Studii de lingvistică 7, 2017, 113 - 124

Un corpus DIY pour l'étude du roumain en diachronie. Stratégies de constitution et stratégies de recherche

A DIY corpus for investigating historical Romanian.
Its creation and research strategies

Ana Zisman[1]

Abstract: The present paper aims to provide an overview of some of the advantages of creating and working with a DIY corpus, *i.e.* a corpus compiled by the linguist, as groundwork for a PhD thesis. Collected in order to investigate the grammatical and pragmatical behavior in historical Romanian of some so-called parenthetical verbs: *a zice/ a spune* 'to say', *a crede* 'to think', *a şti* 'to know', within 5 types of texts from the 16th/17th to the 20th centuries, this DIY corpus represents a necessary alternative as a database of Romania texts. Although its creation demanded some additional steps (*e.g.* the selection of the texts, which is determined by various diachronical factors), such a corpus proves to be relevant for investigating parenthetical verbs in literary, historical and law texts, as well as in formal and informal letters. In order to do so, the paradigm of the afore-mentioned verbs has to be systematized in relation to a precise word frequency per text type.

Key words: DIY corpus, grammar, pragmatics, parenthetical verbs, Romanian.

1. Introduction

Cet article traite de la problématique du travail, en linguistique diachronique roumaine, avec un « *raw corpus* » (cf. Hunston 2006 : 234), ou DYI (angl. 'do it yourself'), destiné à surprendre, sur 5 types de textes et 4-5 siècles, des disparités ou des similitudes d'ordre grammatical et pragmatique de 4 verbes susceptibles de développer un emploi parenthétique : *a zice/a spune* 'dire', *a crede* 'croire' et *a şti* 'savoir'.

Aucune compilation de documents en ancien roumain n'étant disponible, les documents en format papier mis à part, la constitution

[1] Université Babeş-Bolyai Cluj-Napoca; zismana@yahoo.com.

d'un corpus DIY qui réponde aux besoins d'une telle recherche s'est avérée impérative. Contenant des textes provenant des XVIe-XXe siècles, classés par genre, notre corpus DIY s'est imposé donc comme une nécessité. Il peut cependant arriver qu'un corpus DIY soit constitué comme une alternative à des bases de données permettant, certes, des enquêtes complexes sur des phénomènes linguistiques divers, mais dont la dimension – les plus grands comprennent 100+ millions de mots[2] – ne permettrait pas de restreindre le domaine d'investigation.

Néanmoins, si un tel recueil de textes facilite, pour le roumain, l'accès à des études en diachronie, son organisation est susceptible de soulever parfois des inconvénients. A ce titre, nous nous proposons, dans ce qui suit, de donner un aperçu sur les étapes de création d'un corpus DIY, qui dépend d'une série de variables, comme, entre autres, le choix, la nature et la recherche proprement dite des textes. Vu que tous ces facteurs risquent de ralentir le travail du chercheur, notre deuxième but est de montrer à quel point l'utilité des examens effectués sur un corpus DIY contrebalance les difficultés posées par sa constitution.

1.1. Corpus DIY. Quels objectifs ?

L'idée de la constitution d'un corpus DIY s'accorde aux objectifs de notre recherche :

- *primo* : examiner le comportement grammatical et pragmatique de 4 verbes du roumain : *a zice/a spune*, *a crede* et *a şti*, dans leur *emploi parenthétique* ;
- *secundo* : analyser en diachronie les occurrences de ces verbes extraites de textes écrits du XVIe au XXe siècle. Nous partons de l'hypothèse qu'une telle approche diachronique peut révéler des indices sur l'évolution de certaines constructions parenthétiques ;
- *tertio* : saisir les disparités dans l'usage de ces verbes à travers un filtre chronologique et typologique. Cela implique l'exigence d'opérer avec un corpus diachronique fragmenté en 5 sous-corpus, correspondant à 5 types de textes : littéraire, historique, juridique, épistolaire – lettres formelles et lettres informelles.

1.2. Conditions préliminaires de la constitution du corpus DIY

L'un des avantages du corpus DIY réside, à première vue, dans la possibilité de façonner à son gré et de mieux contrôler la collection des textes, ce qui garantit une meilleure connexité linguiste-corpus↔buts (cf. aussi Vaughan & Clancy 2013 : 7). A ce titre, le

[2] Comme le *Corpus d'anglais american historique*, entre autres : http://corpus.byu.edu/coha.

linguiste doit projeter, dans une étape préliminaire, de réflexion, l'architecture optimale de son corpus DIY.

Pour ce qui est du nôtre, nous l'avons tout d'abord imaginé en partant de deux considérations :

- il fallait que la *dimension* du corpus soit relativement restreinte (2.000.000 mots au maximum) ;
- il fallait également que les documents soient *disponibles* et téléchargeables.

2. Méthodologie

La méthodologie sur laquelle nous appuyons notre recherche comprend deux volets : l'un qui concerne la constitution du corpus DIY, plus précisément la recherche des documents (types de textes) (2.1.), et l'autre, qui concerne l'exploitation des données (verbes parenthétiques) (2.2.).

2.1. Un corpus DIY. Quels textes ?

2.1.1. Méthodes de recherche

Traiter de phénomènes linguistiques en diachronie, notamment à partir de documents rares en ancien roumain, comprend de manière habituelle des risques dont nous rappelons l'insuffisance et le manque d'accès immédiat à des textes ; c'est la raison pour laquelle une investigation préalable devient nécessaire : par exemple, ayant établi notre intérêt pour l'étude de tel verbe dans différents types de textes, nous avons cherché sur Internet, pour chaque siècle (du XVIIe au XXe), des titres de documents appartenant aux types de textes envisagés. Cependant, si cette méthode de recherche par titre nous a facilité l'accès à certains types de textes (les textes littéraires notamment), d'autres types, comme, par exemple, les textes épistolaires y ont échappé. Nous avons alors procédé à la recherche des textes dans des bibliothèques digitales : *Digibuc, Dacoromanica, BCU, archive.org*, en ayant comme critère de recherche, à part le titre, des mots-clés (par exemple, pour obtenir des textes juridiques, nous avons effectué une recherche sur des termes appartenant au registre juridique, comme *lege/ leage/ ledge* 'loi').

Mais à quoi sert cette deuxième méthode de recherche, apparemment faite à l'aveugle ? Son utilité réside, du moins selon notre expérience avec les documents en ancien roumain, dans l'alternative qu'elle offre pour trouver des textes indisponibles moyennant une recherche par titre. Il s'agit, le plus souvent, de textes recueillis dans des éditions nouvelles, mais ayant des titres différents. Par exemple, un texte du XVIIe siècle peut se retrouver dans un volume du XIXe siècle, qui provient d'un autre auteur.

En plus, la recherche par mots-clés/morphèmes peut conduire soit à des textes envisagés intuitivement mais *a priori* indisponibles, soit à des textes nouveaux, qui viennent enrichir la palette des documents déjà disponibles. Notons que sonder un tel genre de terrain, inconnu et diachroniquement distant, à travers des mots-clés, présuppose la prise en compte de toutes les graphies possibles des mots-clés en question. Prenons, par exemple, le cas du verbe *zice* 'il dit' du roumain contemporain : ce verbe est susceptible de ne générer aucun résultat lors d'une recherche dans des textes du XVIIe ou du XVIIIe siècles. Il faudrait donc considérer, pour ce verbe, des variantes attestées comme *(d)zice / gice / dzîce*, ou autres.

2.1.2. Critères de sélection des textes

L'étape qui suit la recherche des textes est celle où le linguiste se voit obligé d'en retenir certains et d'en rejeter d'autres, en fonction des objectifs de son travail. Pour notre part, nous nous sommes confrontée à une indisponibilité générale des textes datant notamment du XVIIe siècle, période pour laquelle nous disposons parfois de peu de textes ou même d'un seul document dans le cas de certains types de textes. Cette limitation a influé sur le choix des textes appartenant aux autres périodes.

Reste d'une importance majeure, du moins pour l'ancien roumain, le *critère graphique* : ainsi, même si notre enquête dans les bibliothèques digitales nous a conduite à des textes remplissant nos conditions (période et appartenance au type de texte souhaité), ils n'ont pas été retenus, car il s'agissait de textes en alphabet cyrillique, et non pas latin, ce qui rendait impossible l'extraction par recherche électronique des occurrences de verbes à analyser. De même, pour la pertinence de notre étude, nous avons retenu uniquement des *textes originaux*, en graphie latine et dans la langue de l'époque, et non pas des versions plus récentes.

S'y ajoute le *critère diatopique*, qui nous amenée à sélectionner des textes provenant de plusieurs régions du territoire roumain. Il faut remarquer que les textes datant du XVIIe et du XVIIIe siècle et provenant de zones géographiques distinctes comportent des disparités d'ordre graphique, i.e. *zice* en Valachie vs *gice* ou *dzice* en Moldavie.

2.1.3. Conversion des documents

Ayant finalement établi un corpus de textes filtrés selon les critères mentionnés ci-dessus, nous avons rencontré un autre inconvénient : la plupart des documents recueillis, surtout ceux en ancien roumain, étaient en version « lecture seule » (*read only*), toute

manipulation des données étant bloquée. Obtenus en version Word, mais aussi sous forme scannée, ces textes ont dû être convertis en format pdf, plus adéquat pour la recherche.

Nous précisons, d'un côté, que le recours à des logiciels consacrés aux recherches lexicométriques, tels Hyperbase ou autres, est exclu pour l'exploitation de notre corpus, car la syntaxe du roumain comporte des constructions qui échappent à l'ordre canonique des constituants. Il s'agit de structures du type V/ VS, dues à des déplacements des constituants au sein de la phrase ; de plus, en ancien roumain, le COD, le sujet et même le verbe peuvent avoir une position inhabituelle, sans mentionner le cas de la double saturation de la transitivité du verbe par un COD / une complétive, ou encore les constructions où les verbes *a zice/a spune* s'accompagnent simultanément de la conjonction *că* 'que' et d'un DD. En vue d'un examen syntaxique, il faut éviter de constituer des corpus exploitables uniquement de manière automatique, pour pouvoir toujours avoir accès aux contextes dont proviennent les occurrences de chaque verbe analysé. Nous précisons, de l'autre côté, qu'établir la formule grammaticale des constructions verbales qui nous intéressent, étudiées en diachronie et dans plusieurs types de textes, prend du temps, ce qui justifie notre choix de ne pas compter sur une exploitation manuelle du corpus.

2.1.4. Pour une recherche quantitative et qualitative sur un corpus DIY

Notre choix d'examiner plusieurs types de textes a posé problème, à cause de la rareté des documents recherchés (parfois introuvables ou inexistants), ce qui nous a obligée de classer les documents trouvés dans des catégories plus souples. En ce sens, prenant comme point de départ la période la plus significative – le XVIIe siècle –, nous avons essayé d'établir certains *patterns* pour les textes des autres périodes.

Ainsi, pour pouvoir exploiter en diachronie un sous-corpus de *textes littéraires* (qui comprendrait, dans le cas d'une analyse en synchronie, de multiples ramifications), nous avons considéré des textes qui font la critique des mœurs, même s'ils ont parfois des formats différents : textes en vers comme le seul texte littéraire en vers du XVIIIe siècle de notre corpus *vs* textes en prose, comme les textes appartenant aux XVIIe, XIXe et XXe siècles. Ce compromis semble affecter l'équilibre dimensionnel du corpus, vu que les textes que nous traitons connaissent des dimensions inégales.

Vu l'objet de notre recherche, à savoir l'étude de 4 verbes parenthétiques dans plusieurs types de textes, ce qui nous intéresse, ce n'est pas le nombre de textes, mais le nombre de mots que comprend

chaque sous-corpus. L'utilité du dénombrement des mots[3], effectué à l'aide des logiciels disponibles en ligne, se voit dans le fait qu'il permet ensuite d'établir l'orientation syntaxique de chaque verbe, pour chaque type de texte.

Nous avons synthétisé les résultats obtenus pour les verbes *a zice/a spune* dans le tableau 1 :

Sous-corpus	**Lettres formelles (XVIe-XXe)**	**Lettres informelles (XVIIe-XXe)**	**Textes historiques (XVIIe-XXe)**	**Textes juridiques (XVIIe-XXe)**	**Textes littéraires (XVIIe-XXe)**
Nombre de mots	≈ 50.000	≈ 40.000	357.026	507.474	490.197
Occurrences de *a zice/ a spune*	91	58	350	997	830
Total corpus DIY	1.354.787				

Tableau 1 : Distribution diachronique des verbes *a zice/a spune* dans 5 types de textes

2.2. Analyse des données dans un corpus DIY

Nous montrerons dans ce qui suit de quelle manière nous avons utilisé notre corpus DIY pour l'étude des verbes *a zice/a spune*, *a crede* et *a şti*. Nous nous sommes proposé d'examiner le paradigme de chacun de ces verbes à l'indicatif présent et au passé composé.

Appartenant à la classe des verbes qu'on range dans la classe des verbes parenthétiques (cf. Urmson 1952), ces verbes développent, en tant que verbes transitifs, un emploi dual : ils peuvent régir des complétives moyennant un élément subordonnant (V *că* Q 'V que Q'), mais peuvent aussi apparaître en position détachée, suite à l'effacement de la conjonction (V, Q). La suppression du complémenteur est en même temps responsable du déplacement du verbe, qui, ayant occupé une *position figée*, antéposée à la complétive, devient *mobile*, pouvant être postposé (Q, V). L'affaiblissement du lien grammatical fait que, le plus souvent, le verbe acquière un rôle discursif (indiquant un certain degré de subjectivité de la part du locuteur) : *Ştiu că vine* 'Je sais qu'il

[3] Comptabiliser les mots des documents datant du XVIIe et du XVIIIe siècles s'est avéré parfois difficile, vu que certains de ces documents s'accompagnent de commentaires qui appartiennent aux auteurs des recueils où ils apparaissent. Pour ce faire nous avons parcouru les textes en vue d'isoler les mots qui nous intéressaient. Pour les textes épistolaires le dénombrement est approximatif.

vient' → *Vine, ştiu* 'Il vient, je sais'.

En prenant appui sur notre corpus diachronique, nous nous proposons d'évaluer les cas de *rection faible* (cf. Blanche-Benveniste 1989), en les comparant avec ceux où le verbe sature sa transitivité par une complétive. Nous présentons, dans ce qui suit, les étapes de notre analyse.

2.2.1. Recherche par sous-corpus et par période

L'étude des 4 verbes qui nous intéressent dans le corpus débute par l'évaluation globale du contexte de notre recherche :

- *primo*, notre corpus DIY est segmenté en sous-corpus correspondant respectivement aux 5 types de textes et aux 5 siècles pris en compte ;
- *secundo*, l'analyse du corpus DIY implique le recours à un logiciel de recherche qui réponde à nos besoins, à savoir faire des observations concernant la *fréquence* des verbes étudiés (cf. Hunston 2006), ainsi que leur *paradigme* – formes conjuguées – et leurs *emplois parenthétique* vs *rectionnel fort*.

Tout cela nous a conduite à effectuer des recherches sur les occurrences de chaque verbe par sous-corpus et par période. Une méthode comme la nôtre, dont nous présentons ci-dessous le prototype (v. Prototype de recherche), simplifie le processus de recherche des occurrences, car, pour chaque siècle, elle permet de traiter simultanément des occurrences d'un verbe dans plusieurs textes.

Prototype de recherche

A zice / a spune	→	*Lettres formelles* → XVIe siècle
		Lettres formelles → XVIIe siècle
		Lettres formelles → XVIIIe siècle
		Lettres formelles → XIXe siècle
		Lettres formelles → XXe siècle

2.2.2. Outils et méthode de recherche

Le travail avec un corpus DIY implique, comme nous l'avons déjà montré, une série d'opérations supplémentaires, qui précèdent le traitement proprement dit des données. Ainsi, après avoir établi la stratégie d'analyse des 4 verbes qui nous intéressent dans notre corpus DIY, il a fallu concevoir un instrument qui nous permette de manœuvrer leurs occurrences, en procédant de la même façon que pour la recherche des textes, à savoir *via* des mots-clés. Notons, cependant, que cette méthode ne vise plus une recherche lexicale, mais un examen des

morphèmes d'un seul et unique verbe. Alors, comme nous avons retenu des textes en format pdf et que nous ne comptions sur aucun autre outil de recherche, nous avons considéré utile de recourir à l'outil que propose le Portable Document Format (pdf). Accessible par le raccourci clavier Ctrl + Shift + F, dans un document pdf ouvert, il effectue des recherches basiques ou complexes par mot / morphèm ou, dans notre cas, dans un dossier comprenant plusieurs textes (v. fig. (1) pour *a zice*, sous-corpus *Textes historiques*, XVIIe siècle) :

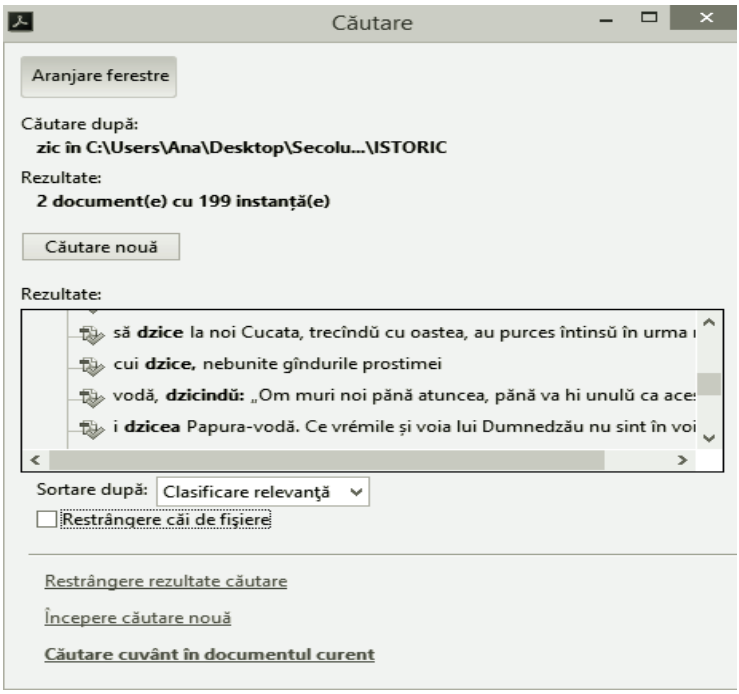

Fig. 1 : Recherche sur sous-corpus par morphème

Dans le cas d'une recherche comme la nôtre, le principal inconvénient du travail avec les données d'un corpus DIY vient de la grande quantité d'information qu'il recueille et qui ne comporte aucune organisation (grammaticale) pouvant offrir des indices sur un possible classement des résultats selon la personne grammaticale, etc. Afin d'atteindre les objectifs de notre recherche, nous avons pris en compte plusieurs variantes d'un même morphème, par exemple *zic/ (d)zic/ gic/ (d)zis* pour *a zice*, *sp/ spu* pour *a spune*, etc.

Le deuxième inconvénient consiste dans le fait que le total des résultats que génère une telle recherche, pour un morphème particulier, ne peut pas offrir un inventaire pertinent en lui-même, vu que le morphème *zice*, par exemple, est contenu tant dans le

présent *zice* 'il dit' que dans le futur *voi/ vei/ va/ vom/ veți/ vor zice* et le conditionnel *aş/ ai/ ar/ am/ ați/ ar zice*, modes et temps grammaticaux qui échappent à nos critères de sélection.

Troisièmement, il est arrivé que la recherche n'affiche aucun résultat pour un morphème donné. Cela s'explique par le fait qu'il y a eu des documents pdf, surtout ceux en roumain des XVIIe et XVIIIe siècles, qui contenaient des caractères graphiques incompatibles avec notre outil de recherche. Notre solution a été de substituer ces textes par d'autres variantes, écrites toujours dans la langue de l'époque.

Vu la distribution chaotique des occurrences de chaque verbe, nous avons jugé impératif de parcourir les apparitions de ces verbes (notamment *a zice / a spune*) dans chaque sous-corpus et d'essayer, pour chaque personne grammaticale, au présent et au passé composé, de comptabiliser : les tendances de subordination par plusieurs conjonctions (le plus souvent les conjonctions *că* et *să* 'que') ; la nature et la position du sujet (SV, V (sujet sous-entendu), VS) ; la disposition pour la rection faible, mais aussi la tendance à s'accompagner d'un discours direct (DD). Cette démarche nous a conduite aux résultats suivants : 45 cas du type SV + *că*, 34 VS + *să*, 5 V en emploi parenthétique, 30 SV + DD. Nous sommes ainsi arrivée à un nombre total d'occurrences, pour chaque verbe, par sous-corpus.

Nous avons pu constater que, au-delà du travail laborieux que suppose une recherche si détaillée, celle-ci permet de gérer l'inventaire des occurrences d'un verbe (parenthétique) pour une personne grammaticale spécifique ou pour toutes les personnes, dans un sous-corpus particulier, tout en permettant également des études contrastives d'un même phénomène dans des types de textes différents.

2.2.3. Distribution des données. Diagrammes et tableaux

Apres avoir recueilli les données pour chaque verbe par sous-corpus et par période, nous avons procédé à leur inscription dans des diagrammes, en opérant avec des statistiques[4]. Ces dernières établissent un rapport quantitatif entre le nombre des occurrences d'un verbe et le nombre des mots que comprend un certain sous-corpus. Néanmoins, bien qu'opportune pour le calcul de la fréquence des formes conjuguées des 4 verbes étudiés, cette méthode ne saurait conduire à des résultats concrets concernant les emplois d'un verbe donné, sa disposition combinatoire ou les effets discursifs qu'il produit.

S'est imposé, par conséquent, la mise en place d'un modèle de distribution complexe qui puisse rendre compte de toutes les fonctions grammaticales (ordre des constituants, etc.) et pragmatiques de chacun des 4 verbes, sous forme de tableau. A ce titre, pour les verbes

[4] Pour l'exploitation en diachronie des corpus par le biais des statistiques v. aussi Sinclair (1995), Hilpert (2006 : 243-256), Baron *et al.* (2009 : 41-67), entre autres.

a zice/a spune, nous avons proposé des tableaux par type de texte et par période, tels que le tableau 2 :

Formes
Mode : indicatif
Temps : présent / autre
Personne, nombre
DI – rection forte (+ *că*/*să*)
Rection faible
Emplois particuliers/ Observations
Total

Tableau 2 : Modèle de distribution des informations pour les verbes *a zice/a spune* par sous-corpus et par période

Pour synthétiser, nous dirons que, dans le contexte de notre recherche, le succès du travail avec des tableaux réside dans le fait qu'ils permettent de :

- fournir des informations détaillées sur : les formes conjuguées des verbes, les cas de subordination (dénombrement en fonction de la nature de l'élément subordonnant), les cas de parenthéticité, etc.
- identifier et faire des observations sur les emplois que nous jugeons comme *particuliers*, par exemple, les emplois citationnels des verbes *a zice/a spune*, etc.
- saisir des différences ou des analogies entre les différents types de comportements d'un verbe, suivant ou non la personne ou le temps, dans le même sous-corpus, pour une période particulière, ou, diachroniquement, dans plusieurs.

2.2.4. Interprétation des données

Après le recueil des données suit l'analyse proprement dite de ces données. Nous rappelons qu'il s'agit d'examiner en diachronie le comportement dans différents types de textes de quelques verbes qui développent un emploi parenthétique (*a zice/ a spune*, *a crede* et *a şti*).

2.2.5. Résultats

À titre d'exemple, nous exposons ci-dessous les résultats qu'a fournis l'examen des occurrences de *a zice/a spune* (première personne, indicatif présent) dans notre corpus DIY contenant 5 types de textes. Ces résultats ont majoritairement confirmé nos hypothèses : pour chaque verbe, chaque sous-corpus enregistre des fluctuations par rapport à sa fréquence, sa tendance à la subordination, les cas

de parenthéticité, son potentiel combinatoire, ses valeurs modales. Le bilan de ces fluctuations pour *a zice/a spune* à la première personne est visible dans le tableau 3[5] :

Type de texte	**Occ.**	**Mots / sous-corpus**	**Rection forte**	**Emplois parenth.**	**Autres emplois**	**Ordre constit.**	**Valeurs**
Lettres formelles (XVIIe-XXe s.)	21	≈50.000	5 *să* 6 *că* 1 *precum că* 1 *cum că*	2 V 1 *precum* 1 *iarăşi* 1 *cum*	1 V+faux DD 2 DD + *că*	15 V 5 SV 1 VS	Évidentiel Épistémique Performatif
Lettres informelles (XVIIe-XXe s.)	2	≈ 40.000	2 *că*	0	0	2 V	Epistémique
Textes littéraires (XVIIe-XXe s.)	99	490.197	16 *că* 7 *să* 1 *cum că*	27	10 express. au subj. 34 faux DD 4 + advers.	47 SV 27 V 25 VS	***Zic*** 'je dis' Épistémique Évidentiel Métadiscursif Marqueur d'approx. Volitif Acte de langage : <conseiller> ***Spun*** 'je dis' Entre évidentiel et epistémique Persuasif Acte de langage : <ordonner>
Textes historiques (XVIIe-XXe s.)	27	357.026	10 *că* 8 *să*	7	2 VS	20 V 4 VS 3 SV	Épistémicité réduite
Textes juridiques (XVIIe-XXe s.)	12	507.474	0	12	0	12 V	Métadiscursif Interjectif Acte de langage : <prévenir>

Tableau 3 : Inventaire diachronique de *a zice/ a spune* (première personne) dans 5 types de textes

[5] Lorsque V_{1Sg} apparaît à côté d'une proposition complétive, leur lien étant soit marqué graphiquement (:), soit absent, il s'agit, d'après nous, d'un type particulier de discours direct que nous appelons *fausse* ou *pseudo-citation* (faux DD).

3. Conclusions

Par la présente étude, nous nous sommes proposé de donner un aperçu sur un type particulier de corpus – le corpus DIY – et sur son applicabilité aux besoins d'une recherche doctorale. En ce sens, nous avons vu que, à la différence des bases de données déjà constituées, la conception d'un tel corpus nécessite des étapes supplémentaires et une méthodologie qui repose sur plusieurs critères : la nature des textes qui le composent, leur accessibilité, leur pertinence, etc. Construit essentiellement comme un instrument efficace, le corpus DIY reflète davantage la contribution du linguiste à la constitution des corpus, car celui-ci façonne cette collection de textes selon ses intentions. Les corpus DIY assurent une liberté dont les avantages contrebalancent, d'après nous, ses inconvénients. Ainsi, nous avons vu, à travers l'étude de 4 verbes parenthétiques en roumain des XVIIe-XXe siècles, dans 5 types de textes, que le travail avec un seul type de texte par siècle n'empêche pas de tirer des conclusions appropriées sur le comportement grammatical et/ou pragmatique de ces verbes. Cela s'explique par le fait que, à part la création d'une base de documents telle un corpus DIY, il faut envisager, notamment pour faire des analyse syntaxiques et discursives, des méthodes optimales d'exploitation des données qui en proviennent. Par conséquent, nous avons brièvement exposé la nôtre, qui implique aussi bien des outils personnalisés de recherche que des outils de dénombrement des mots des documents par type de texte.

Toutes ces opérations, auxquelles s'ajoutent des statistiques, des schémas et des tableaux, nous ont permis de conclure qu'un corpus DIY, comme le nôtre, même s'il ne comprend que 1.354.787 mots, peut générer des résultats validant nos hypothèses de recherche.

Références bibliographiques

Baron, A. *et al.* (2009), "Word frequency and key word statistics in historical corpus Linguistics", in Ahrens, R., Antor, H. (eds), *Anglistik: International Journal of English Studies*, 20/1, p. 41-67.

Blanche-Benveniste, C. (1989), « Constructions verbales 'en incise' et rection faible des verbes », in *Recherches sur le français parlé*, 9, p.53-73.

Hilpert, M. (2006), "Distinctive collexeme analysis and diachrony", *Corpus Linguistics and Linguistic Theory*, 2/2, p. 243-256.

Hunston, S. (2006), "Corpus Linguistics", *Linguistics*, 7/2, p. 215-244.

Sinclair, J. (1995), *Corpus Concordance Collocation*, Oxford University Press.

Urmson, J. O. (1952), "Parenthetical verbs", *Mind*, 61/244, p. 480-496.

Vaughan, E., Clancy, B. (2013), "Small Corpora and Pragmatics", in Romero-Trillo, J. (eds), *Yearbook of Corpus Linguistics and Pragmatics*, vol. 1, Springer, Dordrecht, p. 53-73.