerte*, por la acción del *agua* y el *viento*.
Una *tormenta* causa daños como materiales *inundaciones, muertes, aludes.*
Una inundación causa daños sobre todo debido *lluvias* y *desbordamientos.*

Desde la perspectiva de la lingüística aplicada, esta información nos permite recopilar los miembros de cada categoría semántica y, por lo tanto, definirlas con mayor precisión. A medio plazo, esperamos que esto nos ayude a predecir las traducciones de los verbos; si tenemos un repertorio de los sustantivos que aparecen más frecuentemente en los argumentos de los verbos un subdominio dado y conocemos las categorías a las que pertenecen esos sustantivos, será posible aislar de manera automática, a partir de un enunciado, los N de los argumentos, la categoría a la que pertenecen y, por lo tanto, su estructura actancial. A partir de ahí, y puesto que las estructuras

actanciales son equivalentes entre las lenguas (tabla 1), podremos obtener una lista de los verbos equivalentes en otras lenguas.

4. Conclusión y perspectivas

Hemos partido de la hipótesis de que es posible establecer equivalencias entre los verbos de distintas lenguas de un ámbito de especialidad tomando como punto de referencia su estructura actancial. Para definir esta estructura y poder verificar nuestra hipótesis, necesitamos en primer lugar contar con un repertorio de las clases conceptuales más frecuentes de cada subdominio de las Ciencias Ambientes.

El objetivo de este artículo ha sido exponer la metodología que seguimos para establecer la clasificación de las categorías semánticas más recurrentes de los diversos ámbitos de las Ciencias Ambientales. En concreto, hemos explicado cómo aislamos los verbos que aparecen con cada término de manera automática mediante la herramienta SketchEngine. Una vez que sabemos cuáles son los verbos asociados a cada verbo, estudiamos en el corpus los sustantivos de sus argumentos. En primer lugar, estas listas de sustantivos se clasifican de manera manual e intuitiva en categorías semánticas. A continuación, con el objetivo de asentar la identidad de cada categoría sobre principios sólidos y objetivos, definimos las características propias de cada una mediante una serie de tests distribucionales. Por último, averiguamos cuáles son los miembros que pertenecen a cada categoría del subdominio. Para ello, nos basamos en la hipótesis de que los sustantivos que comparten una misma distribución pertenecen a la misma clase semántica. Así, buscamos patronos recurrentes en el corpus en los que aparecen sustantivos en cuestión y observamos cuáles se comportan de la misma manera desde un punto de vista sintáctico-semántico. Hemos ilustrado esto con los patronos de causalidad. Las lexicalizaciones de las relaciones conceptuales sirven para averiguar los N que forman parte una estructura argumental. De esta manera, conseguimos tener criterios sobre las características de cada categoría y logramos refinar los sustantivos que pertenecen a cada una de ellas.

Las perspectivas de este trabajo son numerosas. La clasificación de los términos de los subdominios de las Ciencias Ambientales constituye una primera línea de trabajo y una ardua labor. Nuestro objetivo a largo plazo, una vez que obtengamos esta tipología y verifiquemos la hipótesis de la equivalencia de verbos basada en su estructura argumental, será establecer un sistema automático que permita la traducción automática de los verbos dentro de este campo de especialidad.

Miriam BUENDÍA-CASTRO
Pilar LEÓN-ARAÚZ
Beatriz SÁNCHEZ-CÁRDENAS

Referencias bibliográficas

Baker, Collin F. / Fillmore Charles J. / Lowe, John B. The Berkeley FrameNet Project. Proceedings od the 17th international conference on Computational Linguistics, volume 1, pages 86-90.

Bosque, Ignacio / Violeta Demonte (eds), 1999, Gramática descriptiva de la lengua española. Madrid: Real Academia Española, Espasa Calpe, § 1.1-7.

Buendía, M. 2013. *Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources*, Thèse de Doctorat, Granada, Universidad de Granada.

Flaux Nelly / Danièle Van de Velde, 2000. *Les noms en français : esquisse de classement,* Paris, Ophrys.

García-Miguel, J.M./ F. González Domínguez / G. Vaamonde 2010. «ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanishv, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta (Malta), 17-23 de mayo [http://www.lrec-conf.org/proceedings/lrec2010/pdf/859_Paper.pdf]

Gross, G. 1994. «Classes d'objets et description des verbes», *Langages*, 115, 15-30.

Kilgarriff A. / P. Rychly / P. Smrz / D. Tugwell. «The Sketch Engine», Proceedings EURALEX 2004, Lorient, France.

Kipper S. / K. 2005. *VerbNet: a Broad-coverage, Comprehensive Verb Lexico* ‹http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf›

Miller G. A. 1990. «WORDNET : An on-line lexical database», *International Journal of Lexicography*, 3(4).

Vers une nouvelle histoire de la linguistique romane : le projet d'un *Dictionnaire HIstorique des COncepts Descriptifs de l'Entité Romane* (D.HI.CO.D.E.R.)

D.HI.CO.D.E.R. est un travail collectif en histoire de la linguistique romane qui entend contribuer à une remise à plat des éléments historiographiques fondateurs de cette discipline.

1. L'historique

Le projet D.HI.CO.D.E.R. est le fruit d'une réflexion initiée au sein de l'UMR HTL/CNRS 7597 de Paris 7, à partir du constat que l'ensemble des langues romanes était scientifiquement complexe à délimiter et définir et surtout qu'il était assez rarement appréhendé comme tel, dans sa globalité, quelle que soit l'entité convoquée.

Lors du colloque « Romania : réalité(s) et concepts [1]» organisé en octobre 2011 grâce aux équipes d'accueil Romania et CELJM de l'Université Nancy 2 avec le soutien du laboratoire d'Histoire des Théories Linguistiques (UMR HTL/CNRS 7597) et de la Société d'Histoire et d'Epistémologie des Sciences du Langage (SHESL), il s'est dégagé nettement le besoin de reconsidérer un certain nombre de concepts définis très diversement selon les discours nationaux ou régionaux sur l'entité romane et l'intérêt de revisiter en nuances l'historiographie de la linguistique romane en ouvrant la question notamment à d'autres aires linguistiques que les aires française et allemande que l'on privilégie habituellement dans le récit de cette discipline.

Fort de ces motivations, une équipe[2] s'est constituée en octobre 2012 autour du projet D.HI.CO.D.E.R. hébergé par le laboratoire A.T.I.L.F/CNRS. L'objectif est de recenser les concepts qui ont permis de décrire l'entité romane et de constituer la linguistique romane depuis le XIXe siècle jusqu'à aujourd'hui.

---

1 Chabrolle-Cerretini, Anne-Marie,(ed), 2012. *Romania : réalité(s) et concepts*, Actes du colloque organisé à l'Université Nancy 2, Limoges, Editions Lambert et Lucas, à paraître.

2 Les membres actuels du projet sont : Marta Andronache (ATILF), Bérengère Bouard (Université de Lorraine-ATILF), Laure Budzinski (ATILF), Cyril De Pins (Paris VII), Anne-Marie Chabrolle-Cerretini (Responsable du projet, Université de Lorraine-ATILF), Narcís Iglesias (Université de Gérone), Gilles Petrequin (ATILF), Christophe Rey (Université de Picardie-LESCLap), Jean-Loup Ringenbach (ATILF), Gilles Siouffi (Paris IV). L'équipe est appelée à solliciter des collaborateurs, mais, dès à présent, ses membres réunissent des

2. Une nouvelle histoire de la linguistique romane

La première caractéristique de D.HI.CO.D.E.R. est de porter sur l'ensemble que constituent les langues romanes. Nous considérons l'entité romane comme cadre de notre étude. D.HI.CO.D.E.R. a ainsi pour objectif de recenser les concepts qui ont permis de décrire cette entité romane depuis le XIXème siècle. Implicitement, nous induisons que la linguistique romane dont nous cherchons à étudier les étapes constitutives est bien celle qui depuis ses débuts s'est donnée pour objet les langues romanes, toutes les langues romanes prises ensemble. Pourtant nous savons parfaitement que le champ disciplinaire de la linguistique romane est beaucoup plus desserré. Depuis la recherche de M. Glessgen (2000) dans les ouvrages qui entrent dans le champ, (49 au total, depuis L. Diefenbach en 1831 à R. Posner en 1996) d'un contenu fondamental commun susceptible de devenir le socle d'une définition de la linguistique romane, nous pouvons dire que celle-ci est un champ d'une relative cohérence en ce qui concerne les langues considérées, même si souvent ce ne sont pas toutes les langues romanes qui sont prises en compte, mais une seule ou plusieurs réunies par groupe géographique. Il est très rare néanmoins que l'entité soit questionnée pour elle-même. M. Glessgen confirme aussi des traditions depuis toujours nationales qui s'accentuent de plus en plus pour des raisons institutionnelles, les poids politiques et culturels des diverses langues. Enfin, il réaffirme que les termes « linguistique » et « romane », chacun pris séparément ont des extensions différentes, réunis, la diversité des acceptions est notable et qu'il n'y a guère de consensus. Ainsi, confrontés à une définition instable, nous ne partons pas d'une définition *a priori* de la linguistique romane, mais d'un objet d'étude, l'entité romane, qui sera notre entrée dans la relecture des textes. La définition de la linguistique romane ne constitue pas davantage un but en soi, mais bien l'une des questions auxquelles le projet apportera des éclaircissements.

La seconde caractéristique de D.HI.CO.D.E.R. est de s'intéresser aux métadiscours et non de porter sur les langues elles-mêmes. C'est logiquement un travail qui se fonde sur une étude des textes théoriques, grammaticaux, les dictionnaires et les correspondances. Paul Veyne dit avec justesse qu'en histoire, nous avons affaire à des récits, nous sommes dans de la diégèse et non la mimesis, nous racontons, nous ne montrons pas :

> L'histoire est un récit d'événements : tout le reste en découle. Puisqu'elle est d'emblée un récit, elle ne fait pas revivre, non plus que le roman ; le vécu tel qu'il ressort des mains de l'historien n'est pas celui des acteurs ; c'est une narration, ce qui permet d'éliminer certains faux problèmes. Comme le roman, l'histoire trie, simplifie, organise, fait tenir un siècle en une page et cette synthèse du récit est non moins spontanée que celle de notre mémoire, quand nous évoquons les dix dernières années que nous avons vécues.Veyne (1971, 14)

---

compétences complémentaires, aires linguistiques (roumaine, italienne, française, espagnole, catalane, portugaise, occitane) spécialistes de différentes périodes de la linguistique romane et de la philologie, des compétences en histoire des idées, méta-lexicographie et bibliographie, théories descriptives ainsi que des compétences en dictionnairique.

Ce sont bien aux récits que nous nous attachons. En effet, par les jeux d'écriture, les enjeux scientifiques et/ou nationaux, l'histoire de la linguistique romane est devenue une histoire simple, linéaire, construite souvent avec les mêmes références, ce qui semble assez loin de la réalité de terrain. De plus, il devient urgent d'intégrer toutes les aires linguistiques romanes afin de ne plus laisser croire que tout s'est presque joué dans un huis clos entre Allemands et Français. Il s'agit de faire ressortir les débats qui ont eu lieu, les contextes d'émergence des idées et des concepts descriptifs de cette entité romane. Ces analyses devraient nous conduire au constat que les concepts n'ont pas forcément partout les mêmes définitions ni à toutes les époques (la notion de « roman » par exemple). L'enjeu de D.HI.CO.D.E.R. est de remettre à plat l'historiographie privilégiée.

3. Une reconstitution d'ensembles conceptuels

Il s'agit de reconstruire des ensembles conceptuels vérifiés et attestés par une étude nouvelle des textes. Les parties du dictionnaire sont celles de l'histoire de la linguistique romane reconsidérée à partir de ces nouveaux ensembles. En effet, il y a généralement consensus sur des débuts de la linguistique romane dans les années 1836-44 avec la parution de la *Grammaire des langues romanes* de F. Diez, puis nous nous arrêtons généralement aux années 1880, puis à la période 1930-1950. D.HI.CO.D.E.R. reconsidère cette périodisation et se positionne par rapport à une notion comme celle de « paradigme [3] » éminemment discutée en histoire des sciences, mais aussi s'appuie sur la notion de « texte fondamental » que l'on peut distinguer de celle de « texte pilier [4] » qui désigne un texte qui n'a pas été écrit dans le but de fonder une école, une tradition, mais qui a exercé ce rôle pour des raisons multiples, notamment celle de s'être révélé plus adapté à cette fonction qu'un autre texte.

D.HI.CO.D.E.R. qui se fonde sur une étude des textes, exploités en vue de constituer un dictionnaire numérique, fera la part belle aux extraits textuels et aux références bibliographiques. Le travail de préparation du dictionnaire (entrée du dictionnaire) se fait sous Oxygen alors que la base bibliographique se constitue avec le logiciel libre Zotero permettant un travail collectif d'élaboration et à terme le partage des données avec l'ensemble des romanistes. La métalangue de D.HI.CO.D.E.R. est le français. Les fiches des concepts retenus apparaîtront avec une entrée du concept en français, le concept en langue d'origine, des listes ordonnées chronologiques des attestations textuelles et lexicographiques dans les différentes langues romanes,

[3] Chabrolle-Cerretini, Anne-Marie, à paraître, « De la constitution des paradigmes en histoire de la linguistique romane : un enjeu du D.HI.CO.D.E.R. », Actes du premier colloque de l'équipe D.HI.CO.D.E.R. *Paradigmes et concepts pour une histoire de la linguistique romane*, 11 avril 2013, Chabrolle-Cerretini Anne-Marie (eds), Limoges, Editions Lambert et Lucas.

[4] Colombat, Bernard/ Fournier, Jean-Marie/ Puech, Christian, 2010. *Histoire des idées sur la langage et les langues*, Paris, Klincksieck.

les sens linguistiques et éventuellement non linguistiques, des commentaires et la bibliographie. Un index multilingue des concepts sera également proposé.

4. Conclusion

Sans viser l'exhaustivité, le D.HI.CO.D.E.R. entend rassembler des données élargies par rapport à celles sur lesquelles s'est construit le récit de l'histoire de la linguistique romane. Cette collecte minutieuse effectuée selon des critères méthodologiques et théoriques inspirés par nos motivations scientifiques, permettra de toute évidence de montrer que les concepts ont des définitions très variées selon les aires linguistiques et qu'il y a bien des récits nationaux voire régionaux.

Elle rendra possible des éclairages complémentaires et sans doute salutaires sur la discipline. Elle permettra sans doute aussi de se questionner sur le champ disciplinaire aujourd'hui et sa définition, s'il a réellement posé et répondu à ses problématiques qui lui sont propres à savoir comment pense-t-on l'entité romane.

Université de Lorraine/ATILF Anne-Marie CHABROLLE-CERRETINI

Références bibliographiques

Bähler, Ursula, 2004. *Gaston Paris et la philologie romane*, Droz.

Bahner, Werner, 1984. « Continuité et discontinuité dans la linguistique romane de la première moitié du XIXème siècle », *Beiträge zur Romanischen Philologie* XXIII, Heft 1.

Bahner, Werner, 1986. « Quelques problèmes méthodologiques dans l'historiographie de la linguistique romane », *Actes du XVIIIe Congrès International de linguistique et de Philologie romanes*, Université de Trèves, Tübingen, Tome VII,

Chabrolle-Cerretini, Anne-Marie, 2009. « La linguistique romane : un champ épistémologique pour penser la diversité linguistique aujourd'hui ? » Colloque International : *La romanistique dans tous ses états.* Organisé par l'EA 739 Dipralang, la collaboration du Cerc et de Redoc. Université de Montpellier III, Centre Universitaire de Béziers. 15-17 mai 2008. Paris, L'Harmattan.

Chabrolle-Cerrretini, Anne-Marie, 2013. « Le paradigme « unité/diversité » des langues dans les textes fondateurs de la linguistique romane du XIX^{e} siècle. », Colloque International *Romania : réalité(s) et concepts*, Nancy 2, 6-7 Octobre 2011, sous le direction d'A-M Chabrolle-Cerretini, Limoges,Editions Lambert et Lucas, à paraître.

Glessgen, Martin-Dietrich, 2000. « Les manuels de linguistique romane, source pour l'histoire d'un canon disciplinaire », *Romanistisches Kolloquium* XIV, Tübingen, Gunter Narr Verlag.

Müller, Bertrand, 1994. « Critique bibliographique et construction disciplinaire : l'invention d'un savoir-faire », *Genèses*, 14.

Oesterreicher, Wulf, 2000. « L'étude des langues romanes », *Histoire des idées linguistiques*, sous la direction de Sylvain Auroux, Tome 3, Sprimont, Mardaga.

Veyne, Paul, 1971. *Comment on écrit l'histoire*, Paris, Editions du Seuil.

CLRE. Corpus lexicographique roumain essentiel. 100 dictionnaires de la langue roumaine alignés au niveau de l'entrée

1. Introduction

Le projet *CLRE, Corpus lexicographique roumain essentiel. Les dictionnaires de la langue roumaine alignés au niveau de l'entrée* (2010–2013), continue une série d'initiatives qui visent l'informatisation de la lexicographie roumaine et qui se sont déroulées au sein de l'Institut de Philologie Roumaine « A. Philippide », de l'Académie Roumaine de Iasi. Ce dernier projet, CLRE, a été financé par le Conseil National de la Recherche Scientifique (CNCS) de Roumanie. Le collectif de recherche a été formé par des lexicographes de l'Institut de Philologie Roumaine « A. Philippide », de l'Académie Roumaine de Iasi (c'est-à-dire Elena Tamba, Ana Catană-Spenchiu, Marius Clim) et par des informaticiens de l'Université « Alexandru Ioan Cuza » de Iasi (Marius Răschip et Mădălin Ionel Patraşcu). Pendant cette période on a envisagé de créer un outil pour aider en particulier les lexicographes qui rédigent le *Dictionnaire-trésor de la langue roumaine*. CLRE représente un important outil pour les spécialistes et facilitera beaucoup le travail des lexicographes, parce que ce corpus offre surtout une perspective historique de la lexicographie roumaine et, implicitement, de la langue roumaine. Cette perspective signifie que le lexicographe aura la possibilité de voir comment un mot du vocabulaire roumain a évolué, s'il y a une variation sémantique, orthographique, ortho-épique ou morphologique. Ce corpus inclut divers types de dictionnaires : des dictionnaires généraux, des dictionnaires auxiliaires (de néologismes, d'étymologie, d'orthographie, orthoépie et de morphologie) et des dictionnaires spéciaux (c'est-à-dire des encyclopédies ou d'autres dictionnaires spécialisés, qui sont de grande importance pour la perspective diachronique sur la langue).

Mais cet outil sera également disponible au grand public qui aura accès à un grand nombre de dictionnaires, qui ont été inclus dans la base de données et alignés au niveau de l'entrée. De la Bibliographie de DLR (*Dictionnaire-trésor de la langue roumaine*) on a choisi 100 dictionnaires – 150.000 pages de dictionnaire. Cela constitue une première pour la langue roumaine, parce que ce corpus est à présent le plus grand et on peut toujours ajouter des nouveaux dictionnaires.

Ainsi, ce projet a eu pour buts :

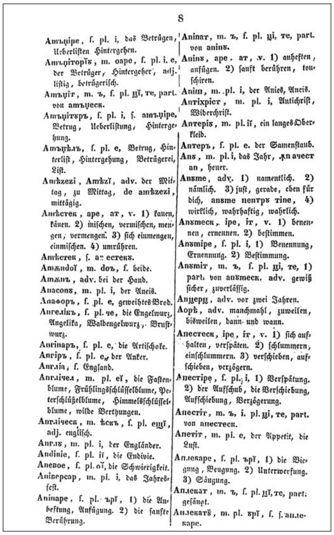
a) la réalisation d'une base de données qui comprenne les dictionnaires essentiels de la Bibliographie du DLR, alignés au niveau de l'entrée. La Bibliographie du *Dictionnaire de la langue roumaine* contient plus de 4000 titres des textes représentatifs pour la culture roumaine et elle est utilisée pour extraire des citations pour exemplifier les sens des mots-titre ;
b) l'élaboration des logiciels qui permettent la consultation interactive de ce corpus, qui peut constituer un cadre moderne de travail et de recherche lexicographique ;
c) la réalisation d'une liste de mots quasi-exhaustive, pour la langue roumaine, à partir du corpus aligné ;
d) l'augmentation de la visibilité de l'activité de recherche des professionnels linguistes et informaticiens, dans le domaine de la langue roumaine, pour promouvoir les moyens informatiques de traitement linguistique créés dans le projet.

2. La digitalisation des dictionnaires

En ce qui concerne les dictionnaires qui font partie du corpus CLRE, ils ont subi un processus de digitalisation, qui signifie qu'ils ont été numérisés à l'aide d'un scanner en V et puis traités du point de vue informatique, à l'aide d'un logiciel de reconnaissance optique des caractères (Abby Fine Reader 9) pour reconnaître les entrées et pour ensuite indexer les résultats dans une base de données XML.

Cette étape a généré de nombreux problèmes à cause du type d'alphabet utilisé, mais également à cause du format du dictionnaire.

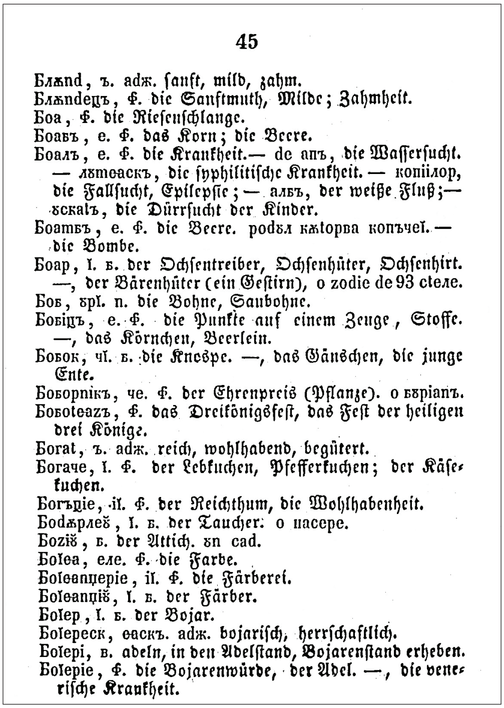
Pour la langue roumaine, les dictionnaires ont été écrits en utilisant trois types d'alphabets : l'alphabet cyrillique, l'alphabet de transition ou l'alphabet latin. Les premiers deux types ont été un véritable défi pour les informaticiens, vu que le logiciel de reconnaissance optique a été presque inefficient pour ces textes. On verra, dans les deux figures suivantes, des captures de ces dictionnaires.



8

Amъцipe, f. pl. i, das Betrügen, Ueberlisten Hintergehen.
Amъцiтopïв, m. oape, f. pl. i. e, der Betrüger, Hintergeher, adj. listig, betrügerisch.
Amъцiт, m. ъ, f. pl. цï, те, part. von amъцeck.
Amъцiтвpъ, f. pl. i, f. amъцipe, Betrug, Ueberlistung, Hintergehung.
Amъцѣлъ, f. pl. e, Betrug, Hinterlist, Hintergehung, Betrügerei, List.
Amѣzezi, Amѣzï, adv. der Mittag, zu Mittag, de amѣzezi, mittägig.
Amѣcтек, ape, aт, v. 1) kauen, käuen. 2) mischen, vermischen, mengen, vermengen. 3) sich einmengen, einmischen. 4) umrühren.
Amѣcтек, f. an естекв.
Amжndoï, m. doъ, f. beide.
Amжнъ, adv. bei der Hand.
Anacons, m. pl. i, der Aneis.
Anaфopъ, f. pl. e, geweihtes Brod.
Anгeлiкъ, f. pl. че, die Engelwurz, Angelika, Waldengelwurz, Brustwurz.
Anгinapъ, f. pl. e, die Artischoke.
Anгipъ, f. pl. e, der Anker.
Anглia, f. England.
Anглiчeл, m. pl. eï, die Fastenblume, Frühlingsschlüsselblume, Peterschlüßelblume, Himmelsschlüsselblume, wilde Berthungen.
Anглiческ, m. ѣскъ, f. pl. ецï, adj. englisch.
Anглв, m. pl. i, der Engländer.
Andinie, f. pl. iï, die Endivie.
Aneвoe, f. pl. oï, die Schwierigkeit.
Aniвepcap, m. pl. i, das Jahresfest.
Aninape, f. pl. ърï, 1) die Anheftung, Anfügung. 2) die sanfte Berührung.
Aninaт, m. ъ, f. pl. цi, те, part. von aninв.
Aninв, ape. aт, v. 1) anheften, anfügen. 2) sanft berühren, touschiren.
Aniш, m. pl. i, der Anies, Aneis.
Anтixpicт, m. pl. i, Antichrist, Widerchrist.
Anтepiв, m. pl. iï, ein langes Oberkleid.
Anтepъ, f. pl. e. der Samenstaub.
Anв, m. pl. i, das Jahr, ꙟn aчесt an, heuer.
Anвme, adv. 1) namentlich. 2) nämlich. 3) just, gerade, eben für dich, anвme пентрв тine, 4) wirklich, wahrhaftig, wahrlich.
Anвmeск, ipe, iт, v. 1) benennen, ernennen. 2) bestimmen.
Anвmipe, f. pl. i, 1) Benennung, Ernennung. 2) Bestimmung.
Anвmiт, m. ъ, f. pl. цi, те, 1) part. von anвmeск. adv. gewiß sicher, zuverlässig.
Anцерц, adv. vor zwei Jahren.
Aopѣ, adv. manchmahl, zuweilen, bisweilen, dann und wann.
Aпеcтеск, ipe, iт, v. 1) sich aufhalten, verspäten. 2) schlummern, einschlummern. 3) verschieben, aufschieben, verzögern.
Aпеcтipe, f. pl. i, 1) Verspätung. 2) der Aufschub, die Verschiebung, Aufschiebung, Verzögerung.
Aпеcтiт, m. ъ, i. pl. цi, те, part. von aпеcтеск.
Aпeтiт, m. pl. e, der Appetit, die Luft.
Aплекape, f. pl. ърï, 1) die Biegung, Beugung. 2) Unterwerfung. 3) Säugung.
Aплекат, m. ъ, f. pl. цï, те, part. gesäugt.
Aплекатв, m. pl. врï, f. f. aплекape.

Figure 1 : Andrei Iser, *Vocabular românesc-nemţesc.* Braşov, I. Gött, 1850.



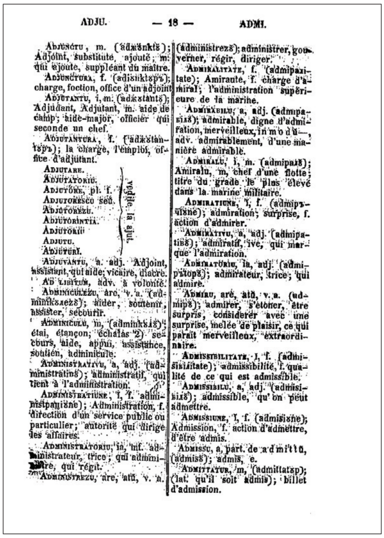
45

Блжnd, ъ. adж. sanft, mild, zahm.
Блжndeцъ, ф. die Sanftmuth, Milde; Zahmheit.
Боа, ф. die Riesenschlange.
Боавъ, e. ф. das Korn; die Beere.
Боалъ, e. ф. die Krankheit.— de апъ, die Wassersucht. — лъmeaскъ, die syphilitische Krankheit. — копiiлор, die Fallsucht, Epilepsie; — албъ, der weiße Fluß; — вскаlъ, die Dürrsucht der Kinder.
Боambъ, e. ф. die Beere. podвл кжtорва копъчеї. — die Bombe.
Боар, ї. в. der Ochsentreiber, Ochsenhüter, Ochsenhirt. —, der Bärenhüter (ein Gestirn), o zodie de 93 стеле.
Боб, врї. n. die Bohne, Saubohne.
Бобiцъ, e. ф. die Punkte auf einem Zeuge, Stoffe. —, das Körnchen, Beerlein.
Бобок, чї. в. die Knospe. —, das Gänschen, die junge Ente.
Боборniкъ, че. ф. der Ehrenpreis (Pflanze). о вврianъ.
Боботеazъ, ф. das Dreikönigsfest, das Fest der heiligen drei Könige.
Богаt, ъ. adж. reich, wohlhabend, begütert.
Богаче, ї. ф. der Lebkuchen, Pfefferkuchen; der Käsekuchen.
Богъцie, ії. ф. der Reichthum, die Wohlhabenheit.
Бodжрлеў, ї. в. der Taucher. o пасере.
Бoziў, в. der Attich. вn cad.
Боiea, еле. ф. die Farbe.
Боieaнцepie, iї. ф. die Färberei.
Боieaнцiў, ї. в. der Färber.
Боiep, ї. в. der Bojar.
Боiepeск, eacкъ. adж. bojarisch, herrschaftlich.
Боiepi, в. adeln, in den Adelstand, Bojarenstand erheben.
Боiepie, ф. die Bojarenwürde, der Adel. —, die venerische Krankheit.

Figure 2 : S. Petri, *Vocabular portativ românesc-nemţesc şi nemţesc-românesc.* Sibiu, G. de Clozius, 1861.

Le processus de reconnaissance optique des caractères n'a pas offert des résultats utilisables, parce que chaque dictionnaire en cyrillique a son propre type de caractères. Cette diversité nécessitera une adaptation du logiciel de reconnaissance optique pour chaque œuvre lexicographique. Et aussi, comme on peut y voir dans les deux captures ci-dessus, vu qu'il s'agit de dictionnaires bilingues, on a des caractères spécifiques pour chaque langue, ce qui rend la tâche du logiciel ABBYY, de reconnaissance des caractères, presque impossible. Dans le cas de dictionnaires en alphabet cyrillique et de transition, on a opté pour la translittération manuelle des entrées. De cette manière on peut aligner les entrées de ces dictionnaires avec les autres dictionnaires. Mais l'utilisateur spécialisé aura également à sa disposition la capture numérisée de l'entrée du mot choisi. Ainsi il pourra utiliser toutes les informations contenues dans ces dictionnaires.

Le format des dictionnaires a constitué tout un autre défi pour les informaticiens, qui ont dû créer un logiciel de délimitation des entrées assez flexible.



ADJU. — 18 — ADMI.

ADJUNCTU, m. (аджѕnktѕ); Adjoint, substitute, ajouté; m. qui ajoute, suppléant du maître.
ADJUNCTURA, f. (адѕnktѕpъ); charge, foction, office d'un adjoint.
ADJUTANTU, i, m. (аджѕtantѕ); Adjudant, Adjutant, m. aide de camp; aide-major, officier qui seconde un chef.
ADJUTANTURA, f. (аджѕtantѕpъ); la charge, l'emploi, office d'adjutant.
ADJUTARE. ADJUTATORIU. ADJUTÓRE, pl. f. ADJUTORESCU seu. ADJUTOREZU. ADJUTORINTIA. ADJUTORIU. ADJUTU. ADJUTURI. — vedile la ajut.
ADJUVANTU, a. adj. Adjoint, assistant, qui aide; vicaire, diacre.
AD LIBITUM, adv. à volonté.
ADMINICULEZU, are, v. a. (adminikѕлezѕ); aider, soutenir, assister, secourir.
ADMINICULU, m, (adminikѕлѕ); étai, étançon; échalas 2) secours, aide, appui, assistance, soutien, adminicule.
ADMINISTRATIVU, a, adj. (administratiвѕ); administratif, qui tient à l'administration.
ADMINISTRATIUNE, i, f. (administpaцiѕne); Administration, f. direction d'un service public ou particulier; autorité qui dirige les affaires.
ADMINISTRATORIU, ia, mf. administrateur, trice; qui administre, qui régit.
ADMINISTREZU, are, atu, v. a. (administrezѕ); administrer, gouverner, régir, diriger.
ADMIRALITATE, f. (admipaлitate); Amiraute, f. charge d'amiral; l'administration supérieure de la marine.
ADMIRABILU, a, adj. (admipaбiлѕ); admirable, digne d'admiration, merveilleux, in m o d u —, adv. admirablement, d'une manière admirable.
ADMIRALU, i, m. (admipaлѕ); Amiralu, m, chef d'une flotte; titre du grade le plus élevé dans la marine militaire.
ADMIRATIUNE, i, f. (admipъцiѕne); admiration; surprise, f. action d'admirer.
ADMIRATIVU, a, adj. (admipatiвѕ); admiratif, ive, qui marque l'admiration.
ADMIRATORIU, ia, adj. (admipъtopѕ); admirateur, trice; qui admire.
ADMIRU, are, atu, v. a. (admipѕ); admirer, s'étoner, être surpris, considerer avec une surprise, melée de plaisir, ce qui parait merveilleux, extraordinaire.
ADMISSIBILITATE, i, f. (admisiбiлitate); admissibilité, f. qualité de ce qui est admissible.
ADMISSIBILU, a, adj. (admisiбiлѕ); admissible, qu'on peut admettre.
ADMISSIUNE, i, f. (admisiѕne); Admission, f. action d'admettre, d'être admis.
ADMISSU, a, part. de a d m i t t u, (admisѕ); admis, e.
ADMITTATUR, m, (admittatѕp); (lat. qu'il soit admis); billet d'admission.

Figure 3 : Nifon Bălăşescu, *Dicionariu româno-frances*. Volum I. Fascicul I. Bucuresci, Tipografia Mitropolitului Nifon, 1859.

Dans ce cas on remarque que l'auteur du dictionnaire a soumis neuf entrées à un autre mot-titre. Mais à cause de cela, ces neuf entrées manquent d'information morphologique, à une seule exception, et on a aussi des problèmes à les délimiter.

AFLUIREA REZERVIȘTILOR 30

AFLUIREA REZERVIȘTILOR. Deplasarea rezerviștilor la unitățile la care sînt repartizați potrivit planului de mobilizare a cetățenilor cu obligații militare. Poate fi : **directă**, cînd deplasarea se face din comune sau de la organizațiile socialiste la unități ; **prin puncte de adunare**, cînd cetățenii respectivi se grupează mai întîi în anumite locuri dinainte stabilite, de unde se trimit la unități.

AGENT DE TRANSMISIUNI. Persoană prin care se transmit diferite mesaje, verbale sau scrise, între diferiți comandanți (șefi) militari, realizînd, în acest fel, legătura între aceștia. Un **A.d.T.** poate fi militar sau luptător din formațiunile de apărare și acționează pe jos sau transportat (cu motocicleta, auto, avion, elicopter).

AGENȚIE DE CAMPANIE A BĂNCII NAȚIONALE A REPUBLICII SOCIALISTE ROMÂNIA. Instituție bancară organizată de Banca Națională, cu atribuții de executare de casă pentru unități și M.U., potrivit dispozițiilor tehnice primite de la filiala de campanie și dispozițiilor organului financiar al ordonatorului de credite pe lîngă care este repartizată.

AGRAFĂ. Material genistic din sîrmă curbată și ascuțită la ambele capete, folosit la prinderea sîrmei (ghimpate) pe pari sau țăruși, pentru realizarea gardurilor și rețelelor de sîrmă.

AGRESIUNE ARMATĂ. După textul „Definiției agresiunii armate" adoptat de Comitetul special al O.N.U. la 15 aprilie 1974 și însușit de Adunarea Generală în același an agresiunea este „folosirea forței armate de către un stat împotriva suveranității, integrității teritoriale sau independenței politice a unui stat sau în orice mod incompatibil cu Carta Națiunilor Unite". Statul care a recurs primul la forța armată, în contradicție cu Carta O.N.U., este calificat agresor. Se consideră **A.A.** acel act comis cu intenție, de o anumită gravitate, care se deosebește de așa-numitele „incidente minore" (de pildă, incidente sporadice de frontieră). Cazurile tipice de **A.A.** sînt : invadarea sau atacarea teritoriului unui stat, bombardarea, blocarea porturilor sau coastelor sale de către

Figure 4 : *Lexicon militar*. Bucureşti, Editura Militară, 1980.

Dans cette figure on remarque une autre spécificité lexicographique, c'est-à-dire qu'il n'y a pas un seul mot-titre, mais un syntagme, une construction fixe ou une expression. Dans ce cas il est impossible d'englober tous les mots dans une entrée et le lexicographe choisira le mot-titre pour pouvoir aligner ce dictionnaire de cette façon.

17 ALGOL

canonice sînt unice, ceea ce rezultă simplu din modul de construire al acestora. Formele canonice sînt importante pentru că permit să se precizeze, fără a fi necesară determinarea valorilor pentru toate combinațiile, dacă două expresii sînt echivalente (reprezintă aceeași funcție). Astfel, considerînd două expresii scrise în forma disjunctivă, pentru a putea vedea că ele sînt echivalente, se deduce mai întîi forma canonică disjunctivă cu ajutorul proprietăților expresiilor booleene. Dacă expresiile obținute conțin aceiași mintermeni, expresiile date sînt echivalente. Să utilizăm acest procedeu pentru a demonstra că expresiile $f(x, y, z)$, $E(x, y, z)$ sînt echivalente. Întrucît $E(x, y, z)$ constituie o formă canonică disjunctivă, este necesar să deducem forma canonică pentru $f(x, y, z)$; se obține expresia: $f(x, y, z) = xy + y\bar{z} + \bar{x}yz = xy \cdot 1 + y\bar{z} \cdot 1 + \bar{x}yz = xy(z + \bar{z}) + y\bar{z}(x + \bar{x}) + \bar{x}yz = xyz + xy\bar{z} + \bar{x}y\bar{z} + xy\bar{z} + \bar{x}yz = \bar{x}yz + \bar{x}y\bar{z} + xy\bar{z} + xyz$, care este echivalentă cu $E(x, y, z)$. Cîteva exemple de funcții de două variabile întîlnite frecvent în informatică sînt: $f_1 = x + y$, $f_2 = x \cdot y$, $f_3 = x \cdot y$, $f_4 = x + y$, $f_5 = x\bar{y} + \bar{x}y = x \oplus y$, $f_6 = xy + \bar{x}\bar{y}$. Funcțiile f_1 și f_2 reprezintă operațiile SAU, respectiv ȘI definite anterior. Funcția f_3 constituie operația ȘI-NU, iar f_4 operația SAU-NU (NICI). Funcțiile f_5, f_6 reprezintă respectiv operațiile SAU-EXCLUSIV (sumă modulo doi) și COINCIDENȚĂ. Din expresiile lui f_5 și f_6 rezultă că funcția f_5 este 1 dacă $x \neq y$, iar f_6 este 1 dacă $x = y$ (deci cînd valorile lui x și y coincid) și că f_5 și f_6 sînt complementare. Deoarece orice funcție booleană admite două forme canonice, exprimate cu ajutorul operațiilor ȘI, SAU, NU, rezultă că acestea pot fi utilizate pentru reprezentarea tuturor funcțiilor. Se numește set complet, sau funcțional complet, de operații mulțimea operațiilor ce permit reprezentarea oricărei funcții booleene; deci operațiile ȘI, SAU, NU formează un set complet. Dacă ținem seama de relațiile lui de Morgan, care permit obținerea operației SAU (ȘI) cu ajutorul operației ȘI(SAU) și al negației, rezultă că perechile SAU, NU și, respectiv ȘI, NU constituie, de asemenea, seturi complete. Întrucît funcțiile logice sînt generate cu ajutorul unor circuite al căror cost depinde de expresiile acestora, se urmărește obținerea unei expresii cît mai simple. Obținerea expresiei cu forma cea mai simplă este numită minimizare, existînd metode eficiente pentru minimizarea funcțiilor booleene. *(P.D.).*

algebră logică (engl.: boolean algebra), **algebră booleană.**

ALGOL (**ALGO**rithmic **L**anguage), limbaj de programare de nivel înalt destinat, în primul rînd, aplicațiilor cu caracter științific. Este primul limbaj de programare elaborat prin cooperare internațională și definitivat, în varianta **ALGOL**-60 în 1960, odată cu publicarea binecunoscutului document „Raport Revizuit asupra Limbajului Algoritmic **ALGOL**-60". **ALGOL**-60 este cunoscut sub trei forme: limbajul de referință, limbajul de lucru și limbajul de publicare. Limbajul de referință corespunde definiției din Raportul Revizuit. În această formă, limbajul

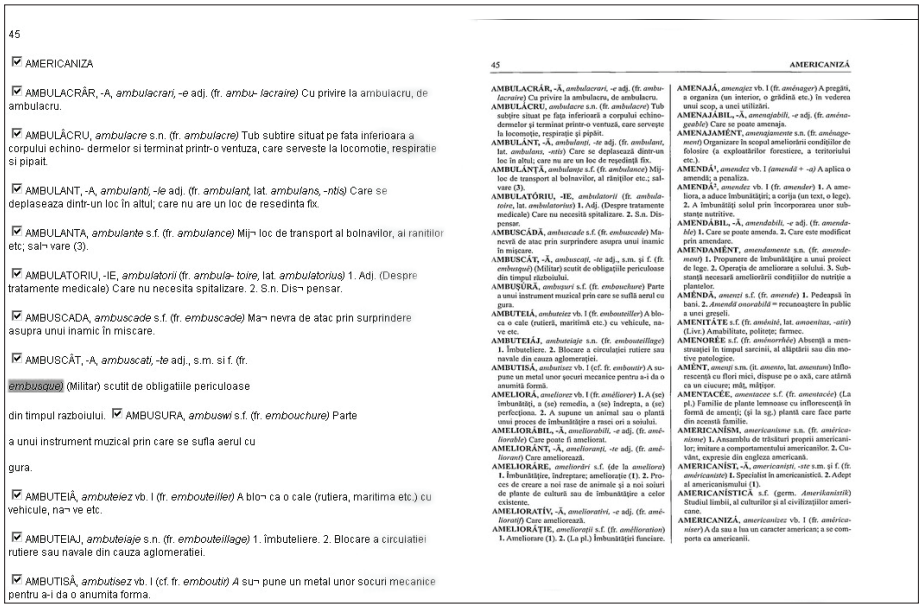
2

Figure 5: *Dicționar de informatică*. București, Editura Științifică și Enciclopedică, 1981.

Dans ce dictionnaire il y a une différenciation entre deux types d'entrées. La première est écrite avec des minuscules, ce qui dénote qu'on a une expression ou un syntagme qui utilise un mots-titre déjà défini au-dessus. Dans la deuxième situation, le mot-titre est utilisé aussi dans le texte de la définition de la même manière formelle. Vu que le mot-titre est écrit plusieurs fois, cela pose des problèmes lors de la délimitation des entrées, le logiciel pouvant les considérer comme des entrées distinctes.

3. La validation de la segmentation des dictionnaires

Ce processus a été réalisé par les lexicographes, mais également avec l'aide de quelques volontaires qui possédaient des connaissances philologiques assez solides.



45

☑ AMERICANIZA

☑ AMBULACRÂR, -A, *ambulacrari, -e* adj. (fr. *ambu- lacraire)* Cu privire la ambulacru, de ambulacru.

☑ AMBULÂCRU, *ambulacre* s.n. (fr. *ambulacre)* Tub subtire situat pe fata inferioara a corpului echino- dermelor si terminat printr-o ventuza, care serveste la locomotie, respiratie si pipait.

☑ AMBULANT, -A, *ambulanti, -le* adj. (fr. *ambulant,* lat. *ambulans, -ntis)* Care se deplaseaza dintr-un loc în altul; care nu are un loc de resedinta fix.

☑ AMBULANTA, *ambulante* s.f. (fr. *ambulance)* Mij¬ loc de transport al bolnavilor, al ranitilor etc; sal¬ vare (3).

☑ AMBULATORIU, -IE, *ambulatorii* (fr. *ambula- toire,* lat. *ambulatorius)* 1. Adj. (Despre tratamente medicale) Care nu necesita spitalizare. 2. S.n. Dis¬ pensar.

☑ AMBUSCADA, *ambuscade* s.f. (fr. *embuscade)* Ma¬ nevra de atac prin surprindere asupra unui inamic în miscare.

☑ AMBUSCÂT, -A, *ambuscati, -te* adj., s.m. si f. (fr.

embusqué) (Militar) scutit de obligatiile periculoase

din timpul razboiului. ☑ AMBUSURA, *ambuswi* s.f. (fr. *embouchure)* Parte

a unui instrument muzical prin care se sufla aerul cu

gura.

☑ AMBUTEIÂ, *ambuteiez* vb. I (fr. *embouteiller)* A blo¬ ca o cale (rutiera, maritima etc.) cu vehicule, na¬ ve etc.

☑ AMBUTEIAJ, *ambuteiaje* s.n. (fr. *embouteillage)* 1. Îmbuteliere. 2. Blocare a circulatiei rutiere sau navale din cauza aglomeratiei.

☑ AMBUTISÂ, *ambutisez* vb. I (cf. fr. *emboutir) A* su¬ pune un metal unor socuri mecanice pentru a-i da o anumita forma.

AMBULACRÁR, -Ă, *ambulacrari, -e* adj. (fr. *ambulacraire)* Cu privire la ambulacru, de ambulacru.
AMBULĂCRU, *ambulacre* s.n. (fr. *ambulacre)* Tub subțire situat pe fața inferioară a corpului echinodermelor și terminat printr-o ventuză, care servește la locomoție, respirație și pipăit.
AMBULÁNT, -Ă, *ambulanți, -te* adj. (fr. *ambulant,* lat. *ambulans, -antis*) Care se deplasează dintr-un loc în altul; care nu are un loc de reședință fix.
AMBULÁNȚĂ, *ambulanțe* s.f. (fr. *ambulance)* Mijloc de transport al bolnavilor, al răniților etc.; salvare (3).
AMBULATÓRIU, -IE, *ambulatorii (fr. ambulatoire,* lat. *ambulatorius)* **1.** Adj. (Despre tratamente medicale) Care nu necesită spitalizare. **2.** S.n. Dispensar.
AMBUSCÁDĂ, *ambuscade* s.f. (fr. *embuscade)* Manevră de atac prin surprindere asupra unui inamic în mișcare.
AMBUSCÁT, -Ă, *ambuscați, -te* adj., s.m. și f. (fr. *embusqué)* (Militar) scutit de obligațiile periculoase din timpul războiului.
AMBUȘÚRĂ, *ambușuri* s.f. (fr. *embouchure)* Parte a unui instrument muzical prin care se suflă aerul cu gura.
AMBUTEIÁ, *ambuteiez* vb. I (fr. *embouteiller*) A bloca o cale (rutieră, maritimă etc.) cu vehicule, nave etc.
AMBUTEIÁJ, *ambuteiaje* s.n. (fr. *embouteillage*) **1.** Îmbuteliere. **2.** Blocare a circulației rutiere sau navale din cauza aglomerației.
AMBUTISÁ, *ambutisez* vb. I (cf. fr. *emboutir*) A supune un metal unor șocuri mecanice pentru a-i da o anumită formă.
AMELIORÁ, *ameliorez* vb. I (fr. *améliorer*) **1.** A (se) îmbunătăți, a (se) remedia, a (se) îndrepta, a (se) perfecționa. **2.** A supune un animal sau o plantă unui proces de îmbunătățire a rasei ori a soiului.
AMELIORÁBIL, -Ă, *ameliorabili, -e* adj. (fr. *améliorable*) Care poate fi ameliorat.
AMELIORÁNT, -Ă, *amelioranți, -te* adj. (fr. *améliorant*) Care ameliorează.
AMELIORÁRE, *ameliorări* s.f. (de la *ameliora*) **1.** Îmbunătățire, îndreptare; ameliorație (1). **2.** Procedeu de creare a noi rase de animale și a noi soiuri de plante de cultură sau de îmbunătățire a celor existente.
AMELIORATÍV, -Ă, *ameliorativi, -e* adj. (fr. *amélioratif*) Care ameliorează.
AMELIORÁȚIE, *ameliorații* s.f. (fr. *amélioration*) **1.** Ameliorare (1). **2.** (La pl.) Îmbunătățiri funciare.
AMENAJÁ, *amenajez* vb. I (fr. *aménager*) A pregăti, a organiza (un interior, o grădină etc.) în vederea unui scop, a unei utilizări.
AMENAJÁBIL, -Ă, *amenajabili, -e* adj. (fr. *aménageable*) Care se poate amenaja.
AMENAJÁMÂNT, *amenajamente* s.n. (fr. *aménagement*) Organizare în scopul ameliorării condițiilor de folosire (a exploatărilor forestiere, a teritoriului etc.).
AMENDÁ[1], *amendez* vb. I (*amendă + -a*) A aplica o amendă; a penaliza.
AMENDÁ[2], *amendez* vb. I (fr. *amender*) **1.** A ameliora, a aduce îmbunătățiri; a corija (un text, o lege). **2.** A îmbunătăți solul prin încorporarea unor substanțe nutritive.
AMENDÁBIL, -Ă, *amendabili, -e* adj. (fr. *amendable*) **1.** Care se poate amenda. **2.** Care este modificat prin amendare.
AMENDAMÊNT, *amendamente* s.n. (fr. *amendement*) **1.** Propunere de îmbunătățire a unui proiect de lege. **2.** Operație de ameliorare a solului. **3.** Substanță necesară ameliorării condițiilor de nutriție a plantelor.
AMÊNDĂ, *amenzi* s.f. (fr. *amende*) **1.** Pedeapsă în bani. **2.** *Amendă onorabilă* = recunoaștere în public a unei greșeli.
AMENITÁTE s.f. (fr. *amènité,* lat. *amoenitas, -atis*) (Liter.) Amabilitate, politețe; farmec.
AMENOREÉ s.f. (fr. *aménorrhée*) Absență a menstruației în timpul sarcinii, al alăptării sau din motive patologice.
AMÊNT, *amenți* s.m. (lt. *amento,* lat. *amentum*) Inflorescență cu flori mici, dispuse pe o axă, care atârnă ca un ciucure; mâț, mițișor.
AMENTACÉE, *amentacee* s.f. (fr. *amentacée*) (La pl.) Familie de plante lemnoase cu inflorescență în formă de amenți; (și la sg.) plantă care face parte din această familie.
AMERICANISM, *americanisme* s.n. (fr. *américanisme*) **1.** Ansamblu de trăsături proprii americanilor; imitare a comportamentului americanilor. **2.** Cuvânt, expresie din engleza americană.
AMERICANÍST, -Ă, *americaniști, -ste* s.m. și f. (fr. *américaniste*) **1.** Specialist în americanistică. **2.** Adept al americanismului (1).
AMERICANÍSTICĂ s.f. (germ. *Amerikanistik*) Studiul limbii, al culturilor și al civilizațiilor americane.
AMERICANIZÁ, *americanizez* vb. I (fr. *américaniser*) A da sau a lua un caracter american; a se comporta ca americanii.

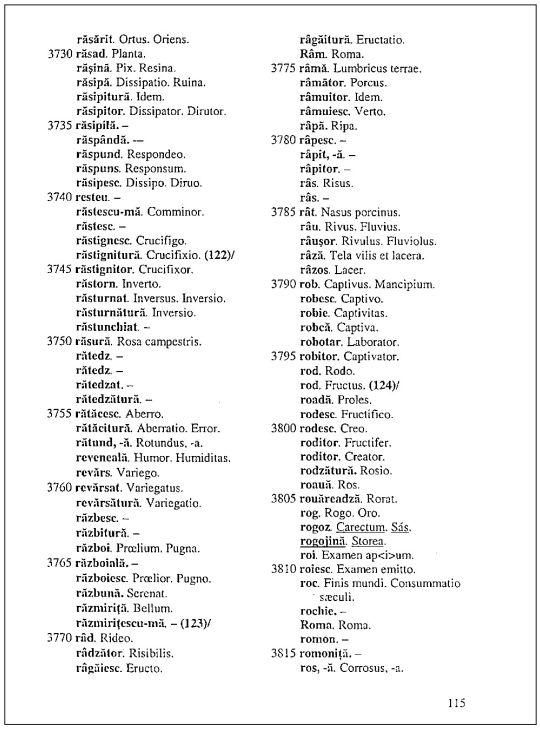
Figure 6 : La validation de la segmentation en entrées d'une page d'un dictionnaire

Dans cette étape, les informaticiens ont créé un logiciel qui essaie de reconnaître chaque entrée d'après ses caractéristiques formelles. Ce logiciel signale les entrées qui sont considérées comme véritables, mais il souligne également les autres mots qui peuvent être envisagés comme des entrées. Le lexicographe doit valider chaque entrée à son tour. Par cette démarche on valide les entrées correctes, mais, en même temps, chaque entrée est corrigée pour pouvoir les aligner facilement. Pour chaque page validée, le lexicographe a accès également à la page numérisée, pour pouvoir corriger les erreurs. S'il y a des mots qui ne peuvent pas être validés comme des entrées, alors le lexicographe peut choisir de valider partiellement la page respective. Après ça, l'informaticien traitera de nouveau cette page pour reconnaître les entrées restantes.

4. La segmentation des entrées de dictionnaire

Un autre élément distinct et définitoire pour le projet CLRE c'est le fait qu'il met à disposition des utilisateurs la variante numérisée du dictionnaire, pour pouvoir vérifier chaque entrée. Ainsi, chaque entrée est segmentée de façon automatique après l'opération de validation réalisée par les lexicographes.

Mais l'exactitude de la segmentation est liée, bien sûr, au format des dictionnaires. Et pour cela les informaticiens ont dû adapter leur logiciel pour toute une variété des formats. En général, le validateur statistique utilise les suivants éléments pour faire la segmentation : l'alignement à gauche et à droite du texte et la police (le style de la police– italique, souligné, gras –) et les types des caractères (majuscules, minuscules).

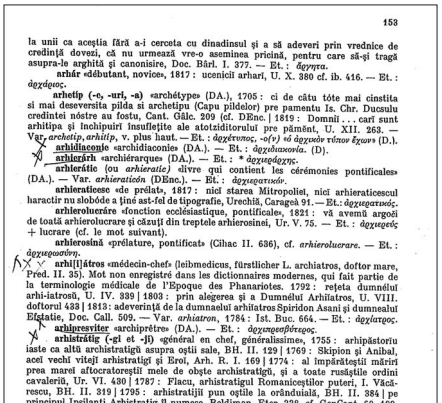


răsărit. Ortus. Oriens.
3730 **răsad**. Planta.
răşină. Pix. Resina.
răsipă. Dissipatio. Ruina.
răsipitură. Idem.
răsipitor. Dissipator. Dirutor.
3735 **răsipită**. –
răspândă. –
răspund. Respondeo.
răspuns. Responsum.
răsipesc. Dissipo. Diruo.
3740 **resteu**. –
răstescu-mă. Comminor.
răstesc. –
răstignesc. Crucifigo.
răstignitură. Crucifixio. **(122)**/
3745 **răstignitor**. Crucifixor.
răstorn. Inverto.
răsturnat. Inversus. Inversio.
răsturnătură. Inversio.
răstunchiat. –
3750 **răsură**. Rosa campestris.
rătedz. –
rătedz. –
rătedzat. –
rătedzătură. –
3755 **rătăcesc**. Aberro.
rătăciturâ. Aberratio. Error.
rătund, -ă. Rotundus, -a.
revenealâ. Humor. Humiditas.
revârs. Variego.
3760 **revârsat**. Variegatus.
revârsătură. Variegatio.
răzbesc. –
răzbiturâ. –
război. Prœlium. Pugna.
3765 **războială**. –
războiesc. Prœlior. Pugno.
răzbună. Serenat.
răzmiriţă. Bellum.
răzmiriţescu-mă. – **(123)**/
3770 **râd**. Rideo.
rădzător. Risibilis.
răgâiesc. Eructo.
râgâiturâ. Eructatio.
Râm. Roma.
3775 **râmă**. Lumbricus terrae.
râmător. Porcus.
râmuitor. Idem.
râmuiesc. Verto.
râpă. Ripa.
3780 **râpesc**. –
râpit, -ă. –
râpitor. –
râs. Risus.
râs. –
3785 **rât**. Nasus porcinus.
râu. Rivus. Fluvius.
râuşor. Rivulus. Fluviolus.
râză. Tela vilis et lacera.
râzos. Lacer.
3790 **rob**. Captivus. Mancipium.
robesc. Captivo.
robie. Captivitas.
robcă. Captiva.
robotar. Laborator.
3795 **robitor**. Captivator.
rod. Rodo.
rod. Fructus. **(124)**/
roadă. Proles.
rodesc. Fructifico.
3800 **rodesc**. Creo.
roditor. Fructifer.
roditor. Creator.
rodzătură. Rosio.
roauă. Ros.
3805 **rouăreadză**. Rorat.
rog. Rogo. Oro.
rogoz. Carectum. Sás.
rogoĭină. Storea.
roi. Examen ap<i>um.
3810 **roiesc**. Examen emitto.
roc. Finis mundi. Consummatio sæculi.
rochie. –
Roma. Roma.
romon. –
3815 **romoniţă**. –
ros, -ă. Corrosus, -a.

115

Figure 7 : *Dictionarium Valachico-Latinum. Primul dicţionar al limbii române*. Studiu introductiv, ediţie, indici şi glosar de Gh. Chivu. Bucureşti, Editura Academiei Române, 2008 [cca 1650]

Dans ce dictionnaire, considéré comme le premier dictionnaire de la langue roumaine, l'auteur a réalisé aussi une numérotation des mots introduits et a marqué cela en faisant une note tous les cinq mots et en inscrivant le nombre correspondant. Le résultat est que le logiciel de segmentation reconnaît les chiffres, et non pas les mots, comme étant les entrées.



153

la unii ca aceştia fără a-i cerceta cu dinadinsul şi a să adeveri prin vrednice de credinţă dovezi, că nu urmează vre-o aseminea pricină, pentru care să-şi tragă asupra-le arghită şi canonisire, Doc. Bâri. I. 377. — Et. : ἄρχυρα.

arhár «débutant, novice», 1817 : ucenicii arhari, U. X. 380 cf. ib. 416. — Et. : ἀρχάριος.

arhetip (-o, -uri, -a) «archétype» (DA.), 1705 : ci de căţu tóte mai cinstita si mai deserverita pilda si archetipu (Capu pildelor) pre pamentu Is. Chr. Ducsulu credintei nóstre au fostu, Cant. Gâlc. 209 (cf. DEnc. | 1819 : Domnii . . . carî sunt arhitipa şi închipuiri însufleţite ale atotziditoruluí pre pământ, U. XII. 263. — Var. *archetip, arhiti-p*, v. plus haut. — Et. : *ἀρχέτυπος, -ον(ν) τό ἀρχικόν τίπτον ἔχων* (D.).

arhidiaconíe «archidiaconie» (DA.). — Et. : *ἀρχιδιακονία*. (D).

arhierárh «archiérarque» (DA.). — Et. : * *ἀρχιερέαρχης*.

arhierátic (ou *arhieratic*) «livre qui contient les cérémonies pontificales» (DA.). — Var. *arhieraticón* (DEnc.). — Et. : *ἀρχιερατικόν*.

arhieraticéscе «de prêtre», 1817 : nici starea Mitropoliei, nici arhieraticescul haractir nu slobóde a ţine ast-fel de tipografie, Urechiă, Caragea 91. — Et.: *ἀρχιερατικός*.

arhieroluerâre «fonction ecclésiastique, pontificale», 1821 : vă avemŭ argoši de toată arhieroluerare şi căzuţi din treptele arhierosinei, Ur. V. 75. — Et. : *ἀρχιερεύς* + lucrare (cf. le mot suivant).

arhierosînă «prêlature, pontificat» (Cihac II. 636), cf. *arhierolucrare*. — Et. : *ἀρχιερωσύνη*.

arhi[i]átros «médecin-chef» (leibmedicus, fürstlicher L. archiatros, doftor mare, Pred. II. 35). Mot non enregistré dans les dictionnaires modernes, qui fait partie de la terminologie médicale de l'Époque des Phanariotes. 1792 : reţeta dumnélui arhi-iatrosŭ, U. IV. 339 | 1803 : prin alegerea şi a Dumnélui Arhiiatros, U. VIII. doftorul 433 | 1813 : adeverinţă de la dumnealui arhiiatros Spiridon Asani şi dumnealui Efstatie, Doc. Call. 509. — Var. *arhiatron*, 1784 : Ist. Buc. 664. — Et. : *ἀρχίατρος*.

archiereíter «archiprêtre» (DA.). — Et. : *ἀρχιερατεῖτερος*.

arhistrátig (-gi et -ji) «général en chef, généralissime», 1755 : arhipăstoriu iaste ca altŭ archistratigŭ asupra oştii sale, BH. II. 129 | 1769 : Skipion şi Aníbal, acei vechi viteji arhistratigi şi Eroi, Arh. R. I. 169 | 1774 : al împărăteştii măriri prea marei aflocratoreştii mele de obşte archistratigŭ, şi a toate rusăştile ordini cavalerieŭ, Ur. VI. 430 | 1787 : Flacu, arhistratigul Romanicestilor puteri, I. Văcărescu, BH. II. 319 | 1795 : arhistratiji pun oştile la crânduială, BH. II. 384 | pe principul Ipsilanti Arhistratig II numesc, Beldiman, Eter. 338, cf. GenCant. 60, 199, 273, etc. Pour un ex. dans Asachi cf. Haneş, L. lit. 46. — Et. : *ἀρχιστράτηγος*. Au XVII[e] s. le même mot, venu par la filière du slave ecclésiastique, a désigné les chefs des anges, les archanges (cf. DA.). La var. *arhistratég* est un néologisme d'origine française (= archistratège). Cf. aroum. *arhistatigó (a. Mihail*, Geagea 314).

arhitécton architecte, 1828 : făcăndusе un tahmin de către un architecton, Ur. VII. 171 | 1837 : şi spre siguranţie am dat chizăş pentru noi pe dumnealui architectonul, Iorga, Scris. Meşt. 94. A la même époque on a aussi *arhitector*, 1813 : va fi îngrijirea acésta asupra Domniei Sale vel Vornic de politie alcătuinduse şi arhitectorŭ al oraşului, Ur. I. 198, cf. 1824 : Bibl. R. IV. 97. Cette dernière forme est la relatinisation du terme italien vieilli *architettore* (cf. DA.). Déjà en 1793 on a *arhitect* (= all. Architekt, fr. *architecte*, cf. Pred. I. 206). — Et. : *ἀρχιτέκτων*.

arhitectonie «architecture» (DA.). — Et. : *ἀρχιτεκτονία*.

arhizugráv chef des peintres», 1787 : să mai adoge obligaţiunea, ca atunci când vre-unulŭ din zugravii cei învăţaţi desévârşiţi vor locui unŭ lucru din cele mai mari, archi-zugravulŭ să aibă a rândui lângă acelŭ zugrav învăţatŭ şi calfă, séŭ ucenici din cei mai proşti. U. III. 601. — Et. : *ἀρχιζωγράφος*. L'alternance *zugrav* (Murnu) ~ *zograv* montre nettement qu'il s'agit d'un emprunt récent où la voyelle radicale s'est mieux conservée, cf. aroum, *zuyráf* (-*hi*, Dal.).

Figure 8 : *Gáldi Ladislas, Les mots d'origine néogrecque en roumain à l'époque des Phanariotes*. Budapest, 1939.

Il y a aussi des dictionnaires, en particulier des dictionnaires anciens, qui ne sont pas faciles à trouver et qui en plus présentent de « traces » d'utilisation. Dans ce cas, c'est presque impossible, pour le logiciel, de reconnaître une entrée, si devant l'entrée il y a toute sorte de notations faites par de divers utilisateurs. Dans ce cas, c'est le validateur humain qui décide où commence l'entrée et qui corrige les éventuelles erreurs.

Les dictionnaires spécialisés posent de divers problèmes ayant trait à leur spécificité. Il y a toute sorte de marques d'orthographe, de morphologie et bien sûr, l'accent.

Après la validation, on essaye d'améliorer les résultats du logiciel de reconnaissance optique des caractères en passant par un nouveau processus de reconnaissance des caractères des dictionnaires analysés.

Les deux dernières étapes du projet ont envisagé l'alignement des entrées des dictionnaires et les possibilités d'interroger cette base de données.

En ce qui concerne l'alignement des dictionnaires, après avoir fini le processus de segmentation des entrées à l'aide de l'interface spécialement créée, on a réalisé un alignement primaire de tous les mots des dictionnaires et ensuite fait un alignement avec l'eDTLR[1]. Les entrées alignées sont incluses dans la base de données qui

[1] Le projet national *eDTLR. Le Dictionnaire (Trésor) de la Langue Roumaine en format électronique* a été financé par le Conseil National du Management des Projets (CNMP), pour la période 2007-2010 et il représente l'un de plus importants projets de type collaboratif des

permettra faire des recherches ultérieures, en utilisant des divers critères. Ainsi, on peut chercher dans un seul dictionnaire ou dans tous les dictionnaires introduites en CLRE, mais on peut effectuer également une recherche selon d'autres critères : on peut chercher les mots commençant avec la même lettre ou qui incluent un certain groupe de lettres. Le résultat de la recherche permet la visualisation de l'entrée du dictionnaire, tout d'abord en format image. Sur la même page, on peut voir aussi la variante reconnue par le logiciel ABBYY, mais elle ne peut pas être utilisée dans le cas des dictionnaires en alphabet cyrillique ou de transition. L'utilisateur peut voir également les informations concernant le dictionnaire qui a servi de source pour le mot en question.

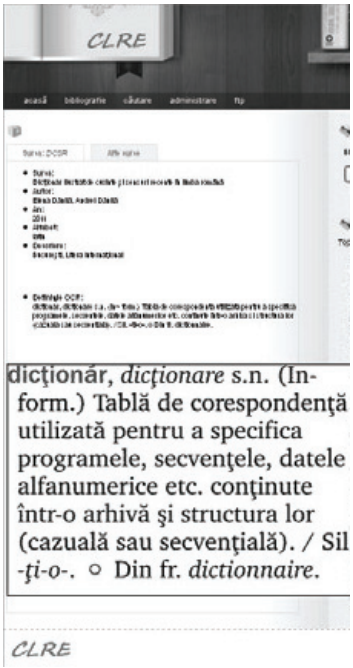


Figure 9 : Une capture d'écran de l'article « dictionnaire » d'un lexique écrit en alphabet latin.

dernières années, qui entremêle l'expérience des lexicographes et des informaticiens pour créer des outils électroniques. Le projet a été coordonné par dr. Dan Cristea de la Faculté d'Informatique de l'Université « Alexandru Ioan Cuza » de Iasi.

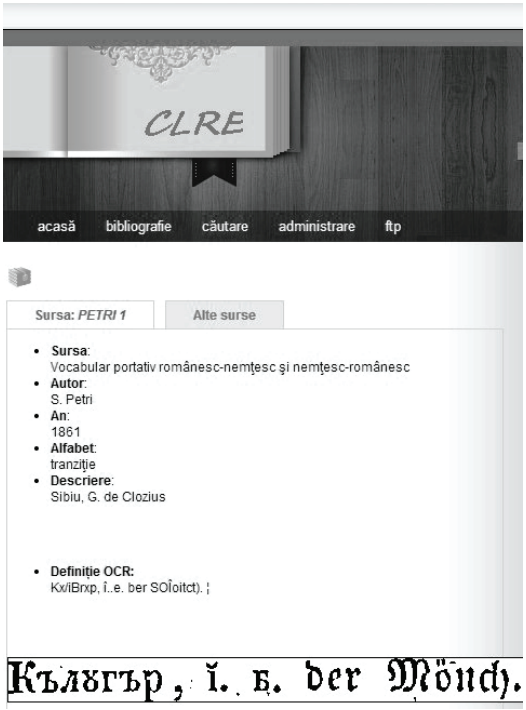


Figure 10 : Une capture d'écran de l'article « moine » d'un dictionnaire écrit en alphabet cyrillique.

5. Conclusions

Tant le spécialiste que l'utilisateur non initié peuvent consulter virtuellement, à l'aide du CLRE, une multitude de dictionnaires qui, autrement, ne seraient disponibles que dans les bibliothèques.

Ce projet est lié avec d'autres projets en ligne réalisés en Roumanie (eDTLR et la troisième édition du dictionnaire RDW[3], version informatisée) mais également avec le processus de réalisation du *Dictionnaire de la Langue Roumaine*, soutenu par l'Académie Roumaine. Il y aussi une liaison entre CLRE et d'autres projets internationaux : par exemple avec DÉRom[2] (le CLRE étant considéré comme source de documentation) ou avec *ENeL : European Network of e-Lexicography* qui est une action de la ligne de financement COST et qui cherche à créer un réseau paneuropéen des ressources lexicographiques pour l'usage académique et pour le public. Ces collaborations servent à confirmer l'importance de cette ressource électronique pour la langue roumaine.

[2] *Dictionnaire Étymologique Roman* (DÉRom) (www.atilf.fr/DERom/).

L'information accessible par l'intermédiaire du CLRE peut être comparée avec celle offerte du modèle des corpus lexicographiques crées pour d'autres langues : *Le rayon des dictionnaires*, http : //www.atilf.fr/ – collection de dictionnaires informatisés français, du XVI[e] jusqu'au XX[e] siècle ou *Das Wörterbuchnetz* ; *Nuevo tesoro lexicográfico de la lengua española*, http : //buscon.rae.es/ntlle/SrvltGUILoginNtlle – base de données qui inclut les versions facsimilées de tous les dictionnaires édités et publiés par la Real Academia Española ; http : //germazope.uni-trier.de/Projects/WBB/_– réseau de dictionnaires de langue allemande, créée à l'Université Trier d'Allemagne, etc.

De cette façon, CLRE offre aux spécialistes intéressés par l'étude de la langue roumaine, aux romanistes, mais aussi aux spécialistes informaticiens qui désirent analyser le traitement du langage naturel à partir de la langue roumaine, la désambiguïsation sémantique, etc. – un instrument de travail extrêmement utile.

L'Académie Roumaine, Filiale de Iasi — Marius-Radu CLIM
L'Académie Roumaine, Filiale de Iasi — Elena TAMBA,
L'Institut de Philologie Roumaine « A. Philippide » — Ana CATANĂ-SPENCHIU
L'Institut de Philologie Roumaine « A. Philippide » — Mădălin PĂTRAŞCU

Références bibliographiques

Atkins, B.T.S., Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford, Oxford University Press.

Clim, Marius / Dănilă, Elena / Haja, Gabriela, 2008. « Premise ale informatizării cercetării lexicografice academice româneşti », in : *Limba română. Dinamica limbii, dinamica interpretării*, Bucureşti, Editura Universităţii din Bucureşti, 585-591.

Dănilă, Elena, 2013. «Corpus lexicographique roumain essentiel. Les dictionnaires de la langue roumaine alignés au niveau de l'entrée», in : Herrero, Casanova / Calvo Rigual, Cesareo (ed.), 2013. *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas 6-11 septiembre 2010, Valencia* (6-11 septembre 2010), Berlin / New York, La maison d'édition Walter de Gruyter, volum VIII, 125-134.

Dănilă, Elena / Clim, Marius-Radu / Catană-Spenchiu, Ana, 2011. «Towards a Romanian Lexicographic Corpus», *Philologica Jassyensia*, An VII, Nr. 2 (14), 191-198.

Tamba, Dănilă Elena / Clim, Marius-Radu / Pătraşcu, Mădălin / Catană-Spenchiu, Ana, 2012. «The Evolution of the Romanian Digitalized Lexicography. The Essential Romanian Lexicographic Corpus», in : Fjeld, Ruth Vatvedt, Torjusen, Julie Matilde, (ed.), 2012. *Proceedings of the 15 th EURALEX International Congress*, 7-11 august 2012, Oslo, Press Reprosentrales, UiO, 1014-1017 ; on-line : http : //www.euralex.org/proceedings-toc/euralex_2012/ .

La realizzazione del *Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan (DiTMAO)*: problemi di organizzazione della conoscenza medico-farmaceutica attestata nei manoscritti in occitano antico

1. Introduzione

Come ho illustrato nei precedenti convegni CILPR in collaborazione con il collega Guido Mensching (Corradini / Mensching 2010 e Corradini / Mensching 2013), al lavoro di carattere critico-editoriale delle fonti manoscritte in occitano medievale che trattano argomenti medico-farmaceutici abbiamo recentemente aggiunto un'attività che, mediante una conveniente organizzazione dei dati, permetta la fruizione del ricco ed eterogeneo materiale isolato all'interno del corpus.

Il progetto, che è in corso di realizzazione nei centri di Colonia, di Gottinga e di Pisa, è stato finanziato dalla DFG[1]. Si tratta dell'elaborazione di un sistema lessicografico sul Web nel quale, se da un lato il lemma mantiene il tradizionale ruolo ai fini della consultazione dell'archivio testuale (e questa modalità costituirà la base del dizionario che sarà in seguito prodotto su stampa), dall'altro il valore semantico di ciascun lemma è strutturato in maniera appropriata per offrire un'ulteriore e, come vedremo, utile chiave di accesso ai dati.

Per raggiungere tale obiettivo è necessario eseguire una puntuale descrizione del dominio in oggetto, e a tal fine si è ritenuto utile adottare un approccio ontologico, per la corretta strutturazione e applicazione del quale abbiamo usufruito delle competenze informatiche che sono state sviluppate presso l'Istituto di Linguistica Computazionale di Pisa (ILC-CNR).

---

[1] Si tratta del progetto *An XML-based Information System for Old Occitan Medical Terminology* finanziato dalla DFG (Deutsche Forschungsgemeinschaft). Equipe: Università di Colonia: Prof. Dr. Gerrit Bos, Veronica Roth; Università Georg August di Gottinga: Prof. Dr. Guido Mensching, Julia Zwink, Anja Weingart, Danielle Friedrich; Università di Pisa: Prof. M.Sofia Corradini, Erminio Maraia; ILC-CNR di Pisa: Prof. Andrea Bozzi, Emiliano Giovannetti.

2. Il corpus testuale

Il corpus di referenza è già stato puntualmente descritto nei contributi citati sopra, e per esso si rimanda anche all'Appendice in calce e al poster «Les termes romanes en graphie hébraïque pour le DiTMAO» presentato in questo medesimo congresso CILPR da Roth, Weingart, Zwink.

E' indispensabile, tuttavia, sottolineare ancora una volta:

(a) che si tratta di fonti di tipologia eterogenea, sia a causa del genere testuale, sia a causa del differente alfabeto (latino ed ebraico) mediante il quale sono veicolati i testi;

(b) che la documentazione sulla quale è effettuato lo spoglio delle voci si fonda su un rigore filologico. Essa è rappresentata, infatti, da edizioni condotte a partire dalle fonti manoscritte, per la maggior parte prodotte a nostra cura[2]; di ogni entrata lessicale, inoltre, sono state prese in considerazione tutte le varianti attestate dalla tradizione di ciascun testo oltre che, naturalmente, le varianti che un termine può presentare nelle differenti opere scritte in entrambi gli alfabeti, in relazione a contesti differenti (ad es. morfologici, di ambito dialettale, etc.). Esemplificano la prima situazione:
   - le differenti lezioni per "pleurite" dei due codici che trasmettono l'*Anatomia Porci*: *pleuresis* (ms. B) e *pleverzis* (ms. C) (Corradini 2006, 480 e 489);
   - le differenti lezioni per "abrotano" di due dei manoscritti che contengono il Sefer ha-Shimmush, libro 29: 'BRWṬNWM (*abrotonum*), ShS1, Shin 3, ms. O e 'BRWṬ'NWM (*abrotanum*), ShS1, Shin 3, ms. V) (Bos / Hussein / Mensching / Savelsberg 2011, 496-497).

Per le varianti legate a diversi ambiti geografici sono significative, per esempio, le realizzazioni di un medesimo termine relative, rispettivamente, all'area pirenaica e a quella linguadociana, com'è il caso delle forme indicanti le "prugnole acerbe": *aranhons* (Thes XXIX 42) e 'GRYNŠ (**agrenas*) (ShS2 Ox.Add.22, Alef 10: Corradini 1997, 89 e 283 e Bos / Hussein / Mensching / Savelsberg in prep.).

3. Strutturazione del sistema

Al di là dei principi fondamentali di carattere linguistico e filologico che sono alla base di tutto il progetto, è opportuno descrivere qui nei dettagli il sistema lessicografico adottato, anche in considerazione delle novità introdotte. Occorre, inoltre, esporre i problemi con i quali ci siamo dovuti misurare in conseguenza della eterogeneità linguistica delle fonti prese in esame, ed anche sottolineare i vantaggi che talvolta provengono proprio dalla loro comparazione. Nei precedenti contributi sopra citati sono già stati descritti alcuni aspetti ineludibili quando ci si accinga a produrre un dizionario basato sullo spoglio di un corpus complesso come quello del DiTMAO. Ciò comporta inevitabilmente, per esempio, difficoltà di interpretazione semantica originata dall'evoluzione dei significati o, a monte di essa, incertezze di

[2] In una seconda fase di avanzamento del lavoro sarà preso in considerazione anche il lessico estratto da edizioni esistenti, qualora esse siano state condotte secondo rigorosi criteri filologici.

lettura sovente dovute a problemi di natura paleografica. Si pensi, per esempio, al termine "istmo": mentre nelle lingue moderne esso possiede, in anatomia, il significato generico di «nom qu'on donne à certaines parties étroites du corps humain» (FEW (4, 822b, s.v. *isthmus),* in a.occ è utilizzato per indicare più precisamente lo "spazio che è compreso fra trachea ed esofago", come attestano le forme *hismon* (Anat.Por, ms. B) e *yssmol* (Anat.Por, ms. C) (Corradini 2006, 479 e 488; Corradini / Mensching 2013, 117-118). Un altro caso è quello della sequenza *a cosarce coll(um/am) que es dit a(n) gelot* (ms. C, f. 4r) che pone di fronte ai problemi di una corretta lettura e del controllo della corrispondenza sinonimica. Mentre il primo è stato risolto mediante comparazione interna e con la fonte latina, tanto da poter proporre l'interpretazione *a[loe], sarcecoll(am) que es dit a(n)gelot,* è proprio dal raffronto con la lista di sinonimi in alfabeto ebraico estratti dal *Sefer ha-Shimmush* (Alef 40) che è stato possibile verificare l'attendibilità della relazione fra *sarcecoll(am* e *a(n)gelot* (Corradini 1997, 263; Corradini / Mensching 2010, 203; Bos / Hussein / Mensching / Savelsberg 2011, 121).

Dal punto di vista pratico, gli aspetti problematici concernono la scelta dei criteri più adatti a organizzare e a rendere interrogabili tutti gli elementi del patrimonio lessicale in oggetto, sia quelli relativi al significante (variazione grafica, fonetica, morfologica, etc.), sia quelli collegati al senso. In quest'ottica, occorre precisare che rivestono una particolare importanza i legami logici che intercorrono fra i significati di ciascuna entità, e la natura di tali legami. Si tenga presente, inoltre, che il dizionario è concepito in funzione di un bacino di utilizzatori potenzialmente molto vasto; essi, in relazione ai propri ambiti di competenza, hanno esigenze diversificate, ora di tipo linguistico, ora filologico, ora storico-culturale o storico-scientifico, oppure riferibili ad una combinazione fra essi.

In seguito a tali considerazioni e per rendere possibile la presentazione dei dati in maniera organica e formale, si è ritenuto indispensabile arricchire la componente costituita, come da prassi, dall'ordinamento alfabetico, mediante una distribuzione degli elementi grammaticali basata su una classificazione tassonomica dei ruoli che le forme linguistiche ricoprono, soprattutto quando di esse sono attestate varianti nei due differenti alfabeti. Queste situazioni, delle quali quella citata delle varianti grafiche costituisce solo uno dei casi meno problematici, hanno potuto essere meglio evidenziate poiché le singole unità (forme e lemmi) sono state distribuite in una struttura omogenea ed esplicita, la cui organizzazione è mutuata dai cosiddetti sistemi ontologici di dominio, molto noti in ambito informatico e del Web semantico.

Parallelamente, per quanto riguarda il versante semantico, è stata concepita una seconda struttura (schema onomasiologico), ancora fondata sui principi dell'ontologia, e condotta in accordo con le classificazioni e le tassonomie delle scienze che vi sono rappresentate (medicina, botanica, etc.). Le due strutture, come vedremo, sono correlate l'una all'altra e, dopo essere state popolate[3] con le unità linguistiche appar-

[3] "Popolare/popolazione" sono termini adottati dagli informatici e si riferiscono al processo di classificazione delle singole unità secondo i valori dello schema ontologico che descrive la conoscenza strutturata del dominio al quale esse appartengono.

tenenti al dominio medico-farmaceutico considerato, sono in grado di rendere possibili interrogazioni incrociate e selettive.

In questo processo di organizzazione sono stati affrontati nell'insieme due aspetti che risultano strettamente collegati:

- il primo riguarda la definizione delle classi e delle sottoclassi da considerare, l'individuazione dei reciproci rapporti e le proprietà di ciascuna di esse, cioè la descrizione degli schemi secondo i quali le informazioni sono organizzate sia nel versante grammaticale che in quello semantico o, più precisamente, concettuale;
- il secondo aspetto è relativo alla conversione delle informazioni in un sistema informatico capace di rappresentare al meglio l'insieme dei dati.

Descriverò di seguito alcuni dei caratteri più importanti relativi alla definizione delle classi e delle sottoclassi nel versante grammaticale[4]; per quanto riguarda l'aspetto semantico e lo sviluppo del sistema informatico rinvio alla comunicazione presentata in questa stessa sessione da Bozzi e Luzzi[5].

3.1. Il significante: classi e sottoclassi e relative proprietà

3.1.1. Nello schema grammaticale si è definita una classe 'mot vedette'[6], coincidente con la terminologia medico-botanica dell'antico occitano, che possiede un certo numero di attributi, o proprietà. Alcune di esse non si discostano da quelle che tradizionalmente accompagnano le entrate dei lessici (traduzione, categoria grammaticale, numero, genere, etc.), anche se al riguardo si impongono alcune precisazioni. Per esempio, quando la forma di un lemma scelta come entrata lessicale[7] del dizionario non è attestata nel corpus testuale, è prevista la possibilità di segnalarla mediante un opportuno indicatore. Inoltre, all'interno delle categorie grammaticali è stata aggiunta quella di 'syntagme', etichetta utilizzata per indicare espressioni da considerarsi come forme lessicali semplici (per es. *acara bacaras, agnus castus, blacte bisantie,* etc.).

D'altronde, altre proprietà che abbiamo attribuito alla classe 'mot vedette' sono specifiche come:

- 'alphabet', la cui indicazione è necessaria per ciascuna entrata a causa della natura del corpus, caratterizzato da testi veicolati da due alfabeti differenti (latino ed ebraico)[8];

---

4 La tassonomia è organizzata nella struttura ontologica espressa in forma pressoché definitiva nella fig. 1 dell'Appendice.

5 Bozzi / Luzzi, «Un'ontologia per il DiTMAO (*Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan)*».

6 Si precisa che la lingua di redazione del lessico è il francese, ed in tale lingua, di conseguenza, è espressa la terminologia descrittiva.

7 A questo proposito si precisa che ai fini della lemmatizzazione si è scelta di privilegiare, quando presente, la variante più vicina all'etimologia del termine in questione; sono state opportunamente segnalate, inoltre, le informazioni relative a lemmi polisemici e a lemmi omografi.

8 I termini in alfabeto ebraico compaiono nel lessico affiancati dalla traslitterazione in caratteri latini secondo la convenzione utilizzata nella *Encyclopaedia Judaica* (16 vol., Jerusalem / New

- 'langue', proprietà che è indispensabile introdurre perché esistono forme attestate nel corpus che, pur non appartenenti al sistema linguistico occitano, è necessario accogliere nel vocabolario in quanto di uso comune nella tipologia dei testi considerati. Si tratta di termini latini (per es. *amidum, balsamita, calamus aromaticum*), catalani (per es. *anpressecs,* di fronte ad occ. *persega, pressega; panicalt* di fronte a occ. *panicaut*) oppure di voci ibride latino-occitane, rappresentate sovente da forme di latino scorretto, compresi i casi in cui di questa lingua viene conservato il morfema flessionale (per es. *sitirions*, di fronte a lt. *satyrion*; *simbra*, di fronte a lt. *sisimbrium* oppure *polvera musci, olibani, camphore, feniculi, gumi arabicum,* etc.) (Corradini 2001, 179-180; Mensching / Savelsberg 2004, 75);
- 'références', che rappresentano il collegamento fra lemmi e testi, tramite i contesti ove il termine in questione compare.

Alcune proprietà del lemma (che, in una concezione gerarchica della struttura lessicografica, ne rappresenta il 1° livello) costituiscono a loro volta delle sottoclassi (2° livello) e possiedono proprietà specifiche, come 'variantes', 'sous lemmes', 'synonymes', 'voir aussi'; alle proprietà di ciascuna sottoclasse si accede, nella versione digitale, mediante l'indicatore (→)[9]. A titolo di esempio si riportano i dati di popolamento dello schema relativo alla parola *aloe*:

Mot vedette :	aloe
• traduction :	aloès
• alphabet:	lt.
• langue:	a. occ.
• catégorie gram:	n.
• genre:	m.
• nombre:	s.
• variantes (→) :	'LW'Y (aloe), 'LW'N (aloen)
• nom scientifique :	Aloe vera L.
• références :	Ric 65 (P); Let2 55(A); Thes VII.30 (C); Ag. thes IV.1 (C)
• usage (date/période) :	XII^{ème}
• correspondances:	aram. 'YLWW', 'LWW'; ar. ṢBR ; a. cat. aloe(n) // gr. ἀλόη, lt. m. aloe
• lien ontologique :	botanique
• bibliographie :	Ric (P); Let2 (A); Thes; Ag.thes; ShS1 Alef 23 // Sin 227; RL1:57b; FEW 24:345b
• sous lemmes (→) :	lin aloe
• synonymes (→) :	-
• voir aussi (→) :	-
• observations :	-

York, 1971-1972), e dalla trascrizione coerente con le norme del sistema linguistico occitano. Per tali aspetti si veda Bos / Hussein / Mensching / Savelsberg (2011) 4-5 e 47-52. Nel presente contributo sono utilizzate solamente le forme in traslitterazione e in trascrizione.

[9] La visualizzazione delle informazioni si ottiene selezionando il simbolo della freccia. Nella versione cartacea queste specifiche saranno visualizzate nella pagina nella forma richiesta dalle modalità di stampa perché i dati dell'archivio informatico sono rappresentati in linguaggio XML che, opportunamente interpretato, consente qualunque tipo di *mise en page*.

3.1.2. All'interno delle 'sottoclassi', quella definita 'variantes' riveste un'importanza fondamentale ai fini della registrazione di tutta la variazione di forma che caratterizza il 'mot vedette'. Se da un lato, infatti, le entrate lessicali sono il risultato del processo di lemmatizzazione, dall'altro ciascuna di esse deve essere la chiave per accedere alla globalità delle forme attestate nei manoscritti, caratterizzate da variazioni grafiche, grafico-fonetiche (frequentemente rappresentative delle differenti *scriptae* adottate dai codici), morfologiche e di alfabeto.

Nella fase di preparazione del dizionario è stata compilata una scheda per ciascuna variante. Ciascuna di esse ha in parte i medesimi attributi del lemma, alcuni dei quali non vengono riproposti ed altri sì, come "références", deputato a registrare le attestazioni della variante in questione. Le varianti possiedono anche proprietà particolari come 'typologie', tramite la quale si può descrivere il tipo di variazione che intercorre fra la forma considerata e l'entrata lessicale.

Si vedano di seguito gli esempi relativi ai dati di popolamento dello schema di alcune varianti. Quella afferente al lemma *ansens* se ne discosta solo per una variazione di tipo grafico-fonetico; quella del lemma *aloe*, invece, differisce anche per l'impiego dell'alfabeto:

Mot vedette : *ansens*
- variante : *aisens*
  (→)
  - typologie : grapho-phonetique

Mot vedette : *aloe*
- variante : *'LW'N (aloen)*
  (→)
  - typologie : d'alphabet et grapho-phonetique

L'esempio seguente, relativo ancora al lemma *ansens*, rende conto della formalizzazione adottata per descrivere una situazione più complessa, e cioè quella in cui una variante in alfabeto ebraico è coincidente con una medesima forma attestata dal corpus in alfabeto latino. Per esprimere tale situazione è stata prevista la tipologia 'variante de variante':

Mot vedette : *ansens*
- variante : *'YŠNS (aisens)*
  (→)
  - typologie : d'alphabet, grapho-phonetique et variante de variante

3.1.3. Un'altra sottoclasse è la proprietà del 'mot vedette' denominata 'sous lemmes', che designa voci che ne rappresentano diverse determinazioni (come, per es., lemma aygua con i sottolemmi *aygua roza, aygua arden, ayga mil,* etc.; lemma budel con i sottolemmi *budel cular, budel gran*) e che quindi è opportuno lemmatizzare sotto un'entrata principale unica.

Essa è a sua volta provvista di proprietà, fra le quali 'variantes' et 'synonymes', che costituiscono, dunque, il 3° livello di strutturazione dei dati lessicali.

Di seguito si riportano le istanze relative al popolamento delle proprietà del sottolemma *lin aloe* (lemma *aloe*) e quelle relative al popolamento delle proprietà delle sue varianti espresse in alfabeto latino e in alfabeto ebraico:

Mot vedette : *aloe*
- sous lemme : *lin aloe*
  (→)
  - traduction : bois d'aloès
  - alphabet: lt.
  - langue: a. occ.
  - catégorie gram: syntagme
  - variantes (→) lignum aloe ; *LYGN' 'LW'Š* (ligna aloes)
  - nom scientifique : Aloe vera L.
  - références : Ric 79 (P)
  - usage (date/période) : XII
  - correspondance: a. cast. *lignaloe* ; a. fr. *lignaloe* // lt. m. *lignum aloes*
  - lien ontologique : botanique (partie des plantes)
  - bibliographie : Ric (P); ShS1 // Sin 124; FEW 5:333b
  - synonymes (→) : lignum amarum

Mot vedette : *aloe*
- sous lemme : *lin aloe*
  (→)
  - variante : *lignum aloe*
    (→)
    - alphabet: lt
    - catégorie gram: syntagme
    - typologie de variante : grapho-phonetique
    - références : Thes XXV.5

Mot vedette : *aloe*
sous lemme : *lin aloe*
(→)
- variante : *LYGN' 'LW'Š (ligna aloes)*
  (→)
  - alphabet: hebr.
  - catégorie gram: syntagme
  - typologie de variante : d'alphabet et grapho-phonetique
  - références : ShS1 Qof 19 (ms. O)

3.1.4. Lemmi e sottolemmi possiedono anche la proprietà 'synonymes', campo che è atto a rendere conto della ripartizione diatopica del termine, qualora essa esista. E' il caso, per esempio, della realizzazione guascone *blanladre* per *libar* "elleboro", o

delle forme *brona* e *aromeze* attestate nella Haute Garonne rispettivamente per le più comuni *abrotanum* "abrotano" e *lapassa* "lapazio".

3.1.5. L'indicazione 'voir aussi' rinvia un lemma (o un sottolemma) ad un altro lemma (o a un sottolemma) che non ne costituisce un sinonimo, ma che ad esso è semanticamente legato, come nel caso, per esempio, di un verbo con un sostantivo del medesimo ambito concettuale (*afestolir* e *festola, afolar* e *afolament*), di un fitonimo con la denominazione del relativo frutto (*amenlier* e *amela*), di una malattia col malato che ne soffre (*gota aretica* e *aretich, ydropizia* e *itropic)*, etc.

3.2. La struttura semantica-concettuale

Con le proprietà denominate 'sous lemmes' e 'synonymes' e con i legami indicati mediante il campo 'voir aussi' si entra nella dimensione semantica, poiché le informazioni ad esse legate possiedono valenza nel versante concettuale. La considerazione di tale aspetto ha dunque spinto me e il collega di Gottinga a ipotizzare una forma innovativa di rappresentazione dei dati che fosse orientata anche da questo punto di vista.

Era necessario, infatti, progettare una struttura atta a superare il concetto di 'lemmi' in quanto unità semantiche separate, capace, cioè, di raccogliere in un unico contenitore logicamente valido tutte le entrate lessicali che afferiscono ad un medesimo ambito concettuale. Ai fini della ricerca in questa tipologia molto particolare di testi, infatti, risulta indispensabile poter raggruppare la molteplicità dei termini che ruota attorno ad un medesimo argomento, venendo a conoscenza, nel medesimo tempo, dell'insieme dei caratteri formali che li riguardano.

Del resto, di fronte all'organizzazione per ordinamento alfabetico che caratterizza fin dalla nascita la lessicografia di ambito scientifico in ambiente greco-romano, già nell'antichità esistono alcuni esempi che mostrano come il contenuto, all'interno di ciascuna lettera, poteva essere disposto secondo logiche differenti. Anche tralasciando l'uso per cui sovente nelle opere di contenuto religioso per ogni lettera dell'alfabeto i termini erano indicati secondo una precisa gerarchia che rifletteva l'importanza rivestita da ciascuna entità in tale contesto, si può citare la particolare forma sistematica adottata da Dioscoride nei cinque libri della sua *Materia medica*, dove i semplici sono raggruppati secondo le proprietà specifiche che rinviano alle indicazioni terapeutiche, oppure l'organizzazione dei capitoli dedicati ai medicamenti composti all'interno dei compendi dei semplici, o ancora il caso dei rimedi che sono raggruppati secondo la malattia da curare proposti dagli antidotari, soprattutto in ambiente arabo (Gutiérrez Rodilla 2007, cap. 1-4).

Così come quelle compilazioni antiche, prevedendo tipologie differenti di consultazione a seconda degli ambienti e delle finalità, cercavano di organizzare il materiale nella maniera più consona alle diverse funzioni sottese alla nozione di lessico, ugualmente noi riteniamo che una rappresentazione ontologica nella quale proiettare l'insieme lessicale dei sistemi noemici relativi ai domini della botanica e della medicina possa rappresentare una valida struttura capace di soddisfare utenti con esigenze

disparate. Abbiamo concepito, dunque, un sistema in grado di interrogare la base dei dati che preveda, come chiave di accesso, un concetto o un tema generico, oppure una relazione fra due concetti, oppure ancora una relazione fra più elementi. I risultati che si ottengono utilizzando solamente 'index verborum' o 'index lemmatum', infatti, possono non essere esaustivi perché vi è il rischio elevato che il testo descriva uno stesso tema con l'uso di parole diverse da quelle utilizzate per l'interrogazione. Le possibilità offerte da raggruppamenti di termini che sono associati su base concettuale, di contro, favorisce il loro reperimento e anche la loro comprensione, perché consente di rilevare minime sfumature di significato e mette in condizione di stabilire, con notevole dettaglio, le relazioni che li legano o li distinguono.

E' per tale motivo che si è ritenuto indispensabile definire un'ulteriore proprietà della classe 'lemma', cioè 'lien ontologique', che traghetta l'utente nella dimensione concettuale. In questa prospettiva è il dominio stesso della terminologia medico-botanica dell'antico occitano che costituisce una 'classe', provvista delle proprie 'sottoclassi'[10].

Le informazioni concernenti la semantica e la dimensione concettuale vanno ad aggiungersi a quelle relative al significante. Tutte assieme, opportunamente collegate dal sistema informatico progettato, concorrono a fornire all'utilizzatore la possibilità di estrarre dai dati lessicali attestati nel corpus informazioni di tipologia differente, sia in relazione a ciascun termine, sia in relazione ai legami fra alcuni di essi, usufruendo della strutturazione effettuata sulla complessità dei rapporti che caratterizza la terminologia medico-botanica dell'antico occitano.

L'esempio delle informazioni che ruotano attorno al lemma *acacia* è significativo. Il termine, infatti, oltre a possedere un valore polisemico, è parte di un'articolata rete di rapporti formali e semantici[11]:

1° significato: "gomma arabica"
varianti: - di alfabeto e grafica: 'Q'SY'(*acassia*) (ShS1 Alef 22);
sinonimi: - *goma arabica* (Ric3. 3), che possiede la variante grafica *goma arabiqua* (Thes XXVI.3), la variante di alfabeto GWMH 'R'BYQ' (*goma arabica*) (ShS1 Qof 7, ms.O;), la variante di alfabeto e grafico-fonetica GWMY 'RBYQWM (*gumi arabicum*) (ShS1 Qof 7, ms. P). E' sottolemma di *goma*;
- *clasa* (Thes XXIX. 13).

2° significato: "succo di prugnole acerbe"
varianti: - grafica: *acasia* (Ashb f. 97r);
- grafico-fonetica: *gacia* (Ric1 T f. 129v);
- d'alfabeto e grafica: 'QSY'(*acassia*) (ShS1 Alef 19);

[10] Per la descrizione di questi aspetti rinvio alla comunicazione di Bozzi / Luzzi indicata sopra alla nota 5.

[11] Nell'esempio sono riportate solo alcune delle proprietà relative al lemma in questione. Per una visione d'insieme dei contenuti si veda la fig. 2, in Appendice.

sinonimi: - *suc pru agre* (Herb 14) ;
- *suc d'aranhons* (Thes XXIX.42), che possiede la variante grafico-fonetica *suc de ranhons* (Thes XVII.7). E' sottolemma di *aranhon*, che possiede il sinonimo *prunelha del boys,* con la variante di alfabeto e grafico-fonetica PRYN' ŠLWDYG' (*pruna silvatica*) (ShS1 Alef 19, ms.V);
- *suc d'armiges* (Thes XXXIII. 5), che è sottolemma di armiges.

3° significato: "acacia"
sinonimi: - 'ŠPYN' GBSY'QH (**spina Aegyptiaca*) (ShS1 Alef 19, ms. V).

Università di Pisa Maria Sofia CORRADINI

Bibliografia

André, Jacques, 1985. *Les noms des plantes dans la Rome antique,* Paris, Les Belles Lettres

André, Jacques, 1991. *Le vocabulaire latin de l'anatomie,* Paris, Les Belles Lettres.

Bos, Gerrit / Mensching, Guido, 2000. «Macer Floridus. A Middle Hebrew fragment with Romance elements», *Jewish Quarterly Review* 91, 17-51.

Bos, Gerrit / Mensching, Guido, 2001. «Shem Tov Ben Isaac, Glossary of botanical terms, nrs. 1-18», *Jewish Quarterly Review* 92, 21-40.

Bos, Gerrit / Mensching, Guido, 2005. «The Literature of Hebrew Medical Synonyms: Romance and Latin Terms and their Identification», *Aleph* 5, 169-211.

Corradini, Maria Sofia, 1991. «Sulle tracce del volgarizzamento di un erbario latino», *Studi Mediolatini e Volgari* 37, 31-132.

Corradini, Maria Sofia, 1997. *Ricettari medico-farmaceutici medievali nella Francia meridionale*, Firenze, Olschki.

Corradini, Maria Sofia, 2001. «Per l'edizione del corpus delle opere mediche in occitanico e in catalano: nuovo bilancio della tradizione manoscritta e analisi linguistica dei testi», *Rivista di Studi Testuali* 3, 127-195.

Corradini, Maria Sofia, 2006. «Due testimoni occitanici della *Anatomia porci* attribuita a Cofone salernitano», in Beltrami, Pietro / Capusso, Maria Grazia / Cigni, Fabrizio / Vatteroni, Sergio (ed.), *Studi di Filologia romanza offerti a Valeria Bertolucci Pizzorusso*, Pisa, Pacini, 463-492.

Corradini, Maria Sofia (in prep.) : *La ricezione salernitana dell'anatomia di Galeno in tre redazioni volgari (occitanica, catalana e oitanica).*

Corradini, Maria Sofia / Mensching, Guido, 2010. «Les méthodologies et les outils pour la rédaction d'un *Lexique de la terminologie médico-botanique de l'occitan du Moyen Age*», in Iliescu, Maria / Danler, Paul / Siller, Heidi M. (ed), *Actes du XXVè Congrès International de Linguistique et de Philologie Romanes (CILPR), Innsbruck, 3-8 septembre 2007*, Berlin, De Gruyter, VII, 200-208.

Corradini, Maria Sofia / Mensching, Guido, 2013. «Nuovi aspetti relativi al *Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan (DiTMAO)*: creazione di una base di dati integrata con organizzazione onomasiologica», in Casanova Herrero, Emili / Calvo Rigual,

Cesareo (ed.), *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas*, *Valencia 6-11 septiembre 2010*, Berlin, De Gruyter, VIII, 113-124.

DAG = Baldinger, Kurt, 1975ss. *Dictionnaire onomasiologique de l'ancien gascon,* Tübingen, Niemeyer.

DAO = Baldinger, Kurt, 1975-2007. *Dictionnaire onomasiologique de l'ancien occitan*, Tübingen, Niemeyer.

DETEMA = Herrera, María Teresa, 1996. *Diccionario español de textos médicos antiguos,* 2 vol., Madrid, Arco Libros.

DOM = Stempel, Wolf-Dieter / Stimm, Helmut, (ed.), 1996-. *Dictionnaire de l'occitan médiéval,*Tübingen, Niemeyer.

FEW = Wartburg, Walter von, 1922-. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, 25 vol., Leipzig/ Bonn/Bâle, Teubner/Klopp/ Zbinden.

Gutiérrez Rodilla, Bertha, 2007. *La esforzada reelaboraciòn del saber*, San Millan de la Cogolla, Cilengua.

LEI = Pfister, Max / Schweickard, Wolfgang (ed.), 1984ss. *Lessico etimologico italiano*, Wiesbaden, Reichert.

Mensching, Guido, 1994. *La Sinonima delos nonbres delas medeçinas griegos e latynos e arauigos*, Madrid, Arco Libros.

Mensching, Guido / Savelsberg, Frank, 2004. «Reconstrucció de la terminologia mèdica occitano-catalana dels segles XIII i XIV a través de llistats de sinònims en lletres hebrees», in: *Actas del I congrés de l'estudi dels Jueus en territori de llengua catalana, Barcelona-Girona 2001*, Barcelona, Universidad, 69-81.

ShS1 = Bos, Gerrit / Hussein, Martina / Mensching, Guido / Savelsberg, Frank, 2011. *Medical Synonym Lists from Medieval Provence: Shem Tov ben Isaac of Tortosa,* Sefer ha - Shimmush. *Book 29. Part 1: Edition and Commentary of List 1 (Hebrew - Arabic - Romance/Latin)*, Leiden, Brill.

ShS2 = Bos, Gerrit / Hussein, Martina / Mensching, Guido // Savelsberg, Frank, in prep. *Medical Synonym Lists from Medieval Provence: Shem Tov ben Isaac of Tortosa,* Sefer ha - Shimmush. *Book 29. Part 1: Edition and Commentary of List 2*.

Trotter, David / Rothwell, William, 2007. *Anglo-Norman Dictionary.* ‹www.anglo-norman.net›.

Vernay, Henri, 1991. *Dictionnaire onomasiologique des langues romanes,* Tübingen, Niemeyer.

Appendice

Corpus testuale

1° tipologia (Corradini 1991, 1997, 2001, 2006, e in prep.)

Manoscritti:

Auch, Arch. Département du Gers I 4066 (= A)

Bâle, Bibliothèque de l'Université D II 11(= B)

Cambridge, Trinity College 903 (= T)

Chantilly, Musée Condé 330 (= C)

Paris, Bibliothèque Sainte Geneviève 1029 (= S. Gen.)

Paris, Bibliothèque de l'Arsenal 8315 (= Ars.)
Princeton, University Library, Garrett 80 (= P)
Ravenne, Biblioteca Classense 215 (= Rav.)
Bordeaux, Bibliothèque municipale 355
Paris, BnF nouv. acq. l. 317
Paris, BnF f. 14974
Rodez, Bibliothèque municipale 60

Testi

Erbario di Odo de Meudon in quattro redazioni, di cui una in rima
Liber de simplici medicinae
Erbario
Lettera di Ippocrate a Cesare in due redazioni e in due testi abbreviati
Thesaur des paubres di Pietro Ispano (= Thes)
Rimedi per le febbri di Pietro Ispano (= Febr)
Antidotarium Nicolai (= Ant.Nic)
Vertutz de l'ayga ardent
Ricettario in due redazioni (mss. P e T)
Ricettario in due redazioni (mss. B e P)
Primo *Ricettario* del ms. A
Secondo *Ricettario* del ms. A
Ricettario del ms. S.Gen.
Livre des raisons di Peyre de Serras
Ricettario di Raimon de Castelnou
Anatomia del maiale (= Ant.Por)
Anatomia umana in due redazioni
Trattato sulle urine (fram. del ms. B)
Epistola Aristotelis ad Alexandrum (fram. del ms. B)

2° tipologia (Bos / Mensching 2000, 2001, 2005 ; Bos / Mensching / Hussein / Savelsberg 2011 e in prep.)

Manoscritti:

Parigi, BN héb. 1163, fol. 191r ss.
Vaticano Ebr. 550, fol. 132r ss.
Oxford, Hunt Donat 2, fol. 277v ss.
Parma, Biblioteca Palatina 2646, fol. 35r-45r
Parma, Biblioteca Palatina 3041
Parma, Biblioteca Palatina 3043, fol. 265r-266v
Berlino, Staatsbibliothek Preußischer Kulturbesitz Qu 835

Berlino, Staatsbibliothek Preußischer Kulturbesitz 24, fol. 72r-73v
Monaco, d.B., Bayerische Staatsbibliothek Cod. hebr. 19
Monaco, d.B., Bayerische Staatsbibliothek, Cod. hebr. 245, 155r-167r
Monaco d. B., Bayerische Staatsbibliothek Cod. hebr. 295
Oxford, Bodleian MS Poc. 353
MS. Mich. Add. 22 (Neubauer 2129), fol. 5v-16r

Testi :

Lista di sinonimi (ebr.-arabo-occ./lat.) del *Sefer ha-Shimmush* (=ShS1)
Lista di sinonimi (occ./lat.-arabo-ebr.) del *Sefer ha-Shimmush* (= ShS2)
Lista di sinonimi (lat.-arabo-occ.)
Lista di sinonimi (occ.- arabo)
Lista di sinonimi (arabo - lat.- occ./cat. - ebr.)
Versione ebraica della lista di sinonimi *Alphita*, con elementi occitanici
Versione ebraica (frammento) dell'*Erbario* di Odo de Meudon (Macer Floridus)
Versione ebraica del "Zad al-musafir" (*Viaticum*) di Ibn a-Jazzar, tradotta da Moses Ibn Tibbon, libro VII, cap. 7-30, con elementi occitanici (edizione in corso)

FIGURE

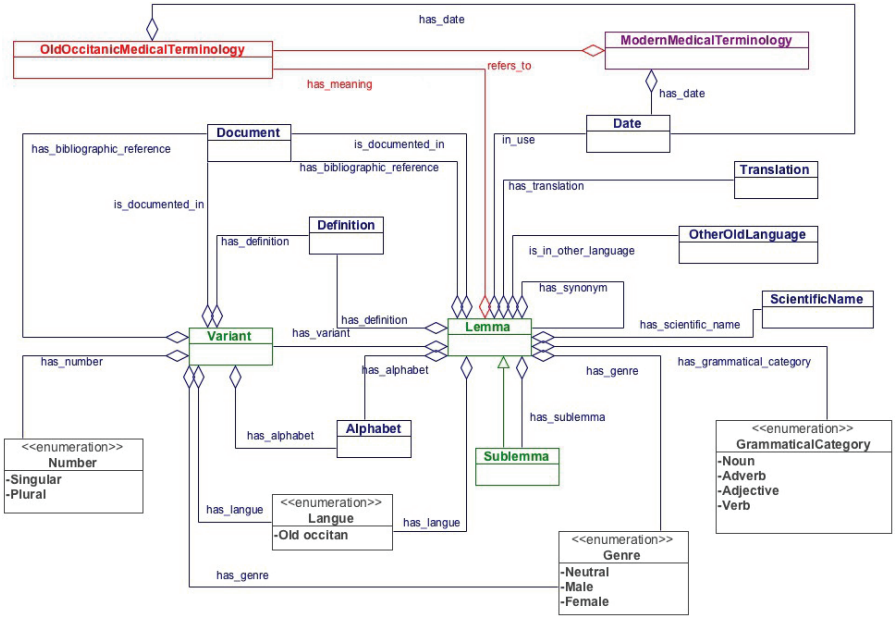


Fig. 1

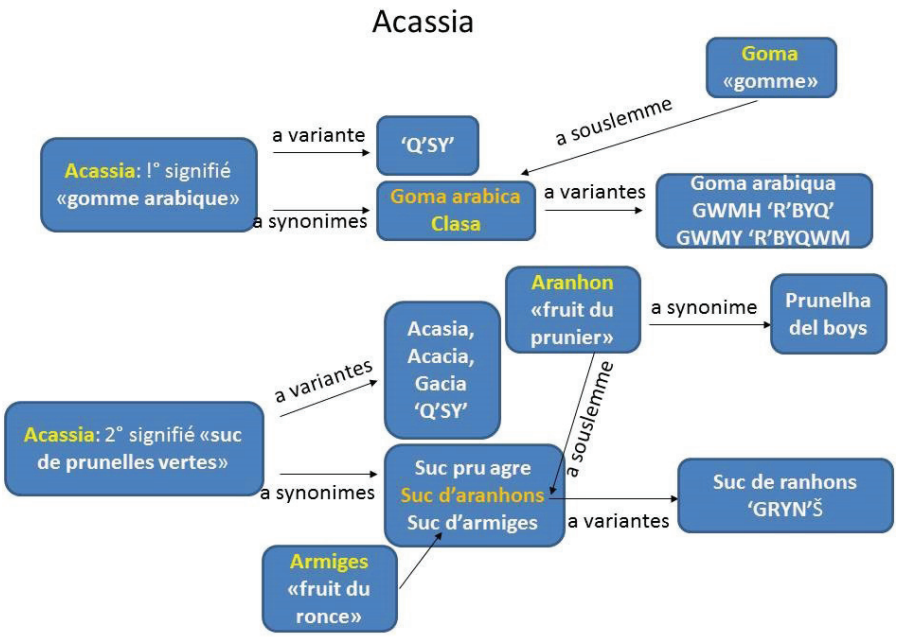

Acassia
Goma
«gomme»
a souslemme
a variante
'Q'SY'
Acassia: !° signifié
«gomme arabique»
a synonimes
Goma arabica
Clasa
a variantes
Goma arabiqua
GWMH 'R'BYQ'
GWMY 'R'BYQWM
Aranhon
«fruit du
prunier»
a synonime
Prunelha
del boys
Acasia,
Acacia,
Gacia
'Q'SY'
a variantes
Acassia: 2° signifié «suc
de prunelles vertes»
a synonimes
Suc pru agre
Suc d'aranhons
Suc d'armiges
a souslemme
Suc de ranhons
'GRYN'Š
a variantes
Armiges
«fruit du
ronce»

Fig. 2

Estudio lexicográfico para el tratamiento automático de la lengua: diccionario bilingüe francés-español de los sustantivos predicativos de las <ayudas económicas>

El objetivo de nuestro trabajo de investigación es la elaboración de un diccionario electrónico bilingüe (francés-español) de los sustantivos predicativos de las <ayudas económicas>, que permita el tratamiento automático de los textos de este campo de especialidad.

Tradicionalmente, la descripción de las lenguas de especialidad se ha reducido a la terminología, acordando una gran prioridad (casi exclusividad) a los sustantivos, sin prestar atención a la articulación de éstos con el resto de elementos de la frase. Este tipo de trabajo es poco útil desde un punto de vista del tratamiento automático de la lengua.

Para que un ordenador sea capaz de reconocer y de generar textos, necesita una descripción de la lengua completamente explícita y coherente desde el punto de vista morfológico, sintáctico y semántico. No es suficiente con las reglas generales de la gramática, la morfología y la sintaxis. Es necesario hacer una descripción fina de todas las propiedades lingüísticas de cada palabra. Estas descripciones serían formateadas en léxicos, de ahí la importancia de los diccionarios electrónicos.

Nuestro principal propósito es la inscripción del sustantivo dentro del marco de la frase, y poder así predecir, de manera automática, las colocaciones verbales. Esta investigación se sitúa pues en un cruce de caminos entre la lingüística, la informática y la inteligencia artificial

1. Marco teórico

El marco teórico en el que se inscribe nuestro trabajo es esencialmente el método de las clases de objeto, desarrollado en el laboratorio LDI[1], aunque también con una importante influencia de la Teoría Sentido-Texto (especialmente el *Dictionnaire Explicatif et Combinatoire*[2] y el *Lexique Actif du Français*[3]).

---

1 *Lexiques, Dictionnaires, Informatique* (UMR 7187, CNRS-Université Paris 13).

2 DEC (Mel'čuk, Igor (ed.), 1984-19994. Dictionnaire explicatif et combinatoire du français contemporain, Montréal, 4 vol.).

3 LAF ((Mel'čuk, Igor / Polguère, Alain (ed.), 2007. Lexique actif du français, Bruxelles, 1vol.).

Este enfoque, además de dar cuenta de una manera fina, rigurosa y sistemática de los hechos de la lengua, se adapta perfectamente a las especificidades de la herramienta informática.

Pasaremos ahora a hacer un breve resumen de las nociones de base de la teoría de las clases de objetos.

1.1. El predicado

El modelo de las clases de objetos postula que una frase está constituida por un predicado, unos argumentos y unos actualizadores.

Un predicado es una relación entre varias entidades (los argumentos). Esta relación se representa de la manera siguiente:

F(x)

F representa la función o relación (el predicado) y *X* es la variable (el/los argumento/s).

Por ejemplo, en la frase simple *Luis come une manzana* el predicado verbal *comer* establece una relación entre dos elementos: el que come (*Luis*) y el que es comido (*una manzana*).

El predicado puedes ser un verbo, como acabamos de ver en el ejemplo precedente, pero también puede aparecer en la frase bajo la forma de:

- un nombre : *Luis brinda ayuda a Mario ;*
- un adjetivo : *Luis es feliz* ;
- una preposición : *El libro está sobre la mesa*.

Un predicado debe ser definido por su serie de argumentos más larga. Para delimitar esta serie hemos de ser capaces de distinguir un argumento de un circunstancial. La diferencia reside en que el argumento es inducido por el propio significado del predicado y el circunstancial no.

1.2. La frase

El esquema de la frase elemental es, entonces, el siguiente:

Predicado (arg1, arg2, arg3…)

Y en el caso de nuestro ejemplo anterior sería:

comer (Luis, manzana)

El paso de esta noción abstracta de relación entre dos elementos a una frase real implica dos operaciones:

- la linearización : el orden de los argumentos es fundamental para determinar el significado. *Luis* satura la posición del primer argumento y *manzana* satura la del segundo ;

- la actualización : consiste en la expresión del tiempo y del aspecto para los predicados y la determinación para los argumentos.

Una de las principales dificultades a las que se enfrenta el tratamiento automático de las lenguas naturales es la polisemia. Para solventar este problema, la teoría de las clases de objetos considera la frase como unidad mínima de análisis, ya que el contexto lingüístico elimina la ambigüedad, es decir, que es el contexto lingüístico el que determina el significado. Para ejemplificarlo, observemos las tres frases siguientes que ilustran los diferentes significados de *descendre*:

- *Luc descend les escaliers* (‘bajar’)
- *Luc descend le criminel* (’matar’)
- *Luc descend deux bouteilles de vin* (‘beber’)

De tal modo, en nuestra descripción de la lengua, y en consecuencia en la elaboración de un diccionario electrónico, no podemos separar el léxico (las entradas del diccionario), la sintaxis (la combinatoria) y la semántica (el resultado de los elementos léxicos organizados de una manera determinada dentro del marco de la frase). Estos tres niveles están agrupados bajo la noción de ‘empleo’: el empleo de una palabra dentro de una frase.

1.3. El empleo

« Un emploi de prédicat est constitué de l’ensemble des propriétés qui sont spécifiques de ce prédicat »[4]. Estas propiedades son las siguientes:

- un esquema de argumentos: la esfera de argumentos regida por el predicado en cuestión ;
- un significado: la definición del esquema de argumentos despeja toda ambigüedad en el significado ;
- una forma morfológica: los predicados pueden tener variantes morfológicas[5] ;
- una actualización: la inscripción del esquema de argumentos dentro del discurso; los determinantes actualizan los argumentos; en cuanto a los predicados, la actualización depende de su naturaleza morfológica: los predicados verbales son conjugados y los predicados nominales y adjetivales son actualizados por los verbos soporte ;
- un sistema aspectual: las diferencias aspectuales ayudan a diferenciar empleos ;
- unas transformaciones: como la tematización, la pronominalización, la pasivación, etc. ;
- un campo de especialidad: las palabras pueden adoptar significados diferentes en función del campo de especialidad en el que son utilizadas ;
- un nivel de lengua: el predicado puede ser definido por el nivel de lengua que le caracteriza[6].

Así, la descripción de la lengua en términos de empleo resuelve el problema de la polisemia.

---

4 Gross (2010, 108).

5 Se trata de los predicados polimórficos. Por ejemplo, *dese-* : *Luis desea regresar a su país*; *Luis alberga el deseo de regresar a su país; Luis está deseoso de regresar a su país*.

6 Es una propiedad a tener en cuenta, por ejemplo, en la traducción.

1.4. La clase de objeto

En un nivel inmediatamente superior a los empleos, se sitúan las clases de objeto. « Une classe d'objets est un ensemble de substantifs, sémantiquement homogènes, qui détermine une rupture d'interprétation d'un prédicat donné, en délimitant un emploi spécifique »[7]. Una clase de objeto es, pues, un conjunto de empleos que operan las mismas restricciones sintáctico-semánticas.

Las clases de objetos están organizadas jerárquicamente en hiperónimos e hipónimos[8]. Ya se habían elaborado anteriormente organigramas léxicos de dependencia hiperonímica e hiponímica[9]. La novedad que aportan las clases de objetos es el método de elaboración de esta clasificación jerárquica y, en consecuencia, sus resultados y aplicaciones. Se trata de un método rigurosamente lingüístico, ya que tiene en cuenta las definiciones nocionales y el comportamiento sintáctico-semántico de los empleos. Además, recordemos que la descripción en clases de objetos está basada en la observación en el corpus de las construcciones sintácticas y de las combinaciones léxicas posibles.

Las clases de objetos son pues descriptores sintáctico-semánticos, es decir, exclusivamente lingüísticos adecuados a las especificidades y a las limitaciones de la máquina.

2. Elaboración de los diccionarios

Una vez descrito el marco teórico, pasamos a explicar las tres grandes etapas que hemos seguido en el proceso de elaboración de nuestros diccionarios. La primera etapa ha consistido en identificar los sustantivos predicativos de las <ayudas> basándonos en su significado tal como está especificado en las obras lexicográficas. Seguidamente hemos elaborado una tabla de análisis adaptada al marco teórico en el que se inscribe este trabajo y la hemos aplicado a todos los sustantivos seleccionados en la etapa anterior. En último lugar, hemos establecido las relaciones de equivalencia entre los sustantivos predicativos de ambas lenguas.

2.1. Recolección de palabras

Nuestro trabajo ha versado únicamente sobre el vocabulario del francés de Francia y del español de España.

Dentro de la nomenclatura de estos diccionarios encontramos tanto palabras de la lengua general (*propina, limosna*) como términos de lenguas de especialidad (*renta agraria*). Consideramos que la discriminación tradicional entre lengua general y lengua de especialidad no está fundamentada de un punto de vista lingüístico. Una

[7] Gross (2008, 121).

[8] Los hiperónimos son sub-clases que permiten un análisis más fino de las restricciones de selección.

[9] Por ejemplo, WordNet, FrameNet, etc.

lengua de especialidad no es más que el uso de una lengua natural para dar cuenta de conocimientos especializados[10]. Por eso decidimos describir conjuntamente, con los mismos útiles de análisis, todas estas palabras.

De este modo, en el proceso de recolección de palabras para el diccionario utilizamos fuentes de diferente naturaleza, como bases de datos de palabras[11], diccionarios de lengua general[12], diccionarios especializados[13], diccionarios combinatorios y de relaciones léxicas[14] y sitios web especializados.

2.2. Elaboración y aplicación de una tabla de análisis

Los criterios para la elaboración de una tabla de análisis han sido seleccionados por su compatibilidad con los principios de análisis del método de las clases de objetos. Hemos seleccionado dos tipos de propiedades: configuracionales y combinatorias. Respecto a las primeras, hemos tenido en cuenta el número de argumentos, su modo de estructuración en la frase y su naturaleza[15]. En cuanto a las propiedades combinatorias[16], hemos extraído los verbos soporte y los verbos predicativos apropiados seleccionados por los sustantivos predicativos[17].

Así, hemos aplicado esta tabla de manera sistemática a toda nuestra lista de sustantivos predicativos, observados en contexto, es decir, dentro del marco de la frase. Este estudio sistemático nos ha permitido, por una parte, separar los diferentes empleos de una palabra en entradas independientes en el diccionario, y, por otra parte, agrupar bajo una misma clase de objeto todos los empleos que comparten el mismo esquema argumental y que seleccionan las mismas colocaciones verbales.

2.3. Correspondencias entre las entradas de ambos diccionarios

En las dos primeras etapas de elaboración de los diccionarios hemos estudiado los sustantivos predicativos de cada lengua de manera independiente. No es hasta el último momento cuando relacionamos estas dos descripciones que hemos elaborado.

Evidentemente, las correspondencias se han establecido entre los sustantivos predicativos y no entre los colocativos, siendo condición *sine qua non* que ambos empleos compartan la misma clase de objeto. Estas permiten que cada colocación

---

10 Lerat (1995).

11 FRANTEXT para el francés y CREA para el español.

12 Por ejemplo, el *DRAE* y el *PRob*.

13 Por ejemplo, el *Dictionnaire des assurances sociales* y el *Dictionnaire économique et financier.*

14 Por ejemplo, el diccionario REDES y el *Dictionnaire des combinaisons des mots* de Dictionnaires Le Robert.

15 Estas propiedades están relacionadas con la operación de linearización evocada anteriormente.

16 Propiedades relacionadas, en este caso, con la operación de actualización.

17 La parte de las propiedades combinatorias está fuertemente inspirada del trabajo de Inès Sfar y Taouffik Massoussi (2009).

sea reconocida en la lengua de origen y generada en la lengua de destino por su diccionario correspondiente.

No se trata pues de un diccionario bilingüe, sino de dos diccionarios monolingües coordinados.

3. Descripción de los diccionarios

Nuestros diccionarios contienen 22 campos. Pasemos pues a describir cada uno de ellos.

3.1. La entrada

Cada entrada corresponde a un empleo predicativo diferente. Hasta ahora los diccionarios contienen 750 entradas para el francés y 200 para el español[18].

Las entradas comprenden tanto palabras simples como compuestas. Estas últimas plantean un problema de análisis específico al soporte informático: para un sistema informático una palabra es una serie de caracteres entre dos espacios. Es necesario pues anotar esos espacios que no representan rupturas sintácticas.

Los sustantivos de ayudas financieras a menudo aparecen en el discurso bajo forma de un acrónimo. Para que la máquina no se bloquee ante estos, hemos considerado necesario incluirlos en la descripción del empleo.

A	B	C	D	E	F	G	H	I	J	K
Nom	**Catégorie grammaticale**	**Classe d'objet**	**N0**	**N1**	**N2**	**N3**	**N4**	**Lo**	**Acronyme**	**espagnol**
13ème mois	nf	aide financière: prime	coll_hum<entre	à<hum>	de<mont>	pour Vinf		dr oit d		extra / extraordinaria / paga extra / paga extraordinaria
allégement fiscal	nm	aide financière: décompte	coll_h	à<	de<m	sur		dr		desgravación fiscal
allègement fiscal	nm	aide financière: décompte	coll_h	à<	de<m	sur		dr		desgravación fiscal
allocation chômeurs âgés	nf	aide financière: allocation régulière	coll_h	à<	de<m	po	po	pr	ACA	subsidio por desempleo

3.2. La categoría gramatical

Indicación sobre la parte del discurso a la que pertenece la entrada y el género. Algunas entradas comprenden también indicaciones del número del sustantivo por ser palabras usadas exclusivamente o generalmente en plural.

[18] En un principio, este estudio iba a versar exclusivamente sobre el francés y para ello analizamos unas 900 palabras. Posteriormente, pensamos extender este estudio al español, con el objetivo de establecer correspondencias entre ambas lenguas. Comenzamos con una pequeña muestra de unas 200 palabas en español para verificar si obtendríamos resultados satisfactorios, antes de emprender un estudio más vasto.

3.3. La clase de objeto

La clasificación de los sustantivos predicativos en clases de objetos es un trabajo ímprobo, ya que hay que estudiar cada empleo de manera independiente, verificar en el contexto el esquema de argumentos y localizar las colocaciones verbales que le acompañan. Esta tarea no puede ser informatizada. Hasta el momento, solo el cerebro humano está capacitado para determinar si un elemento forma parte o no de una clase de objeto determinada.

La clase <aide financière> está organizada en las siguientes subclases[19]:

- < aide financière : allocation ponctuelle> *aide á la reprise ou à la création d'entreprise, aide hospitalière ;*
- < aide financière : allocation régulière > *subsidio por desempleo / allocation de chômage ;*
- < aide financière : allocation régulière> bourse d'enseignement> *beca de doctorado / bourse de thèse ;*
- < aide financière : allocation viagère> *jubilación / retraite ;*
- < aide financière : avance> *avance, anticipo ;*
- < aide financière : commission illicite> *sobre / enveloppe ;*
- < aide financière : compensation>: *indemnización por despido procedente / indemnité légale de licenciement ;*
- < aide financière : décompte> *deducción fiscal / abattement fiscal ;*
- < aide financière : donation> *limosna / aumône ;*
- < aide financière : emprunt>[20] *emprunt, emprunt national ;*
- < aide financière : prêt> *microcrédito / microcrédit ;*
- < aide financière : prime> *extraordinaria / 13ème mois ;*
- < aide financière : subvention> *subvención / subvention.*

A	B	C	D	E	F	G	H	I	K
Nom	**Catégorie gra**	**Classe d'objet**	**N0**	**N1**	**N2**	**N3**	**N4**	**Domaine**	**espagnol**
abattement	nm	aide financière: décompte	coll_hum <vendeur>	à <hum, coll_hum:acheteur>	de <montant, partie_montant>	sur <dette_N1>		commerce	deducción
allocation de recherche	nf	aide financière: allocation régulière: bourse d'enseignement	coll_hum <Etat, CNRS>	à <hum:doctorant>	de <montant>	pour <période_tps>	pour Vinf <but:act:étudier>	enseignement	beca de investigación
assurance automobile	nf	aide financière: compensation	coll_hum <compagnie d'assurances>	à <hum:assuré, tiers>	de <montant>	pour Vinf, N <cause:act, état, évén>		assurances	seguro de coche

3.4. El esquema de argumentos

La descripción del esquema de argumentos se encuentra distribuida en los siguientes campos:

---

19 Mantenemos el metalenguaje en francés, porque así ha sido utilizado en los diccionarios.

20 Esta clase no existe en español, ya que esta lengua no dispone de un sustantivo predicativo con el esquema de argumentos de la clase <emprunt>: N0: deudor / N1: acreedor, como sí que existe en francés *faire un emprunt*.

- *[N0]* corresponde a la posición de sujeto ;
- *[N1]* corresponde a la posición de primer complemento ;
- *[N2], [N3], [N4]* corresponden a las posiciones del segundo, tercer y cuarto complementos respectivamente[21].

Los argumentos de cada empleo son descritos desde un punto de vista morfológico, sintáctico y semántico. Así, en la descripción del esquema de argumentos se especifica:

- la construcción: el modo de estructuración de los argumentos respecto al empleo predicativo; el predicado está definido por su serie más larga de argumentos[22] ;
- la distribución morfo-sintáctica: se describe el tipo de construcción que ocupa las posiciones argumentales (GN, completiva o infinitivas) ;
- la distribución sintáctico-semántica: las particularidades semánticas de los argumentos (clases o hiperclases).

A modo de ejemplo, observemos el esquema de argumentos de la clase <prêt>:

N0	N1	N2	N3	N4
hum <créancier>	a <hum, coll_ hum:débiteur>	de <montant>	a, con <taux>	a <période_tps>

Para que un empleo de una palabra pertenezca a la clase <prêt> necesita un acreedor (N0), una persona o un colectivo de personas (N1) que pide(n) el préstamo, una cantidad de dinero (N2) concedida *a/con* una tasa de interés (N3) *a* un plazo de tiempo (N4).

A	C	D	E	F	G	H	I	K
Nom	**Classe d'objet**	**N0**	**N1**	**N2**	**N3**	**N4**	**Domaine**	**espagnol**
crédit bancaire	aide financière: prêt	coll_hum <créancier:banque>	à <hum, coll_hum:débiteur>	de <montant>	à, avec <taux>	sur, pour <période_tps>	banque	crédito bancario
escompte	aide financière: décompte	hum, coll_hum <vendeur>	à <hum, coll_hum:acheteur>	de <montant, partie_montant>	sur <dette_N1>		commerce	descuento / rebaja
golden parachute	aide financière: compensation	coll_hum <entreprise>	à <hum:travailleur:salarié:cadre dirigeant>	de <montant>	pour Vinf, N <cause:act, état, évén>		droit du travail	paracaídas dorado

3.5. El campo semántico

El campo semántico en el que una entrada determinada adquiere un significado determinado. Por ejemplo, *prima*, a parte del campo semántico del trabajo o de la economía, puede pertenecer al campo de la familia.

[21] No hemos encontrado ningún sustantivo predicativo de <ayuda> que rija más de 4 argumentos.

[22] Esto no quiere decir que sea necesario saturar todas estas posiciones en el discurso.

A	C	D	E	F	G	I	K
Nom	**Classe d'objet**	**N0**	**N1**	**N2**	**N3**	**Domaine**	**espagnol**
prime	aide financière: décompte	hum, coll_hum <vendeur>	à <hum, coll_hum:acheteur>	de <montant, partie_montant>	sur <dette_N1>	commerce	prima
prime	aide financière: compensation	coll_hum	à <hum>	de <montant>	pour Vinf, N <cause:act,	général	prima
prime	aide financière: prime	coll_hum	à <hum:travailleur, stagier>	de <montant>	pour Vinf <cause:servi	finances / travail	prima
prime	aide financière: subvention	org_coll <Etat>	à <hum; org_coll:entreprise>	de <montant>	pour Vinf <but:project:	économie	prima

3.6. La correspondencia en el otro idioma

Como ya hemos señalado, las correspondencias se establecen entre los sustantivos predicativos y no entre sus colocativos. Para que pueda establecerse esta correspondencia, ambos sustantivos deben pertenecer a la misma clase de objeto y por ende, compartir el mismo esquema de argumentos. Recordemos que el significado de un predicado está determinado por el contexto lingüístico en el que aparece (su esquema de argumentos). Así, para poder establecer una relación de equivalencia entre dos predicados de ambas lenguas, la analogía ha de ser total en cuanto al número de argumentos y la naturaleza semántica de los mismos.

A	C	D	E	F	G	I	K
Nom	**Classe d'objet**	**N0**	**N1**	**N2**	**N3**	**Domaine**	**espagnol**
subside	aide financière: allocation régulière	hum, coll_hum	à <hum, coll_hum>	de <montant>	pour <période_tps>	protection sociale	subsidio
subvention	aide financière: subvention	coll_hum	à <hum, coll_hum>	de <montant>	pour Vinf <but:project:act>	culture / politique économique	subvención
retraite progressive	aide financière: allocation régulière	coll_hum <Etat>	à <hum:travailleur:salarié:temps partiel:senior>	de <montant>	pour <période_tps>	protection sociale	jubilación flexible / jubilación parcial

3.7. Verbos soporte activos y pasivos

La definición de las clases de objetos está basada esencialmente en el esquema de argumentos de cada predicado. Los verbos soporte activos y pasivos, por su parte, nos han ayudado a afinar esta primera clasificación. En ocasiones, predicados que compartían el mismo esquema de argumentos, no compartían, sin embargo, los verbos que los actualizan. No podemos pues definir estos predicados dentro de la misma clase de objeto, sino que pertenecen a clases diferentes.

Los verbos soporte activos toman como sujeto el argumento N0 y como primer complemento el argumento N1, mientras que los verbos soporte pasivos toman como sujeto N1 y como primer argumento N0.

A	C	K	M	N
Nom	**Classe d'objet**	**espagnol**	**Verbes supports actifs**	**Verbes supports passifs**
assurance-incendie	aide financière: compensation	seguro de incendio	accorder, allouer, verser	bénéficier de
allocation de chômage	aide financière: allocation régulière	paro / prestación contributiva / prestación	accorder, attribuer, donner, octroyer, offrir, payer, verser	avoir, bénéficier de, jouir de, obtenir, percevoir, recevoir, toucher
bourse	aide financière: allocation régulière: bourse	beca	accorder, attribuer, offrir, proposer, verser	bénéficier de, décrocher, être muni de, être titulaire de, jouir de, obtenir, percevoir, recevoir, toucher, voir

3.8. Verbos soporte aspectuales

Los sustantivos predicativos de las ayudas son sustantivos de <acción>. Entre todas las posibilidades aspectuales, hemos elegido las más fructíferas y pertinentes para los predicados de <acción> y más específicamente para la clase de las <ayudas económicas>. De este modo, nuestros diccionarios contienen actualizadores aspectuales referidos al progreso de la acción (aspectos incoativo, progresivo y terminativo) y a su frecuencia (aspectos iterativo e iterativo-intensivo).

El aspecto inherente de los sustantivos predicativos determina los actualizadores que estos van a seleccionar. Así, un sustantivo de la clase <aide financière : allocation ponctuelle>, cuyo aspecto inherente es puntual, es incompatible con los actualizadores aspectuales de tipo incoativo.

A	C	O	P	Q	R	S
Nom	**Classe d'objet**	**Verbes supports aspectuels : incohatif**	**Verbes supports aspectuels : itératif**	**Verbes supports aspectuels : itératif-**	**Verbes supports aspectuels : progress(**	**Verbes supports aspectuels : terminatif**
aide à la double résidence	aide financière: allocation régulière			augmenter, revaloriser	N1N0 conserver, garder	couper, suspendre
aide au développement	aide financière: prêt	N1N0avoir accès à, conclure, contracter, prendre, recourir (à), souscrire (à)			débloquer, reconduire, renouveler	bloquer,suspendre. N1 acquitter, amortir, rembourser, solder
aide à l'embauche dans les TPE	aide financière: subvention		débloquer, reconduire	accroître, augmenter, majorer		suspendre, éliminer, supprimer, annuler

3.9. Los operadores apropiados

Los sustantivos predicativos pueden aparecer en posición de argumento dentro de frases complejas. En este caso, pues, estos sustantivos (predicados de 1er orden) son seleccionados de manera apropiada por otros predicados de nivel superior (predicados de 2º orden).

El estudio de las unidades léxicas en contexto nos ha permitido observar que existen ciertos verbos predicativos que acompañan frecuentemente a los sustantivos de <ayuda económica>, de ahí que los llamemos *apropiados*. Se trata de los verbos de <petición> (*pedir, solicitar*), los verbos de <puesta en marcha> (*instaurar*), los verbos de <denegación> (*rechazar, denegar*) y los verbos de <disminución> (*recortar, rebajar*).

A	C	T	U	V	W
Nom	**Classe d'objet**	**Opérateurs appropriés : verbes de <demande>**	**Opérateurs appropriés : verbes de <mise en place>**	**Opérateurs appropriés : verbes de <refus>**	**Opérateurs appropriés : verbes de <diminution>**
apport	aide financière: donation: général	demander, solliciter		N0N1 refuser	réduire
bakchich	aide financière: prime	demander, exiger, réclamer, solliciter	créer, instaurer, instituer		
subvention	aide financière: subvention	demander, réclamer, solliciter	voter	limiter, plafonner, réduire, refuser, retirer	

* * *

El interés de esta descripción no reside únicamente en su aplicación al tratamiento automático de las lenguas naturales, sino en el estudio mismo de la lengua; la reflexión que este estudio nos lleva a hacer sobre el comportamiento sintáctico-semántico de las unidades léxicas analizadas.

En cuanto al tratamiento automático, hemos podido verificar que esta descripción de la lengua ayuda a resolver los problemas de la polisemia y de la fijación. Además de permitir el tratamiento automático de la sinonimia, la derivación, la anáfora, la metáfora y la paráfrasis.

Como resultado obtendremos una herramienta útil para la documentación automática, la extracción o localización de información, la clasificación y categorización de documentos, la ayuda en la redacción, etc.

En cuanto a las perspectivas para mejorar estos diccionarios, la descripción de los adjetivos apropiados a estos predicados nominales podría ayudarnos a afinar la clasificación en clases de objetos. Sin embargo, esto podría plantear problemas para establecer las correspondencias de sustantivos en ambas lenguas. Cuanto más fina es la clase, más difícil es encontrar su clase correspondiente en la otra lengua.

Universidad de Salamanca Elena DIEGO HERNÁNDEZ

Referencias bibliográficas

Bernard, Yves / Colli, Jean-Claude / Lewandowski, Dominique, 1975. *Dictionnaire économique et financier*, Paris, Éditions du Seuil.

Bosque, Ignacio / Maldonado, Concepción, 2004. *REDES. Diccionario combinatorio del español contemporáneo*, Madrid, SM.

CREA : *Corpus de Referencia del Español Actual*. ‹www.rae.es›.

DRAE = Real Academia Española, 2001. *Diccionario de la Real Academia Española*, Madrid, Editorial Espasa Calpe, 2vol.

FRANTEXT, ATILF - CNRS & Université de Lorraine. ‹www.frantext.fr›

Grezka, Aude / Buvet Pierre-André, 2007. «Élaboration d'outils méthodologiques pour décrire les prédicats du français», *Linguisticae Investigationes* 30, nº 2, 217-245.

Gross, Gaston, 2008. «Les classes d'objets», *LALIES* 28, 111-165.

Gross, Gaston, 2010. «La notion d'emploi dans une grammaire de prédicats», *Cahiers de lexicologie* 96, 97-115.

Le Fur, Dominique, 2007. *Dictionnaire des combinaisons des mots: les synonymes en contexte*, Paris, Dictionnaires Le Robert.

Le Pesant, Denis / Mathieu-Colas, Michel, 1998. «Introduction aux classes d'objets», *Langages* 131, 6-33.

Lerat, Pierre, 1995. *Les langues spécialisées*, Paris , Presses Universitaires de France.

Mel'čuk, Igor, 1997. *Vers une linguistique Sens-Texte : leçon inaugurale,* Paris, Collège de France.

Mel'čuk, Igor / Class, André / Polguère, Alain, 1995. *Introduction à la lexicologie explicative et combinatoire*, Bruxelles, Éditions Duculot.

PRob = Rey Debove, Josette / Rey, Alain, 2011. *Le nouveau Petit Robert*, Paris, Dictionnaires Le Robert.

Sournia, Jean-Charles, 1973. *Dictionnaire des assurances sociales*, Paris, Maison et Cie Éditeurs.

Sfar, Inès / Massousssi, Taouffik, 2009. «Description des prédicats nominaux : de la langue générale aux langues spécialisées», *Neophilologica* 21, 62-81.

Il *DiVo* (*Dizionario dei Volgarizzamenti*). Nuovi strumenti per lo studio delle traduzioni dal latino nell'italiano delle origini*

«seguendo il volgarizzarsi e il diffondersi
della cultura medievale e classica, specialmente, noi troveremo
il sentiero ascoso che va da Dante teologo al Petrarca filologo»
(Concetto Marchesi)

1. Per lo studio del «classicismo medievale»

Il progetto *DiVo* (*Dizionario dei volgarizzamenti*), ideato e diretto da Elisa Guadagnini e Giulio Vaccaro, mira ad uno studio complessivo del lessico dei volgarizzamenti italoromanzi dei testi classici e tardo-antichi, fornendo allo stesso tempo alla comunità scientifica gli strumenti per affrontarne l'analisi dal punto di vista filologico e linguistico[1].

Il *DiVo* nasce dalla consapevolezza che il lessico di traduzione dovrebbe godere di un'attenzione 'speciale': per il peso dei volgarizzamenti nella documentazione volgare antica, per la rarità delle attestazioni di determinati lemmi o significati che trovano in questa tipologia testuale la loro prima apparizione (fino ad arrivare all'integrazione nel sistema linguistico), per il trattamento filologico del dato, spesso legato al rapporto biunivoco che s'instaura all'atto della traduzione tra il testo di partenza e il testo d'arrivo, in particolare tra il lemma latino e il traducente volgare, secondo le differenti modalità di resa attraverso il prestito diretto, il calco semantico o la riformulazione volgare. Si tratta di una posizione speciale che risulta ben chiara alla luce dell'esperienza di redazione di voci per il *TLIO* (*Tesoro della lingua italiana delle origini*), a margine del quale già da tempo era stato predisposto uno strumento, la *Bibliografia dei volgarizzamenti*, per agevolare una comprensione filologicamente più avvertita e approfondita della documentazione risalente ai testi di traduzione[2].

---

* Nel quadro di una comune elaborazione, si debbono a Diego Dotto i §§ 1-2 e a Cristiano Lorenzi il § 3.

[1] Il progetto, ospitato dall'Opera del Vocabolario Italiano del CNR e dalla Scuola Normale Superiore di Pisa, è finanziato dal MIUR-Ministero dell'Istruzione, dell'Università e della Ricerca, nell'àmbito del FIRB-Futuro in ricerca 2010.

[2] Questa consapevolezza era già presente nella lessicografia della Crusca, se Salviati e Borghini dedicavano pagine illuminanti sull'opportunità di accogliere la documentazione dei volgarizzamenti, mentre, curiosamente, essa si è andata affievolendo nella lessicografia otto-

La focalizzazione sui testi classici e tardo-antichi sollecita inoltre un forte interesse linguistico, storico-letterario e culturale per illuminare la stagione che, con una formula di Gianfranco Folena, potremmo designare con l'etichetta di «classicismo medievale»: essa va intesa come un contenitore complessivo dei diversi momenti di ricezione, o meglio di riscoperta e di rielaborazione dei testi classici nel medioevo, alle soglie dell'umanesimo; rispetto al quale il classicismo medievale ha coordinate ed esiti ben distinti, non solo sul piano cronologico[3]. Si aggiunga che lo studio del classicismo medievale avverrebbe sul piano della tipologia testuale in cui più vivo è il contatto con il 'classico', il testo di traduzione.

Per fare questo, il progetto prevede la costruzione di due strumenti, un sistema di banche dati che supporti ricerche sui testi a partire dal volgare e dal latino, il *corpus DiVo* e il *corpus CLaV*o (cf. § 2), e un repertorio di schede sulla tradizione dei testi inclusi nei *corpora*, collegato alle banche dati, ma anche indipendente, *DiVo DB* (cf. § 3).

Entrambi gli strumenti, che sono già accessibili in rete, sono ancora in costruzione e in quanto tali vanno giudicati almeno al momento (ottobre 2013), ma saranno completati entro il settembre 2014. Una versione stabile dei *corpora* sarà rilasciata nel marzo 2014, tenendo conto che, essendo strumenti *online*, essi potranno essere comunque aggiornati con nuove eventuali acquisizioni. I due anni seguenti saranno dedicati alle analisi lessicali.

In chiave comparatistica romanza, il progetto svolge un percorso per molti aspetti parallelo, nelle premesse come negli sviluppi, a quello promosso e diretto da Frédéric Duval per il francese medievale: la *Base de civilisation romaine (XII^{e}-XV^{e} s.)*.

2. Le banche dati: il corpus *DiVo* e il corpus *CLaVo*

Il *corpus DiVo* raccoglie i volgarizzamenti italoromanzi due-trecenteschi dei testi classici e tardo-antichi, con limite fissato all'opera di Boezio (quindi al VI secolo) e con vincolo all'esaustività per i testi classici. L'esaustività si riferisce alle edizioni giudicate affidabili alla luce delle ricognizioni per la compilazione delle schede di *DiVo DB*; d'altra parte proprio questo lavoro ha imposto o ha reso opportuno l'avvio di lavori preparatori per sostituire alcune delle edizioni esistenti o arrivare alla pubblicazione di testi inediti (cf. § 3)[4].

---

novecentesca (cf. Guadagnini 2013). Il quadro non è diverso nella lessicografia storica francese, in cui la problematizzazione del «lexique de la civilisation romaine» è stata oggetto di attenzione solo recentemente (cf. Duval 2006).

[3] Cf. Folena (1956, XXXVII-XXXVIII): nel discorso di Folena, la formula, mutuata dalla storia dell'arte, si riferisce alla stagione dei volgarizzamenti fiorentini dei testi classici poetici della prima metà del secolo XIV, ma può essere applicata all'intera produzione dei volgarizzamenti dei classici per i secoli XIII e XIV, con l'ovvia avvertenza che una *reductio ad unum* non è possibile.

[4] Cf. Dotto (2013). A fronte dello stato editoriale dei volgarizzamenti, alquanto deficitario, la natura del *corpus DiVo*, in quanto corpus settoriale, ha suggerito una certa inclusività, tesa

Come tutti gli altri *corpora* dell'OVI, esso è consultabile gratuitamente e liberamente in rete, con aggiornamento quadrimestrale, attraverso il software *GATTO 3.3* (*GattoWeb* nella versione *online*) all'indirizzo ‹divoweb.ovi.cnr.it›.

Per i testi classici, occorre precisare che il corpus comprende anche un ristretto numero volgarizzamenti di opere di autori greci, che sono evidentemente mediate da traduzioni (e tradizioni) più tarde, mediolatine. Si tratta per es. dell'*Etica Nicomachea* di Aristotele o delle *Favole* di Esopo: la loro inclusione muove dalla considerazione che anche in questi volgarizzamenti esiste un contatto, sia pure indiretto, con l'universo culturale classico. Non fanno parte del corpus, invece, i volgarizzamenti che derivano da apocrifi, com'è il caso frequente dei testi mediolatini falsamente attribuiti ad Agostino o a Girolamo.

In questa prospettiva, sono inclusi anche i volgarizzamenti che non derivano direttamente dal latino, ma da un'altra varietà romanza, tipicamente il francese, come per es. il volgarizzamento Gaddiano delle *Eroidi*, o da un'altra varietà italiana antica, come per es. l'*Istoria di Eneas* di Angelo di Capua. Non sono esclusi dal corpus neppure le compilazioni o i rimaneggiamenti di materia classica, che, pur contenendo solo frammenti effettivamente tradotti, a causa del loro contenuto risultano significativi per lo studio del classicismo medievale. A questa tipologia appartengono per es. il rimaneggiamento dell'*Eneide* di Guido da Pisa all'interno della *Fiorita d'Italia* (o *Fiore d'Italia*, titolo più diffuso, ma non sostenuto dalla tradizione manoscritta), in cui il rapporto di fedeltà con il testo latino è limitato esclusivamente ad alcune sezioni del testo (in particolare la parte finale), o le redazioni anonime dei *Fatti dei Romani*, ampia compilazione di storia romana di origine antico francese, che contamina fonti diverse (Sallustio, Lucano, Svetonio, ecc.).

Fanno parte del corpus anche i commenti in formato di chiose e i glossari che spesso accompagnano i volgarizzamenti dei testi classici nella tradizione manoscritta. Questa documentazione risulta estremamente utile a livello lessicografico e più in generale storico-culturale per l'analisi della distanza tra la cultura classica e quella medievale e per le reti onomasiologiche che in particolare le chiose di tipo lemmatico permettono di tracciare.

***

Le caratteristiche salienti del corpus sono due:

a) l'associazione paragrafo per paragrafo del testo latino al testo volgare (con un interesse prevalente sui volgarizzamenti dei classici, che saranno associati esaustivamente, mentre l'associazione riguarderà solo una parte dei volgarizzamenti di testi tardo-antichi), in modo da evidenziare al meglio il rapporto tra il testo di partenza e quello d'arrivo;

ad accogliere le edizioni che garantissero una buona affidabilità sul piano della sostanza, accettando consapevolmente, viceversa, lo scarso rispetto della forma del testo, quindi della *facies* fonomorfologica, com'è normale, per es., nella prassi editoriale ottocentesca.

b) un sistema di lemmatizzazione e d'iperlemmatizzazione finalizzato ad agevolare l'interrogazione del corpus nei punti 'sensibili' per lo studio del classicismo medievale[5].

A partire da queste premesse, questi i dati riferiti all'ultimo aggiornamento (giugno 2013), che presentiamo in rapporto al corpus di riferimento, il *corpus OVI dell'Italiano antico*, rispetto al quale il *DiVo* contiene già ora più della metà delle occorrenze provenienti da testi 'nuovi' (cioè non inclusi nel *corpus OVI*):

	corpus DiVo	corpus OVI
N. di testi	119	2.314
N. di occorrenze	4.444.056	23.160.300
N. di forme diverse	145.919	467.190
N. di lemmi	2.256	116.534
N. di occ. lemmatizzate	3.1136	365.3328
N. di iperlemmi	13	–
N. di occ. iperlemmatizzate	3.436	–

Tab. 2

	corpus DiVo 'base'	corpus DiVo 'nuovo'
N. di testi	53	66
N. di occorrenze	2.141.176	2.302.880
N. di forme diverse	100.773	82.029

Tab. 1

Alla fine del lavoro di acquisizione di testi (marzo 2014), il *corpus DiVo* supererà con ogni probabilità i 150 testi e soprattutto i 5.500.000 di occorrenze, per cui esso si consoliderà come uno strumento essenziale per l'integrazione del *corpus OVI*.

Una situazione di questo tipo si giustifica in primo luogo con il fatto che solo da alcuni anni il corpus di riferimento dell'italiano antico ha cominciato ad inserire testi

[5] Il sistema di annotazione è ancora in una fase sperimentale (cf. Dotto 2012): l'obbiettivo è quello di marcare, in particolare con l'iperlemmatizzazione, alcune aree lessicali di àmbito tecnico e 'storico' utili per sondare le forme della comprensione del mondo antico da parte della cultura medievale (con lessico 'storico' intendiamo il lessico privo di una prosecuzione volgare a causa della scomparsa del referente). L'iperlemma è una categoria sovraordinata in grado di raccogliere lemmi che presentano una qualche affinità, semantica, grammaticale, ecc.: così nel *corpus DiVo* all'iperlemma «Culto (pagano)» saranno collegati i nomi delle cariche, degli uffici sacerdotali, dei luoghi di culto e degli oggetti sacri (per es. *auguriato*, *polliere*, *tirso*, ecc.) o i verbi che indicano attività specifiche o anche generiche ma connesse all'àmbito religioso (per es. *auspicare*, *instaurare*, *vaticinare*, ecc.).

posteriori al 1375, visto che tutta la documentazione posteriore alla morte di Boccaccio era stata esclusa in un primo momento; in secondo luogo con la comparsa di edizioni di testi allora inediti, che fatalmente non potevano essere prese in considerazione; infine con la natura settoriale del *corpus DiVo*, per cui, a differenza del *corpus OVI*, esso può contenere anche le cosiddette «edizioni di riserva», cioè edizioni di cui sarebbe auspicabile la sostituzione per i limiti dei criteri adottati e che nondimeno risultano sufficientemente affidabili almeno sul piano dell'attestazione lessicografica[6].

***

Così nella Fig. 1, vediamo come si presenta nella visualizzazione per contesto singolo l'occorrenza della forma *lettiera* in uno dei volgarizzamenti dei *Remedia amoris*. Nella fascia in alto trovano spazio le informazioni bibliografiche sul testo (con l'indicazione, attraverso l'abbreviazione «LAT», della lingua del testo volgarizzato, in modo da distinguere i volgarizzamenti diretti dal latino da quelli con intermediario francese o italoromanzo; inoltre aprendo la scheda bibliografica, è possibile attraverso un link esterno accedere alla relativa scheda di *DiVo DB*); in quella centrale il testo con l'occorrenza del lemma *lettiera*, cui è associato anche l'iperlemma «Cultura materiale»; in quella in basso, che è la più significativa, sono collocati le eventuali chiose associate al testo (che sono interrogabili) e il testo latino corrispondente, dal quale è immediatamente ricavabile che *lettiera* è la traduzione del lat. *lectica* (il volgarizzatore è quindi ricorso a un equivalente volgare a fronte di un lemma, *lectica*, che appartiene al lessico 'storico'). Il brano latino è inoltre accompagnato da note che ambiscono a ricostruire il dettato del testo servito per il volgarizzamento, sia attraverso lo spoglio degli apparati dell'edizione latina di riferimento (per es. la lezione *il giovano v'era* dipenderà dalla lezione testimoniata in alcuni codici della tradizione latina dei *Remedia*, *aderat iuvenis* v. 663, a fronte del testo critico *aderam iuveni*), sia attraverso le ipotesi che l'editore del testo volgare o l'*équipe* del *DiVo* possono formulare (così l'editrice del testo volgare, Vanna Lippi Bigazzi, a fronte del testo *e lle mane delle mane* congettura una possibile lezione – o lettura da parte del volgarizzatore – *et manus e manibus* v. 667 – poiché essa non trova riscontro negli apparati delle edizioni latine, è racchiusa tra due simboli di uguale).

[6] Per la tripartizione delle edizioni, tra edizioni «affidabili», «di riserva» e «inaffidabili» presso l'Ufficio filologico dell'OVI, cf. De Robertis (1985).

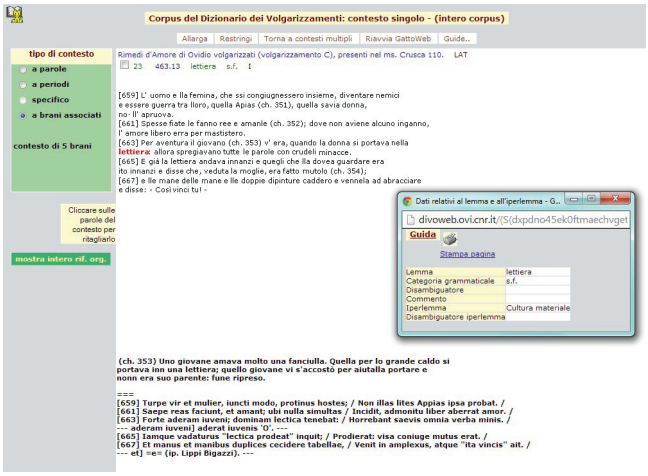


Fig. 1

Accanto al *corpus DiVo*, è stato realizzato il *corpus CLaVo* (*Corpus dei classici latini volgarizzati*), che permette l'interrogazione, per il momento solo per forme, ma da marzo 2014 anche per lemmi, dei testi classici latini presenti nel *corpus DiVo*, con l'associazione paragrafo per paragrafo del testo volgare a quello latino (sono previste alcune aggiunte legate al valore storico-culturale o linguistico del testo, come per es. il volgarizzamento di Boezio di Alberto della Piagentina o le diverse versioni dei *Dialoghi* di Gregorio Magno)[7].

Una particolarità del *CLaVo* è l'ordinamento dei testi, che non segue la cronologia delle opere latine, ma quella dei volgarizzamenti: in questo modo si è inteso dar conto di come siano cambiate nel corso del XIII e XIV secolo le modalità traduttive.

3. La bibliografia filologica (*DiVo DB*) e le nuove acquisizioni

La bibliografia filologica parte dal presupposto che il testo e la sua trasmissione sono elementi non dissociabili né in una visione strettamente filologica né in una visione lessicografica. La bibliografia si propone dunque come uno strumento complementare rispetto al corpus, in grado di fornire per ciascun volgarizzamento tutte le informazioni utili tanto sul testo di riferimento adottato, quanto sullo stato degli studi. A tal fine per ciascuna opera latina e volgare inclusa nel corpus è stata allestita un'apposita scheda (al momento quelle pronte e consultabili sono oltre 150), costruita secondo il modello *TLIon* (*Tradizione della Letteratura Italiana online*), progetto sviluppato dal 2001 da Claudio Ciociola: la bibliografia filologica del *DiVo*, infatti, è

[7] Il *corpus CLaVo* potrebbe essere sfruttato utilmente anche dai filologi classici, considerate le possibilità d'interrogazione del software *Gatto 3.3* e l'estensione di un corpus che conterrà oltre 40 testi per 1.324.727 occorrenze e 121.534 forme distinte.

stata elaborata a partire da *TLIon software* © 4.0 e fa parte della rete *TLIon net*, al pari di altri importanti progetti italiani nati negli ultimi anni che si avvalgono dello stesso software[8].

Le schede preparate per *DiVo*, oltre ad essere raggiungibili tramite link dal corpus, sono liberamente consultabili *online* all'indirizzo ‹tlion.sns.it/divo›. Dalla pagina iniziale, che descrive il progetto, è dunque possibile accedere all'elenco delle schede pubblicate attraverso il pulsante contrassegnato dalla dicitura *DiVo DB*. Nella nuova schermata (Fig. 2) in alto si trova una maschera che consente di effettuare ricerche libere su tutte le schede (per autori, opere, incipit, lingue, manoscritti, bibliografia, ecc.), mentre più in basso sono presenti gli indici alfabetici (delle opere, delle schede, ecc.) che permettono di rintracciare agevolmente tutto il materiale immesso nel sito.

Fig. 2

Passiamo dunque a una breve presentazione della struttura delle schede. Quelle delle opere latine associate ai volgarizzamenti sono sintetiche e contengono soltanto informazioni sull'autore, sulla compilazione e sul genere dell'opera, oltre all'indicazione dell'edizione di riferimento, accompagnata – quando necessario – da bibliografia filologica supplementare.

[8] Per il progetto *TLIon* si veda www.tlion.it, a cui si rimanda anche per i dati tecnici del software. Gli altri progetti che sfruttano le potenzialità di *TLIon software* © 4.0 sono i seguenti: *ENAV – Edizione Nazionale degli Antichi Volgarizzamenti dei testi latini nei volgari italiani* (‹www.ilritornodeiclassici.it/enav›); *ENSU – Edizione Nazionale dei testi della Storiografia Umanistica* (‹www.ilritornodeiclassici.it/ensu›); *CASVI – Censimento, Archivio e Studio dei Volgarizzamenti Italiani* (‹casvi.sns.it›); *SALVIt – Studio, Archivio e Lessico dei Volgarizzamenti italiani* (‹www.salvit.org›); *CSC – Corpus dei Serventesi Caudati* (‹tlion.sns.it/csc›).

Le schede volgari, invece, sono molto più ampie e sono articolate in quattro sezioni principali:

(a) *Notizie generali*: contiene le informazioni generali sull'opera, inclusa datazione, localizzazione linguistica e tipologia testuale;
(b) *Tradizione dell'opera*: include tutti i testimoni manoscritti e a stampa noti attraverso la bibliografia pregressa (ma in alcuni casi sono state fatte apposite ricerche dai redattori delle schede); ciascun manoscritto è accompagnato da una descrizione che fornisce i dati materiali principali del testimone, la datazione, il copista e, ove attingibili, incipit ed explicit del volgarizzamento in esame;
(c) *Storia della tradizione*: costituisce il cuore della scheda, dato che reca l'intera storia della tradizione del volgarizzamento: il curatore fornisce un quadro generale sulla tradizione manoscritta e a stampa dell'opera, descrivendo lo stato attuale degli studi sull'argomento; viene inoltre individuata l'edizione di riferimento;
(d) *Bibliografia*: si trovano tutte le informazioni bibliografiche, articolate per punti (edizione di riferimento, altre edizioni importanti, repertori e cataloghi, altra bibliografia filologica).

Come è facile immaginare, i lavori preparatori per la realizzazione delle schede al servizio del corpus hanno prodotto già qualche tangibile risultato, di cui si darà qui brevemente conto in una rapida panoramica. In primo luogo, vanno menzionate le acquisizioni emerse dalle indagini condotte dai redattori delle schede sul materiale manoscritto: i codici, quando possibile, sono stati descritti accuratamente, per lo più sulla base delle informazioni ricavabili dalla bibliografia pregressa, ma talvolta anche a séguito di nuovi esami autoptici del reperto. Tra i non pochi casi degni di nota, segnalo qui la dettagliata descrizione con tavola analitica del codice Panciatichiano 56 della Biblioteca Nazionale di Firenze (sec. XIV ex.-XV in.), recante la seconda redazione del volgarizzamento delle *Epistole morali a Lucilio* di Seneca[9]. Il Panciatichiano è peraltro un manoscritto dalla storia affascinante, poiché fu utilizzato dal Salviati degli *Avvertimenti* prima e dalla Crusca poi proprio per citare il Seneca volgare.

L'allestimento del *corpus DiVo* ha posto inoltre un delicato problema: quello della datazione di molte opere. Come noto, numerosi volgarizzamenti antichi presentano datazioni incerte, mancando indizi che permettano collocazioni cronologiche più precise. Spesso poi le edizioni sette e ottocentesche forniscono datazioni alte (soprattutto relative al sec. XIV) anche per testi in verità molto più tardi (si pensi al volgarizzamento delle *Tusculane*, traduzione cinquecentesca di Fausto da Longiano, ma considerato trecentesco da Francesco Zambrini)[10], poiché si fondano su perizie

[9] La descrizione, che si può visualizzare all'interno della scheda del volgarizzamento delle *Epistole*, è a cura di Cristiano Lorenzi Biondi, che sta anche approntando per il *corpus* una trascrizione dal codice dell'intero volgarizzamento, dato che la seconda redazione – che costituisce la vulgata del testo, ed è dunque fondamentale per la storia della sua ricezione – risultava ancora inedita.

[10] Cf. Zambrini (1884[4], 269). Della traduzione di Fausto da Longiano (sul quale cf. la voce eponima di Pignatti [1995]) non abbiamo edizioni moderne, per cui si deve ricorrere alla cinquecentina (Fausto da Longiano [1544]).

linguistiche e codicologiche piuttosto imprecise, non di rado condizionate ideologicamente dalla volontà di fornire ai lettori un testo del 'buon secolo'. Compito dei curatori delle schede è stato dunque anche quello di verificare, per quanto possibile, tali datazioni e di proporne eventualmente altre più attendibili.

Tra le numerose nuove collocazioni cronologiche emerse, un caso particolarmente interessante – anche perché si tratta di una notevole retrodatazione – è quello del volgarizzamento delle *Collazioni dei Santi Padri* di Cassiano: sul fondamento del manoscritto base utilizzato dall'editore ottocentesco Telesforo Bini (Lucca, Biblioteca Statale, 1637), datato 1442[11], il testo si riteneva della prima metà del Quattrocento o al limite genericamente trecentesco (così per il *Grande Dizionario della Lingua Italiana* e per il *TLIO*), mentre si tratta in realtà di un volgarizzamento molto più antico, risalente ancora al Duecento, poiché è conservato, tra gli altri, dal manoscritto I V 8 della Biblioteca Comunale degli Intronati di Siena, databile al sec. XIII ex.[12]. Si aggiunga poi che nella stessa biblioteca è presente un secondo testimone (I VI 38) datato 1387.

Le edizioni sette-ottocentesche pongono talvolta un altro problema: l'assenza di indicazioni sul manoscritto di base utilizzato per la trascrizione. Ciò capita ad esempio nell'ed. Mattioli (1888) del quattrocentesco volgarizzamento delle *Confessioni* di Agostino, attribuito senza fondamento a Giovanni da Salerno (morto nel 1388): l'editore non dà alcuna informazione sul testimone da lui prescelto, riportando solo la sottoscrizione con data 1453[13]. Il codice, stimato perduto anche da De Luca e Baldassarri[14], è stato invece ritrovato: si tratta del ms. 856 della Biblioteca Nazionale Centrale Vittorio Emanuele II di Roma.

L'analisi della storia della tradizione che trova posto nelle schede ha come fine primario l'indagine più propriamente filologica allo scopo di individuare l'edizione di riferimento per il corpus, ovvero il testo con maggiori garanzie di documentazione testuale. In un certo numero di casi, a fronte di edizioni datate ritenute scarsamente affidabili (soprattutto a causa di evidenti interpolazioni o arbitrari interventi da parte dell'editore), si è reso necessario promuovere nuove edizioni proprio in vista dell'inclusione nel *corpus DiVo*. Così è avvenuto per i volgarizzamenti delle orazioni *Pro Marcello* e *Pro rege Deiotaro* a opera di Brunetto Latini, dei quali è stata allestita una nuova edizione critica da parte di chi scrive a sostituzione dell'ed. Rezzi (1832). L'edizione, che ha tenuto conto di tutta la tradizione manoscritta delle due 'dicerie' (una decina di testimoni), è già interrogabile *online*[15].

---

[11] Cf. Bini (1854); il codice è datato e sottoscritto a c. 226*v*: «Compiuto di scrivere a di xxviiij° di gienaio M cccc xlij», come riporta lo stesso Bini (1854, 316).

[12] Per la descrizione e la datazione del manoscritto cf. Manetti / Savino (1990, 164-166). In ragione di questo importante ritrovamento, è stato deciso di approntarne un'edizione a uso interno a cura di Andrea Felici, che sarà inserita a breve nel *corpus DiVo*.

[13] Cf. Mattioli (1888, 300).

[14] Cf. De Luca (1954, 499) e Baldassarri (1995, 256).

[15] I volgarizzamenti delle due orazioni circolano pressoché sempre appaiati e le dinamiche di trasmissione degli errori si muovono secondo le stesse direttrici, tanto che si dovrà pensare a

Infine, si dovrà segnalare l'immissione nel corpus anche di alcuni testi sinora inediti, come ad esempio il volgarizzamento della *Consolatio ad Polybium* di Seneca, conservato in attestazione unica nel ms. Laurenziano plut. 76.61. Le altre due consolazioni in volgare (a Marcia e a Elvia), tràdite oltre al Laurenziano da altri tre codici (Vaticano Urbinate lat. 1142, Vaticano Rossiano 401 e Casanatense 117), erano già note e edite fin dall'Ottocento[16], mentre quella a Polibio, certamente della stessa mano, era stata solo segnalata da Lara Nicolini nella scheda del codice contenuta nel catalogo della mostra di manoscritti ed edizioni su Seneca del 2004, ma mai pubblicata. La trascrizione dal codice Laurenziano, ad opera di chi scrive, è dunque consultabile *online* nel *corpus DiVo*[18].

Queste sono quindi alcune delle principali acquisizioni che *a latere* del progetto *DiVo* stanno via via vedendo la luce, ma altre è lecito attendersi col prosieguo del lavoro: in un campo come quello dei volgarizzamenti, dove ancora molto resta da scavare, specie in àmbito filologico, ogni indagine apre infatti nuovi orizzonti e infiniti àmbiti di ricerca.

Firenze, Istituto Opera del Vocabolario Italiano (CNR) Diego DOTTO
Pisa, Scuola Normale Superiore Cristiano LORENZI

Bibliografia

Baldassarri, Guido, 1995. «Letteratura devota, edificante e morale», in: Malato, Enrico (ed.), *Storia della Letteratura Italiana*, Roma, Salerno Ed., 14 vol., vol. III, 201-326.

Base de civilisation romaine (XII^e-XV^e s.): Duval, Frédéric (ed.), *Base de civilisation romaine (XII^e-XV^e s.)*, CNRTL CNRS-ATILF. ‹www.cnrtl.fr/lexiques/civirom/›.

Bibliografia dei volgarizzamenti: Artale, Elena (ed.), *Bibliografia dei volgarizzamenti [del corpus TLIO]*. ‹http://tlio.ovi.cnr.it/BibVolg/›.

un'unica tradizione per entrambi i testi. I testimoni sono i seguenti: Firenze, Biblioteca Medicea Laurenziana, Conventi Soppressi 122 (solo la *Pro Marcello*); Firenze, Biblioteca Nazionale Centrale, II II 23; Firenze, Biblioteca Riccardiana, 1538; ivi, 1563; Kórnik, Polska Akademia Nauk, Biblioteka Kórnicka, 633; London, British Library, Additional 16437; Milano, Biblioteca Ambrosiana, I 166 inf.; Napoli, Biblioteca Nazionale «Vittorio Emanuele III», XIV D 17; Città del Vaticano, Biblioteca Apostolica Vaticana, Chigiano L VII 267; Venezia, Biblioteca Nazionale Marciana, Italiano VIII 26 (= 6090); ad essi va inoltre aggiunta la testimonianza della *princeps* Corbinelli (1568), che deriva da un codice oggi perduto. Il testo delle due orazioni si può leggere (accompagnato da un'ampia nota filologica) anche in Lorenzi, 2013.

16 Spezi (1866), che fondava la propria edizione sul Vaticano Urbinate.

17 Cf. De Robertis / Resta (2004, 283-284).

18 Il testo è stato contestualmente edito a stampa: cf. Lorenzi (2012).

Bini, Telesforo (ed.), 1854. *Volgarizzamento delle collazioni dei SS. Padri del venerabile Giovanni Cassiano*, Lucca, Giusti.

Corbinelli, Jacopo (ed.), 1568. *L'Ethica d'Aristotile ridotta in compendio da ser Brunetto Latini*, Lione, Per Giovanni de Tornes.

corpus CLaVo: Burgassi, Cosimo / Dotto, Diego / Guadagnini, Elisa / Vaccaro, Giulio (ed.), *Corpus dei Classici Latini Volgarizzati*. ‹clavoweb.ovi.cnr.it/›.

corpus DiVo: Burgassi, Cosimo / Dotto, Diego / Guadagnini, Elisa / Vaccaro, Giulio (ed.), *Corpus del Dizionario dei Volgarizzamenti*. ‹divoweb.ovi.cnr.it›.

corpus OVI: Artale, Elena / Larson, Pär (ed.), *Corpus OVI dell'Italiano antico*. ‹gattoweb.ovi.cnr.it›.

De Luca, Giuseppe (ed.), 1954. *Prosatori minori del Trecento*, I. *Scrittori di religione*, Milano / Napoli, Ricciardi.

De Robertis, Domenico, 1985. «L'Ufficio filologico dell'Opera del Vocabolario, il suo impianto, il suo lavoro», in: *La Crusca nella tradizione letteraria e linguistica italiana*, Firenze, presso l'Accademia, 443-451.

De Robertis, Teresa / Resta, Gianvito (ed.), 2004. *Seneca: una vicenda testuale. Mostra di manoscritti ed edizioni, (Firenze, Biblioteca Medicea Laurenziana, 2 aprile-2 luglio 2004)*, Firenze, Mandragora.

DiVo DB: Guadagnini, Elisa / Vaccaro, Giulio (ed.), *DiVo – Bibliografia filologica*. ‹tlion.sns.it/divo›.

Dotto, Diego, 2012. «Note per la lemmatizzazione del *corpus DiVo*», *BOVI* 17, 339-366.

Dotto, Diego, 2013. «Notizie dal *DiVo*. Un primo bilancio sulla costituzione del corpus», in: Larson, Pär / Squillacioti, Paolo / Vaccaro, Giulio (ed.), *«Diverse voci fanno dolci note». L'Opera del Vocabolario Italiano per Pietro G. Beltrami*, Alessandria, Ed. dell'Orso, 71-93.

Duval, Frédéric, 2006. «Pour la révision des mots de civilisation romaine du Trésor de la langue française (informatisé)», in: E. Buchi (ed.), *Actes du Séminaire de méthodologie en étymologie et histoire du lexique (Nancy/ATILF, année universitaire 2005/2006)*, Nancy, ATILF (CNRS / Université Nancy 2/UHP). ‹www.atilf.fr/atilf/seminaires/Seminaire_Duval_2006-11.pdf›.

Fausto da Longiano (ed.), 1544. *Le Tusculane di M. Tullio Cicerone recate in italiano*, Venezia, appresso Vicenzo Vaugris al segno d'Erasmo.

Folena, Gianfranco (ed.), 1956. *La istoria di Eneas vulgarizata per Angilu di Capua*, Palermo, Centro di studi filologici e linguistici siciliani.

Guadagnini, Elisa, 2013. «Notizie dal *DiVo*. Parole tradotte e lessicografia dell'italiano», in: Larson, Pär / Squillacioti, Paolo / Vaccaro, Giulio (ed.), *«Diverse voci fanno dolci note». L'Opera del Vocabolario Italiano per Pietro G. Beltrami*, Alessandria, Ed. dell'Orso, 59-70.

Lorenzi, Cristiano, 2012. «Un volgarizzamento inedito della 'Consolatio ad Polybium' (ms. Laurenziano Plut. 76.61)», *BOVI* 17, 221-243.

Lorenzi, Cristiano, 2013. «Le orazioni 'Pro Marcello' e 'Pro rege Deiotaro' volgarizzate da Brunetto Latini», *SFI* 71, 19-77.

Manetti, Roberta / Savino, Giancarlo, 1990. «I libri dei Disciplinati di Santa Maria della Scala di Siena», *Bullettino senese di storia patria*, 97, 122-192.

Mattioli, Nicola (ed.), 1888. *Antico volgarizzamento delle Confessioni di S. Agostino*, Roma, Tip. poliglotta della S. C. de propaganda fide.

Pignatti, Franco, 1995. Voce «Fausto (Fausto da Longiano)», in: *Dizionario Biografico degli Italiani*, Roma, Istituto della Enciclopedia Italiana, vol. XLV, 394-397.

Rezzi, Lugi Maria (ed.), 1832. Le *tre orazioni di Marco Tullio Cicerone dette dinanzi a Cesare per M. Marcello, Q. Ligario e il re Dejotaro volgarizzate da Brunetto Latini*, Milano, Fanfani.

Spezi, Giuseppe (ed.), 1866. *Volgarizzamento inedito della Consolazione di Lucio Anneo Seneca ad Elvia ed a Marcia*, Roma / Torino, Poliglotta-Pontificia.

Zambrini, Francesco (ed.), 1884[4]. *Le opere volgari a stampa dei secoli XIII e XIV*, Bologna, Zanichelli.

Détection, extraction automatique et analyse des lexèmes construits par la préfixation en *non-* en français

1. Introduction

Ce travail vise à mettre à contribution linguistique théorique et Traitement Automatique des Langues (TAL), dans le cadre d'une thèse consacrée à l'étude du préfixe[1] de négation *non-*. En français contemporain, le préfixe *non-* peut s'adjoindre à des bases nominales (p.ex. NON-CONFORMITÉ, NON-RESPECT) et adjectivales (p.ex. NON-FERREUX, NON-PRODUCTIF), ce qu'illustrent (1) et (2) :

(1) Rien n'est plus facile que de constater la conformité de l'écriture d'un texte, ou sa non-conformité, avec
l'orthographe légale. (*TLFi, s.v.* 'non-conformité')

(2) Le luxe consiste essentiellement dans les dépenses non-productives, quelle que soit d'ailleurs la nature de ces
dépenses. (*TLFi, s.v.* 'non(-)')

J'expose dans cet article la méthodologie employée pour détecter, extraire automatiquement et analyser les lexèmes construits sur base nominale et adjectivale à l'aide du préfixe *non-* en français dans un corpus ouvert, à savoir les pages en français indexées par le moteur de recherche *Google*™. Un de mes objectifs est de comparer ces données avec les lexèmes préfixés par *non-* attestés dans le *Trésor de la Langue Française informatisé* (*TLFi*)[2], l'hypothèse sous-jacente étant que les dictionnaires ne reflètent que très partiellement l'état actuel du français (cf. notamment Dal & Namer 2012). En effet, dans le cas qui m'intéresse ici, il semble que les dictionnaires ne recensent qu'une infime partie des lexèmes préfixés par *non-* effectivement utilisés par les locuteurs et la préfixation par *non-* ne paraît soumise qu'à peu de contraintes (concernant par exemple les bases nominales, voir Dugas 2012). D'autre part, ce travail me permet de disposer d'un lexique aussi exhaustif que possible des lexèmes formés par la préfixation en *non-*, dans la perspective de l'examen des caractéristiques de cette préfixation et des lexèmes qu'elle permet de produire.

L'article est organisé en trois temps. Dans un premier temps, je décris la façon dont a été constitué mon corpus et comment s'est faite la détection des formes can-

---

1 Par commodité, je parlerai dans cet article du « préfixe » non-, même si son statut préfixal est sujet à questionnement lorsqu'il s'adjoint à des adjectifs (Dugas 2013).

2 ‹http://atilf.atilf.fr/›

didates. La seconde partie propose une analyse des fréquences obtenues. En conclusion, je rappelle les principaux points de ce travail et j'esquisse quelques perspectives de recherche.

2. Constitution du corpus et détection des formes

2.1. Constitution du lexique de référence

Mon lexique de référence est *Morphalou*[3], lexique ouvert des formes fléchies du français (Romary *et al.* 2006), construit à partir de la nomenclature du *TLFi*. J'ai extrait de ce lexique les formes nominales et adjectivales, soit 82 067 formes : 59 334 formes attestées comme noms et 22 733 formes attestées comme adjectifs. Il faut noter ici que parmi ces 82 067 formes, 6197 possèdent une double catégorisation et sont attestées la fois comme noms et comme adjectifs. Enfin, le lexique *Morphalou* contient 190 lexèmes déjà préfixés par *non-* : 147 noms (p.ex. NON-ACCEPTATION, NON-CONFORMITÉ, NON-INGÉRENCE), 30 adjectifs (p.ex. NON-PRÉDICATIF, NON-REMBOURSABLE, NON-VÉNÉNEUX), et 13 lexèmes avec la double catégorisation nom et adjectif (p.ex. NON-COMMUNISTE, NON-JUIF, NON-LIBÉRAL).

2.2. Génération des lexèmes potentiels préfixés par non- à partir du lexique de référence

L'exploitation du lexique *Morphalou* m'a permis de générer une liste de formes candidates au moyen du composant de prédiction morphologique « non ». Le préfixe *non-* a l'avantage de ne pas posséder d'allomorphes en français contemporain, contrairement à d'autres préfixes : c'est le cas par exemple de *in-,* très productif, qui a pour allomorphes *il-, im-, ir* : LÉGAL > ILLÉGAL, MANGEABLE > IMMANGEABLE, RATIONNEL > IRRATIONNEL ; c'est aussi le cas du préfixe *a-*, qui possède l'allomorphe *an-* devant voyelle : ORGANIQUE > ANORGANIQUE, ENCÉPHALE > ANENCÉPHALE. Ainsi, le préfixe *non-* a une forme stable, quelle que soit le lexème auquel il s'adjoint. De plus, il n'entraîne pas de modification morphologique de sa base. La génération des formes candidates est donc facilitée, puisque la chaîne de caractères à ajouter en début de mot est invariable, et qu'il n'est pas besoin d'adapter la chaîne de caractères composant la forme de base.

Toutefois, la question s'est posée de l'orthographe à choisir pour la génération des lexèmes potentiels : *non-* est-il soudé graphiquement à sa base, ou doit-il en être séparé par un trait d'union, ou par une espace typographique ? La plupart des préfixes du français sont graphiquement soudés à leur base (*a-* : AMORAL ; *in-* : IMPOSSIBLE ; *pré-* : PRÉCOLOMBIEN, etc.) le préfixe *non-* est particulier puisque dans certaines attestations il est séparé de sa base par un trait d'union ou par une espace. Le *TLFi* n'a pas une position claire sur la présence du trait d'union dans les formes en *non-* : certes, « le trait d'union est quasi systématique pour les substantifs; dans le corpus

[3] ‹http://www.cnrtl.fr/lexiques/morphalou/›

littéraire du *TLF*, proportion de l'ordre de 1 ou 2 % d'exceptions » ; toutefois, pour les adjectifs « la règle est l'absence du trait d'union, mais il y a des hésitations. Souvent aussi, notamment dans le discours philosophique, le trait d'union marque une liaison conceptuelle » (*TLFi, s.v.* 'non(-)').

Dans la pratique, si l'on observe de plus près les formes attestées, on rencontre trois cas de figure : (i) présence d'un trait d'union ; (ii) *non-* détaché de la séquence droite, mais sans trait d'union ; (iii) *non-* soudé graphiquement à la séquence droite. On remarquera que le trait d'union est plus systématique devant les bases nominales (p.ex. NON-LINGUISTIQUE, NON-REMBOURSEMENT) que devant les bases adjectivales (p.ex. NON-PRÉDICATIF, NON-SYNDICAL). Au sein de la classe adjectivale, on trouve plus souvent un trait d'union devant les adjectifs typiquement adjectivaux, c'est-à-dire les adjectifs simples et les adjectifs construits par suffixation (p.ex. NON-COMMERCIAL, NON-FERREUX), que devant les formes apparentées à des participes passés (p.ex. NON-ADAPTÉ, NON-DÉFINI). Il est possible que les locuteurs perçoivent les noms précédés de *non-* comme des unités lexicales, autrement dit comme des mots à part entière, où *non-* est un véritable préfixe. L'orthographe plus variable (présence moins systématique du trait d'union) des adjectifs précédés de *non-* témoignerait alors du fait que le statut de ces formes est peu clair pour le locuteur, *a fortiori* lorsque l'adjectif est apparenté à un participe passé, où la frontière entre catégorie adjectivale et verbale et entre lexique et syntaxe paraît brouillée.

D'autre part, les fluctuations orthographiques s'observent non seulement entre différentes formes, mais aussi pour une même forme : ainsi on pourra trouver à la fois dans les corpus des attestations de *non-progrès, non progrès* et *nonprogrès* (Dugas 2012). Toutefois, l'examen d'un échantillon de données a révélé que la fréquence des formes où *non-* est graphiquement soudé à sa base était négligeable (de l'ordre de quelques centaines d'occurrences) et par conséquent, il a été décidé que seules les formes où *non-* est lié à sa base par un trait d'union et celle où il est séparé de sa base par une espace typographique devaient faire l'objet d'une requête. Étant donné qu'une requête comportant un trait d'union (par exemple « non-linguiste ») permet aussi de recueillir des formes sans trait d'union (« non linguiste »), le format de requête retenu fut le suivant : « *non*-mot »; ceci est illustré dans le tableau 1 ci-après.

Formes attestées dans Morphalou	*Requêtes*	*Formes collectées par le moissonneur*
linguistique	non-linguistique	non-linguistique; non linguistique
lingule	non-lingule	non-lingule ; non lingule
lingère	non-lingère	non-lingère ; non lingère
linier	non-linier	non-linier ; non linier
liniment	non-liniment	non-liniment ; non liniment

Tableau 1. Format de requête et formes collectées.

2.3. Vérification automatique de l'attestation sur Google™ des lexèmes générés

Une fois les lexèmes générés à partir du corpus de référence, j'ai utilisé un moissonneur (Berland & Grabar 2001) qui a automatisé l'interrogation des pages en français indexées par le moteur de recherche *Google™*. L'interrogation des pages s'est faite en avril et mai 2013. Les requêtes ont été protégées avec des guillemets («») afin que ce soit l'ensemble « *non*-mot », et uniquement cet ensemble, qui soit l'objet de la requête (évitant ainsi que, par exemple, une requête de format « non-linguistique » recueille des formes contenant « linguistique » mais pas « non »). Le moissonneur m'a donc permis de repérer toutes les formes graphémiques commençant par la chaîne de caractères « non » générées à partir du corpus de référence, où « non » peut être suivi d'une espace ou d'un trait d'union.

Le moissonneur m'a fourni deux types de résultats: une sortie qui indiquant la fréquence ainsi que la première URL d'attestation (3), et une autre sortie n'indiquant que la fréquence (4) :

(3) google|lang_fr|2|»»»»non-linguistique»»» - Recherche Google»|Environ 35 000|
google|lang_fr|»»non-linguistique»»|1|http://eurofle.wordpress.com/2008/08/19/le-linguistique-et-le-non- linguistique-en-intercomprehension/
google|lang_fr|2|»»»»non-lingule»»» - Recherche Google»|Environ 0|
google|lang_fr|»»non-lingule»»|0

(4) non-linguistique|35 000
non-lingule|0

Ainsi, au moment de l'interrogation par le moissonneur, « non-linguistique » a recueilli 35 000 attestations tandis que « non-lingule » n'en a recueilli aucune.

2.4. Traitement du bruit

Il était nécessaire de procéder à un tri manuel afin de filtrer les formes candidates. En effet, une requête avec le mot « non » génère un bruit important. Un travail antérieur portant sur les lexèmes nominaux préfixés par *non-* a montré que sur un corpus de 1 601 562 mots issu de la Toile, seules 3.53% des occurrences de « non » correspondaient à des noms préfixés par *non-* (Dugas 2012). Je vais ici indiquer les principales causes de bruit (cf. Fradin *et al.* 2008 et Hathout *et al.* 2009 pour une discussion générale sur l'apprêt des données issues de la Toile).

Une des particularités du préfixe *non-* en français est qu'il est homographe de l'adverbe de phrase *non,* qui est une forme libre employée en syntaxe. D'autre part, comme indiqué plus haut, les requêtes ont été protégées avec des guillemets ; mais il a fallu prendre en compte le fait que l'espace entre les guillemets vaut pour tout caractère alphanumérique, dont la ponctuation. Par conséquent, le bruit a majoritairement été causé par les formes syntaxiques, que le robot d'extraction ne peut pas discriminer (5), et par la ponctuation (6) :

(5) La propagande a un effet de vérité mobilisateur, mais non émancipateur.
(6) Je m'attendais plutôt à mâcher un bout de poulet sec, mais non. Surprenant.

Deuxièmement, malgré une requête sur les pages françaises uniquement, les données recueillies par le robot d'indexation contenaient quelques occurrences de la forme *non* qui, sans être des emprunts, appartiennent à d'autres langues que le français. Il s'agit notamment de formes de l'anglais, de l'italien et du latin, où existe une forme *non* homographe du *non* français.

D'autre part, le serveur de reconnaît pas les accents. Ainsi, par exemple, une requête de forme « non sûr » a recueilli *non sûr* mais également *non sur*, où *sur* sans accent circonflexe est une préposition, et où *non* est un adverbe. De la même façon, les fréquences des formes obtenues avec la requête « non épouvante » (où *non-* s'adjoint au nom ÉPOUVANTE) comprennent également les occurrences de *non épouvanté* (où *non-* s'adjoint à l'adjectif ÉPOUVANTÉ).

Enfin, les productions langagières issues de la Toile peuvent contenir de nombreuses coquilles et fautes d'orthographe, ce dont il a également fallu tenir compte. Fait bien connu, cette fluctuation orthographique est particulièrement avérée dans les contextes non normés, où les scripteurs ont tendance à s'exprimer très rapidement, sans relecture : messages postés sur les forums de discussion, commentaires de lecteurs sur les sites de presse, etc. Par exemple, la forme *non épaule* dans un énoncé comme *Charles Aznavour chante Au creux de non épaule à l'Olympia* est un faux positif. Au contraire, les articles de journaux et les textes publiés sur des sites administratifs ou officiels (sites gouvernementaux, etc.) contiennent peu de coquilles. D'autre part, et pour faire le lien avec le point précédent, on note une forte tendance à l'omission des accents ou à leur emploi inapproprié (p.ex. *non-progres* ou *non-progrés* pour *non-progrès*), mais cela n'a aucune incidence sur le bruit étant donné que les accents ne sont pas reconnus.

Étant donné l'ampleur du bruit et l'impossibilité de le corriger, du fait de la quantité considérables de données, le choix a été fait de répartir les statistiques obtenues dans 12 ensembles de fréquences : < 500 occurrences ; 500 occ. > 1000 occ. ; 1000 occ. > 5000 occ. ; 5000 occ. > 10 000 occ. ; 10 000 occ. > 50 000 occ. ; 50 000 occ. > 100 000 occ. ; 100 000 occ. > 500 000 occ. ; 500 000 occ. > 1 million occ. ; 1 million occ. > 5 millions occ. ; 5 millions occ. > 10 millions occ. ; 10 millions occ. > 50 millions occ ; 50 millions occ. > 100 millions occ. ; > 100 millions occ. L'analyse des fréquences a ensuite reposé sur la comparaison des statistiques de chaque ensemble.

3. Analyse des fréquences

3.1. Les fréquences

Le tableau 2 indique les fréquences d'attestation sur *Google™* des lexèmes préfixés générés.

Fréquences	*Noms*	*Ratio/N*	*Adjectifs*	*Ratio/Adj*	*Total N+Adj*	*% N+Adj*
≥ 100M	7	0.01%	3	0.01%	10	0.01%
≥ 50M	9	0.01%	0	0%	9	0.01%
≥ 10M	38	0.06%	13	0.05%	51	0.06%
≥ 5M	45	0.07%	11	0.04%	56	0.06%
≥ 1M	213	0.36%	121	0.53%	334	0.41%
≥ 500 000	222	0.37%	116	0.51%	338	0.41%
≥ 100 000	1153	1.94%	617	2.71%	1770	2.16%
≥ 50 000	1020	1.71%	460	2.02%	1480	1.80%
≥ 10 000	3942	6.64%	1948	8.56%	5890	7.18%
≥ 5000	2227	3.75%	1121	4.93%	3348	4.08%
≥ 1000	5488	9.25%	2557	11.24%	8045	9.80%
≥ 500	2569	4.32%	1116	4.90%	3685	4.49%
< 500	*42 401*	*75.45%*	*14 650*	*64.4%*	*57051*	*69.52%*
Total	*59 334*	*100%*	*22 733*	*100%*	*82 067*	*100%*

Tableau 2. Fréquence sur *Google™* des lexèmes préfixés générés (12 ensembles).

Le tableau 3 donne un compte-rendu simplifié des résultats, regroupés en 4 grands ensembles de fréquences :

Fréquence	*Noms*	*Ratio/N*	*Adjectifs*	*Ratio/Adj*	*Total N+Adj*	*% N+Adj*
≥ 500 000	534	0.9%	264	1.2%	798	1%
≥ 50 000	2173	3.7%	1077	4.7%	3250	3.9%
≥ 5000	6169	10.4%	3069	13.5%	9238	11.3%
< 5000	*50 458*	*85%*	*18 323*	*80.6%*	*68 781*	*83.8%*
Total	*59 334*	*100%*	*22 733*	*100%*	*82 067*	*100%*

Tableau 3. Fréquence sur *Google™* des lexèmes préfixés générés (4 ensembles).

On constate que la plupart des formes générées sont très peu fréquentes : 69.52% des lexèmes générés recueillent moins de 500 occurrences (tableau 2), et 83.8% des lexèmes générés recueillent moins de 5000 occurrences (tableau 3). Une comparaison de ces statistiques avec celles que l'on obtient sur *Google™* pour les lexèmes déjà

préfixés par *non-* attestés dans le *TLFi* (tableau 4, ci-dessous) montre que ces derniers recueillent un plus grand nombre d'attestations sur la Toile que les formes préfixées générées : 43.7% des lexèmes préfixés par *non-* attestés dans le *TLFi* comptent plus de 5000 attestations, contre seulement 11.3% des lexèmes préfixés générés. Ceci va à l'encontre de l'hypothèse selon laquelle les dictionnaires ne reflètent pas l'état actuel de la langue. Toutefois, on peut objecter que l'écart de fréquence entre les lexèmes générés et les lexèmes déjà préfixés par *non-* dans le *TLFi* est relativement minime et que les lexèmes attestés dans le *TLFi* ne recueillent pas des fréquences très élevées.

Fréquence	*Noms*	*Ratio/N*	*Adjectifs*	*Ratio/Adj*	*Total N+Adj*	*% N+Adj*
≥ 500 000	20	13.6%	5	11.6%	25	13.2%
≥ 50 000	50	34%	*19*	*44.2%*	69	36.3%
≥ 5000	*66*	*44.9%*	17	39.5%	*83*	*43.7%*
< 5000	11	7.5%	2	4.7%	13	6.8%
Total	*147*	*100%*	*43*	*100%*	*190*	*100%*

Tableau 4. Fréquence sur *Google™* des *non*-N et des *non*-Adj attestés dans le *TLFi*.

D'autre part, 25% des *non*-N et des *non*-Adj générés ne sont pas attestés dans *Google™* au moment de la requête, alors que leur forme non préfixée est attestée. Il s'agit principalement de mots construits (≈ 95% du total des lexèmes générés qui ne recueillent aucune occurrence). Le tableau 5 donne des exemples de ces lexèmes générés qui ne recueillent aucune occurrence sur *Google™*, et indique la fréquence, sur ce même corpus, des bases de ces lexèmes. Ainsi, alors que NON MULTICOQUE ne recueille aucune attestation, MULTICOQUE est tout à fait fréquent puisqu'il recueille 1 670 000 occurrences.

Lexème préfixé	*Fréquence sur* Google™	*Lexème-base*	*Fréquence sur* Google™
NON-MULTICOQUE	0	MULTICOQUE	1 670 000
NON-ORANGERAIE	0	ORANGERAIE	458 000
NON-ENFIÉVRÉ	0	ENFIÉVRÉ	197 000
NON-BALEINIER	0	BALEINIER	194 000
NON-INTERSYNDICAL	0	INTERSYNDICAL	166 000
NON-AFFECTUOSITÉ	0	AFFECTUOSITÉ	139 000
NON-DÉGRINGOLANT	0	DÉGRINGOLANT	125 000
NON-VIDÉOTRANSMISSION	0	VIDÉOTRANSMISSION	113 000

Tableau 5. Exemples de fréquences des formes générées vs. fréquences des bases.

Le cas des lexèmes générés ne recueillant aucune occurrence est tout à fait intéressant. La productivité du préfixe *non-*, entendue comme l'aptitude de la préfixation en *non-* à former de nouveaux lexèmes (cf. Corbin 1987) a été soulignée à plusieurs reprises dans la littérature (Jespersen 1917, Zimmer 1964, Gaatone 1971, Hamawand 2009, Dugas 2012 entre autres), suggérant, d'une part, qu'il n'y aurait pas de restriction à la préfixation en *non-*, c'est-à-dire que n'importe quel lexème (quelles que soient ses caractéristiques morphologiques et phonologiques) pourrait être l'input d'une préfixation par *non-* ; d'autre part, que plus un lexème est fréquent, plus sa forme négative en *non-* serait fréquente. Or la non-attestation de certaines formes en *non-* est un argument en faveur de l'existence de contraintes à la préfixation en *non-* en français contemporain. Tout d'abord, on peut légitimement supposer que les lexèmes les moins fréquents, parce qu'ils appartiennent soit à un domaine de langue spécialisé, soit à un dialecte régional, ou parce que leur référent n'est plus courant aujourd'hui, ont peu de probabilité d'apparaître avec le préfixe *non-*. Ainsi, ne recueillent aucune occurrence des lexèmes comme NON-ABERROGRAPHE, NON-GOUTTIER, NON-GRISOUMÈTRE. Toutefois, NON-ACIÉRISTE, NON-ACROBATISME, NON-AFRICANISME, NON-GIFLEUR, NON-MULTICOQUE, NON-ORANGERAIE, NON-PRÉFIGURATEUR, dont les lexèmes bases sont pourtant *a priori* plus fréquents que ne le sont ABERROGRAPHE, GOUTTIER et GRISOUMÈTRE, ne sont pas attestées dans *Google*^{TM,}. Se pose donc la question des facteurs qui entrent en jeu dans leur apparente incapacité à prendre la préfixation en *non-* - mais ceci dépasse les limites de cet article.

3.2. Tri manuel et examen des formes les plus fréquentes

Dans un second temps, je me suis focalisée sur les 798 formes recueillant plus de 500 000 occurrences et j'ai procédé à leur tri manuel. Les formes éliminées l'ont été dans les cas suivants :

(i) Lorsque le moissonneur a retenu le nombre d'occurrences de la forme suggérée par *Google*^{TM}. Il arrive en effet que le moteur de recherche suggère des formes dont la fréquence est beaucoup plus importante que la forme qui fait l'objet de la requête : dans mon cas, il s'est agi par exemple de *non-gage* pour *non-aiguage*, *non-adhésive* pour *non-athésie*, *non-fiction* pour *non-coction*. Parfois, la forme suggérée était un lexème négatif préfixé par *in-* : *incertitude* à la place de *non-certitude*, *inquiétude* à la place de *non-quiétude*, *invertébré* à la place de *non-vertébré*. Ceci est d'ailleurs intéressant du point de vue de la fréquence respective, pour une même base, d'un dérivé en *non-* et d'un dérivé en *in-*. D'autre part, le moteur de recherche a parfois renvoyé une forme dont la première chaîne de caractère était *in,* non pas le préfixe de négation, mais le préfixe homographe à valeur inchoative : par exemple, *incarnation* à la place de *non-carnation*, *incinération* pour *cinération*, *infloresence* pour *non-florescence*.

(ii) Les formes qui ont un homographe très fréquent en anglais, italien etc. Il s'agit notamment de *non-agricultural, non-business*, *non-living-room*, *non-credo, non-mi, non-perché*.

(iii) Les formes qui correspondent très souvent à des expressions en syntaxe, par exemple *non car, non-même, non merci, non-pour-soi, non-son*. Ces groupes de mots peuvent être employés en réponse à une question, où *non* est employé comme adverbe, par exemple *viens, on tombe amoureux ? non car je suis déjà amoureux de toi*.

À l'issue de ce tri manuel, 108 des 798 formes en *non-* recueillant plus de 500 000 occurrences ont été retenues. Elles ont été classées selon les caractéristiques morphologiques de leur base, i.e. selon que la base était (i) elle-même dérivée par suffixation sur base verbale (noms portant les suffixes *-ade, -age, -ance, -ée, -ment, -sion, -tion, -ure* ; adjectifs portant les suffixes *-eur, -able*, *-ible, -uble*, *-if, -oire*), (ii) dérivée sur base adjectivale (noms portant les suffixes *-ité*, *-té*, *-eté*, *-eur, -esse, -ise*, *-ice*, *-ion, -erie, -ie, -itude*, *-étude*), (iii) dérivée sur base nominale (noms portant les suffixes *-ade, -age, -ance, -aie, -aille, -at, -erie, -ier, -ure* ; adjectifs portant les suffixes *-aire, -al, -el, -esque, -eux, -ien, -ier, -ique, -u, -ais, -ain, -an, -oïde, -ois, -ote*), (iv) apparentée à un participe passé ou présent, (v) non construite en synchronie. Le tableau 6 donne un aperçu des résultats obtenus : les formes en *non-* les plus fréquentes possèdent soit une base apparentée à un participe, soit une base simple.

Morphologie	*Fréquences*	*Exemples*
apparentés à des participes	41	non-adapté, non-communiqué non-défini, non-renseigné
simplex	33	NON-FICTION, NON-RESPECT NON-DISPONIBLE, NON-TOXIQUE
déverbaux	19	NON-FUMEUR, NON-UTILISATION NON-NÉGLIGEABLE, NON-NÉGOCIABLE
dénominaux	6	non-commercial, non-contractuel, non-ferreux, non-professionnel
désadjectivaux	4	non-conformité, non-franchise, non-responsabilité, non-violence
autres	5	NON-HOMOLOGUE, NON-LOF

Tableau 6. Exemples de formes les plus fréquentes sur *Google*^{TM} (> 500 000 occ.) et leurs caractéristiques morphologiques.

Enfin, il m'a paru intéressant de comparer les fréquences des lexèmes générés avec la fréquence de leur base (c'est-à-dire sans le préfixe *non-*). Par exemple, comme l'indique le tableau 7, NON-CONTRE-INDICATION est plus fréquent que CONTRE-INDICATION puisque le rapport de la forme préfixée à la forme non préfixée est supérieur à 1. D'autres formes préfixées sont relativement fréquentes (quoique moins fréquentes que leur base), comme NON-CUMULABLE (rapport de 0.78), NON-FERREUX (rapport de 0.68).

[4] Livre des Origines Français, qui répertorie les origines des chiens de race français, ici utilisé avec valeur d'adjectif pour désigner un chien (non) répertorié au LOF.

Lexème préfixé (1)	*Lexème non préfixé (2)*	*Ratio (1)/(2)*
NON-CONTRE-INDICATION	CONTRE-INDICATION	1.75
NON-CUMULABLE	CUMULABLE	0.78
NON-LOF[1]	LOF	0.76
NON-FERREUX	FERREUX	0.68
NON-LUCRATIF	LUCRATIF	0.63
NON-FUMEUSE	FUMEUSE	0.61

Tableau 7. Ratio forme préfixée/forme non préfixée.

On note toutefois que la quasi-totalité (102 sur 108) des lexèmes préfixés ont une fréquence inferieure à 50% de celle de leur forme non préfixée (i.e. avec un ratio inférieur à 0,5).

3. Conclusions et perspectives de recherche

Dans cet article ont été retracées les différentes étapes d'un travail conjuguant méthodes de TAL et linguistique théorique et consistant à détecter, à extraire automatiquement et à analyser les lexèmes construits sur base nominale et adjectivale à l'aide du préfixe *non-* en français dans les pages en français indexées par *Google*™. Le principal apport de ce travail est d'illustrer le fait les données dictionnairiques ne reflètent qu'un état limité de la langue, puisque les lexèmes en *non-* effectivement employés par les locuteurs ne se limitent pas à ceux qui sont attestés dans le *TLFi*. Toutefois, les résultats obtenus suggèrent l'existence de limites à la capacité de la préfixation en *non-* à créer de nouveaux lexèmes, étant donné d'une part que 25% des formes générées n'étaient pas attestées dans le corpus examiné au moment de la requête, et que, d'autre part, les lexèmes générés apparaissent à des fréquences variables.

Je dispose à l'issue de ce travail d'un corpus extensif de lexèmes préfixés par *non-* sur base nominale et adjectivale, qui va me permettre de rendre compte des éventuelles contraintes morphologiques, sémantiques et phonologiques qui entrent en jeu dans leur formation, et de proposer une interprétation sémantique de ces dérivés. J'espère ainsi pouvoir approfondir la description et l'analyse de ce type de préfixation négative, qui n'a encore jamais fait l'objet d'un examen systématique en morphologie.

UMR STL 8163, Université Charles de Gaulle - Lille 3 Edwige DUGAS

Références bibliographiques

Berland, Sophie / Grabar, Natalia, 2001. « Construire un corpus Web pour l'acquisition terminologique », *Terminologie et intelligence artificielle* (TIA), 44-54.

Corbin, Danielle, 1987. *Morphologie dérivationnelle et structuration du lexique*, volume 1, Lille, Presses Universitaires de Lille.

Dal, Georgette / Namer, Fiametta, 2012. « Faut-il brûler les dictionnaires ? Ou comment les ressources numériques ont révolutionné les recherches en morphologie », *SHS Web of Conferences* 1, 1261-1276.

Dugas, Edwige, 2013. « [non(-)Adj] sequences in contemporary French: morphological negation, syntactic negation, or in between? », Communication faite lors du Symposium *Morphologie et ses interfaces*, Université de Lille 3, 12-13 septembre 2013.

Dugas, Edwige, 2012. *La négation en morphologie : le cas des formes nominales en* non- *en français*, Mémoire de Master 2 (non publié), Université de Lille 3.

Fradin, Bernard et al., 2008. « Remarques sur l'usage des corpus en morphologie », *Langages* 171, 34-59.

Gaatone, David, 1971. *Étude descriptive du système de la négation en français contemporain*, Genève, Librairie Droz.

Hamawand, Zeki, 2009. *The semantics of English negative prefixes*, Equinox Publications.

Hathout, Nabil et al., 2009. « La collecte et l'utilisation des données en morphologie », in : Fradin, Bernard / Kerleroux, Françoise / Plénat, Marc (ed.), *Aperçus de morphologie du français,* Presses universitaires de Vincennes, 267–287.

Jespersen, Otto, 1917. « Negation in English and other languages », in : *Selected Writings of Otto Jespersen*, Routledge, 3-151.

Muller, Claude, 1991. *La Négation en français : syntaxe, sémantique et éléments de comparaison avec les autres langues romanes*, Genève, Librairie Droz.

Romary, Laurent et al., 2004. « Standards going concrete: from LMF to Morphalou », in : Zock, Michael / Saint-Dizier, Patrick (ed), *COLING 2004 Enhancing and using electronic dictionaries*, Geneva, 22-28.

Zimmer, Karl, 1964, « Affixal negation in English and other Languages: an Investigation of Restricted Productivity », Supplement to *Word* 20 (2).

Une édition au format TEI de la première traduction française de *La Cité de Dieu* par Raoul de Presles (1371-1375)

1. Introduction

La présente contribution décrit notre mise en œuvre de la TEI (Text Encoding Initiative) pour l'édition d'un texte volumineux. Nous prétendons que cette méthode qui suppose de s'abstraire de la forme matérielle que prend l'édition peut être profitable à la qualité de l'édition produite. Pour justifier notre point de vue, nous présentons la forme papier de certaines éditions critiques (la nôtre en tant que résultat relève de cette description), la forme TEI plus abstraite que les éditeurs produisent, les outils permettant d'améliorer le codage et la transformation vers la forme papier.

1.1. Le projet

La traduction française de *La Cité de Dieu* de Raoul de Presles que nous éditons à l'ATILF grâce au financement européen du programme de recherche ERC - Starting Grant - intitulé « Histoire du lexique politique français »[1] se justifie par l'observation que 40%, autrement dit près de la moitié, de l'actuel vocabulaire politique français émergent aux XIV^e^ et XV^e^ siècles. Cette éclosion du vocabulaire politique s'explique par le fait que Charles V, dans la prolongation de l'œuvre entreprise par son père Jean le Bon, a fait traduire en français de nombreux textes de l'Antiquité gréco-latine portant sur l'art de gouverner et que, pour exprimer ou plutôt pour transposer la pensée politique en langue vernaculaire française, les traducteurs, gênés par la déficience du lexique français, ont dû recourir à un processus de création lexicale. On voit donc apparaître à cette époque de nombreux néologismes sémantiques pour la plupart empruntés aux auteurs latins et grecs. Sont nés à cette époque des mots tels que *anarchie*, *aristocratie*, *démagogie*, *oligarchie* empruntés au grec et *magistrat*, *politique, suffrage* empruntés au latin.

1.2. Le texte

Parmi les traductions commanditées par Charles V, celle du *De Civitate Dei contra paganos*, un ouvrage écrit au V^e^ siècle de notre ère par saint Augustin et longtemps considéré par les historiens comme la première oeuvre politique des Temps

[1] La présentation générale du programme est visible sur le site de l'Atilf à l'adresse suivante : ‹http://www.atilf.fr/›.

modernes, occupe une place centrale pour notre connaissance du lexique politique français actuel. On y rencontre, comme le montrent nos premiers travaux[2], une mine de lexèmes politiques relevant de la néologie formelle, parmi lesquels on peut citer *dictateur, monarche, monarchie, proletaire, proscription, tresor publique*, etc. Or la dernière édition du texte de Raoul de Presles date de 1531. Comme il n'existe pas à ce jour d'édition moderne de référence de cette traduction, nous avons été amenés à établir nous-mêmes cette édition pour avoir ensuite accès à la langue du texte et pouvoir offrir ensuite cet accès à l'ensemble de la communauté scientifique.

1.3. Le manuscrit de base

Le manuscrit que nous éditons est celui que Raoul de Presles a offert au monarque vers 1376 : le manuscrit français BnF, fr. 22912 et 22913. Le choix de ce manuscrit, parmi les cinquante-huit manuscrits qui nous sont parvenus de ce texte et qui ont été copiés entre 1376 et la fin du XV^{e} siècle, a été guidé par le fait qu'il s'agit, nous l'avons vu, d'un manuscrit royal composé à une date très proche de la date d'écriture (entre 1371 et 1375). Il est de plus complet, soigné et très proche des autres manuscrits (il présente très peu de variations lexicales par rapport aux autres manuscrits), de surcroît, il comporte en son tout début une liste d'autorités unique.

Le texte d'Augustin est déjà considérable en taille (22 livres), celui de Raoul de Presles l'est encore plus. Le texte latin occupe entre 250 et 300 folios dans la plupart des manuscrits consultés tandis que le la traduction française, dans la mesure où elle est accompagnée de riches gloses explicatives, couvre pas moins de 894 folios. Assez fréquemment, et probablement en raison des neuf siècles qui séparent l'auteur du traducteur, Raoul de Presles, qui cherche à éclairer le lecteur du XIV^{e} siècle, double ou triple, voire plus, le texte qu'il traduit par ses gloses-commentaires. Pour exemple, le texte de la traduction du chapitre 22 du livre II occupe un folio recto-verso complet, alors que les gloses remplissent trois folios recto-verso complets et deux colonnes de plus. Il va donc de soi que notre projet d'édition d'un texte aussi long est un projet à long terme et que ce temps long justifie l'investissement dans des outils d'aide à l'édition même spécifiques à notre projet[3].

1.4. Le travail d'édition

Notre modèle d'édition, cherche à être au plus près des règles éditoriales de base mises au point par l'École nationale des chartes dans son ouvrage tripartite intitulé *Conseils pour l'édition des textes médiévaux*[4]. Parmi les choix d'édition possibles de ce

[2] Cf. Cerrito / Stumpf, 2013 et les chapitres 9 à 11 dans *Sciences et savoirs sous Charles V*, sous la direction d'Olivier Bertrand, Paris, Champion, à paraître.

[3] Nous ne prétendons aucunement que nos outils (en particulier ceux qui calculent la forme papier) soient réutilisables directement pour un autre projet.

[4] *Conseils pour l'édition des textes médiévaux*, fasc. I- III, *Textes littéraires*, publ. par l'École nationale des chartes, Paris, Comité des travaux historiques et scientifiques, École nationale des chartes, 2001-2002.

texte, nous avons décidé de donner à lire un texte qui observe le plus grand respect du manuscrit de base, sans reproduire les abréviations, la ponctuation, les majuscules de la copie, mais en intervenant en cas de lacune ou de texte fautif. Bien entendu, toute correction au manuscrit est justifiée sur la base des manuscrits de contrôle. Il a été décidé aussi que la *varia lectio* devra se limiter aux variantes lexicales qui présentent une alternative sémantique intéressante ou qui attirent l'attention sur une difficulté du texte. Pour qu'elle soit complète, nous voulions que cette édition s'accompagne à la fois de commentaires relatifs à l'identification des citations et des sources et, comme il se doit pour l'édition d'un texte littéraire, d'un *index nominum*[5].

Pour établir au mieux l'édition de ce texte imposant, nous avons opté pour un encodage TEI. Néanmoins, le choix de publier l'édition sur support papier a été impulsé par l'idée de léguer à la communauté scientifique un document pérenne. Car comme le souligne très justement M. Burghart, dans sa communication intitulée « Digital Editions as the Myth of Sisyphus » (Burghart 2011), beaucoup d'éditions électroniques ne sont pas maintenues sur le long terme[6].

Dans ce qui suit, nous aborderons la question des contraintes que nous nous imposons pour l'édition sur support papier et celle de sa réalisation par l'intermédiaire du format d'encodage TEI. Puis, avant de montrer comment ce format d'encodage permet un travail d'édition de qualité si on met aussi en œuvre le précieux outil de lemmatisation LGeRM (Souvay / Pierrel 2009 et Souvay / Bazin 2013, 56), nous présenterons la manière dont se fait la transformation de la version TEI à la forme papier.

2. La forme papier d'une édition critique

Une édition critique comme la nôtre se présente formellement comme un texte sur lequel portent des notes. Le texte « principal », immédiatement accessible est le texte édité, c'est-à-dire le texte du manuscrit de base amendé lorsqu'il est jugé fautif à partir de la collation de plusieurs manuscrits préalablement sélectionnés. Chaque correction opérée au manuscrit de base donne lieu à une note justificative. Par ailleurs, d'autres notes servent à associer au texte édité la *varia lectio*. On a donc essentiellement le texte édité, très proéminent, et accessoirement des références, via des notes, à des manuscrits de contrôle et de variantes.

Les notes qui justifient les choix éditoriaux en cas de correction de faute de copiste, doivent identifier le segment de texte édité, identifier son équivalent considéré comme fautif dans le manuscrit qui sert de base à l'édition, et enfin mentionner les manuscrits sur lesquels s'appuie la correction. Pour exemple, là où la leçon *Comment* du

---

[5] Un glossaire est prévu. Nous n'avons qu'ébauché les solutions techniques qui devront conduire à sa production pour le second volume. À ce moment du projet nous aurons en effet couvert les 10 premiers livres.

[6] Une revue des projets TEI (cf. Burghart / Rehbein 2012) montre que 97 % des projets encodés en TEI ne visent que la publication sur le WEB.

texte édité est mise pour la leçon fautive *romment* du manuscrit (fol. 48 v°)[7], une note sera donnée pour consigner la correction ; elle doit comprendre le repérage du segment concerné (numéro de la ligne numérotée du texte édité où est attesté ce segment), l'identification du segment (la leçon du manuscrit) et les sigles des manuscrits qui ont permis la correction à la leçon jugée fautive.

Les notes introduisant les variantes de lecture entre le texte établi par l'éditeur scientifique et les témoins du texte doivent elles aussi identifier le segment du texte édité d'une part, et sa contrepartie dans d'autres manuscrits, d'autre part. Pour exemple, là où la leçon du manuscrit reprise dans le texte édité porte la forme ancienne *herites* «hérétiques», les copies plus tardives portent la forme moderne *heretiques* correspondante, une note sera donnée pour consigner la variante ; selon le même principe que pour les notes d'édition, elle doit comprendre le repérage du segment concerné, l'identification du segment suivi de la ou des leçons divergentes du ou des manuscrits accompagnés de leurs sigles, en guise de référence.

L'objet central de l'édition est le texte édité, et on met en relation des segments de cet objet central avec des segments soit du manuscrit de base lorsqu'il est différent de celui de l'édition, soit des manuscrits qui fournissent des variantes au texte. D'un point de vue abstrait, on a donc essentiellement des mises en correspondances de segments de textes.

3. L'encodage en TEI

La TEI est une organisation internationale qui produit des recommandations et des schémas d'encodage (en XML) de toutes sortes de textes (pièces de théâtre, dictionnaires, etc.). Les recommandations sont organisées en chapitres plus ou moins étroitement liés aux grandes formes de textes. L'un des chapitres est spécifiquement dédié aux éditions critiques. La TEI évolue sous le contrôle de ses utilisateurs, par conséquent, l'encodage en TEI des éditions critiques a été conçu par des personnes qui réalisent des éditions critiques.

L'encodage d'un texte en TEI se veut, autant que possible, indépendant de la forme matérielle que prendra l'édition ; nous entendons par là que le même fichier TEI servirait tout autant à produire une édition papier qu'une édition électronique. Parler d'ailleurs d'une édition électronique est presque un abus de langage si on se réfère par exemple à la « Queste del saint Graal »[8] dont « le texte peut s'afficher sous un triple format – version courante, version diplomatique, version facsimilaire » et ce de façon dynamique.

---

[7] Il s'agit non pas du "roman" de Nicholas Trevet mais du "commentaire" des tragédies de Sénèque qu'il a fait, comme on peut le voir dans le passage concerné extrait du livre II au chapitre 7 : *« Et se tu veulx veoir de ceste matiere plainnement, voy Seneque en sa seconde tragedie, avecques le Comment de Travet. »*.

[8] Voir ‹http://portal.textometrie.org/txm/›.

Pour arriver à un tel résultat, il va de soi que l'encodage ne peut qu'être abstrait par rapport à une mise en forme particulière. Le cœur de l'encodage repose donc sur la notion de mise en parallèle de segments dont nous avons parlé précédemment[9].

Pour en donner un aperçu, nous allons brièvement parcourir les façons d'encoder :

- Les caractéristiques du texte transcrit (abréviations, régularisations, etc.)
- Les opérations d'édition (corrections par substitution de leçons et par ajout au texte)
- Les variantes au texte édité

3.1. Les caractéristiques du texte transcrit

Bien que certaines caractéristiques de la transcription telles que les abréviations n'apparaissent pas sur la version papier que nous visons, celles-ci ont été encodées. Pour le confort de la lecture, elles sont résolues dans le texte édité mais, pour pouvoir lever des doutes sur leurs résolutions, nous avons jugé utile d'en garder trace dans le texte encodé. Ainsi, selon qu'il s'agit d'une abréviation par signes spéciaux (barre de nasalité, lettre suscrite, *p* barré, *p* barré courbe, etc.) ou par suspension (*mons.* pour *monseigneur*), on encode de deux manières distinctes. Dans le premier cas on encode avec la balise <ex> et dans le second avec<choice><abbr><expan>. On encode par exemple :

<ex>con</ex>sacréz » pour noter le mot « consacréz » (lisible en ignorant les balises)

pour signifier que dans le manuscrit une abréviation est mise respectivement pour le segment « con » (en l'occurrence cette abréviation est un neuf tironien, mais l'encodage ne le mentionne pas) et on encode :

<choice><abbr>mons</abbr><expan>monseigneur</expan></choice> pour « monseigneur ».

On signifie ainsi que « mons » est l'abréviation de la forme étendue *monseigneur.*

On encode également les pieds de mouche, les rubrications, les ajouts postérieurs, les lettrines, les notes marginales, etc. (ces éléments contrairement aux abréviations sont visibles sur papier sous formes respectives de rupture de paragraphe, titre, texte normal, lettrine et note marginale).

En ce qui concerne la segmentation des mots liée à la plasticité de l'écriture au Moyen Âge, nous encodons à la fois la forme du manuscrit et la forme éditée. Ainsi, nous encodons :

<choice><sic>tresbon</sic><reg>tres bon</reg></choice>

et

<choice><sic>du quel</sic><reg>duquel</reg></choice>

---

[9] Ceci est d'autant plus vrai que la TEI est notre format d'entrée : les éditeurs produisent directement de la TEI. On fait donc usage de la « Parallel Segmentation Method » (‹http : //www.tei-c.org/›).

même si, dans la forme papier que nous produisons, les régularisations *tres bon* et *duquel* sont silencieuses.

3.2. Les opérations d'édition

Les éditeurs scientifiques peuvent s'éloigner de la transcription du manuscrit de base de deux façons : soit ils substituent à une leçon manifestement fautive du manuscrit de base une autre leçon, soit ils ajoutent du matériau nécessaire pour le sens du texte mais oublié dans la copie de base.

Ainsi la substitution d'une leçon à une autre s'encode par : <choice><sic></sic><corr></corr></choice>. Par exemple :

```
<choice><sic>Comment</sic><corr source=»#C1 #P17 #P11 #P31»>Romment</corr>
</choice>
```

où l'éditeur a décidé (<choice>) que la leçon fautive (<sic>) *Romment* doit être corrigée (<corr>) en *Comment*, d'après les manuscrits (attribut source) $C^1P^{17}P^{11}$ et P^{31}. (Pour la version papier, voir *infra* § 4, ex. 1).

L'ajout de matériau correspond au balisage suivant :

```
<w>phil<supplied reason=»manque» source=»#C1 #P17 #P11 #P31»>os</supplied>ophes</w>
```

qui indique que nous sommes en présence d'un mot (<w>), à savoir *philosophes*, dont le segment *os* a été ajouté (<supplied>) parce qu'il manque dans le manuscrit (reason=»manque»), sur la foi (source) des manuscrits $C^1P^{17}P^{11}$ et P^{31}. (Pour la version papier, voir *infra* § 4, ex. 2).

3.3. Les variantes

En ce qui concerne la *varia lectio*, on notera que chaque lieu variant est codé avec la balise <app>, suivie de la leçon retenue par l'éditeur scientifique balisée avec <lem> elle-même suivie par autant de <rdg> qu'il y a de leçons différentes et complétée par l'attribut 'wit' pour indiquer les identifiants des témoins portant cette leçon. Ainsi, on encode :

```
<app><lem>herites</lem><rdg wit=»#M1 #éd1486»>heretiques</rdg></app>
```

pour signaler qu'on veut mettre dans l'apparat critique (<app>) comme variante de *herites* (<lem>) la leçon *heretiques* (<rdg>) tirée d'un manuscrit et d'une édition auxquels on réfère par des sigles. (Pour la version papier, voir *infra* § 4, ex. 3).

3.4. Conclusion sur l'encodage

Une des caractéristiques de l'encodage en TEI, si on le compare à la forme de l'édition, est de ne pas mettre en avant la forme que l'on veut éditer : il suffit d'oublier le contenu des 'supplied' et de prendre en compte la partie 'sic' des corrections pour lire la transcription du manuscrit au lieu de celle que donne à lire l'éditeur et qui

peut s'éloigner du texte du manuscrit. Inversement, on peut lire (en faisant les choix opposés) l'édition du texte établi en ignorant les parties de la transcription estimées comme fautives. Formellement, l'encodage est donc neutre vis-à-vis du texte édité et de la transcription du manuscrit : il se contente de les mettre en parallèle.

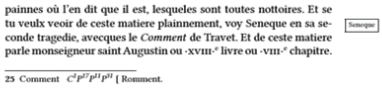
4. Le passage de la version TEI à la forme papier de l'édition

Nous avons développé des outils pour passer de la version TEI à la version papier. Nous nous contentons d'illustrer les cas de figure précédemment évoqués.

Voici, comme aperçu, un court extrait de la version TEI que nous avons retenu pour visualiser la correction de *Romment* en *Comment* :

```
Et se tu veulx veoir de ceste matiere plainnement, voy <rs type="anthroponyme" key="Sénèque"
    >Seneque</rs> en <note type="note_en_marge_manuscript">Seneque</note>sa seconde tragedie,
 avecques le <choice><sic>Romment</sic>
    <corr source="#C1 #P17 #P11 #P31">
        <rs type="source" key="Commentaire sur les Tragédies de Sénèque">Comment</rs>
    </corr></choice> de <rs type="anthroponyme" key=" Nicolas Trevet ">Travet</rs> .Et de ceste
 matiere <w><ex>par</ex>le</w> monseigneur saint <rs type="anthroponyme" key="Augustin, saint"
    >Augustin</rs> ou <num type="ordinal">XVIII</num> livre ou <num type="ordinal">VIII</num>
 chapitre.
```

Et son résultat dans la version papier[10]



À présent, concentrons-nous sur la manière dont sont signalés dans l'édition papier les deux types de notes de bas de page : les notes d'édition (premier étage de note) et les notes de variantes (deuxième étage de note).

Notes d'édition :

(1) « 25 Comment $C^1P^{17}P^{11}P^{31}$ [Romment. »

L'identification du segment correspond au numéro de la ligne numérotée du texte édité où est attesté le segment *Comment* visé et cité ; les sigles des manuscrits[11] qui suivent sont ceux qui ont permis la correction à la leçon jugée fautive du manuscrit de base que l'on donne juste après. Les sigles auxquels on réfère trouvent leur développement dans l'entête de la TEI (TEI Header).

(2) « 20 phil[os]ophes $C^1P^{17}P^{11}$ P^{31} [os *mq.* »

Comme précédemment, la note commence par l'identification du segment visé, reproduit sous la forme « phil[os]ophes » où le segment *os* est rétabli entre crochets droits,

[10] On notera dans l'encodage les balises <rs> ; ces balises sont destinées à la production des index. Dans l'extrait cité ici, apparaissent (pour l'index) des références à Sénèque, au *Commentaire sur les tragédies de Sénèque*, à Nicolas Trevet et à saint Augustin.

[11] Respectivement : Chantilly, Bibl. du Château, 322 ; Paris, BnF, fr.170-171 ; Paris, BnF, fr. 23-24 et Paris, BnF, fr. 20105.

d'après les quatre manuscrits qui donnent la forme correcte, parce que *os* manque '*mq*' dans le manuscrit de base.

Notes de variantes :

(3) « 7 herites / heretiques *M*[1] *éd.1486*. »

Comme toujours, la note commence par l'identification du segment *herites* visé et cité, puis suivi de la variante accompagnée des sigles du manuscrit et de l'édition qui nous offrent ces variantes[12]. Cette note se trouve en bas de page, dans le deuxième étage de notes.

5. Encodage en TEI et amélioration de la qualité de l'édition

Nous prétendons que l'encodage en TEI améliore la qualité de l'édition papier produite. Selon nous, cette amélioration de l'édition est relative à :

- La forme (la forme papier est plus systématique que si elle avait été directement produite à l'aide d'un traitement de texte)
- Au fond (indépendamment même de la forme papier, un certain nombre d'erreurs ne sont pas possibles en passant par un encodage TEI)
- Au processus d'édition lui-même qui, en ce qu'il est rendu plus confortable limite les possibilités d'erreurs.

5.1. Amélioration de la forme

Comme nous l'avons mentionné à propos des corrections, l'édition sur papier doit mettre en parallèle le segment édité (et corrigé) et le même segment dans le manuscrit. Pour ce faire, la note de bas de page justifiant la correction doit reprendre dans son intégralité le segment édité ; dans la mesure où la note est entièrement fabriquée sur la base de l'encodage en TEI (l'éditeur scientifique n'a aucune prise là-dessus), sa fabrication est systématique. À l'opposé, si les éditeurs utilisaient un traitement de texte classique, il serait à leur charge de produire des notes de la forme :

segment du texte [segment originel

et

segment du texte / segment variant

Notons que pour ce faire, le risque de se tromper dans la recopie du « segment du texte » n'est pas négligeable. En outre, toujours au chapitre de l'amélioration de la forme, le fait qu'une source médiévale ou antique soit indexée en tant que source fait qu'elle apparaît en italique dans le texte ; de même, l'encodage des pieds de mouche génère le retour à la ligne, celle des rubrications l'affichage du titre, celles des notes marginales les vignettes marginales, etc.

[12] C'est-à-dire Mâcon, Bibl. mun., 1 et Abbeville, Jehan du Pré et Pierre Gerard, 1486.

Du point de vue de la forme, nous souhaitons mettre l'accent sur la fabrication des index, particulièrement délicate pour un texte aussi imposant et aux sources aussi nombreuses que variées, puisque les textes s'étalent « depuis la latinité classique jusqu'à des auteurs presque contemporains de Raoul de Presles » (Bertrand, *Introduction*, p. 87). On notera tout d'abord que pour ces raisons de taille et de diversité, l'exhaustivité que nous visions nous a amenés à ne pas fondre *l'index nominum* en un index unique, mais à le scinder en trois index distincts : l'index des noms de personnes (personnages historiques ou mythiques), l'index des œuvres citées et l'index des toponymes et des noms de peuple. Pour l'édition papier, le choix a été fait de classer les entrées par ordre alphabétique à la forme la plus fréquente du texte et de les faire suivre des formes secondaires également classées par ordre alphabétique, puis d'indiquer les références aux pages où elles sont attestées et enfin de donner la forme de référence accompagnée d'une notice succincte qui sert à identifier clairement la personne, le lieu ou la source. Pour ce texte médiéval qui présente une grande variété de graphies anciennes pour une même entité, se pose, outre le problème de l'exhaustivité, celui du repérage de la forme la plus fréquente.

L'éditeur se contente d'indexer les formes du texte par des clefs de regroupement (cf. exemple ci-dessus pour Sénèque, saint Augustin et Nicolas Trevet). Le calcul aboutissant à la forme papier de l'index est ensuite entièrement automatique et la clef d'indexation n'apparaît pas dans l'index produit. Dans l'exemple donné, les entrées à l'index sont : Seneque, Augustin et Travet parce qu'il se trouve que la clef choisie par les encodeurs est la forme la plus fréquente dans le texte ; pour la clef Ciceron au contraire, l'entrée est à Tulle. Notons enfin, que des renvois sont prévus (dans les fichiers d'index), mais que ces renvois n'apparaissent que si la forme de renvoi est effectivement attestée dans le texte. Le bénéfice du travail d'encodage sur une forme abstraite est ici évident : les éditeurs n'ont pas à se préoccuper de ce que sera l'entrée d'index ; tout ce qu'ils ont à faire est de choisir une clef identifiante de la personne, de l'œuvre ou du lieu pour le nom propre qu'ils ont sous les yeux[13].

5.2. Amélioration du fond

Par amélioration du fond, nous entendons amélioration de l'encodage en TEI qui évidemment se traduit par une amélioration de l'édition papier. Le fichier TEI est produit sous la contrainte d'un schéma XML. Concrètement, cela signifie que des vérifications syntaxiques sont opérées sur le fichier TEI. Ainsi, les sigles des manuscrits auxquels les éditeurs réfèrent sont vérifiés et de façon générale, beaucoup d'erreurs sont filtrées à l'édition même. Il est hors de question de les mentionner toutes ; elles sont aussi diverses que l'obligation d'avoir des variantes et un lemme dans un lieu variant ou d'avoir une rubrique en tête d'un chapitre.

[13] De plus, les clefs ont été régulièrement recensées et ajoutées au schéma d'encodage. En pratique, les éditeurs choisissent donc leurs clefs dans des listes déroulantes et ne créent de nouvelles clefs qu'en cas d'absence dans la liste considérée (noms de personnes ou oeuvres citées ou toponymes / noms de peuples).

5.3. Amélioration du processus d'édition

Nous avons vu précédemment que la forme papier de l'édition met en avant le texte édité alors que la forme TEI est neutre et se contente de mettre en parallèle la transcription du manuscrit de base et les corrections apportées à cette transcription. Une conséquence en est que si un éditeur veut produire une édition critique directement à partir d'un traitement de texte, il devra commencer par transcrire (comment faire autrement ?) puis faire passer (là où il corrige) une partie de son texte dans des notes de bas de page. En revanche, l'éditeur d'une édition TEI commence lui aussi par transcrire le texte du manuscrit de base, mais il se contente dans un second temps de ne faire qu'annoter sa transcription par des corrections portant sur cette transcription : le caractère neutre de l'encodage TEI permet cette mise en avant de la transcription durant la phase d'encodage (la symétrie dans le fichier TEI entre transcription et édition permet cette proéminence de la transcription pour le travail tout autant que la proéminence du texte édité pour la mise sur papier). De fait, l'encodage est ainsi beaucoup plus fluide et respecte le processus qui enchaîne transcription → édition → encodage des variantes → ajout de commentaires à valeur historique[14]. Notons d'ailleurs que les outils de production de la version imprimable sont utilisés tout au long de ce processus pour permettre une relecture pour correction plus aisée.

6. Outils mis en œuvre pour améliorer l'édition

Au-delà des outils produisant la version pdf, le fait que la TEI soit un format entièrement ouvert (i.e. parfaitement décrit) permet d'adapter ou de réaliser des outils de contrôle de l'édition. Le plus important des outils informatiques utilisés dans notre cas a été LGeRM qui complète avantageusement les ouvrages de lexicographie et permet de faire des vérifications en amont et en aval. Ce lemmatiseur, grâce au formulaire de recherche des formes dans la base des textes médiévaux disponibles à l'ATILF qui lui est intégré, pointe tous les mots ou formes curieuses et toutes les formes qui lui sont inconnues, ce qui permet à l'éditeur de retourner au texte et de corriger les éventuelles coquilles et les erreurs de transcription ou, le cas échéant, de faire un choix éditorial réfléchi, c'est-à-dire de maintenir ou, au contraire, de corriger dans l'édition cette leçon du manuscrit.

7. Conclusion

Il peut sembler curieux, voire hasardeux, lorsqu'on vise essentiellement la version papier d'une édition critique de faire un détour par une forme d'encodage abstraite. Au premier abord, on pourrait penser que réaliser l'édition via un outil de type traitement de texte (plus ou moins sophistiqué selon la forme visée) constitue la meil-

[14] Ces commentaires consignés dans le troisième étage de notes, sont réservés aux sources d'Augustin et de Raoul de Presles et comportent des notes explicatives concernant pour l'essentiel l'histoire romaine.

leure solution. Notre expérience de l'édition d'un gros texte est que ce détour par un encodage en TEI se justifie non seulement par la possibilité d'utiliser la version TEI à d'autres fins (une édition électronique par exemple) mais également par le fait que cette forme :

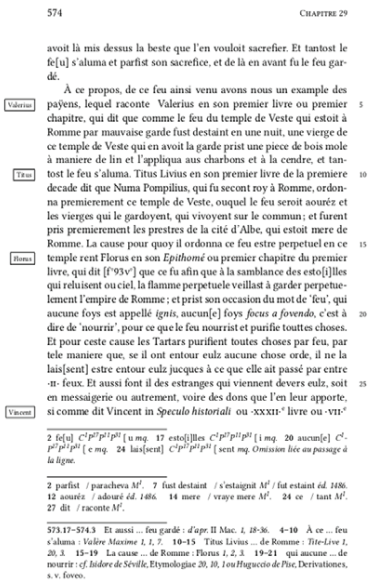
- rend mieux compte que d'autres du processus allant de la transcription à l'édition
- permet des contrôles automatiques améliorant l'édition
- permet un rendu sur papier de meilleure qualité grâce au caractère systématique de production des notes de bas de page

Nous ne saurions nous prononcer quant au bénéfice d'un tel encodage pour des textes beaucoup plus courts (le bénéfice lié aux utilisations multiples du texte demeurerait), l'écriture d'outils de conversion à une forme imprimable n'est peut-être pas rentable dans ce cas ; en ce qui concerne *La Cité de Dieu* en revanche, la taille du texte nous fait supposer qu'*a contrario* l'usage d'un outil travaillant directement sur la forme (un traitement de texte pour parler clair) était infiniment plus risqué.

ATILF-CNRS, Université de Lorraine

Bertrand GAIFFE
Béatrice STUMPF

8. Annexe (La Cité de Dieu, dir. par O. Bertrand, p.574)



574 CHAPITRE 29

avoit là mis dessus la beste que l'en vouloit sacrefier. Et tantost le fe[u] s'aluma et parfist son sacrefice, et de là en avant fu le feu gardé.

Valerius À ce propos, de ce feu ainsi venu avons nous un example des païens, lequel raconte Valerius en son premier livre ou premier chapitre, qui dit que comme le feu du temple de Veste qui estoit à Romme par mauvaise garde fust destaint en une nuit, une vierge de ce temple de Veste qui en avoit la garde prist une piece de bois mole à maniere de lin et l'appliqua aus charbons et à la cendre, et tantost le feu s'aluma.

Titus Titus Livius en son premier livre de la premiere decade dit que Numa Pompilius, qui fu secont roy à Romme, ordonna premierement ce temple de Veste, ouquel le feu seroit aouréz et les vierges qui le gardoyent, qui vivoyent sur le commun ; et furent pris premierement les prestres de la cité d'Albe, qui estoit mere de Romme. La cause pour quoy il ordonna ce feu estre perpetuel en ce

Florus temple rent Florus en son *Epithomé* ou premier chapitre du premier livre, qui dit [f°93v°] que ce fu afin que à la samblance des esto[i]lles qui reluisent ou ciel, la flamme perpetuele veillast à garder perpetuelement l'empire de Romme ; et prist son occasion du mot de 'feu', qui aucune foys est appellé *ignis*, aucun[e] foys *focus a fovendo*, c'est à dire de 'nourrir', pour ce que le feu nourrist et purifie touttes choses. Et pour ceste cause les Tartars purifient toutes choses par feu, par tele maniere que, se il ont entour eulz aucune chose orde, il ne la lais[sent] estre entour eulz jucques à ce que elle ait passé par entre ·II· feux. Et aussi font il des estranges qui viennent devers eulz, soit en messaigerie ou autrement, voire des dons que l'en leur apporte,

Vincent si comme dit Vincent in *Speculo historiali* ou ·XXXII·^{e} livre ou ·VII·^{e}

2 fe[u] $C^{1}p^{1}l^{1}p^{1}l^{1}p^{1}l$ [u *mq.* 17 esto[i]lles $C^{1}p^{1}l^{1}p^{1}l^{1}p^{1}l$ [i *mq.* 20 aucun[e] C^{1}-$p^{1}l^{1}p^{1}l^{1}p^{1}l$ [e *mq.* 24 lais[sent] $C^{1}p^{1}l^{1}p^{1}l^{1}p^{1}l$ [sent *mq. Omission liée au passage à la ligne.*

2 parfist / paracheva M^{1}. 7 fust destaint / s'estaignit M^{1} / fut estaint *éd. 1486.* 12 aouréz / adouré *éd. 1486.* 14 mere / vraye mere M^{1}. 24 ce / tant M^{1}. 27 dit / raconte M^{1}.

573.17–574.3 Et aussi ... feu gardé : *d'apr.* II Mac. *1, 18-36.* 4–10 À ce ... feu s'aluma : *Valère Maxime 1, 1, 7.* 10–15 Titus Livius ... de Romme : *Tite-Live 1, 20, 3.* 15–19 La cause ... de Romme : Florus *1, 2, 3.* 19–21 qui aucune ... de nourrir : *cf. Isidore de Séville*, Etymologiae *20, 10, 1 ou Huguccio de Pise*, Derivationes, s. v. foveo.

Références bibliographiques

Bertrand, Olivier, sous presse. La Cité de Dieu *de saint Augustin traduite par Raoul de Presles (1371-1375), édition du ms BnF, fr. 22912*, Paris, Champion, « Linguistique : traduction et terminologie » 1, volume 1, tome 1.

Burghart, Marjorie, 2011. « Digital Editions as the Myth of Sisyphus », TEI Members Meeting and Conference, Würtzburg, 2011, ‹www.zde.uni-wuerzburg.de/tei_mm_2011/abstracts/abstracts_micropapers/#c249141›.

Burghart, Marjorie / Rehbein, Malte, 2012. « The Present and Future of the TEI Community for Manuscript Encoding », Journal of the Text Encoding Initiative, 2012. ‹http://jtei.revues.org/372›.

Cerrito, Stefania / Stumpf, Béatrice, 2013. « Le lexique politique dans la traduction de la Cité de Dieu de saint Augustin par Raoul de Presles (1371-1375) », in : Ligas, Pierluigi / Frassi, Paolo (ed), *Lexiques Identités Cultures*, Vérone, QuiEdit, 197-220.

Gaiffe, Bertand / Stumpf, Béatrice, 2011. « A large scale critical edition : first translation od St Augustine's City of God by Raoul de Presle », TEI Members Meeting and Conference, Würtzburg, 2011, ‹www.zde.uni-wuerzburg.de/en/tei_mm_2011/abstracts/abstracts_papers/›.

Rochebouet, Anne, 2009. « D'une pel toute entiere sans nulle cousture. La cinquième mise en prose du Roman de Troie, édition critique et commentaire », thèse de doctorat de l'université Paris 4, sous la dir. de Gilles Roussineau, 2009.

Souvay, Gilles / Pierrel, Jean-Marie, 2009. « LGeRM, Lemmatisation des mots en moyen français », *TAL* (Traitement automatique des langues), 50, 149-172. ‹http://halshs.archives-ouvertes.fr/halsbbhs-00396452›.

Souvay, Gilles / Bazin, Sylvie, 2013. « Le Dictionnaire du Moyen Français : la version DMF 2010 », in : Casanova Herrero, Emili / Calvo Rigual, Cesáreo (ed.), *Actes du XXVI^e Congrès International de Linguistique et de Philologie Romane* (Valencia, 6-11 septembre 2010), Berlin, W. de Gruyter, 55-65.

Les ‘petits recueils’ de fabliaux : présence, composition, perspectives

Le projet de recherche *Lire en contexte à l’époque prémoderne : enquête sur les manuscrits de fabliaux* (2010-2014) vise à étudier de façon systématique, par le biais du croisement suivi des démarches – analyse en détail, vue panoramique –, l’ensemble des témoins contenant ou ayant contenu au moins un fabliau, suivant la définition du *corpus* établie par le NRCF[1]. Dans ce cadre, j’ai été amené à m’occuper directement de témoins, complets ou fragmentaires, contenant à présent un ou quelques fabliaux. Il était convenu de laisser à d’autres le soin d’étudier la plupart des manuscrits anglo-normands, qui semblent obéir à des logiques de composition en partie distinctes, ainsi que certains des témoins davantage marqués par la figure d’un auteur, aussi n’est restée dans mon carquois qu’une douzaine et demi d’unités[2], représentant tout de même presque la moitié du *corpus*. Il est apparu rapidement que la plupart de ces manuscrits, tels qu’ils avaient été appréhendés ou exploités jusqu’ici, transmettaient une image partielle, voire biaisée de leur composition, ainsi que de leur structure d’origine. Certes, il ne s’agit nullement des grands recueils du genre, qui demeurent sous les feux des projecteurs depuis un demi-siècle au moins[3]. Il est plutôt question de recueils modestes, pour ce qui est de leur contribution au genre, et au profil souvent banal, dont l’intérêt n’est toutefois pas mince : d’abord, par rapport à d’autres constellations de textes et aux spécificités de leurs traditions ; ensuite, parce qu’ils nous renseignent efficacement sur la circulation des fabliaux en dehors des grands recueils narratifs, en contexte d’hétérogénéité textuelle, sur un arc chronologique étendu et au sein de typologies livresques variées. Or, dans de nombreux cas, cette arrière-garde a été négligée à tel point que la physionomie actuelle et primitive elle-même des diffé-

---

1 Les prémisses et objectifs du projet ont été esquissés par Trachsler (2010), dans le droit fil du vœu exprimé, avec la clairvoyance habituelle, par Rychner (1960, vol. 1, 140-141).

2 Audenarde, KerkarchiefWalburga, 32 (sigle : *n*), Chantilly, Bibl. du Château, 475 (*T*), ancien Chartres, Bibl. mun., 620 (*f*), Le Mans, Méd. Louis Aragon, 173 (*m*), Lyon, Bibl. mun., 5495 (*V*^{bis}), Nottingham, University Library, WLC/LM/6 (*G*), Paris, Bibl. de l’Arsenal, 3114 (*e*) et 3527 (*p*), Paris, Bibl. nationale de France, fr. 375 (*a*), fr. 1635 (*L*), fr. 2173 (*K*), naf. 934 (*q*) + Troyes, Méd. du Grand Troyes, 1511 (*k*), naf. 5094 + Clermont-Ferrand, Arch. dép. du Puy-de-Dôme, 1 F 2 (*i*), Rome, Bibl. Casanatense, 1598 (*N*), Turin, Bibl. Nazionale Universitaria, L.II.14 (*r*) et ancien L.V.32 (*U*).

3 Je fais, bien entendu, référence aux mss Paris, BnF, fr. 837 (*A*), fr. 19152 (*D*), fr. 1593 (*E*), Berne, Bibl. de la Burgeoisie, 354 (*B*), Berlin, Staatsbibl. – Preußischer Kulturbesitz, Hamilton 257 (*C*) et aux travaux marquants ou récents, entre autres, de Rychner (1960), van den Boogaard (1984 : d’ici notre étiquette commode de ‘petits recueils’) et Busby (2002, vol. 1, 64-73, 437-463).

rents témoins reste dans le flou. On imagine aisément le niveau de dégâts qu'une telle méconnaissance peut engendrer dans le cadre de la recherche en cours.

Plusieurs cas de figure se présentent. Je me bornerai à en examiner cinq[4].

Le premier cas, le plus éclatant, est celui – heureusement exceptionnel – de la disparition du témoin de la littérature critique et des éditions de référence. C'est ce qui est arrivé, à la fin du XX^e^ siècle, à un témoignage ancien et précieux, la version laconique du fabliau du *Vilain Asnier* (18 vers) contenue dans le ms. 173 de la Méd. Louis Aragon du Mans (f. 110r)[5]. Un simple incident technique au sein de l'entreprise du NRCF en a provoqué l'oubli presque complet dans la production critique[6]. Jusqu'au vol. 6 (1991), le NRCF enregistrait le manuscrit en tant que témoin du *Vilain Asnier* (sigle : *m*), mais sous une cote erronée (5495), doublant celle du fragment de Lyon (Bibl. mun., 5495 [V^{bis}])[7]. De toute évidence, dans une liste préparatoire, les témoins étaient ordonnés selon le lieu de conservation, et le témoin du Mans (Mans [Le]) suivait immédiatement celui de Lyon. L'édition du *Vilain Asnier* approchant à grands pas, suivant le plan de publication (vol. 8 [1994]), on s'est aperçu que le témoin était introuvable avec cette cote, et on l'a évacué : l'*Errata* du vol. 7 (1993) sanctionne formellement la suppression du ms. du Mans[8] et l'édition du fabliau se fonde uniquement sur le témoignage du ms. *D*. Enfin, l'*Appendice* et le *Supplément bibliographique* du vol. 10 (1998) ne reviennent pas sur l'incident[9]. Bien entendu, le manuscrit est toujours resté à sa place, au Mans. Ce qui est captivant dans cette histoire d'arrière-cuisine, et qui donne à réfléchir, c'est que l'existence et la valeur de cette version du fabliau n'ont jamais été effacées des travaux critiques concernant les autres textes du recueil, notamment le *Commentaire sur le Cantique des Cantiques*. Au contraire, elles y ont été mises en valeur, en bénéficiant de mises au point éclairées[10].

---

4 Les détails et précisions des cas présentés sont réservés au volume collectif, dirigé par Olivier Collet, Francis Gingras et Richard Trachsler, qui réunira l'ensemble des résultats obtenus au sein du projet, et dont la publication est en cours.

5 Sur les coordonnées spatio-temporelles de ce recueil, cf. Hasenohr (1976).

6 Aucun soupçon qu'il existe une autre version que celle du NRCF (8, 207-214, 374-375) ne révèlent, par ex., Levy (2006) ni Whalen (2007, 151). Se soustraient à cet oubli généralisé Pearcy (2007, 221-223) et Ménard (1982, 199-200), qui imprime notre version dans les *Addenda aux notes* de la première édition (1979), où elle avait été oubliée.

7 NRCF (vol. 6, XXII).

8 NRCF (vol. 7, 454) : « Dans l'*Inventaire des fabliaux* 92. (149) Le Vilain Asnier Dm : supprimer m ; dans l'*Inventaire des manuscrits* : supprimer m Le Mans, Bibl. mun., 5495 92 ». Ce n'est donc pas « its context » qui « precluded it from inclusion in NRCF », ainsi que le voudrait Willem Noomen, d'après Pearcy (2007, 222).

9 L'*Appendice* (NRCF [vol. 10, 305-313, 383-385]) est entièrement consacré au deuxième témoin de la pièce *Des Putains et des Lecheors* et à son texte excellent, négligé auparavant (vol. 6, 145-153, 335), mais remis à l'ordre du jour par Straub (1993) et Ménard (1997) ; le *Supplément* (NRCF [vol. 10, 315-335]) ne s'attarde pas sur notre fabliau.

10 Cf., par ex., Hunt (1980), Zink (2001), Paradisi (2009, 90-95, 126-128). Pearcy (2007, 221-223) ne semble pas connaître l'analyse avisée de Michel Zink, qui lui aurait épargné nombre d'imprudences.

Le second cas concerne un volume complètement détruit au cours des bombardements de la Seconde Guerre mondiale (ancien ms. 620 de la Bibl. mun. de Chartres [*f*]). Dans les travaux postérieurs concernant ses textes, les négligences et approximations se multiplient[11]. Pourtant, on peut en reconstituer le contenu et, dans ses grandes lignes, la composition matérielle à l'aide des descriptions publiées avant 1944[12]. En plus, l'histoire et l'étendue des dégradations auxquelles il a été exposé à l'époque moderne, avant 1944 et avant même les débuts de la prise en compte des manuscrits comme objets d'étude, ont beaucoup à nous apprendre. Par ex., qu'*Auberee* n'était probablement pas le seul fabliau transmis par ce recueil : ses 67 vers finaux se conservaient au f. 129 uniquement parce qu'on désirait préserver le début de l'*Isopet*, qui occupe le *verso* ; autrement, ce feuillet aurait été arraché, comme l'ont été les quatorze ou quinze feuillets qui précédaient. Or, le fabliau de l'entremetteuse ne pouvait remplir qu'entre quatre et six feuillets. Il s'ensuit que les huit à onze feuillets précédents contenaient autre chose. Au vu de la logique des mutilations, il n'est pas hasardeux d'avancer que, si ces textes devaient paraître déplacés parmi les œuvres d'édification, à celui qui en a ôté brutalement le support, et si *Auberee* en faisait partie, il pouvait bien s'agir de fabliaux.

Le troisième cas, plus classique, est celui du démembrement d'un volume pour en faire des gardes ou des renforts de couverture. C'est ce qui s'est produit à l'abbaye de Clairvaux au tout début du XVI^e^ siècle, à partir d'un recueil d'envergure dont André Vernet a collecté quelques pièces[13] : Paris, BnF, naf. 934, f. 9-10, Troyes, Méd. du Grand Troyes, 1511 (garde) et, peut-être, 2139 (contre-garde), plus un fragment soutiré à Troyes par Guglielmo Libri et porté à présent disparu. Or, ce travail précieux n'a laissé aucune trace dans la plupart des études et des éditions concernant les fabliaux[14]. On a d'ailleurs dénommé les deux fragments avec deux sigles différents, *k* (Troyes, MGT, 1511) et *q* (BnF, naf. 934)[15], et on continue de les traiter comme deux fragments n'ayant aucun rapport entre eux, ni avec les autres pièces du puzzle. Pourtant, la richesse des pistes dégagées par la recherche érudite est le plus souvent inestimable. C'est le cas, pour rester dans notre enceinte, du ms. L.V.32 de la Bibl. Nazionale Universitaria de Turin (*U*), entièrement brûlé dans l'incendie survenu en 1904. Jusqu'ici, on reconstituait l'agencement de cet important recueil datant des dernières années du XIII^e^ siècle à partir de la description donnée par Auguste Scheler en 1867,

11 Cf., par ex., Spiele (1975, 148) : perte des v. 145-438 de la *Bible*, à la suite de l'ablation de quatre feuillets entre les f. 49 et 50. La récente mission de restauration, de reproduction numérique et d'étude des manuscrits de la Bibl. mun. de Chartres (cf. Grousson [2010]) a confirmé que le volume est entièrement détruit.

12 Notamment Meyer (1894), ainsi que les travaux des érudits chartrains (Pierre-Alexandre Gratet-Duplessis, Maurice Jusselin etc.).

13 Vernet (1973). Le spécialiste fit connaître ses découvertes en 1951, au sein de la Société Nationale des Antiquaires de France (séances du 10 et 17 janvier 1951 : compte rendu détaillé dans le *Bulletin de la Société Nationale des Antiquaires de France* 1950-51, 146-147, 148-149).

14 Mais il n'a pas échappé à Cobby (2009, 8).

15 NRCF (vol. 1, XX à vol. 10, XXII).

et on récupérait les deux tiers de sa cinquantaine de textes grâce à la transcription de Georges-Jean Mouchet (ms. BnF, Moreau 1727)[16]. Or, des fouilles ciblées m'ont permis de corriger ou de préciser la description de Scheler à l'aide d'une table détaillée du recueil, rédigée probablement par Mouchet lui-même en guise d'aide-mémoire (BnF, Moreau 1725, f. 1-3), et des notes et transcriptions prises par Paul Meyer au cours de ses missions dans les bibliothèques d'Italie (BnF, naf. 23955-23971). Au vu de la perte définitive du volume et du manque de reproductions photographiques, ainsi que des compétences exceptionnelles du chartiste, une annotation au crayon s'avère être capitale, et infléchira dorénavant notre regard sur ce recueil disparu : « Je ne sais sur quoi se fonde Scheler pour dire qu'il y a diverses mains. C'est la même » (BnF, naf. 23965, f. 566r).

Le quatrième cas est moins courant, mais illustre bien les dégradations auxquelles nos recueils ont été exposés à l'époque moderne sur le marché du livre ancien. Le ms. 3114 de l'Arsenal de Paris (*e*) se compose de dix-huit feuillets et de six textes brefs, dont le fabliau de *La Dame escoillee*. Il serait tentant, au vu des développements récents de celle que certains dénomment la 'New Codicology', de disserter sur l'assemblage des six textes dans ce mince témoin et sur les rapports qu'ils entretiennent entre eux, ou encore – mais ce serait beaucoup demander –, sur l'unité et le sens de ce *libellus*[17] ayant brillamment échappé à la fois à l'insertion en recueil et à l'épreuve du temps. Mais l'étude des textes et de leur tradition m'a fait découvrir que ces dix-huit feuillets ont été découpés au XVIII^e^ siècle d'un recueil assez imposant qui ne comptait, à l'origine, pas moins de deux douzaines de textes (de longueur et de teneur extrêmement variables) et que leur réunion dans le volume actuel de l'Arsenal date de la même époque. Puisque j'ai pu identifier le gros du recueil originel, lui aussi conservé dans une bibliothèque parisienne, et par conséquent localiser et dater l'ensemble avec une précision inhabituelle (ancien diocèse de Soissons, années 1290), de tout autres horizons d'analyse s'ouvrent aux chercheurs[18].

---

[16] La description qui importe ici est contenue dans le dernier segment de Scheler (1866-67). Bien entendu, cette étude n'a pas toujours été exploitée correctement : par ex., les responsables de la toute nouvelle édition du *Lai du Conseil* (Grigoriu / Peersman / Rider [2013, 8]) ne s'aperçoivent pas du fait que le recueil de Turin conservait, certes, les v. 1-335 de la pièce (f. 233r-234v), mais aussi ses derniers vers (858-862), en tête du f. 180r (ils ont été transcrits, mais non identifiés, par Scheler [1866-67, 26]), et que l'éparpillement et le caractère fragmentaire de ces composantes est dû aux altérations brutales dont le ms. a fait l'objet au cours de son existence. En dernier lieu, cf. Braccini (2001 : mais les imprécisions y foisonnent), Capusso (2005), Capusso (2006).

[17] J'entends par là un ensemble de cahiers non reliés, ou bien « couverts d'une simple feuille de cuir, de parchemin ou d'étoffe et […] cousus très légèrement » (Vezin [1997, 64]), conçu comme une entité relativement autonome et dotée de cohérence au niveau du contenu.

[18] L'une de ces pistes – la production de manuscrits en langue vernaculaire dans l'ancien diocèse de Soissons au XIII^e^ siècle, au-delà des livres de dévotion mieux connus (cf., par ex., Stones [2013, vol. 1, 38, 68 ; vol. 2, 480-485]) – a été parcourue par moi-même lors des journées d'études qui se sont tenues à l'Université de Montréal du 24 au 26 octobre 2013, cf. Giannini (sous presse).

Le dernier cas nous laisse entrer de plain-pied dans l'atelier du projet. Grâce aux travaux d'Alison Stones et de François Avril, on sait depuis une vingtaine d'années que l'Arsenal 3527 (*p*) – un recueil pieux axé sur la dyade classique *Vie des Pères-Miracles de Nostre Dame* de Gautier de Coinci (la première *Vie*, sauf les contes 14 et 29, et un choix du premier livre des *Miracles*) – a été illustré par deux artistes[19]. L'enlumineur principal, dénommé le Maître de Bute, a travaillé sur au moins une quinzaine d'autres manuscrits, tous identifiés, entre le milieu des années 1270 et 1290 environ ; son collaborateur sur cinq autres volumes. Parmi ces derniers se trouve le *Perceval* de Mons (Bibl. Centrale de l'Université de Mons, R 2/C 331/206), où le roman de Chrétien de Troyes est précédé de deux prologues et suivi de ses continuations. Personne, à ma connaissance, ne s'est depuis penché sur ce *cluster*, afin de vérifier si les données codicologiques, paléographiques, scriptologiques etc., confirment ou infirment les conclusions de spécialistes de l'enluminure et, en cas de réponse affirmative, pour mieux appréhender cette importante unité de production de livres latins et vernaculaires. Or, le copiste de l'Arsenal 3527, à l'écriture autant nette qu'inconstante, est aussi celui du *Perceval* de Mons ; organisation des cahiers, dimensions, mise en page, réglure et mise en texte sont aussi les mêmes ; les dispositifs régissant la décoration et l'illustration dans les deux volumes coïncident. Cela établi, il ne s'agit pas de deux volets d'un même recueil, mais de deux produits d'un système de production standardisé, où interviennent de surcroît les mêmes artisans[20]. Ce qui importe, désormais, c'est moins de rectifier certaines méprises – ainsi des localisations discordantes (Arras [le *Comte de Poitiers* de l'Arsenal 3527] / Tournai [v. 1-2000 du *Perceval* de Mons]) établies pour des copies exécutées en réalité par le même copiste –[21] que de travailler enfin sur l'ensemble des textes qui étaient à disposition au sein de la même unité de production. Les spéculations du chercheur prendront désormais de la hauteur.

Ce balayage de fond va nous permettre d'esquiver les embûches les plus grossières, certes. Mais quelles tendances fondamentales et indications typologiques se dégagent de l'examen des 'petits recueils', un travail à tel point minutieux et morcelé qu'il risque de se perdre dans l'irréductibilité des cas d'espèce ? Il est trop tôt pour tirer des conclusions et dégager des pistes d'interprétation. On se bornera donc à quelques constats, non impressionnistes, mais nécessitant tout de même davantage d'enquêtes ciblées et, surtout, d'affinage.

Le premier constat frôle la banalité[22]. Aux XIII^e^ et XIV^e^ siècles, le fabliau peut intégrer n'importe quel véhicule de transmission vernaculaire apte à assimiler sa forme versifiée, et il ne semble souffrir de cloisonnement ou d'interdit d'aucun genre,

---

19 Stones (1993, 243-250), Stones (1995), Avril (1998, 292-293, 297-298), Stones (2013, vol. 1, 38-39, 63-64, 65 ; vol. 2, 278-285, 289-296, 332-334).

20 Sans qu'il soit nécessaire d'avoir recours à la notion, commode mais périlleuse, d'atelier. Cf. la mise en garde de Collet (2013).

21 Dees (1987, 522, 524). La localisation tournaisienne est acceptée par Busby (1993, XXIII).

22 À la lumière de réflexions telles celles, par moments malicieusement excessives, de Varvaro (2001).

dû à son ton ou à son contenu[23]. Le recueil contemporain, qu'il soit à tonalité profane, didactique ou pieuse, plus ou moins axé sur un auteur, lui ouvre naturellement ses portes, sans condition apparente : le cas d'*Auberee* dans le manuscrit de Chartres, flottant entre la *Bible* d'Herman de Valenciennes, les prières à la Vierge et la *Vie de sainte Marguerite*, et celui de l'Arsenal 3527, où *Sacristain* précède la *Passion* dite *des jongleurs* au sein d'un recueil de contes dévots et miracles, en sont des exemples saisissants ; cette véritable bibliothèque dominée par les romans en vers qu'est la seconde unité codicologique du ms. BnF, fr. 375 (f. 34-163 [*a*])[24], où un strapontin est néanmoins réservé, et à deux reprises (f. 295v-296r, 344r-v, plus le résumé du f. 34r)[25], à *La vieille Truande*, en est un spécimen de plus ; même au sein des recueils de textes en prose, tels le fr. 12581 de la BnF (*Queste del Saint Graal*, *Tresor* de Brunet Latin, *Bible du XIII^e siècle* etc.), le fabliau, déployé pour l'occasion à longues lignes, arrive à se frayer un chemin (*Les Tresces*). De ce point de vue, le fabliau est un texte comme les autres, sans connotation particulière établie d'avance, si ce n'est la liberté de mouvement que lui garantit sa brièveté.

On observe ensuite que sa position dans le recueil semble obéir à deux tendances de fond suffisamment claires : elle est soit marginale, dans tous les sens de cet adjectif (à la marge, secondaire, décalé, adventice etc.), soit non marquée, lorsque le fabliau est incorporé dans un *continuum*. La position marginale est caractéristique des apparitions les plus anciennes du genre – comme dans le ms. du Mans, que l'on date prudemment du milieu du XIII^e siècle, malgré son allure archaïque, due à l'emprise du modèle monastique et à sa localisation périphérique, où *Le Vilain asnier* fait surface sous forme d'apologue conclusif mettant en garde le lecteur du *Commentaire sur le Cantique des Cantiques*, ou dans la section du ms. WLC/LM/6 de Nottingham (*G*) contenant les fabliaux de Gautier le Leu et d'autres, décalée et vraisemblablement postérieure au gros du recueil, qui daterait du deuxième quart du XIII^e siècle, d'après les dernières estimations[26], ou encore, dans le recueil épique très ancien (Angleterre, début du XIII^e siècle) dont deux bifeuillets ont échappé à la dispersion (BnF, naf. 5094 et Clermont-Ferrand, Arch. dép. du Puy-de-Dôme, 1 F 2), le dernier desquels porte une version de *Cele qui fu foutue et desfoutue* ajoutée tardivement par une main adventice[27]. Mais cet emplacement n'est pas l'apanage des témoignages les plus

[23] Même les pièces lyriques et théâtrales, qui ont emprunté des voies de transmission et de conservation à part (cf. Hasenohr [1990, 329-340]), ne semblent pas exclure sèchement la contiguïté du fabliau, comme le témoigne le cas, certes singulier, du ms. 12581 de la BnF (*X*) et de ses intermèdes lyriques.

[24] Sur ce « manuscrit de métier ou de confrérie, dans lequel chacun pouvait apprendre (lire, copier) le ou les textes qu'il désirait adjoindre à son répertoire ; où chacun, grâce à Perrot de Nesle, trouvait non seulement le texte mais son mode d'emploi », cf. Gasparri / Hasenohr / Ruby (1993, 119-125, 144-146 [146]), puis Hasenohr (1999, 43-44, 45, 46-48).

[25] Le doublet n'est pas isolé, au sein de la tradition manuscrite des fabliaux : *Le Vilain de Baillel* est transcrit aux f. 239v-240r, puis 255r-v du ms. BnF, fr. 12603 (*F*).

[26] Renseignement fourni par Maria Careri et Patricia Stirnemann, qui révisent ainsi la datation (premier quart du XIII^e siècle) suggérée par Stones (2010, 41-46, 49-56).

[27] Cf. en dernier lieu Di Luca (2015).

reculés, car il persiste et fait surface régulièrement, jusqu'au XIV^e^ siècle : par ex., dans l'attachant recueil épique, sur fond d'histoire sacrée, conservé maintenant à Turin (BNU, L.II.14 [*r*]), datant de 1311 ou peu après, *La Housse partie* s'agrippait, avant l'incendie de 1904, aux derniers feuillets de ce volume colossal et somptueux[28]. Bref, la marginalité, liée sans doute à double fil à la brièveté, pourrait constituer un trait durable, quoique non totalisant, de l'histoire de la tradition manuscrite du genre, au-delà des facteurs chronologiques et évolutifs.

En revanche, lorsque le fabliau plonge dans une suite plus ou moins hétérogène de textes, sa nature et sa position ne sont pas marquées, la plupart du temps, et il a tendance à se fondre dans le décor : dans l'Arsenal 3527, où les textes sont agencés comme les grains d'un chapelet, sans hiérarchisation ni brisure nette, la première partie de *Sacristain*, fabliau échevelé et burlesque, est mise sur le même pied que les contes de la *Vie des Pères* et les miracles de Gautier de Coinci ; dans la seconde section du ms. 2173 de la BnF (et dans sa copie tardive, le Bodmer 113 de Cologny [*l*]), les fabliaux font suite aux *Fables* attribuées à Marie de France, sans écart ou balise décelables, et deux d'entre eux (*Celui qui bota la Pierre* et *La Coille noire*) arrivent même à se faufiler au sein du fablier (f. 78v-79r, 92r-93r), là encore, avec nonchalance[29] ; les trois fabliaux de Jean de Condé conservés par le recueil organique de Rome (Bibl. Casanatense, 1598 [*N*]) se coulent parfaitement dans le florilège de l'œuvre du ménestrel de la cour du Hainaut offert ici, ainsi que dans les structures très homogènes de son support matériel. Le fabliau ne trouble donc nullement cet « effet de continuité qui gomme ou estompe l'individualité des constituants d'un recueil et assimile leur diversité à la 'globalité' et à la linéarité d'un texte unique », effet courant dans la plupart des recueils littéraires des XIII^e^ et XIV^e^ siècles, que la critique moderne peine à reconnaître[30]. Cela dit, si l'on se penche sur les articulations internes des recueils, des connexions préférentielles se dessinent – je pense, par ex., à la contiguïté récurrente des genres vernaculaires relevant, de près ou de loin, de l'apologue –, et viendra le moment de se consacrer à leur exploration.

La troisième observation part du fait qu'on répète, un peu trop souvent, que les fabliaux sont rarement illustrés[31]. Ce qui est à peu près vrai, mais susceptible de faire

[28] Je me permets de renvoyer à ma contribution récente, Giannini (2012).

[29] Fait qui n'est pas relevé par Pickens (2007), étude au titre pourtant prometteur.

[30] Azzam / Collet / Foehr-Janssens (2010, 19). Cf. déjà Hasenohr (1990, 277-283).

[31] Cf., par ex., Busby (2002, vol. 1, 457) ou Meneghetti / Bertelli / Tagliani (2012, 79), où la remarque appelle toutefois des rectifications : « quasi nessuno dei testimoni manoscritti dei fabliaux è dotato di immagini, anche se alcuni riservano degli spazi, rimasti vuoti, alla decorazione : le uniche due eccezioni sono appunto il fr. 2173 e un altro testimone, il codice 113 della Biblioteca Bodmeriana (Cologny-Genève) ». Dans les manuscrits suivants, un ou plusieurs fabliaux sont illustrés : Berne, BB, 354, Cologny, Fondation Bodmer, Bodmer 113, Oxford, Bodleian Library, Douce 111 (*o*), Paris, Bibl. de l'Arsenal, 3525 (*S*) et 3527, BnF, fr. 2188 (*j*), fr. 2173, Rothschild 2800 (*g*), peut-être le recueil détruit de Turin, BNU, L.V.32 (cf. Scheler [1866-67, 2]) ; le fr. 24432 de la BnF (*P*) porte des réserves destinées à l'illustration des textes, parmi lesquelles une consacrée à *Boivin de Provins*. D'autres recueils présentent une ou plusieurs illustrations, mais les fabliaux qu'ils préservent en sont dépourvus, pour dif-

perdre de vue l'essentiel, c'est-à-dire la cohérence des partitions et le suivi des dispositifs. De façon générale, les fabliaux profitent de la décoration et de l'illustration en place dans la section ou dans l'unité codicologique qui les héberge, sans écart par rapport aux textes qui les entourent : dans la seconde section ms. fr. 2173 de la BnF (et, bien entendu, dans sa copie de la Fondation Bodmer), les fabliaux sont accompagnés de dessins coloriés au même titre que les fables, qui constituent le texte majeur de la section ; de même, tous les textes de l'Arsenal 3527 jouissent d'une lettre historiée censée fournir un instantané du récit qui s'ensuit, et *Sacristain* ne s'en prive pas (est peint le coup de gourdin fatal de Guillaume) ; le ms. 475 de Chantilly (*T*), un autre recueil pieux reposant sur l'axe *Vie des Pères-Miracles* de Gautier de Coinci, est décoré, dans ses tronçons les plus anciens, de façon sobre mais accomplie et somme toute élégante, avec des lettres puzzle rouges et bleues au début des textes et des lettres filigranées à l'intérieur, dans la ligne de la production vernaculaire contemporaine au coût maîtrisé, et les fabliaux et dits n'échappent pas à la règle. Encore une fois, dans les 'petits recueils', les fabliaux sont des textes comme les autres, et ils sont traités comme tels. Ensuite, on ne peut pas nier que leur position volontiers marginale n'a pas encouragé les efforts des artisans au sujet du suivi de l'ornementation, et il est certain que les recueils qu'ils intègrent, entre la deuxième moitié du XIII^{e} siècle et le début du suivant, sont façonnés assez souvent sur un modèle typologique qui, pour des raisons essentiellement budgétaires, dirait-on aujourd'hui, renonce d'emblée à l'enluminure et privilégie la décoration sérielle à la plume.

Le dernier constat est une invitation à jouer constamment en sourdine. Une portion consistante de notre *corpus* (Chantilly, Bibl. du Château, 475, BnF, fr. 1635, fr. 2173, fr. 25545 etc.) est représentée par des recueils formés à partir de *libelli* assemblés, complétés et reconfigurés suivant des trajectoires diverses, parfois contradictoires, le plus souvent difficiles à démêler avec certitude. Les volumes se situant, à des emplacements et selon des modalités à chaque fois changeantes, entre le recueil cumulatif et le recueil composite sont donc nombreux, tandis que les recueils organiques dont l'intégralité est garantie – les seuls, à la rigueur, qui s'avèrent être « légitimement utilisables pour traiter des questions liées à la réception des œuvres » –[32] ne sont pas légion. Cette situation ne peut qu'inciter à la prudence, à tous les égards. Même lorsque la mise en recueil des *libelli* s'est produite – et ce n'est pas, là non plus, le cas de tous les livrets, ainsi que pourrait l'attester l'exemple, tout à fait exceptionnel, du ms. fr. 2188 de la BnF (*Trubert*) –, ils n'ont pas eu la vie facile. Au Moyen Âge déjà, ils ont fait l'objet d'altérations diverses, allant de la mutilation à la reprise : la silhouette modeste et le caractère inachevé de beaucoup d'entre eux ont sans doute poussé aux

férentes raisons : ancien Chartres, Bm, 620, Le Mans, Méd. Louis Aragon, 173, Londres, British Library, Add. 10289 (*Y*), Nottingham, UL, WLC/LM/6, Paris, Bibl. de l'Arsenal, 3524 (*R*), BnF, fr. 837, fr. 1446 (*W*), fr. 1553 (*J*), fr. 1593, fr. 12581 (*X*), fr. 12603 (*F*), Turin, BNU, L.II.14 (s'ajouterait le fr. 2168 de la BnF [*H*], si l'on incluait les schémas de l'*Image du monde*).

[32] Hasenohr (1999, 39).

manipulations les plus diverses[33]. Mais c'est à l'époque moderne que revient le gros des dégâts. La pandémie du démembrement et du remploi des manuscrits médiévaux qui a déferlé sur l'Europe aux XVI^e^ et XVII^e^ siècles et qui fut particulièrement intense dans le cadran nord-oriental du domaine d'oïl[34], a laissé des séquelles au cœur de notre *corpus*, où les fragments récupérés d'anciennes couvertures foisonnent : outre les pièces conservées (à présent ou autrefois) à Troyes, Clermont-Ferrand et Paris, déjà mentionnées, d'autres ont fait surface à Lyon (Bibl. mun., 5495) et à Audenarde (Kerkarchief Walburga, 32). Ensuite, vinrent les grands bibliophiles du XVIII^e^ siècle (les Paulmy, La Vallière etc.) et les pratiques douteuses auxquelles s'adonnèrent les marchands de livres qu'ils soudoyaient – l'Arsenal 3114 en est un effet direct –[35], puis les pilleurs de la trempe d'un Guglielmo Libri – si l'on songe au fait qu'à Troyes, il s'attaqua même aux gardes anciennes –, sans oublier les dommages occasionnés entre-temps par la pruderie qui animait certains conservateurs ou lecteurs, à la sensibilité bien moderne – le recueil de Chartres en fut largement victime, avant que l'aviation alliée ne lui assène le coup de grâce. Cela, simplement, pour rappeler qu'il n'est pas sûr qu'on puisse imaginer une quelconque proportion, quantitative, bien sûr, mais également qualitative, entre ce que nous avons sous les yeux aujourd'hui et ce qui circulait effectivement lorsque le genre s'épanouissait. Mais ça, c'est de l'histoire connue.

Université de Montréal Gabriele GIANNINI

Références bibliographiques

Avril, François, 1998. « Manuscrits », in : *L'Art au temps des rois maudits, Philippe le Bel et ses fils (1285-1328)*, Paris, Éditions de la Réunion des musées nationaux, 256-334.

Azzam, Wagih / Collet, Olivier / Foehr-Jannsens, Yasmina, 2010. « Mise en recueil et fonctionnalités de l'écrit », in : Collet, Olivier / Foehr-Jannsens, Yasmina (ed.), *Le recueil au Moyen Âge. Le Moyen Âge central*, Turnhout, Brepols, 11-34.

Braccini, Mauro, 2001. « Unica e esemplari creduti irrecuperabili dopo l'incendio della Biblioteca Nazionale di Torino : un ulteriore controllo sulla copia settecentesca del cod. L.V.32 », *SMLV* 47, 191-204.

[33] Un exemple éclairant de recueil issu d'un système de production par *libelli* et soumis à de multiples campagnes de démembrement et de réassemblage est offert par le ms. BnF, fr. 25545 (I). Cf. Collet (2008, 303-311).

[34] Cf. en dernier lieu les remarques introductives de Giannini / Nieus / Palumbo (2014).

[35] L'opération inverse, mais parfois consécutive au dépeçage, la réunion d'éléments disparates sous la même reliure (recueil factice), ne fut pas moins pratiquée aux XVII^e^-XVIII^e^ siècles, au sein des mêmes réseaux, et s'avère tout autant insidieuse, comme l'a montré Short (1988, 11-12).

Busby, Keith (ed.), 1993. Chrétien de Troyes, *Le Roman de Perceval ou le Conte du Graal*, Tübingen, Niemeyer.

Busby, Keith (2002), *Codex and Context. Reading Old French Verse Narrative in Manuscript*, Amsterdam/New York, Rodopi, 2 vol.

Capusso, Maria Grazia, 2005. « Note sulla tradizione manoscritta del *Lai du Conseil* », *SMLV* 51, 27-57.

Capusso, Maria Grazia, 2006. « La copia settecentesca del *Lai du Conseil* (Paris, BNF, Collection Moreau 1727) », in : Sommariva, Grazia (ed.), *‹Amicitiae munus›. Miscellanea di studi in memoria di Paola Sgrilli*, La Spezia, Agorà, 1-18.

Cobby, Anne E., 2009. *The Old French Fabliaux. An Analytical Bibliography*, Woodbridge, Tamesis.

Collet, Olivier, 2008. «'Textes de circonstance' et 'raccords' dans les manuscrits vernaculaires : les enseignements de quelques recueils des XIII^e^-XIV^e^ siècle», in : Van Hemelryck, Tania / Colombo Timelli, Maria (ed.), *‹Quant l'ung amy pour l'autre veille›. Mélanges de moyen français offerts à Claude Thiry*, Turnhout, Brepols, 299-311.

Collet, Olivier, 2013. « Les 'ateliers de copistes' aux XIII^e^ et XIV^e^ siècles : errances philologiques autour du *Chevalier qui faisait parler les cons* », in : Corbellari, Alain *et al.* (ed.), *‹Philologia ancilla litteraturae›. Mélanges de philologie et de littérature françaises du Moyen Âge offerts au Professeur Gilles Eckard par ses collègues et anciens élèves*, Genève, Droz, 61-72.

Dees, Anthonij, 1987. *Atlas des formes linguistiques des textes littéraires de l'ancien français*, Tübingen, Niemeyer.

Di Luca, Paolo, 2015. « Deux fragments anglo-normands de la Chanson d'Aspremont : description et étude de P4 et C », in : Ailes, Marianne J. *et al.* (ed.), *Epic Connections / Rencontres épiques. Proceedings of the Nineteenth International Conference of the Société Rencesvals*, Oxford, 13-17 August 2012, Edimbourg, Société Rencesvals British Branch, vol. 1, 191-214.

Gasparri, Françoise / Hasenohr, Geneviève / Ruby, Christine, 1993. « De l'écriture à la lecture : réflexion sur les manuscrits d'*Erec et Enide* », in : Busby, Keith *et al.* (ed.), *Les Manuscrits de Chrétien de Troyes*, Amsterdam/Atlanta, Rodopi, vol. 1, 97-148.

Giannini, Gabriele, 2012. « Poser les fondements : lieu, date et contexte (essai sur le recueil L.II.14 de Turin) », *Études françaises* 48/3, 11-31.

Giannini, Gabriele / Nieus, Jean-François / Palumbo, Giovanni, 2014. « Nouveaux fragments du *Merlin en prose* et de sa *Suite Vulgate* (Namur, Archives de l'État, Arch. eccl. 1664) », *Le Moyen Âge* 120, 673-711.

Giannini, Gabriele, sous presse. « L'Arsenal 3114 et la production de manuscrits en langue vernaculaire dans l'ancien diocèse de Soissons (1260-1300 environ) », in : Giannini, Gabriele / Gingras, Francis (ed.), *Les Centres de production des manuscrits vernaculaires au Moyen Âge*, Paris, Classiques Garnier.

Grigoriu, Brînduşa E. / Peersman, Catharina / Rider, Jeff (ed.), 2013. *Le Lai du Conseil*, Liverpool, University of Liverpool. <www.liv.ac.uk/media/livacuk/cultures-languages-and-area-studies/liverpoolonline/Le_Lai_du_Conseil.pdf>.

Grousson, Mathieu, 2010. « Les manuscrits renaissent de leurs cendres », *Le journal du CNRS* 242, 36.

Hasenohr, Geneviève, 1976. Compte rendu de Pickford, Cedric E. (ed.), 1974. *The Song of Songs. A Twelfth-Century French Version Edited from Ms. 173 of the Bibliothèque Municipale of Le Mans*, Londres/New York/Toronto, Oxford University Press, *Cahiers de civilisation médiévale* 19, 289-293.

Hasenohr, Geneviève, 1990. « Traductions et littérature en langue vulgaire », in : Martin, Henri-Jean / Vezin, Jean (ed.), *Mise en page et mise en texte du livre manuscrit*, Paris, Éditions du Cercle de la Librairie-Promodis, 229-352.

Hasenohr, Geneviève, 1999. « Les recueils littéraires français du XIII^e siècle : public et finalité », in : Jansen-Sieben, Ria / van Dijk, Hans (ed.), *Codices Miscellanearum (Colloque Van Hulthem, Bruxelles 1999)*, Bruxelles, Association des archivistes et bibliothécaires de Belgique, 37-50.

Hunt, Tony, 1980. « The O.F. Commentary on the *Song of Songs* in MS Le Mans 173 », dans *ZrP* 96, 267-297.

Levy, John F., 2006. « *Le vilain asnier* : a 'perfect little exemplum'», *Reinardus* 19, 107-128.

Ménard, Philippe, 1997. « Une nouvelle version du dit *Des putains et des lecheors* », *ZrP* 113, 30-38.

Ménard, Philippe (ed.), 1998[2]. *Fabliaux français du Moyen Âge*, Genève, Droz.

Meneghetti, Maria Luisa / Bertelli, Sandro / Tagliani, Roberto, 2012. « Nuove acquisizioni per la protostoria del codice Hamilton 390 (già Saibante) », *Ctes* 15, 75-126.

Meyer, Paul, 1894. « Notice sur le ms. 620 (ancien 261) de la Bibliothèque de Chartres », *Bulletin de la Société des anciens textes français* 20, 36-60.

NRCF = Noomen, Willem / van den Boogaard, Nico (ed.), 1983-98. *Nouveau Recueil Complet des Fabliaux (NRCF)*, Assen-Maastricht, Van Gorcum, 10 vol.

Paradisi, Gioia, 2009. *La parola e l'amore. Studi sul* Cantico dei Cantici *nella tradizione francese medievale*, Rome, Carocci.

Pearcy, Roy J., 2007. *Logic and Humour in the Fabliaux. An Essay in Applied Narratology*, Cambridge, Brewer.

Pickens, Rupert T., 2007. « Marie de France in the Manuscripts : Lai, Fable, Fabliau », in : Burr, Kristin L. / Moran, John F. / Lacy, Norris J. (ed.), *The Old French Fabliaux. Essays on Comedy and Context*, Jefferson, McFarland & Co., 174-186.

Rychner, Jean, 1960. *Contribution à l'étude des fabliaux. Variantes, remaniements, dégradations*, Genève, Droz, 2 vol.

Scheler, Auguste, 1866-67. « Notice et extraits de deux manuscrits français de la Bibliothèque royale de Turin », *Le bibliophile belge* 1, 246-279, 343-374; 2, 1-33.

Short, Ian, 1988. « L'avènement du texte vernaculaire : la mise en recueil », in : Baumgartner, Emmanuèle / Marchello-Nizia, Christiane (ed.), *Théories et pratiques de l'écriture au Moyen Âge. Actes du Colloque (Palais du Luxembourg – Sénat, 5 et 6 mars 1987)*, Nanterre, Université Paris X-Nanterre, 1988, 11-23.

Spiele, Ina (ed.), 1975. Li Romanz de Dieu et de sa Mere *d'Herman de Valenciennes chanoine et prêtre (XII^e siècle)*, Leyde, Presse Universitaire de Leyde.

Stones, Alison, 1993. « The Illustrated Chrétien Manuscripts and their Artistic Context », in : Busby, Keith *et al.* (ed.), *Les Manuscrits de Chrétien de Troyes*, Amsterdam/Atlanta, Rodopi, vol. 1, 227-322.

Stones, Alison, 1995. « Stylistic Associations, Evolution, and Collaboration : Charting the Bute Painter's Career », *The J. Paul Getty Museum Journal* 23, 11-29.

Stones, Alison, 2010. « Two French Manuscripts : WLC/LM/6 and WLC/LM/7 », in : Hanna, Ralph / Turville-Petre, Thorlac (ed.), *The Wollaton Medieval Manuscripts. Texts, Owners and Readers*, York, York Medieval Press, 41-56.

Stones, Alison, 2013. *Gothic Manuscripts, 1260-1320. Part one*, Londres, Harvey Miller, 2013, 2 vol.

Straub, Richard, 1993. « *Des Putains et des Lecheors* : la version oubliée du manuscrit *G* », *VR* 52, 164-179.

Trachsler, Richard, 2010. « Observations sur les 'recueils de fabliaux' », in : Collet, Olivier / Foehr-Jannsens, Yasmina (ed.), *Le recueil au Moyen Âge. Le Moyen Âge central*, Turnhout, Brepols, 35-46.

van den Boogaard, Nico, 1984. « La définition du fabliau dans les grands recueils », in : Bianciotto, Gabriel / Salvat, Michel (ed.), *Épopée animale, fable, fabliau. Actes du IV^e colloque de la Société internationale renardienne (Évreux, 7-11 septembre 1981)*, Paris, PUF, 657-668.

Varvaro, Alberto, 2001. « Élaboration des textes et modalités du récit dans la littérature française médiévale », *R* 119, 1-75.

Vernet, André, 1973. « Fragments d'un *Moniage Richeut* ? », in : *Études de langue et de littérature du Moyen Âge offertes à Félix Lecoy par ses collègues, ses élèves et ses amis*, Paris, Champion, 585-597.

Vezin, Jean, 1997. « ‹Quaderni simul ligati›. Recherches sur les manuscrits en cahiers », in : Robinson, Pamela R. / Zim, Rivkah (ed.), *Of the Making of Books. Medieval Manuscripts, their Scribes and Readers. Essays presented to M. B. Parkes*, Aldershot, Scolar Press, 64-70.

Whalen, Logan E., 2007. « Modern Dirty Jokes and the Old French Fabliaux », in : Burr, Kristin L. / Moran, John F. / Lacy, Norris J. (ed.), *The Old French Fabliaux. Essays on Comedy and Context*, Jefferson, McFarland & Co., 147-159.

Zink, Michel, 2001. « Le *Cantique des Cantiques* et le *Vilain ânier* », in : Henrard, Nadine / Moreno, Paola / Thiry-Stassin, Martine (ed.), *Convergences médiévales : épopée, lyrique, roman. Mélanges offerts à Madeleine Tyssens*, Bruxelles, De Boeck, 631-641.

Dictionnaire Électronique de Chrétien de Troyes (*DÉCT*) : fin et suite[1]...

1. Exposé

Le *DÉCT1* est terminé ! C'est avec joie, et non sans une certaine fierté, qu'au nom de mes collaborateurs j'annonce enfin cette nouvelle dans le cadre de la 27e édition du congrès de notre société. Ce dictionnaire est l'un des fruits d'une entente conclue entre le Laboratoire de Français Ancien (LFA) de l'Université d'Ottawa et l'INaLF (Institut National de la Langue Française) en 1995 (Robert Martin étant alors directeur) et renouvelée en 2003, sous le directorat de Jean-Marie Pierrel, l'ATILF ayant pris la suite de l'INaLF. On comprendra que nous ayons tenu à achever ce travail et à le présenter à l'occasion de notre congrès. Gilles Souvay et moi sommes d'ailleurs des habitués de ce genre d'exposé. Jamais deux sans trois : à Innsbruck en 2007, notre communication, avec Hiltrud Gerner, portait le sous-titre *État actuel du projet* ; en 2010 à Valencia, *Révision et élargissement* ; aujourd'hui, ce sera *Suite et fin* ou plutôt, comme l'indique le programme, de façon inattendue et surprenante, *Fin et suite…* ; on en verra la raison au fil de cette communication.

L'équipe de recherche s'est constituée à l'automne 2003, à l'occasion d'une demande de subvention au Conseil de recherches en sciences humaines du Canada. L'ouvrage actuel représente donc une dizaine d'années de travail. À ce stade-ci, l'équipe est constituée d'un rédacteur (Pierre Kunstmann), de trois personnes bénévoles chargées de la révision des articles (deux pour la version française : Hiltrud Gerner et May Plouzeau ; une pour la version anglaise : Ineke Hardy) et, last but not least, d'un informaticien (G. Souvay), qui s'est révélé d'une compétence remarquable et d'un dévouement extrême. Il importe de préciser que le rédacteur, qui est aussi le maître d'œuvre, est seul responsable de toute erreur qui a pu demeurer après le processus de révision ou qui a pu se glisser depuis, lors de modifications ultérieures qui n'ont plus été soumises à nos trois réviseuses.

La première partie de notre communication sera consacrée à l'histoire de la réalisation du dictionnaire ; la deuxième partie s'ouvrira sur un retour du philologique, contenu sinon refoulé au début de nos travaux, et traitera de la mise en place, de la mise en phase, du *DÉCT2* ; la troisième partie consistera en une démonstration, par Gilles Souvay, des fonctionnalités du dictionnaire actuel.

---

[1] Exposé de Pierre Kunstmann, suivi d'une démonstration de Gilles Souvay.

Tous les articles du *DÉCT* sont maintenant en ligne, du premier AAISEMENT («Soin, attention, prévenance») au dernier YVAIN, le héros du *Chevalier au Lion*. Tout un programme, tout un parcours ! Neuf ans de travail depuis l'octroi de la première subvention. Ce projet de recherche, nous l'avons toujours considéré comme un « travail en cours », c'est-à-dire en progression (*work in progress*), qu'il nous fallait présenter au public par étapes suffisamment rapprochées pour susciter et maintenir l'intérêt des utilisateurs, pour nous obliger aussi à avancer à une vitesse raisonnable dans l'élaboration de l'outil. La première version a été lancée officiellement à Nancy en mai 2007 : la base textuelle était déjà complète, le lexique se réduisait à la lettre A. La dernière mise à jour remonte au début du mois ; elle nécessitera encore, du point de vue informatique, quelques ajustements, mais à toutes fins utiles on peut légitimement tenir cet outil pour achevé.

Pour ceux d'entre vous qui ne connaissent pas encore cet ouvrage, il convient de préciser qu'il est fondé sur le principe de navigation : navigation interne, d'abord, entre trois zones (lexique, transcription des romans et base textuelle) ; navigation externe ensuite, grâce aux liens avec les dictionnaires de la même maison, l'ATILF (*DMF*, *FEW* et *TLFi*), mais également avec d'autres dictionnaires (Foerster-Breuer, TL, Gdf, DEAF, AND), tous en ligne à l'exception du Tobler-Lommatzsch. Dans sa démonstration, G. Souvay cliquera sur certains boutons pour nous faire naviguer ; de mon côté, je vais indiquer brièvement comment nous avons structuré les articles. Comme pour le *DMF*, la hiérarchie pour la numérotation des paragraphes est la suivante : au niveau supérieur sont employés les chiffres romains ; puis au fur et à mesure qu'on descend dans le détail, on trouve successivement les lettres majuscules, les chiffres arabes, les lettres minuscules, les tirets et les points. Les chiffres romains sont réservés aux parties du discours (emploi adjectif, emploi substantif, etc.) ainsi qu'à la distinction des constructions verbales (emploi transitif, direct ou indirect, emploi intransitif, emploi pronominal). Entre crochets sont spécifiées les conditions d'emploi ; c'est une balise dont le contenu est libre, c'est-à-dire qu'il n'est pas lié à une liste close d'éléments. Pour chaque paragraphe, nous avons fixé à 10 la limite du nombre d'exemples retenus.

Les exemples sont tirés directement des transcriptions des cinq romans effectuées à partir du manuscrit *P* (la copie du scribe Guiot). Pour partir du document brut et éviter toute intervention critique donc subjective, on aurait pu songer à fonder le dictionnaire sur des transcriptions diplomatiques, comme celles d'*Érec* ou d'*Yvain* qu'on peut consulter sur le site du LFA ; mais le traitement automatique aurait été impossible. Force nous a été de procéder d'emblée à deux interventions sur le texte : segmentation et ponctuation à la moderne. Interventions minimales certes, mais déjà critiques car elles peuvent nécessiter une certaine réflexion et entraîner un risque d'erreur. Après avoir payé notre dû à l'ordinateur, nous sommes allés plus loin et avons pensé signaler à l'utilisateur les fautes grossières du copiste ; celles-ci ont alors été marquées d'un astérisque et suivies d'une correction (ce que nous avons supposé être la « bonne » forme). La plupart du temps, ces indications venaient d'une compa-

raison avec les autres manuscrits. Le philologique est revenu en force ! J'en donnerai un exemple (détail peut-être, mais dont on aurait tort de négliger la valeur) :

> Mate et dolante et sopiranz,
> Monte la reïne et si dist
> An bas, por ce qu'an ne l'oïst :
> « Ha ! rois*[l. amis], se vos ce seüssiez,
> Ja, ce croi, ne l'otroiesiez,
> Que Kex me menast un seul pas ! » La 206-211

La leçon *rois* que présente le seul manuscrit *P* n'a aucun sens dans ce contexte ; nous suggérons *amis*, leçon (isolée aussi) du manuscrit de Chantilly, suivant ainsi l'exemple des éditions de Méla et de Poirion, lequel déclare en note que la version de *Ch* « donne un sens plus subtil et plus profond, avec son allusion à Lancelot (« *amis*) ».

Le souci de rigueur pour la préparation du *DÉCT* (coller absolument à nos transcriptions, lesquelles par un simple clic peuvent être vérifiées sur les images du manuscrit) a donc été quelque peu compensé (mais point du tout effacé) par la volonté d'informer le lecteur sur la qualité des leçons du copiste. Ceci s'est fait de différentes façons : dans le lexique par des corrections suggérées dans les citations ou par des observations introduites sous forme de commentaire d'exemple ou de remarque ; dans les transcriptions en XML (qui sont téléchargeables) par insertion dans diverses balises (*choice*, *gap*, *note*). Cela exige encore un certain réglage, mais cela nous pousse surtout à passer rapidement au *DÉCT2*, c'est-à-dire à notre seconde phase.

Le *DÉCT2* constituera une version enrichie par l'apport de quatre éléments nouveaux :

1. L'ajout des variantes des autres manuscrits. Deux cas sont à distinguer :

- celui des mots apparaissant dans toutes les balises *occurrence* des exemples figurant dans les articles du *DÉCT1*. On a prévu une balise *variante* pour insérer localement tous les mots lexicaux différents trouvés dans la varia lectio.
- celui des nouveaux lemmes, des nouvelles acceptions, des nouvelles constructions syntaxiques. On rédigera alors un autre article ou, dans un article du *DÉCT1*, de nouveaux paragraphes.

Ces variantes seront repérées non pas tant à la relecture des articles de la première phase qu'à la relecture des romans au fil des textes, à commencer par *Yvain*, à partir du travail de Kajsa Meyer.

2. L'ajout des mots grammaticaux (plus de 500 lemmes) : ils avaient été laissé de côté lors de la première phase, quoique bien identifiés déjà dans la base textuelle. On consacrera à chacun un article en précisant sens et emplois. Le travail de M.-L. Ollier apportera une aide précieuse ; nous nous appuierons aussi sur le modèle des articles du *DMF* dans ce domaine.

3. L'ajout des synonymes et des antonymes : le repérage s'effectuera suivant les lignes directrices de ma communication au colloque *La « logique » du sens, Autour des propositions de Robert Martin* (Metz 2011).

4. L'ajout d'un classement sémantique : nous partirons du cadre offert par WordNet et des possibilités qu'il présente. Nous utiliserons en particulier le double jeu de définitions français / English qu'offre le *DÉCT*, dictionnaire bilingue.

2. Démonstration

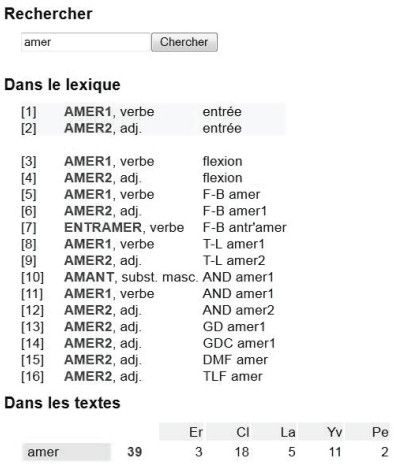
Celle-ci a consisté à montrer tout d'abord le lien aux images du manuscrit stockées sur gallica à l'adresse http://gallica.bnf.fr/ark:/12148/btv1b84272526.

Elle a présenté l'évolution des interfaces qui utilise désormais des onglets :

- au niveau de l'article :

- au niveau des fonctionnalités :

On a lancé une réflexion sur un nouveau mode d'entrée dans la base : en tapant un mot on recherche toutes les informations que l'on peut obtenir. Pour la forme *amer* par exemple :



Les textes vont prochainement entrer dans Frantext et dans la BFM (Base de Français Médiéval). Le *DÉCT* sera cité dans le *DMF* version 2015.

Pierre KUNSTMANN
Gilles SOUVAY

Références bibliographiques

DÉCT: Dictionnaire Électronique de Chrétien de Troyes, http://www.atilf.fr/dect, LFA/Université d'Ottawa ATILF/Université de Lorraine.

Kunstmann, Pierre / Gerner, Hiltrud / Souvay, Gilles, 2010, « Dictionnaire électronique de Chrétien de Troyes : état actuel du projet », in : Iliescu, Maria / Siller-Runggaldier, Heidi / Danler, Paul (ed.) : *Actes du XXV^e Congrès International de Linguistique et de Philologie Romanes*, Berlin / New York, Walter de Gruyter, vol. 2, 185-192.

Kunstmann, Pierre, 2011, « Synonymie et antonymie : des propositions de Robert Martin à la préparation de la seconde phase du *Dictionnaire Électronique de Chrétien de Troyes* », in : Duval, Frédéric (ed.) *La « logique » du sens, Autour des propositions de Robert Martin*, Metz, Université Paul Verlaine (*Recherches Linguistiques* 32), 301-317.

Kunstmann, Pierre / Gerner, Hiltrud / Souvay, Gilles, 2013, « Le Dictionnaire Électronique de Chrétien de Troyes (*DÉCT1*) : révision et élargissement », in : Casanova, Emili / Calvo, Cesáreo (ed.) : *Actas del XXVI Congreso Internacional de Lingüística y Filología Románica*, Berlin / New York, Walter de Gruyter, 2013, vol. 8, 195-204.

Statistiques et motifs séquentiels : le cas des suites [Nom préposition Nom], *mot à mot, pétale par pétale*

La contribution se propose d'évaluer l'outil statistique pour identifier et analyser les suites *Nom préposition Nom*, avec pour critère restrictif que les deux substantifs sont identiques[1].

L'enjeu essentiel de cette recherche est double : il est, d'une part, d'observer le rôle structurel des prépositions dans ces unités, à la fois sur le plan lexical et sémantique ; il est d'autre part d'évaluer le rendement de ces structures comme outils de classifications et, en particulier, comme marqueurs génériques. La base qui sert de corpus exploratoire à cette analyse permet en effet ce type de recherches, car elle est constituée d'un extrait de *Frantext* catégorisé qui comprend quatre groupes génériques ainsi répertoriés : deux ensembles narratifs d'une part, les mémoires (109 textes) et les romans (740 textes), la poésie (191 textes) et le théâtre (210 textes) d'autre part, soit près de 85 millions de mots. Les genres principaux sont ainsi représentés.

Le premier travail consiste à repérer les suites recherchées. Il s'avère que toutes les prépositions ne sont pas aptes à entrer dans ce type de constructions. Une réflexion pourra s'engager sur le statut des prépositions éligibles et sur leur portée sémantique. Le rendement des structures est évalué en observant ensuite les variations de substantifs de part et d'autre des prépositions : la préposition peut varier tandis que les noms restent identiques ; l'expression conserve globalement le même sens (*mot à mot, mot pour mot* ; *bouche à bouche, bouche contre bouche*) ; la préposition peut être la même tandis que les substantifs peuvent varier (*nuit après nuit, soir après soir* : le sens exprime ici l'idée de succession ; *rendre haine contre haine, corps à corps, nez contre nez* : le sens se module en l'idée d'échange, d'opposition, de contact).

L'analyse a aussi à composer avec les structures lexicalisées ; on s'intéressera aux critères qui autorisent une interprétation sur le degré de figement des structures, situées entre l'expression lexicalisée enregistrée en langue et la création lexicale. Entre phraséologie et style d'auteur, conçu comme appropriation des unités linguistiques, entre langue et discours, ces unités seront évaluées en fonction de leur productivité dans le corpus d'étude. L'observation des hapax – ou unités de fréquence 1 – sera

[1] Nos remerciements vont à G. Purnelle (Université de Liège ; LASLA) pour son aide quant à l'élaboration des résultats statistiques présentés dans ce travail, qui est un extrait d'une recherche plus ample en cours (en particulier sur le rendement de la théorie du motif, JADT 2012).

engagée comme possible critère distinctif de l'innovation lexicale. L'étude amorcera par ce biais une délimitation et une possible définition des unités lexicalisées.

Comme le corpus est constitué de quatre ensembles génériques, il sera loisible de voir enfin comment ces unités phrastiques se distribuent et si une répartition générique peut être proposée.

1. L'unité Nom préposition *Nom*

1.1. Observation du corpus

On commence par l'observation des données expérimentales fournies par le corpus d'étude afin de dégager quelles sont les prépositions éligibles, autrement dit susceptibles de relier immédiatement deux substantifs identiques, chacun étant dépourvu de déterminant.

Dans le corpus d'étude, des occurrences de la suite ternaire recherchée ne se rencontrent qu'avec les prépositions : *à*, *après*, *contre*, *par*, *pour*, *sur* ainsi qu'avec les structures plus complexes *de … à* et *de … en*[2].

Structure = N préposition N	Total	Exemple d'expression	Structure = N préposition N	Total	Exemple d'expression
à	2997	mot à mot	d(e) … en	4222	de page en page
après	247	Heure après heure	par	969	goutte par goutte
contre	253	cœur contre cœur	pour	398	coup pour coup
d(e) … à	372	d'ami à ami	sur	663	pierre sur pierre
			Total	10121	

Figure 1- Les structures éligibles

Toute réalisation d'une structure NpN à partir d'un substantif particulier est appelée *expression ou unité* : *heure après heure*, *d'ami à ami*. Sur le plan terminologique, on pourrait réutiliser le terme de « motif »[3] séquentiel ou de « cadre collocationnel »[4] dans le sens d'un patron syntaxique fixe.

---

[2] La préposition *selon* ne fournit qu'une seule occurrence : le « pli selon pli » de Mallarmé dans le sonnet *Remémoration d'amis belges* : « Que se dévêt pli selon pli la pierre veuve ».

[3] Le motif est défini de manière un peu abstraite dans certains travaux comme « un sous-ensemble ordonné [d'un ensemble] (E) formé par l'association récurrente de n éléments de l'ensemble (E) muni de sa structure linéaire ». D. Longrée, X. Luong, S. Mellet, 2008. « Les motifs : un outil pour la caractérisation topologique des textes », *9e JADT*, 733-744, <lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf>.

[4] A. Renouf et J. Sinclair, 1999. *English corpus Linguistics : Studies in Honour of Jan Svartvik*, Longman, Chapter Collocational Framaeworks in English, 128-143.

Sur le plan formel, les structures NpN peuvent en effet entrer dans la catégorie du motif syntaxique, qui est une des catégories identifiées du motif, à côté de celles constituées par les suites phonologiques ou encore métriques. La structure étudiée se caractérise par le fait que les substantifs de part et d'autre de la préposition sont en théorie réversibles puisqu'ils sont identiques dans leur forme ; cependant, l'ordre d'apparition des occurrences dans la phrase a forcément une incidence sur le sémantisme du lexème induisant une vectorisation du système. D'une occurrence à l'autre, le sens se module ; la répétition induit une variation de sens. La disposition des unités dans l'espace du texte oriente la structure en l'inscrivant dans la temporalité de la lecture cursive.

Le corpus d'étude fournit 10 122 occurrences de la structure ternaire choisie. Indépendamment de la partition générique, les expressions les plus fréquentes sont :

côte à côte	1107	nez à nez	165
goutte à goutte	359	jour par jour	150
coup sur coup	315	de main en main	133
de minute en minute	223	de porte en porte	124
corps à corps	192	tête à tête	108
de place en place	186		

Figure 2- Les expressions les plus fréquentes

1.2. Les unités et le paramètre sémantique

La plupart des prépositions sont polyvalentes et polysémiques. C'est le sens des substantifs de part et d'autre qui détermine le sens de la relation sémantique réalisée entre les noms et, par conséquent, le sens global de l'expression.

L'examen des expressions NpN du corpus permet d'établir une liste de huit expressions que le paramètre sémantique[5] distingue :

Succession : décomposition du procès en éléments

> Nous y secouons nos cigarettes: nous sommes lavés, nous avons le temps ; le temps chaud et stagnant, le temps qui passe *minute à minute*, sans bruit, sans fièvre, en chuchotant. (A. Sarrazin, *L'Astragale*)

[5] Sur 10121 occurrences, 11 n'entrent pas dans ces 8 catégories et sont donc soustraites des effectifs.

Accumulation : addition d'objets

La mère, toujours malade, comptait *sou à sou*. Elle tremblait devant lui, autant que nous. (R. Martin du Gard, *Les Thibault, l'Été 1914*).

Mon grand vizir est à la guerre où il remporte d'ailleurs *victoire après victoire* sur les Russes (M. De Grèce, *La Nuit du Sérail*)

Distance :

J'ai tendu des cordes *de clocher à clocher* ; des guirlandes *de fenêtre à fenêtre* ; des chaînes d'or *d'étoile à étoile*, et je danse. (A. Rimbaud, *Illuminations*)

Déplacement - mouvement

Je marche *de surprise en surprise*, *d'ahurissement en ahurissement*. (P. Benoit, *L'Atlantide*)

Relation :

D'où Nana tombait-elle ? Et des histoires couraient, des plaisanteries chuchotées *d'oreille à oreille*. (É. Zola, *Nana*)

Mille cris se croisaient *de baraque à baraque*. (P. Adam, *L'Enfant d'Austerlitz*)

Contact :

J'ai l'horreur de la fraternité qui établit des contacts *de peau à peau* (J. Genet, *Miracle de la rose*)

Disait Hérode à Sigognac, qui cheminait *botte à botte* avec lui (Th. Gautier, *Le Capitaine Fracasse*)

Opposition :

On se battit *corps à corps* sous les arbres (V. Hugo, *Les Misérables*)

J'ai lutté pendant quatre années jour et nuit, *corps à corps*, *astuce contre astuce*, avec ce génie infernal (Ponson du Terrail, *Rocambole*)

Échange :

Violence contre violence. Chacun donne ici sans mesure sa force. (Aragon, *Les beaux Quartiers*)

	à	après	Contre	de … à	de … en	par	pour	sur	Total
Accumulation	5	28	2	0	0	15	0	253	303
Contact	1790	0	190	2	0	0	0	51	2033
Déplacement	0	0	0	25	4222	0	0	0	4247

	à	après	Contre	de … à	de … en	par	pour	sur	Total
Distance	1	0	0	6	0	0	74	0	81
Échange	0	0	11	0	0	6	324	0	341
Opposition	171	0	50	0	0	0	0	0	221
Relation	30	0	0	329	0	0	0	0	359
Succession	994	219	0	5	0	948	0	359	2525
Total	2991	247	253	367	4222	969	398	663	10110

Figure 3 - Fréquences des structures et des sémantismes

En conservant soit la préposition, soit le substantif comme invariant d'une expression, et en faisant varier l'autre élément, on obtient les deux cas de figure suivants au service d'un processus interactif de constitution du sens :

> Une structure qui combine une même préposition avec des noms différents peut prendre des sens différents : la préposition n'a donc pas un sens stable et invariant, donné en langue, mais un sens toujours modulable selon les combinaisons où elle entre, son sens s'ajustant au substantif répété de part et d'autre. La préposition « contre » par exemple peut exprimer quatre types de relations sémantiques : l'*Opposition* : « argument contre argument » ; le *Contact* : « bord contre bord » ; l'*Échange* : « donnant donnant, enfant contre enfant » ; l'accumulation « bloc contre bloc ».

La polysémie de la structure N *contre* N est ainsi démontrée.

De manière symétrique, des substantifs identiques reliés par des prépositions différentes peuvent constituer une structure de même sens : par exemple, entre les expressions *lutter âme à âme* et *lutter âme contre âme* ou encore entre *bouche à bouche* et *bouche contre bouche* peut être postulée une équivalence de sens. Un même sémantisme est réalisé par plusieurs structures.

Les statistiques établissant des pourcentages valideront-elles ces hypothèses de travail ?

Structure	à	après	contre	d(e) … à	d(e) … en	par	pour	sur
Accumulation	0,17%	11,34%	0,79%			1,55%		38,16%
Contact	59,85%		75,10%	0,54%				7,69%
Déplacement				6,81%	100,00%			

Structure	à	après	contre	d(e) … à	d(e) … en	par	pour	sur
Distance	0,03%			1,63%			18,59%	
Échange			4,35%			0,62%	81,41%	
Opposition	5,72%		19,76%					
Relation	1,00%			89,65%				
Succession	33,23%	88,66%		1,36%		97,83%		54,15%

Figure 4 – Quels sémantismes pour les structures identifiées ?

Ce tableau présente les différents sémantismes associés à une même structure.

- Les structures les plus polysémiques sont *à* (6 sémantismes lui sont associés : l'accumulation, le contact, la distance, l'opposition, la relation, la succession), *de … à* (5 sémantismes : le contact, le déplacement, la distance, la relation, la succession) et *contre* (4 sémantismes : l'accumulation, le contact, l'échange, l'opposition).
- En revanche, *De … en* est monosémique et indique le déplacement.
- Chaque structure a un sémantisme nettement dominant, intrinsèque, sauf *à* qui cumule le *Contact* et la *Succession*, *contre,* associé au *Contact* et à l'*Opposition* et *sur,* associé à *l'Accumulation* et la *Succession*.

Structure	à	après	contre	d(e) … à	d(e) … en	par	pour	sur
Accumulation	1,65%	9,24%	0,66%			4,95%		83,50%
Contact	88,05%		9,35%	0,10%				2,51%
Déplacement				0,59%	99,41%			
Distance	1,23%			7,41%			91,36%	
Échange			3,23%			1,76%	95,01%	
Opposition	77,38%		22,62%					
Relation	8,36%			91,64%				
Succession	39,37%	8,67%		0,20%		37,54%		14,22%

Figure 5 – Par quelles structures sont réalisés les sémantismes ?

Ce tableau présente les différentes réalisations concrètes autrement dit les structures associées à un même sémantisme.

- Les sémantismes qui se réalisent au moyen des structures les plus variées sont l'*Accumulation* et la *Succession* qui sont toutefois sémantiquement proches, car ils sont liés en particulier par l'idée de succession chronologique qu'ils impliquent tous deux : il y a polyvalence des structures, à des degrés divers.
- Chaque sémantisme a une structure privilégiée, sauf la *Succession*, qui présente des affinités, à parts à peu près égales, avec les structures articulées autour de *à* ou de *par* ; la primauté d'une structure est moins nette pour l'*Opposition*. Par exemple, le sémantisme du contact privilégie les structures avec *à* ; ceux de la distance et de l'échange, les structures avec *pour*.

La synonymie des structures est vérifiée au travers des résultats statistiques, mais elle est plutôt limitée, si on prend en considération les sémantismes dominants associés à une structure, et ne porte de manière significative que sur deux paires de structures, comme le montre l'analyse arborée, figure 6.

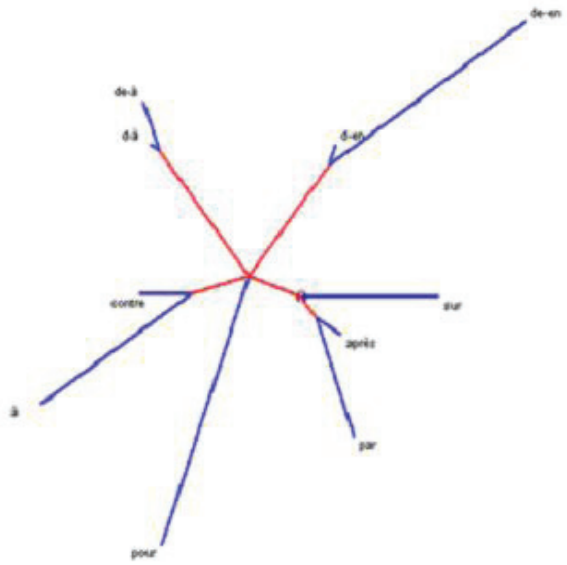


Figure 6 - Les structures évaluées selon les sémantismes associés

L'analyse arborée manifeste le rapprochement sur une même branche des prépositions *par* et *après* (succession), *à* et *contre (contact)*. On peut en déduire que les expressions N *par* N et N *après* N d'une part et N *à* N et N *contre* N d'autre part se rejoignent par un sémantisme commun.

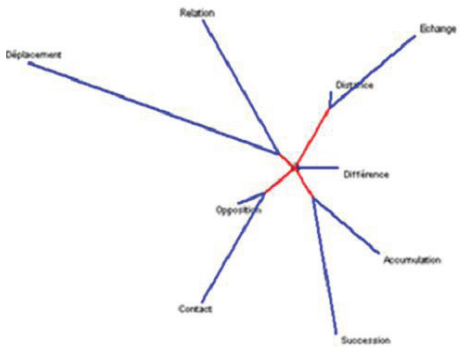


Figure 7 - Les sémantismes évalués selon les structures associées

On peut faire deux observations : la polysémie, partant la polyvalence partielle des structures d'un côté, la tendance à la spécialisation de l'autre. L'étude des structures NpN peut donc être envisagée soit à partir des 8 structures, soit à partir des 8 sémantismes.

La fréquence et la productivité des structures sont deux paramètres qui règlent l'appréciation de leur rendement.

2. Fréquence et productivité des unités

2.1. Les résultats contrastifs et relatifs

Pour évaluer la productivité des structures NpN, on peut calculer la proportion des occurrences des structures NpN sur les effectifs totaux de chaque préposition dans le corpus (*Mémoires, Roman, Poésie, Théâtre* de 1830 à nos jours). De ces proportions, trois observations peuvent être tirées : les prépositions les plus productives en NpN - dont on retient les effectifs - sont *contre* et *en* ; la préposition *à* est si fréquente et présente une telle polyvalence qu'elle produit moins de structures NpN ; enfin la productivité d'une préposition n'est pas identique selon les genres. On reviendra sur ce point en dernière partie de l'étude.

On peut aussi établir une comparaison entre la fréquence des prépositions et le nombre d'expressions différentes produites par les structures NpN. On calcule alors d'abord la proportion des effectifs de chaque préposition sur le total des effectifs des 7 prépositions, ensuite la proportion des différentes expressions générées par chaque préposition sur le nombre total des expressions.

Prépositions			Structures		
à	1341642	40,46%	à / de … à	347	15,37%
après	75393	2,27%	après	122	5,41%
contre	44305	1,34%	contre	114	5,05%
en	772517	23,29%	de … en	1083	47,98%
par	295617	8,91%	par	220	9,75%
pour	437191	13,18%	pour	137	6,07%
sur	349623	10,54%	sur	234	10,37%
Total	3316288			2257	

Figure 8 - Les proportions – tableau préparatoire au graphique de la figure 9

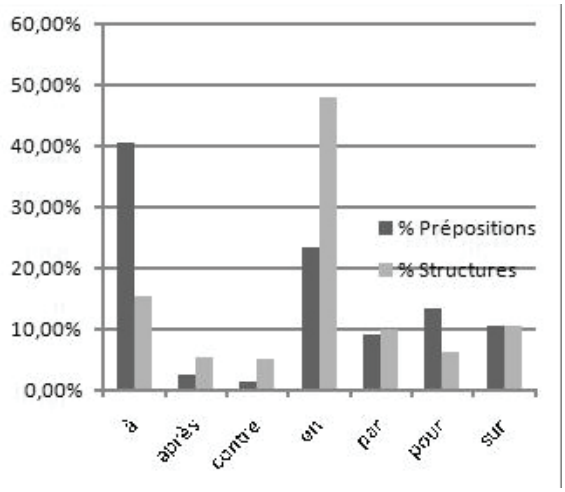


Figure 9 - La productivité des structures

À et *en* sont les deux prépositions les plus fréquentes dans le corpus, mais leur productivité est inversement proportionnelle : *à* génère beaucoup moins d'expressions différentes que *en* dans la structure *de … en*. Comme pour *à*, les autres écarts sont le signe d'une spécialisation partielle des structures qui confine à la lexicalisation ou au figement : *œil pour œil*, *corps à corps*, ou, à l'inverse, d'une productivité : *après, contre*.

La structure choisie se situe entre langue et discours. La part doit être faite en effet entre les lexies ou unités phraséologiques introduites dans le discours et les créations de séquences ternaires sur un patron linguistique préconstruit. L'analyse croise des structures qu'on pourrait dire figées puisqu'elles sont constituées en langue, comme syntagmes adverbiaux. La lexicalisation de la structure se vérifie notamment lorsqu'elle repose sur un transfert normalisé du sens propre au sens figuré :

(1) Ce n'est pas que Combeferre ne fût capable de combattre, il ne refusait pas de prendre *corps à corps* l'obstacle et de l'attaquer de vive force et par explosion. (V. Hugo, *Les Misérables*)

Dans d'autres exemples, le jeu sur l'emploi du sens concret ou figuré de la structure procède au défigement discursif de la structure, susceptible de créer un effet.

(2) Tout le monde descend l'escalier ; toi, tu déchireras tes draps de lit, tu en feras, *brin à brin*, une corde, puis tu passeras par ta fenêtre, et tu te suspendras à ce fil sur un abîme. (V. Hugo, *Les Misérables*)

(3) Aux quatre vents, quelle tempête s'amasse sur le lac du genre humain ! N'est-ce pas la création sans foi qui se détache *brin à brin* des mains du créateur, et tombe dans l'abîme, comme le chapelet d'un prêtre d'Arménie tombe à ses pieds. (É. Quinet, *Ahasverus*)

La contextualisation de la structure permet d'appréhender celle-ci : elle est susceptible d'être envisagée comme lexie quand l'emploi métaphorique est neutralisé par l'usage comme dans l'exemple (2). Sa place entre langue et discours est cependant préservée dans tous les cas, car le contexte est toujours à même de remotiver une figure lexicalisée.

2.2. La création lexicale

La création de nouvelles expressions est un fait de discours susceptible d'être interprété comme fait de style. La création peut se réaliser par analogie à partir de syntagmes figés en langue. Par exemple, si *d'égal à égal* est identifié comme locution adverbiale par le *TLF*, *groin à groin* est attesté en discours, mais n'apparaît pas dans l'article *groin* du dictionnaire. Le sens de l'expression se comprend par référence à d'autres expressions structurellement équivalentes et usuelles avec lesquelles elle forme un paradigme et par sa contextualisation ; mais l'expression n'est pas grammaticalisée. La fréquence des hapax dans un corpus donné peut être proposée comme indice de la productivité des structures.

Il est difficile d'établir facilement et avec certitude quelles suites existent en langue et lesquelles sont des créations ; un critère statistique robuste peut être choisi comme préalable :

– les structures les plus fréquentes atteignent un degré de figement élevé jusqu'à se grammaticaliser pour certaines et à fonctionner le plus souvent comme locutions adverbiales ;
– inversement, moins une expression est attestée, plus elle échappe à la lexicalisation.

De ce postulat, on déduira que les hapax peuvent fonctionner comme des indices de productivité ; pour une structure, compter le nombre d'expressions de fréquence 1 dans le corpus établit une nouvelle mesure entre le nombre d'hapax et le total des expressions de la structure. Le décompte se porte sur le « vocabulaire » pour déterminer la proportion des hapax sur le nombre d'expressions différentes, autrement dit pour calculer combien de structures sont hapax et combien ne le sont pas.

On pourrait supposer que la proportion des expressions qui sont hapax est directement liée à la fréquence des structures : plus une structure est fréquente, plus le nombre de ses occurrences qui sont des hapax est important. C'est un fait avéré, mais avec les exceptions significatives qui indiquent une plus grande capacité à produire des hapax pour une expression et le phénomène inverse pour une autre. Il s'avère que la structure avec *sur* produit plus d'hapax que ne le laisserait prévoir sa fréquence relative, confrontée aux autres structures. La structure avec *à* manifeste un déficit en hapax ce qui signifie qu'elle entre davantage dans des structures figées.

On peut coupler cette observation des hapax avec une distribution par genres pour avancer sur l'idée proposée au départ que la structure NpN est un possible marqueur générique.

Les paramètres observés jusqu'à présent, sémantisme et productivité, peuvent être utilisés pour une approche de la distribution des structures NpN selon les genres.

3. Les structures et les genres

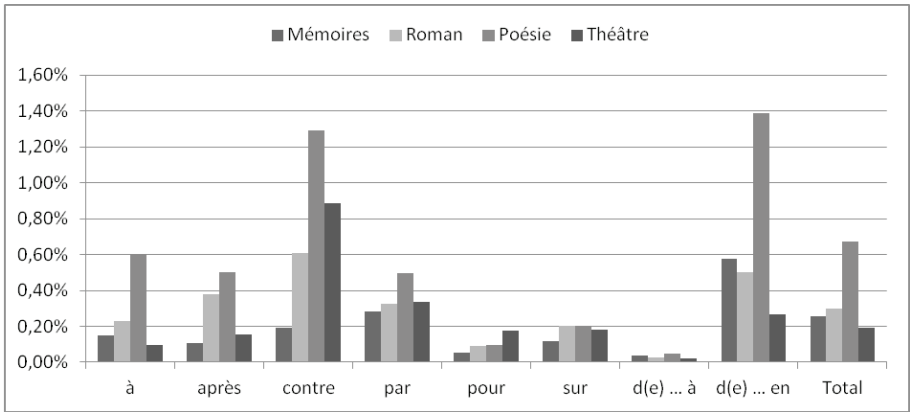


Figure 10 – Distribution des structures selon les genres

La figure 10 montre que la poésie est le genre qui mobilise la plus grande partie des effectifs de ses prépositions pour des structures NpN : cette affirmation est vraie pour *à*, *après*, *contre*, *par* et *en*.

Ceci est donc un premier indice d'une « fonction » générique de la structure NpN comme possible motif : il est plus spécifique à la poésie.

L'analyse factorielle permet la visualisation à la fois des structures et des genres : on voit ainsi sur le graphique 11 que certains genres affichent des affinités avec certaines structures en particulier.

Autour du roman gravitent les structures *après, sur* et *contre*, tandis que la poésie attire *de…en*, les mémoires *de…à* et le théâtre *pour* et *par.*

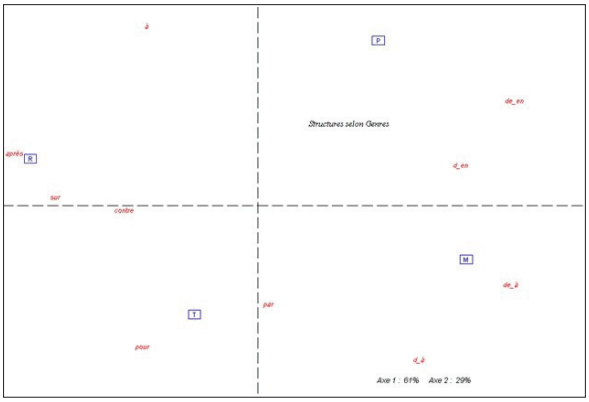


Figure 11 – Les structures selon les genres

L'observation peut être précisée en détaillant les données et en distinguant hapax et non-hapax (graphique 9) :

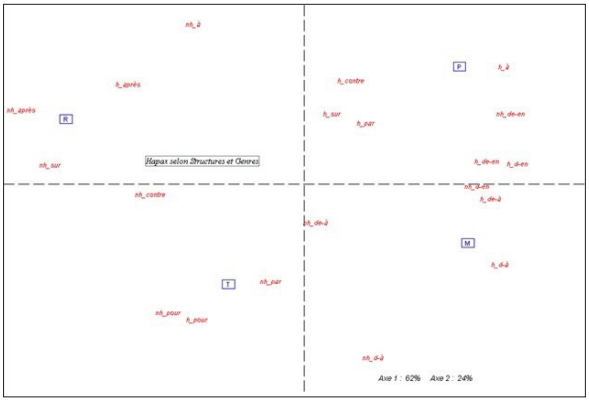


Figure 12 – Hapax et non-hapax selon les genres

On voit que la plupart des structures, dans les occurrences d'expressions hapax et non-hapax, ont le même comportement par rapport aux genres, à l'exception de la poésie, qui attire les occurrences hapax de plusieurs structures : *contre, sur, par, à,* et *de … en*.

Ceci constitue un deuxième indice de l'affinité d'un genre parmi les autres pour le motif NpN : la poésie est le genre le plus propice à la création d'expressions nouvelles à partir des structures *contre, sur, par, à et de…en*.

L'affinité des genres et des sémantismes est représentée dans le graphique 13.

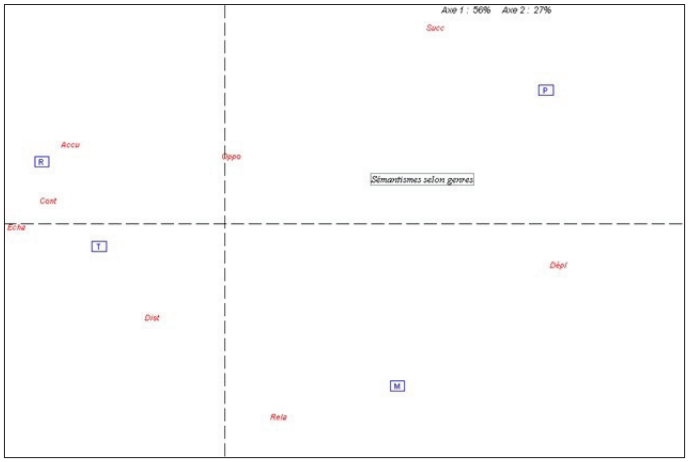


Figure 13 – Les sémantismes selon les genres

Cette AFC permet de visualiser les sémantismes privilégiés par chaque préposition. On peut supposer que le roman est caractérisé par le sémantisme du *Contact* en raison de son affinité avec la structure *contre*. La poésie se rapproche de la *Succession* – sémantisme réalisé par le plus grand nombre de structures sans qu'aucune n'affirme sa prédominance. C'est bien un sémantisme propre à la poésie.

4. Conclusion

Cette analyse sur les structures NpN a permis d'amorcer une réflexion sur le statut des prépositions et leur rôle structurel dans l'unité lexicale et sémantique étudiée.

Un processus interactif de constitution du sens de l'unité a pu être démontré. Le sens de la préposition est modulé en fonction des noms environnants ; les structures sont polysémiques et, symétriquement, un même sémantisme peut être exprimé par plusieurs structures.

Toutefois, la plupart des structures ont un sémantisme dominant, affirmant une spécialisation d'emploi. Chaque sémantisme a une structure privilégiée.

La productivité des structures a été évaluée à l'aune des hapax qu'elles produisent. *À* accuse un déficit en hapax, entrant dans la constitution d'expressions figées tandis que *sur* entre dans le plus grand nombre d'expressions nouvelles. *Sur* est plus productif quant à l'innovation lexicale.

La suite NpN peut être proposée comme un possible marqueur générique puisqu'a été observée une distribution nette entre les genres. La poésie, en particulier, se distingue par les traits suivants : c'est le genre qui mobilise le plus grand nombre de ses prépositions dans des structures NpN ; c'est le genre le plus propice à la création d'expressions nouvelles à partir des structures *contre*, *sur*, *par*, *à*, *de…en* et se trouve associée au sémantisme de la succession. Pour poursuivre l'analyse de ces suites NpN, il serait opportun d'établir un autre corpus non littéraire afin d'observer le comportement de la suite NpN dans l'un et l'autre ensemble et de voir si la suite peut fonctionner comme indice de littérarité[6].

Université. Nice Sophia Antipolis
CNRS, BCL, UMR 7320 Véronique MAGRI-MOURGUES

Bibliographie

D. Longrée, X. Luong, S. Mellet, 2008. « Les motifs : un outil pour la caractérisation topologique des textes », 9e JADT, p. 733-744. ‹http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf›.

A. Renouf et J. Sinclair, 1999. *English corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, Chapter Collocational Framaeworks in English, 128-143.

[6] Idée intéressante suggérée par David Trotter, lors de la présentation orale de ce travail.

Dictionnaire du patois de Bagnes (DPB)

Le but de cette communication[1] est de présenter :

- le projet lexicographique *Dictionnaire du patois de Bagnes* (*Valais*) : un projet né des longues relations entre le Glossaire des patois de la Suisse romande et la population patoisante de la commune de Bagnes, dont le produit final prendra la forme d'un dictionnaire illustré, assorti d'un index inverse français - patois et d'une banque de données sonore ;
- la base de données relationnelle FileMaker Pro conçue et développée dans son cadre, qui constitue l'un de ses sous-produits et sert actuellement de plateforme de rédaction.

1. Le projet lexicographique

1.1. Pertinence

Le projet lexicographique *Dictionnaire du patois de Bagnes* (*Valais*)[2] a pour objectif de décrire le lexique du patois de Bagnes, une variété du francoprovençal valaisan de type savoyard[3].

Le patois était, il y a un siècle, la langue première de l'ensemble de la population de la commune de Bagnes. Le français y était essentiellement limité à sa fonction de langue de l'écrit, de la communication suprarégionale et de la scolarité. En 1880, Jules Gilliéron notait Gilliéron (1881, p. III) :

> Le français, qui a fait de rapides progrès dans les cantons de Genève, de Vaud, de Neuchâtel et de Berne, et qui y a refoulé en partie les dialectes locaux, peut donc être encore regardé comme une langue étrangère dans le Valais. Sion, seul, le chef-lieu du canton, l'a adopté partiellement comme langue courante ; partout ailleurs, le patois règne encore comme unique langue parlée.

---

[1] Merci beaucoup à Jean-Claude Bliss, Eric Flückiger et Gisèle Pannatier pour leur relecture attentive de ce texte.

[2] Mandat de la Société des patoisants de Bagnes. Garantie financière de la Commune de Bagnes. Conception scientifique et informatique: Raphaël Maître. Équipe actuelle : rédaction : Eric Flückiger, Raphaël Maître, Gisèle Pannatier ; développement de la base de données : Jean-Claude Bliss, Raphaël Maître ; iconographie et enregistrement sonore : Société des patoisants de Bagnes. Coordination : Jean-Pierre Deslarzes, Raphaël Maître. Partenaires institutionnels : Glossaire des patois de la Suisse romande et Centre de dialectologie de l'Université de Neuchâtel ; Société des patoisants de Bagnes. Comité scientifique : Hervé Chevalley, Andres Kristol. Voir www2.unine.ch/dialectologie/page-8179.html.

[3] Pour la distinction entre les types savoyard et épiscopal du francoprovençal valaisan, cf. Jeanjaquet (1931).

Une génération plus tard, le français s'était introduit dans le quotidien des familles bagnardes Casanova (1971, 208) :

> En 1910 déjà, la plus grande partie des parents enseignent d'abord le français à leurs enfants. Cela n'exclut pas cependant qu'ils leur parlent patois par la suite, dans les villages supérieurs surtout et aux garçons en particulier.

Devenu marginal dans les usages réels, il est aujourd'hui la langue maternelle d'un petit pourcentage des habitants du district d'Entremont[4]. Tous les patoisants sont aussi francophones, et les occasions de communiquer en patois se raréfient. Malgré cela, ou, au contraire, à la faveur du sentiment de perte découlant de cette évolution, le parler de la communauté est investi dans les représentations collectives d'une valeur identitaire à certains égards plus forte que le français : langue des racines, il a valeur d'emblème d'une civilisation millénaire et de l'identité locale. En tant que langue de proximité et de l'affectivité, il est vécu comme irremplaçable par ses locuteurs natifs. Par son lexique – senti comme le plus adéquat pour l'expression du relief, des lieux, des phénomènes atmosphériques locaux – et par les noms de lieux, il enracine la population dans son environnement. Une étude sur la situation linguistique d'une autre commune valaisanne, Évolène, a pu mettre en évidence que la majorité des résidents, même non locuteurs et même non originaires du lieu, identifient le patois comme un repère identitaire crucial[5]. À Bagnes, l'investissement pour la mise en valeur du patrimoine, et du patois en particulier, est important, comme en témoigne, entre autres, la publication de l'ouvrage *Les noms de lieux de la commune de Bagnes. Toponymie illustrée* (Fellay / Dumoulin / Deslarzes 2000 ; ci-après *Toponymie illustrée*)[6], réalisé par la Société des patoisants et publié grâce au soutien de la Commune – déjà. C'est dans ce mouvement et ce contexte culturel que s'incrit la réalisation de notre dictionnaire.

1.2. Matériaux

L'exceptionnel corpus lexicographique bagnard résulte, pour l'essentiel, d'une collaboration de longue durée entre les lexicographes du Glossaire des patois de la Suisse romande[7] et les locuteurs natifs du patois de Bagnes. Dès sa fondation en 1899, le Glossaire lance, en vue de l'élaboration du Glossaire, une vaste enquête lexicologique par correspondance sur tout le territoire romand[8]. Il jouit, pour la seule mais grande commune de Bagnes en Valais (plus de 4'000 habitants en 1900, disséminés sur 5 km^2

---

4 Ce pourcentage s'élevait à 6,6% (662 personnes) en 1990, selon le recensement fédéral, et représentait environ six fois la moyenne suisse romande. Cf. Kristol (1998), Maître (2003), Pannatier (2002).

5 Cf. Maître / Matthey (2003).

6 Cf. RAGPSR n° 101-102 (1999-2000, 36).

7 Ci-après : Glossaire. Voir www.gpsr.ch.

8 Cf. BLGPSR, n° 1692.

de territoire habitable[9]), de l'engagement de trois correspondants patoisants[10]. En une décennie, ceux-ci remplissent, en plus ou moins grande partie, 229 questionnaires thématiques, lexicaux ou grammaticaux. Leurs réponses sont consignées sur des fiches de papier qui comportent chacune, typiquement, un mot patois, son sens, un exemple illustratif et sa traduction. Les correspondants appliquent les instructions qu'ils ont reçues pour la transcription phonétique du patois, ou adoptent leurs propres règles, pour graphier l'image acoustique des mots de leur langue maternelle. Au moment de l'enquête, le patois de Bagnes déploie pleinement son lexique dans tous les domaines de la vie quotidienne : habitat, culture, élevage, météorologie, psychologie, organisation sociale, spiritualité, etc., autant de domaines investigués méthodiquement par les questionnaires thématiques du Glossaire. Les résultats de l'enquête du Glossaire reflètent de manière extraordinairement précise et nuancée le dialecte de cette commune, tel qu'il était en usage au début du vingtième siècle parmi les habitants, jeunes et vieux, de ses villages.

Cette abondante récolte est complétée par de riches contributions de la même époque, dont un lexique complet (environ 7'600 fiches) de Louis Courthion[11] et de nombreux travaux de Maurice Gabbud[12]. À eux deux, ces derniers ont livré trois quarts des matériaux bagnards. S'y ajoutent encore un nombre appréciable d'études dialectologiques (l'enquête pionnière : Cornu [1877] ; la description grammaticale la plus détaillée : Bjerrome [1957]) et de dépouillements divers. Au bout du compte, la part bagnarde du fichier central du Glossaire constitue un sous-ensemble qu'on peut estimer à une cinquantaine de milliers de fiches. Elle est traitée, au fil des fascicules, dans les articles du GPSR, conformément à sa vocation première.

1.3. Le lancement

Maurice Casanova, natif de Bagnes et rédacteur au Glossaire de 1968 à 1990, est confronté jour après jour à l'extrême richesse des matériaux du fichier central en provenance de sa commune. Encouragé par la Société des patoisants de Bagnes, il a le désir de leur donner un traitement lexicographique particulier, et, en 1981, la Commission philologique du Glossaire accepte que ceux-ci puissent être réunis en vue d'être publiés en un bloc[13]. C'est ainsi que les 50'000 fiches bagnardes disséminées dans les 1'300 boîtes du fichier central du Glossaire sont photocopiées une à une par Josée Casanova, épouse de Maurice, mandatée par la Société des patoisants, puis classées par ordre alphabétique dans un nouveau fichier de 36 boîtes (voir fig. 1). Ensemble elles forment le Fichier DPB, le matériau brut de notre projet.

---

9 Source : <www.bagnes.ch>.

10 Maurice Charvoz, du Châble (cf. BLGPSR, n° 1786) ; Maurice Gabbud, de Lourtier (cf. BLGPSR, n° 1787) ; Maurice-Auguste Perraudin, de Lourtier également (cf. BLGPSR, n° 1789).

11 Cf. BLGPSR, n° 1792.

12 Cf. BLGPSR, n° 1211, 1507, 1787, etc.

13 Cf. RAGPSR n° 83 (1981, 7).

Fig. 1. Le fichier DPB (36 boîtes), déposé au deuxième étage du Glossaire des patois de la Suisse romande.

Mais Maurice Casanova décède en 1995, avant que le projet ne démarre.

Il faut attendre le début des années 2000 pour que le projet soit réanimé, et que la Société des patoisants numérote les boîtes et les paquets (environ 10’000 paquets ; voir fig. 2).

Fig. 2. La boîte n° 2, ouverte.

Les travaux rédactionnels commencent en 2005. Le travail est réparti entre deux équipes : les dialectologues se chargent de la rédaction ; les locuteurs natifs (membres de la Société des patoisants) préparent l'iconographie et l'enregistrement sonore des en-têtes, tout en relisant les articles rédigés.

La graphie de la *Toponymie illustrée*, élaborée par les auteurs de cet ouvrage avec Maurice Casanova[14], est reprise et rendue phonologiquement univoque. À cette fin, des tests phonologiques complets sont menés avec les locuteurs, leurs résultats sont confrontés aux descriptions existantes et, surtout, mis en adéquation autant que possible avec les codes et les habitudes graphiques de nos différentes sources. La graphie bagnarde présente deux qualités déterminantes : elle est lisible pour toute personne intéressée, sans aucune formation préalable, car, faisant largement recours à l'orthographe du français, elle est intuitive ; et elle répond aux besoins de la démarche scientifique. En outre, elle est déjà connue d'une partie du public intéressé depuis la parution de la *Toponymie illustrée.*

1.4. Du fichier DPB au dictionnaire : options méthodologiques

Au début du projet, deux méthodes rédactionnelles opposées nous semblaient réalistes : la rédaction paquet par paquet, ou la rédaction en base de données FileMaker Pro[15]. Une comparaison des avantages et inconvénients de chacune nous a amenés à privilégier la seconde[16].

1.4.1. La rédaction paquet par paquet

La méthode de rédaction la plus simple aurait consisté à saisir un paquet – le premier de la première boîte – et à rédiger son contenu à l'aide d'un logiciel de traitement de texte, puis de traiter de même le paquet n° 2, et ainsi de suite, alphabétiquement, jusqu'au dernier paquet de la dernière boîte.

Cette méthode paraissait linéaire, facile à planifier et peu gourmande en investissement, mais ses inconvénients l'ont disqualifiée. Citons-en deux :

– la surexposition aux occurrences surprises, c'est-à-dire à l'apparition d'un mot dans des exemples cités sous d'autres mots : si ces occurrences échappent à l'attention du rédacteur, le dictionnaire est lacunaire ; si le rédacteur les repère, il sera tenté ou forcé de rouvrir l'article déjà rédigé pour l'amender ;
– l'isolement des dossiers de mots : comment accéder, par exemple, dans un fichier consistant en un simple empilement alphabétique de dossiers de mots bruts, dénué de tout accès aléatoire (non alphabétique) aux fiches et de tout véritable système de renvois, aux éventuels dérivés

[14] Cf. Fellay / Dumoulin / Deslarzes (2000, 19-22).

[15] Voir ‹www.filemaker.fr›.

[16] En réalité, nous n'avons pas correctement évalué (ni n'étions en mesure de le faire, vu le caractère innovant de cette méthode) la masse de travail nécessaire pour relever les défis posés par le développement de notre base de données lexicographique.

et composés d'un mot, etc., si ce n'est en remontant – parfois non sans peine – aux paquets sources du fichier central du Glossaire ?[17]

1.4.2. La rédaction en base de données

Nous avons donc décidé, en 2005, de créer une base de données relationnelle, dans laquelle nous avons d'abord saisi tous les matériaux en les phonologisant. Une base relationnelle relativement statique suffisait pour cette opération de stockage. 20'000 exemples ont été introduits, et plus de 100'000 liens les relient aux lemmes qui les composent. Si bien que chaque exemple est disponible sous cinq lemmes en moyenne, au lieu de ne l'être que sous un seul – celui pour l'illustration duquel il a été produit[18].

Puis la base est devenue de plus en plus dynamique, pour que toutes les tâches rédactionnelles puissent être réalisées en elle au moyen de scripts (plus de 450 scripts actuellement) exécutés par l'intermédiaire de boutons (dont le dernier lance l'exportation de l'article). La base est encore en cours de développement et le sera sans doute jusqu'à la fin du projet, tant elle colle aux données qu'elle stocke et gère, et tant le développement de la base est un travail lexicographique au même degré que l'est la rédaction des articles.

3. Structure du dictionnaire

3.1. Macrostructure

La nomenclature réunit appellatifs et noms propres en une série alphabétique unique. Elle est structurée en deux niveaux : celui des lemmes simples, qui forment son ossature ; celui des locutions grammaticales, qui leur sont subordonnées. Le corpus compte, dans son état actuel, environ 12'000 lemmes simples et 750 locutions.

[17] Au contraire du fichier DPB, le fichier central du Glossaire est une gigantesque base de données manuelle (doublée depuis 1998 d'une base de données numérique), qui, grâce à ses chemins d'accès multiples (étymologique, lexical, morphologique, phonétique, sémantique, géographique et historique, entre autres) et à ses systèmes de renvois (internes et externes) perfectionnés, fournit au rédacteur un profil très complet du mot traité et une large vue d'ensemble de son réseau linguistique.

[18] On constate aujourd'hui que les exemples sont cités dans un tiers des cas sous le lemme pour l'illustration duquel ils ont été produits ; il le sont approximativement autant de fois sous d'autres lemmes. À ce jour, où 10'000 exemples sont cités, 8'000 le sont une seule fois ; 1'750 le sont 2 fois ; 200 le sont 3 fois ; 20 le sont 4 fois ; et 2 le sont 5 fois.

3.2. Microstructure

Les articles se subdivisent en quatre parties principales :

3.2.1. Paragraphe formel

Le paragraphe formel est introduit par le lemme. À sa suite sont présentées :

- les variantes phonologiques ;
- les formes fléchies : féminin des noms et adjectifs, troisième personne de l'indicatif présent et participe passé des verbes ;
- parfois, les variantes de sandhi.

Une forme attestée par une seule source est accompagnée de l'indication de cette source. Une forme signalée comme nouvelle, désuète ou peu usitée est accompagnée de sa marque d'usage.

3.2.2. Rubrique étymologique

Chaque mot est expliqué en une formule succincte, dans la mesure du possible.

- Un dérivé est rattaché à son mot-base. Pour les formations calquées sur le français, on donne le mot-base patois ainsi que la formation française source.
- Pour un emprunt (du français ou d'autres langues), on indique la langue d'origine et le mot source.
- Pour un mot autochtone d'origine latine, on donne l'étymon ; s'il est d'origine autre, on donne seulement la langue.

3.2.3. Structure sémantique

Les sens du mot sont ordonnés selon une numérotation hiérarchisée. En fonction des matériaux disponibles, chaque sens peut être complété par une description, une note encyclopédique, etc., et illustré par des syntagmes, exemples contextualisés, formules plus ou moins figées : proverbes, dictons, devinettes, expressions, etc.

3.2.4. Appareil de renvois

Un commentaire linguistique concis clôt l'article. Il comprend un nombre variable de rubriques, dont les plus fréquentes sont :

(a) Synonymie et antonymie.

(b) Références lexicographiques :

- celle des renvois aux articles de la tranche publiée du GPSR (laquelle s'étend actuellement de *a* à *gova*) ;
- celle des renvois, depuis les mots-bases, aux en-têtes FEW.

3.3. Articles illustratifs

Les articles ci-dessous illustrent ce qui précède.

*dichyon**i**ro*

{du français *dictionnaire*}

N. m. ◆ Dictionnaire. ◊ *Rad**â** on mo din o dichyon**i**ro*, consulter un mot dans le dictionnaire [GAB].

RÉFÉR. *GPSR* sous *dictionnaire*, *FEW* 3 sous *dictio*. (GP)

*distr**è*** [GAB] FÉM. *distr**è**ssa*

{de distr**è**rre}

Adj. ◆Qui manque d'attention. ◊ ***I**tre distr**è***, être distrait [GAB].

RÉFÉR. *GPSR* sous *distrait*. (GP)

*distr**è**rre* [GAB]

{du français *distraire*}

V. trans. 1° ◆Détourner de son objet l'attention. ◆ Réflexivement Se laisser gagner par l'inattention, être porté à la distraction. ◊ *I game**ï**n in k**ou**la son indyabl**ô** po së distr**è**rre*, les gamins à l'école ont un fort penchant à se distraire [GAB]. 2° ◆Faire oublier une peine, un souci, consoler. ◆ Réflexivement Se détendre pour oublier momentanément ses préoccupations. ◊ *Sai pâ avou**i** kô së distr**è**rre*, il ne savait pas avec qui se distraire [GAB].

SYNON. *dënoy**e***. RÉFÉR. *GPSR* sous *distraire*, *FEW* 3 sous *distrahere*. (GP)

*Farr**on**da*

N. pr. ◆ Nom donné à des vaches dont la robe noire ou rouge est tachetée de blanc sur les épaules et à la naissance de la queue. ◊ *A grôssa Ley**on** ë a dòy**in**ta Farr**on**da*, la grande *Ley**on*** et la petite *Farr**on**da* [GAB]. ◊ Fô lachy**e** agot**â** Farr**on**da, *plak**â** d'â-y ârry**â***, il faut laisser tarir *F.*, cessez de la traire [COU].

RÉFÉR. *GPSR* sous *farǫnda*. (EF)

*fëss**â***

{du latin FISCELLA}

N. f. ◆Grand moule en bois, en forme de boîte rectangulaire, dans lequel on met le sérac à l'alpage ; son fond est percé de petits trous pour l'égouttage et une de ses parois est amovible.

RÉFÉR. *GPSR* sous *faisselle*, FEW 3 sous *fiscella*. (EF)

*fœuf**è**le*

{du latin FALCICULA}

N. f. ◆Faucille. ◊ *Tal**in** da fœuf**è**le*, tranchant de la faucille [GAB]. ◊ *E fœuf**è**le di blô*, la faucille pour couper les blés, un peu plus grande que la faucille ordinaire [GAB]. ◊ *Inplëy**e** a fœuf**è**le*, employer la faucille [GAB]. ◊ *Kop**â** avoui a fœuf**è**le*, couper avec la faucille [GAB].

RÉFÉR. *GPSR* sous *faucille*, *FEW* 3 sous *falcīcula*. (EF)

*fouat**on*** [GAB] VAR. *folat**on***

{remonte au latin FOLLE}

N. m. 1° ◆Lutin, esprit follet auquel on attribuait naguère une foule de phénomènes. ◊ On croit que le *fouat**on*** suscite des tourbillons qui emportent le foin sec sur le point d'être rentrés, qu'il s'en prend aux animaux domestiques dont il fait gonfler les mamelles, qu'il a parfois une influence maléfique sur les personnes. À la Sainte-Agathe (5 février), chaque famille fait bénir à l'église quelques poignées de sel qu'on distribue aux bestiaux afin de leur épargner l'approche des *fouat**on*** et du démon. ◊ *É folat**on** no-z a to-t éparpëly**a** sé fin*, le *folat**on*** a éparpillé tout notre foin [COU]. 2° ◆AU FIG. 1. ◆Personne toujours pressée, qui ne tient pas en place. 2. ◆Sobriquet d'un homme qui croyait à l'existence du *fouat**on***. 3° ◆Tourbillon d'air.

RÉFÉR. *GPSR* sous *fòlatọn* 1, *FEW* 3 sous *follis*. (EF)

*intarv**â*** 3E SG. *int**è**rvë* P.P. *intarv**ô***

{du latin INTERROGARE}

V. trans. indir. ◆Interroger qqn, lui demander un renseignement, lui poser une question. ◊ *A rin intarv**ô** a nyon*, il n'a interrogé personne [GAB]. ◊ *I fô y intarv**â** se sâ dë nov**ë**lë*, il faut lui demander s'il a des nouvelles [GAB]. ◆Employé absolument et suivi d'une proposition interrogative. ◊ *Moùch**eu** Prëzid**an**, yoù vo dëm**an**do pâ se vo baly**e** a koùm**an** dë danshly**e** a kramintr**an**, yoù v**e**nyo intarv**â** tan kë kan n'in-n in o drai*, Monsieur le Président, je ne viens pas vous demander la permission de danser à carnaval, je viens demander jusqu'à quelle date nous en avons le droit [COU].

RÉFÉR. *FEW* 4 sous *ĭntĕrrŏgare*. (EF)

*int**è**rva*

{de *intarv**â***}

N. f. ◆ Demande de renseignements, d'information. ◊ *D'int**è**rva dë koùrey**œu***, de l'interrogation de curieux, réponse par laquelle on éconduit un questionneur jugé indiscret.◊ EXPR. PROV. ◊ *Pë-r int**è**rva on va a R**o**ma*, on va loin, on arrive partout à force de poser des questions, litt. par interrogation on va à Rome.

SYNON. *inf**ò**rma*. (EF)

*kamiy**on*** [CAS]

{du français *camion*}

N. m. ◆Camion. ◊ *Sa tsar**o**pa dë Bar**on** ai te pâ rapistol**ô** o mékan**i**ke du kamiy**on** at**o** on tro dë fi d'arts**ô***, cette canaille de Baron avait rafistolé les freins de son camion avec un bout de fil de fer [CAS].

RÉFÉR. *FEW* 23 sous « charrette ». (EF)

*kamyoùn**â**dzo* [GAB]

{de *kamiy**on***, d'après le français *camionnage*}

N. m. ◆(PEU USITÉ) Camionnage. (EF)

*kapoùn**â*** 3E SG. *kap**où**në* P.P. *kapoùn**ô***

{de *kap**on***}

V. intr. 1° ◆ Perdre courage, ne pas oser entreprendre ce qu'on s'était proposé de réaliser. ◊ *M'a fal**u** kapoùn**â***, il m'a fallu renoncer [GAB]. ◊ ***O**mo ! Py**a**rro, kap**où**na pâ* ! hardi ! Pierre, ne capitule pas ! [COU]. 2° ◆ Renoncer à accomplir quelque chose que l'on avait commencé et qui demande par trop d'efforts ou de peine. ◊ *S'toù vœu **i**tre on bon dy**â**blo, të fô pâ kapoùn**â***, si tu veux être un brave, il ne faut pas abandonner [GAB]. ◊ *Sâ d**a**ma ai vol**u** f**ï**re assinchy**on** du Konb**eïn**, meïn kan ë zu in**ô** pë o mait**in**, a kapoùn**ô***, cette dame avait voulu faire l'ascension du Grand Combin, mais à mi-chemin, elle y a renoncé [COU]. 3° ◆ Céder le pas (sur la route).

RÉFÉR. *GPSR* sous *caponner*, *FEW* 2 sous *capo*. (GP)

*n**eïn**të* [PER]

{de l'italien *niente* « rien »}

Pron. indéf. ◆ Rien. ◊ *In sâ n**eïn**të*, il n'en sait rien [PER]. (EF)

*soùss**an**ta* (devant voy. *soùss**an**ta-z* [PER] ; *soùss**an**t'* [PER])

{du bas-latin SEXANTE}

Adj. num. card. ◆ Soixante. Peut être coordonné par *ë* ou *y'* à un adj. num. card. compris entre *on* et *nœu*. ◊ *Soùss**an**t' ë dou*, soixante-deux [PER]. ◆ Pronominalement ◊ *N'in fi oùna gr**an**ta kondyu**in**te dë soùss**an**ta dë no po trén**â** a k**on**tse du born**é***, nous avons formé une longue chaîne de soixante personnes pour traîner le bassin de la fontaine [GAB].

RÉFÉR. *FEW* 11 sous *sĕxagĭnta*. (RM)

*trav**è***

{du français *travers*}

I. N. m. 1° ◆ Étendue, DANS : *Trav**è** dë tin*, période d'une certaine importance. ◊ *Ë ét**ô** mal**â**do on gran trav**è** dë tin*, il a été malade durant plusieurs mois au moins [COU]. 2° ◆ Travers, défaut. ◊ *A on krouè trav**è***, il a un vilain défaut [COU].

II. Élément de loc.

*in trav**è*** [GAB] VAR. *in tray**è*** Adv. ◆ En travers, en biais, transversalement. ◊ *A sâ kr**ò**ye abit**u**de dë së vreye in trav**è** u mét**in** du tsem**eïn***, il a la mauvaise habitude de se mettre en travers au milieu du chemin [GAB].

*dë trav**è*** VAR. *dë tray**è*** [BJE] Adv. 1° ◆ De travers, en biais. ◊ *V**ai**dë vo sé tsoùm**eïn** kë va in**ô** dë trav**è**?* voyez-vous ce chemin qui monte là en biais ? [COU]. 2° ◆ À rebours, à l'envers ; mal. ◊ *E fi to dë trav**è***, il fait tout à rebours [COU]. ◊ *To më va dë trav**è***, tout se passe mal, je suis dans la déveine [GAB]. 3° Adjectivement, DANS : *On rëg**â** dë trav**è***, un regard malveillant ou suspicieux.

*dë trav**è** dë* [COU] Prép. 1° ◆ Au travers de, à travers. ◊ *Keïn ë te kë va in**ô** dë trav**è** di tsan?* qui est la personne qui monte à travers champs ? [COU]. 2° ◆ Au rebours de, à l'opposé de. ◊ *E fi to dë trav**è** dë sin k'on i rëkoùm**an**dë*, il fait tout à l'opposé de ce qu'on lui recommande [COU].

*a trav**è** dë* [COU] Prép. ◆ À travers. ◊ *Y'ô-y é yu pass**â** **u**tre a trav**è** da tyœudr**ë**ya*, je l'ai vu passer à travers la coudraie [COU]. ◊ *Aprey**an**dë toù pâ dë pass**â** a trav**è** da gran g**o**le?* n'appréhendes-tu pas de traverser la mer ? [COU].

RÉFÉR. *FEW* 13/2 sous *transvĕrsus*. (EF)

5. Avancement du projet

À ce jour, le fichier manuscrit est entièrement phonologisé et saisi. Le corpus net qui en résulte réunit (dépouillements et renvois compris) de plus de 14'000 dossiers de mots, 50'000 fiches de mots (formes ou sens) et près de 25'000 exemples traduits. Environ 6'000 articles, qui couvrent la moitié de la nomenclature, sont rédigés, ont été relus par les partenaires patoisants et en partie corrigés. Les travaux rédactionnels s'achèveront fin 2015, et la publication est prévue en 2016.

Glossaire des patois de la Suisse romande /
Centre de dialectologie et d'étude du français régional,
Université de Neuchâtel (Suisse) Raphaël MAÎTRE

Références bibliographiques

Bjerrome, Gunnar, 1957. *Le patois de Bagnes (Valais),* Stockholm, Almqvist & Wiksell [= Romanica Gothoburgensia VI].

BLGPSR = Gauchat, Louis / Jeanjaquet, Jules, 1912 et 1920. *Bibliographie linguistique de la Suisse romande,* Neuchâtel, Attinger.

Casanova, Maurice, 1971. « Rapport de la communication de Mme Rose Claire Schüle : ' Comment meurt un patois ' », in : Marzys, Zygmunt / Voillat, François (ed.), *Actes du Colloque de dialectologie francoprovençale organisé par le Glossaire des patois de la Suisse romande, Neuchâtel, 23-27 septembre 1969*, Neuchâtel / Genève, Droz, 207-213.

Cornu, Jule, 1877. « Phonologie du Bagnard », *Romania* 6, 369-427.

Fellay, Willy / Dumoulin, Hilaire / Deslarzes, Jean-Pierre, 2000. *Les noms de lieux de la commune de Bagnes. Toponymie illustrée,* Bagnes, Commune.

Fluckiger, Eric, 2002. « Les enquêtes lexicologiques du Glossaire des patois de la Suisse romand », in : *Actes de la conférence annuelle sur l'activité scientifique du Centre d'études francoprovençales : lexicologie et lexicographie francoprovençales, Saint-Nicolas, 16-17 décembre 2000,* Aoste, BREL, 23-40.

Gillliéron, Jules, 1881. *Petit Atlas Phonétique du Valais Roman (sud du Rhône)*, Paris, Champion.

Jeanjaquet, Jules, 1931. « Les patois valaisans : caractères généraux et particularités », *RLiR* 7, 23-51.

Kristol, Andres, 1998. « Que reste-t-il des dialectes gallo-romans de Suisse romande ? », in : Eloy, Jean-Michel (ed.), *Évaluer la vitalité. Variétés d'oïl et autres langues. Actes du Colloque international 'Évaluer la vitalité des variétés régionales du domaine d'oïl ', Amiens, 29-30 novembre 1996*, Amiens, Université de Picardie, 101-114.

Maître, Raphaël, 2003. « La Suisse romande dilalique », *Vox Romanica* 62, 170-181.

Maître, Raphaël / Matthey, Marinette, 2003. « Le patois d'Évolène aujourd'hui … et demain ? », in : Boudreau, Annette / Dubois, Lise / Maurais, Jacques / McConnell, Grant (eds), *Colloque international sur l'Écologie des langues*, Paris, L'Harmattan [Sociolinguistique], 45-65.

Pannatier, Gisèle, 2002. « Les patois valaisans », *micRomania* 43, 3-9.

RAGPSR = Glossaire des patois de la Suisse romande, 1900-. *Rapport annuel,* Neuchâtel, Attinger puis Genève, Droz.

Sites internet

Commune de Bagnes : *Site officiel.* ‹www.bagnes.ch›.

Centre de dialectologie et d'étude du français régional de l'Université de Neuchâtel : *Dictionnaire du patois de Bagnes.* ‹www2.unine.ch/dialectologie/page-8179.html›.

Glossaire des patois de la Suisse romande (GPSR) : *Site officiel.* ‹www.gpsr.ch›.

FileMaker Pro : *Site officiel.* ‹www.filemaker.fr›.

L'évolution de Frantext : quelles modifications et quels usages pour un Frantext 2 ?

La renommée de la base de données textuelles Frantext (www.frantext.fr), développée et maintenue à l'ATILF, anciennement INALF, n'est plus à faire. En témoigne le nombre important d'utilisateurs, et ce malgré l'accès payant dû à des contraintes éditoriales imposées par les droits d'auteur. À peu près deux cent cinquante bibliothèques universitaires de par le monde, de l'Europe à l'Asie et aux Amériques, à la suite d'une souscription annuelle, fournissent à leurs usagers la possibilité d'accéder à ce corpus constitué d'œuvres écrites directement en français[1]. Ses utilisateurs les plus enthousiastes sont principalement des linguistes, des littéraires, des professionnels de la documentation et des journalistes.

Nous présenterons les transformations qui ont été envisagées pour la base dans le cadre d'un projet qui fut appelé Frantext 2 et qui a eu cours de 2009 jusqu'à présent. Le but était de proposer une version améliorée de Frantext, visant à la création d'un outil plus souple et plus performant, des voies nouvelles d'exploration et d'exploitation de corpus échantillonnés, constitués à partir de critères scientifiques. Tout d'abord, nous exposerons l'historique de Frantext, sa conception et son fonctionnement, pour pointer ensuite les objectifs et les usages attendus de Frantext 2, qui seraient exploitables en linguistique comme dans d'autres domaines.

1. Descriptif et historique de la base

Frantext actuel est un corpus d'à peu près 4500 œuvres en français, couvrant une période étendue de l'histoire des textes, du Moyen Âge au XXIe siècle. La plus ancienne est *La Chanson de Roland*, datée d'environ 1125, les plus récentes sont quatre textes publiés en 2012, dont *Une vie brève* de Michèle Audin. 1470 auteurs sont représentés. La base est associée à un moteur de recherche, Stella, qui permet

[1] Actuellement, la base comporte également des traductions de différentes langues, y compris du grec et du latin, vers le français, provenant de diverses époques (de la Renaissance, grande époque des traductions et des adaptations en langue vernaculaire, à la periode contemporaine). Le retrait de ces textes a été envisagé, mais pas finalisé, et ce malgré les erreurs d'attribution de certaines œuvres que cette pratique des traductions et des adaptations a pu entraîner. Ainsi, plusieurs œuvres traduites, et non seulement des traductions pratiquées à la Renaissance où la notion d'auteur était encore peu développée, ont longtemps figuré dans la base sous le nom du traducteur uniquement (par exemple, des œuvres de Faulkner, Calderon de la Barca ou Dino Buzzati ont été attribuées à Camus). Nous avons pu corriger certaines erreurs ; d'autres corrections sont en cours.

d'interroger un corpus sélectionné préalablement d'après certains critères (auteurs, titres ou encore périodes, genres littéraires[2]). Corpus à dominante littéraire, Frantext contient à peu près 90 % d'œuvres littéraires et 10 % d'ouvrages techniques. La moitié du corpus couvre la période 1900-2012. Ce double déséquilibre qui penche en faveur de la littérature et de la modernité, alors que la base est utilisée surtout par des linguistes qui s'intéressent à une variété de registres et d'états de langue, serait bien sûr à corriger à l'avenir, en élargissant les collaborations à des projets visant des corpus philosophiques, scientifiques ou techniques.

La base fut conçue initialement pour constituer un réservoir d'exemples et de contextes qui allaient alimenter le *Trésor de la langue française*. Si on peut situer vers 1957 la naissance du projet du *Trésor de la langue française*, l'origine de la base Frantext qui accompagne et constitue un support de ce vaste projet dictionnairique peut être située dans les années 1960-1970. Frantext est devenu une base de données accessible au public, via internet, à partir de 1998. À l'époque de sa création, dans les années 1970, Frantext était pionnier : l'idée de construire une base de données à partir d'un ensemble de textes ayant subi un traitement automatisé représentait une révolution dans le paysage de la culture numérique. L'intérêt permanent que présente Frantext et qui le distingue toujours d'autres bases consiste dans la combinaison d'un grand corpus et d'un moteur de recherche complexe.

Cette base permet de faire des recherches à différents niveaux[3] : retrouver une citation exacte et son auteur ; rechercher les occurrences d'un terme ou d'une expression dans un corpus d'œuvres sélectionnées ou sur tous les textes de la base ; calculer des fréquences d'usage ou des études de voisinage ; rechercher des listes de mots ; faire des tris de vocabulaire ; effectuer des recherches étendues sur la langue française : lemmes, expressions régulières... La base offre aux utilisateurs avertis la possibilité de constituer leurs propres 'grammaires' de recherche. La représentativité de certains auteurs, parmi les plus connus dans les lettres françaises, dans l'ensemble de la base pourrait certes être critiquée mais il faut rappeler que cet aspect est dû à la conception d'origine de Frantext, pensé comme un réservoir de textes des plus variés où des auteurs quasi anonymes avaient aussi leur place, textes destinés à fournir une grande diversité de définitions et d'exemples au *Trésor de la langue française*. Une autre contrainte concerne l'affichage des résultats. Conçu tout premièrement comme

[2] Les genres littéraires indiqués à ce jour dans la base proposent uniquement des points de repère généraux ; certains sont carrément erronés ou correspondent à un état ancien de l'étude des genres. L'un des objectifs de Frantext 2 allait dans le sens d'une identification juste du genre d'une œuvre donnée et d'une réflexion sur la question des genres littéraires, de leur entrecroisement et de leurs voisinages.

[3] Des exemples d'utilisation de la base (des 'grammaires' d'utilisation) ainsi qu'un document qui résume les différentes possibilités de recherche se trouvent sous la rubrique 'documentation' à l'adresse www.frantext.fr. Deux didacticiels sont disponibles sous la rubrique 'Ressources didactiques' et une aide en ligne est proposée au cours des différentes étapes. Un intéressant et pratique guide d'utilisation a été proposé également par la Bibliothèque universitaire d'Avignon, disponible à l'adresse <http://www.bu.univ-avignon.fr/uavg/Formation_publics/bases%20shs/frantext_depliants_pdf.pdf>.

un outil d'analyse textuelle, soumis aux contraintes d'affichage restreint de fragments textuels (moins de 300 mots par résultat), Frantext ne permet pas d'afficher une œuvre dans son intégralité et pour le moment il ne semble pas envisageable de l'ouvrir à un affichage plus large, malgré les efforts de l'équipe qui œuvre dans ce sens.

Une prise en compte accrue de la diachronie a été récemment proposée par l'introduction d'une flexion médiévale (qui permet une recherche des formes anciennes des mots de la langue française en s'appuyant sur les lemmes du *Dictionnaire du Moyen Français*, 1330-1500) et d'une flexion du XVIIe (en vue de recherches des formes des mots du XVIIe, en prenant appui sur les travaux du projet IMPACT[4], initié en partenariat avec la BnF, Paris).

Aujourd'hui encore, Frantext est le support de plusieurs projets. Il est utile d'évoquer ainsi la jonction entre Frantext et le *Dictionnaire du Moyen Français* (http://www.atilf.fr/dmf/) et leur enrichissement mutuel. De manière semblable, Frantext CTLF (Corpus de Textes Linguistiques Fondamentaux) est issu de la collaboration avec l'ENS Lyon (disponible en accès libre à l'adresse http://ctlf.ens-lyon.fr/). Frantext constitue de même une ressource d'avenir, sur laquelle s'appuient grandement de nouveaux projets, comme le projet Ressource Lexicale Informatisée d'Envergure sur le Français (RELIEF[5], développé à l'ATILF). Frantext sera intégré à l'équipex ORTOLANG, piloté également à l'ATILF. Cette base représente ainsi – elle l'a toujours été – une ressource transversale qui répond à de nouveaux besoins de constitution et d'exploitation de corpus et, pour cette même raison, elle doit évoluer à son tour.

2. Objectifs des métamorphoses de Frantext 2

La nécessité d'améliorer cette base, déjà singulière si on la compare à Gallica (qui propose une numérisation en mode image) ou à d'autres corpus francophones, paraît indéniable. Les modifications proposées en vue de réaliser la version Frantext 2 tiennent aux usages et aux objectifs qui se sont précisés, notamment au perfectionnement des corpus et des procédés d'interrogation par des critères plus nombreux et plus fins. En comparaison des grands corpus, à l'exemple de Google books, qui couvrent des périodes, des collections et des bibliothèques entières, Frantext peut être vu comme une petite bibliothèque dont le contenu est sélectionné soigneusement pour servir à des projets de recherche portant sur des corpus d'auteurs précis ou sur des problématiques théoriques[6]. L'enrichissement du corpus est constant et bien spectaculaire par rapport au rythme que Frantext avait connu auparavant (depuis peu, quelques centaines de nouveaux titres par an), grâce aux progrès de la numérisation

[4] La coordination de ce projet au sein de l'ATILF est assurée par Gilles Souvay, également responsable de l'interface de Frantext.

[5] RELIEF est un projet prioritaire de l'ATILF en lexicographie synchronique, officiellement lancé en juin 2011, dont la finalité est de développer une ressource lexicale – le Réseau Lexical du Français (RLF).

[6] Sous la responsabilité de Véronique Montémont, un large corpus de journaux personnels et d'autobiographies permet de faire des recherches sur cette thématique d'actualité.

et du balisage automatique. Les collaborations à des projets de linguistique historique ont permis de changer la répartition des œuvres par période de l'histoire, en augmentant le nombre de textes du Moyen Âge, des XVIe, XVIIe et XVIIIe siècles. En élargissant les catégories de textes à englober dans la base, selon des souhaits exprimés ponctuellement par certains utilisateurs intéressés par des phénomènes qui traduisent les préoccupations de la contemporanéité – comme des discours oraux, des discours électoraux, des discours de réception (des prix Nobel, à l'Académie…), des rapports des agences internationales de cotation économique –, Frantext sera susceptible de faciliter les recherches de spécialistes de domaines de plus en plus variés : histoire, anthropologie politique, droit, économie…

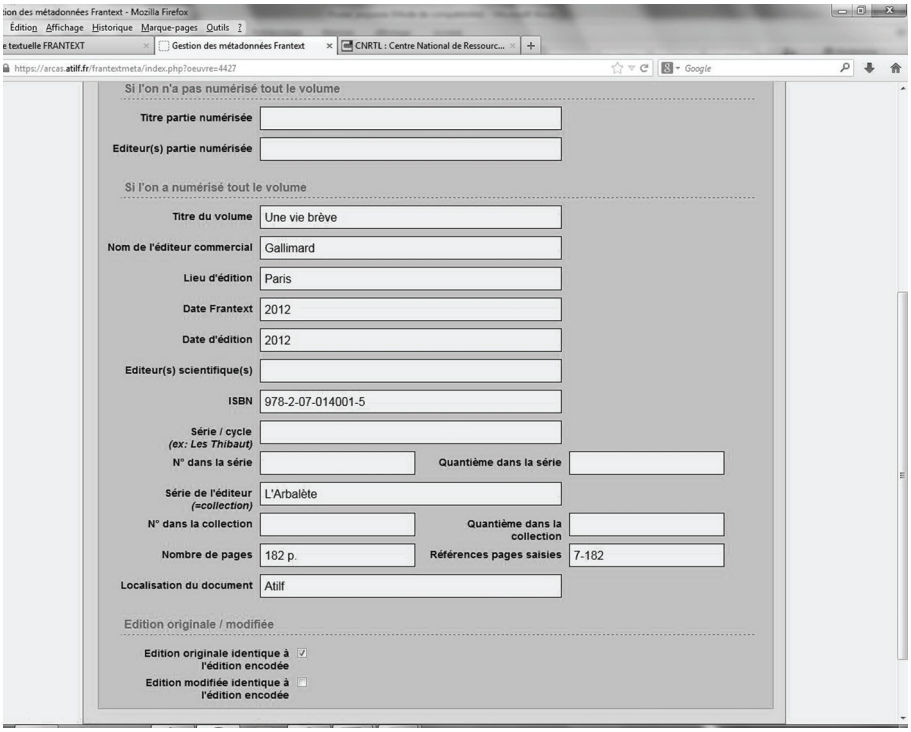
Un changement important concerne les métadonnées et l'encodage des textes. Tous les textes nouvellement entrés dans Frantext (et, dans la perspective de Frantext 2, tous les textes qui y sont déjà) sont encodés en XML-TEI, ce qui facilite l'interopérabilité avec d'autres ressources. En vue de l'un des principaux objectifs de Frantext 2, de fournir un supplément de renseignements sur les auteurs, les œuvres, les genres littéraires et le contenu des textes, il est prévu que les textes soient équipés de 'métadonnées' référencées et précises. Ces descripteurs plus nombreux et plus précis, nommés 'métadonnées', permettront d'obtenir des renseignements détaillés sur les textes interrogés. Pour une bonne partie des textes, notamment ceux qui ont rejoint la base depuis 2009, ces renseignements sont regroupés sur la plate-forme Frantext, avant-poste de la base, accessible pour l'heure uniquement au personnel qui renseigne les données et qui œuvre pour l'amélioration de la ressource[7].

Ces métadonnées permettront de retrouver des renseignements sur les éditions utilisées en vue de la numérisation (selon le choix du chercheur chargé du projet respectif), sur l'édition originale et l'édition définitive, sur les genres littéraires ainsi que sur les thématiques abordées par les textes. Grâce à Frantext 2, les utilisateurs pourront développer des recherches en fonction de critères multiples (simples ou combinés : dates biographiques de l'auteur, dates de parution des œuvres ; références bibliographiques complètes comme : éditeur scientifique, éditeur commercial, collection, prix, paratexte, publications anthumes ou posthumes, date de création pour le théâtre). La création d'un formulaire multicritères de constitution de corpus et d'une interface ergonomique à niveaux d'interrogation multiples représente l'un des objectifs principaux de Frantext 2. À l'avenir, la base sera également dotée d'un nouveau moteur de recherche et idéalement la base catégorisée en parties de discours sera agrandie.

L'image qui suit montre le travail qui est fait sur la plate-forme Frantext dans laquelle les interrogations possibles concernent les renseignements sur : l'édition

[7] Il s'agit d'Isabelle Clément, chargée de la numérisation et du balisage, de moi-même, qui ai renseigné les métadonnées, de Sylvain Briat, qui a développé d'un point de vue informatique la plate-forme sous la direction d'Etienne Petitjean, responsable du service informatique. La conception de la plate-forme et de ses nombreuses rubriques détaillées résulte, quant à elle, des quant à elle issue des propositions de l'équipe élargie de Frantext sous la direction de Véronique Montémont qui a piloté le projet.

numérisée ; l'édition modifiée ; l'édition originale, chacune se déclinant en : titre du volume ; éditeur commercial ; lieu d'édition ; date d'édition ; éditeur scientifique ; ISBN ; cycle ; collection ; nombre de pages ; paratexte (dédicace, épigraphe, préface, postface…). Tous ces renseignements sont réunis dans une base de métadonnées qui ressemble à un catalogue d'une grande précision et qui permettra d'accéder par la suite aux textes auxquels correspondent ces métadonnées. La présence de ces données, loin d'être de l'ordre de l'insignifiant, est une garantie de la qualité des résultats générés par la recherche. Un livre mal rangé dans une bibliothèque serait difficile sinon impossible à retrouver. Le même principe de netteté et de précision dans le classement et l'indexation guide l'utilisation et le renseignement des métadonnées.



L'optimisation de Frantext est la préoccupation permanente de ceux qui travaillent à son amélioration en termes de corpus, de procédures d'interrogation ou de système d'exploitation. Cette union de points de vue variés, allant de la précision du référencement des métadonnées à la diversification des corpus et à l'accessibilité d'utilisation, menée par des documentalistes, des techniciens en numérisation et balisage, des informaticiens et des chercheurs qui définissent des besoins d'interrogation et des corpus prioritaires, est censée répondre au besoin de perfectionnement de la base.

Les objectifs scientifiques sont multiples : tout premièrement, la création d'une ressource pluridisciplinaire, réalisable par l'enrichissement du corpus vers les domaines

sociologique, historique, juridique, médical et l'élargissement des exploitations autres que linguistiques (littéraire, stylistique, sémiotique, génétique littéraire). Il s'agira ensuite de la conservation des états diversifiés de la langue (en synchronie, diachronie, diatopie).

Frantext 2 sera utile dans l'établissement d'éditions critiques[8] ou philologiques, et le chercheur pourra citer en s'appuyant sur Frantext, vérifier le lexique, les cooccurrences… Se présentera aussi la possibilité de faire des interrogations sur les états de langue d'un même texte, en comparant plusieurs versions, éditées à des époques différentes. Il s'agit notamment d'auteurs pré-modernes. Plusieurs pièces de Corneille, comme *Le Cid*, *Médée*, *L'Illusion comique*, par exemple, sont disponibles dans la base dans plusieurs versions (respectivement celles de l'édition princeps, éditée à l'époque contemporaine dans le respect de l'orthographe et de la graphie originales ; de l'édition de maturité des *Œuvres* de 1660 ; de l'édition définitive des *Œuvres* de 1682, dernière édition publiée du vivant de Corneille, la plus censurée et la plus épurée), elles-mêmes éditées à des moments différents (dans la deuxième moitié du XIXe siècle quand la philologie et l'édition critique connaissent un intérêt croissant, dans la deuxième moitié du XXe siècle, par la « Société des Textes français modernes » ou encore dans des éditions courantes, modernisées à l'usage de tout un chacun). La comparaison des tribulations du texte d'une pièce facilitera l'établissement d'une nouvelle édition critique.

L'étude des diverses variétés du français sera abordable dans des corpus qui toucheront des communautés linguistiques extérieures à la France : Suisse, Belgique, Maghreb, Afrique subsaharienne, Québec, Antilles, en augmentant les corpus d'œuvres provenant de ces régions. Le corpus pourra être établi soit en cherchant une maison d'édition non métropolitaine, soit en interrogeant la fiche-auteur qui indiquera la nationalité ou les nationalités multiples, ou encore les changements de nationalité. Il serait en effet possible d'ouvrir à des problématiques de recherche nouvelles à partir de l'enrichissement des notations concernant les fiches-auteur : non seulement date de naissance et de décès pour situer chronologiquement l'écrivain, mais aussi lieux de naissance et de décès qui permettront aux sociologues, par exemple, d'étudier les déplacements (du centre vers les périphéries ou l'inverse, de la capitale vers la province ou l'inverse), les migrations, les exils, les légitimations culturelles et littéraires. L'étude de l'origine de l'auteur, pour aller au-delà de la notion proposée par Marthe Robert dans *Roman des origines et origines du roman*, renverra au multiculturalisme, au bilinguisme, à l'interculturel, aux métissages culturels, aux questions de légitimation de certains auteurs : auteurs-femmes, auteurs venus d'ailleurs, auteurs de la périphérie.

[8] Frantext a servi par exemple à l'établissement d'une édition critique d'un texte d'Ancien Régime : Claude Favre de Vaugelas, *Remarques sur la langue françoise*, ed. Zygmunt Marzys, Genève, Droz, 2009. Voir l'avant-propos de l'éditeur scientifique, p. 7, où il indique des qualités et des défauts de Frantext.

L'introduction de mots-clés issus d'une indexation fine des contenus fournira une ouverture vers le cœur du livre, ce qui permettra d'ajouter aux corpus des documents qui ne sont pas connus a priori. La diversité des mots-clés proposés ira du contenu détaillant des thèmes et des problématiques vers la forme, les techniques d'écriture (l'emploi de l'oralité, des dialogues, de l'intertextualité), de même que vers les genres et sous-genres littéraires. Frantext permettra ainsi l'étude des genres du discours et des genres littéraires (en interrogeant des genres particuliers : poèmes en prose, tragi-comédie, drame…) ou encore l'analyse des genres hybrides qui naissent au contact de plusieurs formes génériques. En partant du postulat que les structures textuelles varient en fonction de la généricité, la recherche d'informations sur les régularités et les différences des textes conduira à une typologie des genres textuels. Enfin, l'hyper-naviguer entre métadonnées et textes est souhaitable, chaque élément de la base étant lié aux autres et permettant l'accès à des renseignements complémentaires.

Frantext est pour l'heure un instrument de travail qui se décline en plusieurs formules : Frantext intégral, Frantext catégorisé, Frantext démonstration, Frantext Agrégation, Frantext Moyen Français, Frantext AFNOR et Frantext CTLF, Frantext « textes libres de droits »[9]. Comme son nom l'indique, Frantext intégral contient la totalité des textes des sous-bases Frantext, y compris les textes médiévaux (depuis juin 2013). Frantext catégorisé comporte une sélection de 1940 œuvres en prose des XIXe et XXe siècles, qui ont fait l'objet d'un codage grammatical, 'catégorisé' signifiant étiqueté grammaticalement selon les parties du discours pour permettre des recherches syntaxiques, des requêtes sur les codes grammaticaux. La base Frantext démonstration est en accès libre mais elle contient un nombre restreint de textes (35 titres allant d'une œuvre de Buffon de 1778 à une œuvre de Vidal de la Blache de 1921) ; comme ces textes ne tombent plus sous l'incidence des droits d'auteur, ils peuvent être affichés intégralement. Les textes au programme des concours de l'Agrégation et de l'École Normale Supérieur sont proposés dans deux bases spécifiques destinés à faciliter le travail des candidats aux concours. Une bibliographie accompagne Frantext intégral et réunit les références (assez sommaires pour l'instant) sous la rubrique 'Bibliographie'.

Une nouvelle formule pourrait voir le jour, suite aux améliorations apportées sur le plan des références bibliographiques et de la description des contenus textuels : Frantext référencé qui comporterait les œuvres dont les métadonnées intégrales ont été saisies et qui permettrait des interrogations nouvelles.

La version Frantext 2 sera une base de données équipée d'un moteur de recherche perfectionné, assortie d'un portail bibliographique, réunissant une base de métadonnées et la base de données elle-même, permettant l'hypernavigation entre les deux bases, offrant une interface de consultation la plus claire possible, à tous les niveaux : constitution des corpus de travail, affichage, ergonomie des fonctions de recherche. Les requêtes gagneraient en précision, notamment pour ce qui est des références, très

[9] Cette version, disponible sur le site du CNRTL : ‹http://www.cnrtl.fr/corpus/frantext/›, comporte 500 textes couvrant la période du XVIIIe au XXe siècle et autorise l'accès non payant aux textes complets. Elle est pour l'heure la seule à le proposer.

vagues pour l'heure et souvent difficiles à identifier, mais aussi pour ce qui est des contenus textuels. Intrinsèquement conçue pour servir de réservoir d'informations à plusieurs approches : littéraire, lexicale, linguistique, sociologique, historique…, la base sera accessible à nombre d'utilisateurs. Si les linguistes, les lexicographes, les stylisticiens ont toujours mis en évidence toutes les possibilités de recherche offertes par la base, si les littéraires ont apprécié la rapidité d'accès aux références, la possibilité de retrouver des occurrences et de comparer les différents usages, d'établir des corpus de travail sur certaines thématiques et problématiques avec efficacité en peu de temps, d'autres usagers pourront l'apprécier également : des bibliothécaires ou des documentalistes, des professeurs de l'enseignement secondaire et supérieur, des historiens, des juristes, des sociologues, des médecins…

Lucia MANEA

Bibliographie

Béhar, Henri, 1996. *La Littérature et son golem*, Paris, Champion.

Brunet, Étienne, 1986. *Méthodes quantitatives et informatiques dans l'étude des textes : colloque international, Université de Nice, 5-8 juin 1985, en hommage à Charles Muller*, Genève, Slatkine ; Paris, Champion.

Lemercier, Claire / Zalc, Claire, 2008. *Méthodes quantitatives pour l'historien*, Paris, La Découverte, coll. Repères.

Montémont, Véronique, 2012. « Comment bâtir une ressource lexicale : l'exemple de Frantext », *in* Ioan Negrutiu, Jean-Paul Bravard *et al.* (ed.), *Les Ressources* (Journées Scientifiques de l'Institut Universitaire de France, Lyon, 30-31 mai 2011), Presses Universitaires de Saint-Étienne.

Montémont, Véronique, (à paraître). « How To Explore A Digitalized Autobiographical Corpus : The Case Of Frantext », « 3rd Global Conference ‹ Digital Memories › », coordonnée par Daniel Riha, Prague, 16-18 mars 2011 (http: //www.inter-disciplinary.net/wp-contnt/uploads/2011/02/montemontdmpaper.pdf).

Montémont, Véronique, 2008. « Discovering Frantext », *in* Jan Auracher & Willie van Peer (ed.), *New Beginnings in Literary Studies*, Newcastle, Cambridge Scholars Publishing, 89-107.

Montémont, Véronique / Bernard, Pascale, hiver 2010-2011. « Voyage au cœur du langage : le Trésor de la langue française et Frantext », *Culture et Recherche* 124, 34-35.

Naumann, Bastian, 2004. *Der Textcorpus Frantext*, Studien arbeit, GRIN Verlag.

Rastier, François, 2001. *Arts et science du texte*, Paris, PUF.

Pierrel, Jean-Marie, 2003. « Un ensemble de ressources de référence pour l'étude du français : TLFi, Frantext et le logiciel Stella », *Revue québécoise de linguistique*, vol. 32, 155-176.

Pierrel, Jean-Marie / Buchi, Éva, 2009. « Research and Resource Enhancement in French Lexicography : the ATILF Laboratory Computerised Resources », in : Bruti, Silvia / Cella, Roberta / Foschi Albert, Marina (ed.), *Perspectives on Lexicography in Italy and Europe*, Newcastle-upon-Tyne, Cambridge Scholars Publishing, 79-117.

Bases textuelles et sites

Base textuelle FRANTEXT, ATILF - CNRS & Université de Lorraine. Site internet : ‹http://www.frantext.fr›

CNRTL : ‹http://www.cnrtl.fr/corpus/frantext/›

Dictionnaire du Moyen Français. Site internet : ‹http://www.atilf.fr/dmf/›

Frantext CTLF (Corpus de Textes Linguistiques Fondamentaux). Site internet : ‹http://ctlf.ens-lyon.fr/›

Une base de données au service de la toponymie corse

1. Préambule : la toponymie de la Corse

Les toponymes de la Corse ont une forme orale qui s'inscrit dans le cadre dialectal traditionnel de l'île. Depuis le Moyen-Âge, le toscan a offert, de par son statut véhiculaire, sa graphie aux noms de lieux insulaires dans la documentation écrite. Traditionnellement, les locuteurs, en possession des deux codes, étaient en mesure de restituer dans la forme corse les noms qui étaient graphiquement toscans. Dans les énoncés en langue française ou en contexte officiel, jusqu'à quelques années de ça, la forme toscane était employée dans la plupart des cas pour les noms des villes, régions et hameaux principaux.

L'équilibre corse-toscan s'est progressivement rompu avec la francisation linguistique de l'île. Partant de la forme toscane du nom traditionnel, le contact avec le français conduit à une altération des toponymes : la prononciation tend à s'effectuer de plus en plus en fonction de la valeur des graphèmes du français. On opposera ainsi les noms prononcés selon la tradition corse[1] et toscane, face à une prononciation inspirée du français, fait qui n'était, jusqu'à une vingtaine d'années de ça, réservé qu'aux noms des principales villes de l'île. Par exemple :

	Graphie toscane	forme corse	forme altérée
(a)	*Murato*	[mur'atu]	[myrato]
(b)	*Salice*	[u z'aldʒɛ]	[salis]
(c)	*Macinaggio*	[matʃin'aɟu]	[masinaʒ]
(d)	*Ferringule*	[fɛr̄ 'iŋgulɛ]	[farinɔl]

Selon les schémas prosodiques du français, le déplacement d'accent sur la dernière syllabe tend à se généraliser (exemples *supra*) et les noms à s'abréger, particulièrement les noms composés :

	Graphie toscane	forme corse	forme altérée
(e)	*Pietra Serena*	[p'etra z'erɛna]	[pjetra]
(f)	*Bustanico*	[bust'aniku]	[bys]

[1] Nous donnons ici un transcription moyenne de la prononciation corse.

On observe également des réinterprétations de toponymes, selon un exemple connu *e Canne* (formé depuis canna 'roseau') devenu *les Cannes*, ou encore des traductions comme *San Fiurenzu* > *Saint Florent*, *Stretta di u Soli* > *Rue du Soleil*. En outre, l'urbanisation importante que connaît la Corse s'accompagne bien entendu d'un développement important de néotoponymes, souvent en français (*Les Résidences de… Les Jardins de… Les Terrasses de…*).

Ces évolutions, liées au contact des variétés, sont accrues par l'abandon des campagnes, les mutations démographiques et le recul constant de la pratique linguistique du corse comme vernaculaire, ainsi que la perte du toscan comme variété véhiculaire. Ces faits conduisent à une altération profonde voire à la disparition du patrimoine toponymique de l'île.

Du côté des études scientifiques relatives à la toponymie corse, au-delà du faible nombre d'études d'intérêt notable, le bilan dressé (Medori 2009) a montré qu'il était indispensable de disposer de données fiables, issues de la tradition orale, pour mener des analyses scientifiques sur les noms de lieux de la Corse.

Par ailleurs, les collectivités locales – particulièrement la *Collectivité Territoriale de Corse* (CTC) – se montrent désireuses, depuis quelques années, de disposer d'un outil répondant à des besoins divers tels que l'adressage, la signalétique ou les documents d'urbanismes, mais surtout la valorisation de la toponymie corse qui fait partie des actions en faveur de la langue corse dans le cadre de la *Cartula di a Lingua Corsa* (mise en œuvre par la CTC auprès d'institutions et de partenaires privés).

2. Le CESIT Corsica : création d'une base de données toponymiques

La situation que nous avons décrite succinctement, a conduit à la création, en 2009, d'une association Loi 1901, le *Comité d'Etudes Scientifiques et Informatiques de la Toponymie de la Corse*, CESIT Corsica. Cette association a été fondée par des scientifiques (linguistes, informaticiens, historiens), et par des amateurs d'onomastique; elle s'est dotée d'un Conseil Scientifique afin d'assurer la qualité de sa démarche. Elle s'est donnée pour but d'inventorier, par le biais d'enquêtes de terrain réalisées auprès de locuteurs natifs, les toponymes traditionnels de la Corse. Le recueil des toponymes s'appuie sur la cartographie des données cadastrales (cadastre napoléonien et rénové), et de l'IGN. Si le travail de l'association s'articule essentiellement autour du recueil des toponymes oraux, certains membres de l'association dédient leur temps aux recherches en archives.

Après un séjour d'étude auprès de l'*Atlante Toponomastico del Piemonte Montano* à Turin grâce au soutien de la *Corsica Ferries*, le chantier de la base de données toponymiques a été entamé. Conçue sur la plateforme DynMap pendant l'année 2010 par la société *Cyrnéa Info Géographie*, avec le concours d'un financement de la *Collectivité Territoriale de Corse*, elle a été mise en ligne en juillet 2011 à l'adresse suivante : <http://www.cesitcorsica.org>.

Désormais, nous allons présenter brièvement les contenus de la base de données[2]. Celle-ci accueille le dépouillement des noms de lieux qui sont géolocalisés. En préambule, il faut préciser que cinq catégories de toponymes sont distinguées:

- microtoponymes ;
- macrotoponymes ;
- noms de communes ;
- petits choronymes (vallées, chaînes montagneuses) ;
- choronymes (régions).

Pour des raisons techniques, les microtoponymes sont matérialisés par des points, tandis que les macrotoponymes et les choronymes sont matérialisés par des polygones. Les fonds cartographiques, rendus accessibles par une convention avec l'*Office de l'Environnement de la Corse* sont

- le Top 25 (IGN) ;
- le cadastre rénové ;
- la photographie aérienne de 2002.

Pour chaque toponyme saisi, les informations consignées sur les fiches sont les suivantes (macrostructure) :

- forme phonique et orthographique ;
- ethnonyme corse pour les lieux habités (régions, communes, hameaux, quartiers) ;
- informations INSEE (nom de commune, toponyme et référence) ;
- métadonnées : noms de l'enquêteur, de l'informateur, et du collecteur pour les sources écrites (cadastres, archives), dates ;
- données cadastrales et du Plan terrier de la Corse (formes et références dont numéros de parcelles ou de rouleau) ;
- données d'archives (publiées et inédites) ;
- analyse linguistique (classification, étymologie, reconstruction du lemme et du signifié originels) ;
- référent ;
- documents sonores, textuels, iconographiques.

Le public n'a accès qu'à des fiches simplifiées[3] avec nom corse, nom du cadastre napoléonien, ethnonymes. Il faut préciser que lorsque la forme corse d'un toponyme n'a pas été recueillie oralement, et quelle qu'en soit la raison (aléas de l'enquête ou perte du nom dans la mémoire collective), c'est le nom du cadastre napoléonien qui figure sur l'en-tête de la fiche ; il est noté avec un astérisque pour marquer le fait qu'il n'est pas attesté à l'oral (**Pietre Ritte*, **Aja al Chioso*).

---

2 Prochainement en accès multilingue : français, corse, italien et anglais.

3 Le site fait actuellement l'objet d'un travail d'ergonomie afin d'améliorer l'accessibilité et la maniabilité des données par le public. Il est à noter que le public peut aussi suggérer des toponymes par le biais d'un forum avec carte interactive

Les modules de travail internes sont, comme le montre la liste *supra*, beaucoup plus complexes que ne le suggère l'accès externe. L'analyse linguistique des toponymes fait toujours l'objet d'essais et de réflexions qui seront soumis prochainement au conseil scientifique pour évaluation. L'objet est, à terme, de pouvoir conduire des recherches complexes telles que les :

1. continuateurs d'un étymon ;
2. correspondants onomastiques d'un élément lexical y compris s'il a disparu de l'usage ;
3. formes onomastiques qui correspondent à un signifié d'appellatif originel.

La géolocalisation des données et le développement à venir de modules de symbolisation devraient permettre d'appréhender la base de données comme un outil géolinguistique. Il devrait être possible, ainsi, d'établir par les traces toponymiques[4], l'ancienne extension d'une unité lexicale par comparaison avec des cartes d'atlas linguistiques.

3. Quelques résultats

En raison de la forte dégradation de la situation linguistique de la Corse, le CESIT Corsica a choisi de faire des enquêtes de terrain sa priorité. A ce jour, au niveau quantitatif, une soixantaine de communes a été couverte (en totalité ou partiellement)[5], et les données sont saisies progressivement. Les 5000 premières fiches ont fait l'objet d'une première phase de correction et la vérification de chaque forme phonique et graphique est en cours par un dépouillement de contrôle des enregistrements.

D'un point de vue qualitatif, voire scientifique, le collectage des toponymes oraux, confronté à l'analyse linguistique, permet de réviser certaines données offertes par les cadastres, révisions confortées également par l'observation des référents. On peut citer par exemple :

	Cadastre		Forme corse		Étymon
(g)	*La Mosa*	*vs*	*Lamosa*	<	LAMA REW 4862 'marais' ou 'ronce' ;
(h)	*Areto*	*vs*	*Laretu*	<	LAURUS REW 4943 'laurier' ;
(i)	*Campo ai Piedi*	*vs*	*Campu li Peri*	<	PĬRUS REW 6525 'poirier';
(j)	*Acqua Degna*	*vs*	*Acque Tigne*	<	TĪNEA REW 'misérable' ;
(k)	*Pentanelli*	*vs*	*Pantanelli*	<	*PALTA REW 6177 'vase, marais'.

Concernant la dimension diachronique, quelques faits peuvent déjà être mentionnés, telle l'observation de l'extension géographique du type lexical *valdu* 'forêt' aujourd'hui employé dans une aire géolinguistique restreinte du centre ouest (BDLC) alors que les attestations toponymiques montrent sa présence sur tout le territoire

[4] Dans cette perspective voir Pfister (1999) et des applications pour la Corse notamment in Medori (2013).

[5] La Corse compte 366 communes.

insulaire, ce qui est d'ailleurs démontré par la documentation médiévale[6]. On peut observer aussi la substitution d'une forme lexicale à une autre, la forme ancienne étant figée par les noms de lieux. C'est le cas de *Salettu* < SALĬCTUM REW 7534 'saulaie', dont P. Aebischer (1963, pp. 169-181) a pu démontrer qu'il s'est éteint au Moyen-Âge, spécialement en Toscane. On peut évoquer également le témoignage de l'usage passé de **àciaru* < lat. *ACER (LEI s.v.) 'érable sycomore' qui a disparu de la pratique linguistique contemporaine. Enfin, du point de vue du signifié de l'appellatif originel, les données toponymiques suggèrent la reconstruction de certains signifiés comme *'source' pour toponymes de type *Polla* (signifié attesté en Italie, DELI s.v. *pollare*) ou *Prunu,* les substantifs corses *polla* (avec dérivés) et *prunu* connaissant aujourd'hui d'autres signifiés.

4. Conclusion

La base de données du CESIT Corsica répond à un besoin scientifique et présente un potentiel important en terme de transfert vers la société civile, puisqu'elle est amenée à être utilisée dans le cadre de révisions cartographiques, de signalétique, d'adressage mais aussi, par les analyses qui devraient se développer prochainement, en terme d'outil pour la connaissance du territoire (géographie historique, urbanisme). Sur le plan scientifique, l'absence de données fiables était jusqu'alors préjudiciable aux études toponymiques relatives à l'île. Le CESIT Corsica espère, par la mise à disposition du produit de ses enquêtes, contribuer à la connaissance du trésor onomastique de l'île ainsi qu'à sa conservation – restitution, la démarche du CESIT répondant en tout point aux recommandations de l'UNESCO pour la sauvegarde des noms géographiques.

CESIT Corsica

Stella RETALI-MEDORI
Francescu Maria LUNESCHI

Références bibliographiques

Aebischer, Paul, 1963. *Miscelánea Paul Aebischer*, Barcelona, Instituto internacional de cultura románica, Biblioteca filológica-histórica 9.

BDLC = *Banque de Données Langue Corse*, éd. par Marie José Dalbera-Stefanaggi *et al*. ‹bdlc.univ-corse.fr›.

Dalbera-Stefanaggi, Marie-José / Retali-Medori, Stella (ed.), 2013. *Castagni è puddoni, le lexique corse de la castanéiculture*.

DELI = Cortelazzo, Manlio / Zolli, Paolo, 1999. *Dizionario Etimologico della Lingua Italiana*, Bologna,

Zanichelli.

Pfister, Max, 1999, « L'importanza della toponomastica per la storia della lingua nella Galloromania e nell'Italoromania », *RIOn* V/2, 449-464.

[6] Voir Dalbera-Stefanaggi / Retali-Medori (2013).

Retali-Medori, Stella, 2009. «Toponymie corse: études et matériaux», *RIOn* XV/1, 71-88.

Retali-Medori, Stella, 2013. « Éléments gallo-italiens et gallo-romans dans les parlers corses », *RLiR* 77, 121-138.

Le projet « Lire en contexte à l'époque prémoderne. Enquête sur les manuscrits de fabliaux »

La question de la mise en recueil de la littérature française médiévale est au cœur de la recherche depuis les origines de la discipline et les célèbres *Notices et Extraits des manuscrits de la Bibliothèque du Roi* (1787-1933). Mais depuis quelques décennies, le mouvement s'est amplifié et l'on constate que la volonté de ramener l'attention de l'œuvre éditée vers son support de transmission originel a entraîné une tendance parallèle, qui replace le texte copié dans son contexte manuscrit. Le potentiel intellectuel du sujet est tel qu'il a dépassé depuis longtemps le simple effet de mode. Le projet « Lire en contexte à l'époque prémoderne. Enquête sur les manuscrits de fabliaux » le confirme.

Ce projet de recherche international a débuté en 2010 et trouvera sa conclusion en 2014. Il est dirigé conjointement par Olivier Collet (Université de Genève), Francis Gingras (Université de Montréal) et Richard Trachsler (Université de Zürich)[1]. Il met en action trois équipes composées de chercheurs débutants ou avancés en littérature médiévale, histoire de la littérature, codicologie et philologie. Par son sujet et ses méthodes, il s'inscrit dans la lignée des travaux pionniers de Sylvia Huot (1987), Ian Short (1988), Pamela Gehrke (1993), Marc-René Jung (1995), Geneviève Hasenohr (1999) et Keith Busby (2002) et, par son corpus, dans celle du *Nouveau Recueil Complet des Fabliaux* (1983-1998, désormais NRCF). Il présente ainsi un triple avantage. Celui de se consacrer entièrement au sujet de la mise en recueil, ce qui reste finalement relativement rare pour le domaine francophone[2] ; d'apporter une intéressante synthèse et de permettre un état de la question sur un sujet dont l'étude a débuté il y a plus de 25 ans maintenant ; et de compléter l'œuvre éditoriale du NRCF en procurant un volume consacré aux manuscrits de fabliaux et à leur description codicologique[3].

Le corpus compte 43 recueils différents en taille, en support, en date de confection et surtout en contenu. En effet, l'une des particularités des manuscrits de fabliaux

---

1 A l'ouverture du projet, R. Trachsler travaillait à l'Université de Goettingen et dirigeait la partie allemande du projet.

2 En effet, ce type de recherche a jusqu'à présent plutôt été le fait d'études ponctuelles sur l'un ou l'autre manuscrit, au profit d'un article. On peut citer toutefois les ouvrages collectifs suivants : Leroux (2007), Foehr-Janssens / Collet (2010), van Hemelryck / Thiry (2010).

3 Le *Nouveau Recueil Complet des Fabliaux* est l'édition critique synoptique de tous les fabliaux de langue française. Il ne compte pas moins de dix volumes ; aucun toutefois ne décrit les manuscrits dans lesquels ces matériaux sont consignés.

réside dans le fait que ces textes littéraires ne sont jamais conservés seuls[4], mais qu'ils côtoient des genres aussi nombreux que variés, passant du romanesque au satirique ou au liturgique. C'est ce qui rend cet objet d'étude aussi riche que complexe, et qui de fait a nécessité la création d'une base de données pour la saisie et le traitement des informations codicologiques, philologiques, paléographiques et iconographiques. Elle a été développée par Mohan Halgrain (Université de Neuchâtel) à partir du logiciel Filemaker Pro et porte le nom de « Ecodex ».

La base possède une interface claire et fonctionnelle qui permet autant de saisir de nouvelles données que de procéder à des recherches transversales et/ou cumulatives. Elle est construite sur un principe d'arborescence et s'organise autour de plusieurs 'bibliothèques' : manuscrits, textes, mains, styles ornementaux, bibliographie. En pratique, chaque manuscrit possède une fiche qui lui est dédiée et dont les différents onglets permettent de décrire toutes les spécificités. La page principale renseigne sur les données générales (lieu de conservation, sigle, reliure, date de confection, cotes anciennes et histoire du manuscrit). Les autres sur l'organisation matérielle du manuscrit, les foliotations, les textes contenus, les mains et l'iconographie notamment. Le point fort de cet outil est de permettre une vision globale des manuscrits, médiévale et moderne, puisqu'il répertorie autant l'histoire des volumes et leurs interventions ultérieures, les liens de fabrication ou de tradition textuelle que la bibliographie spécifique. C'est le même principe organisationnel qui sous-tend le formulaire de recherche. Ce dernier permet d'interroger la base par niveau (selon que l'on souhaite obtenir une réponse générale sur les manuscrits ou spécifique sur leurs unités), en incluant, excluant ou cumulant les éléments de recherche.

Les premiers bénéfices de ce travail sont déjà visibles. 2013 voit la publication du 48e volume de la revue des *Etudes françaises* qui présente le travail de nombreux membres du projet sur la question de la lecture en contexte des manuscrits de fabliaux. C'est encore cette année que se déroulent les deux volets du colloque organisé par le projet : les 24 au 26 octobre 2013 à Montréal sur le thème « Les Centres de production de la littérature vernaculaire » et les 12-13 décembre à Zürich avec « Les Fabliaux et le recueil complet ». Il faudra attendre fin 2014 pour la publication d'un volume conclusif. Il contiendra en introduction des réflexions poussées sur les manuscrits de fabliaux et leur mise en contexte, puis laissera la place à l'édition des notices codicologiques de tous les témoins du corpus. La base devrait être disponible dans le courant de l'année 2015. Elle se placera comme un intéressant complément à la publication papier, car elle mettra à disposition des chercheurs la globalité des données récoltées ainsi que l'interface de recherche qui leur permettra de combiner ces informations selon leurs propres besoins et intérêts.

Université de Genève Gaëlle MOREND JAQUET

[4] A l'exception du ms. Paris, Bibliothèque Nationale de France, Français 2188, qui conserve le seul Trubert. Rappelons toutefois que cette œuvre est un fabliau singulier, dont la pertinence générique est sujette à controverse.

Références bibliographiques

Busby, Keith, 2002. *Codex and Context: Reading Old French Verse Narrative in Manuscript*, Amsterdam, Rodopi, Faux titre (221 - 222), 2 vol.

Collet, Olivier, Gingras, Francis, Trachsler, Richard (ed.), 2012. « Lire en contexte : enquête sur les manuscrits de fabliaux », *Etudes françaises*, 48/3.

Foehr-Janssens, Yasmina, Collet, Olivier (ed.), 2010. *Le recueil au Moyen Age. Le Moyen Age central*, Turhout, Brepols, Texte, codex & contexte, 8.

Gehrke, Pamela, 1993. *Saints and Scribes. Hagiography in its Manuscript Context*, Berkeley, University of California Press.

Hasenohr, Geneviève, 1999. « Les recueils littéraires français du XIII[e] siècle : public et finalité », in Jansen-Sieben, R. et van Dijk, H. (ed.), *Codices Miscellanearum, Colloque Van Hulthem Bruxelles 1999*, Bruxelles, Archives et bibliothèques de Belgique, 37-50.

Hemelryck, T. van, Thiry, C. (ed.), 2010. *Le recueil au Moyen Age. La fin du Moyen Age*, Turnhout, Brepols, Texte, codex & contexte, 9.

Huot, Sylvia, 1987. *From song to book. The poetics of writing in old French lyric and lyrical poetry*, New York, Cornell University Press.

Institut royal de France (ed.), 1787-1933. *Notices et extraits des manuscrits de la Bibliothèque du Roi : et autres bibliothèques [...], faisant suite aux Notices et extraits lus au comité établi dans l'Académie des inscriptions et belles-lettres*, Paris, Imprimerie royale, puis Impr. impériale, puis Impr. nationale, 42 vol.

Jung, Marc-René, 1995. *La légende de Troie en France au Moyen Âge: analyse des versions françaises et bibliographie raisonnée des manuscrits*, Tübingen-Basel, Francke, Romanica Helvetica 114.

Leroux, Xavier (ed.), 2007. « La mise en recueil des textes médiévaux », *Babel*, 16.

Noomen, Willem, Boogaard, Nico van den (ed.), 1983-1998. *Nouveau Recueil Complet des Fabliaux*, Assen, Van Gorcum, 10 vol.

Short, Ian, 1988. « L'Avènement du texte vernaculaire: la mise en recueil », in Baumgartner, Emmanuèle et Marchello-Nizia, Christiane (ed.), *Théories et pratiques de l'écriture au moyen âge : actes du colloque, Palais du Luxembourg-Sénat, 5 et 6 mars 1987,* Paris, Université de Paris X Nanterre, Littérales 4, 11-24.

Le corpus multilingue InterCorp et les possibilités de son exploitation

L'objectif de cet article est de présenter un nouvel outil de recherche linguistique, le corpus multilingue InterCorp (www.korpus.cz / intercorp), et les possibilités de son utilisation. Après avoir montré la composition du corpus et le cadre institutionnel de tout le projet dans la section 1, nous allons préciser dans la section 2 les procédés techniques de la constitution du corpus et détailler les fonctions les plus importantes du moteur de recherche qui permet de l'exploiter. Pour terminer, nous donnerons dans la Conclusion quelques exemples de recherches déjà réalisées sur le corpus InterCorp, tout en soulevant les problèmes méthodologiques les plus pertinents qui pèsent sur toute exploitation de corpus parallèle.

1. La composition du corpus et son cadre institutionnel

Le corpus multilingue InterCorp implique actuellement 39 langues, parmi lesquelles des langues romanes – le catalan, le français, l'espagnol, l'italien, le portugais et le roumain. Le catalan n'a été intégré au corpus InterCorp qu'en 2013, mais grâce au dynamisme des coordinateurs et grâce à l'aide du Ministère de l'Éducation nationale de la Principauté d'Andorre, cette section se développe rapidement. Les langues slaves sont également bien représentées : nous y trouvons les sections biélorusse, bulgare, croate, macédonienne, polonaise, russe, slovaque, slovène, serbe, ukrainienne et tchèque. L'anglais ne manque naturellement pas dans ce corpus multilingue, mais d'autres langues germaniques y sont également présentes : l'allemand, le danois, le néerlandais, le norvégien et le suédois. En outre, deux langues difficiles à traiter par les outils du TALN, l'arabe et l'hindi, ont été récemment ajoutés à la liste des langues pour lesquelles le corpus InterCorp offre des données linguistiques ajoutées[1].

InterCorp est un ensemble de corpus parallèles, c'est-à-dire un ensemble de paires composées de textes originaux et de leurs traductions respectives, et les textes sont

[1] Voici la liste complète des langues représentées dans le corpus InterCorp (par ordre alphabétique) : l'albanais, l'allemand, l'anglais, l'arabe, le biélorusse, le bulgare, le catalan, le croate, le danois, l'espagnol, l'estonien, le finnois, le français, le grec, l'hebreu, l'hindi, le hongrois, l'italien, l'islandias, le japonais, le letton, le lithuanien, le macédonien, le malais, le maltais, le néerlandais, le norvégien, le polonais, le portugais, le roumain, le russe, le slovaque, le slovène, le serbe (en alphabets latin et cyrillique), le suédois, l'ukrainien, et le tchèque – langue pivot. L'ajout d'autres langues, telles que le chinois et le turque, le romani, le vietnamien est aussi envisagé.

alignés au niveau des phrases. Pour des raisons techniques ainsi que conceptuelles, la langue tchèque est la langue pivot de l'ensemble du corpus – chaque texte doit avoir sa contrepartie tchèque. En outre, tous les textes sont alignés d'après la segmentation du texte tchèque, ce qui donne au corpus l'homogénéité nécessaire. Néanmoins, dans les recherches sur corpus, il est possible de laisser le tchèque de côté et d'effectuer la recherche seulement sur les langues « étrangères », par exemple sur l'anglais, l'allemand, le catalan, l'espagnol, le finnois, le français et l'italien[2]:

[FR] *« Moi, se dit le petit prince, si j'avais cinquante-trois minutes à dépenser, je marcherais tout doucement vers une fontaine … »* (Antoine de Saint-Exupéry, *Le Petit prince*)

[ALL] *„Wenn ich dreiundfünfzig Minuten übrig hätte", sagte der kleine Prinz," würde ich ganz gemächlich zu einem Brunnen laufen…"* (trad. par Leitgeb, Grete; Leitgeb, Josef)

[ANG] *"As for me," said the little prince to himself, "if I had fifty-three minutes to spend as I liked, I should walk at my leisure toward a spring of fresh water."* (trad. par Katherine Woods)

[CS] *Kdybych já měl padesát tři minuty nazbyt, řekl si malý princ, šel bych docela pomaloučku ke studánce…* (trad. par Zdeňka Stavinohová)

[HIN] „अगर मेरे पास तिरपन मिनिट बिताने को होते „ छोटे राजकुमार ने सोचा, , तो मैं धीरे - धीरे एक जलाशय की ओर चल पड़ता ... ।» (trad. par कशिोर बलवीर, जगवं श)

[IT] *„Io», disse il piccolo principe, «se avessi cinquantatré minuti da spendere, camminerei adagio adagio verso una fontana…»* (trad. par Nini Bompiani Bregoli)

[RO] *«Eu, îşi spuse micul prinţ, dacă aş avea de irosit cincizeci si trei de minute, aş porni în linişte spre o fântână.»* (trad. par Benedict Corlaciu)

Comme nous l'avons signalé ci-dessus, la langue tchèque se trouve en tant que langue pivot au centre du corpus aussi pour des raisons conceptuelles : en effet, tout le projet est centré en particulier sur les recherches contrastives du tchèque par rapport à d'autres langues. Le projet, financé dans son intégralité par le Ministère de l'Éducation nationale tchèque, est réalisé par l'Institut du Corpus national tchèque (www.korpus.cz) qui fait partie de la Faculté des Lettres de l'Université de Prague. L'Institut du Corpus national tchèque a été fondé déjà en 1994 afin de créer le corpus de référence pour la langue tchèque, et cet objectif a été atteint en 2000, quand le corpus SYN2000 a été rendu librement accessible sur Internet. Dans sa partie principale, synchronique, le Corpus national tchèque contient actuellement plus d'un milliard de mots et il est constamment actualisé. Le corpus central est complété par plusieurs corpus spécifiques, par exemple le corpus diachronique, des corpus oraux, des corpus de correspondance ou des textes linguistiques, etc. Depuis 2005, InterCorp, le corpus parallèle, fait également partie de ce vaste projet.

Le corpus InterCorp est le résultat de la coopération de plusieurs établissements et organismes en République tchèque. La préparation des textes en langues étrangères

[2] Le moteur de recherche permet aussi des recherches unilingues sur le sous-corpus choisi ; il suffit de choisir seulement une langue dans le formulaire initial.

est assurée par les coordinateurs qui appartiennent en général aux différents départements linguistiques de la Faculté des Lettres de l'Université Charles à Prague. Quant aux outils informatiques nécessaires pour la préparation et l'exploitation du corpus, ils sont souvent issus de la coopération avec des spécialistes en TALN d'autres établissements, par exemple l'Institut de la linguistique théorique et computationnelle de la Faculté des Lettres de l'Université à Prague (http://utkl.ff.cuni.cz), l'Institut de la linguistique formelle et appliquée de la Faculté des Mathématiques et de la Physique de la même Université (http://ufal.mff.cuni.cz/) ou la Faculté de l'Informatique de l'Université Masaryk de Brno (http://www.fi.muni.cz). L'Institut du Corpus national tchèque est ensuite responsable de la coordination de toutes les activités, du stockage central des données, de leur traitement informatique et de leur mise en ligne via le moteur de recherche (‹http://kontext.korpus.cz›).

La première version du corpus InterCorp a été mise en ligne en 2008, mais depuis, le corpus ne cesse de croître : après la dernière actualisation en mai 2015, l'ensemble du corpus InterCorp contient 1 597 462 625 mots. Le noyau du corpus est constitué par les textes littéraires, ce qui représente la spécificité du projet en comparaison avec d'autres corpus parallèles, tels que OPUS ou Hansard par exemple. Le sous-corpus littéraire, contenant actuellement presque 280 millions de mots, est complété par quatre grands ensembles de textes. Premièrement, il s'agit de textes journalistiques tirés de serveurs multilingues *Project Syndicate* (www.project-syndicate.org) et *Presseurop / Voxeurop* (www.voxeurop.eu) en dix langues (51 177 809 mots au total)[3]. En deuxième lieu, il est possible d'effectuer les recherches sur les textes juridiques de l'Acquis communautaire de l'Union européenne, disponibles dans toutes les langues de communication des états membres (450 498 235). Le troisième ensemble de textes n'a été ajouté qu'en avril 2013 – il s'agit du corpus EuroParl, constitué par les débats au Parlement européen (277 951 947 mots). La dernière collection, intégrée au corpus en 2014, contient des sous-titres de films (Subtitles, 539 060 969 mots).

[3] Les collections *Presseurop* et *Project Syndicate* sont disponibles en allemand, anglais, espagnol, français et italien ; le russe est inclus seulement dans le *Project Syndicate*, et le néerlandais, le polonais, le portugais et le roumain font partie uniquement de *Presseurop*.

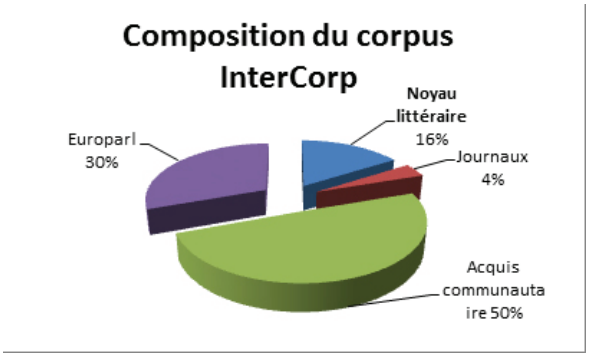


Figure 1 – Composition du corpus InterCorp (par types de textes)

Figure 1 montre les données pour l'ensemble du corpus InterCorp, mais il faut préciser que les différentes langues ne sont pas représentées dans le corpus de manière égale, comme le révèle *Figure 2* ci-bas. Parmi les langues les mieux représentées, nous trouvons par exemple l'espagnol : ce sous-corpus contient plus de 100 millions de mots dans sa totalité, dont 17 millions dans le noyau littéraire. Dans cette section du corpus, nous pouvons effectuer des recherches par exemple sur plusieurs textes de Gabriel Garcia Marquez ou de Mario Vargas Llosa. L'italien et le français ne sont pas en reste – leurs sous-corpus comptent presque 90 millions de mots (le français) et 65 millions de mots (l'italien), dont par exemple le français 9 millions de mots dans le noyau littéraire. Dans ces sections, l'utilisateur peut effectuer des recherches par exemple dans les textes de Patrick Chamoiseau, Albert Camus ou Amélie Nothomb pour le français, et dans des œuvres de Umberto Eco ou Alessandro Baricco pour l'italien.

Les différences de taille entre les sections linguistiques dépendent tant du taux d'activité des différents coordinateurs que de la disponibilité des textes – par exemple le nombre de traductions du portugais en tchèque (et vice versa) est assez limité, et ce sous-corpus doit ainsi être complété par des traductions d'une troisième langue, par exemple des livres de J.R.R. Tolkien ou de J.K. Rowling. Il faut également préciser que le noyau dit « littéraire » contient aussi des textes appartenant à d'autres genres, par exemple des essais (en français des textes d'Albert Camus ou de Saint-Exupéry) ou des textes scientifiques, comme ceux de Georges Duby[4]. Dans la section française, nous pouvons trouver également les versions parallèles de cinq volumes des aventures d'Astérix et Obélix[5].

[4] Avant la fin de l'année 2013, des textes de Michel Foucault, Henriette Walter et Ferdinand de Saussure ont été ajoutés au corpus.

[5] En définissant le corpus de travail, l'utilisateur peut préciser les genres qu'il désire inclure dans sa recherche.

Langue	Noyau	Journaux	Acquis	Europarl	Subtitles
ar	34 325	0	0	0	0
be	2 152 724	0	0	0	0
bg	5 240 831	0	13 816 405	9 083 403	0
ca	4 632 696	0	0	0	0
da	3 016 838	0	21 679 997	13 915 841	14 429 778
de	27 681 897	6 207 922	21 723 929	13 089 209	8 366 765
el	0	0	25 069 611	15 403 662	23 714 597
en	15 488 167	6 488 284	24 207 801	15 580 109	52 101 283
es	17 475 748	7 140 829	27 001 343	15 885 394	36 378 715
et	0	0	15 962 544	10 899 550	10 296 031
fi	3 426 226	0	16 455 144	10 175 256	15 097 653
fr	9 170 042	7 321 278	27 351 591	17 178 444	25 961 848
he	0	0	0	0	16 221 237
hi	408 616	0	0	0	0
hr	15 479 547	0	0	0	19 092 559
hu	5 387 533	0	19 176 514	12 306 692	21 239 634
is	0	0	0	0	1 584 758
it	7 247 545	3 359 150	24 849 477	15 489 468	14 653 613
ja	0	0	0	0	113 320
lt	358 253	0	18 392 644	11 212 864	557 961
lv	1 336 888	0	18 744 927	11 688 597	280 117
mk	3 741 900	0	0	0	1 877 210
ms	0	0	0	0	3 520 701
mt	0	0	14 133 133	0	0
nl	9 961 680	3 269 635	24 746 144	15 563 231	29 362 826
no	4 815 797	0	0	0	0
pl	17 516 332	2 378 025	20 627 627	12 811 143	26 572 483
pt	2 393 287	3 369 337	28 602 556	16 484 692	43 391 919
ro	3 432 615	2 737 807	8 199 565	9 446 369	34 128 511

ru	3 337 545	3 174 152	0	0	6 885 753
sk	7 401 998	0	19 222 784	12 734 444	5 134 150
sl	900 221	0	19 645 598	12 240 548	17 024 593
sq	0	0	0	0	2 003 579
sr	8 823 894	0	0	0	20 776 850
sv	8 138 161	0	20 585 800	13 840 373	14 693 861
tr	0	0	0	0	21 190 828
uk	5 054 034	0	0	0	246 059
vi	0	0	0	0	1 473 591
cs	84 718 325	5 731 390	20 303 101	12 922 658	50 688 186
TOTAL	278 773 665	51 177 809	450 498 235	277 951 947	539 060 969

Figure 2 – Composition du corpus InterCorp (par langues en nombre de mots)

Les textes pour le corpus sont choisis en fonction de plusieurs critères. Le premier critère limitant le choix de textes est la date de la création de l'original – étant donné le caractère synchronique du corpus, seuls les textes écrits après la seconde guerre mondiale sont admis. Néanmoins, pour certains textes appartenant au patrimoine littéraire universel, des exceptions ont été autorisées ; c'est ainsi que nous pouvons trouver dans le corpus *Le Voyage au bout de la nuit* de Ferdinand Céline ou des textes de l'auteur tchèque Karel Čapek, qui est mort en 1938. Deuxièmement, les coordinateurs sont encouragés à choisir pour le corpus les textes susceptibles d'être traduits en plusieurs langues, pour favoriser les intersections entre les différentes sections linguistiques. Ainsi, il n'est pas surprenant de trouver parmi les textes les mieux représentés dans le corpus des best-sellers anglais, tels que *Le Seigneur des anneaux* ou *Harry Potter*. Cependant, parmi les textes les mieux « cotés », nous trouvons aussi un texte français, *Le Petit Prince* d'Antoine de Saint-Exupéry, qui est représenté dans le corpus en 25 versions linguistiques. Plusieurs textes tchèques sont également présents dans un grand nombre de versions linguistiques, des romans de Milan Kundera (en particulier *La Plaisanterie* ou *L'Insoutenable légèreté de l'être*) et le *Brave soldat Chvéïk* de Jaroslav Hašek[6].

[6] Le corpus InterCorp est ouvert à la coopération avec d'autres projets : à titre d'exemple, en 2013, l'auteur du corpus ASPAC (*The Amsterdam Slavic Parallel Aligned Corpus*, cf. Waldenfels 2006) a fourni ses textes parallèles au projet InterCorp, ce qui a permis d'améliorer leur alignement et de les rendre accessibles via le moteur de recherche *NoSketch Engine*. De même, le Corpus national tchèque a accueilli et rendu disponible le corpus de sorabe (DOTKO et HOTKO).

2. Procédés techniques de la constitution du corpus et son exploitation

Le traitement informatique des textes choisis pour le corpus est effectué en plusieurs étapes, mais la procédure est différente dans le cas des textes du noyau littéraire et ceux des collections de textes, telles que Presseurop ou Europarl.

Les textes destinés au noyau littéraire sont océrisés et entièrement relus par des collaborateurs du projet, dans la plupart des cas étudiants universitaires de la langue concernée. Les textes sont ensuite exportés à l'aide d'un macro de Visual Basic dans un format quasi-XML qui identifie les frontières entre les paragraphes et phrases et transforme les caractères spéciaux (&, <, >) en entités. Pour le tchèque, nous utilisons le segmentateur basé sur des règles, pour les autres langues, il s'agit d'un algorithme à l'apprentissage non-supervisé (*Punkt sentence tokenizer*, cf. Kiss / Strunk, 2006). Dans l'étape suivante, le coordinateur de la section linguistique concernée enregistre le texte dans la base de données rassemblant tous les textes du corpus et lui assigne les informations bibliographiques nécessaires.

Le coordinateur principal du projet effectue ensuite l'alignement automatique de la paire de textes donnée, à l'aide du logiciel *hunalign* (cf. Varga et al., 2005, http://mokk.bme.hu / rerources / hunglishcorpus /), et il enregistre le résultat dans l'éditeur de textes parallèles nommé *InterText*. Ce logiciel, créé directement pour le projet InterCorp (cf. Vondřička, 2010, http://wanthalf.saga.cz / intertext), permet aux collaborateurs du projet de corriger l'alignement automatique appliqué ultérieurement aux textes :

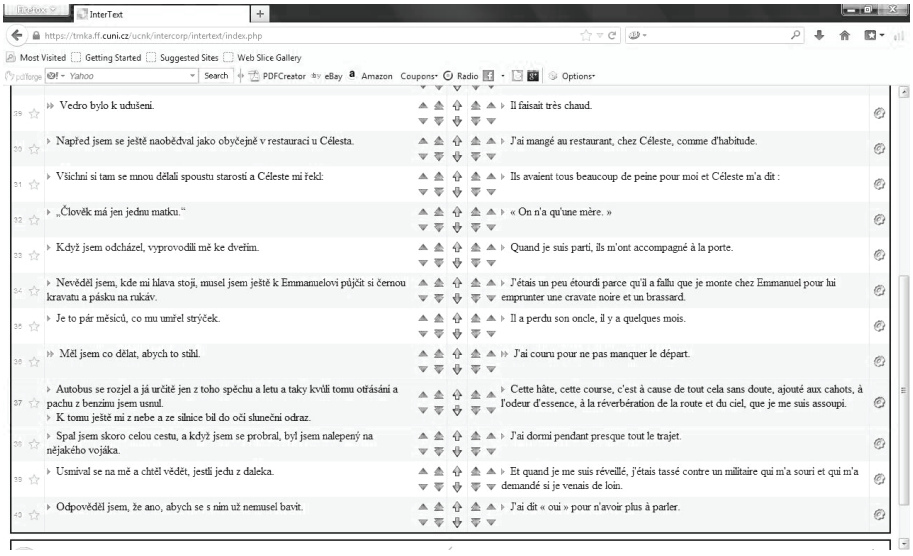


Figure 3 : L'éditeur de textes parallèles InterText

En ce qui concerne les textes des collections (*EuroParl*, *Acquis communautaire*, *Project Syndicate, Presseurop et Subtitles*), la relecture et la correction semi-manuelle de l'alignement automatique n'y sont pas appliquées. Par conséquent, les textes de ce type peuvent être ajoutés au corpus plus rapidement, en plus grande quantité et à un coût considérablement moins élevé que les textes du noyau littéraire, mais le résultat final est proportionnellement moins fiable. De plus, les collections de Presseurop / Syndicate et de l'Acquis communautaire présentent un inconvénient limitant leur utilité pour les recherches contrastives, parce qu'elles ne permettent pas d'identifier la direction de la traduction, c'est-à-dire la langue source. Pour les recherches contrastives et traductologiques, cette lacune représente un défaut majeur, étant donné les spécificités de la langue de traduction. Les techniciens participant au projet tentent actuellement de remédier à ce problème en identifiant la langue source à partir des données bibliographiques incluses directement dans les textes, et cette information devrait être disponible dans la prochaine version pour les textes journalistiques et pour Europarl (pour l'Acquis communautaire, la procédure est envisagée).

Les textes issus des alignements automatique ainsi que semi-automatique sont ensuite exportés et les liens d'alignement sont stockés dans un fichier séparé. Finalement, les textes alignés sont dotés de la lemmatisation et de l'étiquetage morphologique. Toutes les langues impliquées dans le projet InterCorp sont lemmatisées, à l'exception du bulgare, du hongrois et du néerlandais[7]. Pour l'annotation morphologique, le logiciel le plus utilisé dans le corpus est TreeTagger (www.ims.uni-stuttgart.de / projekte / corplex / TreeTagger), mais le tchèque et le slovaque sont annotés à l'aide de l'annotateur *morče* (http://ufal.mff.cuni.cz / morce), qui est capable de refléter la richesse morphologique de ces langues (il offre 15 positions de sous-ensembles de marques morphologiques, cf. Hajič, 2004). Les textes sont enfin reliés avec les données bibliographiques et indexés par le logiciel (gestionnaire de corpus) Manatee (cf. Rychlý, 2007). La nouvelle version du corpus est mise en ligne approximativement tous les six mois. L'actualisation qui a eu lieu en avril 2013, a apporté deux changements importants : le corpus a été considérablement élargi et le moteur de recherche utilisé pour l'exploitation du corpus est désormais *NoSketch Engine*, offrant un large éventail de possibilités de tri et de traitement statistique des concordances[8].

L'accès au corpus est disponible à partir du site ‹http://kontext.korpus.cz› et il est gratuit après inscription. Pour s'inscrire et obtenir le mot de passe, il suffit de remplir le formulaire en ligne où l'on s'engage à ne pas utiliser le corpus à des fins commerciales et à signaler toute publication que l'on réalise grâce aux données issues d'InterCorp (www.korpus.cz / english / dohody.php). Le même mot de passe est également valide

[7] La lemmatisation du néerlandais est prévue pour la prochaine actualisation du corpus.

[8] L'ancien concordancier, nommé Park, est toujours disponible sur ‹http://korpus.cz / Park / login›. Dans l'évaluation des données statistiques issues des deux logiciels, il faut prendre en considération le fait que les calculs dans Park sont opérés sur le nombre de mots, tandis que ceux de NoSketch Engine sont basés sur le nombre de tokens (positions), incluant non seulement les mots, mais aussi la ponctuation. En linguistique de corpus, la deuxième approche est la plus courante.

pour le Corpus national tchèque et pour quatre larges corpus unilingues contenant chacun plus d'un milliard de mots – les WaCky corpora pour l'allemand, le français, l'italien et l'anglais britannique (cf. Baroni et al., 2009)[9]. Avant la fin de l'année 2013, nous avons ajouter aussi l'accès au corpus *Est républicain* dont les textes ont été rendus disponibles par le Centre national de Ressources Textuelles et Lexicales (‹http://www.cnrtl.fr / corpus / estrepublicain›). Tous les corpus unilingues sont lemmatisés et dotés de l'annotation morphologique.

La recherche sur InterCorp s'effectue en trois étapes, comme dans la plupart des corpus :

(1) La définition du sous-corpus de travail

L'utilisateur peut choisir deux ou plusieurs langues qu'il veut analyser, par exemple le français, l'anglais et le finnois, et définir le sous-corpus de travail en fonction de plusieurs critères :

(a) le genre du texte (roman, drame, texte juridique, scientifique, etc.) ;
(b) la langue source et la distinction original / traduction ;
(c) les textes concrets.

Le sous-corpus de travail peut être sauvegardé pour que l'utilisateur puisse y revenir à chaque session.

(2) La formulation de la requête

La requête dans le corpus InterCorp peut porter sur un mot simple (*chat*) ou sur un lemme (*chanter*), mais aussi combiner les expressions régulières et les marques morphologiques dans une expression CQL (*Corpus Query Language*). Par exemple, pour chercher les gérondifs en français, il est possible de profiter des marques morphologiques assignées par TreeTagger ([word="en"] []{0,2} [tag="VER:ppre"]), ou utiliser les expressions régulières ([word="en"] []{0,2} [word=".*ant"])[10]. Les requêtes dans les autres langues choisies peuvent rester libres, mais il est également possible de préciser que les phrases équivalentes en anglais doivent contenir la forme en –ing ([word=".*ing"]) ou l'équivalent systémique présumé du gérondif en tchèque, le transgressif présent ([tag="Ve.*"])[11]. Les informations affichées dans la concordance obtenue indiquent les données bibliographiques concernant les textes sources, mais

[9] La lemmatisation du néerlandais est prévue pour la prochaine actualisation du corpus.

[10] Avant de formuler la requête impliquant l'annotation morphologique, il est nécessaire de consulter les listes de marques (*tagsets*) pour les langues choisies (cf. la liste ‹http://www.korpus.cz / intercorp / ?req=page:info›). La définition et l'utilisation d'une même marque ainsi que la segmentation peuvent varier d'une langue à l'autre (cf. Rosen, 2010 et 2012).

[11] Pour créer la requête en tchèque sans avoir à connaître les listes de marques pour toutes les 15 positions de l'annotation morphologique, l'utilisateur peut profiter de l'application nommée *klikátko*, qui lui permet de sélectionner les marques directement d'après leur description (‹http://utkl.ff.cuni.cz / ~skoumal / morfo / ?lang=cs›).

aussi des données statistiques, concrètement la fréquence absolue et relative (i.p.m.) de l'élément analysé.

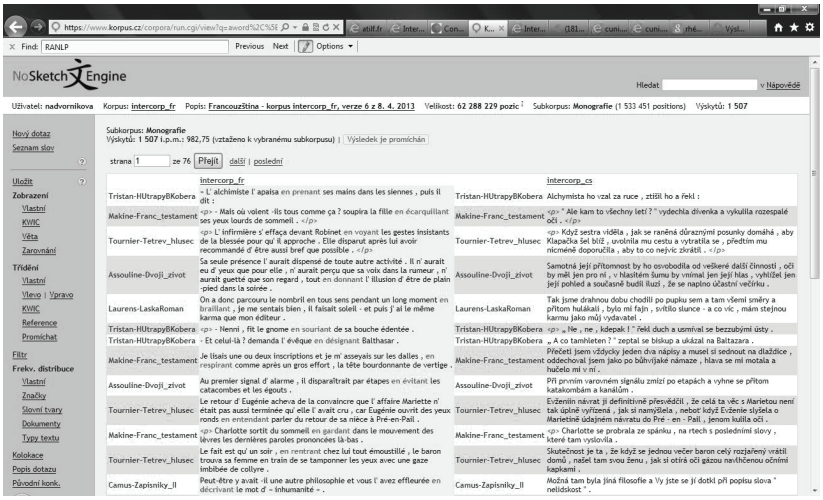


Figure 4 : Exemple de concordance dans le corpus InterCorp

(3) Le tri des résultats et leur traitement statistique

NoSketch Engine permet de trier et filtrer les résultats visionnés dans la concordance en fonction de plusieurs critères, par exemple par leur contexte de droite ou de gauche ou par les propriété de KWIC (forme, lemme, marque morphologique). Le concordancier offre également un grand nombre de possibilités de traitement statistique des données et de calcul de collocations (t-score, MI-score, logDice, etc.). La distribution de fréquence visualise de manière claire les fréquences absolue et relative dans les différents documents du corpus :

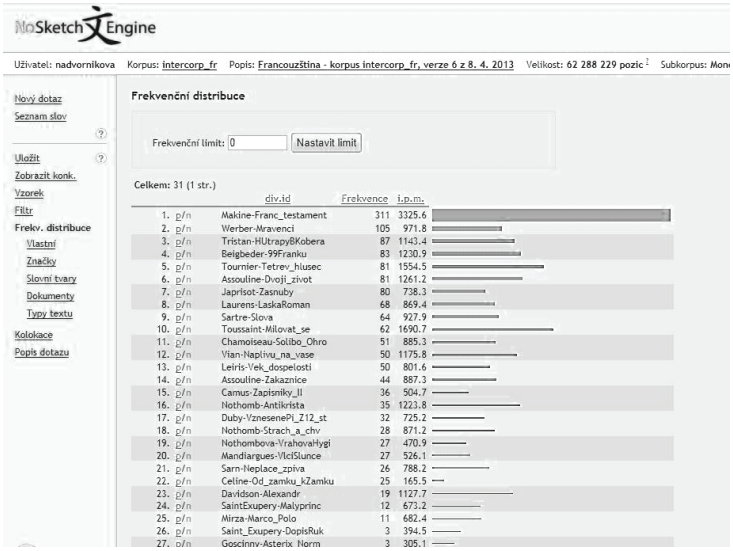


Figure 5 : Visualisation des fréquences relative et absolue de la distribution du gérondif dans le sous-corpus français

A la fin de la session, l'utilisateur peut télécharger la concordance finale dans un fichier Excel et continuer ses analyses.

3. Conclusion

Le corpus parallèle InterCorp était initialement destiné en particulier aux linguistes, mais la communauté de ses utilisateurs est actuellement bien plus variée : les traducteurs l'utilisent comme source d'inspiration en cas de problème, les enseignants de langues profitent de cet outil informatique pour faire varier leurs méthodes en classe et les étudiants en tchèque langue étrangère dispersés aux quatre coins du monde y cherchent les solutions qu'il serait impossible ou trop long de trouver dans les grammaires et les dictionnaires. Par ailleurs, l'Institut du Corpus national tchèque s'ouvre de plus en plus au large public et organise des stages d'initiation au travail sur corpus non seulement à Prague, mais aussi dans d'autres villes en République tchèque.

Néanmoins, les linguistes restent le noyau du groupe des utilisateurs du corpus et réalisent des recherches contrastives sur la plupart des langues représentées dans le corpus[12]: P. Čermák et P. Štichauer ont par exemple comparé les constructions factitives en espagnol, en italien et en tchèque (Čermák / Štichauer, 2010) ou le préfixe *re-* et le suffixe *-ble / -bile* à la lumière de leurs équivalents tchèques (voir Čermák, 2013 et

[12] L'Institut du Corpus national tchèque a déjà organisé deux colloques internationaux présentant des recherches effectuées sur InterCorp (voir Čermák / Corness / Klégr, 2010 et Čermák, 2011) et un autre colloque s'est tenu en septembre 2014 (http://www.korpus.cz / kl2014 /).

Bratánková / Štichauer, 2011). En allemand, T. Káňa et H. Pelouškova ont étudié les diminutifs ou la construction modale *ohne* + *zu* en comparaison avec leurs équivalents en tchèque (Káňa / Pelouškova, 2009 et 2011 et Čermák / Nádvorníková *et al.* 2015) ; et la section anglaise s'est concentrée sur l'analyse des prépositions (Malá / Šaldová / Klégr, 2010) ou l'ordre de mots et la portée de rhématiseurs (Martinková, 2012). L'atout principal du corpus consiste en particulier dans sa capacité d'offrir au chercheur de larges données authentiques qui lui permettent de vérifier les informations présentées comme incontestables dans les grammaires et les dictionnaires[13]. Le corpus InterCorp a ainsi permis de mettre en doute la tradition prétendant l'équivalence systémique entre le gérondif français et le transgressif présent tchèque : les données tirées du corpus montrent que l'équivalent français le plus fréquent de cette forme verbale tchèque n'est pas le gérondif, mais le participe présent (Nádvorníková, 2010b). Même les étudiants se lancent dans des recherches lexicales et grammaticales ponctuelles, explorant par exemple l'emploi de connecteurs (*en effet*) ou du conditionnel passé.

Tout en approfondissant leurs recherches sur le corpus multilingue InterCorp, les linguistes se rendent compte des contraintes méthodologiques multiples qui pèsent sur son exploitation. Premièrement, ils doivent prendre en considération les différences de taille entre les sections linguistiques du corpus ainsi que les spécificités des genres qui y sont inclus (romans, textes juridiques de l'Union européenne, etc.). De plus, les linguistes doivent accepter le fait qu'un corpus parallèle ne peut jamais être balancé et représentatif, parce que certains types de textes ne sont presque jamais traduits ; et l'oral est quasiment exclu de ce domaine[14]. Deuxièmement, les recherches sur corpus parallèles se heurtent au problème donné par la spécificité de la langue de traduction (*translationese*) qui tend à la normalisation et à la simplification (cf. par exemple Olohan, 2004, 91-104 ou Baker, 1996, 176-177). Les linguistes réagissent à ce problème d'une part en vérifiant les résultats sur les larges corpus unilingues, d'autre part en respectant strictement la direction de la traduction (la langue source)[15]. L'analyse bi-directionnelle, profitant des deux directions de la traduction, s'avère ainsi très prometteuse (cf. Malá, 2013). L'intégration de ce respect pour les spécificités de

---

[13] L'Institut d'études romanes de la Faculté des Lettres de l'Université à Prague a entamé récemment un vaste projet impliquant quatre langues romanes (l'espagnol, le français, l'italien et le portugais) dans la recherche sur cinq sujets, à savoir le préfix *re-*, le suffixe *-ble/ -bile*, les constructions inchoatives et factitives, et le gérondif. Les résultats de cette recherche sur corpus ont été publiés en 2015.

[14] Les textes du corpus Europarl ne reflètent pas la production orale spontanée et les textes sont très probablement post-édités. Le corpus parallèle de sous-titres de films pourrait refléter davantage la langue parlée authentique. (voir Čermák / Nádvorníková 2015).

[15] Pour le français, il est possible de vérifier les résultats sur FRANTEXT (www.frantext.fr) ; dans le cas du gérondif, cette procédure a révélé que la fréquence relative de cette forme verbale était plus élevée dans le noyau littéraire du corpus InterCorp (2 066 i.p.m.) que dans le corpus de 192 romans publiés après 1950 et disponibles dans FRANTEXT (1 629 i.p.m.). L'analyse plus détaillée des textes a ensuite démontré que cette différence était liée surtout à l'idiolecte d'un seul auteur, A. Makine, où la fréquence relative du gérondif s'élève à 5 135 i.p.m. (cf. Nádvorníková, 2012, 165-166).

la langue de traduction est le fruit de la coopération avec les traductologues, qui offre à la linguistique de corpus contrastive les méthodes d'analyse plus fines.

Les projets d'avenir pour le corpus multilingue InterCorp ne devraient donc pas être concentrés seulement sur l'élargissement du corpus et l'amélioration de son moteur de recherche (par exemple par l'ajout de l'alignement par mots), mais les linguistes utilisant les données parallèles devraient se lancer dans un débat méthodologique sérieux qui leur permettrait de profiter pleinement des vastes champs d'exploration que les corpus multilingues leur ouvrent.

Université Charles à Prague Olga NÁDVORNÍKOVÁ

Références bibliographiques

Baker, Mona, 1996. « Corpus-based translation studies: The challenges that lie ahead », in : Somers, H. (ed.), *Terminology, LSP and Translation: Studies in languagae engineering, in honour of Juan C. Sager*, Amsterdam, John Benjamins, 175-186.

Baroni, Marco / Bernardini, Silvia / Ferraresi, Adriano / Zanchetta, Eros, 2009. « The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora », *Language Resources and Evaluation*, 43 (3), 209-226.

Bratánková, Leontýna / Štichauer, Pavel, 2011. « Italský iterativní prefix *ri-* a jeho české protějšky », in : Čermák, F. (ed.), *Korpusová lingvistika Praha 2011 – 1 InterCorp*, Praha, Nakladatelství Lidové Noviny, 136-143.

Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), 2010. *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny.

Čermák, František (ed.), 2011. *Korpusová lingvistika Praha 2011 – 1 InterCorp*, Praha, Nakladatelství Lidové noviny, 2011.

Čermák, František / Rosen, Alexandr, 2012. « The case of InterCorp, a multilingual parallel corpus », *International Journal of Corpus Linguistics*, 13, 3, 411-427. ‹http://utkl.ff.cuni.cz / ~rosen / public / 2012_intercorp_ijcl.pdf›.

Čermák, Petr, 2007. « Acerca de los corpora paralelos: el proyecto Intercorp », *Verba, Anuario Galego de Filoloxia* (Santiago de Compostela), 34, 377-382.

Čermák, Petr, 2012. « *Al + infinitivo*: estudio contrastivo checo-español », *Acta Universitatis Carolinae – Philologica*, 2 / 2009 – *Romanistica Pragensia* XVIII, 2010, 49-64.

Čermák, Petr, 2013. « Las posibilidades de estudio ofrecidas por los corpus paralelos: el caso del prefijo español *re-* », *Acta Universitatis Carolinae – Philologica*, 2 / 2013 – *Romanistica Pragensia XIX* (*Les langues romanes à la lumière des corpus linguistiques*), 2013, 123-136.

Čermák, Petr / Štichauer, Pavel, 2010. « Španělské a italské kauzativní konstrukce *hacer / fare + sloveso* a jejich české ekvivalenty », in : Čermák, F. / Kocek, J. (et al.), *Mnohojazyčný korpus Intercorp: Možnosti studia*, Praha, Nakladatelství Lidové Noviny, 70-90.

Čermáková, Anna / Fárová, Lenka, 2010. « Keywords in Harry Potter and their Czech and Finnish translation equivalents », in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 177-188.

Čermák, Petr / Nádvorníková, Olga *et al*., 2015. *Románské jazyky a čeština ve světle paralelních korpusů* [*Les Langues romanes et le tchèque à la lumière des corpus parallèles*], Praha. Karolinum.

Hajič, Jan, 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, Praha, Karolinum.

Káňa, Tomáš / Pelouśková, Hana, 2011. *Deutsch und Tschechisch im Vergleich: Korpusbasierte linguistische Studien* II, Brno, Masarykova univerzita.

Káňa, Tomáš / Pelouśková, Hana, 2009. *Deutsch und Tschechisch im Vergleich. Kopusbasierte linguistische Studien* I, Brno, Masarykova univerzita.

Kiss, Tibor / Strunk, Jan, 2006. « Unsupervised multilingual sentence boundary detection », *Computational Linguistics*, 32 (4), 485-525.

Malá, Markéta, 2013. « Translation counterparts as markers of meaning. The case of copular verbs in a parallel English–Czech corpus », *Languages in Contrast*, 13(2), 170-192.

Malá, Markéta / Šaldová, Pavlína / Klégr, Aleš, 2010. « English equivalents of the Czech preposition *v / ve* from the point of view of the 'open-choice principle' and the 'idiom principle' », in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 2010, 118-137.

Martinková, Michaela, 2012. Subject-operator inversion after sentence-initial *only* seen through its Czech equivalents, *Linguistica Pragensia*, 22 (2), 79-97.

Nádvorníková, Olga, 2010a. « Les corpus parallèles : L'Espace pour l'analyse contrastive », *Études Romanes de Brno*, 31 (1), 7-27.

Nádvorníková, Olga, 2010b. *The French* gérondif *and its Czech equivalents*, in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 2010, 83-96.

Nádvorníková, Olga, 2012. *Korpusová analýza faktorů sémantické interpretace francouzského gérondivu* [Thèse de doctorat], Filozofická fakulta Univerzity Karlovy v Praze, dir. H. Loucká.

Olohan, Maeve, 2004. *Introducing Corpora in Translation Studies*, London / New York, Routledge.

Rosen, Alexandr, 2010. « Morphological tags in parallel corpora », in : Čermák, František / Corness, Patrick / Klégr, Aleš (ed.), *InterCorp: Exploring a Multilingual Corpus*, Praha, Nakladatelství Lidové noviny, 2010, 205-234. ‹http://utkl.ff.cuni.cz / ~rosen / public / unitags_pap.pdf›.

Rosen, Alexandr, 2012. « Mediating between incompatible tagsets », in : Ahrenberg, Lars et al. (ed), *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC), Volume 10 of NEALT Proceedings Series*, Tartu, Northern European Association for Language Technology, 53-62.

Rosen, Alexandr / Vavřín, Martin, 2012. « Building a multilingual parallel corpus for human users », in : Calzolari, N. et al. (ed), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC´12)*, Instanbul, European Language Resources Association (ELRA), 2447-2452.

Rychlý, Pavel, 2007. « Manatee / Bonito – a modular corpus manager », in : *Ist Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, Masarykova univerzita, 65-70.

Svášek, Martin, 2007. *Définition, élaboration et exploitation d'un corpus parallèle bidirectionnel français – tchèque tchèque – français* [Thèse de doctorat], Filozofická fakulta Univerzity Karlovy v Praze (dir. V. Petkevič) / INALCO (LALIC – CERTAL), Paris (dir. P. Pognan).

Varga, Dániel et al., 2005. « Parallel corpora for medium density languages », in : *Proceedings of RANLP 2005*, Borovets, Bulgaria, 590-596. ‹http://www.ldc.upenn.edu / Catalog / docs / LDC2008T01 / ranlp05.pdf›.

Vavřín, Martin / Rosen, Alexandr, 2008. « InterCorp: A Multilingual Parallel Corpus Project », in : Proceedings of the International Conference Corpus Linguistics - 2008, St. Petersburg State University, 97-104. ‹http://utkl.ff.cuni.cz / ~rosen / public / 2008_intercorp_peterburg.pdf›.

Vondřička, Pavel, 2010. « TCA2 - nástroj pro zpracovávání překladových korpusů » [TCA2 - a tool for processing translation corpora], in : Čermák, F. / Kocek, J. (et al.), *Mnohojazyčný korpus Intercorp: Možnosti studia*, Praha, Nakladatelství Lidové Noviny, 225-231.

Waldenfels von, Ruprecht, 2006. « Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alingment », in : Brehmer, B. / Zdanova, V. / Zimny, R. (ed.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, *Volume 9*, München, Verlag Otto Sagner, 123-138.

Cet article a été réalisé grâce au soutien financier du projet Český národní korpus (LM2011023), *soutenu par le Ministeère de l'Éducation nationale tchèque dans le cadre de l'activité* « Projekty velkých infrastruktur pro VaVal ».

Ferdinand de Saussure e la linguistica romanza. Un'applicazione web per l'edizione elettronica dei manoscritti[1]

1. L'edizione digitale dei manoscritti di Ferdinand de Saussure

Il progetto di edizione digitale dei manoscritti di Ferdinand de Saussure (PRIN 2008), al quale hanno collaborato quattro unità (Università della Calabria, Università di Firenze, ILC-CNR Pisa e Università di Salerno), si è posto come duplice obiettivo l'accesso a un ampio corpus di manoscritti saussuriani e un superamento della pura e semplice trasposizione in formato elettronico di edizioni cartacee[2]. Il gruppo di ricerca dell'Istituto di Linguistica Computazionale (ILC-CNR, Pisa), guidato da Andrea Bozzi, ha messo a punto i seguenti strumenti:

- Thesaurus computazionale della terminologia saussuriana[3];
- Banca dati per la gestione bibliografica e catalografica;
- Descrizione ontologica delle entità fisiche (manoscritti), strutturazione esplicita della base di conoscenza della teoria linguistica saussuriana e inserimento sperimentale di un campione di istanze al fine di validare lo schema ontologico proposto;
- Applicazione web per l'analisi filologico-computazionale[4].

La piattaforma per l'analisi filologica e critico-testuale dei manoscritti è stata personalizzata e adattata alle particolari esigenze dell'editore e dello studioso saussuriano[5]. Per una prima verifica delle sue funzionalità sono stati scelti due testi che, oltre a differire dal punto di vista del contenuto, ci mettono di fronte a problemi ecdotici diversi: del primo è disponibile l'edizione critica, mentre l'altro è per la maggior parte ancora inedito. La scelta di impiegare come primo 'banco di prova' il manoscritto Ms. fr. 3955/1 (Bibliothèque de Genève, Fonds F. de Saussure), pubblicato da Maria

---

1 Il lavoro presentato è frutto di una ricerca e di un impegno comuni. Tuttavia si precisa che a Luca Pesini vanno attribuiti i §§ 1-5, ad Andrea Bozzi il § 6 e il § 10, ad Angelo Del Grosso i §§ 7-9.

2 I risultati del progetto sono esposti in Gambarara / Marchese (2013).

3 Il primo thesaurus-lessico elettronico della terminologia linguistica saussuriana, attraverso una rappresentazione strutturata che ha richiesto anzitutto la creazione di un'ontologia lessicale di dominio, è accessibile all'indirizzo ‹http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=917/vers=ing›. Su questo tema si veda il contributo di Ruimy / Piccini / Giovannetti / Bellandi (2013).

4 Cf. Murano / Pesini (2013) e Del Grosso / Marchi (2013).

5 Sui problemi della filologia digitale si veda Bozzi (in corso di stampa).

Pia Marchese nel 2002, non è casuale ed è stata dettata dalla precisa intenzione di saggiare le potenzialità del sistema e di individuarne punti di forza e limiti. Il testo di *Théorie des sonantes*, in cui Saussure raccoglie le sue critiche alla teoria di Johannes Schmidt, riveste un particolare interesse per l'indoeuropeistica ed è un ottimo esempio delle problematiche che deve affrontare lo studioso/editore degli autografi saussuriani: cancellature, aggiunte interlineari e marginali, ripensamenti ed esitazioni costellano ogni pagina del manoscritto e testimoniano il travaglio intellettuale che caratterizza in genere la produzione scientifica del maestro ginevrino. Compito dell'editore è proprio quello di far emergere l'avvicendarsi delle diverse fasi redazionali in appunti che rispecchiano un pensiero ancora *in fieri*, spesso ellittico, lacunoso e denso di aporie. Il secondo manoscritto del quale si sta approntando l'edizione sulla piattaforma filologico-computazionale è dedicato esclusivamente a questioni di linguistica romanza e di esso ci occuperemo più in dettaglio.

2. Il Ms. fr. 3956 (Bibliothèque de Genève)

Il Ms. fr. 3956, intitolato *Notes sur l'étymologie des noms de lieux de la Suisse romande et sur les patois romands et chablaisiens* (Bibliothèque de Genève, Fonds F. de Saussure) e per la maggior parte inedito, contiene diciotto *dossiers* (i fascicoli n. 2 e 3 e i fascicoli n. 6 e 7 sono riuniti in una sola busta). Dopo la morte di Saussure questi appunti, relativi a due principali filoni di ricerca (toponomastica storica della Svizzera romanza e inchieste dialettologiche condotte nei dintorni di Ginevra negli anni 1901-1903) furono consegnati a Jules Ronjat, al quale gli editori del *Cours de linguistique générale*, Charles Bally e Albert Sechehaye, affidarono l'incarico di visionare il manoscritto e di selezionare le parti che, secondo il suo autorevole giudizio, meritavano di essere pubblicate[6]. L'insigne romanista riordinò le carte e aggiunse di suo pugno chiose, commenti e critiche, rimarcando in molti casi l'indubbio interesse che questo materiale presenta per linguisti e dialettologi. Per dare un'idea del contenuto del Ms. fr. 3956 ne riportiamo qui di seguito l'indice con l'aggiunta delle osservazioni autografe di Ronjat (indicate con la sigla « J.R. »):

(1) Résumé d'une communication à la Société d'histoire et d'archéologie de Genève, 29 janvier 1903, 5 ff.
(2-3) Toponymie d'Oron, 27 ff. ; « prêt à imprimer, sauf à tenir compte des observations que j'ai notées sur fiches insérées dans cette chemise. J. R. ».
(4) Exposé général : Genthod et Ecogia, 17 ff.
(5) Carouge, Joux – Jorat – Jura, 9 ff. ; « prêt à imprimer ».
(6-7) Appendices sur Joux – Jura, 24 ff. ; « À publier (cf. mes notes ci-incluses). J. R. ».
(8) Inscription à Gex, etc. Note sur la retraite de Xénophon. Note et lettre sur un soupçon d'espionnage à Segny, 17 ff. ; « Pièces diverses me paraissant mériter publication. J. R. ».

[6] Ronjat era stato coinvolto anche nella revisione del manoscritto del *Cours de linguistique générale*, come sappiamo da una nota degli editori : « nous exprimons aussi nos plus vifs remerciements à M. Jules Ronjat, l'éminent romaniste, qui a bien voulu revoir le manuscrit avant l'impression, et dont les avis nous ont été précieux » (*CLG/de M*, Préface, 8).

(9) Patois de diverses localités du Pays de Vaud, 40 ff. ; « À verser au Glossaire ».
(10) Cahiers d'enquête sur les patois romands (et quelques-uns de Chablais), 106 ff. ; « Du plus haut intérêt. A verser dans le Glossaire où ils auront leur utilisation naturelle. J. R. »
(11) Sur le traitement de *o* et *u* latins, 18 ff. ; « intéressant, à verser au Glossaire. J. R. ».
(12) Avers et revers. Polémique avec Eugène Demole, 12 ff.
(13) Notes sur l'idiotisme *en faire à sa tête*. Empros (coupure de journal). Les armaillis des Colombettes (copie avec quelques interprétations phonétiques). « Noms de lieux genevois en *-nex* rattaché à germ. *-nâh* [étrange !] J. R. », 6 ff.
(14) Matériaux de valeur très douteuse. Genevoisismes, 12 ff. ; « Je les ai collés sur des fiches, sans quoi on aurait fini par les perdre. Colle à la gomme arabique, peut partir en mouillant. À offrir à l'œuvre du Glossaire des patois de la Suisse romande. J. R. ».
(15) Echantillons de patois fribourgeois, 1 carnet, 21 ff.
(16) Patois français, 3 ff.
(17) Noms de lieux en *-ens* et *-ingen*, 8 ff.
(18) La communication à la Société de linguistique. Note sur le patois vaudois ou fribourgeois de la Suisse romande, 18 ff. : « est faite de seconde main, principalement d'après le carnet noir *Échantillons du patois fribourgeois*…caporal Alb. Müller et accessoirement par des morceaux patois écrits dans des journaux avec des systèmes d'écriture incohérents. C'est une œuvre de débutant qui ne présente qu'un intérêt istoriq. (*sic)* en montrant que de bonne eure (*sic)* F. de S. dirigeait son attention de ce côté. Elle est fort incomplète et offre un défaut absolu de localisation. On ne sait jamais de quel patois il est question là-dedans et à chaque instant F. de S. est arrêté par le manque de renseignements sur les problèmes qui se posent. Je verserais la communication F. de S. et le carnet Müller à l'œuvre du *Glossaire*. J. R. »

Scorrendo l'indice appare evidente che il manoscritto contiene molto materiale prezioso per chi intenda approfondire un filone piuttosto trascurato del pensiero linguistico saussuriano, rimasto finora in secondo piano rispetto agli studi di ambito indoeuropeistico e agli scritti di linguistica generale. Esso conserva infatti una ricca messe di dati dialettali raccolti personalmente dallo stesso Saussure in una serie di inchieste sul campo, che copre una vasta area della Svizzera romanza, con qualche 'sconfinamento' oltre la frontiera francese. Nel Ms. fr. 3956 abbiamo attestazioni di prima mano non soltanto di singoli termini ma anche di espressioni, frasi e addirittura frammenti di conversazione spontanea, registrati dalla viva voce dei parlanti secondo un sistema di trascrizione fonetica messo a punto *ad hoc* da Saussure[7]. Salta immediatamente all'occhio il carattere contingente ed estemporaneo di questi appunti, buttati giù alla svelta e non di rado, secondo la consueta prassi scrittoria di Saussure, su supporti 'di fortuna', come le carte intestate dell'Hôtel Terminus di Losanna e dell'Hôtel de la Gare di Moudon.

[7] Il sistema di trascrizione, illustrato in Ms. fr. 3956, fascicolo 9, f. 2r, è meno raffinato rispetto a quello impiegato dai collaboratori del *GPSR* (1924, t. I, 14-17).

3. Tra linguistica storica e ricerca sul campo

Saussure intraprende una campagna di scavo linguistico che lo porta, lontano dalle austere aule universitarie, sullo stesso terreno battuto dai ricercatori del *GPSR* (*Glossaire des patois de la Suisse romande*), la grande impresa fondata nel 1899 da Louis Gauchat. Proprio nei primi anni del secolo, quando Saussure compie le sue esplorazioni nei dintorni del lago di Ginevra, i collaboratori del *GPSR* sono impegnati in un'inchiesta a tappeto sulla fonetica dei 'patois' svizzeri, che arriva a toccare circa 400 località: « [l]a première année des travaux (1899) a été consacrée à des relevés phonétiques faits dans toutes les parties de la Suisse romande, ainsi qu'au recrutement de correspondants disposés à collaborer à la grande enquête par questionnaires prévue pour les années suivantes »[8].

Ovviamente non era possibile setacciare in modo capillare l'intero territorio studiato con rilevamenti fatti di persona nelle varie località. Fino al 1911 l'impresa prosegue quindi per corrispondenza, tramite l'invio di questionari: « [n]os patois étant fort disséminés et très différents les uns des autres, nos forces n'auraient pas suffi pour en recueillir sur place les nombreuses variétés. C'est pourquoi nous avons eu recours à une enquête systématique par questionnaires, adressés à des correspondants répartis sur tout le territoire »[9]. Nell'arco di un ventennio l'archivio del *GPSR* arriva a raccogliere circa un milione e mezzo di *fiches* manoscritte, in cui confluiscono, oltre ai materiali del questionario, « le vocabulaires régionaux déjà existants, le produit des investigations faites sur le terrain par les rédacteurs et leurs auxiliaires, le dépouillement des textes anciens et modernes ». Nel 1924 esce il primo fascicolo di questa monumentale opera, ancora oggi in corso di pubblicazione, alla quale, a partire dal 1903, contribuisce anche il linguista Ernest Muret con una ricerca sui toponimi dialettali, registrati dalla viva voce dei parlanti. Questa indagine, protrattasi per decenni (in totale si contano 120000 *fiches* per circa 150000 nomi di luogo) è particolarmente preziosa poiché « n'a pas seulement complété et rectifié les données de nos meilleures cartes par le dépouillement de tous les plans cadastraux et le relevé sur place des noms, mais elle a aussi recueilli ces derniers, partout où cela était encore possible, sous la seule forme vraiment authentique, celle du patois local »[10].

Nel 1901 anche Saussure intraprende la sua personale inchiesta sui *patois* della Svizzera romanza e dell'Alta Savoia; in lui, tuttavia, non si scorge un interesse puramente dialettologico, l'intenzione di documentare il vernacolo di una minuscola comunità che traspare, ad esempio, ne *L'unité phonétique dans le patois d'une commune* di Gauchat (1905). In Saussure è sicuramente preponderante la lettura in diacronia dei dati dialettali, che servono al linguista per ricostruire etimologie e scrivere (oppure riscrivere) le regole di mutamento fonetico, sempre applicando il metodo storico-ricostruttivo. Gauchat dà invece un taglio decisamente sincronico al suo lavoro

[8] *GPSR* (1924, t. I, 6).
[9] *GPSR* (1924, t. I, 7).
[10] *GPSR* (1924, t. I, 9).

al microscopio sul dialetto di Charmey (distretto di Gruyère, Canton Friburgo) e raggiunge una finezza di analisi tale da cogliere gli aspetti di eterogeneità presenti addirittura al livello del *patois villageois*; l'opinione corrente che vedeva nel 'villaggio' un sistema linguistico uniforme e monolitico, privo di differenziazione e stratificazione interna, veniva ad essere smentita dai fatti: mettendo in luce le differenze di pronuncia rilevabili fra compaesani di età diverse, il linguista di Neuchâtel mostrava quanto fosse illusorio il tentativo di stabilire « un type charmeysan »[11]. L'interesse per la dimensione sociale del vernacolo emerge anche dagli appunti di Saussure, attento alle minime differenze di pronuncia fra gli informatori. Tuttavia nella sua rielaborazione dei dati dialettali si avverte sempre lo spirito del 'neogrammatico', che si pone come obiettivo principale l'individuazione delle leggi fonetiche e interpreta ogni fenomeno in una prospettiva eminentemente storica. Saussure compie un'autentica operazione di archeologia linguistica, volta a far luce in particolare sugli esiti delle vocali ū, ŭ, ō e del dittongo AU nelle parlate della Svizzera romanza.

L'interesse per questo problema di fonetica storica nasce da una ricerca sull'origine del toponimo *Oron*, antico centro del Canton Vaud, capoluogo dell'omonimo distretto. All'etimologia di Oron sono dedicati il secondo e il terzo fascicolo del Ms. fr. 3956, che contengono gli appunti per la prima delle tre comunicazioni tenute da Saussure alla *Société d'Histoire et d'Archéologie* di Ginevra tra il 1901 e il 1904: di questo primo contributo, intitolato « Le nom de la ville d'Oron à l'époque romaine » e letto nella seduta del 28 marzo 1901, venne pubblicato un *résumé* sul *Journal de Genève* del 7 aprile 1901 (ripubblicato in Saussure 1922, 604-605)[12]. Dovettero passare quasi vent'anni prima che Louis Gauchat mettesse mano agli appunti (27 fogli) per trarne l'unico articolo postumo di Saussure, « Le nom de la ville d'Oron à l'epoque romaine : Étude de Ferdinand de Saussure† publiée et annotée par L. Gauchat », pubblicato nel 1920.

4. Saussure etimologo

Vale la pena di riassumere nelle sue linee principali l'articolo dedicato al nome di Oron. Nell'*Itinerarium Antonini* una tappa del percorso tra Milano e Magonza è indicata col nome *Bromago* (forma all'ablativo locativo di *Bromagus*): essa è situata sulla strada che dal passo del Gran San Bernardo (*Summo Pennino*) porta a Martigny (*Octoduro*) e poi a Villeneuve (*Penne Locos*). Le *stationes* che s'incontrano

[11] Gauchat (1905, 53) : «Rigoureusement il n'y a pas d'unité dans le parler de Charmey, parce que les générations ne sont pas d'accord, et cette unité est d'autant moins une réalité que d'autres villages peuvent être arrivés au même point de l'évolution linguistique».

[12] Il Ms. fr. 3956 contiene altri quattro *dossiers* di argomento toponomastico, in cui sono raccolti gli appunti per la seconda conferenza tenuta alla *Société d'Histoire et d'Archéologie*, intitolata «Origine de quelques noms de lieux de la région genevoise» (29 gennaio 1903). Si tratta del fascicolo n. 4 su «Genthod et Ecogia» e dei n. 5, 6 e 7 su «Carouge, Joux – Jorat – Jura», che sono stati pubblicati da Arsenijevič (1998). Il breve contributo originario, apparso anch'esso in forma di *résumé* sul *Bulletin de la Société d'Histoire et d'Archéologie de Genève*, tome II, 342 è stato ripubblicato nel *Recueil* (Saussure 1920, p. 605).

dopo *Penne Locos* sono nell'ordine: *Vibisco* (Vevey), *Bromago*, *Minnodunum* (Moudon) *Aventiculum Helvetiorum* (Avenches). Fin dal XVI secolo gli eruditi locali e gli archeologi, sulla base di una semplice assonanza, avevano identificato *Bromagus* col moderno villaggio di Promasens, tra Vevey e Moudon. Saussure scarta immediatamente questa ipotesi, insostenibile dal punto di vista della fonetica storica, e spiega che la forma *Bromago* è frutto di un errore paleografico per *Uromago*, lezione tramandata da due manoscritti dell'*Itinerarium*[13]. Il toponimo celtico *Uromagus* ha significato trasparente ed è traducibile con "campo dell'uro": ūrus è il nome latino del "bue selvatico" (in francese anche *aurochs*) mentre l'elemento *-magus* indica il "campo". Nei toponimi di questo tipo, accentati sulla terzultima, le due sillabe finali erano destinate a cadere, secondo gli esempi *Noviómagus* > *Noyon*; *Tornómagus* > *Tournon; Mosómagus* > *Mouzon; Rigómagus* > *Riom; Argantómagus* > *Argenton*. Analogamente la regolare evoluzione fonetica di *Urómagus* è *Oron* (*Ourón* nel dialetto locale); questa ricostruzione è inoltre confortata da considerazioni di ordine storico: Oron-la-ville, pur essendo un villaggio di modeste dimensioni, ha mantenuto per secoli una notevole importanza come centro amministrativo; capoluogo di distretto e sede di tribunale ancora agli inizi del Novecento, in passato era stato residenza del balivo bernese, dei conti di Gruyère e ancor prima della famiglia feudale d'Oron, « [e]t ainsi en remontant de proche en proche il paraît également clair que la seigneurie d'Oron trouve la raison historique dans le vieux castellum romain »[14].

Una possibile obiezione a quest'ipotesi era rappresentata dall'attestazione del toponimo nella forma *Auronum*, menzionata per la prima volta nel 516, in un atto di donazione con cui il re burgundo Sigismondo confermava l'abbazia di San Maurizio d'Agauno nel possesso di varie località del Canton Vaud. Tuttavia Saussure, ricostruendo con l'aiuto di Victor van Berchem la storia documentaria degli *Acta concilii agaunensis*, arriva a stabilire che il testo in questione, trasmesso da copie risalenti al dodicesimo secolo, « remonte tout au plus à l'an mille, en prenant une moyenne »[15]: *Auronum* deve quindi derivare, secondo una prassi molto diffusa tra gli scribi dell'epoca, da una falsa latinizzazione di *Oron*, il cui legame coll'originario *Uromagus*, all'altezza cronologica considerata, risultava da tempo obliterato per via dell'erosione fonetica della finale *-ago*. Restava infine da risolvere una difficoltà di carattere fonetico relativa alla quantità della vocale iniziale: la presenza di ūrus "bue selvatico" come primo elemento del composto rende infatti necessario postulare un originario ūrómagus con ū, che solleva « un problème au première moment insoluble ». Proprio per venire a capo della questione e verificare quale fosse l'autentica pronuncia dialettale del toponimo studiato, Saussure decide di intraprendere la sua ricerca sul campo:

[13] La *Tabula Peutingeriana* riporta ancora un'altra variante: « *uiromagus,* faute greffée sur *vromagus*, confirmation indirecte, mais trè solide, de *uromagus* ».

[14] Saussure (1920, 7).

[15] Saussure (1920, 8).

> Ce problème m'a déterminé à faire ce par quoi on devrait toujours commencer dans les recherches de ce genre, c'est-à-dire d'aller voir tout simplement sur place comment on dit Oron dans le propre langage du pays. Non pas selon l'écriteau de chemin de fer qui indique la station d'Oron aux voyageurs, mais selon ce qui se prononce par tradition authentique dans le patois de la contrée. À ma grande surprise, j'ai constaté que sur toute l'étendue du pays, non seulement dans la vallée de la Broye, de Palézieux, jusqu'à Payerne, mais encore en-dehors, par exemple à Forel, Mézières, Siviriez, et partout sans exception on dit uniquement : *Ouron* [16].

Il problema della vocale iniziale di *Ouron* era da spiegare nel quadro del sistema fonologico dialettale, che presenta condizioni diverse rispetto al francese comune. Nel *patois* del Canton Vaud AU, ŏ, ō, ŭ protoniche confluiscono in *o* (Saussure cita i seguenti esempi: AURÍCULA > *orol'e*; NŎVÉLLU > *novi*; PLŌRÁRE > *pl'ora*; CŬBÁRE > *kova*); ū presenta invece esiti differenti in base alla posizione dell'accento: ū tonica si sviluppa in *ü* (*mü* 'muro'; *pü* 'puro'; *dü-düra* 'duro-dura'), mentre ū protonica ha come regolare succedaneo *ou* (MŪRÁLIA > *moural'e*; EPŪRÁRE > *èpoura*; fr. *durée* > *douraye*). I dati dialettali confermavano quindi l'identificazione *Uromagus-Oron*: « [i]l ne reste plus qu'à conclure: *Ouron* est la propre forme réclamée par les celtiste. Le patois vient à la rencontre de leur opinion et la confirme »[17].

5. Saussure dialettologo

L'indagine sul campo, che porta Saussure a perlustrare numerosi villaggi nei dintorni di Ginevra, nel Canton Vaud, nel Chiablese e nell'Alta Savoia, pur restando una ricerca *a latere*, in posizione ancillare rispetto al lavoro di ricostruzione storica ed etimologica che si svolge nel chiuso della biblioteca, ne costituisce sicuramente il complemento indispensabile. I dati raccolti occupano tre ricchi *dossiers* (n. 8, 9 e 10) per un totale di 163 fogli: generalmente troviamo come intestazione il nome del punto d'inchiesta, che può essere accompagnato da qualche cenno sulle prime impressioni ricevute dal visitatore al suo arrivo in paese; le notizie sugli intervistati sono di solito molto scarne e si riducono all'indicazione dell'età, del sesso o del luogo d'origine, mentre in una minoranza di casi sono registrati nome, cognome, età, luogo di nascita ed eventuali spostamenti di residenza avvenuti nel corso della vita. Suddivisi per informatore si susseguono elenchi più o meno lunghi di parole (nomi comuni e toponimi), frasi e addirittura frammenti di parlato spontaneo, trascritti a volte con accuratezza, a volte in tutta fretta. A un primo sguardo gli appunti danno nel loro insieme un'impressione di asistematicità, poiché Saussure non si serve di un questionario prestabilito. È comunque evidente che tutte le domande mirano a individuare gli esiti fonetici di AU, ŏ, ō, ŭ, ū. In un *dossier* a parte (n. 11), intitolato « Sur le traitement de *o* et *u* latins », una scelta dell'abbondante materiale accumulato nel corso delle inchieste è sistemata in una serie di tabelle a campi incrociati, che riportano nell'intestazione delle colonne i nomi delle località indagate e nell'intestazione delle righe le parole del francese comune di cui si è cercato

[16] Saussure (1920, 10).

[17] Saussure (1920, 11).

il corrispondente dialettale (per esempio *couvée, goulu, souvent, poulain* per ŭ; *oie, joie, pauvre, chou* per AU).

Per quanto concerne il metodo di rilevamento dei dati, Saussure si richiama ad accorgimenti fondamentali, ad esempio quando avverte che è indispensabile evitare domande dirette contenenti il termine di lingua, che potrebbe influenzare l'informatore nelle sue risposte:

> Les personnes qui prétendront avoir entendu d'un paysan quelconque « Oron » sont victimes d'une erreur. Il ne faut naturellement pas commencer par demander comment s'appelle *Oron*. La réponse sera sûrement le nom français. Mais il faut demander par exemple : les tours du château d'Oron, en attirant l'attention sur les tours pour être sûr que Oron suive selon la naturelle force du dialecte. Dans ce cas je défie qu'on entende jamais autre chose que : lɐ twa doʊ tsati d'ʊrõ[18].

Altre osservazioni interessanti riguardano i problemi legati alla trascrizione fonetica. Proprio negli stessi anni, tra 1902 e 1910, si pubblicavano i diciassette volumi dell'*Atlas linguistique de la France*, nato dalla collaborazione fra il linguista Jules Gilliéron, che ne fu l'ideatore e il promotore, ed Edmond Edmont, un cultore di dialetti dall'orecchio finissimo, il quale eseguì materialmente le inchieste tra il 1897 e il 1902. Il sistema adottato da Saussure differisce sia da quello noto col nome di Gilliéron-Rousselot sia da quello impiegato dal *GPSR*. Negli appunti si sottolinea l'importanza di una trascrizione che sia al contempo leggibile e priva di ambiguità:

> D'après le double principe de la lecture facile à vue d'œil et de la non-ambiguïté phonologique, je considère comme très utile dans ce patois la suppression totale des deux signes *u* et *e* qui ne peuvent engendrer que perpétuelles confusions pour le lecteur, sinon même pour le notateur comme j'oserai dire que c'est le cas au moins en ce qui concerne *e*, et même pour le plus attentif des notateurs s'il n'a pas rompu une fois pour toutes avec cette malheureuse lettre *e*. Quand une lettre a 2 significations possibles dont aucune n'est tout à fait ineffaçable dans l'usage, il ne faut pas user de compromis avec ce signe, mais le retrancher radicalement de notre alphabet phonétique[19].

Saussure si sofferma spesso a commentare le abitudini di pronuncia degli informatori, registrando meticolosamente le realizzazioni individuali di certi fonemi. Ad esempio nel verbale dell'inchiesta svolta a Grande-Rive, località situata sulla riva del Lago Lemano nei pressi di Évian-les-Bains (Alta Savoia), sono segnalate le differenze fra tre parlanti nella conservazione dei timbri originari delle vocali atone finali[20]:

> Dans la prononciation de Grande Rive pas de peine à distinguer chez le premier individu *a* de *ɛ*, *e* dans les finales atones. Mais chez les 2 autres, un *a* final[e] atone est une chose qu'ils *perçoivent* comme *a* (en s'opposant vivement à ce qu'on écrive autre chose) mais qui est très

[18] La nota (Ms. fr. 3956, 11, 10v) è riportata nell'articolo postumo pubblicato a cura di Gauchat (cf. Saussure 1920, 10).

[19] Ms. fr. 3956, 9, 12r.

[20] La presenza di vocali atone finali di suono pieno, insieme alla conservazione del ritmo parossitonico, è un tratto distintivo del francoprovenzale di contro alla forte tendenza del francese all'ossitonia generalizzata.

semblable pour moi à *ɛ* ou *e*, et cependant à l'épreuve on voit que leur différence est réelle, car ils ne permettent ni *a* au pluriel *fɛne*, ni *e* au sing. *fɛnɛ* qui sonne presque idem[21].

A Montagny-près-Yverdon, villaggio del Canton Vaud, Saussure intervista una coppia di anziani coniugi, « M. François Bourgeois, 86 ans, de Montagny (né 1816) » e sua moglie, rilevando esiti diversi del dittongamento di ŏ, ĕ nei continuatori di MŎLA(M) e PĔTRA(M):

> Observations sur Bourgeois (Montagny) - voir ci-dessus. […] Dans un ou 2 mots il peut y avoir mélange avec les réponses de la femme du dit F. Bourgeois, native de Chavornay (mais ayant passé sa vie à Montagny.) Dans le mot pour *meule de moulin*, il dit *mōla* (*ō* ouvert), et elle dit : *ma^{o}la*. Il dit *pyo̯ra* pour *pierre*. Elle dit *pyɛra*[22].

Nel corso delle sue inchieste Saussure non trascura i giudizi metalinguistici dei parlanti, che possono riguardare la percezione dei confini dialettali[23] oppure la presenza di varianti fonetiche connotate dal punto di vista sociolinguistico. Per esempio un informatore sessantacinquenne di Grande Rive riferisce che nella stessa località coesistono due pronunce distinte della parola *verre*: « *õ vɐ̯re* un verre dans la partie de Grande Rive du côté d'Evian - *õ vaire* dans la partie Est du même village ». Saussure registra questo fenomeno di microvariazione e aggiunge: « [p]as improbable, car tous les pêcheurs sont du côté Est, et le village a bien 250 mètres de longuer »[24].

Altro esempio di microvariazione è quello tra due pronunce del toponimo *Moudon*, una innovativa con vocale atona nasalizzata, l'altra conservativa con vocale atona dittongata. Questo fatto, legato all'appartenenza generazionale, risulta evidente alla consapevolezza metalinguistica di un intervistato di Promasens, nel Canton Friburgo, ed è confermato dallo stesso Saussure:

> L'homme de Promasens (originaire de Morlens) qui dit que les jeunes, ou ceux qui ont appris après coup le patois, disent *Moõdõ* au lieu de *Moʊ̯dõ* paraît être dans le vrai. Et c'est général, même hors du nom de Moudon. Mais je croirais que c'est une transformation en

---

21 Ms. fr. 3956, 10, 2v.

22 Ms. fr. 3956, 10, 58v. Saussure raccoglie dall'intervistato varie parole e frasi volte a verificare gli esiti fonetici delle vocali velari latine: « un bon roti de mouton *õ^{m} bõ r^{o}ti dɛ mo̯tõ* »; « ma femme est allée à Yverdon acheter quelque chose pour le ménage *ma fɐ̯na ɛzalāya a övérdõ (ou ɔ̃vérdõ) atθetā o̯kyɛ po lo mɛnādzo* »; « Cet homme est-il de Moudon ou bien d'Oron *sɐ̯ omo…dɛ Moʊ̯dõ o bɛ̃ d'ʊ̄rõ* »; « oie oison *dɛz ʊ̄yɛ* ». Inoltre vengono trascritte le forme dialettali dei toponimi Jura-Jorat-Joux: « Le Jura *lo Džüra* - le Jorat *Dzora* - la vallée de Joux *la vală̆ dɛ Dzŏ - la joux — ?- na dzorăt^{a}a* déscente de bois sur les pentes » (Ms. fr. 3956, 10, 58r). Anche la moglie di Bourgeois si rivela una preziosa fonte di informazioni : « Je désespérais de trouver la forme patoise de moulin qui est partout (sur Fribourg comme sur Vaud) *moʊ̯lɛ̃*, quand une bonne femme de Montagny (v. plus haut) m'a dit *mo̯yɛ̃*, forme dont je suis s'autant plus sûr qu'elle ne l'a pas prononcée sur ma question, mais à propos de *farine* etc. » (Ms. fr. 3956, 10, 74v).

23 Come nel caso di un informatore di Boëge, residente a Evian-les-Bains « depuis longtemps », il quale « [d]it qu'on se comprend jusqu'à Vienne en Dauphiné, et jusqu'au Piage près de Vienne mais qu'au Piage existerait une limite d'incompréhensibilité » (Ms. fr. 3956, 10, 8r).

24 Non avendo verificato di persona questo dato, Saussure adotta un atteggiamento prudenziale e annota « à ce qu'il prétend » (Ms. fr. 3956, 10, 2r).

> cours même hors de l'influence française ; on pourrait croire que c'est l'influence française parce que toute *diphtongue* est contraire au français, tandis que presque toute voyelle nasale est diphtonguée chez le Vaudois qui parla français : de sorte que celui qui a l'habitude du français répugne à *oʊ* mais pas à *oõ*. Cependant on voit même dans le canton de Fribourg à des endroits où tout le monde parle patois le *õ* gagner chez les jeunes : ce qui est même précieux pour distinguer dans certains cas *aʊ* de *ʊ*, *o* dans les protoniques : le *aʊ* faisant *õ* (ou *aõ*)[25].

Segnaliamo infine un curioso incidente che vide coinvolto Saussure durante le sue escursioni dialettologiche e del quale resta traccia nelle carte inedite del Ms. fr. 3956. Egli stesso - non è chiaro se più indispettito o divertito - riferisce di essere stato vittima di un imbarazzante equivoco nel villaggio francese di Segny, situato ai piedi del massiccio del Jura, a poca distanza dalla frontiera col Canton Ginevra. Agli occhi degli abitanti il fatto che un forestiero si aggirasse per le vie del paese, facendo strane domande e prendendo appunti, doveva essere sembrato quantomeno insolito. La presenza dell''intruso' aveva finito per destare un sospetto e un allarme tali, che l'innocuo linguista fu scambiato addirittura per una spia[26] :

> 20 Nov. 1901
>
> À Segny, recherchant le patois, je fus accusé d'espionnage par le cantonnier de l'endroit. Bientôt rumeur dans toute l'auberge, et le cantonnier qui m'avait livré les quelques mots ci-dessus répétait sans cesse : « Je me suis laissé entraîner…je me suis laissé entraîner - comme ayant livré la patrie sans le vouloir ». Ils se retirent presque menaçants. Êtes-vous *autorisé*, telle est la grande question. « Vous prenez des paysans tout à fait pour des niais, et puis si le garde champêtre était là je vous sommerais de montrer vos papiers »[27].

Nello stesso *dossier* si conserva anche il brogliaccio di una lettera indirizzata al sottoprefetto di Gex in cui Saussure esprime il proprio disappunto per l'accoglienza ricevuta a Segny. A differenza delle righe precedenti, nelle quali l'episodio è narrato con un sorriso ironico, la missiva, lasciata incompiuta e verosimilmente mai spedita, è scritta in pieno stile 'burocratico' e in un tono formale e compassato. Ne riportiamo parte del testo, che risulta in più punti lacunoso e necessita di alcune integrazioni:

> Dans le but scientifique de connaître les patois du pays de Gex, qui sont un complément des patois romands suisses actuellement objet d'un ouvrage d'ensemble, je me suis rendu [à Prévessin et à] Segny, pensant naturellement faire cette enquête comme je l'avais faite depuis des années soit dans le Canton de Vaud soit dans la Haute Savoie en rencontrant de la part des habitants une prévenance[28]. [J'ai été] l'objet dans les localités de votre ressort d'une régulière manque de complaisance approchant de l'hostilité, et je me suis rendu compte à la fin que cette hostilité provenait d'une suspicion d'espionnage très pardonnable de la part

---

25 Ms. fr. 3956, 10, 68.

26 Jules Ronjat, nei suoi commenti critici, osserva: « note et lettre sur un soupçon d'espionnage à Segny […] non seulement curieuse comme détail de mœur, mais importante par l'annonce d'un ouvrage d'ensemble (ligne 5 de la lettre) ».

27 Ms. fr. 3956, 9, 7v.

28 A questo punto è leggibile una frase cancellata: « À mon étonnement, j'ai été pris … pour un espion ! ». Per dare senso alla frase successiva è necessaria l'integrazione [J'ai été]. Nel seguito della lettera il testo è lasciato in sospeso in diversi punti.

> de villageois qui ne comprennent rien à ces interrogations linguistiques. Comme la question régulière était de savoir si j'étais muni […] Je n'ai pas tardé à m'apercevoir qu'un véritable obstacle que je n'avais jamais prévu jusqu'alors ni dans mes […] s'élevait […] Cet obstacle, si absurde qu'il semble, est celui du soupçon d'espionnage. Je n'en veux nullement aux populations très simples qui […][29].

6. Utilizzo dell'applicazione per l'edizione del Ms. fr. 3956

L'edizione digitale, rispetto a quella cartacea, apporta diversi vantaggi allo studio dei manoscritti saussuriani: la base di dati, immediatamente accessibile, può infatti essere continuamente aggiornata e resa più facilmente fruibile attraverso la tecnologia informatica. La piattaforma rende possibili le seguenti operazioni:

(1) visualizzazione simultanea della riproduzione (facsimile) del manoscritto e della relativa trascrizione;
(2) creazione di un apparato critico-filologico strutturato in molteplici livelli di annotazione;
(3) elaborazione di indici lessicali che, attraverso ricerche per lingua o per forma, permettano un'indagine approfondita sui contenuti e sul metalinguaggio saussuriano;
(4) ricerche multilivello, i.e., testo, note critico-filologiche, paratesto, extratesto.

Per quanto concerne il primo aspetto è superfluo sottolineare che per lo studioso è fondamentale disporre, oltre che delle informazioni essenziali relative al manoscritto (collocazione, descrizione materiale, datazione, contenuto etc.), di riproduzioni ad alta risoluzione che, da un lato, agevolino la trascrizione (spesso complicata dalla presenza di simboli o disegni) per chi si occupa dell'edizione e, dall'altro, consentano a chi voglia semplicemente consultare il testo un immediato confronto fra trascrizione e foglio originale.

Diversamente dall'edizione a stampa, in cui vari tipi di informazione (note filologiche vere e proprie, commenti di carattere linguistico o ermeneutico, rimandi ad altri testi) si trovano generalmente relegati nelle note a piè di pagina senza una strutturazione interna, nell'edizione digitale è possibile distinguere diverse tipologie di annotazione e creare un apparato critico-filologico articolato su più livelli. Le annotazioni sono state suddivise in:

(1) *notes théoriques*, per i commenti critici al testo (di tipo linguistico o filologico);
(2) *personne*, per le persone citate;
(3) *bibliographie*, per le opere citate;
(4) *glossaire*, contenente i termini che l'editore considera importanti;
(5) *notes critiques au texte par l'auteur*, in cui vengono segnalati gli interventi di Saussure sul testo, suddivise in:
    (a) *biffure* per la semplice cancellatura;
    (b) *ajout* per le aggiunte marginali o interlineari;
    (c) *variante* per varianti di parole o di frasi nel rigo;
    (d) *restitution* per il ripristino di porzioni di testo prima cancellate;

[29] Ms. fr. 3956, 9, 8r/v.

(6) *notes critiques au texte par l'éditeur*, per gli interventi operati dall'editore, suddivise in:
   (a) *correction* per gli emendamenti;
   (b) *intégration* per l'integrazione di lettere o parole;
   (c) *développement des abréviations* per lo scioglimento delle abbreviazioni (l'indice lessicale terrà conto anche dei casi in cui Saussure ricorre alla forma abbreviata di una parola);
(7) *notes supplémentaires*, categoria a disposizione dell'editore per aggiungere informazioni libere.

Considerato che Saussure, in manoscritti come Ms. fr. 3955/1 (*Théorie des sonantes*), lavora su più lingue col metodo comparativo-ricostruttivo, per lo studioso può rivelarsi particolarmente utile un indice lessicale preciso e diversificato. Questo strumento consente di individuare tutte le parole di una determinata lingua presenti all'interno del manoscritto (per *Théorie des sonantes* le lingue indicizzate, oltre ovviamente al francese, sono greco, latino, tedesco e sanscrito, con l'aggiunta delle forme ricostruite dell'indoeuropeo); inoltre, tramite la ricerca "per forma", si ottiene il totale delle occorrenze di una singola parola, della quale è possibile visualizzare il contesto.

7. Elaborazione degli indici lessicali

L'indicizzazione del Ms. fr. 3955/1 ha consentito di generare due tipi di risorse indicate come indice 'backend' e indice 'frontend'. Il primo tipo definisce una struttura ausiliaria attraverso la quale il programma applicativo è in grado, tramite un'interrogazione ('query'), di reperire con rapidità ed efficienza l'insieme dei documenti che rispondono ai criteri di ricerca. Il secondo tipo, invece, è una lista di parole suddivise sulla base di uno specifico criterio (per esempio la lista dei lemmi, dei termini di dominio, delle parole per lingua etc.), grazie alla quale l'utente dell'applicazione riesce ad avere un quadro sinottico del lessico, ad ottenere le concordanze delle parole usate nel manoscritto e a visualizzare i luoghi in cui occorre il termine cercato.

Gli indici 'frontend' elaborati nel corso del progetto, suddivisi per lingua e ordinati alfabeticamente secondo un criterio lessicografico, comprendono tutte le forme attestate e corredate col relativo numero di occorrenze. In una prima fase il sistema ha segmentato ciascuna sequenza di caratteri in 'token', che sono stati analizzati singolarmente e classificati secondo la presunta lingua di appartenenza. Questo processo, eseguito con tecniche semi-automatiche, ha generato per ogni lingua rilevata nel documento un gruppo di parole ordinate alfabeticamente. L'identificazione di una parola e il suo corretto inserimento in un determinato gruppo è riconducibile ad un problema di classificazione, dove date n risorse e m etichette ('labels'), si definiscono algoritmi e modelli allo scopo di assegnare a ciascuna delle n risorse l'etichetta corretta tra le m disponibili. Nella classificazione di una risorsa possono essere impiegate diverse strategie: alcune prevedono l'uso di 'thesauri' e vocabolari, altre ricorrono a modelli statistici in base ai quali si stabilisce la probabilità che una data risorsa appartenga ad un particolare gruppo. Nel progetto di edizione digitale dei manoscritti

saussuriani la creazione degli indici non si è avvalsa di metodi statistici a causa della mancanza di risorse annotate per questo specifico scopo; sono stati impiegati, invece, vocabolari ed euristiche che hanno permesso di identificare la lingua di ciascuna parola del testo. Il processo si è articolato essenzialmente in due operazioni: l'analisi del 'range' Unicode dei caratteri per l'assegnazione dell'alfabeto di appartenenza e la richiesta automatica ad un dizionario online per l'indicazione della lingua[30]. Infine si è resa necessaria una verifica manuale sia per correggere l'errata attribuzione della lingua sia per attribuirne una in caso di fallimento della procedura. In futuro si prevede di dotare il sistema di un lemmatizzatore, che permetta di raffinare le ricerche sul testo. Attraverso questa estensione si possono accomunare tutte le forme flesse di una parola ed elaborare indici più precisi, che tengano conto dei composti, delle polirematiche e di sintagmi rilevanti dal punto di vista del metalinguaggio saussuriano.

8. Trattamento delle immagini

Ogni foglio del manoscritto, estratto ed analizzato nel lavoro di manipolazione delle fonti testuali, è accompagnato dalla relativa immagine della risorsa originale. Allo stato attuale le immagini sono memorizzate all'interno del contesto dell'applicazione web ma è prevista l'adozione di un'apposita piattaforma capace di velocizzarne la gestione e il trasferimento ad alta risoluzione. Le riproduzioni fotografiche, consegnate in forma 'Tagged Image File Format' (TIFF) con una risoluzione sia verticale che orizzontale di 400 dpi e una dimensione media di 50 megabyte per immagine, sono state sottoposte ad un'opportuna riconversione e riduzione della risoluzione in un formato più adatto al trasferimento web. Ciò ha permesso di ridurre la dimensione delle immagini di due ordini di grandezza, abbattendo il tempo di trasferimento dall'ordine dei minuti (2-3 minuti circa con connessioni ADSL) all'ordine dei secondi (1-2 secondi circa), senza, per questo, perdere la leggibilità del facsimile.

9. Tecnologie impiegate e architettura del sistema

I problemi affrontati nel corso del progetto hanno richiesto la conoscenza di un vasto sistema di tecniche, tecnologie, 'framework' e librerie, legate strettamente al linguaggio di marcatura XML e al paradigma object oriented di sviluppo tramite Java. Un sistema di gestione XML garantisce flessibilità nella costruzione e nello scambio delle risorse primarie, sebbene le prestazioni siano inferiori rispetto ad un sistema di gestione dei dati di tipo relazionale. Le risorse sono memorizzate e conservate all'interno di un sistema XML nativo (eXist-db), il quale permette da un lato la manipolazione e l'interrogazione di documenti strutturati e dall'altro una facile integrazione con la libreria software, sviluppata da Apache, utilizzata per l'indicizzazione dei dati testuali (Lucene).

[30] Si è scelto di ricorrere a *Wiktionary* non solo per la sua accessibilità in rete e per il fatto che si tratta di una risorsa compilata manualmente ma anche per la disponibilità di un'interfaccia di interrogazione JSON ideale per la creazione di un modulo di recupero dei dati rapido e agile.

Il 'framework web' adottato risponde alle ultime specifiche dello standard JEE 6 (JSR-316) e le tecnologie da esso proposte sono state utilizzate per implementare l'ambiente di filologia digitale e computazionale: le linee-guida e le specifiche JSF2, le Facelets come tecnologia per la costruzione delle interfacce e i Managed Bean come strumento programmatico di controllo.

Al momento l'architettura sviluppata non supporta le specifiche Entity Java Bean (EJB) e Context and Dependency Injection (CDI) poiché l'approccio implementato per la persistenza dei dati non risulta adeguato al modello rappresentabile attraverso di essi. In fine, la libreria di componenti grafici e di interazione Primefaces è stata impiegata come estensione delle possibilità JSF di base, grazie alla natura Ajax e jQuery che implementa nativamente

10. Futuri sviluppi del progetto

Il lavoro in corso sul Ms. fr. 3956 renderà possibile un ulteriore sviluppo della piattaforma web nelle seguenti direzioni:

(1) Collegamento tra gli indici e il thesaurus del lessico saussuriano;
(2) Gestione avanzata delle riproduzioni digitali tramite server ottimizzati per la persistenza delle immagini (IIP);
(3) Integrazione di un modulo per la gestione di Open Linked Data;
(4) Associazione tra porzioni di testo e porzioni di immagine;
(5) Interventi di miglioramento dell'interfaccia a disposizione dell'utente (per esempio attivando l'evidenziazione dei risultati della ricerca sia sul testo che sulle immagini);
(6) Perfezionamento del modulo di ricerca con l'aggiunta di informazioni morfosintattiche o riguardanti la distanza tra le parole e il grado di affidabilità del testo ricostruito.
(7) Ampliamento delle categorie comprese negli indici e creazione di moduli per la lemmatizzazione.

Per lo studio del Ms. fr. 3956 le estensioni a cui si accenna in (7) rivestono un particolare interesse. Data la presenza di una copiosa messe di termini dialettali (in trascrizione fonetica o ortografica) sembra opportuno aggiungere al novero delle lingue (che per ora comprende soltanto francese, tedesco, greco, latino, sanscrito e termini ricostruiti dell'indoeuropeo) l'etichetta 'terme dialectale', che potrebbe articolarsi al suo interno su base geografica a vari livelli (in raggruppamenti più vasti come 'patois vaudois', 'patois fribourgeois', 'patois chablaisien' etc. oppure per singoli punti d'inchiesta). Inoltre, considerando che il manoscritto contiene centinaia di nomi di luogo, la presenza di una categoria 'toponyme' renderebbe possibili indagini mirate sulla toponomastica storica della Svizzera romanza. Ovviamente i parametri di ricerca devono potersi incrociare fra loro: nel caso di Oron le attestazioni del toponimo nella forma del francese comune saranno indicizzate sia sotto 'français' che sotto 'toponyme'; analogamente le varianti dialettali, in trascrizione fonetica (*ʊrõ*) e ortografica (*Ouron, Ourón*), rientreranno sia nella categoria 'terme dialectale' che in quella 'toponyme'; infine anche le forme latine (*Uromagus*, *Urómagus, Auronum* e le

variae lectiones come *uiromagus, bromago*) e gli etimi (ūrómagus)[31] riceveranno una duplice marcatura e compariranno sia nell'indice delle parole latine sia nell'indice toponomastico.

Scuola Normale Superiore di Pisa — Luca PESINI
Istituto di Linguistica Computazionale-CNR, Pisa — Andrea BOZZI
Istituto di Linguistica Computazionale-CNR, Pisa — Angelo Mario DEL GROSSO

Bibliografia

Arsenijevič, Mirolad, 1998. «Manuscrit inédit de Ferdinand de Saussure à propos des noms de Genthod, Ecogia, Carouge et Jura», *CFS* 51, 275-287.

Bozzi, Andrea (in corso di stampa). «La filologia computazionale per l'edizione elettronica dei documenti digitali: il modello e il metodo di CoPhi», in Morace, Aldo Maria / Caocci, Duilio (ed.), *La metamorfosi digitale. Scritture, archivi, filologia*, Pisa, ETS.

CLG/de M = *Saussure, Ferdinand de, Cours de linguistique générale* [Lausanne, 1916], ed. Tullio de Mauro, Paris, 1973.

Del Grosso, Angelo Mario / Marchi, Simone, 2013. «Una applicazione web per la filologia computazionale: un esperimento su alcuni scritti autografi di Ferdinand de Saussure», in: Gambarara, Daniele / Marchese, Maria Pia (ed.), *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, Alessandria, Edizioni dell'Orso.

Engler, Rudolf, 1980. «Linguistique 1908: un débat-clef de linguistique géographique et une question de sources saussuriennes», in : Koerner, Konrad (ed.), *Progress in Linguistic Historiography*, Amsterdam, John Benjamins, 257-270.

Engler, Rudolf, 2000. «La géographie linguistique», in : Auroux, Sylvain (ed.), *Histoire des idées linguistiques*, Sprimont, Mardaga, vol. III (L'hégémonie du comparatisme), 239-252.

Frýba-Reber, Anne-Marguerite / Chambon, Jean-Pierre, 1995-1996. «Lettre et fragments inédits de Jules Ronjat adressés à Charles Bally (1912-1918)», *CFS* 49, 9-63.

Gambarara, Daniele / Marchese, Maria Pia (ed.), 2013. *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, Alessandria, Edizioni dell'Orso.

Gauchat, Louis, 1905. *L'unité phonétique dans le patois d'une commune*, Halle, Niemeyer.

GPSR = Gauchat, Louis et al., 1924ss. *Glossaire des patois de la Suisse romande*, Neuchâtel / Paris.

Joseph, John E., 2012. *Saussure*, Oxford, Oxford University Press.

Murano, Francesca / Pesini, Luca, 2013. «L'edizione digitale dei manoscritti di Ferdinand de Saussure: l'analisi dei requisiti e il caso di Théorie des sonantes», in: Gambarara, Daniele / Marchese, Maria Pia (ed.), *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, Alessandria, Edizioni dell'Orso.

[31] L'etichetta 'terme reconstruit', sebbene a rigore dovrebbe applicarsi anche agli etimi latini, è riservata ai termini ricostruiti dell'indoeuropeo.

Ruimy, Nilda / Piccini, Silvia / Giovannetti, Emiliano / Bellandi, Andrea, 2013. «Lessicografia computazionale e terminologia saussuriana», in: Gambarara, Daniele / Marchese, Maria Pia (ed.), *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, Alessandria, Edizioni dell'Orso.

Saussure, Ferdinand de, 1920. «Le nom de la ville d'Oron à l'epoque romaine : Étude de Ferdinand de Saussure† publiée et annotée par L. Gauchat», *Anzeiger für schweizerische Geschichte/Indicatore di storia svizzera/Indicateur d'histoire suisse* 18, 1-11.

Saussure, Ferdinand de, 1922. *Recueil des publications scientifiques de Ferdinand de Saussure*, eds. Bally, Charles et Gautier, Léopold, Genève, Payot-Droz.

Saussure, Ferdinand de, 2002. *Théorie des sonantes. Il manoscritto di Ginevra BPU Ms. fr. 3955/1*, ed. Maria Pia Marchese, Padova, Unipress, Quaderni del Dipartimento di Linguistica dell'Università di Firenze, Studi 5.

Le lexique électronique de la terminologie de Ferdinand de Saussure : une première

1. Introduction

La terminologie de Ferdinand de Saussure a depuis toujours fait l'objet de nombreuses études. Parmi les plus remarquables, il convient de citer le petit lexique contenu dans l'ouvrage exégétique de Godel (1957) « Les sources manuscrites du cours de linguistique générale », ainsi que l'étude la plus complète à ce jour, le « Lexique de la terminologie saussurienne » de Engler (1968) dans lequel sont définis, et souvent commentés, 441 termes clés de la pensée saussurienne. Concernant la terminologie, seules des questions ponctuelles ont été traitées depuis lors. C'est pourquoi, à la lumière des acquisitions théoriques liées à la découverte des manuscrits de l'Orangerie, il a semblé opportun d'actualiser ces études. Ainsi, en mettant à profit les potentialités offertes par l'informatique dans le domaine des humanités et l'état de l'art de la Lexicographie Computationnelle, le premier lexique électronique de la terminologie linguistique saussurienne a vu le jour.

2. Un modèle ontologique

L'utilisation des nouvelles technologies de l'information dans le domaine des sciences humaines a permis de stocker d'énormes quantités de données linguistiques qui ne se prêtent cependant à être exploitées et partagées qu'au prix d'une organisation préalable et d'une représentation rigoureusement structurée et formalisée. Ceci explique l'intérêt croissant que suscitent, depuis près de trente ans, la définition et la formalisation d'ontologies lexicales.

Pour le terme 'ontologie'[1], nous adopterons ici la définition de Studer (1998) qui englobe celles de Gruber (1993) et de Borst (1997) « An ontology is a formal, explicit specification of a shared conceptualization »[2]. Les ontologies offrent en effet une modélisation des connaissances : représentation formalisée et structurée des concepts caractérisant un domaine, de leurs propriétés et restrictions, de leurs relations, et des instances de ces concepts. À partir de ces données, elles permettent la formulation

---

1 Le terme 'ontologie', né dans le domaine de la philosophie, y est défini comme l'étude des propriétés les plus générales de l'être.

2 « Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée ».

de requêtes, la réalisation d'inférences, ainsi qu'une navigation intelligente dans les bases de données textuelles annotées par leurs concepts.

Cet aspect de support à la recherche d'information conceptuelle a motivé notre choix d'un modèle basé sur une ontologie lexicale pour la construction de ce lexique. L'adaptation d'un modèle existant nous est d'autre part apparue comme un choix judicieux, tant en termes d'économie de temps et de coût qu'en terme de qualité des résultats. Après un examen attentif des principaux modèles lexicaux à structure ontologique, conçus pour le développement de grandes ressources lexicales informatisées, nous avons opté pour le modèle SIMPLE (Lenci, 2000).

SIMPLE, inspiré du modèle WordNet et plus tard inspirateur du standard ISO LMF, est un modèle sémantique basé sur les principes fondamentaux de la théorie du Lexique Génératif (Pustejovsky, 1995). Il a permis la création de 12 lexiques sémantiques pour autant de langues européennes (Ruimy, 2003). Ce modèle offre un réseau conceptuel articulé en 157 classes sémantiques, reliées par des liens hiérarchiques et non hiérarchiques. Chaque sens d'un terme y est décrit au moyen de relations et de traits sémantiques ; l'information relationnelle, très riche et en grande partie fondée sur une relecture et extension de la structure des *qualia*[3], incorpore également les liens synonymiques, dérivationnels et de polysémie logique. La richesse expressive de ce modèle, la finesse de ses descriptions, son niveau de granularité descriptive et sa flexibilité nous ont ainsi incités à entreprendre sa customisation en vue de construire la ressource lexicale que nous présentons ici.

3. Le modèle Simple_FdS[4]

3.1. L'ontologie de domaine

La connaissance spécifique de notre domaine d'intérêt a été modélisée dans l'ontologie de domaine Simple_*FdS,* en fonction de son objectif, de la typologie des utilisateurs, des types de recherche qui seront effectuées et des réponses attendues ; ceci, conformément aux principes de conception de l'ontologie générale SIMPLE.

Il existe différentes méthodologies de construction manuelle d'une ontologie[5]. L'approche 'middle-out' ou centrifuge, qui a présidé à la création de cette ontologie, nécessite l'identification des concepts essentiels à la connaissance du domaine, suivie

[3] La structure des *qualia* est composée de quatre rôles permettant respectivement de situer l'entité dénotée parmi d'autres (rôle formel), d'en indiquer la fonction (rôle télique), l'origine (rôle agentif) et la composition (rôle constitutif).

[4] Dans les différentes sous-sections, nous ne fournirons que quelques éléments concernant l'ontologie, les relations sémantiques et les propriétés des termes du lexique. Une description complète est consultable à l'adresse : ‹www.ilc.cnr.it/saussure_prg/papers/Specifications_Linguistiques_Lexique.pdf›.

[5] L'approche 'top-down' ou descendante qui opère par spécialisation des concepts les plus généraux ; l'approche 'bottom-up' ou ascendante où l'on procède à une généralisation des concepts les plus spécifiques ; l'approche 'middle-out' ou centrifuge.

de la recherche de concepts plus généraux et plus spécifiques en vue de la création du schéma conceptuel global. Notre point de départ a consisté en un noyau de termes significatifs, extraits des vocabulaires définis par Godel et Engler, d'où ont émergé les concepts les plus représentatifs de la pensée de Saussure. Une fois structurés en liens hiérarchiques et non hiérarchiques, ces concepts, situés à un niveau intermédiaire du schéma ontologique, ont été généralisés et spécialisés. Parallèlement, la structure interne des classes conceptuelles ainsi que les propriétés et restrictions des relations les unissant ont été définies : propriétés obligatoires ou optionnelles ; valeurs autorisées et cardinalité.

L'ontologie Simple_*FdS* est composée de 43 concepts hiérarchisés, répartis sur 4 niveaux de profondeur (cf. fig. 1).

À son sommet, et regroupés sous une classe nommée 'Thing', racine de la hiérarchie, figurent 4 classes sémantiques qui reflètent les 4 rôles de la structure des qualia, conformément au fondement théorique du modèle lexical. Il s'agit des classes 'Entity' (rôle formel), pour les entités dont la nature concrète ou abstraite n'est pas déterminée (*entité*), 'Telic', 'Agentive' et 'Constitutive'. Ces trois dernières classes permettent de regrouper des entités dont l'hyperonyme ne peut être déterminé et qui instancient principalement une dimension de sens télique (*objet_linguistique*), agentive (*origine_du_langage*) ou constitutive (*système*).

Seules quelques-unes des classes conceptuelles de l'ontologie seront à présent brièvement décrites.

Dans la classe 'Convention' sont regroupées les lois, les conventions en général, mais aussi celles qui régissent les phénomènes linguistiques (*Lautverschiebung*) ; 'Domain_of_Knowledge' héberge les différents sous-domaines (*signologie*). La classe 'Property', où sont regroupées les propriétés générales (*faculté*) ainsi que les propriétés plus spécifiques des éléments linguistiques (*arbitraire*), a pour sous-hiérarchie un ensemble de classes à usage exclusif des adjectifs, permettant d'en représenter les propriétés physiques, psychologiques, relationnelles et temporelles. Les activités mentales sont rassemblées dans trois classes : 'Cognitive_Event' (*abstraction*) et ses sous-classes : 'Cause_Change_Cognitive_Event', pour les activités mentales impliquant un changement et un résultat (*analyse_subjective*) et 'Mental_Creation' pour les processus mentaux de création (*paraplasme*). 'Speech_Act' encode les termes liés à l'acte de la parole (*circuit_de_la_parole*). 'State' rassemble les états (*synchronie*) et subsume les classes 'Relational_State', pour les termes dénotant une relation stative (*ablaut*) et 'Identificational_State', pour une relation d'identité ou d'opposition (*homorganie*) ; tandis que les instances de 'Change_of_State' sont des termes exprimant des changements d'état (*diachronie*). 'Representation', classe multidimensionnelle qui incorpore une dimension agentive et une dimension télique, regroupe des entités ayant une fonction évocatrice (*carré_linguistique*). 'Linguistic_System' rassemble des termes dénotant la langue en tant que code et en tant qu'idiome, ainsi que les différentes langues (*sanscrit*). Les instances de 'Relational_Entity' sont, quant à elles, des termes clés de la pensée saussurienne qui dénotent toujours des entités relationnelles (*terme*).

La classe ‘Linguistic_Entity’ réunit des termes généraux concernant les unités linguistiques soit composites soit irréductibles résultant d’un processus de délimitation ou d’articulation (*unité*). Elle subsume six classes conceptuelles caractéristiques du domaine, parmi lesquelles ‘Basic_Element’, qui regroupe des termes dénotant les éléments irréductibles des différents niveaux de description linguistique (*consonne*) ; ‘Complex_Element’, pour les éléments composites des différents niveaux de description linguistique (*syntagme*) ; enfin deux classes particulièrement représentatives de la pensée de Saussure : ‘Concrete_Element’, dont les instances sont des termes dénotant la partie perceptible, concrète du signe linguistique (*figure_vocale*) et ‘Mental_Element’, qui regroupe les termes dénotant la partie conceptuelle, psychique, mentale du signe linguistique (*valeur_significative*).

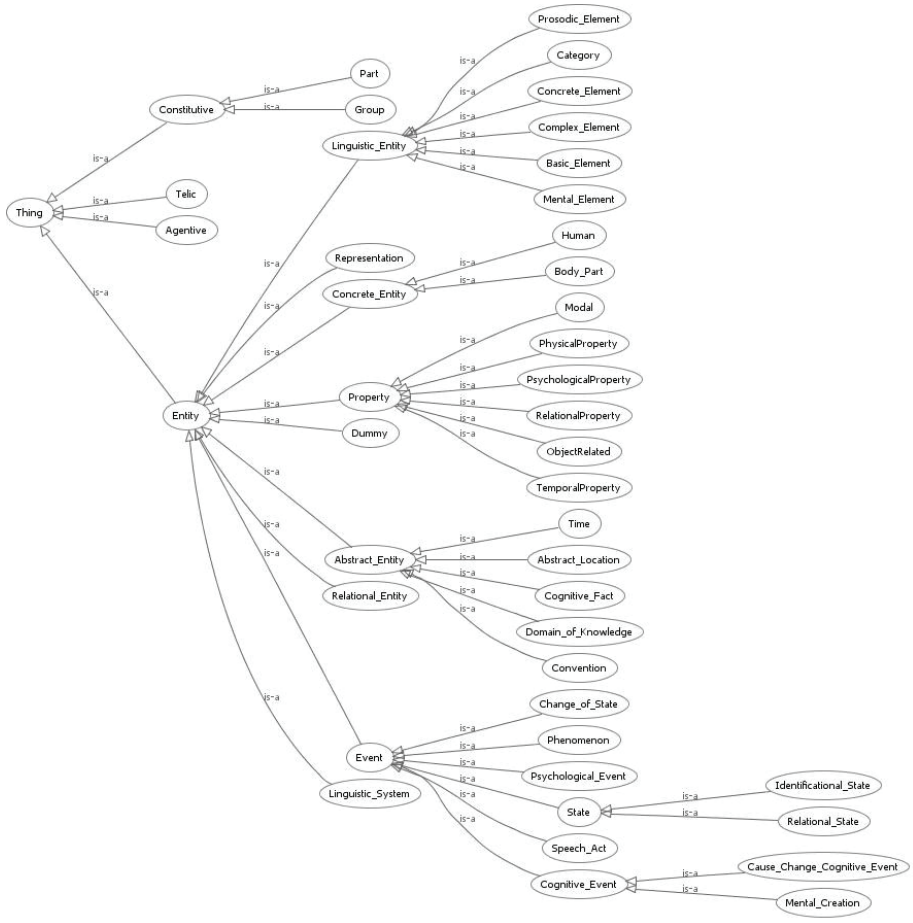


Fig.1. Représentation graphique de l’ontologie Simple_*FdS*

3.2. Les relations sémantiques

Les relations sémantiques constituent le cœur du modèle SIMPLE. La plupart d'entre elles sont le fruit d'une relecture et d'une extension de la structure des *qualia* du Lexique Génératif, la '*Extended Qualia Structure*'. Sous forme de triplets <*unité sémantique source, relation, unité sémantique cible*>, elles permettent de représenter de manière structurée la multi-dimensionnalité du sens d'un mot et de décrire finement le réseau de liens que les instances des différentes classes ontologiques entretiennent, tant au plan paradigmatique qu'au plan syntagmatique. Les relations classiques de synonymie, antonymie et de dérivation morphologique complètent ce large éventail de liens.

La flexibilité du modèle de référence a permis de conserver uniquement les relations jugées pertinentes à notre domaine de la connaissance et d'en créer de plus spécifiques afin d'exprimer des liens qui caractérisent l'organisation conceptuelle propre à ce domaine. Voyons ici quelques-unes des relations sémantiques les plus représentatives de la pensée saussurienne.

La relation 'duality' relie certains termes clés des réflexions saussuriennes, comme *langue* et *parole*, *synchronie* et *diachronie*. Pour le lecteur expert, le nom de cette relation évoque le célèbre manuscrit de l'Orangerie intitulé "De l'essence double du langage", et en particulier la note pour le second cours de linguistique générale [6]. Cette relation implique la forte interaction entre les deux concepts en jeu, contrairement à la traditionnelle interprétation dichotomique de ces couples, héritage de la conception structuraliste des niveaux hiérarchiques.

Liée à la précédente, la relation 'directlyImplies' relie, par exemple, *signifié* à *signifiant*. Il s'agit là d'une dualité non antinomique ; les deux entités entretiennent un rapport d'implication réciproque en ceci qu'elles constituent les deux faces indissociables du signe linguistique[7].

La relation 'belongsTo' permet de situer un concept, soit dans une perspective synchronique ou diachronique, soit dans la sphère d'appartenance à la *langue* ou à la *parole*. Ce sont là deux oppositions capitales. D'une part, la distinction entre « la linguistique de la langue et la linguistique de la parole »[8]. D'autre part, celle entre deux différents ordres de faits : la linguistique synchronique, qui a pour centre l'état d'une langue et les rapports entre des « termes coexistants », et est concentrée sur la

---

[6] « le langage est réductible à cinq ou six DUALITÉS ou *paires de choses* » (ELG, p. 298).

[7] « Il est aussi vain de vouloir considérer l'idée hors du signe que le signe hors de l'idée » (ELG, p. 44).

[8] Distinction que Hjelmslev définit la « thèse primordiale » du CLG et qui semble constituer la « première vérité » sur laquelle se fonde toute la linguistique générale, (Riedlinger, 1911 (SM 30)). La langue, vue comme la partie passive du circuit de la parole, le code social qui organise le langage et représente l'instrument fondamental pour l'exercice de la faculté du langage, se distingue de la parole, l'acte individuel, momentané du sujet parlant utilisant le code.

perspective du sujet parlant et la linguistique diachronique, centrée sur les rapports entre des « termes successifs » et fortement liée au facteur temps[9].

Les relations 'hasPreviousDenomination' et 'hasSubsequentDenomination' permettent de saisir et de formaliser les changements terminologiques typiques de Saussure. S'élevant contre l'inexactitude de certains termes traditionnels dans le domaine de la linguistique, il tente d'en définir de manière rigoureuse des concepts centraux afin que celle-ci accède véritablement au statut de science. À cet effet, il confère des sens nouveaux à des mots existants, change la dénotation de certains concepts au fil du temps, forge des néologismes dont certains, jugés inadéquats, sont abandonnés par la suite.

Dans cette optique, deux autres relations ont été créées. La première, 'hasNear-Synonym' relie des termes caractérisés par une sémantique lexicale très semblable (*diachronie* et *évolution*)[10]. La seconde, 'hasOtherDenomination', permet de signaler l'introduction d'un terme nouveau, souvent créé *ad hoc*, pour indiquer un concept précédemment exprimé de façon différente. Ainsi, *morphologie* rebaptisée *théorie des signes*[11].

3.3. Propriétés exprimées par des couples attribut-valeur

Dans sa version actuelle, le modèle Simple_*FdS* est caractérisé par 30 propriétés appartenant à trois catégories différentes : les propriétés sémantiques, caractérisant soit une classe conceptuelle tout entière soit une instance spécifique ; les propriétés relatives à l'usage des termes ; les propriétés morphosyntaxiques (catégorie grammaticale).

Les propriétés sémantiques apportent un complément d'information concernant la portée sémantique des termes. Représentant des caractéristiques transversales à la hiérarchie des classes, elles favorisent le regroupement d'unités lexicales, indépendamment de leur classification ontologique. De plus, elles permettent de représenter des dimensions de sens, clairement perçues dans le contenu sémantique d'un mot, mais cependant trop imprécises pour être exprimées par un terme, au sein d'une relation sémantique.

Voici quelques-uns des traits sémantiques spécifiques des classes nominale et verbale :

'Mental' désigne la nature typiquement mentale d'une entité ou d'un événement (*image*, Mental yes). 'Intentionality' indique la présence ou l'absence d'intentionnalité

---

9 « Tous les objets à étudier forment deux domaines : le ou les champs synchroniques, le champ diachronique, à quoi correspondent des perspectives et des méthodes différentes. Dans le champ diachronique, la méthode se dédouble ; dans le champ synchronique, l'unique méthode est l'observation de ce qui est ressenti par les sujets parlants. » (R 84-85) (Godel, 72).

10 « Il ya un certain nombre de termes à peu près synonymes…: *histoire, évolution altération* – et on peut proposer aussi le terme de *diachronie*. » (D232).

11 « Le vrai nom de la morphologie serait : la théorie des signes, et non des formes » (3293 §2).

dans un événement, selon la distinction chère à Saussure entre les activités conscientes opérées sur la langue par le linguiste et les activités inconscientes typiques de la masse parlante (*analyse_objective,* Intentionality yes ; *analyse_subjective,* Intentionality no). 'AgentiveF' et 'TelicF' indiquent respectivement une dimension de sens agentive ou télique inexprimable au niveau lexical dans une relation sémantique (*innovation,* Agentive yes ; *symbole*, Telic yes). Il en est de même pour 'ResultingState' qui signale la présence d'un résultat imprécisable, conséquence d'une action (*altération*, ResultingState yes).

À ces propriétés sémantiques s'ajoute un ensemble de traits spécifiques des adjectifs. Six 'Meaning Component' précisent l'information fournie par la classe sémantique (*cosystématique* : RelationalProperty, MeaningComponent : SetMembership). D'autres traits marquent la distinction entre adjectifs extensionnels et intensionnels ; indiquent leur graduabilité ; la durée temporelle d'une propriété physique ; et pour les adjectifs modaux, le contrôle humain sur un événement ainsi qu'une nécessité, une prédiction, ou une possibilité d'actualisation.

Les propriétés concernant l'usage des termes fournissent des indications relatives à la période d'attestation ou d'apparition d'un terme (*temps_homogène*, AttestationPeriod 1910-1911) ; sa fréquence, au moyen de valeurs non numériques (*circuit_de_la_parole,* Frequency hapax) ainsi que la référence à l'ouvrage dans lequel il apparaît (*terme*, Source N 20).

4. La base de connaissance lexicale

Les moyens descriptifs qui ont été présentés ont permis la création, à ce jour, de 500 entrées lexicales[12] ayant un contenu informatif particulièrement riche. Chaque entrée décrit une seule acception d'un lexème simple ou complexe et fournit une représentation structurée de la définition que Saussure lui-même donne de ce terme ou, à défaut, de celles qu'en proposent Godel et Engler. Dans les exemples d'entrées illustrés (cf. tableau 1), l'on peut observer, outre les susdites définitions en langage naturel, la présence d'informations concernant la classe ontologique d'appartenance du terme, le domaine d'usage, la période d'attestation, la fréquence d'occurrence, la source, les collocations, ainsi qu'un ensemble de relations qui établissent un lien entre le sens décrit et un certain nombre d'événements ou d'entités qui lui sont strictement liés.

[12] Dont 379 noms, 112 adjectifs et 9 verbes.

id = "uni-spatialité" **Saussure definition** = " « Principe de l'uni-spatialité (si l'on considère le sôme) ayant pour conséquence dans le sème la divisibilité par tranches (toujours dans le même sens et par coupures identiques) [l'opposition entre deux sèmes se règle au moyen de tranches allant dans le même sens et n'arrivant qu'une à une] au lieu de la divisibilité pluriforme qu'on aurait par exemple dans un système visuel direct (= écriture idéographique). » (3317.2) ; « Il faut s'astreindre à dire uni-spatialité du signe linguistique chaque fois, afin de faire sentir que ce n'est pas un caractère général du sème. » (3318.1) " **Definition** = " «Linéarité du signifiant.» (Engler, 51) " **Features**=" **SemanticType** Property **SuperType** Entity **Domain** Linguistics **PoS** Noun" **SemRelations:** **Isa:** proprieté (Property) **isPropertyOf:** signifiant2 (Concrete Element) **isPropertyOf:** signe4 (Concrete Element) **isPropertyOf:** sôme (Concrete Element) **isPropertyOf**: discours (Speech_Act) **isPropertyOf:** chaîne_de_la_parole (Speech_Act) **isPropertyOf:** syntagme (Complex Element) **isPropertyOf:** parole (Speech Act) **Concerns**: divisibilité (Property) **hasSynonym**: linéarité (Property) **hasOtherDenomination**: succession_temporelle (Property) **deajectivalNoun**: spatial (Physical Property) **belongsTo**: parole (Speech Act)	**id = "parole"** **Saussure definition** = " « Dans le langage, la langue a été dégagée de la Parole, elle réside dans [ ] l'âme d'une masse parlante, ce qui n'est pas le cas pour la parole. » (ELG, 333) ; « Quand on défalque du Langage tout ce qui n'est que Parole, le reste peut s'appeler proprement la Langue et se trouve ne comprendre que des termes psychiques, le nœud psychique entre idée et signe, ce qui ne serait pas vrai de la parole. » (ELG, p. 334) " **Definition** = " « langage moins la langue » (CLG, Engler, 38) « Tout ce qui est amené sur les lèvres par le besoin du discours et par une opération particulière. » (IR2.24) ; « Acte de l'individu réalisant sa faculté du langage au moyen de la convention sociale qui est la langue. » (IIR6-7) ; « Partie active et individuelle [du langage], mais où il faut distinguer : 1°) l'usage des facultés en général en vue du langage (phonation, etc.) ; 2°) l'usage individuel du code de langue selon la pensée individuelle. » (D178 ; cf. 184 J) (Godel, 271) " **Collocations** = "parole effective ; parole potentielle ; circuit de la parole ; linguistique de la parole" **Features** = "**SemanticType** Speech_Act **SuperType** Act **Domain** Linguistics **EventType** Process **Intentionality** Yes **Sound** Yes **AttestationPeriod** 1811-1911 **PoS** Noun " **SemRelations:** **Isa:** exécution (Speech_Act) **hasInstrument** : langue2 (Linguistic_System) **isTypicalOf** : sujet_parlant (Human) **hasProperty:** linéarité (Property) **hasProperty:** uni-spatialité (Property) **duality**: langue2 (Linguistic_System) **hasSynonym:** langue_discursive (Speech_Act) **hasSynonym:** discours (Speech_Act) **Constitutive**: langage (Property)

Tab. 1 Deux entrées lexicales : *uni-spatialité* et *parole*

5. Migration d'Access à Protégé-OWL

Il existe actuellement deux versions de la base de données lexicales. Dans la première version, le lexique est mémorisé dans une base de données relationnelle Microsoft Access. Celle-ci est gérée par une interface implémentée dans le langage de programmation VBA et articulée en une série de masques pour la création, la consultation et la modification des entrées lexicales. Le système permet en outre d'imprimer des rapports, d'exporter des parties spécifiques du lexique et d'effectuer des contrôles de cohérence.

Dans le but de satisfaire aux impératifs de standardisation et d'interopérabilité, la ressource lexicale a été successivement traduite de manière semi-automatique en langage OWL (Web Ontology Language), standard du W3C (World Wide Web Consortium) pour la représentation et le partage d'ontologies sur le Web. La version 4.1 de la plateforme Protégé[13] a été utilisée pour la consultation et la révision du lexique OWL. Ce système intègre des plugins pour l'interrogation des données, tels que DL Query et OWL2Query, qui basent leur fonctionnement sur des raisonneurs sémantiques. D'autres plugins facilitent la phase de consultation et de révision de l'ontologie. Protégé permet en outre d'effectuer, toujours grâce à des moteurs de raisonnement, nombre de tests de cohérence de l'ontologie.

[13] ‹protege.stanford.edu/›

Bien que les deux versions du lexique électronique (Access et OWL) soient encore actives à ce jour, la désactivation de la version Access est prévue dès que seront développés et intégrés dans Protégé des plugins open source dédiés aux fonctionnalités manquantes de répertoriage, d'exportation et de vérification. Ces nouveaux plugins permettront de personnaliser l'environnement Protégé pour la création et la consultation de lexiques électroniques.

Cette activité prévoit par ailleurs la production de composants supplémentaires pour la version en ligne de Protégé (Web Protégé[14]), qui ne requiert aucune installation sur ordinateur et est utilisable à travers un navigateur commun.

6. Interrogation du lexique

Sur le lexique, un grand nombre de requêtes peuvent être effectuées très simplement en partant d'une donnée mémorisée : une relation, un couple attribut-valeur ou une unité sémantique ; données interrogeables tant individuellement que simultanément. Il est possible d'extraire des groupes de termes partageant une ou plusieurs propriétés, selon des critères établis par le spécialiste pour satisfaire aux exigences de sa recherche. Effectuées dans un environnement formalisé garantissant la complétude des données extraites, le résultat de ces requêtes offre une vision exhaustive et approfondie de la nature componentielle et relationnelle du sens des mots.

Voici, à titre d'exemple, quelques-unes des nombreuses requêtes pouvant être effectuées :

- Quels sont les termes définis comme étant une faculté, un état ou un acte cognitif ?
- Quels sont les termes définis comme étant des créations résultant d'un processus d'association ?
- Quels sont les termes appartenant au domaine de la diachronie qui ont changé de dénomination au fil du temps ?
- Quels sont les termes dont la période d'attestation se situe entre 1893 et 1894 ?

Outre l'utilisation des plugins de l'éditeur Protégé, et afin de guider l'utilisateur peu familier avec les outils informatiques dans la formulation des requêtes, une interface conviviale de questions-réponses a été récemment développée à l'ILC. Elle se présente comme un ensemble de modèles de requêtes ('templates') en langage naturel contrôlé (Schwitter, 2010). Plusieurs types de requêtes peuvent être formulées à partir d'un seul 'template', grâce à des menus déroulants permettant de sélectionner, parmi les éléments proposés, les critères de la requête.

Chaque 'template' est composé de trois parties :

- une partie fixe qui caractérise un modèle d'interrogation (ex. : 'quels sont') ;
- une partie variable permettant de choisir dans un menu déroulant le type d'élément du lexique sur lequel effectuer la requête, ex. : 'relation sémantique' ou 'propriété' (cf. fig. 2,

[14] ‹protegewiki.stanford.edu/wiki/WebProtege›

requête 1) ; ‘propriété sémantique’, ‘propriété morphosyntaxique’ ou ‘propriété d’usage du terme’ (cf. requête 2) ;

– une seconde partie variable permettant le choix d’une certaine propriété et de sa valeur (ex. : ‘Intentionality’, ‘no’) ; d’une certaine relation avec ou sans indication de son terme objet (ex. : ‘belongsTo’, ‘*diachronie*’) ou d’un certain terme du lexique (ex. : ‘*diachronie*’).

L’utilisateur, guidé par le modèle choisi, peut donc ainsi composer une requête en utilisant exclusivement le langage naturel.

Supposons qu’il souhaite savoir quels sont les faits intentionnels appartenant au domaine de la diachronie. Le modèle de requête sera le suivant, les choix qu’il devra opérer étant entre crochets angulaires :

Quels sont les termes caractérisés par la propriété sémantique <attribut> <valeur> et par la relation sémantique <nom de la relation> dont l’objet est <second terme de la relation> ? (cf. fig. 2, requête 4)

L’attribut sélectionné sera la propriété ‘Intentionality’, sa valeur ‘yes’ ; la relation, ‘belongsTo’ et le second terme, ‘*diachronie*’.

La composition des requêtes est un processus incrémental. En sélectionnant la valeur désirée dans chaque menu, celles contenues dans les menus déroulants successifs sont recalculées en conséquence. Reprenant l’exemple précédent, si l’utilisateur sélectionne dans <attribut> la propriété sémantique ‘Intentionality’, seules les valeurs associées à cette propriété apparaîtront dans le menu déroulant <valeur>. Le choix de la valeur ‘yes’ restreindra, à son tour, la liste des relations sémantiques : seules seront affichées dans le menu <nom de la relation> les relations liées aux termes ayant la propriété ‘Intentionality yes’. Enfin, le choix d’une des relations affichées, ici ‘belongsTo’, restreindra plus encore la liste des termes sur lesquels opérer un choix dans <second terme de la relation> puisque seules seront proposées les unités lexicales associées aux deux caractéristiques précédemment sélectionnées.

Les différentes typologies de requêtes sont illustrées dans la figure 2.

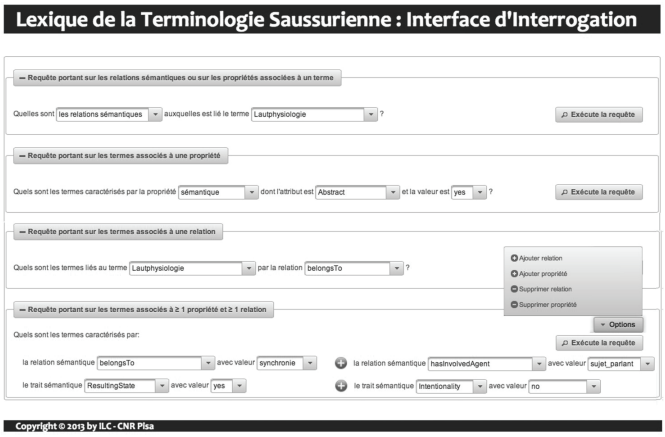


Fig. 2 Interface conviviale pour l’interrogation du lexique

Observons plus en détail le point 4 de la figure 2. Il s'agit là de la représentation formelle de la requête suivante :

Quels sont les termes dénotant des activités inconscientes (<Intentionality> <no>), typiques du sujet parlant (<hasInvolvedAgent> <sujet_parlant>), qui ont un résultat (<ResultingState> <yes>) et qui appartiennent au domaine de la synchronie (<belongsTo> <*synchronie*>) ?

La figure 3 illustre la visualisation sous forme de tableau du résultat de cette requête. L'utilisateur peut classer par ordre alphabétique le contenu de chaque colonne, afin de mettre tour à tour en relief différents aspects intéressants des données extraites et conserver le résultat de sa requête en l'exportant au format pdf.

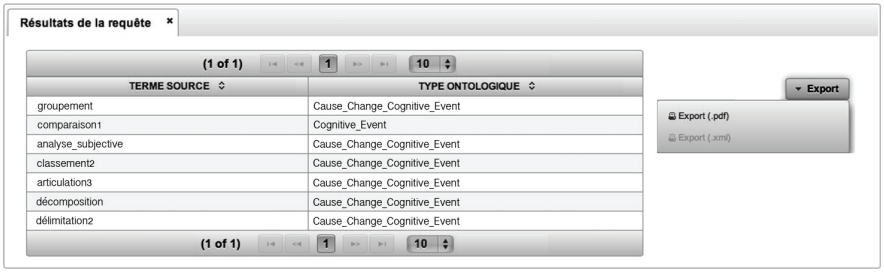


Fig. 3 Visualisation des résultats d'une requête

Il apparaît donc clairement que même un utilisateur non expert en recherche d'informations peut ainsi formuler tout à fait aisément des requêtes précises et complexes en restreignant progressivement le champ de recherche par la sélection de propriétés et relations liées aux termes du lexique.

6.1. Technologies utilisées pour l'interface d'interrogation

L'interface décrite a été développée comme application Web et est accessible sur le site de l'ILC[15]. Le dorsal de l'application a été développé avec la technologie Java EE. À partir des données contenues dans le lexique, JENA[16] construit le modèle ontologique adéquat en tenant compte de la connaissance inférée par son algorithme de raisonnement. Les modèles de requêtes sont traduits par le logiciel en requêtes SPARQL[17] qui interrogent ce graphe et fournissent à l'utilisateur les réponses désirées, grâce également aux informations inférées par le raisonneur. Le frontal de l'application, qui implémente les modèles en langage naturel assisté pour les requêtes a

15 ‹www.ilc.cnr.it/›.

16 *Framework* Java pour les applications web sémantique. ‹jena.apache.org›.

17 Langage de requêtes pour des données représentées sous forme de triples RDF ‹www.w3.org/TR/rdf-sparql-query/›.

été, quant à lui, développé avec des technologies Java Server Faces (JSF)[18] et Primefaces[19].

7. Conclusion

Ce lexique a été créé dans le but d'offrir un instrument de recherche lexicale performant pour les études saussuriennes. Il a pour ambition de permettre une étude rigoureuse de la terminologie du grand linguiste genevois en fournissant une structuration multidimensionnelle des concepts du domaine de référence et une représentation formelle et hautement structurée de la sémantique lexicale de chacun des termes caractéristiques de sa pensée, ainsi qu'une vision d'ensemble du réseau de liens qui les unissent. Cet outil pourrait permettre une nouvelle approche de l'œuvre saussurienne et contribuer de manière déterminante à en éclairer certains aspects originaux.

CNR-Pise, Istituto di Linguistica Computazionale *Antonio Zampolli* — Silvia PICCINI
Computazionale *Antonio Zampolli* — Emiliano GIOVANNETTI
Computazionale *Antonio Zampolli* — Andrea BELLANDI
Computazionale *Antonio Zampolli* — Nilda RUIMY

Références bibliographiques

Engler, Rudolf, 1968. *Lexique de la terminologie saussurienne*, Utrecht-Anvers, Spectrum, Comité international permanent des linguistes. Publication de la commission de terminologie.

Francopoulo, Gil et al., 2007. « Lexical Markup Framework: ISO standard for semantic information in NLP lexicons », *Proceedings of Lexical-semantic and Ontological Resources*, Workshop at Gesellschaft für linguistische Datenverarbeitung, Biennal Spring Conference, Tübingen.

Godel, Robert, 1957. *Les sources manuscrites du Cours de linguistique générale de Ferdinand de Saussure*, Genève, Droz.

Gruber, Tom Robert, 1993. « A translation approach to portable ontologies », *Knowledge Acquisition* 5 (2), 199-220.

Guarino, Nicola et al., 2009. « What Is an Ontology ? », in : Staab, Steffen et Studer, Rudi, (ed.), *Handbook on Ontologies*, second edition, Berlin-Heidelberg, Springer Verlag, 1-17.

Lenci, Alessandro et al., 2000. « SIMPLE: A General Framework for the development of Multilingual Lexicons », *International Journal of Lexicography* 13 (4), Special issue 'Dictionaries, Thesauri and Lexical-Semantic Relations', 249-263.

Pustejovsky, James, 1995. *The Generative Lexicon*, Cambridge MA, The MIT Press.

[18] ‹www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html›.

[19] ‹primefaces.org/›.

Ruimy, Nilda et al., 2003. « A computational semantic lexicon of Italian: SIMPLE », in : Zampolli, A. et al. (ed.), *Computational Linguistics in Pisa, Special Issue,* Pisa-Roma, IEPI, Vol. XVIII-XIX, II, 821-864.

Saussure, Ferdinand de, 1916. *Cours de linguistique générale*, Bally, Charles et Sechehaye, Albert (ed.), Lausanne-Paris, Payot.

Saussure, Ferdinand de, 1968-1974. *Cours de linguistique générale*, Engler, Rudolf (ed.), Vol. 1-2, Wiesbaden, Otto Harrassowitz.

Saussure, Ferdinand de, 1922. *Recueil des publications scientifiques de Ferdinand de Saussure*, Bally, Charles et Gautier, Léopold (ed.), Lausanne, Payot.

Saussure, Ferdinand de, 2002. *Écrits de linguistique générale*, Bouquet, Simon / Engler, Rudolf (ed.), Paris, Gallimard.

Schwitter, Rolf, 2010. « Controlled Natural Languages for Knowledge Representation », *Proceedings of COLING 2010*, Beijing, China, 1113-1121.

Sowa, John F., 2002. « Building, Sharing, and Merging Ontologies ». ‹www.jfsowa.com/ontology/ontoshar.htm›.

Studer, Rudi, Benjamins, V. Richard, Fensel, Dieter, 1998. « Knowledge engineering: principles and methods », *Data & Knowledge Engineering*, Elsevier Ltd, Volume 25, Issues 1-2, 161-197.

Vossen, Piek, 1999. *Eurowordnet. A multilingual database with lexical semantic networks*, Dordrecht, Kluwer.

Apache Jena Project! Apache Jena™. ‹jena.apache.org›.

RDF Vocabulary Description Language 1.0: RDF Schema. ‹www.w3.org/TR/rdf-schema/›.

SPARQL Query Language for RDF W3C Recommendation 15 January 2008. ‹www.w3.org/TR/rdf-sparql-query/›.

Mise en ligne, mise à jour et mise en réseau du *Französisches Etymologisches Wörterbuch*

La mise en ligne et la mise en réseau des ressources lexicographiques constituent un chantier actuellement très important en Europe : le CNRTL[1] à Nancy, le projet GiGaNT[2] à Leiden ou le projet Wörterbuchnetz[3] à Trèves en sont quelques exemples. L'un des moyens mis en œuvre consiste à 'rétroconvertir' un dictionnaire imprimé en dictionnaire électronique en y insérant, de façon semi-automatisée, des balises XML. Cette opération s'avère particulièrement délicate dans le cas d'ouvrages aux structures complexes, dont on veut rendre compte sans en cacher les imperfections.

Parmi les ouvrages à informatiser de cette manière se trouve le FEW, œuvre incontournable dans le domaine de la linguistique française et romane et en même temps reconnue comme difficilement exploitable. Il est depuis longtemps acquis que l'informatisation du FEW, à condition de respecter les structures et la complexité de l'oeuvre, permettra de résoudre les problèmes d'accessibilité et ouvrira de nouveaux itinéraires de consultation. Par ailleurs, le FEW constitue un « dizionariotetto » (Buchi/Renders 2013) qui serait bénéfiquement mis en réseau avec des dictionnaires en ligne tels que le *Dictionnaire étymologique de l'ancien français* (DEAF) ou l'*Anglo-Norman Dictionary* (AND).

La présente contribution a pour objectif premier de faire le point sur l'avancement du projet d'informatisation des 25 volumes du *Französisches Etymologisches Wörterbuch* (FEW). Nous présentons ensuite brièvement un projet récent, fortement lié au précédent et soutenu en Belgique par le Fonds National de la Recherche Scientifique sous l'intitulé « Étude des modalités de mise à jour et mise en réseau du *Französisches Etymologisches Wörterbuch*, dans le cadre de son informatisation, avec d'autres ressources linguistiques notamment belgoromanes ».

1. Informatisation du FEW et mise en ligne

L'informatisation du FEW et sa mise en ligne à travers une interface d'interrogation sont attendues depuis longtemps par la communauté scientifique : elles constituent ensemble la solution la plus efficace aux difficultés d'accès de l'ouvrage. L'objectif est double : d'une part, faciliter la recherche d'informations ciblées (étymons ou

[1] Cf. Pierrel / Buchi 2009 ; ‹www.cnrtl.fr/›.

[2] ‹www.inl.nl/onderzoek-a-onderwijs/projecten/gigant›

[3] ‹woerterbuchnetz.de›

lexèmes bien sûr, mais aussi références bibliographiques, suffixes etc.) ; d'autre part, faciliter la lecture d'un article grâce à diverses aides (plan de l'article, lien vers les abréviations des sigles, etc.).

Pour qu'un texte, quel qu'il soit, puisse être interrogé en ligne, trois étapes sont nécessaires : la mise en forme du texte, son enrichissement via des métadonnées et, enfin, la création d'outils de requête permettant d'interroger le texte via ces métadonnées. Il en va de même pour un dictionnaire : il s'agit (1) d'acquérir le texte (caractères spéciaux inclus) sous forme électronique ; (2) d'identifier sa structure et son contenu de façon explicite au moyen de métadonnées, de façon à permettre des requêtes sur les divers types d'information voulus ; (3) de donner accès au texte et aux métadonnées via une interface d'interrogation.

Idéalement, ces trois étapes se succèdent dans le temps, chacune permettant la suivante. En ce qui concerne le FEW cependant, la résolution de ces trois étapes n'a pas eu lieu dans l'ordre indiqué. Le problème de l'identification des informations a été étudié et résolu avant que ne soient trouvés des financements pour réaliser la première étape, qui est aussi la plus coûteuse. Le projet a donc été pensé de façon à ce que son avancement ne dépende pas de l'acquisition du texte entier des 25 volumes du FEW. Actuellement (juillet 2013), seule une acquisition partielle est en cours. La troisième étape reste, quant à elle, encore à réaliser. L'ensemble du projet fait l'objet d'une collaboration étroite entre l'ATILF (CNRS/Université de Lorraine) et le service de *Linguistique du français et dialectologie wallonne* de l'Université de Liège, dont nous faisons partie en tant que chargée de recherches F.R.S.-FNRS. L'acquisition du texte sous forme électronique bénéficie également de l'aide de la Fondation pour le FEW (Fondation pour le dictionnaire étymologique du français du professeur von Wartburg) et du Center for Digital Humanities de Trèves (http://kompetenzzentrum.uni-trier.de).

Les choix posés ne sont pas sans avoir d'influence sur la manière dont le FEW informatisé sera interrogé et utilisé. Nous expliquons ci-dessous l'avancement du travail et les solutions adoptées en ce qui concerne chacune de ces trois étapes, dans l'ordre qui est celui de leur résolution : l'identification des informations, l'acquisition du texte et, enfin, la création d'une interface.

1.1. Identification des informations

Si la deuxième des trois étapes précitées a été résolue en premier, c'est parce que l'ensemble du projet d'informatisation nécessitait une étude préalable de faisabilité, qui s'est clôturée en 2011 sous la forme d'une thèse de doctorat en cotutelle Liège-Nancy (Renders 2015). Les solutions adoptées pour l'identification des divers composants du discours 'fewien' sont le résultat de cette thèse, constituée d'une partie plutôt linguistique (métalexicographique) et d'une partie plutôt informatique (algorithmique). Il s'agissait en effet de proposer une modélisation du discours fewien avant d'étudier la façon dont cette modélisation, formalisée en XML, pourrait être intégrée de façon automatisée dans le texte lexicographique.

La modélisation du FEW, telle qu'elle a été proposée et adoptée, présente une caractéristique essentielle. Alors que la thèse d'Éva Buchi présentait les structures du FEW selon un point de vue interne (Büchi 1996), ces mêmes structures sont appréhendées, pour la modélisation, selon le point de vue de l'utilisateur. En effet, à partir du moment où l'objectif est de résoudre des problèmes d'utilisation, il nous paraît évident qu'il faut avoir une idée de l'identité des utilisateurs du dictionnaire, de ce qu'ils y cherchent, de la façon dont ils le consultent, des difficultés qu'ils rencontrent et des fonctionnalités qu'ils voudraient y trouver. Des discussions avec plusieurs spécialistes et avec les membres de la rédaction du FEW, ainsi que les réponses à un questionnaire distribué en 2007 à l'occasion du XXV^e^ CILPR, ont permis d'obtenir un aperçu des pratiques actuelles d'utilisation du FEW et de connaître les souhaits des utilisateurs dans l'optique d'un FEW informatisé, souhaits qui sont en relation étroite avec les difficultés qu'ils rencontrent dans la consultation de la version imprimée. L'étude a mis en évidence les parcours suivis par les utilisateurs au sein du dictionnaire et, surtout, deux modes d'utilisation de l'ouvrage, correspondant à deux dimensions du FEW : une utilisation du FEW comme thesaurus d'unités lexicales d'une part, comme recueil d'articles monographiques d'autre part. La vision du FEW comme thesaurus concerne notamment les parcours de consultation et les voies d'entrée dans le dictionnaire ; la vision du FEW comme recueil de monographies vise davantage les parcours de lecture et la mise en relation des informations au sein d'un article. Ces deux dimensions présentent une complémentarité essentielle, dans le sens où privilégier l'une des deux au détriment de l'autre est source d'approximations et d'erreurs. À partir de ce constat, nous avons pu construire une modélisation du FEW intégrant les deux dimensions précitées et respectant à la fois les structures du dictionnaire et les attentes de l'utilisateur. Le modèle identifie une trentaine de types d'information explicites ou implicites, situés à différents niveaux du FEW (de l'infrastructure à la superstructure, cf. Büchi 1996). La formalisation du modèle a été réalisée au moyen d'un schéma XML.

Il restait à trouver le moyen d'intégrer ce formalisme au sein du texte lexicographique de la façon la plus automatisée possible, un balisage manuel étant impossible du fait de la densité de l'ouvrage. C'était l'objectif de la partie algorithmique de la thèse, qui a effectivement conduit à l'automatisation complète du balisage. Cette automatisation est possible grâce à une séquence d'algorithmes qui, chacun, balise un des types d'information retenus (étymon, lexème, sigle bibliographique, datation, commentaire, champ documentaire, appel de note etc.). Les algorithmes de balisage utilisent des informations syntaxiques (fournies via les balises insérées par les algorithmes qui les précèdent dans la séquence) et des informations lexicales (fournies via des listes de mots-clés provenant, entre autres, du *Complément* au FEW). Un mécanisme de 'chaînes virtuelles' (cf. Briquet/Renders 2010) facilite la détection de ces informations en permettant de rendre invisibles certains éléments XML lors de la recherche de chaînes de caractères. Le tout s'intègre dans un logiciel dit 'de

rétroconversion' implémenté en java[4]. Les résultats fournis sont plus que satisfaisants, puisque la plupart des informations sont reconnues à cent pour cent dans le corpus de test, constitué d'une centaine d'articles concernés par le projet ANR DETCOL[5]. Les erreurs et oublis éventuels sont repérables grâce à un mécanisme d'alertes et de vérification visuelle intégré à la fin de chaque article XML produit. Enfin, le bon fonctionnement du logiciel ne nécessite pas que la totalité du FEW soit disponible : tout article ou échantillon d'articles fourni au logiciel, à condition qu'il soit muni d'un balisage minimal (uniquement typographique) et codé en Unicode, se voit en quelques secondes doublé d'une copie XML totalement balisée, dans laquelle toutes les informations voulues sont identifiées.

La modélisation du discours fewien et le logiciel de balisage permettent d'effectivement insérer dans chaque article du FEW les métadonnées adéquates : l'étude de faisabilité résoud ainsi la deuxième des trois étapes précitées. La mise en pratique nécessite toutefois que soit préalablement disponible une version électronique brute de l'article, ce qui nous ramène à la première étape, celle de l'acquisition du texte.

1.2. Acquisition du texte

L'opération consistant à acquérir une version électronique brute du texte lexicographique peut s'effectuer par saisie manuelle ou par océrisation, c'est à dire par l'application, sur un scan du document, d'un logiciel de reconnaissance optique. Quelle que soit la solution retenue, elle requiert également que soit choisie une police de caractères permettant le codage de tous les glyphes présents dans le dictionnaire.

Les tests d'océrisation réalisés sur le FEW ont fourni des résultats qui, malgré leur qualité (jusqu'à 96% de reconnaissance), nécessitaient ensuite un effort important de correction manuelle, plus coûteux en ressources temporelles et humaines que la solution d'une saisie directe. Cette dernière, proposée par le Center for Digital Humanities de Trèves, consiste à effectuer une double saisie, puis à comparer à l'aide d'un logiciel automatique les deux versions obtenues et à vérifier le résultat uniquement là où il y a divergence. Ce traitement offre un résultat excellent, voire parfait. Le coût pour la totalité du FEW s'élève à plusieurs dizaines de milliers d'euros. Heureusement, le logiciel de rétroconversion ayant été conçu de façon à pouvoir traiter le dictionnaire par ensemble d'articles, il est possible d'avancer en fonction des financements obtenus. Actuellement (juillet 2013), trois volumes, à savoir les volumes 16, 17 et 19, ont pu, avec l'aide de la Fondation pour le FEW, être scannés à l'ATILF et fournis au Center for Digital Humanities de Trèves, qui opère la double saisie et la vérification.

---

[4] L'implémentation en java du logiciel a été réalisée par Cyril Briquet, ingénieur de recherche à l'ATILF entre octobre 2008 et décembre 2009.

[5] Pour une présentation du projet ANR-DETCOL (Développement et Exploitation Textuelle d'un Corpus d'Œuvres Linguistiques), voir http://ctlf.ens-lyon.fr/documents/de_anr_workshop-corpus.asp

La question de la police de caractères est, quant à elle, directement liée à la présence de caractères spéciaux et à la question de leur codage : la police utilisée doit en effet contenir l'ensemble des caractères du texte. En ce qui concerne le FEW, un relevé a fait apparaître la présence de nombreux caractères phonétiques non standards, c'est-à-dire non répertoriés dans Unicode et, donc, non systématiquement pris en charge par les polices disponibles sur le marché. La présence de ces caractères non standards empêche d'une part l'acquisition, d'autre part la visualisation sur écran, du texte fewien. Pour lever le premier obstacle, il a été décidé de coder les caractères non standards en tant qu'entités XML (telles que « &few-e-ouvert-accent », par exemple). Cette solution permet de poursuivre l'informatisation du FEW depuis la phase d'acquisition jusqu'au balisage complet des articles. Seule la dernière étape du processus, à savoir la mise en ligne du dictionnaire et sa visualisation sur écran, nécessitent la création d'une police contenant l'ensemble des caractères fewiens. Le relevé de l'ensemble de ces caractères, accompagnés de leur codage Unicode (éventuellement dans la zone privée pour les caractères non standards) se trouve dans Renders 2015 ; ce relevé constitue une base technique pour la future police du FEW, dont la réalisation a été confiée en 2013 à l'Atelier National de Recherche Typographique (http://www.anrt-nancy.fr/).

1.3. Mise en ligne et interface d'interrogation

La dernière étape du processus, consistant à mettre en ligne les articles balisés et à les rendre interrogeables via un moteur de recherche, est prise en charge par l'ATILF. À l'heure où nous écrivons ces lignes, un prototype existe, mais il nécessite encore du développement. Soulignons que le FEW informatisé sera mis gratuitement à la disposition de la communauté des chercheurs : cette volonté, dont on peut se réjouir, s'inscrit dans le courant actuel de l'Open Access.

En attendant la création d'une interface d'interrogation en ligne, l'ATILF a prévu de rendre disponibles les scans en version image des volumes du FEW. Les trois volumes en préparation seront bientôt accessibles en format image via le site de l'ATILF (www.atilf.fr/few) ; les autres devraient suivre assez rapidement.

2. Mise à jour et mise en réseau

Parallèlement à l'acquisition, l'informatisation et la mise en ligne du texte complet du FEW, un nouveau projet de recherche a vu le jour, qui anticipe l'avenir et se penche sur deux questions : la mise à jour du FEW et sa mise en réseau avec d'autres ressources lexicographiques pour lesquelles il constitue une référence obligée. Ces deux problématiques sont liées, puisque la mise en réseau doit rendre disponible une partie des ajouts et corrections actuellement consignés dans des ouvrages annexes (cf. Greub 2012) ou dans des ressources en ligne (TLF-Étym, DEAF ou AND, pour ne citer qu'eux). Les possibilités d'informatisation de ressources belgoromanes corrigeant le FEW, telles que l'*Atlas Linguistique de la Wallonie* (ALW), font partie de l'étude.

Le corpus d'articles choisi pour tester les possibilités de mise à jour du FEW comprend notamment les articles du volume 19 (Orientalia), qui présentent l'intérêt d'avoir fait l'objet d'actualisations et de corrections systématiques (Arveiller 1999) et de pouvoir être mis en relation avec des notices du projet TLF-Étym consacrées aux arabismes du français (Baiwir 2010). Au-delà des possibilités et des choix informatiques, il s'agit d'étudier la façon dont la mise en relation du FEW avec des ressources en ligne susceptibles elles-même d'évoluer peut influencer les modalités de mise à jour du FEW lui-même.

Cette étude n'en est qu'à son début, mais elle a déjà permis de dégager des constats nets. La question de la mise à jour du FEW, tout comme celle de son informatisation et de sa mise en ligne, nécessite de cerner d'abord les parcours traditionnels suivis par les utilisateurs qui veulent vérifier l'actualité d'une information donnée par le FEW. Elle nécessite également que soit identifié le contenu possible de ces mises à jour : s'agit-il de compléments, de corrections, d'ajouts, de refontes ? Quelles parties du discours lexicographique sont concernées ? Ce n'est qu'ensuite que seront envisagées les possibilités de parcours informatiques, de façon à ce que le FEW informatisé permette les parcours traditionnels et les facilite. L'utilisateur pourrait par exemple, à partir d'un article en ligne, avoir accès à tous les compléments qui y sont apportés ailleurs, via des liens directs (uri, pour les ressources en ligne) ou via des références bibliographiques (pour les ressources disponibles uniquement en version imprimée). Dans le cas d'articles refondus, le parcours serait inversé : l'utilisateur partirait de l'article mis à jour et pourrait, s'il le désire, consulter l'ancienne version archivée. Rappelons que toute version d'un article, même obsolète, doit absolument être conservée, dans l'optique de permettre la vérification des sources d'une recherche : toute version est en effet susceptible d'avoir été utilisée et citée par un chercheur.

Nous pourrions peut-être aller plus loin encore et envisager une intégration des corrections et ajouts au sein même du FEW. En effet, la question de la mise à jour pourrait, en définitive, se résumer à ceci : les mises à jour du FEW peuvent-elles intégrer le FEW = avec le risque de perturber les structures du discours initial – ou doivent-elles rester à l'écart de l'ouvrage originel, si cohérent dans ses qualités et ses défauts? L'étude y répondra ; qu'il nous soit déjà permis d'affirmer qu'énoncée comme telle, et malgré sa réelle pertinence, cette question relève avant tout d'une conception 'papier' de l'objet dictionnairique. Par ailleurs, il ne sera jamais question de « mettre à jour le FEW », mais de mettre à jour, soit un article du FEW, soit une unité lexicale du FEW (pourvue de toutes les informations y associées) : ici comme précédemment, il est nécessaire, nous semble-t-il, de distinguer les deux dimensions du FEW, en tant que thesaurus et en tant que recueil de monographies. Cette distinction est peut-être la clé d'une mise à jour intégrée qui soit à la fois respectueuse des structures de l'ouvrage et susceptible de répondre malgré tout aux attentes des utilisateurs.

La question de la mise en réseau du FEW avec d'autres ressources informatisées devrait, quant à elle, se résoudre plus simplement, l'"adresse FEW' faisant partie du programme lexicographique de la plupart des ouvrages concernés. Il restera à définir pour le FEW informatisé des adresses pérennes qui puissent servir à la fois de référence et de lien hypertextuel.

3. Conclusion

Certes, le FEW n'est pas encore disponible en version électronique. Derrière ce constat désolant, se cache une grande activité : à l'heure où nous écrivons ces lignes, le FEW est scanné à Nancy, le texte de trois volumes est en cours d'acquisition à Trèves, le logiciel de balisage est opérationnel et attend d'être utilisé à Liège. Ces opérations devraient voir très bientôt leur achèvement. Si tout se passe comme prévu, l'année 2014 verra également la création de l'interface d'interrogation à l'ATILF et, espérons-nous, la création de la police de caractères par l'Atelier National de Recherche Typographique, ce qui permettra la mise en ligne effective des volumes 16, 17 et 19 du FEW. L'informatisation des volumes suivants dépendra de nouveaux financements, non encore obtenus. En attendant, nous espérons que la mise en ligne d'une partie du FEW permettra de fournir à la communauté scientifique un exemple, critiquable, de ce que peut être le FEW informatisé, mis à jour et mis en réseau.

FNRS / Université de Liège Pascale RENDERS

Références bibliographiques

ALW = Remacle, Louis / Legros, Élisée et al., 1953–. *Atlas Linguistique de la Wallonie. Tableau géographique des parlers de la Belgique romane d'après l'enquête de Jean Haust et des enquêtes complémentaires* (10 volumes), Liège.

AND = Rothwell, W. / Gregory, S. / Trotter, D. A. (dir.), 2005[2] [1977–1992[1]]. *Anglo-Norman Dictionary*. ‹www.anglo-norman.net›

Arveiller, Raymond, 1999. *Addenda au FEW XIX (Orientalia)*, Tübingen, Niemeyer.

Baiwir, Esther. *Refonte des notices 'étymologie et histoire' des articles suivants du Trésor de la Langue Française informatisé (TLFi) : achour (2010), alezan (2010), antari (2010), coufique (2010), fez (2010), ketmie (2010), amin (2010), sumac (2010)*. ‹www.atilf.fr/tlf-etym›.

Buchi, Éva / Renders, Pascale, 2013. « 46. Gallo-romance I : Historical and etymological lexicography », in : *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume : Recent developments with special focus on computational lexicography*, Grouws, Rufus et al. (ed.), De Gruyter, Berlin/New York.

DEAF = Baldinger, Kurt et al., 1971–. *Dictionnaire étymologique de l'ancien français*, Québec/Tübingen/Paris, 1971–. ‹http://www.deaf-page.de›

FEW = Wartburg, Walther von et al., 1922–2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes (25 vol.)*, Bonn / Heidelberg / Leipzig-Berlin / Bâle, Klopp / Winter / Teubner / Zbinden.

Briquet, Cyril / Renders, Pascale, 2010. « A virtualization-Based Retrieval and Update API for XML-Encoded Corpora », in : *Proceedings of Balisage: The Markup Conference 2010*, Balisage Series on Markup Technologies, vol. 5.

Buchi, Éva, 1996. *Les Structures du 'Französisches Etymologisches Wörterbuch'. Recherches métalexicographiques et métalexicologiques*, Tübingen, Niemeyer.

Complément = Chauveau, Jean-Paul / Greub, Yan / Seidl, Christian, 2010. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Complément*, Strasbourg, Éditions de linguistique et de philologie, Bibliothèque de Linguistique Romane, Hors Série 1.

Greub, Yan, 2012. « Sur le FEW », *The Aberystwyth Colloquium. Present and future research in Anglo-Norman (21-22 juillet 2011)*, Aberystwyth, Anglo-Norman Online Hub, 187-190.

Pierrel, Jean-Marie / Buchi, Éva, 2009. « Research and Resource Enhancement in French Lexicography : the ATILF Laboratory's Computerised Resources », in : Bruti, Silvia / Cella, Roberta / Foschi Albert, Marina (ed.), *Perspectives on Lexicography in Italy and Europe*, Newcastle upon Tyne, Cambridge Scholars Publishing, 79-117.

Renders, Pascale, 2015. *L'informatisation du Französisches Etymologisches Wörterbuch. Modélisation d'un discours étymologique*, Strasbourg, Éditions de linguistique et de philologie.

TLF-Étym = Collectif, 2005. *Projet TLF-Étym: mise à jour des notices étymologiques du Trésor de la langue française informatisé. Dossier de présentation*, Nancy, ATILF/CNRS/Université Nancy 2/UHP, ‹http://www.atilf.fr/tlf-etym› ; Voir aussi Buchi, Éva, 2005. « Le projet TLF-Étym (projet de révision sélective des notices étymologiques du Trésor de la langue française informatisé) », *Estudis romànics* 27, 569-571.

Términos médico-botánicos del occitano antiguo en grafía hebrea para el DiTMAO : casos problemáticos

1. Introducción

Este artículo tiene como finalidad mostrar algunos de los casos más problemáticos que surgieron durante la lematización de términos en grafía hebrea en el proyecto DiTMAO[1] (Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan). El proyecto consiste en la elaboración de un diccionario de consulta del occitano antiguo basado en la terminología médico-botánica (véase también los artículos de María Sofía Corradini y de Andrea Bozzi en el presente volumen). El trabajo de lematización se realiza, por su operatividad, con el sistema *Reperio* (www.reperio.it), puesto que permite la publicación tanto en formato electrónico (web) como en formato de libro impreso. El DiTMAO se está constituyendo con un corpus de textos medievales del ámbito médico-botánico, en el que se utilizan tanto textos escritos en grafía latina, como textos escritos en grafía hebrea (cf. Corradini / Mensching 2010, 2013). Éstos últimos son principalmente listas de equivalencias, tradicionalmente llamadas listas de sinónimos, y a causa de las corresponcias indicadas en otras lenguas (hebreo-árabe-latín), son de muchísima importancia para el estudio de dicha terminología del occitano antiguo. La datación de los manuscritos que constituyen el corpus abarca los siglos XIII, XIV y XV[2], con una temática de lo más diversa y que comprende desde recetarios, herbarios, tratados monográficos y las mencionadas listas de sinónimos, hasta pasajes relacionados de forma más amplia con la práctica sanadora. Los textos redactados en occitano antiguo y escritos en grafía latina han sido editados por Maria Sofia Corradini (1997, 2001).

Puesto que este artículo enfoca la lematización de los términos del DiTMAO en grafía hebrea, no vamos a tomar en cuenta los manuscritos existentes en grafía latina[3].

---

[1] El DiTMAO (Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan) está financiado por la DFG (Deutsche Forschungsgemeinschaft) y es un proyecto colaborativo entre la Universität zu Köln (Investigador principal: Gerrit Bos, colaboradora: Veronika Roth), el Instituto di Linguistica Computazionale Antonio Zampolli (Investigador principal: Andrea Bozzi, colaboradores: Emiliano Giovannetti, Damiana Luzzi), la Università di Pisa (Investigadora principal: Maria Sofia Corradini) y la Georg-August-Universität Göttingen (Investigador principal: Guido Mensching, colaboradoras: Anja Weingart y Julia Zwink).

[2] Véase Corradini (1997, 14-21), ShS1 (55-58) y Bos/Mensching 2005,177).

[3] Para una descripción detallada de los manuscritos, puede consultarse las ediciones antes señaladas de Corradini (1997) y (2001).

Los textos en grafía hebrea que se reúnen en el corpus del diccionario, son, como hemos apuntado, listas de sinónimos, pero también se han incluido dos tratados médicos en lengua hebrea que contienen voces en occitano antiguo. El primero de estos textos, es el fragmento de una traducción al hebreo del herbario latino *Macer floridus* (véase la edición en Bos / Mensching 2000) y el segundo, es una traducción hebrea del compendio árabe *Zād al-musāfir wa-qūt al-ḥādir* de Ibn al-Ǧazzār por Moisés Ibn Tibbon con el título *Ṣedat ha-Derakhim* (la edición en Bos et.al. está en preparación; véase también Mensching / Zwink en imprenta). Entre los manuscritos en grafía hebrea que están enumerados en (1), se encuentran las listas de sinónimos contenidas en el libro 29 del *Sefer ha-Shimmush* de Shem Tov ben Isaak de Tortosa, que es la mayor fuente lexicográfica de términos médico-botánicos en occitano antiguo.

(1) Textos en grafía hebrea.
  (a) Listas que forman parte de repertorios médico-botánicos.
    (i) *Sefer ha-Shimmush* de Shem Tov ben Isaak de Tortosa, lista 1 (hebreo-árabe-occitano/latín; ed.: Bos et.al. 2011) [=ShS1]
      - París , Bibl. Nat. Héb. 1163, ff. 191r-195r [= ShS1 P]
      - Vaticano Ebr. 550, pp. 132r-153r [= ShS1 V]
      - Oxford, Bodleian Hunt Donat 2, ff. 278r-283rr [= ShS1 O]
    (ii) *Sefer ha-Shimmush* de Shem Tov ben Isaak de Tortosa, lista 2 (occitano/latín-árabe-hebreo; ed.: en preparación) [=ShS2]
      - París , Bibl. Nat. héb. 1163, pp. 195r-198r [= ShS2 P]
      - Vaticano Ebr. 550, pp. 153v-163v [= ShS2 V]
      - Oxford, Bodleian Hunt Donat 2, ff. 284a-290a [= ShS2 O]
  (b) Listas anónimas.
    (i) Versión hebrea de la lista de sinónimos «Alphita» (cf. Bos / Mensching 2005:201-208)
      - Parma Bibli. Palat. 3043, ff. 265r-266v [= Alphita]
    (ii) Lista latín-árabe-francés u occitano. (cf. Bos / Mensching 2005:172-177)
      - Múnich Bayer. Staatsb. 245, ff. 155r-167r, [= Múnich I]
    (iii) Lista árabe-latín-francés u occitano
      - Múnich Bayer. Staatsb. 245, ff. 167v-175r, [= Múnich II]
    (iv) Lista árabe-(hebreo)-occitano o latín (cf. Bos / Mensching 2005:172-177)
      - Oxford, Bodleian L. MS Mich. Add. 22, ff. 5v-16r [= Oxford]
  (c) Tratados.
    (i) Versión hebrea del Macer Floridus (fragmento, ed. Bos / Mensching 2000)
      - Berlín Staatsb. Preussischer Kulturbesitz Qu 834 (cat. Steinschneider 241), ff. 72a-73b [= MF]
    (ii) Ṣedat ha-Derakhim de Moisés Ibn Tibbon (edición Bos et.al. en prep., cf. Mensching / Zwink en impr.)
      - Berlín, Staatsb. Preussischer Kulturbesitz Qu 835 (cat. Steinschneider 239), ff. 115a-137b
      - Múnich, Bayer. Staatsb. Cod. hebr. 295, ff. 150b-160a
      - Oxford, Bodleian MS Poc. 353, ff. 70b-80b
      - Múnich, Bayer. Staatsb. Cod. hebr. 19, ff. 184b-206a

Bos / Mensching (2005,170-171) diferencian entre tres tipos de listados de sinónimos escritos en grafía hebrea denominadas como A, B y C. Bajo el tipo A se se agrupan los listados que no contienen material en hebreo. Las listas tipo B y C contienen material hebreo organizado de la siguiente forma: El tipo B está ordenado por el lema de otro idioma, sea latín, romance o árabe. El tipo C está ordenado por el lema hebreo. Un ejemplo del tipo A son las dos listas anónimas en el manuscrito Múnich Bayer. Staatsb. 245, que no contienen elementos hebreos. La primera lista de este manuscrito-Múnich I, está ordenada por el lema latino seguido por sinónimos árabes y romances. La segunda lista, la Múnich II, está ordenada por el lema árabe seguido por sinónimos romances o latinos. De la misma forma, el manuscrito abreviado como Oxford, también anónimo, está organizado como el manuscrito Múnich II, pero contiene junto a los sinónimos romances o latinos algunas voces en lengua hebrea. Por eso, forma parte de las listas del tipo B. Otro representante del tipo B, tal vez el más conocido, es la versión en caracteres hebreos del Alphita, donde aparecen, principalmente, términos en latín a los que - en algunos casos - se añadieron sinónimos en romance y explicaciones en hebreo.

El libro 29 del *Sefer ha-Shimmush* contiene dos listados: el primero es el editado en Bos et.al. 2011 [= ShS1] que pertenece al tipo C, ya que está ordenado por el lema y el alfabeto hebreo, seguido por términos en árabe y romance o latín. El segundo listado [= ShS2], cuya edición se está realizando en curso del proyecto, está ordenado por el lema romance o latino, seguido por sinónimos en romance o latín , términos en árabe y, a veces, también en hebreo, por lo que forma parte del tipo B.

Aparte de describir la forma de organización de las listas, queremos llamar la atención sobre el uso de las lenguas vernáculas. Tal es su importancia, que en las listas de la tradición hebrea se incluyeron términos en lenguas nativas de manera sistemática –a partir del siglo XIII– mucho antes que en las listas de la tradición latina[4] El uso de lenguas vernáculas se puede relacionar con la situación histórica de los judíos durante la Edad Media. Se supone que la emigración de judíos procedentes de Al-Ándalus a zonas como Cataluña, así como a otras zonas como el Lenguadoc o Provenza, provocó el desuso y desconocimiento del árabe, lo que supuso una dificultad añadida para entender los términos técnicos en árabe. Estas circunstancias, y el hecho de que los nombres de plantas varíen de un lugar a otro, no fue sólo la motivación para la creación de numerosas traducciones de obras médicas al hebreo y al latín y de las listas de sinónimos, sino también para el empleo general de términos vernáculos. Estos términos fueron vitales para identificar correctamente las diferentes plantas y medicamentos, disminuyendo así el riesgo de su utilización de forma equivocada con consecuencias fatales (cf. introducción a ShS1, pág. 6).

De esta forma, los textos en grafía hebrea representan una fuente importante para el léxico médico-botánico del occitano antiguo y un desafío lexicográfico debido a la grafía hebrea (cf. Mensching 2006, 2009, Mensching / Savelsberg 2004, Mens-

[4] Cf. ShS1, 7-8 , Gutiérrez Rodilla (2007, 290-291).

ching / Bos 2011, Mensching / Zwink en imprenta). En Corradini / Mensching (2010) y (2013) se presentan el enfoque lexicográfico y el método empleado para la lematización. Podemos resumir, de forma breve, este método de la forma siguiente: Si una voz está documentada en grafía latina y en grafía hebrea, una de las variantes en grafía latina constituirá el lema, mientras que las variantes documentadas en grafía hebrea, indicadas en un sistema de transliteración[5], se clasificará como variantes alfabéticas del lema occitano. La mayoría de los casos que se desvían de este método consisten en: 1. Palabras en grafía hebrea que son variantes morfológicas o gráfico-fonéticas de un lema occitano; 2. Palabras en grafía hebrea que son variantes de una variante de un lema occitano; 3. Voces reconstruidas e hipotéticas; 4. Palabras que se componen de un elemento occitano y otro hebreo. Aquí cabe mencionar también la importancia de lo que los autores antiguos indicaron como «sinónimos» en hebreo y árabe para la reconstrucción de algunas palabras. Por eso, éstos son también incluidos en el DiTMAO, marcados como «equivalente en otro idioma» puesto que no son sinónimos en un sentido moderno. La discusión de los casos problemáticos, que es el tema central del presente artículo, se expone en el tercer punto. Con el fin de explicar la transcripción y a la interpretación de las voces en grafía hebrea, se presenta en el párrafo siguiente una entrada en el diccionario que no reviste mucha problemática.

2. Una entrada ejemplar

Se puede considerar una entrada tipo aquella voz en grafía hebrea que corresponde exactamente con la voz en grafía latina, es decir, donde la única diferencia es el uso de un alfabeto diferente. En (2) se reproduce una entrada del *Sefer ha-Shimmush* (ShS1)

(2) אסא ב«ה סא בו«ל הטרינ (ShS1 Alef 5-P, pág.: 95)[6]
ʾSʾ ár. ʾS en lengua vernácula NYRṬH

El lema hebreo (ʾSʾ- אסא) está identificado por un término árabe (ʾS - סא), introducido por ירגהב (*bə-hagarī*, abreviado ב«ה - BʿH) que significa ʿen árabeʿ. La expresión *bə-laʿaz* (זעלב, abreviado ב"ל - B"L) que fue traducida como ʿen lengua vernáculaʿ y que significa literalmente ʿcualquier lengua que no sea el hebreoʿ, que, en nuestro caso, se refiere al occitano antiguo, introduce el término en lengua vernácula que es NYRṬH.

A fin de comprender como fue transcrita esta locución, conviene repasar someramente algunas nociones generales sobre la ortografía de los términos romances

[5] La transcripción de las letras hebreas usado para en el ShS1 (véase p.: 5): Alef (א - ʾ), Bet (ב - B), Gimel (ג - G), Gimel (ג- Ǧ), Dalet (ד - D), He (ה - H), Waw (ו - W), Zayin (ז - Z), Ḥet (ח - Ḥ), Ṭet (ט - Ṭ), Yod (י - Y), Kaf (כ ך - K), Lamed (ל - L), Mem (מ ם - M), Nun (נ ן - N), Samekh (ס - S), Ayin (ע - ʿ), Pe (פ ף - P), Ṣade (צ ץ - Ṣ), Qof (ק - Q), Resh (ר - R), Shin (ש - Š), Tav (ת - T).

[6] Las abreviaturas de la documentación se leen de la siguiente manera: Alef 5 refiere a la quinta entrada de la letra Alef y P es la abreviatura para el manuscrito París , Bibli.Nat. Héb. 1163 (véase el listado en (1)). En la edición del ShS 1 esta entrada se encuentra en la página 59.

en grafía hebrea[7]. Habría que señalar que las vocales no son representadas necesariamente pero en el Corpus se encuentran frecuentemente representadas a través de *matres lectionis*. En los textos de nuestro corpus se ha utilizado la letra Yod (י - Y) para representar las vocales /i/ y /e/, la letra Waw (ו - W) para representar las vocales /o/ y /u/. La letra Alef (א - ʾ) puede indicar varios fenómenos[8]: En posición inicial indica la presencia de una vocal (en la mayoría de los casos la vocal /a/). Si está al final o en medio de la palabra, indica normalmente la presencia de la vocal /a/. También la letra He (ה - H) representa la vocal /a/ pero sólo cuando ésta se sitúa en posición final. En (3) se ejemplifica la transcripción al alfabeto latino y en (4) la interpretación de la transcripción.

El ejemplo (3) se lee de derecha a izquierda, siguiendo las reglas del alfabeto hebreo. Bajo el (4a) hemos invertido la transliteración, de forma que se pueda leer de izquierda a derecha (éste es el orden que usaremos de aquí en adelante).

(3) (a) נ י ר ט ה
(b) H Ṭ R Y N

(4) (a) N Y R Ṭ H
(b) n e r t a

En nuestro ejemplo (4) se utiliza Yod para representar /e/ y He para la /a/ final. En los otros manuscritos del ShS1, los denominados O y V, se utilizó la letra Alef para representar la /a/ final (אטרינ - NYRṬʾ). La alternancia de He y Alef es una mera variación gráfica sin influencia en la pronunciación. Para representar /k/ y /t/ se usan frecuentemente las letras Qof (ק - Q) y Ṭet (ט - Ṭ), respectivamente. Las letras hebreas הטרינ - NYRṬH se pueden leer como la palabra *nerta*, que significa como los otros sinónimos dados en (2), 'mirto' o 'murta' (Myrtus communis L.)[9] La palabra *nerta* está documentada en nuestro corpus de textos en grafía latina (p. ej.: Corradini 1997, 179), de esta forma, aparte de ser documentado en textos occitanos, no cabe ninguna duda de que se trata de un término occitano (cf. FEW: 98-99). Respecto a su clasificación, tampoco existen problemas. La voz en grafía hebrea no se diferencia en ningún aspecto, sea morfológico o grafo-fonético, de la forma en grafía latina que se clasifica como lema. La única diferencia es el uso del alfabeto. Es decir, de esta forma, la palabra en grafía hebrea será clasificada como variante alfabética del lema.

---

7 Para informaciones más detalladas y una bibliografía amplia véase ShS1, 47-52.

8 La letra Alef, en posición inicial puede representar todas las vocales. Por ejemplo la /e/ en ʾŠPRMʾ - esperma (cf. ShS1, 91) y la /a/ ʾRMWLŠ - armols (véase el párrafo 3.2.1). Además existen las combinaciones de Alef inical con Yod o Waw. La combinación ʾY se interpreta como /i/ o /e/ como en ʾYNGYLŠ - enguilas (véase 3.1). Las vocales /o/ o /u/ se representan con la combinación ʾW como en ʾWRRYGʾ- orega (cf. MF, 34) Un ejemplo para la interpretación en posición final e intermedia es la voz MWŠQʾDʾ - muscada (véase el párrafo 3.4.). En posición intermedia entre dos vocales, Alef señala que estas dos vocales deben ser pronunciadas separadamente, p. ej.: ʾMYʾWŠ - ameos. (cf. ShS1, 337)

9 Cf. ShS1, 95.

3. Casos problemáticos

En muchos casos el término en grafía hebrea no es una simple transcripción del lema al alfabeto hebreo, tal y como hemos visto en la entrada tipo del punto anterior, sino que se distingue por otros aspectos. En los siguientes ejemplos veremos los casos problemáticos que se han encontrado.

3.1. Palabras en grafía hebrea que son variantes de un lema occitano

En el corpus de los textos en grafía latina se encuentra la palabra *anguila* (es decir ˈanguilaˈ -Anguilla anguilla-) que está documentada en Corradini (1997, 246) y que figura como el lema, puesto que es la forma mas frecuente (cf. ShS1, 431). En los manuscritos del ShS1 se descubrieron las formas listadas en (5):

(5) (a) שליגניא - ʾYNGYLŠ (ShS1 Ṣade 4-P, pág.: 431)
Léase “enguilas”

(b) שילגניא - ʾYNGLYŠ (ShS1 Ṣade 4-O, pág.: 431)
Léase “enguilhas” o “enguiles” (cat.)

(c) שיליגייא - ʾYYGYLYŠ (ShS1 Ṣade 4-V, pág.: 431)
Quizás “enguilhas” o “enguiles” (cat.)?[10]

Las dos letras iniciales ʾY (Alef y Yod) muestran que, en este caso, hay que leer enguila en vez de anguila[11]. Aparte de la variación gráfico-fonética, los tres ejemplos son formas del plural tal y como indica la letra final Š. El problema de interpretación aparece causado por la penúltima letra Y de los ejemplos (5a) y (5b), puesto que al interpretarse en combinación con la última letra Shin (ש - Š) se puede identificar la forma del plural en /-es/, lo que indicaría la palabra catalana enguiles. Otra posibilidad sería interpretar la letra Y junto con la antepenúltima letra Lamed (ל - L), lo que puede indicar una pronunciación palatal de la <l> (cf. ShS1, 431). Sabemos que en la Edad Media no había normas establecidas para indicar la articulación palatal de la <l> y <n> ni en los textos en grafía latina, ni tampoco en los textos en grafía hebrea. De esta forma, se puede identificar tres maneras de señalar una /l/ palatal mediante el uso de las letras <ll> , <lh> o simplemente <l>. En los textos en grafía hebrea se añade muchas veces Yod para distinguir entre palatales y no palatales (ShS1, 51-52). Como en nuestro corpus están documentadas unas variantes occitanas (en grafía latina) con la /l/ palatal, p.ej.: anguilha y anguilla (cf. Corradini 1997, 133-134) y también una variante enguillon (FEW 24.268), optamos a favor de la interpretación de LY (Lamed y Yod) como indicación de un palatal. En función de la documentación de

[10] Es una variante errónea o no documentada. Tal y como está en el manuscrito, permitiría lecturas como **eguilhas*, **eiguilhas*, **aiguilhas* o variantes catalanes corresponientes en *-les*.

[11] La combinación de las letras Alef y Yod se puede interpretar, en principio, como /e/ o como /i/. En el caso de los ejemplos (5a)-(5c) la interpretación con /e/ es la más probable ya que existen variantes como *enguila* en occitano antiguo (cf. FEW 24.567 y Raynouard 1838-1844, vol. 1, 88)

las formas anguilha, anguilla y enguila es posible suponer que las formas enguilha, enguilla también existían, aunque no haya documentación de las formas enguilha o enguilla en grafía latina. De esta forma se puede concluir que las tres variantes del ShS1 se deben clasificar como variantes alfabéticas (hebreo), morfológicas (plural) y grafo-fonéticas (a causa de la vocal inicial /e/ y la /l/ palatal en los manuscritos O y V) del lema occitano anguila.

3.2. Palabras en grafía hebrea que son variantes de una variante del lema occitano

En el último ejemplo hemos visto cómo se introdujeron tres tipos de variantes: un tipo alfabético, uno morfológico y una gráfico-fonético. En este párrafo veremos ejemplos de voces en grafía hebrea que son variantes de una variante del lema.

3.2.1. Variante alfabética de una variante morfológica

En el corpus se identificó el lema *armol* (documentado en Corradini 1997, 137) con el significado de 'armuelle' (Atriplex hortensis L.). El plural *armols,* también está documentado en grafía latina (Corradini 1997, 315) y es variante morfológica del lema *armol*. En el ShS1 aparecen las formas siguientes que corresponden al plural ya documentado y que sólo difieren en el uso del alfabeto hebreo:

(6) (a) שלומרא - ʾRMWLŠ En SHS1 Lamed 6-PV, p.: 286
Léase: "armols"

(b) ילומרא - ʾRMWLṢ En SHS1 Lamed 6-O, p.: 286
Léase: "armols"

Las formas en (6a) y (6b) se incluyen en el DiTMAO clasificadas como variantes alfabéticas de la variante morfológica (*armols*) del lema (*armol*).

3.2.2. Variante alfabética y grafo-fonética de una variante morfológica

A veces, una forma en grafía hebrea no sólo es variante alfabética de una variante morfológica, sino también se distingue de esa variante en aspectos gráfico-fonéticos. En el ejemplo (7a) la palabra BRWṬS tiene una lectura como "brots", que corresponde a la forma del plural *brotz* (documentado en Corradini 1997, 144 y 247). La forma del singular y el lema de la entrada es *brot* (documentado en Corradini 1997, 134 y 349) que significa 'brote de vid'. La forma del plural *brotz* se clasifica como variante morfológica de *brot*. De esta forma, la forma en (7a) es una variante alfabética de la variante morfológica (*brotz*).

(7) (a) סטורב - BRWṬS (ShS1 Lamed 4-PV, pág.: 287)
Léase: «brots»

(b) יטורוב - BWRWṬṢ (ShS1 Lamed 4-O, pág.: 287)
Léase «*borotz»[12]

[12] El asterisco indica que la forma no está documentada en grafía latina.

El manuscrito O contiene la forma en (7b) que no corresponde exactamente a la forma del plural en grafía latina. Entre las primeras letras Bet (ב - B) y Resh (ר - R) se encuentra insertada la letra Waw (ו - W), que es interpretada como vocal epentética (aquí la vocal /o/). El uso de una vocal epentética no es extraordinario, puesto que se encuentra frecuentemente insertada entre las consonantes oclusivas y líquidas (véase ShS1: 52). Además de ser una variante alfabética de la forma de plural (la variante morfológica del lema), la forma **borotz* se clasifica también como variante grafofonética debido a la vocal epentética.

3.3. Voces reconstruidas e hipotéticas

Ahora veremos ejemplos de voces en grafía hebrea que no están documentados ni en el corpus en grafía latina, ni en diccionarios como el FEW o el DAO. En el primer caso son voces compuestas cuyo significado puede ser reconstruido con ayuda de la documentación de sus partes así como de los sinónimos. En el segundo caso, las voces simples quedan opacas, aunque se puede formular una hipótesis sobre su significado mediante los sinónimos en árabe y hebreo.

3.3.1. Voces compuestas con ambas partes documentadas por separado

En el ShS2 se encuentra la voz compuesta ĞYRWPLY DYWYLʾNŠ representada en (8). La voz consiste en las partes *girofle*, *de* y *vilans*.

(8) שנאליויד ילפורי'ג- ĞYRWPLY DYWYLʾNŠ (ShS2 Alef 12-P, pág.: 100)
Léase: "*girofle de vilans"

Como hemos apuntado, el término completo no está documentado, pero sí sus partes: las palabras *girofle* y *vilan* (y evidentemente la preposición *de*). La palabra *girofle*, documentada en nuestro corpus de textos en grafía latina (cf. CB, 144 y 240), significa 'clavo' o 'girofle' (cf. FEW 2.446b) y la palabra *vilan* significa 'campesino' (cf. FEW 14.452). Asimismo, un indicio clave para la reconstrucción del significado son los sinónimos que acompañan el término romance. En (9) se muestran los sinónimos de ĞYRWPLY DYWYLʾNŠ .

(9) (a) ארזא ארקב - ʾZRʾ BQRʾ (ShS2 Alef 12-P, pág.: 100)
Léase: "asara bacara"

(b) יראשא - ʾŠʾRY (ShS2 Alef 12-P, pág.: 100)
Léase: "asari"

Ambos sinónimos están documentados en occitano antiguo con el significado de ásaro (Asarum europaeum L). La forma *acara bacaras* está documentada en Corradini (1997, 318) y la forma *assari* en Corradini (1997, 308). En el Alphita en grafía latina los dos términos figuran como sinónimos de *gariofilus agrestis*:

(10) (a) « Asarus, asara baccara, vulgago, gariofilus agrestis idem » (ed. García Gonzáles 2007, pág.: 150)
(b) « Gariofilus agrestis, asarum idem » (ed. García Gonzáles 2007, pág.: 225)

La palabra *gariofilus* corresponde semánticamente a *girofle* (cf. MLW 2:317). El hecho de que ambos términos aparezcan como sinónimos en el ShS2 refuerza la correspondencia semántica. La entrada del ShS2 en (11) muestra los dos términos: *girofle* seguido por *gariofilus*.

(11) ילפוריג וא שולפוירג בו"ה לפנרק בו"מ משב (ShS2 Gimel 14)
GYRWPLY o GRYWPLWŠ, ár. QRNPL, hebreo bíblico BŠM

Acerca de la expresión *de vilan*, se puede constatar la correspondencia semánticamente con *agrestis* (cf. MLW 1:407). De esta forma, es muy probable que los términos **girofle de vilans* y *gariofilus agrestis* tengan el mismo significado[13], lo que nos lleva a la conclusión de que **girofle de vilans* es un calco de *gariofilus agrestis* y por lo tanto debe significar 'Asarum europaeum L'. En conclusión, proponemos incluir el término **girofle de vilans* en el DiTMAO como forma hipotética en grafía latina que tiene el significado 'Asarum europaeum L'. La forma se clasifica como sublema de *girofle*.

3.3.2. Voces no documentadas en absoluto

En ocasiones, un término que aparece en una lista de sinónimos, no está documentado en ninguna parte, es decir, ni en el corpus de textos médico-botánicos escritos en grafía latina, ni en el FEW o el DAO. En este caso dependemos de los sinónimos árabes y hebreos para averiguar el significado de la palabra. Un ejemplo es la voz QRDYLʾŠ en (12) que se encuentra en ambas listas del ShS.

(12) שאלידרק - QRDYLʾŠ (SHS1 Ayin 7-VO, pág.: 381 y ShS2 Qof-12)
Léase : "*cardelas" o "*cardelhas"

Como en el ShS2 QRDYLʾŠ figura como lema y se sabe que representa una voz en lengua romance (probablemente occitano antiguo), el análisis nos permite suponer que QRDYLʾŠ es el plural de una forma como *cardela o *cardelha – está marcada de forma hipotética ya que no está documentada. Acerca del origen y el significado de esta forma, se presentan los argumentos a favor y en contra de dos análisis posibles. El primer análisis presupone que *cardela o *cardelha son formas derivadas de la palabra occitana card que significa 'cardo'. Desde este punto de vista QRDYLʾŠ debería pertenecer a la entrada de card (probablemente como diminutivo), aunque en contra de este argumento tengamos el significado de los sinónimos árabes y hebreos que figuran en el ShS1 y en el ShS2, y que se representan en (13).

(13) ár.: אבדנה - HNDBʾ (ShS1 Ayin 7, pág.: 381 y ShS2 Qof 12)
heb.: וישלוע - ʿWLŠYN (ShS1 Ayin 7; pág.: 381 y ShS2 Qof 12)

Ambos términos significan 'achicoria' (Cichorium intibus L. y Var.) o 'endibia' (Cichorium endivia L. y Var.). El segundo análisis parte de los significados de los

[13] El término **girofle de vilans* también podría corresponder a un término documentado por ejemplo en italiano *garofani salvatichi* (y variantes dialectales) (véase LEI 12:888,894). Sin embargo esto no es tan probable ya que los sinónimos indican que **girofle de vilans* debe significar Asarum europaeum L. mientras que el término *garofani salvatichi* se refiere a Dianthus carthusianorum L. y otras especies del género Dianthus.

sinónimos y se supone que **cardela* o **cardelha* también tendrían este significado. Es fundamental pensar que esta conclusión podría ser precipitada, puesto que ya existen las palabras occitanas *sicoreia* y *endivia* con el significado de 'achicoria' y 'endibia', ambas perfectamente documentadas (véase ShS1, 382). Por lo tanto, si QRDYLʾŠ es considerado un sinónimo corriente de sicoreia y endivia ¿por qué no aparecen estos términos en las listas? Aunque esta sea una duda razonable, existen más indicios que nos permiten corroborar la conclusión basada en los sinónimos hebreos y árabes. Es sabido que algunos nombres derivados de la palabra latina cardu(u)s o **carda* servían para denominar varias plantas comestibles que, desde el punto de vista de la botánica moderna, no tenían mucho que ver con el cardo (ShS1, 382). Otro indicio esclarecedor es que en occitano moderno existe la forma cardélo, que significa 'cerraja' (Sonchus) (cf. FEW 2:371), una planta muy parecida a la achicoria. También en otras lenguas romances existen palabras similares a **cardela / *cardelha*, aunque no se utilicen necesariamente para denominar a la achicoria o a la endibia, pero que denominan plantas parecidas a estas. También en algunos dialectos italianos (véase LEI 12:79-80) varios nombres derivados de cardu(u)s denominan plantas que pertenecen a la familia de las Cichorieae[14], como por ejemplo:

(14) plantas del género Sonchus (cerraja o lechuga de las liebres):
  (a) siciliano antiguo: *cardella* (formas modernas p. ej.: *cardedda*, *kardéḍḍa*, etc.)
  (b) ligur occidental: *cardela* o *cardelle* o *cardellina*
  (c) ligur alpino: *cardéla*
(15) lig. occ. (Airole): *kardéḷḷe* significa 'diente de león' (Taraxacum officinale)
(16) lig. occ. (Bordighera): *cardella* o *cardellina* significa 'Urospermum dalechampii'
(17) lacial centro-septentrional: *cardellùzzo* se refiere a una 'pianta erbàcea simile al grespino e alla cicoria'

Estos ejemplos son relevantes, puesto que, además, Liguria limita geográficamente con las regiones occitano-hablantes. Tomando en cuenta que las plantas mencionadas tienen un aspecto similar y que en la Edad Media las plantas no fueron clasificadas según criterios botánicos modernos, es muy probable que se usaran los mismos nombres para denominar plantas parecidas, lo que nos hace presuponer que la palabra **cardel(h)a* pueda encuadrarse en estos casos. Así, en base a los indicios que hemos estudiado, optamos por dar más veracidad al segundo análisis. Por ello se propone introducir en el DiTMAO un nuevo lema, la forma hipotética **cardel(h)a* con el significado de 'achicoria' o 'endibia'. La voz QRDYLʾŠ se clasificará como variante alfabética y variante morfológica (plural) de este nuevo lema.

[14] Hay nombres derivados de *cardus* que también se utilizan para otras plantas como por ejemplo el cardo o la carlina, pero aquí nos limitamos a los géneros más afines a la achicoria. Para informaciones taxonómicas sobre las plantas y su pertenecencia a la tribu de las Cichorieae véase: Euro+Med Plantbase - the information resource for Euro-Mediterranean plant diversity. Publicado en internet (http://www.emplantbase.org/home.html).

3.4. Palabras compuestas de un elemento occitano y otro hebreo

Una particularidad de los textos en grafía hebrea es el uso de palabras compuestas por un elemento occitano y otro hebreo. En (18a) se presenta un ejemplo claro de una palabra compuesta por un sustantivo hebreo y un adjetivo occitano y en (18b) el caso contrario, es decir, con un sustantivo occitano modificado por un adjetivo hebreo.

(18) (a) אדאקשום זוגא - ʾGWZ MWŠQʾDʾ (Múnich I Alef 4)
Léase "egoz muscada"

(b) יפא ירריה - ʾPY HRRY (Múnich I Alef 3)
Léase "api harari"

En el caso (18a), la deducción del significado no representa un problema. El término hebreo ʾGWZ (léase "egoz") significa 'nuez' y mediante el sinónimo árabe GWZ BWY ('awz baww')[15] y el occitano NWṢ MWŠQʾDʾ (léase noz muscada)[16] sabemos que toda la expresión significa 'nuez moscada'. De esta forma, el problema reside únicamente en la lematización del término. Para dar solución a esta duda, o bien admitimos una palabra hebrea como lema occitano con el fin de poder clasificar la palabra compuesta como respectivo sublema, o bien clasificamosʾGWZ MWŠQʾDʾ como «equivalente en otro idioma» de noz muscada, un sublema de noz, ya que aparecen como sinónimos en la lista Múnich I. De ambas soluciones a la duda, la segunda parece la más correcta, puesto que bajo la categoría «equivalente en otro idioma» los sinónimos hebreos y árabes serán incluidos en el DiTMAO. Esta solución conlleva la afirmación de que ʾGWZ MWŠQʾDʾ o MWŠQʾDʾson un calco en hebreo medieval, cuyo único indicio (hasta el momento de redactar este artículo) a favor de esta suposición se encuentra en Löw 1928-34 [=LF]. Ahí podemos constatar que ʾGWZ aparece con asiduidad con un adjetivo romance, como muscato o moscada, para denominar a la nuez moscada, donde ʾGWZ por sí solo significa 'nuez' en general o nogal común (Juglans regia L.) (cf. LF 2:29ff.). Hasta el mismo momento de redactar este artículo, el equipo del proyecto no ha tomado una decisión definitiva de la lematización del término.

Aún falta el ejemplo en (18b), donde la palabra está compuesta por el sustantivo occitano ʾPY ('apio' - Apium L.) y el adjetivo hebreo HRRY que significa 'de montaña'/'de monte', así que todo el sintagma se traduciría por "apio de monte/de montaña". A través del sinónimo árabe karafs 'abal' se conoce el significado de la palabra compuesta que sirve para denominar, de hecho, a la planta 'apio de monte' (Peucedanum oreoselinum) (cf. DT 3:61-62). La lematización de este caso parece, a primera vista, menos problemática, puesto que existe el lema occitano api (cf. Corradini 1997, 149, 163 y 270). Según los criterios de lematización, la forma ʾPY es la variante alfabética de api, pero la palabra compuesta ʾPY HRRY no puede ser la variante alfabética de un respectivo sublema, ni siquiera en caso de que éste exi-

[15] Véase también ShS 2 Nun 6.

[16] Véase también p.ej. FEW 19:133 y la entrada nos muscadas en Corradini (1997, 147).

stiera, ya que contiene una palabra hebrea. Aún así, queda la posibilidad de incluirla en el DiTMAO directamente como sublema de api, una solución que tampoco es la ideal desde un punto de vista lexicográfico: hasta ahora nos hemos imaginado un usuario occitanista/romanista, puesto que para un investigador que trabaja con textos hebreos en donde aparecen este tipo de palabras, resulta muy útil un diccionario que facilite este tipo de información. En el caso de (18b) el conocimiento del adjetivo hebreo (HRRY) no es suficiente para deducir el significado del término compuesto. De esta forma, tomando en cuenta el aspecto interdisciplinario, se puede optar por la solución presentada.

4. Conclusión

En este texto se ha tratado de esbozar algunos problemas que plantea la lematización de términos occitanos documentados en grafía hebrea. De la misma forma, se han propuesto soluciones sostenibles para las palabras en grafía hebrea que son variantes nuevas de un lema occitano. Estas soluciones se basan principalmente en las palabras en grafía hebrea que sólo difieren en el uso del alfabeto de una variante ya documentada en grafía latina. Al mismo tiempo se ha puesto de relieve que las *synonyma* son determinantes en la elaboración de un diccionario médico-botánico del occitano antiguo. Así como que mediante los sinónimos se puede averiguar y comprobar el significado de un término opaco, como en el caso de las voces reconstruidas e hipotéticas. Las palabras compuestas de un elemento occitano y otro hebreo representan un verdadero acertijo lexicográfico que, hasta el día de hoy, aun no tiene una solución completamente satisfactoria.

Universidad de Colonia — Veronika ROTH
Universidad de Gotinga — Anja WEINGART
Universidad de Gotinga — Julia ZWINK

Bibliografía

Bos, Gerrit / Mensching, Guido, 2000. «Macer Floridus: A Middle Hebrew Fragment with Romance Elements», *The Jewish Quarterly Review* 91 (1-2), 17-51.

Bos, Gerrit / Mensching, Guido, 2005. «The Literature of Hebrew Medical Synonyms: Romance and Latin Terms and their Identification», *Aleph* 5, 169-211.

Bos, Gerrit / Hussein, Martina / Mensching, Guido / Savelsberg, Frank, 2011. *Medical Synonym Lists from Medieval Provence: Shem Tov ben Isaac of Tortosa, Sefer ha-Shimmush, Book 29*, Leiden, Brill, vol.1. [= ShS1]

Corradini, Maria Sofia, 1997. *Ricettari medico-farmaceutici medievali nella Francia meridionale*. Florence, Leo S. Olschki.

Corradini, Maria Sofia 2001. «Per l'edizione del corpus delle opere mediche in occitanico e in catalano: nuovo bilancio della tradizione manoscritta e analisi linguistica dei testi» , *Rivista di Studi Testuali 3*,127-195.

Corradini, Maria Sofia / Mensching, Guido 2010. «Les méthodologies et les outils pour la rédaction d'un Lexique de la terminologie médico-botanique de l'occitan du Moyen Age», Iliescu, Maria / Siller-Runggaldier, Heidi M. / Danler, Paul (ed.), *Actes du XXV^e Congrès International de Linguistique et de Philologie Romanes (CILPR), Innsbruck, 3-8 septembre 2007*, Berlin, Walter de Gruyter, 200-208.

Corradini, Maria Sofia / Mensching, Guido 2013. «Nuovi aspetti relativi al Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan (DiTMAO): Creazione di una base di dati integrata con organizzazione onomasiologica», Casanova, Emili /Cesareo Calvo (ed.), *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas*, Berlin, Walter de Gruyter.

DAO (Baldinger, Kurt (ed.), 1975- . *Dictionnaire onomasiologique de l'ancien occitan*, Tübingen, Niemeyer)

DT (Dietrich, Albert (ed.), 1988. *Dioscurides Triumphans. Ein anonymer arabischer Kommentar (Ende 12. Jh. n. Chr.) zur Materia medica, arabischer Text nebst kommentierter deutscher Übersetzung*, Göttingen, Vandenhoeck & Ruprecht).

FEW (Wartburg, Walther von, 1922-2002. *Französisches Etymologisches Wörterbuch*. Eine Darstellung des galloromanischen Sprachschatzes, Leipzig/Bonn/Bâle, Teubner/Klopp/Zbinden, 25 vols.)

García González, Alejandro (ed.), 2007. *Alphita. Edición crítica y comentario*, Florencia, SISMEL-Edizioni del Galluzzo.

Gutiérrez Rodilla, Bertha M., 2007. *La esforzada reelaboración del saber. Repertorio médicos de interés lexicográfico anteriores a la imprenta*, San Millán de la Cogolla, Cilengua.

LEI (Pfister, Max / Schweickard, Wolfgang, 1932- . *Lessico etimologico italiano*, Wiesbaden, Reichert.)

LF (Löw, Immanuel, 1928-1934. *Die Flora der Juden*, Viena/Leipzig, R. Löwit, reimprimido Hildesheim, Olms 1967, 4 vols.)

Mensching, Guido, 2006. «Per la terminologia medico-botanica occitana nei testi ebraici: le liste di sinonimi di Shem Tov Ben Isaac di Tortosa», Corradini, Maria Sofia / Periñán, Blanca (eds.), *Giornate di Studio di Lessicografia Romanza. Il linguaggio scientifico e tecnico (medico, botanico, farmaceutico e nautico) fra Medioevo e Rinascimento*. Atti del convegno internazionale (7-8 novembre 2003), Pisa, 93-108.

Mensching, Guido, 2009. «Listes de synonymes hébraïques-occitanes du domaine médico-botanique au Moyen Âge», Latry, Guy (ed.), *La voix occitane. Actes du VIIIe Congrès International d'Études Occitanes*, Burdeos, 509-526.

Mensching, Guido / Bos, Gerrit, 2011. «Une liste de synonymes médico-botaniques en caractères hébraïques avec des éléments occitans et catalans», Rieger, Angelica (ed.), *L'Occitanie invitée de l'Euregio. Liège 1981-Aix-la-Chapelle 2008, Bilan et perspectives (Actes du IXe Congrès International de l'Association Internationale d'Études Occitanes, Aix-la-Chapelle, 24-31 août 2008)*, Aachen, 225-238.

Mensching, Guido / Savelsberg, Frank 2004. «Reconstrucció de la terminologia mèdica occitanocatalana dels segles XIII i XIV a través de llistats de sinònims en lletres hebrees. Edició i anàlisi del vint-i-novè llibre del Sèfer ha-Ximmuix de Xem Tov ben Isaac de Tortosa», *Actas del I congrès de l'estudi dels jueus en territori de llengua catalana*, Barcelona, 69-81.

Mensching, Guido / Zwink, Julia (en imprenta): «L'ancien occitan en tant que langage scientifique de la médecine. Termes vernaculaires dans la traduction hébraïque du Zād al-musāfir wa-qūt al-ḥādir (XIIIe s.)», Actes du IXe Congrès Internationale d'Études Occitanes, Béziers 12th-19th June 2011.

Raynouard, François-Just 1838-1844. *Lexique roman: ou dictionnaire de la langue des troubadours comparée avec les autres langues de l'Europe latine*, 6 vols, Heidelberg, Winter (reimprimido Heidelberg 1970).

Wartburg, Walther von, 1956. «Die griechische Kolonisation in Südgallien und ihre sprachlichen Zeugnisse im Westromanischen», *Von Sprache und Mensch*, Bern, Francke.

Dicionário-Aberto: Construção semiautomática de uma funcionalidade codificadora

1. Introdução

O nosso projeto visa transformar o Dicionário Aberto (DA) num avançado dicionário codificador ou de produção. O que se pretende é que, utilizando as funcionalidades de pesquisa reversa do Dicionário Aberto, o utilizador possa procurar unidades lexicais relacionadas (sinónimos, quase-sinónimos, hiperónimos, hipónimos, merónimos, holónimos, co-ocorrentes, etc.) a partir de uma ou de várias palavras.

Para que isto seja possível, será necessário que o sistema de pesquisa disponibilizado aos utilizadores não use apenas os termos introduzidos pelo utilizador e as respectivas ocorrências nas definições, mas também uma estrutura semântica que providencie relações léxico-conceptuais com os termos introduzidos. Estas relações serão cruzadas e serão calculadas medidas de proximidade, sendo deste modo possível apresentar um conjunto de resultados ordenados por relevância.

Na secção 2 é apresentado o projeto, a sua origem e um pouco da sua história. De seguida, na secção 3, são apresentadas as funcionalidades básicas de pesquisa presentes em praticamente todos os dicionários disponíveis em linha. Por sua vez, a secção 4 apresenta as funcionalidades avançadas de pesquisa, como sejam a pesquisa reversa ou a pesquisa ontológica. Finalmente, na secção 5 são tiradas algumas ilações referentes ao trabalho desenvolvido.

2. O Dicionário Aberto

O DA (disponível na rede, para consulta e para extração automática de informação, em http://www.dicionario-aberto.net, mas também para uso local, de modo aberto e gratuito) iniciou em junho de 2005, como a transcrição da edição de 1913 dos dois volumes do *Novo Diccionário da Língua Portuguesa,* de Cândido de Figueiredo. Para a transcrição do dicionário usou-se uma aplicação *web* criada para ajudar na transcrição de livros para integrarem o acervo do *Projecto Gutenberg* (<www.gutenberg.org/A>). Esta aplicação está disponível num sítio da Internet, chamado PGDP (*Project Gutenberg, Distributed Proofreaders,* <www.pgdp.net>) (Newby & Franks, 2003), onde é apresentado o texto de uma página, resultante de um OCR (*Optical Character Recoginition*) que o utilizador deve ler e corrigir de acordo com a imagem disponibilizada.

O processo de transcrição, efetuado inteiramente por voluntários, terminou em 2010. Durante estes cinco anos o sítio *web* do dicionário tornou forma, e foi incorporando, diariamente, um conjunto de novas palavras que terminavam a passagem pelas várias rondas de revisão.

Desde então o dicionário tem sido sujeito a um conjunto de alterações, como a transformação num formato standard e sintaticamente rico (TEI - Text Encoding Initiative - <www.tei-c.org>), a validação da notação e, mais recentemente, a modernização da ortografia (Simões & Farinha, 2011).

Embora esta última etapa não esteja concluída, por necessitar de validação manual do resultado da ferramenta automática de modernização, o dicionário está disponível livremente não só para a pesquisa em-linha, que será seguidamente apresentada, mas também em diferentes formatos para que possa ser usado em diferentes contextos. Destes vários formatos salientamos o formato StarDict (<www.stardict.org>), usado por exemplo na ferramenta de tradução assistida por computador OmegaT (<www.omegat.org>), o formato SQL (Structured Query Language), para poder ser incorporado numa base de dados local e processada computacionalmente, ou ainda a disponibilização através de um serviço RESTful (Fielding 2000) que permite, por exemplo, o desenvolvimento de aplicações móveis do DA que consultem dinamicamente a versão mais recente disponível.

3. A pesquisa simples

A aplicação web do DA tem evoluído, incorporando diversos tipos de pesquisa: pela entrada do artigo, partes do lema (início, meio ou fim) ou ainda a pesquisa reversa, entre outras funcionalidades. Mais recentemente incorporou-se um sistema semiautomático para a extração de sinónimos/antónimos, hiperónimos/hipónimos e merónimos/holónimos que serve para explorar as relações léxico-conceptuais entre os termos introduzidos na pesquisa (Simões, Iriarte & Almeida, 2012).

Na pesquisa simples, o utilizador já vai encontrar palavras relacionadas com a entrada procurada (como sejam sinónimos, hiperónimos ou merónimos). A discussão sobre como estas palavras são obtidas e os seus potenciais problemas é feita na secção 4.

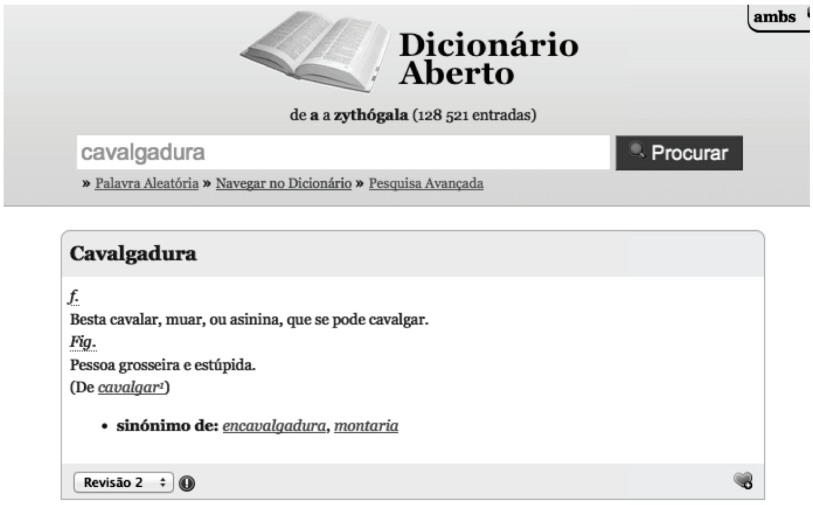


Figura 1
Resultado da pesquisa simples da palavra "cavalgadura"

A pesquisa simples também permite obter um conjunto de palavras ortograficamente semelhantes (Levenshtein, 1966), como mostra a figura 2, úteis para quando se desconhece a grafia exata de um lema.

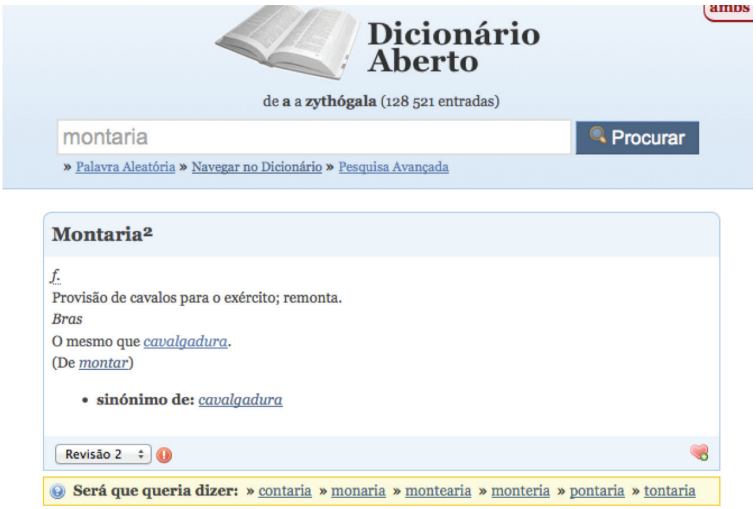


Figura 2
Procura da palavra "montaria" e apresentação de palavras ortograficamente semelhantes

Finalmente, existe uma outra funcionalidade que permite ao utilizador navegar pelas entradas do dicionário de forma sequencial, vendo as palavras ortograficamente próximas ordenadas alfabeticamente. Esta abordagem tenta assemelhar-se ao processo de folhear um dicionário. A figura 3 apresenta esta funcionalidade. Repare-se na possibilidade de obter a palavra seguinte ou anterior (presentadas acima do verbete, com setas para a esquerda e para a direita) e na lista de palavras anteriores e posteriores à entrada atual, apresentada na coluna à esquerda.

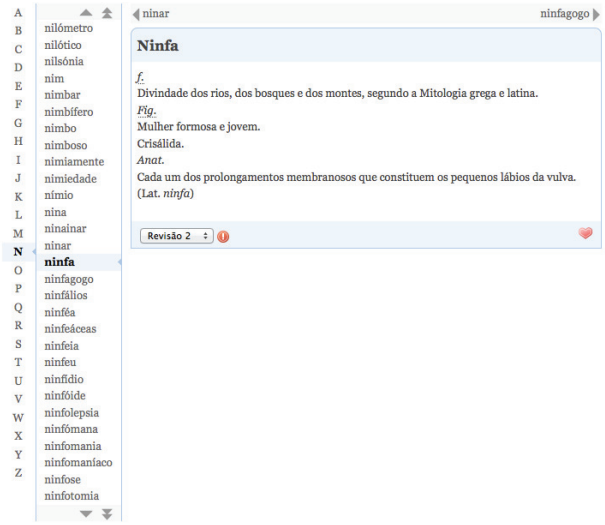


Figura 3
Verbete da palavra “ninfa” com a interface de navegação

Embora estas funcionalidades sejam úteis para grande parte das situações em que um utilizador consulta um dicionário convencional, é possível adicionar outras que tornem o dicionário muito mais útil, funcional e poderoso, como veremos na secção seguinte.

4. A pesquisa avançada

Muito mais interessantes são as funcionalidades avançadas de pesquisa, que transformam o DA num verdadeiro dicionário codificador. Utilizando as funcionalidades de pesquisa reversa ou pesquisa ontológica, por exemplo, o utilizador poderá procurar unidades lexicais relacionadas (sinónimos, quase-sinónimos, hiperónimos, hipónimos, merónimos, holónimos, co-ocorrentes, etc.) a partir de uma ou de várias palavras. Estas funcionalidades avançadas, disponíveis para utilizadores registados, estão a transformar o DA numa ferramenta muito útil para linguistas e investigadores em Processamento da Linguagem Natural.

4.1. A pesquisa por prefixo, infixo e sufixo

Após ter selecionado a "pesquisa avançada", é possível pesquisar fragmentos de lema (início, meio ou fim da palavra), selecionando na interface a palavra "Prefixo", "Infixo" ou "Sufixo". Estes termos, que esperamos poder vir a substituir, não são os mais felizes, uma vez que os resultados não correspondem a estas categorias morfológicas, mas são suficientemente transparentes e, principalmente, curtos para poderem ser utilizados confortavelmente na interface gráfica e entendíveis pelo público em geral.

Esta funcionalidade pode ser uma ferramenta importante para a morfologia (Millán, 1999). Imagine-se, por exemplo, a sua utilidade para o estudo da produtividade dos afixos, como a dos diminutivos em *-inho, -ito - ino*, etc.; a produtividade de determinados sufixos na terminologia científica: *-ato, -eto - ito*; da produtividade e verdadeira sinonímia ("ação ou resultado/efeito de") de sufixos como *–dade/ –ção/ –são, –ança/ –ância*; etc.

O DA pode vir a ser um recurso importante para a investigação linguística e para a elaboração gramáticas e de outros dicionários, uma vez que nos permite descarregar os resultados destas pesquisas. Por exemplo, a figura 4 apresenta um estudo da produção de advérbios em *-mente* a partir de adjetivos em *-vel*.

...	...
Agitável	-
Aglutinável	-
Agradável	Agradavelmente
Agradecível	-
Agricultável	-
Ajuntável	-
Alcançável	-
Alcoolizável	-
Alheável	-
Aliável	-
Alienável	-
Alliável	-
Alterável	-
Amável	Amavelmente
Amigável	Amigavelmente
Amissivel	-
Amoedável	-
Amoldável	-
Amolgável	-
Amorável	Amoravelmente
Amortizável	-
Amotinável	-
Amovível	-
Amparável	-
...	...

Figura 4
Análise do léxico tendo por base listas de palavras descarregadas a partir do Dicionário-Aberto

A pesquisa por sufixos pode também ser usada a modo de dicionário de rimas gráficas. Também é fácil incluir a capacidade de pesquisa inversa: transformando o DA num dicionário inverso, em que a ordenação alfabética dos lemas é feita da direita para a esquerda.

4.2. A pesquisa reversa

A pesquisa reversa transforma os dicionários eletrónicos numa espécie de base de dados conceptual, que funciona a modo de dicionário onomasiológico ou de dicionário de produção ou codificador.

Utilizando as funcionalidades de pesquisa reversa do DA, o utilizador pode procurar unidades lexicais relacionadas (sinónimos, quase-sinónimos, hiperónimos, hipónimos, merónimos, holónimos, co-ocorrentes, etc.) a partir de um conjunto de palavras. Um exemplo das potencialidades como dicionário codificador ou dicionário de produção: introduzindo nas janelas de pesquisa as palavras "endurecer" e "metal" obtemos como resultado a entrada "temperar" (ver figura 5).

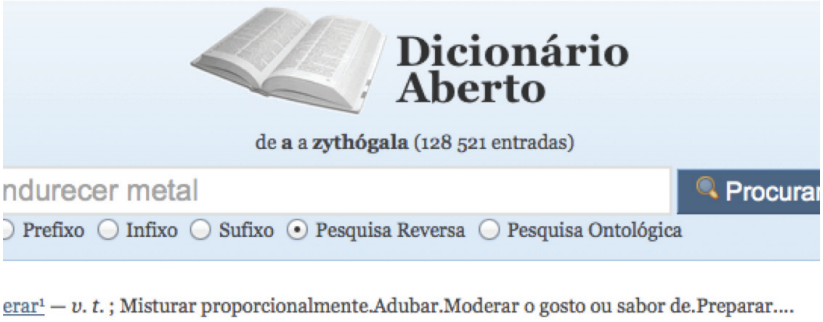


Figura 5

Pesquisa reversa das palavras "endurecer" e "metal", obtendo o resultado "temperar"

Esta pesquisa é efetuada sobre as definições e não sobre os lemas, o que acaba por ser uma limitação, já que só serão encontradas entradas que usem exatamente as palavras usadas na pesquisa. A esse respeito existe um conjunto de melhorias a serem implementadas a este tipo de pesquisa, nomeadamente:

- o uso de um analisador morfológico para que não sejam indexadas as formas das palavras ocorrentes das definições, mas os seus lemas, e para que as palavras procuradas sejam também lematizadas antes de serem usadas na pesquisa. Embora esta funcionalidade seja relativamente fácil de implementar levará a que existam alguns problemas na cobertura do analisador morfológico;
- as palavras pesquisadas podem ocorrer em qualquer sítio da definição. Seria bastante interessante que fosse possível procurar por uma sequência de palavras exata, tal como acontece com os motores de pesquisa na Internet.

4.3. A pesquisa ontológica

Muito mais interessante e prometedor é o que chamamos de "pesquisa ontológica". Esta pesquisa é baseada numa ontologia construída dinamicamente (e portanto, de forma completamente automática). De seguida resume-se a abordagem utilizada na extração automática da ontologia a partir das definições (Simões, Álvaro a Almeida, 2012), e posteriormente é explicado o algoritmo de pesquisa baseado nesta ontologia.

Para a extração automática de relações léxico-conceptuais usa-se um conjunto de regras ou padrões (Hearst 1992). Estes padrões são sequências de palavras que se pretendem encontrar nas definições e que indicam a grande probabilidade da palavra que segue o padrão estar relacionada com o lema da respetiva definição. Isto só é possível graças à regularidade da estrutura das definições.

Seguem-se alguns exemplos de regras (uma por relação léxico-conceptual) que foram usadas na extração de relações:

- Sinonímia (SYN): *o mesmo (ou melhor) que ...* ;
- Antonímia (ANT): *que não é ...* ;
- Hiponímia (HIPO): *espécie de ...* ;
- Hiperonímia (HIPER): *que tem por tipo...* ;
- Meronímia (MERO): *cada uma das partes que formam ...* ;
- Holonímia (HOLO): *composto por...* .

Para esta funcionalidade, também se usam relações calculadas (por exemplo, utilizando a transitividade da relação de hiperonímia). Somos conscientes de que algumas regras podem ser problemáticas (por exemplo, no caso dos sinónimos e quase-sinónimos). Em todo o caso, parece-nos preferível apresentar um conjunto de possíveis falsos sinónimos do que garantirmos a correção e diminuir drasticamente o número de relações da ontologia resultante. Note-se que uma ontologia torna-se mais útil quando há grande diversidade nos relacionamentos existentes.

Assim, definiram-se regras matemáticas para a "completação" da ontologia, inferindo novos relacionamentos a partir dos relacionamentos iniciais:

- a SYN b ⇒ b SYN a – propriedade de simetria entre sinónimos: se *a* é sinónimo de *b,* então *b* também é sinónimo de *a*. Esta relação é bastante produtiva já que em muitas situações de verdadeiros sinónimos o dicionarista definiu a relação de sinonímia apenas num dos verbetes das palavras envolvidas;
- a HIPO b ∧ c HIPO b ⇒ a COHIPO c – a relação de co-hiponímia não pode ser obtida diretamente usando padrões de Hearst, mas pode ser calculada a partir das relações de hiponímia. Assim, se duas palavras *a* e *c* são hipónimos da mesma palavra *b,* então *a* e *c* são co-hipónimos. Esta relação é extremamente útil já que tipicamente relaciona objetos de um mesmo tipo ou classe;
- a HIPO b ∧ b HIPO c ⇒ a HIPO c – a transitividade das relações hierárquicas, como é o caso da hiperonímia ou hiponímia, permite que se possa procurar termos genéricos com base num termo bastante específico ou termos específicos usando termos bastante genéricos. Como exemplo prático, considere-se a pesquisa pelo termo "animal". É pouco provável que se encontrem entradas referentes a animais diretamente relacionados com

> este termo, mas por transitividade poderemos encontrar outras classes como "mamífero" ou "peixe" como hipónimos de "animal", e como hipónimos destes, encontrar entradas correspondentes a animais.

Na pesquisa ontológica, os resultados da pesquisa são as entradas que se relacionam com essas palavras (ordenadas por quantidade de relacionamentos). Repare-se que, ao contrário do que acontece com a pesquisa reversa, alguns dos resultados são entradas que não contêm nenhum dos termos introduzidos pelo utilizador na janela de pesquisa. Assim, no exemplo recolhido na figura 5, o terceiro resultado, entre outros, corresponde a uma entrada que não contém os termos introduzidos pelo utilizador ("veado" e "animal").

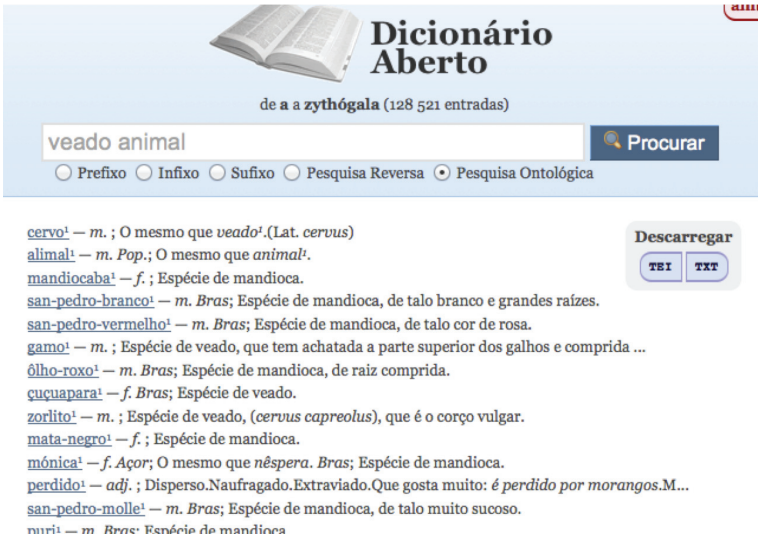


Figura 6

Pesquisa ontológica a partir dos termos "veado" e "animal"

Tal como no caso da pesquisa reversa, a pesquisa ontológica pode beneficiar de algumas novas funcionalidades. Se por um lado a ontologia extraída poderá lucrar em ser expandida usando, por exemplo, uma ontologia externa como o Onto.PT (Gonçalo Oliveira / Gomes, 2012), por outro lado os termos de pesquisa deveriam ser lematizados antes de serem usados no algoritmo de procura. Além disso, a possibilidade de incorporar funcionalidades de pesquisa reversa juntamente com as funcionalidades de pesquisa ontológica seria muito proveitosa.

4. Conclusões

Estamos convictos de que o DA é uma excelente ferramenta que pode ser usada como um dicionário tradicional, mas também como um recurso para tarefas de processamento de linguagem natural e como ferramenta para ajudar na verificação de hipóteses colocadas na investigação linguística e na elaboração gramáticas e de outros dicionários.

O facto de ser um recurso aberto poderá também significar que as suas funcionalidades poderão vir a aumentar em quantidade e em qualidade. Mas, por outro lado, o facto de não ser um projeto financiado limita a possibilidade de contratação de mão de obra especializada para a revisão exaustiva, por exemplo, da modernização da língua. Ou seja, a evolução da qualidade das entradas do DA está limitada à disponibilidade de voluntários.

Por outro lado, e embora a evolução do conteúdo linguístico do dicionário seja lenta, a implementação de algoritmos e a execução de experiências sobre o DA tem sido bastante proveitosa, demonstrando que é possível criar funcionalidades úteis a partir de dicionários convencionais.

Universidade do Minho - Portugal — Alberto SIMÕES
Universidade do Minho - Portugal — Álvaro IRIARTE
Universidade do Minho - Portugal — José João ALMEIDA

Referências bibliográficas

Almeida, José João / Santos, André / Simões, Alberto, 2010. «Bigorna – a toolkit for orthography migration challenges», in: Calzolari, N., et al., (ed.), *Seventh International Conference on Language Resources and Evaluation (LREC 2010),* Valletta, Malta, 227–232.

Fielding, Roy T. 2000. «Representational State Transfer (REST)», *Architectural Styles and the Design of Network-based Software Architectures*, PhD Dissertation, Irvine, University of California.

Gonçalo Oliveira, Hugo / Gomes, Paulo, 2012. «Ontologising semantic relations into a relationless thesaurus», *Proceedings of 20th European Conference on Artificial Intelligence (ECAI 2012)*, Montpellier, France, August 2012. IOS Press, 915-916.

Hearst, Marti, 1992. «Automatic acquisition of hyponyms from large text corpora», *Proceedings of the Fourteenth International Conference on Computational Linguistics,* Nantes, France, Volume 2, 539-545.

Levenshtein, Vladimir Iosifovich, 1966. «Binary codes capable of correcting deletions, insertions, and reversals», *Soviet Physics Doklady* 10, 707-710.

Millán, José Antonio, 1999. «Zigzag, gong, ping-pong, iceberg. donde se descubre que hay diccionarios inversos, y su utilidad manifiesta para el progreso de la humanidad» <jamillan.com/inverso.htm> (último acesso em 1 de julho de 2013).

Millán, José Antonio, 2011. «Dirae: consulte el DRAE como ya no podía hacerlo» <jamillan.com/librosybitios/2011/05/dirae-consulte-el-drae-como-ya-no-podia-hacerlo> (último acesso em 1 de julho de 2013).

Newby, Gregory B. / Franks, Charles, 2003. «Distributed proofreading», *Proceedings of the Joint Conference on Digital Libraries,* 361-363.

Simões, Alberto / Farinha, Rita, 2011. «Dicionário Aberto: Um novo recurso para PLN», *Vice-versa* 16, 159-171.

Simões, Alberto / Iriarte, Álvaro / Almeida, José João, 2012. «Dicionário-aberto – a source of resources for the portuguese language processing», in: Caseli, H. / Villavicencio, A. / Teixeira, A., / Perdigão, F. (ed.) *Computational Processing of the Portuguese Language, Lecture Notes for Artificial Intelligence,* Berlim, Springer, 7243, 121-127.

De la théorie à la pratique. Le projet SOCIODIC : un premier dictionnaire roumain de terminologie sociolinguistique*

Le projet en cours que nous présentons ici a été développé grâce à une collaboration entre l'Université de Pitești (Roumanie) et l'Université de Neuchâtel (Suisse) ; il est financé par le programme SCIEX helvétique, qui favorise les échanges scientifiques entre la Suisse et les pays de l'Est européen (projet n° 12.026).

En comparaison avec les pays anglo-saxons, mais aussi avec la France, la sociolinguistique est moins développée en Roumanie que d'autres disciplines de la linguistique. Le nombre de travaux consacrés à la question est relativement restreint et se limite le plus souvent au domaine roumain.

Notre projet se propose de réaliser pour la première fois en Roumanie et en langue roumaine, une vue d'ensemble de la recherche en sociolinguistique à travers l'élaboration d'un dictionnaire terminologique de cette discipline. Nous visons à créer un instrument de travail pour les chercheurs et les étudiants roumains, qui leur facilitera l'accès aux connaissances actuelles et qui contribuera en même temps à créer en roumain le métalangage – la terminologie – nécessaire au développement de la recherche en roumain.

Notre dictionnaire cherche à recenser et à définir l'essentiel de la terminologie utilisée par les différents courants de la recherche sociolinguistique internationale. Il cherche en particulier à désambiguïser les nombreux termes qui – étant donné l'hétérogénéité croissante de la recherche en sociolinguistique – sont employés avec différentes significations dans les différentes traditions académiques. Les concepts étudiés sont régulièrement exemplifiés et accompagnés par les principales références bibliographiques disponibles dans le domaine ; chaque entrée est aussi accompagnée

* Je tiens à remercier chaleureusement Andres Kristol qui m'a accueillie à l'Université de Neuchâtel, qui a créé pour moi la base de données FileMaker Pro qui sert à l'élaboration du dictionnaire et l'a gracieusement mise à ma disposition, qui a attentivement relu mon texte et discuté avec moi de nombreuses questions évoquées dans ces lignes. Mes remerciements vont également à Madame Sanda-Marina Bădulescu, professeure à l'Université de Pitești : Vă mulțumesc atât pentru discuțiile utile și încurajările permanente cât și pentru asigurarea unor materiale necesare în tot acest demers. Je suis très reconnaissante aussi à Raphael Maître, rédacteur au Glossaire des patois de la Suisse romande (Université de Neuchâtel) pour ses précieux conseils scientifiques qui m'ont beaucoup aidée et aussi pour sa relecture de ce texte. Mes remerciements vont enfin au réseau de mes collègues, locuteurs roumains natifs – non spécialistes en sociolinguistique – auxquels j'ai pu soumettre mes hésitations terminologiques.

de ses équivalents français et anglais (cf. la copie d'écran du masque de saisie reproduite ci-dessous).

Dans ce qui suit, nous illustrerons la spécificité des définitions terminologiques en comparaison avec les définitions lexicographiques, nous présenterons notre méthodologie, la typologie des entrées et toutes les contraintes et difficultés liées à la conception du travail terminologique.

1. Définitions terminologiques vs définitions lexicographiques

Malgré certaines caractéristiques communes dans le fond et la forme, dans la conception des articles (p.ex. absence de l'article défini dans la formulation du lemme), la différence entre les dictionnaires de langue et les dictionnaires terminologiques est profonde. Leur finalité est différente. La terminologie identifie des concepts et leur associe des termes, tandis que la lexicographie décode des unités lexicales et en décrit le sens ou les différentes significations (Vézina et al., 2009 : 6) : « La définition terminologique s'attache à décrire, à énoncer un concept (ou notion) désigné par un terme […] et à le caractériser par rapport à d'autres concepts à l'intérieur d'un système organisé (appelé système conceptuel), tandis que la définition lexicographique cherche à décrire le ou les sens (signifié) d'une unité lexicale ». Dans le domaine de la terminologie, le concept revêt uniquement une dimension désignative ou dénotative, alors que dans la lexicographie générale, le signifié a souvent une dimension connotative et « culturelle qui lui confère une plus grande richesse sémantique, laquelle témoigne entre autres de la mentalité, des croyances, des attitudes, des goûts ou des us et coutumes des locuteurs d'une langue » (Vézina et al., 2009 : 6).

Les vertus d'une bonne terminologie sont bien connues (Neveu, 2009 : 5) : économie (la parcimonie est le garant présumé de l'objectivité), la transparence (les termes requis doivent être univoques), la cohérence (l'absence de contradictions internes). En ce qui nous concerne, parmi les principes définitoires que nous avons pris en compte nous citons les principes de concision, clarté, explicitation, adéquation, non-tautologie, généralisation et abstraction.

2. Méthodologie

Notre projet vise à apporter une contribution sur deux volets. D'une part, il cherche à contribuer au développement, à l'enrichissement et à la standardisation de la terminologie de la sociolinguistique en roumain, et d'autre part, comme nous l'avons déjà mentionné, nous espérons que sur un plan plus concret et plus pratique, notre dictionnaire pourra devenir un instrument de travail pour les étudiants, les enseignants et tous ceux qui s'intéressent à ce domaine de recherche.

Dans un premier temps, nous avons constitué notre nomenclature en dépouillant les travaux spécialisés en français, en anglais et en roumain – actuellement notre bibliographie comprend près de 300 références : dictionnaires terminologiques de

sociolinguistique (Swann et al. 2004, Trudgill 2003, Simonin / Wharton 2013, Moreau 1997), encyclopédies linguistiques (Ammon et al. 2004, Goebl et al. 1996, Holtus/Metzeltin/Schmitt 1996, Holtus/Günter 1990), monographies, revues, articles en ligne ou sur papier, en roumain, anglais, français, allemand, espagnol et italien.

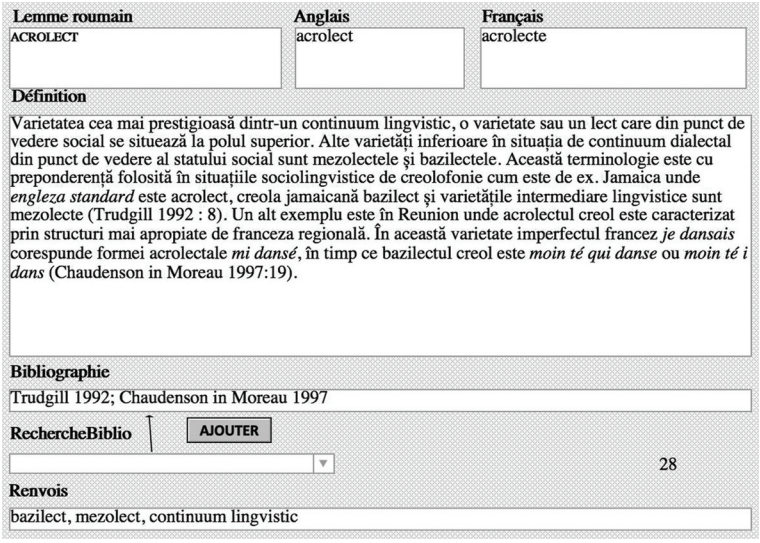
Nous avons cherché à explorer les différentes traditions et approches qui composent la sociolinguistique en rattachant les termes et les concepts à ses différents champs d'étude : standardisation, planification des langues, comportement bilingue et multilingue, stratification sociale du langage, structure de la communication, attitudes envers le langage, pidginisation et créolisation, variation de la langue et changement linguistique. La délimitation de la terminologie retenue n'est pas stricte : elle dépasse les frontières de la sociolinguistique vers la dialectologie, l'anthropologie, la psychologie, la sociologie et la linguistique générale, en fonction des chevauchements terminologiques entre ces disciplines. Nous avons inclus tout d'abord les termes courants, susceptibles d'être rencontrés par les étudiants dans les textes académiques, mais nous tenons également compte d'une terminologie plus pointue, susceptible d'intéresser des chercheurs spécialistes des différents domaines de recherche en sociolinguistique. La gestion de toutes les difficultés liées à l'hétérogénéité des définitions trouvées dans la littérature a constitué un des défis majeurs de notre entreprise. Une petite partie de nos entrées est également consacrée aux sociolinguistes mêmes qui ont fondé cette discipline ou qui ont contribué de manière significative au développement de ses théories et de ses méthodes de recherche (Labov, Hymes, etc.).

L'ensemble du projet – articles rédigés et bibliographie – est géré par une base de données relationnelle FileMaker Pro, qui nous permettra à la fin du projet de générer automatiquement le dictionnaire.

3. Typologie des entrées

Pour la création des fiches individuelles – et donc des articles correspondants de notre futur dictionnaire – nous nous sommes inspirée du format du dictionnaire de terminologie linguistique d'Abraham 1988, qui nous semblait bien correspondre à nos besoins. Chaque fiche se compose en principe de six champs (cf. l'illustration ci-dessous) : une entrée, qui se compose du lemme roumain avec ses correspondances en anglais et en français, la définition proprement dite, les références bibliographiques et des renvois à d'autres entrées. Selon les articles, les définitions sont de longueur très variable : dans certains cas, une définition brève et simple est entièrement suffisante ; dans d'autres cas, une notion a besoin d'être étudiée de manière plus nuancée et plus développée. Le champ «bibliographie» fournit sous forme brève toutes les références bibliographiques utilisées dans la définition ; une section bibliographique détaillée rassemblera à la fin de l'ouvrage toutes les indications bibliographiques présentées à l'intérieur des articles. Le dernier champ de chaque entrée est constitué par les renvois à d'autres entrées, qui sont destinés à aider l'utilisateur à trouver des termes connexes et à mieux comprendre ainsi un groupe de concepts.

Les articles sont strictement classés par ordre alphabétique ; chaque entrée traite un seul terme. Ainsi, *diglossie*, *microdiglossie* et *macrodiglossie* constituent trois articles distincts, avec des renvois de l'un à l'autre, évidemment. La plupart des lemmes sont transparents, ou s'expliquent grâce à la définition. Ce n'est que très rarement nous avons également fourni certaines précisions étymologiques utiles à la compréhension du sens[1].



Copie d'écran d'un article modèle dans la base de données FileMaker

4. Typologie de la création terminologique

Un des principaux objectifs de notre démarche a été de recenser la terminologie déjà existante dans la recherche sociolinguistique en roumain et à la compléter en fonction des principales recherches disponibles dans les autres langues. Notre nomenclature englobe actuellement 521 entrées dont 128 sont effectivement nouvelles, et qui ont été minutieusement analysées du point de vue formel et sémantique. Pour les besoins de notre travail, nous distinguons cinq catégories de néologismes : néologismes graphiques / phonétiques, néologismes morphologiques, néologismes sémantiques, calques et néologismes d'emprunt. L'étude de ces cinq types de néologismes fait l'objet d'un autre travail (Ungureanu 2013 sous presse) ; nous n'en parlerons pas ici.

[1] C'est le cas par exemple pour *Chicano*: le mot *Chicano* (plus rarement écrit *Xicano*) partage une origine commune en nahuatl avec *México* et *mexicano*. *Mexica* [meˈʃika] serait le nom utilisé par les Aztèques pour désigner le 'territoire' (cf. http://fr.wikipedia.org/wiki/ Chicano, consulté le 30.05.2013).

5. Typologie des difficultés

5.1. Difficultés de traduction trilingue (roumain – anglais – français)

Certains termes forgés en anglais surtout, plus rarement aussi en français, sont difficiles à rendre en roumain, non seulement dans la traduction du lemme, mais aussi dans celle des définitions telles qu'on les trouve chez les auteurs qui les ont proposées. Dans ces lignes, nous discuterons quelques exemples significatifs des difficultés que nous avons rencontrées. Dans plusieurs cas, nous avons invité nos collègues proches ou lointains à prendre part à notre réflexion.

5.1.1. angl. translanguaging → roum. comunicare translangajieră

Le terme translanguaging a été proposé par Ofelia García (2009) pour désigner les pratiques multiples discursives dans la perspective des locuteurs plurilingues, souvent dans un contexte didactique, dans l'enseignement des langues ; la notion inclut celle de l'alternance codique. À notre connaissance, aucune langue romane ne connaît actuellement de terme équivalent ; la notion a donné du fil à retordre à plusieurs participants du Congrès du Réseau francophone de sociolinguistique (RFS) de Corse (31 juillet-5 juillet 2013) avec qui nous en avons discuté. Après avoir écarté « translinguisme », « translanguer » et « translanguage », qui ne satisfont pas aux critères traductionnels, nous avons opté pour le syntagme « communication translangagière » (en français)[2], « comunicare translangajieră » (en roumain), qui y satisfait tout en présentant l'avantage de la transparence sémantique.

5.1.2. angl. genderlect → roum. lect de gen

La notion de *genderlect* appartient au domaine des *Gender Studies* ; ce dernier concept a été traduit en français par *études genre*, et en roumain par *studii de gen*. Swann et al. (2004 : 122) définissent *genderlect* comme « a constellation of linguistic features associated either with female or male speaker ». Dans ce cas aussi, nous avons cherché en vain une traduction française ou roumaine dans la littérature disponible ; les travaux français correspondants utilisent constamment l'anglicisme, avec des guillemets ou en italiques, pour marquer l'emprunt non intégré. Fidèle au concept de notre dictionnaire qui cherche à élaborer une terminologie de la sociolinguistique *en roumain*, nous avons décidé de procéder à la création d'un néologisme.

Une première éventualité que nous avons envisagée, était de créer *genolect*, sur le modèle de *sociolecte, régiolecte, topolecte*, *idiolecte,* etc., disponible dans d'autres langues romanes et également introduits en roumain (*sociolect,* etc.). En réalité, une analyse rapide permet de constater que le préfixe *geno-* pourrait induire en erreur quant au sens linguistique voulu : il est occupé par les notions de *race* ou de *gène,* comme par exemple dans *génocide* ou *génotype*, comme le montrent les définitions proposées par le *DEX* 1998 (sous *geno-*) : « care produce, care ia naștere; specie, gen,

[2] Proposition de Raphaël Maître.

fel ; familie, neam ». C'est d'ailleurs sans doute aussi la raison pour laquelle, en anglais, c'est *genderlect* qui s'est imposé et non pas **genolect*.

Sexolecte ? Ce terme apparaît en effet dans certaines publications en français comme équivalent de l'anglais *genderlect* : une simple recherche sur Google donne environ 2050 occurrences. Même s'il est prononçable, le mot fait pourtant tout à fait bizarre en roumain (aucune occurrence pour «sexolect» chez Google), en renvoyant à des connotations sexuelles malvenues. On sait en effet que depuis les années 1980, la notion de *genre* a été préférée dans les sciences sociales, pour pouvoir faire la distinction entre *sexe* comme attribut biologique et *genre* comme attribut social (Swann et al. 2004 : 122). Le sexe/genre a été une variable sociale dans plusieurs études variationnistes (liées à la variation de la langue et au changement de langue) ou dans la sociolinguistique interactionnelle qui a identifié « féminin » et « masculin » dans les styles conversationnels. Ces *lectes* n'étant pas des attributs sexuels, mais des comportements sociaux, « sexolecte » est évidemment à proscrire.

Dans ces circonstances, nous nous sommes tournée vers une troisième option : fr. *lecte de genre* / roum. *lect de gen*. La tournure nous semble satisfaisante d'un point de vue sémantique et formel, même si elle peut légèrement décontenancer à première vue. En effet, on trouve bien les termes *lecte* et de *genre* en sociolinguistique, mais pas les syntagmes *lecte de couche sociale*, *lecte de lieu*, etc. Il ne nous appartient évidemment pas d'introduire en français le syntagme *lecte de genre*, qui n'est pas consacré par l'usage, mais en ce qui concerne le roumain, *lect de gen* nous semble la seule solution viable. En fait, une petite « enquête d'acceptabilité », dans notre petit réseau de locuteurs natifs, a confirmé que la tournure était tout à fait recevable. Pour *genolect*, en revanche, l'impression unanime a été que par sa forme, le mot renvoie toute suite à *genocid*.

5.1.3. fr. alémanique → *roum.* alemanic

Depuis Ferguson (1959), le statut des dialectes alémaniques par rapport à l'allemand « standard », en Suisse, peut être considéré comme un cas classique de diglossie, une diglossie qui, par la suite, a été définie comme *médiale*, les dialectes alémaniques de Suisse pouvant se trouver en position *haute* par rapport à la langue standard, dans certains usages oraux surtout. La situation est différente au sud de l'Allemagne, en Alsace et dans les parties occidentales de l'Autriche, où les dialectes alémaniques sont actuellement menacés. Dans ce cas, ce qui nous a facilité la tâche de la rédaction de cette problématique en roumain, c'est le fait que le néologisme *alemanic* (qui ne se confond évidemment pas avec *german* ou *germanic* 'allemand') est recensé par le DOOM 2 (2005).

5.1.4. roum. alolingv

Le problème de la traduction trilingue, prévue dans le « chapeau » de chacun de nos articles, peut aussi se poser dans le cas où un terme technique spécifique n'est attesté qu'en roumain. C'est le cas par exemple d'*alolingv* : cette forme a été créée en

République de Moldavie pour désigner une personne qui n'est pas de langue maternelle roumaine ou n'utilise pas couramment la langue roumaine. Comme le terme se réfère à une situation sociolinguistique spécifique, nous n'avons pas voulu le traduire par la notion, proche pourtant, d'*alloglot/alloglotte*. Dans un tel cas, nous laissons donc le lemme roumain sans correspondance en français ou en anglais.

5.2. Difficultés d'homogénéisation des définitions

Un autre aspect problématique auquel nous avons été confrontée a été la diversité des approches et des points de vue pour certaines notions, divergences provenant surtout de la différence entre la tradition universitaire française et anglaise. Comment tenir compte de plusieurs définitions, dans le dictionnaire, sans tomber dans un trop long exposé des différentes théories dans lesquels elles s'inscrivent ? Nous avons opté pour la présentation succincte des différents emplois des notions retenues en nous référant aux auteurs qui les ont définies, tout en essayant d'homogénéiser les définitions autant que possible.

5.2.1. Alternance codique

Dans la tradition anglo-américaine, la notion d'*alternance codique* (*code switching*) est liée aux domaines de la linguistique de contact et du bilinguisme (cf. Simonin / Wharton 2013 : 44), alors que dans la recherche française « ce champ d'analyse […] s'est développé tant dans des perspectives sociolinguistiques, interculturelles ou didactiques que linguistiques » (Canut, 2002 : 9).

Dans la littérature on rencontre souvent les définitions de Gumperz (1989 : 57) : « La juxtaposition à l'intérieur d'un même échange verbal de passages où le discours appartient à deux systèmes ou sous-systèmes grammaticaux différents », de Milroy / Muysken (1995 : 7) : « l'utilisation alternée par des bilingues de deux langues ou plus au sein d'une même conversation », de Heller (1988 : 1) : « l'utilisation de plus d'une langue dans le cours d'un même épisode communicatif » ou bien de Winford (2003 : 11) : « comportement des bilingues qui exploitent les ressources des langues qu'ils maîtrisent de diverses manières, pour des buts sociaux et stylistiques, et accomplissent cela en passant d'une langue à l'autre, ou en les mélangeant de différentes manières ».

L'homogénéisation de ces différentes définitions est difficile dans le sens que les différentes traditions définissent *alternance codique* différemment. En tenant compte de ces différentes approches, en partie complémentaires, nous définissons *alternanţă codică / schimbare de cod* comme suit :

> Desemnează capacitatea individului de a trece de la o limbă, un dialect sau un stil la altele, în cursul interacțiunii verbale (Ionescu-Ruxăndoiu / Chițoran 1975:290).
>
> Conform structurii sintactice a segmentelor alternate schimbarea de cod poate fi (Thiam in Moreau 1997:32):

a. intrafazică cînd structurile sintactice care aparțin a două limbi diferite coexistă în interiorul aceleiași fraze

b. interfrazică, este o alternanță de limbi la nivelul unităților mai lungi, de fraze sau fragmente ale discursurilor în vorbirea aceluiași locutor

c. extrafazică când segmentele alternate sunt expresii idiomatice, proverbe etc.

Aceste segmente pot să varieze deci în lungime, pornind de la un item la un enunț, trecând printr-un grup de cuvinte, o propoziție sau o frază cu condiția ca segmentele în cauză să nu fie integrate în sistemul morfo-sintactic și lexical al limbii în care se ține discursul de bază (Laroussi / Marcellesi in Goebl et al. 1996:196). Nu vorbim așadar de schimbare de cod dacă constatăm că un locutor folosește o limbă cu superiorii săi și o limbă când vorbește cu cei apropiați; ca să fie schimbare de cod trebuie ca cele două coduri să fie utilizate în același context (Thiam in Moreau 1997:33). […]

Schimbarea de cod poate fi condiționată de schimbarea de situație (schimbare situațională) sau poate avea o motivare individuală, emfatică sau de contrast (schimbare metaforică).

Schimbarea situațională se bazează pe existența unei relații precis determinate între anumite varietăți ale limbii și anumite caracteristici ale situației de comunicare (subiecte, locuri, participanți, scopuri). Ea poate fi personală, dacă interacțiunea verbală are loc între indivizi legați prin relații personale, sau tranzacțională, dacă indivizii se află într-o relație cu un caracter oficial mai pronunțat.

Schimbarea metaforică presupune nu numai existența unui set comun de norme situaționale, ci și existența aceleiași concepții cu privire la inviolabilitatea lor; altfel intervine riscul înțelegerii greșite sau contradictorii a acestei schimbări (Ionescu-Ruxăndoiu / Chițoran 1975: 290).

5.2.2. Angl. vernacular, fr. vernaculaire

La « paire » vernacular / vernaculaire est particulièrement significative pour les divergences qui existent dans l'utilisation d'une notion apparemment identique dans le monde anglophone et francophone. À beaucoup d'égards, il s'agit ici de véritables « faux amis ». Un regard sur l'étymologie du mot ne nous avance pas beaucoup : dans l'Antiquité, vernaculus signifie « relatif aux esclaves nés dans la maison », puis, de manière figurée « qui est du pays, indigène, national » (TLFi, s.v., qui reproduit les définitions du dictionnaire latin-français de Gaffiot). La forme moderne, savante, a été formée avec l'ajout du suffixe -arius > -aire ; en français, vernaculaire est attesté depuis 1765 avec le sens de « tout ce qui est particulier à un pays ».

En anglais, vernacular a un sens assez large. Le mot renvoie à la notion de « dialect », avec sa définition caractéristique pour la linguistique anglo-saxonne (cf. 5.2.3., ci-dessous) : « language or dialect of a speech community e.g. the vernacular of Liverpool, Jamaica etc. » (Crystal 1985), mais aussi à celle de langue d'usage quotidien: « the current daily speech of a people or geographical area as distinguished from the literary language » (Pei 1966).

Labov (1972) définit vernacular comme « the form of language first acquired, perfectly learned, and used only among speakers of the same vernacular ». Swann et al. (2004 : 327) proposent deux acceptions du terme vernacular comme « variétés non-standard » et « style de discours informel » (casual speech) :

a. Refers to relatively homogenous and well-defined non-standard varieties which are used regularly by particular geographical, ethnic or social groups and which exist in opposition to a dominant (not necessarily related) standard variety (such as, for example, African American Vernacular English in the United states). b. According to William Labov, the most casual speech style in the linguistic repertoire of a speaker. The vernacular is used when talking to friends and family in informal contexts. It is acquired in childhood and is believed to be linguistically more regular than more formal, careful speech styles, which typically show varying degrees of influence from standard varieties or other local high-prestige varieties.

En français, vernaculaire a plusieurs définitions. Le sens le plus connu est celui qui l'oppose à véhiculaire : « langue parlée seulement à l'intérieur d'une communauté, parfois restreinte (par opposition à langue véhiculaire) »[3]. Calvet in Moreau (1997 : 292) propose de réserver l'appellation de langue vernaculaire à « une langue utilisée dans le cadre des échanges informels entre proches du même groupe, comme par exemple dans le cadre familial, quelle que soit sa diffusion à l'extérieur de ce cadre ». Abecassis (2009 : 285) considère que vernaculaire représente « les variétés de français utilisées par les couches inférieures de la population ». Le créoliste Valdman (2000 :1180) parle d'une variété langagière stigmatisée par la société, « associée au sous-prolétariat parisien » qui est issue à l'origine « des parlers des masses ouvrières et paysannes » (Valdman, 1982 : 226). Le Québécois Corbett (1990 : 222) propose : « vernaculaire du français, parlé par la majorité de la population du Québec ».

Dans un tel cas, face aux divergences entre les définitions en anglais et en français (et entre les différentes définitions proposées en français même), nous avons décidé de synthétiser toute l'information et de présenter objectivement les traditions divergentes[4].

5.2.3. Angl. dialect*, fr.* dialecte

Des problèmes comparables surgissent pour la « paire » dialect / dialecte dont les définitions divergent considérablement entre la tradition anglo-saxonne et la linguistique romane. Les dialectes primaires du monde anglo-saxon ayant pratiquement disparu, dialect of English désigne actuellement ce que la linguistique romane appelle variété régionale d'une langue commune (français régional, italien régional). Dans de tels cas, nous pensons que la confrontation des différentes définitions, dans notre dictionnaire, pourra réellement rendre service aux étudiants débutants qui risquent

[3] Cf. *Dictionnaire français Larousse*, en ligne : ‹www.larousse.fr/dictionnaires/francais/vernaculaire/81591?q=vernaculaire# 80626›.

[4] Les limites de place imposées ici ne nous permettent malheureusement pas de donner comme illustration le texte qui se trouvera dans le dictionnaire.

de se trouver insécurisés, dans leurs lectures, par ces contradictions apparentes – et nous avons constaté que même en français, certains chercheurs mal informés commencent à utiliser « dialecte » dans le sens anglo-saxon, au lieu de réserver son emploi aux dialectes primaires du monde latin.

6. Conclusions

Notre dictionnaire balise un espace central de la recherche en sociolinguistique à partir de sa nomenclature. Il enregistre la terminologie en usage dans les principaux travaux de sociolinguistique en anglais, en français, en roumain et – à un degré moindre – dans d'autres langues aussi. Nous souhaitons ainsi offrir au lecteur un outil de travail utile et performant sans éviter les questions ouvertes, et en apportant aussi un éclairage ponctuel sur les notions qui ne posent pas de problèmes définitoires majeurs.

Université de Pitești Cristina UNGUREANU

Références bibliographiques

Abecassis, Michaël, 2009. « La représentativité du français parisien dans le cinéma des années 30-40 », in : Aquino-Weber, D. / Cotelli, S. / Kristol A. (ed.), *Sociolinguistique historique du domaine gallo-roman. Enjeux et méthodologies d'un champ disciplinaire émergeant*, Berne, Lang, 283-317.

Abraham, Werner 1988. *Terminologie zur neueren Linguistik*, Tübingen, Niemeyer 1988

Alby, Sophie, 2013. « Alternances et mélanges codiques », in : Simonin, Jacky et Wharton, Sylvie (éd), *Sociolinguistique du contact: Dictionnaire des termes et concepts*, Lyon, ENS Editions, 43-70.

Ammon, Ulrich et al., 2004. *Sociolinguistics / Soziolinguistik. An International Handbook of the Science of Language and Society / Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*, Berlin / New York, Mouton de Gruyter.

Avram, Mioara, 1990. *Ortografie pentru toți*, București, Editura Academiei

Calvet, Louis-Jean, 1997. « Verlan », in : Moreau, Marie-Louise (ed.), *Sociolinguistique : les concepts de base*, Sprimont, Mardaga, 290-291.

Canut, Cécile, 2002. « Introduction », in : Canut, C. et Caubet D. (ed.), *Comment les langues se mélangent. Codeswitching en francophonie*, Paria, L'Harmattan, 9-19.

Crystal, David, 1985. *A dictionary of linguistics and phonetics*. 2nd edition. New York, Basil Blackwell.

Ferguson, Charles A. 1959. «Diglossia», *Word* 15, 325-340.

García, Ofelia, 2009. *Bilingual Education in the 21st Century: A Global Perspective*, Malden, MA and Oxford, Basil/Blackwell.

Goebl, Hans et al., 1996. *Kontaktlinguistik : ein internationales Handbuch zeitgenössischer Forschung = Contact Linguistics : an International Handbook of Contemporary Research = Linguistique de contact : Manuel international des recherches*, Berlin/New York, de Gruyter.

Gumperz, John, 1989. *Engager la conversation. Introduction à la sociolinguistique interactionnelle*, Paris, Minuit.

Heller, Monica, 1988. *Codeswitching: Anthropological and Sociolinguistic Perspectives*, Berlin, Mouton de Gruyter.

Holtus, Günter/Metzeltin, Michael/Schmitt, Christian, 1998-2005. *Lexikon der Romanistischen Linguistik (LRL)*, Tübingen, Niemeyer.

Labov, William, 1972. *Sociolinguistic Patterns*, Philadelphia, University of Pennsylvania Press.

Milroy, Lesley, and Muysken, Pieter, 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching.* Cambridge, Cambridge University Press.

Moreau, Marie-Louise, 1997. *Sociolinguistique: les concepts de base*, Sprimont, Mardaga.

Neveu, Franck, 2009. *Dictionnaire des sciences du langage*, Paris, Armand Colin.

Simonin, Jackey / Wharton,Sylvie, 2013. *Sociolinguistique du contact: Dictionnaire des termes et concepts*, Lyon, Ecole normale supérieure de Lyon.

Swann Joan et al., 2004. *A dictionary of sociolinguistics*, Edinburgh, Edinburgh University Press.

Trudgill, Peter, 2003. *A Glossary of Sociolinguistics*, Edinburgh, Edinburgh University Press.

Ungureanu, Cristina, 2013. « Élaborer la terminologie de la sociolinguistique en roumain. Études de cas », *Vox Romanica* 72, Éditions Francke Verlag Tübingen, 56-74

Valdman, Albert, 1982. « Français standard et français populaires : sociolectes ou fictions », *French Review* 56(2), 218-27.

Valdman, Albert, 2000. « La langue des faubourgs et des banlieues : de l'argot au français populaire », *French Review* 73(6), 1179-92.

Vézina, Robert et al., 2009. *La rédaction de définitions terminologiques*, Montréal, Office québécois de la langue française.

Winford, Donald, 2003. *An Introduction to Contact Linguistics*, Oxford, Blackwell.

Sites internet

Centre National de Ressources Textuelles et Lexicales. ‹www.cnrtl.fr ›

DEX (*Dicţionarul explicativ al limbii române*) ‹dexonline.ro›

Dictionnaire de français, Larousse. ‹www.larousse.fr/dictionnaires/francais/ ›

DOOM2 (*Dicţionarul ortografic, ortoepic şi morfologic al limbii române*). ‹http://dexonline.ro›

Gaffiot, Félix 1934. Dictionnaire latin-français. Paris, Hachette. ‹ www.lexilogos.com/ latin/gaffiot.php›

TLFi (Trésor de la langue française informatisé). ‹atilf.atilf.fr›

Il VoLIP una risorsa per lo studio della variazione nel parlato della lingua italiana

1. Introduzione

I *corpora* sono una grande risorsa per le indagini linguistiche poiché permettono di costruire e/o validare le proprie spiegazioni su materiale autentico, strutturato, pubblicamente fruibile. Un *corpus* non è infatti una mera collezione di dati, ma un insieme strutturato sulla base di criteri espliciti e controllabili. I lavori basati su *corpora* consentono quindi di condividere il processo di costruzione delle ipotesi teoriche formulate su dati condivisi e non creati *ad hoc*. È questa una condizione necessaria per la realizzazione di protocolli di ricerca comuni, riproducibili e riutilizzabili da più ricercatori.

I *corpora* sono, d'altro canto, essi stessi il frutto di ipotesi linguistiche poiché sono disegnati sulla base di criteri di rappresentazione in scala di una lingua o di una porzione di essa: ogni *corpus* sottintende un modello di lingua e di aggregazione dei dati. Ogni *corpus* implica, infatti, la scelta delle variabili che si ritengono più significative nel condizionare gli usi linguistici e il peso che si assegna a ciascuna di esse; queste scelte producono *corpora* non solo di dimensioni diverse, ma di contenuti molto diversi. Questa scelta può variare in base alle convinzioni teoriche dei compilatori del *corpus* e/o in base ai livelli linguistici che si vogliono indagare o, ancora, in base agli scopi per i quali il *corpus* è concepito. La grande espansione della linguistica dei *corpora* negli ultimi vent'anni, dovuta soprattutto allo sviluppo straordinario degli strumenti elettronici di raccolta, annotazione e elaborazione del materiale linguistico, ha permesso non solo la crescita delle dimensioni dei *corpora*, ma la moltiplicazione dei tipi di *corpora*. Accanto ai *corpora* che potremmo chiamare di riferimento o generali, si possono consultare *corpora* concepiti per indagare specifici ambiti o per usi determinati, quali per esempio quelli didattici (O'Keeffe, McCarthy 2010).

Oltre alle scelte relative alle dimensioni di variazione e al peso ad esse assegnato e quindi ai tipi di testi da inserire, i *corpora* differiscono molto per ciò che riguarda le scelte di trattamento e annotazione, il tipo e il grado di fruizione pubblica del materiale (Baroni 2010). E' opportuno operare una distinzione tra *corpora* di testi scritti e di testi parlati. Benché anche i testi scritti siano sottoposti a trattamenti di varia natura per renderli elaborabili elettronicamente e poi fruibili in modo uniforme dagli utenti, i testi parlati richiedono un passaggio ulteriore: la trascrizione ortografica più

o meno larga. Si tratta di un passaggio non necessario né concettualmente né teoricamente, ma che metodologicamente aumenta la facilità di elaborazione del parlato, rendendolo permanente e non volatile. E' questo un lavoro molto oneroso in termini di tempo e tutt'altro che neutro, che ancora una volta implica scelte teoriche e metodologiche non banali. E' utile ribadire che la trascrizione di un testo parlato deve trasferire attraverso il canale grafico-visivo, normalmente utilizzato per comunicazioni unidirezionali non sincrone, testi perlopiù dialogici sincroni e quindi trovare soluzioni univoche per registrare fenomeni di dinamica comunicativa (presa dei turni, sovrapposizioni ecc.), non segmentali (pause, velocità di eloquio, andamento prosodico, variazioni vocali significanti, come il *self-talk*), disfluenze (autocorrezioni, cambiamenti di progetto ecc.), elementi verbali non lessicali (segnalazioni di assenso, interiezioni varie, forme contratte ecc.), elementi non verbali (elementi vocali, per esempio risate, colpi di tosse, inspirazioni e simili, ma anche espressioni facciali, movimenti del corpo ecc.) e contestuali ritenuti rilevanti (Edwards 1993, Leech *et al*. 1995). I sistemi di trascrizione oggi a disposizione sono molteplici e in alcuni casi molto ricchi: possono includere per esempio molti livelli di codifica, fino a comprendere anche quello cinesico. Purtroppo più una trascrizione è dettagliata e 'completa' meno è leggibile: nuovamente si è di fronte ad una scelta che dipende dal tipo di analisi a cui si è interessati e per la quale il *corpus* è stato concepito.

I *corpora* di scritto e di parlato presentano di fatto delle differenze anche per ciò che riguarda l'accesso al materiale. Mentre i primi di solito rendono disponibile il materiale raccolto e, se interrogabili, si può di norma risalire dall'interrogazione al cotesto e quindi all'intero testo originario, molto raramente questo è possibile con i *corpora* di parlato. Benché esistano *corpora* etichettati a vari livelli, da quello fonetico a quello pragmatico (Lüdeling, Kytö 2008), non tutti sono interrogabili e sono pochi i *corpora* di parlato che danno accesso al materiale audio e ancora meno quelli che permettono qualche forma di interrogazione diretta o indiretta, attraverso la trascrizione ortografica, del materiale audio.

Un *corpus* è quindi prima di tutto il risultato di un insieme di scelte complesse e integrate, fortemente condizionate anche da fattori metodologici e di organizzazione della ricerca. Per questo motivo, così come normalmente abbiamo bisogno di diverse condizioni sperimentali sui cui testare le nostre ipotesi, è utile poter disporre di più *corpora*. Ciò è tanto più vero quando si tratta di parlato, almeno per due ragioni. Innanzi tutto i *corpora* di parlato sono ancora molto pochi se confrontati con quelli di testi scritti. Ciò vale per tutte le lingue, ma ancor di più per le lingue romanze. Anche nella linguistica del *corpus*, la riflessione teorica sulla specificità del parlato è relativamente recente e ancora più recente è il riconoscimento di una rilevanza della variabile diamesica in sé (Baker 2010). Ciò rende necessario uno studio della comunicazione parlata in diverse condizioni enunciative e, probabilmente, anche in diversi contesti settoriali per aver ben chiara l'integrazione tra tutte le componenti del processo di significazione.

In secondo luogo, più *corpora* permettono di prendere in considerazione un maggior numero di contesti diafasici e diatopici. La questione è di grande rilevanza soprattutto per aree linguistiche come quella italiana in cui il parlato non solo è associato al registro informale e colloquiale, ma anche a varietà geografiche locali. È ben nota la storia linguistica italiana caratterizzata fino al secondo dopoguerra da una forte diglossia, con uso del dialetto nelle situazioni informali parlate e uso dell'italiano nello scritto formale. Sebbene dagli inizi degli anni Ottanta ad oggi sia cresciuta la consapevolezza che il canale ha un ruolo suo proprio nel determinare l'uso delle lingue storico-naturali e quindi le caratteristiche dei testi, l'associazione prevalente/preponderante di un canale con certe situazioni e/o varietà diatopiche rende difficile individuare le proprietà attribuibili alle diverse dimensioni di variazione. Ciò ingenera un'indebita equivalenza tra proprietà del canale e registri e varietà regionali o dialetti, tra variazione diamesica, diafasica, diatopica (Voghera 2010a). I tre tipi di variabili vanno evidentemente tenuti separati e un *corpus* di parlato deve prevedere una variazione lungo l'asse sia diafasico sia diatopico.

Partendo da queste considerazioni, da quasi un decennio abbiamo iniziato un percorso di costruzione di una base di dati varia e condivisa del parlato italiano, creando nuove risorse e valorizzando e sviluppando quelle già esistenti, grazie alla creazione di un osservatorio permanente del parlato italiano, il portale Parlaritaliano (‹www.parlaritaliano.it›, Voghera 2010b). Si tratta di un osservatorio permanente del parlato che ha due obiettivi fortemente interrelati. Il primo consiste nell'allargare la base conoscitiva dei principali meccanismi enunciativi e grammaticali della comunicazione parlata attraverso studi basati su *corpora*. Il secondo obiettivo consiste nel valutare come e quanto l'allargamento e l'approfondimento della base documentaria possano contribuire a una migliore comprensione del sistema linguistico nel suo complesso. È evidente infatti che il parlato, ad un primo livello, si caratterizza come un sottoinsieme di strutture linguistiche tipiche non, o solo parzialmente, osservabili in altri contesti; ad un secondo livello, il parlato ci permette di scoprire relazioni tra porzioni diverse della grammatica, altrimenti nascoste, ma comunque centrali nell'architettura generale del sistema.

Agli obiettivi più strettamente linguistici si aggiungono obiettivi informatici rivolti allo studio della struttura dei metadati e dei *database* linguistici, e all'analisi e all'implementazione, anche in ambienti di calcolo ad alte prestazioni, di tecniche di segmentazione automatica per segnali audio-video, con specifiche finalità di supporto alla gestione di *corpora* linguistici multimediali e multimodali.

In questa cornice è nato il VoLIP (Voce del LIP), finanziato dal Ministero italiano dell'Istruzione, dell'Università e della Ricerca, una nuova risorsa linguistica che permette l'ascolto e l'interrogazione dei file audio del *corpus* LIP (De Mauro *et al.* 1993), secondo criteri sociolinguistici, lessicali e morfosintattici. Il VoLIP è stato creato in una prospettiva ecologica delle risorse linguistiche che mira al pieno utilizzo e potenziamento, laddove è possibile, di risorse già disponibili, ma sottoutilizzate per vari motivi. Il VoLIP, mettendo a disposizione i file audio del LIP e rendendoli

interrogabili, ha potenziato l'utilizzazione della risorsa offrendo alla comunità scientifica nuove funzionalità di fruizione non esistenti precedentemente e la possibilità di accesso libero e diretto all'intero *corpus* sul portale Parlaritaliano.it.

2. Il passaggio dal LIP al VOLIP

Come abbiamo già detto, la costruzione del VoLIP ha comportato un processo di valorizzazione di una risorsa già esistente, ma solo parzialmente utilizzata. Ciò ha implicato varie fasi nel trattamento del materiale originario.

2.1. Trattamento dei testi

Il *corpus* LIP è costituito da circa 60 ore di registrazione per un totale di circa 500.000 occorrenze, ricavate da testi raccolti in diverse città italiane e disposti su una scala di progressiva dialogicità e spontaneità. Nella Tabella 1 si dà la distribuzione delle occorrenze in rapporto alle variabili diatopica e diafasica.

Città	Conversazioni faccia a faccia	Conversazioni telefoniche	Interviste Dibattiti	Monologhi	Radio/ TV	Totale
Milano	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Firenze	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Roma	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Napoli	~25.000	~25.000	~25.000	~25.000	~25.000	~125.000
Totale	~100.000	~100.000	~100.000	~100.000	~100.000	~500.000

Tabella 1: composizione del *corpus* LIP

Le trascrizioni ortografiche originarie evitavano qualsiasi intervento di normalizzazione. Nel passaggio dal LIP al VoLIP i materiali audio sono stati digitalizzati in file wav (Windows PCM, 22050Hz, 16 bit) e il riascolto ha inevitabilmente portato a revisioni nelle trascrizioni. Si è quindi deciso di rendere disponibili sia la vecchia sia la nuova trascrizione: ogni porzione revisionata richiama una finestra di *pop-up* con la trascrizione emendata.

2.2. Metadatazione

La scelta di procedure standardizzate, adeguate e sufficientemente accurate di metadatazione, incide notevolmente sulla gestione, fruibilità e diffusione dei *corpora*.

Al fine di arricchire la catalogazione già prevista nel *corpus* LIP, per generi discorsivi e provenienza geografica dei parlanti, ma soprattutto con lo scopo di adottare un formato standardizzato, nel passaggio al VoLIP si è deciso di adottare la metadatazione in formato IMDI (*ISLE Metadata Initiative*[1]).

Tale formato, proposto per descrivere risorse linguistiche multimediali e multimodali, è stato scelto anzitutto per favorire il confronto con altri *corpora* (data la sua diffusione in vari progetti di ricerca su *corpora* di diverse lingue), ma anche per la garanzia della correttezza formale dei metadati grazie ai *tool* di supporto (Broeder *et al*. 2001).

Attualmente è possibile consultare ed interrogare in rete il VoLIP sia in base alla catalogazione adottata nel *corpus* LIP, sia in base ai metadati IMDI. La figura 1 mostra la maschera di interrogazione dei metadati in VoLIP.

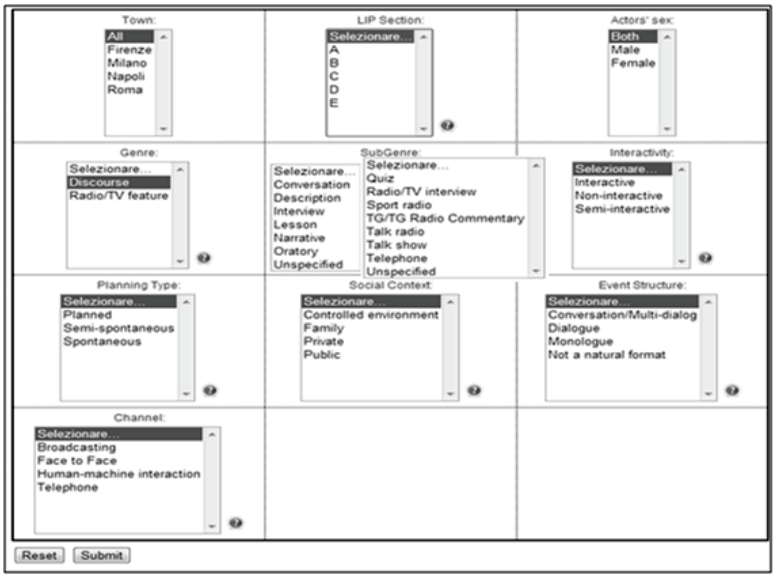


Figura 1: maschera di interrogazione per metadati.

Ciascun menù a tendina corrisponde a un campo interrogabile che consente di filtrare e selezionare tutti i testi che presentano le caratteristiche richieste. È quindi possibile effettuare diversi tipi di interrogazioni, il cui risultato consiste nelle trascrizioni ortografiche associate ai file audio.

La casella *Town* permette di selezionare la provenienza geografica dei parlanti (delle quattro città Firenze, Milano, Napoli, Roma), mentre il campo *Lip section* fa

[1] Per le specifiche adottate, rimandiamo alla documentazione disponibile all'indirizzo: ‹www.mpi.nl/imdi/›.

riferimento ai generi discorsivi della catalogazione adottata nel LIP, riportati nella tabella 1.

Le altre voci in figura corrispondono, invece, a campi dei metadati IMDI. La casella *Actors'sex* consente di filtrare voci maschili o femminili. A *Genre* corrisponde la prima macrocatalogazione tra parlato radiotelevisivo (*Radio/TV feature*) e tutti gli altri tipi (*Discourse*). Per ciascuno dei due, è poi possibile mediante il campo *SubGenre* selezionare un sottotipo specifico: in *Radio/TV feature*, si possono richiedere, ad esempio, interviste radiofoniche o notiziari televisivi; in *Discourse*, è possibile effettuare una ricerca che individui, ad esempio, unicamente lezioni scolastiche o universitarie. Gli altri campi sono relativi al grado di interattività (*Interactivity*) e di pianificazione dei testi (*Planning type*), al contesto in cui avviene lo scambio comunicativo (*Social Context*), alla struttura dell'evento rispetto al numero di partecipanti (*Event Structure*) e al tipo di canale di trasmissione (*Channel*).

La possibilità di incrociare diverse selezioni consente di effettuare ricerche particolareggiate relative a più aspetti contemporaneamente.

3. L'allineamento della voce con le trascrizioni

L'obiettivo principale del VoLIP è quello di dare voce al LIP, cioè rendere disponibile e interrogabile il *corpus* di parlato nella sua interezza[2]. Lo scopo del VoLIP è quindi non solo quello di fornire la versione audio dei testi del LIP, ma di poter fare ricerche che permettano di trovare e ascoltare singole parole in contesto. Abbiamo quindi creato un sistema *online* che permette di ascoltare una data forma all'interno del suo contesto e contemporaneamente leggere la relativa trascrizione ortografica senza che sia necessario scaricarla sul proprio computer.

Ciò si è ottenuto con l'uso di uno strumento di allineamento forzato che ha elaborato le trascrizioni ortografiche e i file audio del *corpus*. La procedura di *Forced Alignment* (FA) fra testo e audio si basa su un approccio semplificato di quanto accade normalmente nel caso del processo di *Automatic Speech Recognition* (ASR). Il caso FA è semplificato dal fatto che il testo è noto, dunque la procedura deve solo ottimizzare la suddivisione del segnale in piccole parti e la loro assegnazione ai singoli eventi fonici che il sistema è in grado di riconoscere contemporaneamente nella loro forma simbolica (testo) e nella loro forma acustica (audio). Il testo da allineare viene quindi suddiviso (tokenizzato) in parole, e per ogni parola un trascrittore automatico grafema > fonema descrive la composizione in eventi simbolici unitari. Il sistema di ASR analizza il segnale audio ogni 20 millisecondi e determina quante porzioni consecutive di segnale possono essere associate ad ogni evento simbolico nella catena. Successive fasi di ottimizzazione completano il processo.

---

[2] Ricordiamo che le trascrizioni originarie dei testi erano state già pubblicate insieme al lessico di frequenza (De Mauro *et al.* 1993) e che sono interrogabili sul sito dell'Università di Graz ‹badip.uni-graz.at/›.

Poiché il materiale originale era di qualità molto varia e raccolto in condizioni audio non ottimali, spesso molto disturbate (mercati, assemblee, classi scolastiche, ecc.), è stato necessario integrare il lavoro svolto tramite il programma di allineamento con un lavoro di divisione manuale a partire dalla trascrizione ortografica. Tale soluzione è consistita nell'individuare porzioni della durata media di 30 secondi. Il lavoro si è svolto in due fasi: nella prima, le trascrizione ortografiche sono state suddivise in porzioni di senso compiuto cercando di rispettare, per quanto possibile, l'intervallo temporale fissato; nella seconda fase si sono ascoltati i file audio e tramite un software appositamente creato, basato sulla libreria SoundManager 2 (‹www.schillmania.com/projects/soundmanager2/›), si sono salvati il tempo di inizio e fine di ogni intervallo precedentemente segnalato sulla trascrizione. Il risultato di queste due fasi è un file di tempi ed un file contenente la trascrizione ortografica divisa in parti. A questo punto, per verificare la correttezza del lavoro manuale e allineare ogni coppia porzione-di-trascrizione/tempi in un unico file che fosse nello stesso formato finale dei file ottenuti in precedenza grazie al software di FA, è stato sviluppato un altro script che fosse anche in grado di rilevare e segnalare differenze tra le porzioni della trascrizione e il numero di intervalli salvati.

Concluse le fasi appena descritte, si è giunti alla fase finale del lavoro. In particolare, è stato sviluppato un innovativo software che guida l'utente nella scelta dei testi. Una volta cercata una forma all'interno del *corpus*, il software permette di visualizzare tutti i testi in cui occorre, di evidenziarne le occorrenze nella trascrizione dei singoli file e di ascoltarle in contesto online. Questa procedura rende molto veloce l'accesso al *corpus* perché non è necessario scaricare alcuno strumento o file. Il programma consente anche di scaricare i contesti audio oppure l'intero file audio corredato di trascrizione per una futura elaborazione[3].

4. Le interrogazioni nel VoLIP

Il VoLIP prevede diversi tipi di interrogazione a partire dalla ricerca per lemmi o per forme, permettendo di ottenere sia le liste di frequenza delle occorrenze sia il contesto dell'elemento lessicale ricercato in trascrizione ortografica e sonora.

4.1. Frequenza di lemmi e forme

Nella ricerca per lemmi si può selezionare una specifica parte del discorso, tra quelle su cui è basata la lemmatizzazione del LIP (*Adjective, Adverb, Article*, *Company name*, *Conjunction*, *Geographic name*, *Interjection*, *Name*, *Noun*, *Onomato-*

[3] Il menù di selezione che appare su ogni occorrenza individuata e che consente la scelta di scaricare il file o una sua porzione, e di ascolto della sezione audio relativa alla forma cercata è basata su jQuery (‹jquery.com/›), libreria creata da John Resig, ed il plugin Audero Context Menu (‹github.com/AurelioDeRosa/Audero-Context-Menu›) creato da Aurelio De Rosa. Infine, l'estrazione vera e propria di una certa porzione di audio di un file WAV è resa possibile grazie alla libreria Audero Wav Extractor (‹github.com/AurelioDeRosa/Audero-Wav-Extractor›) creata da Aurelio De Rosa.

poeia, *Preposition*, *Pronoun, Surname, Verb*), o selezionarle tutte. Sia l'interrogazione per lemmi sia quella per forme offre la possibilità di raffinare la ricerca selezionando uno dei cinque genere discorsivi distinti nel LIP o il loro insieme (si veda Tabella 1).

La maschera di interrogazione è riprodotta in figura 2.

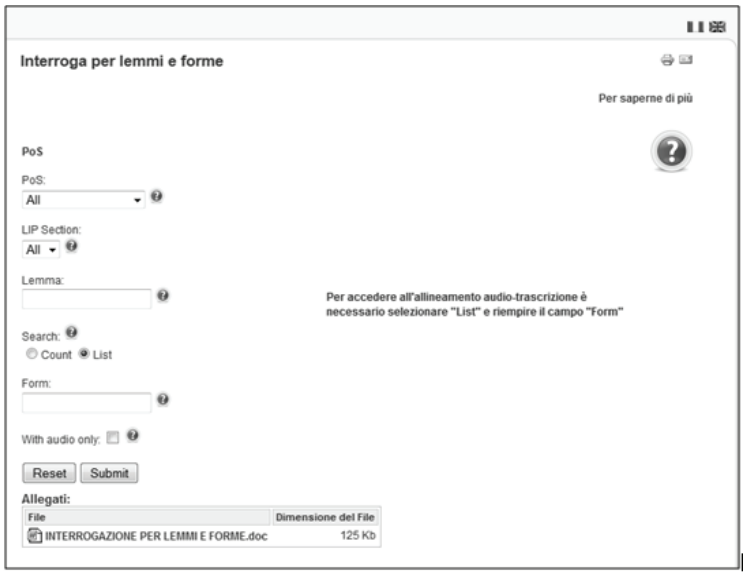


Figura 2: maschera di interrogazione per lemmi e forme.

La casella "PoS" permette di restringere le parti del discorso da ricercare. Una volta digitata nella casella "Lemma" la parola che si intende cercare, la scelta fra le due opzioni "Count" e "List" permette di orientare la ricerca in due direzioni.

L'opzione "Count" consente di ottenere sia il numero totale di occorrenze del lemma sia la loro distribuzione nei cinque generi discorsivi: il conteggio è quello originario del LIP.

Come si può notare dall'esempio (1), se si cerca il lemma correre, viene, viene anche fornita l'indicazione della parte (o delle parti del discorso) a cui può essere ricondotto il lemma cercato.

(1)

Pos	A	B	C	D	E	Total
Verb	45	71	25	22	73	236
Noun	0	0	0	0	2	2

La scelta dell'opzione "List" permette invece di ottenere l'elenco delle forme e la loro distribuzione suddivisa nei cinque generi di discorso (come mostrato nell'esempio numero 2). Sono riportate tra @ le forme locali e/o dialettali.

(2)

Pos	Lemma	Form	A	B	C	D	E	Total
Verb	CORRERE	corre	2	0	1	9	1	13
Verb	CORRERE	corri	1	2	1	0	0	4
Verb	CORRERE	correre	0	0	1	1	1	3
Verb	CORRERE	correndo	0	1	0	1	0	2
Verb	CORRERE	corrono	1	0	0	1	0	2
Verb	CORRERE	correra'	0	1	0	1	0	2
Verb	CORRERE	@curruto@	0	0	1	0	0	1
Verb	CORRERE	correrebbero	0	0	0	1	0	1
Noun	CORRERE	correre	0	0	0	1	0	1
Verb	CORRERE	@core@	0	1	0	0	0	1

4.2. L'ascolto di forme o sequenze

Le interrogazioni possono avere come esito l'ascolto di singole forme o sequenze. L'accesso al contesto dei file audio e di trascrizione avviene attraverso la compilazione del campo "Form", nel quale si può richiedere una delle forme presenti nel *corpus*, cfr. Figura 3

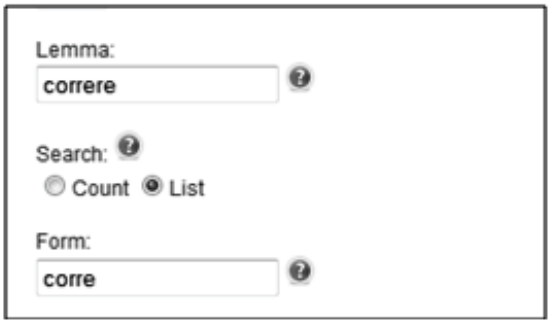


Figura 3: esempio di ricerca di una forma.

Attraverso il campo “Form” si possono ricercare anche sequenze di parole, ad es. “ho dormito”; “i cani”, “e’ una citta’ che”.

La ricerca di una forma permette di visualizzare l’indicazione del numero di occorrenze totali e ripartite nei cinque generi di discorso del LIP, di accedere all’intero testo grafico e audio contenente la forma, e, soprattutto, di accedere direttamente alla porzione di testo e di audio contente la forma ricercata. E’ possibile ricercare anche una sequenza; in questo caso il risultato non fornisce il numero di occorrenze, ma i file audio in cui occorre la sequenza.

Forniamo qui di seguito un esempio di ricerca del segnale audio dei contesti in cui occorre la forme “corre”. La ricerca, illustrata nell’esempio numero (3) dà come risultato il numero delle occorrenze totali della forma e la loro ripartizione nei cinque generi e l’elenco dei file (audio e trascrizioni) in cui l’elemento compare (si veda la figura 4)

(3)

Pos	Lemma	Form	A	B	C	D	E	Total
Verb	Correre	corre	2	0	1	9	1	13

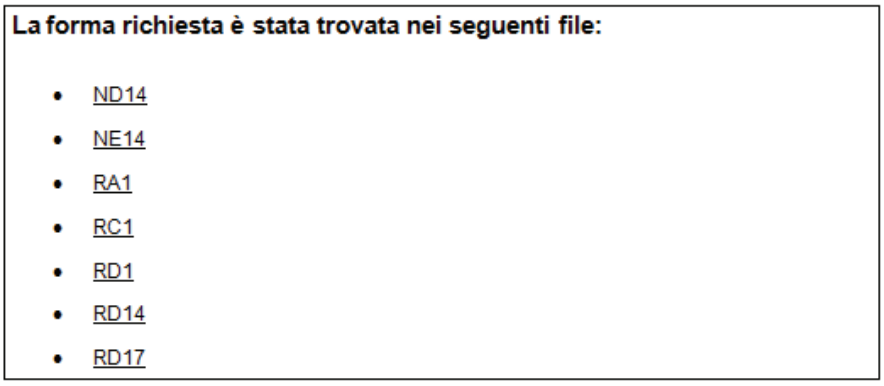


Figura 4: risultato della ricerca della forma “corre” (elenco dei file in cui occorre).

E’ possibile scaricare ciascun file sia in forma di testo che di audio, ma soprattutto individuare velocemente le varie occorrenze della forma (come mostrato in figura 5).

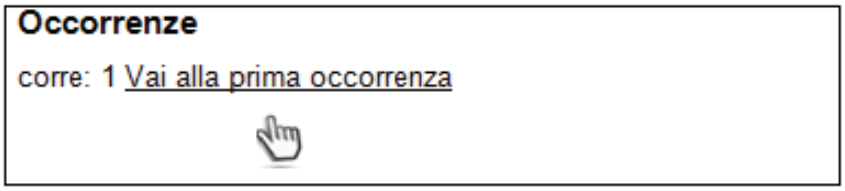


Figura 5: individuazione delle varie occorrenze della forma cercata.

La forma ortografica viene visualizzata in giallo e cliccando sulla forma è possibile ascoltare il contesto in cui appare, grazie all'allineamento tra trascrizione e audio, come mostrato nella figura 6. Oltre ad ascoltare il contesto, è possibile scaricare il frammento di file audio, l'intero file o procedere nella ricerca di altre occorrenze della forma all'interno del file.

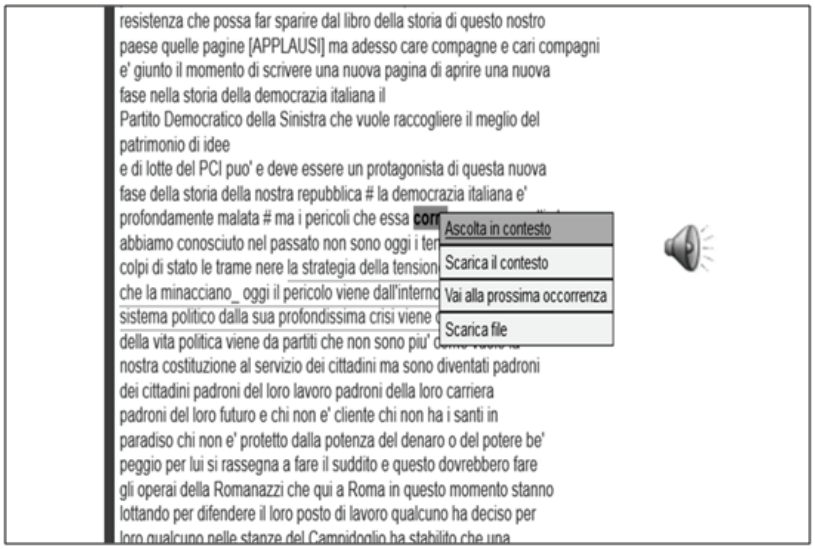


Figura 6: contesto di occorrenza della forma cercata con allineamento tra trascrizione ed audio.

E' ovviamente possibile raffinare la ricerca attraverso l'incrocio dei parametri dei criteri qui brevemente illustrati, in modo, ad esempio, da ottenere informazioni su una data forma, distinguendo le occorrenze in relazione alle parti del discorso a cui possono essere ricondotte. Un tipico esempio è quello della risoluzione dell'omografia tra la prima persona singolare del presente indicativo del verbo "amare" ed il singolare del nome "amo" (si veda l'esempio numero 4).

(4)

Pos	Lemma	Form	A	B	C	D	E	Total
Verb	AMARE	amo	0	3	0	0	0	3
Noun	AMO	amo	0	0	0	0	1	1

Una volta individuata la forma appartenente alla parte del discorso voluta, è possibile osservare e ascoltare i contesti di occorrenza (come illustrato negli esempi 5 e 6).

(5) Nome amo

io lo vado a fare cosi' con un amo in maniera tale praticamente da farvi capire che non e' difficile corrente non ce n'e'

(6) Verbo amo

A: ah senti una cosa ma adesso cosa fai fino all'una stai li' a
B: e' quello che non amo affatto 'st' idea di star qua in
cucina da sola o anche in salotto ad aspettare

5. Ipotesi applicative

Nonostante la grande espansione della linguistica del *corpus* negli ultimi decenni, il numero di *corpora* di parlato è ancora molto limitato. Basti pensare che sia nel *British National Corpus* (‹www.natcop.ox.ac.uk›) sia nel *Corpus de Referencia del Español Actual* (‹www.rae.es›), due tra le più importanti imprese per la costruzione di *corpora* nazionali di riferimento, il rapporto tra testi parlati e scritti è di 1: 10. Ciò dà la misura di quanto il parlato sia poco rappresentato anche quando si parla di lingue che hanno una dimensione di diffusione intercontinentale.

La situazione non è diversa, naturalmente, per l'italiano. Mentre negli ultimi anni si sono sviluppate numerose iniziative di raccolta di *corpora* scritti, i *corpora* di parlato disponibili sono ancora molto poco numerosi (Cresti, Moneglia a cura di 2005; Baroni 2010). Ancora di meno sono quelli che consentono il libero accesso ai materiali audio *online* (Savy, Cutugno 2009). Anche per questo motivo il VoLIP rappresenta una risorsa importante che integra varie funzionalità di ricerca sia nei metadati sia all'interno del *corpus*.

La possibilità di incrociare i criteri della metadatazione IMDI con l'originaria suddivisione in tipi di discorso parlato fatta nel LIP consente di selezionare parlati diversi dal punto di vista diafasico e diatopico e oltre ad avere un'ovvia rilevanza per le ricerche di tipo sociolinguistico, può consentire anche un'utilizzazione didattica del VoLIP.

La ricerca sociolinguistica per forme, lemmi e contesti comparabili si arricchisce del contemporaneo accesso al segnale audio: sebbene il VoLIP non sia stato con-

cepito sul piano tecnico-metodologico per ricerche di tipo fonetico, la sua struttura stratificata ne consente l'utilizzo come *corpus* di controllo o di 'verifica' per analisi socio-fonetiche condotte su altri materiali (parlato di laboratorio, *corpora* speciali, parlato elicitato semi-spontaneo o letto) rispetto ai quali il VoLIP presenta l'indubbio vantaggio di un parlato decisamente spontaneo sotto tutti i punti di vista.

L'allineamento di trascrizione ortografica e audio permette per esempio un lavoro di analisi e decostruzione di ciò che rappresenta uno dei punti più significativi e complicati nell'apprendimento di una lingua da parte di non-nativi: la relazione inversa tra complessità dei contenuti e delle costruzioni morfosintattiche e quella della forma fonetica, in conseguenza del grado di pianificazione e di controllo dell'enunciazione. Infatti, i testi usati nei primi livelli di insegnamento, che coincidono di solito con testi poco pianificati e informali, sono in realtà i più complessi sul piano fonetico perché ricchi di riduzioni foniche e fenomeni di coarticolazione. Al contrario, i testi pianificati, tendenzialmente più articolati sintatticamente e nell'estensione lessicale e semantica, tendono ad essere iperarticolati sul piano fonetico e quindi di più facile comprensione perché si avvicinano maggiormente alla forma fonetica prototipica di fonemi e giunture. Ciò, come si può facilmente sperimentare, fa sì che, dal punto di vista della forma fonica, sia più comprensibile una lezione universitaria piuttosto che una conversazione tra amici. Benché la didattica delle lingue, soprattutto la didattica delle lingue seconde, abbia cercato di sviluppare le abilità di ascolto e parlato, questo genere di questioni non è affrontato in modo sistematico, ma ci si affida piuttosto alla naturalità del parlato e dell'ascolto che è, implicitamente, ritenuta la base sufficiente per garantire il raggiungimento degli obiettivi.

Al contrario, l'insegnamento e l'apprendimento si gioverebbero molto di esercizi che uniscano l'ascolto con attività di osservazione e analisi. Il VoLIP poiché consente interrogazioni mirate che isolano singole forme e piccoli contesti offre spunti utili per la creazione di materiale didattico *on* e *off line*, contribuendo alla costruzione di sillabi più strutturati per la didattica dell'ascolto e del parlato.

Università degli Studi di Salerno — Miriam VOGHERA
Università degli Studi di Salerno — Claudio IACOBINI
Università "Federico II" di Napoli — Francesco CUTUGNO
Università degli Studi di Salerno — Renata SAVY
Università degli Studi di Salerno — Iolanda ALFANO
Università degli Studi di Salerno — Aurelio DE ROSA

Riferimenti bibliografici

Baker, Paul, 2010. *Sociolinguistics and Corpus Linguistics*, Edimburgh, Edimburgh University Press.

Baroni, Marco, 2010. *Corpora di italiano, voce della Enciclopedia dell'Italiano diretta da R. Simone*, Istituto dell'Enciclopedia Italiana G. Treccani.

Broeder, Daan / Offenga, Freddy / Willems, Don / Wittenburg, Peter, 2001. «The IMDI Metadata set, its Tools and accessible Linguistic Databases», *Proceedings of the IRCS Workshop on Linguistic Databases*.

De Mauro, Tullio / Mancini, Federico / Vedovelli, Massimo / Voghera, Miriam, 1993. *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri.

Cresti, Emanuela/Moneglia, Massimo (ed.), 2005. *C-Oral-Rom. Integrated reference corpora for spoken romance languages*, Amsterdam /Philadelphia, Benjamins.

Edwards, Jane A., 1993. «Principles and Contrasting Systems of Discourse Transcription», in: Edwards, Jane A. / Lampert, M. D. (ed.), *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale, N. J., Lawrence Erlbaum Associates, 3-32.

Leech, Geoffrey / Meyers, Greg / Thomas, Jenny (ed.), 1995. S*poken English on Computer. Transcription, Mark-up and Applications*, New York, Longman Publishing, 82-98.

Lüdeling, Anke / Kytö, Merja (ed.), 2008. *Corpus Linguistics. An International Handbook*, Berlin/New York, De Gruyter.

O'Keeffe, Anne / McCarthy, Michael (ed.), 2010. *The Routledge Handbook of Corpus Linguistics*, London / NY, Taylor & Francis Ltd Routledge.

Voghera, Miriam, 2010a. *Lingua parlata, voce della Enciclopedia dell'Italiano diretta da R. Simone*, Istituto dell'Enciclopedia Italiana G. Treccani.

Voghera, Miriam, 2010b. «Parlare italiano: towards a multidimensional description and a multidisciplinary explanation», in: Pettorino, Massimo / Giannini, Antonella / Dovetto, Francesca M. (ed.), *La comunicazione parlata 3*, Napoli, Università degli Studi di Napoli L'Orientale, 603-617.

Savy, Renata / Cutugno, Francesco, 2009. CLIPS. «Diatopic, diamesic and diaphasic variations in spoken Italian», in: Mahlberg, Michaela / González-Díaz, Victorina / Smith, Catherine, *Proceedings of Vth Corpus Linguistic Conference (CL2009)*, 1-24 Liverpool, University of Liverpool.