eDTLR – base de données et instrument pour la recherche lexicographique roumaine

Elena DĂNILĂ

Key-words: *electronic dictionary*, *linguistic resources*, *corpus*, *computerized lexicography*

La société actuelle est caractérisée par la possibilité de communiquer des informations sans avoir des barrières spatiales ou temporelles. Dans ces circonstances, la seule limitation, dépassée partiellement à présent, reste celle linguistique. C'est pourquoi un objectif actuel des politiques européennes est la préservation et la valorisation de l'identité linguistique nationale, vu le fait qu'il y a une tendance générale d'utiliser certaines langues privilégiées par le fait qu'elles possèdent des moyens informatisés de promouvoir. Dans ces conditions, en Roumanie on a commencé à créer des instruments et des ressources électroniques nécessaires pour soutenir la langue et la culture roumaines, au niveau transnational, dans le contexte plus large de la recherche académique fondamentale.

En Roumanie, l'initiative des recherches dans le domaine du traitement automatique du langage naturel appartient aux informaticiens de Bucarest (ICIA[1]) et Iaşi (UAIC – FII[2], IIT[3]). Le fait que les études dans ce domaine, ayant pour objet la langue roumaine, nécessitaient la collaboration avec les linguistes a fait démarrer un dialogue scientifique entre les informaticiens et les linguistes, dialogue matérialisé en quelques projets de recherche communs sur l'informatisation de la recherche linguistique roumaine, pour valoriser les ressources existantes par leur acquisition en format électronique et par la création de nouvelles ressources et nouveaux instrument pour le traitement automatique de la langue.

Pendant les dernières années, des lexicographes et des informaticiens ont travaillé pour la création d'une version informatisée du *Dictionnaire «Trésor» de la Langue Roumaine*, qui représente le plus complexe ouvrage lexicographique créé sous l'égide de l'Académie Roumaine – commencé il y a 105 années (avec les deux séries : DA, la série ancienne, – parue entre 1907 et 1944 et DLR, la série nouvelle, parue entre 1965 et 2010). La série ancienne (DA) a été coordonnée par Sextil Puşcariu, tandis que la série nouvelle a été élaborée dans les trois centres de recherche académique de Bucarest, Iaşi et Cluj-Napoca, sous la coordination initiale

[1] http://www.racai.ro
[2] http://www.info.uaic.ro, www.consilr.info.uaic.ro
[3] http://www.iit.tuiasi.ro

„Philologica Jassyensia", An VI, Nr. 1 (11), 2010, p. 37–46

de certains linguistes renommés, tels Iorgu Iordan, Alexandru Graur et Ion Coteanu, et, dès 2000, par Marius Sala et Gheorghe Mihăilă.

Ce qui est spécifique du point de vue lexicographique au *Dictionnaire de la langue roumaine*, rapporté à d'autres ouvrages lexicographiques édités sous l'égide de l'Académie Roumaine, est déterminé par son caractère de «trésor», ce qui suppose l'enregistrement et le traitement lexicographique de tous les mots de la langue roumaine enregistré dans des textes écrits, dans au moins deux styles fonctionnels, à partir des sources écrites, citées sous chaque définition.

La première édition du *Dictionnaire-trésor de la langue roumaine* vient d'être achevée en avril 2010.

Avant même de l'achèvement de la première édition du Dictionnaire on a réalisé des projets pour trouver une solution pour l'acquisition en format électronique de la variante imprimée de DLR.

Jusqu'à présent, dans le Département de Lexicologie – Lexicographie de L'Institut de Philologie Roumaine «A. Philippide»[4], ont a finalisé quelques projets :

a) *Dicţionarul limbii române (DLR) în format electronic. Studii privind achiziţionarea* [Le Dictionnaire de la Langue Roumaine (DLR) en format électronique. Études sur l'acquisition], projet financé par le Ministère de l'Education et de la Recherche (MEC), par le Conseil National de la Recherche Scientifique de l'Enseignement Supérieur (CNCSIS) (2003–2005) – par ce projet on a vérifié et on a démontré la possibilité de transformer *Le Dictionnaire de la Langue Roumaine* du texte imprimé en texte électronique annoté[5], traité à l'aide d'un programme spécifique, DLRex – un instrument d'acquisition, de traitement et consultation du DLR, à partir d'une heuristique par laquelle on reconnaît les différents champs formels du texte d'un article, ayant pour résultat la possibilité d'identifier automatiquement le texte des définitions, des citations et celui des sigles.

b) *Resurse lingvistice în format electronic: Monumenta linguae Dacoromanorum. Biblia 1688. Regum I, Regum II – Ediţie critică şi corpus adnotat* [Les Ressources linguistiques en format électronique: Monumenta linguae Dacoromanorum. Biblia 1688. Regum I, Regum II – édition critique et corpus annoté], (projet CNCSIS 2006–2007). Dans ce projet on a trouvé une méthode d'acquisition en format électronique des livres anciens de la Bibliographie DLR[6] (en particulier deux livres de la Bible imprimé à Bucarest en 1688, *Regum I, Regum II*) et la création de certains instruments pour indexer et annoter automatiquement, au niveau du mot, des textes roumains anciens.

c) *DLRI. Bază lexicală informatizată. Derivate* [DLRI. Base lexicale informatisée. Des dérivés] (projet CNCSIS 2007–2008). Par ce projet on a proposé la

[4] http://www.academiaromana-is.ro/philippide/pages/lexicografie.html

[5] Le texte annoté est un texte analysé et marqué du point de vue formel, de sorte qu'il peut être consulté, corrigé, modifié, etc. par les lexicographes, à l'aide de l'ordinateur. On a la possibilité d'extraire du format complet une forme destinée seulement à la consultation, qui s'adresse à un public plus large que celui des experts. Pour des détails, voir Haja, Dănilă *et alii*, 2005.

[6] Parallèlement, à Bucarest s'est déroulé un autre projet destinée à l'acquisition en format électronique des livres de la Bibliographie DLR, *CNR – Corpus de referinţă al limbii române pentru constituirea de dicţionare academice* [CNR – Corpus de référence pour la langue roumaine, pour la réalisation des dictionnaires académiques] (projet CNCSIS, 2007–2008; directeur dr. Monica Busuioc).

réalisation d'un échantillon lexicographique formé des dérivés dans la langue roumaine avec les suffixes *-ime* et *-işte*, de la série ancienne DA et de la série nouvelle DLR, et l'unification du point de vue des normes lexicographiques des articles de DA et de DLR.

Parallèlement on démarré la processus de digitalisation dans le Département de Dialectologie et d'Histoire Littéraire du même Institut.

Fondamentalement, ces démarches de digitalisation de la recherche philologique roumaine visent, d'une part, la réalisation d'*instruments* informatisés spécifiques à la recherche philologique (des programmes de traitement/ analyse automatique du texte écrit/ parlé; des interfaces de travail on-line, nécessaires pour valoriser les ressources crées ou existantes) et, d'autre part, la réalisation de *ressources* linguistiques digitalisées (des dictionnaires informatisés, des *corpus* de textes écrits/ parlés).

Ces démarches ont préparé le terrain pour le projet complexe *eDTLR – Dicţionarul tezaur al limbii romane in format electronic* [eDTLR – Le Dictionnaire trésor de la langue roumaine en format électronique].

Les grandes cultures européennes possèdent, depuis plusieurs années, des dictionnaires-trésor et des corpus de textes en format électroniques. Pour mieux comprendre la dimension du *Dictionnaire de la langue roumaine*, on présente quelques données statistiques de celui-ci comparées à d'autres grands dictionnaires européens:

– *Dicţionarul tezaur al limbii române* [Le Dictionnaire trésor de la langue roumaine] (deux séries: DA – 1907–1944, DLR – 1965–2010), 13 tomes en 37 volumes, totalisant plus de 17.500 pages et plus de 175.000 entrées, y compris les variantes; l'élaboration de la variante électronique: 2007–2010;

– *Oxford English Dictionary* (*OED*, http://www.oed.com/) – la première édition – 1928, 20 volumes (la deuxième édition – 1989), 301.100 entrées, 2.412.400 exemples; la première version électronique: 1988;

– *Deutsches Worterbuch der Grimm* (DWB, http://germazope.uni-trier.de/Projects/DWB: 1838-1961), 32 volumes, 350.000 entrées et variantes; l'élaboration de la variante électronique: 1997–2004;

– *Trésor de la Langue Française* (TLF), sec. XIX-XX (http://atilf.atilf.fr/: 1971-1994 – la première édition imprimée), 16 volumes, 100.000 entrées, 270.000 définitions, 430.000 exemples; l'élaboration de la variante électronique: 1990–2004;

– *Diccionario de la lengua espanola de la Real Academia Espagnola* (DRAE, http://buscon.rae.es/draeI/): 1780 – la première édition imprimée; à présent on travaille à la 23$^{-ème}$ édition; 88.500 entrées, 161.962 exemples; l'élaboration de la variante électronique: 1992.

A partir des informations présentées ci-dessous, on observe le fait que le *Dictionnaire de la langue roumaine* se situe, par la conception et la réalisation, parmi les travaux similaires européens et c'est pour quoi sa digitalisation s'est imposée, en tant qu'une étape normale dans l'évolution de la lexicographie roumaine.

Le projet national *eDTLR. Le Dictionnaire (Trésor) de la Langue Roumaine en format électronique*[7] (CNMP, 2007–2010; directeur: prof. univ. dr. Dan Cristea) représente l'un des plus importants projets de type collaboratif des dernières années, qui entremêle l'expérience des linguistes lexicographes et les recherches linguistiques appliquées, assistées par l'ordinateur.

L'initiateur et, en même temps, le coordonnateur est la Faculté d'Informatique de l'Université « Alexandru Ioan Cuza », Iaşi; directeur dr. Dan Cristea. Vu le fait qu'il s'agit d'un projet qui suppose la corroboration de l'expérience des informaticiens et de celle des linguistes lexicographes, avec un rôle important des recherches informatiques, dans le projet se sont impliqués, en tant que partenaires :

– les trois instituts académiques qui ont participé à la rédaction du *Dictionnaire* :

a) l'Institut de Linguistique « Iorgu Iordan – Al. Rosetti », L'Académie Roumaine, Bucarest[8], responsable pour le projet – acad. Marius Sala (par dr. Monica Busuioc) ;

b) l'Institut de Philologie Roumaine « A. Philippide », L'Académie Roumaine, Iaşi[9], responsable pour le projet – dr. Gabriela Haja ;

c) l'Institut de Linguistique et d'Histoire Littéraire « Sextil Puşcariu », L'Académie Roumaine, Cluj-Napoca[10], responsable pour le projet – dr. Rodica Marian ;

– les deux instituts académiques spécialisés dans le traitement du langage naturel:

a) l'Institut de Recherches pour l'Intelligence Artificielle, l'Académie Roumaine, Bucarest, responsable pour le projet – acad. Dan Tufiş ;

b) l'Institut d'Informatique Théorique, l'Académie Roumaine, Iaşi, responsable pour le projet – acad. Horia Neculai Teodorescu ;

– la Faculté de Lettres, de l'Université « Alexandru Ioan Cuza », Iaşi – responsable pour le projet – dr. Eugen Munteanu.

Dans le projet se sont impliqués, en tant que volontiers, des étudiants, des masterands et les doctorands dans le domaine de la philologie et de l'informatique, de plusieurs facultés de spécialité du pays.

Les principaux objectifs du projet sont : la réalisation du format électronique du plus grand ouvrage lexicographique roumain de type académique, *Le Dictionnaire de la langue roumaine* (comprenant les deux séries DA et DLR), la création d'une archive électronique de textes roumaines comprenant toutes les sources dépouillées en vue de l'élaboration du *Dictionnaire* (donc, qui font partie de la Bibliographie du DLR), la réalisation des liaisons entre le dictionnaire électronique et les sources bibliographiques du celui-ci, organisées en tant que base de données, ce que permettra des interrogations complexes du *Dictionnaire* et la continuation des activités d'actualisation et de publication. Donc, la base de données comprendra, d'une part, la version informatisée du dictionnaire actuel, toutes les sources bibliographiques, en format électronique, et, d'autre part, des programmes

[7] Des informations concernant le projet sont publiées à l'adresse: https://consilr.info.uaic.ro/edtlr/wiki/index.php?title=Despre_proiect.

[8] http://www.lingv.ro/

[9] http://iit.iit.tuiasi.ro/philippide/

[10] http://www.acad-cluj.ro/institut_lingvistica_istorie_literara.php.

spécifiques, réalisés dans le projet, des moteurs de recherche, etc. Le Dictionnaire devient, de cette manière, un instrument et une base de données structurée intelligemment, avec des applications salutaires dans l'actualisation des moyens de recherche lexicographique en Roumanie.

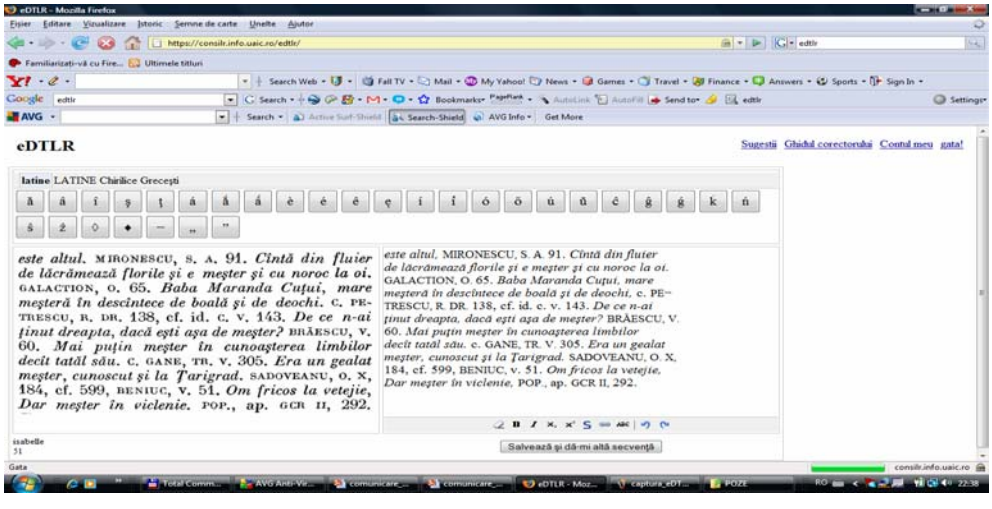
Pour la réalisation du projet on a établi deux étapes principales :

1. Scanner, OCR[11]-iser, corriger le *Dictionnaire de la langue roumaine* (DA + DLR) (approximativement 175.000 pages) et la réalisation du format électronique du *Dictionnaire*.

2. Scanner, OCR-iser, traiter du point de vue informatique la *Bibliographie* du DLR et la réalisation des liaisons entre le dictionnaire électronique et les sources bibliographiques du celui-ci (approximativement 3000 volumes).

Pendant la première étape, la plus problématique activité est celle de corriger le texte reconnu automatiquement après l'avoir scanné. Pour résoudre le mieux possible ce problème on a opté pour une correction « en cascade », opérée, en première phase, par des correcteurs volontiers, non-experts (des étudiants, des masterands, des doctorands de Roumanie – l'Université « Alexandru Ioan Cuza », Iaşi, de l'Université de Bacău, de l'Université de Nord, Baia Mare, de l'Université « Ştefan cel Mare », Suceava –, de Moldavie, etc.). Puis cette corrections a été vérifiée / validée par les experts lexicographes impliqués dans le projet. Ces opérations se font online, sur un site réalisé particulièrement par les informaticiens impliqués dans le projet. Dans la littérature de spécialité, pour les activités collaboratives[12] telles la correction des volontiers de la première phase, il y a déjà un nom : « crowdsourcing » ou « digital sharecropping »[13].

Dans la capture d'écran (screenshot) suivante on peut voir l'interface de correction online pour les volontiers (http://consilr.info.uaic.ro/edtlr/):

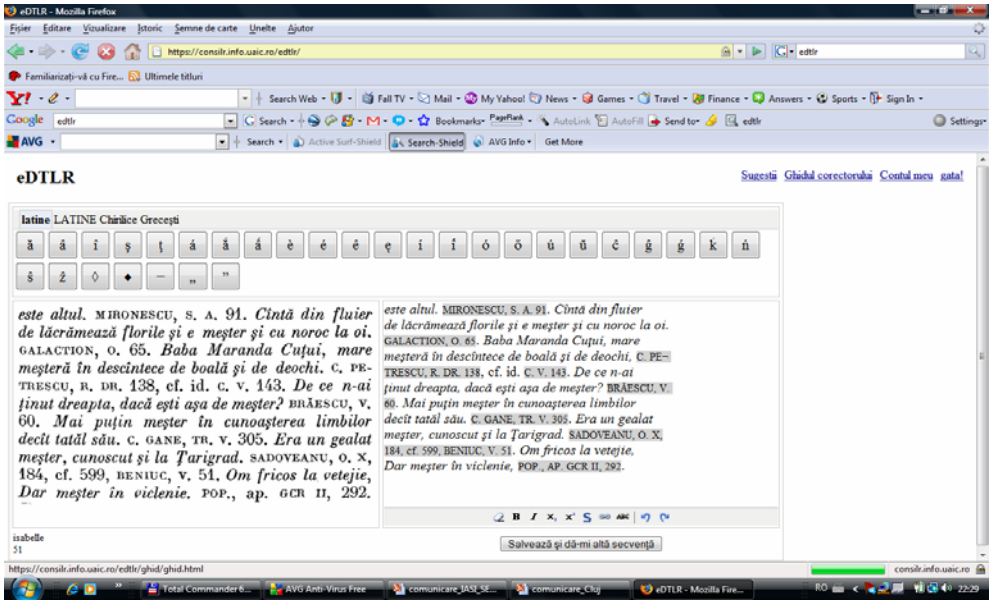


[11] *Optical Character Recognition* permet la transformation du format image (.jpg, .tiff, .gif etc.) en format texte (.doc, .rtf).

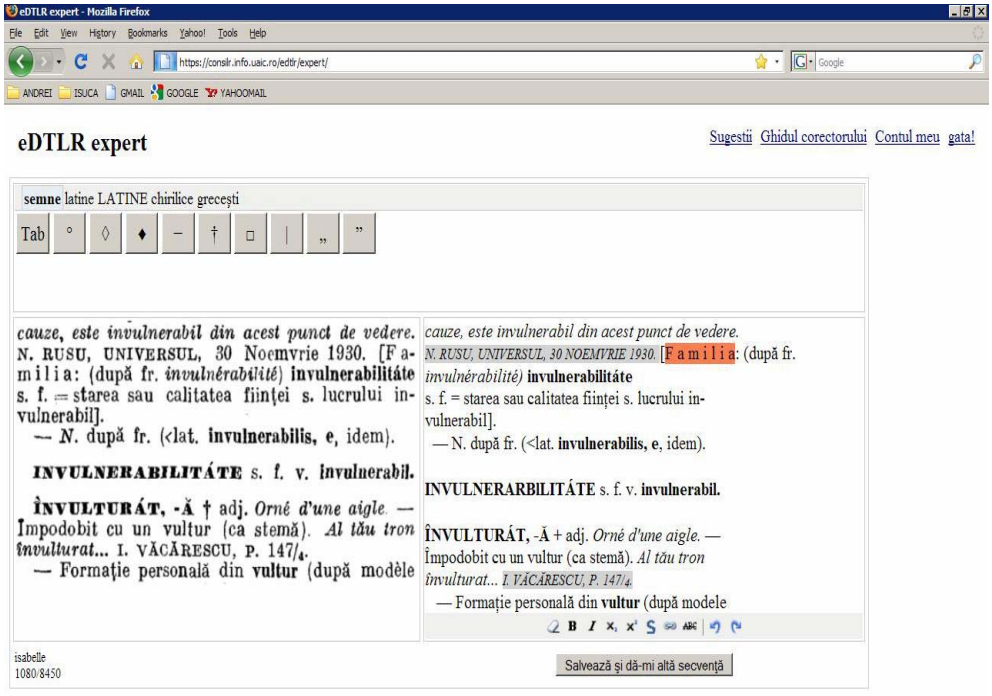
[12] Voir, pour une autre initiative lexicographique collaborative pour la langue roumaine, le projet DEXONLINE, qui représente, à présent, le plus complexe corpus lexicographique roumain on-line. Des informations concernant le projet sont publiées à l'adresse : http://dexonline.ro/

[13] Voir aussi Cristea, Răschip *et alii* 2007: 195–206.

On observe dans la capture d'écran 2 les corrections faites par les volontiers (http://consilr.info.uaic.ro/edtlr/):

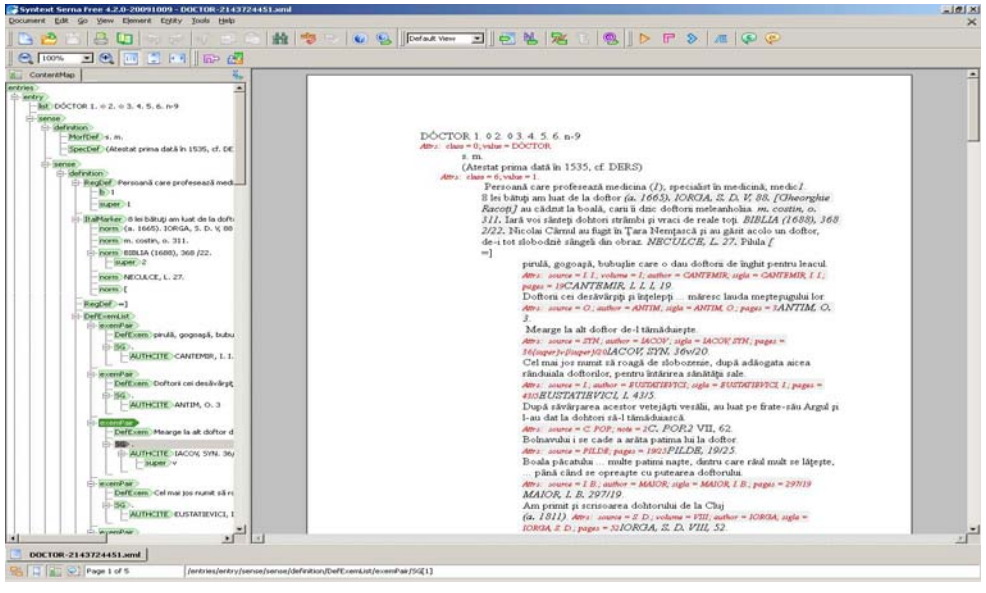


Dans la capture d'écran 3 on peut voir les corrections faites pendant la validation des experts lexicographes (http://consilr.info.uaic.ro/edtlr/):



Après l'achèvement de la correction du texte du point de vue formel, les experts lexicographes vérifieront et corrigeront les structures sémantiques (les «arbres» sémantiques) des entrées, générées après l'opération de parser pour chaque entrée de DA et de DLR dans un format qui permet le traitement du texte avec des moyens informatiques et la rédaction des éditions futures du DLR.

Dans la capture d'écran 4 on peut voir l'interface pour la correction de structure faite par les experts lexicographes (http://consilr.info.uaic.ro/edtlr/):



À la fin de cette année le projet doit être achevé. Jusqu'à présent[14] on a finalisé quelques des activités proposées : on a scanné et OCR-isé les entrées du *Dictionnaire de la langue roumaine* (DA + DLR) ; on a achevé la correction faite par les volontiers. Dès octobre 2008 l'activité de correction faite par les experts a commencé et à présent est en train d'être achevée. Concomitamment, dès mai 2010 on a commencé la correction de structure, réalisé aussi par les experts lexicographes.

Parallèlement, les experts informaticiens ont réalisé, à côté des interface de correction online, le logiciel pour parser le DLR et DA et ils ont préparé l'interface de validation et correction de la structure sémantique des articles générés, donc ils ont raffiné les programmes et les instruments électroniques nécessaires pour le traitement automatique du Dictionnaire et du corpus bibliographique.

On a scanné intégralement les sources DLR existants à l'Institut de Philologie Roumaine «A. Philippide» et quelques sources qui se trouvent à la Bibliothèque Centrale Universitaire „Mihai Eminescu", Iaşi. Ces sources ont été OCR-isées et archivées en tant que base de données. On continue de travailler au programme de

[14] A la dernière éditon de la Conférence *Ressources linguistiques et instruments pour le traitement de la langue roumaine*, Bucarest, 6–7 mai 2010, on a présenté un rapport sur la situation actuelle du projet et on a discuté sur les modalités d'accomplir tous les fins du projet.

reconnaître dans les sources scannées les citations de DA / DLR et au programme qui permet d'extraire des citations des sources scannées. La dernière activité s'est imposée, hors du plan de travail initial du projet, en tant que nécessité dans la collaboration interinstitutionnelle et dans l'évolution de la recherche interdisciplinaire. Plus précisément, après avoir achevé la première édition du *Dictionnaire de la langue roumaine*, l'Académie Roumaine a établi l'actualisation des premiers volumes du Dictionnaire (les lettres A, B, C, F-J), ce qui impose aussi une modernisation des moyens de travail et l'utilisation d'un programme qui permet l'extraction des citations de toutes les sources imprimées inclues dans la Bibliographie du *Dictionnaire*, ce qui représente une solution indispensable.

eDTLR permettra de nouvelles modalités de travail/ étude/ recherche dans la lexicographie roumaine, en incluant sa perspective computationnelle, offrira la seule modalité moderne de compléter et actualiser le *Dictionnaire de la langue roumaine*, offrira la possibilité de consulter interactivement le *Dictionnaire par tous connaisseur* de la langue roumaine de Roumanie ou de n'importe où. eDTLR serra le premier dictionnaire d'une envergure pareille dédiée à la langue roumaine, sur un support électronique. Pour la première fois, les chercheurs pourront trouver les citations directement dans les sources bibliographiques. Pour les experts informaticiens, eDTLR représentera un instrument très important pour les programmes de désambiguïsation sémantique et de traduction automatique.

A notre avis, une étape obligatoire qui suit à la digitalisation du DLR est représentée par la réalisation d'un corpus lexicographique qui inclut eDTLR et les autres dictionnaires essentiels de la langue roumaine (dictionnaires anciens et nouveaux, dictionnaires généraux ou «spéciaux» etc.), alignés au niveau de l'entrée et même du sens. De cette manière, la langue roumaine va se comparer avec les autres langues romanes qui possèdent déjà un corpus pareil (voir le français, l'espagnol, etc.).

Conclusions

Les dictionnaires en format électronique et les corpus de textes, structurés en tant que bases de données, sont importantes de plusieurs points de vue: d'une part, ils permettent la connaissance et la préservation de l'identité d'une culture manifestée dans le plan linguistique et, d'autre part, ils facilitent l'inclusion d'une langue nationale dans le champs d'intérêt de la recherche informatisée des langages naturels, au niveau global.

Les résultats du projet *eDTLR. Dictionnaire de la langue roumaine informatisé* ouvriront une nouvelle étape dans la recherche lexicographique roumaine, ce qui permet l'alignement de la recherche roumaine aux standards internationaux de ce domaine. La version informatisée du dictionnaire et le corpus de textes roumains faciliteront l'accès des spécialistes (linguistes et informaticiens) à un instrument de travail indispensable, longtemps attendu, extrêmement utile pour le développement et la préservation de la langue roumaine du point de vue des (dés)avantages que la globalisation suppose.

Bibliographie

DA = *Dicţionarul limbii române*, tom I–II, Bucureşti, Tipografia ziarului „Universul”, Imprimeria Naţională, 1913–1937.

DLR = *Dicţionarul limbii române*, Serie nouă, tom VI–XIV, Bucureşti, Editura Academiei, 1965–2010.

DWB = *Deutsches Wörterbuch “der Grimm”* = http://germazope.uni-trier.de/Projects/DWB.

TLFi = *Le Trésor de la Langue Française Informatisé* – http://atilf.atilf.fr/

OED = *Oxford English Dictionary* = http://www.oed.com/.

DRAE = *Diccionario de la lengua espanola de la Real Academia Espagnola* = http://buscon.rae.es/draeI/.

Aldea, Dănilă *et alii* 2006: Bogdan-Mihai Aldea, Elena Dănilă, Corina Forăscu, Gabriela Haja, *Dicţionarul limbii române (DLR) în format electronic. Aplicaţii*, in Elena Dănilă, Ofelia Ichim, Florin-Teodor Olariu (eds.), *Comunicare interculturală şi integrare europeană*, Iaşi, Editura Alfa, p. 7–17.

Clim, Dănilă *et alii* 2008: Marius Clim, Elena Dănilă, Gabriela Haja, *Premise ale informatizării cercetării lexicografice academice româneşti*, in *Limba română. Dinamica limbii, dinamica interpretării*, Editura Universităţii din Bucureşti, p. 585–591.

Cristea, Răschip 2008: Dan Cristea, Marius Răschip, *Linking a Digital Dictionary onto Its Sources*, in Marko Tadic, Mila Dimitrova-Vulchanova, and Svetla Koeva (eds.) *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages* (FASSBL 2008), p. 50–52, Dubrovnik, Croatia, September 25–28, 2008.

Cristea, Răschip *et alii* 2007: Dan Cristea, Marius Răschip, Corina Forăscu, Gabriela Haja, Cristina Florescu, Bogdan Aldea, Elena Dănilă, *The Digital Form of the Thesaurus Dictionary of the Romanian Language*, in Corneliu Burileanu, Horia-Nicolai Teodorescu (eds.), *Advances in Spoken Language Technology*, Bucureşti, Editura Academiei Române, p. 195–206.

Curteanu, Moruz *et alii* 2008: Neculai Curteanu, Alexandru-Mihai Moruz, Diana Trandabăţ, *Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing*, in *Proceedings of the COLING 2008 Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, p. 55–63, Manchester, UK, 24 August, 2008.

Dănilă 2007: Elena Dănilă, *Tradiţie şi inovaţie în cercetarea lexicografică românească*, in *Evoluţia şi funcţionarea limbii – perspective normative în noul context european*, Editura Universităţii Suceava, p. 213–215.

Dănilă 2008: Elena Dănilă, *Avantajele informatizării cercetării filologice în studierea limbii române*, in *Româna ca limbă străină – între metodă şi impact cultural*, Iaşi, Casa Editorială *Demiurg*, p. 746–750.

Dănilă, Haja 2009: Elena Dănilă, Gabriela Haja, *Dicţionarul limbii române în format electronic (eDTLR) în perspectiva globalizării*, in « Communication interculturelle et literature », nr. 1 (6), aprilie–mai–iunie, Galaţi, Editura Europlus, p. 269–273.

Haja 2007: Gabriela Haja, *Resurse electronice pentru cercetarea lexicografică românească*, in *Limba română azi*, Iaşi, Editura Universităţii „Alexandru Ioan Cuza”, p. 129–134.

Haja, Dănilă *et alii* 2005: Gabriela Haja, Elena Dănilă, Corina Forăscu, Bogdan-Mihai Aldea, *Dicţionarul limbii române (DLR) în format electronic. Studii privind achiziţionarea*, Editura Alfa, Iaşi, ou la version digitalisée sur www.consilr.info.uaic.ro.

Haja, Forăscu *et alii* 2006: Gabriela Haja, Corina Forăscu, Bogdan-Mihai Aldea, Elena Dănilă, *The dictionary of Romanian Language: steps toward the electronic version*. In *Proceedings of EURALEX 2006*, Torino, Italy, september 2006.

eDTLR – Data Base and Instrument for the Lexicographic Romanian Research

This paper aims at highlighting the changes currently taking place in the Romanian lexicography. Over the last years lexicographers and informaticians worked on the creation of an digitalized version of eDTLR – the *Thesaurus Dictionary of the Romanian Language* and also on the creation of the electronical version of texts containing all the sources from which quotations were extracted in order to create the *Dictionary*. This project closes at the end of 2010. Through its dimensions, eDTLR aligns to important dictionaries of other Roman languages (*Trésor de la langue française informatisé* – TLFi ; *Diccionario de la lengua espanola de la Real Academia Espagnola* – DRAE, *Tesoro della lingua italiana delle originii* and others), large dictionaries, which already have a digitalized version.

Institut de Philologie Roumaine «A. Philippide», Iaşi
Roumanie