UNIVERSITÀ DEGLI STUDI DI UDINE

Corso di Dottorato di Ricerca
in
Scienze linguistiche e letterarie
(XXV ciclo)

TESI DI DOTTORATO DI RICERCA

La forme et l'usage des parémies.

Classification lexico-grammaticale, modélisation de grammaires locales et étude de fréquence

Tutor
Prof. Sergio Cappello
Prof.ssa Mirella Loredana Conenna

Dottorando
Mario Marcon

ANNO ACCADEMICO
2013/2014

Résumé

La présente thèse porte sur la forme et sur l'usage des parémies françaises. Pour mieux réexaminer l'acquis parémiologique sur ces deux aspects, nous mettons en œuvre une rencontre interdisciplinaire et translinguistique entre parémiologie, phraséologie, linguistique de corpus, linguistique computationnelle, lexicologie et syntaxe. Sur l'escorte de la littérature dans ces domaines, nous proposons d'envisager la parémie comme une *Gestalt linguistique* et identifions la *séquence lexico-grammaticale* comme l'unité linguistique servant à saisir les régularités lexico-grammaticales parémiques. La description lexico-grammaticale des parémies suit une approche hybride harrisienne et sinclairienne qui aboutit à l'esquisse d'une *grammaire parémique*. La description lexico-grammaticale supporte la modélisation de requêtes informatiques (expressions régulières et automates à états finis) en vue du repérage des parémies dans des corpus de français contemporain. L'attention finale à l'établissement d'une liste de fréquence des parémies permet d'enrichir la littérature parémiologique de données quantitatives et de nouveaux constats fondés sur des corpus. Notre étude se présente ainsi comme un exemple de *parémiologie linguistique basée sur l'usage* et propose un cadre méthodologique fédérateur valable pour d'autres recherches à venir.

Mots-clés : corpus, français, fréquence, parémie, parémiologie, phraséologie, repérage automatique

Remerciements

Je remercie mes directeurs de thèse. *Nul ne peut servir deux maîtres à la fois*. Moi, je peux. Du moins, j'ai fait tout mon possible pour contredire cette parémie.

Je remercie mes parents et ma famille. *Pas à pas on va loin*. Et on est là tous ensemble.

Je tiens à remercier tous les professeurs, les chercheurs, les collègues, les étudiants et tous les administratifs qui m'ont supporté. Parfois, *il vaut mieux être seul qu'en mauvaise compagnie*. Heureusement, j'ai été en bonne compagnie.

Je remercie tous mes amis. *Loin des yeux,* jamais *loin du cœur*.

On ne peut aller contre son étoile. Franchement, moi, je l'ai à mes côtés. Merci.

À mes parents et à toute ma famille

À Pierre Cadiot
parce qu'au bout du compte, on fait tous partie d'une seule Gestalt

Table des matières

Index des figures

Index des tableaux

INTRODUCTION

Dans une de ses contributions les plus connues consacrées à la sémantique proverbiale, Kleiber écrit :

> « […] nous avons une compétence de ce qu'est sémantiquement (et formellement aussi, mais le problème est un peu différent) un proverbe, même s'il y a des hésitations et des erreurs possibles dans les listes et même si nous ne pouvons pas définir le plus souvent en quoi consiste cette compétence […]. Nous pouvons en effet, plus ou moins bien, faire le départ sémantique entre une phrase qui est un proverbe (ou qui pourrait en être un) et une phrase qui n'en est pas (ou qui ne peut pas le devenir). […] c'est bien une question de moule ou de schème sémantique proverbial qui est en jeu. C'est cette même structure sémantique qui doit nous guider dans une autre manifestation, souvent signalée, de notre compétence sémantique proverbiale, à savoir l'aptitude à fabriquer des phrases […] qui, autant du point de vue formel, le côté le plus visible, que du côté sémantique, peuvent passer pour des proverbes. Cette aptitude à confectionner des proverbes représente à notre avis un des arguments les plus forts en faveur de l'hypothèse d'un sens propre attaché à la catégorie des proverbes. Si l'on peut fabriquer des proverbes, c'est bien que l'on dispose d'un modèle de la structure sémantique des proverbes. […] nous pouvons reconnaître, fabriquer, interpréter ou essayer d'interpréter des proverbes, si nous pouvons décider si telle phrase pourrait ou non devenir un proverbe, c'est parce que nous avons la compétence du sens proverbial, la connaissance (devenue) intuitive d'une structure sémantique générale sur laquelle s'articulent les proverbes particuliers. Et donc on peut – et on doit – essayer de mettre en relief ce sens définitoire unitaire » (2000 : 43-45).

Cette citation nous servira de manifeste et de référence constante tout au long de notre étude. Qu'il soit en discours ou enregistré dans un recueil parémiographique, on sait identifier un proverbe et, en général, une parémie parmi plusieurs combinatoires, parfois de façon impropre par rapport aux classifications des parémiologues. Et « c'est bien une question de moule » qui facilite la segmentation et l'extraction sémantique d'un proverbe en discours, mais aussi la production de combinatoires qui « peuvent passer pour des proverbes ». L'appel au moule ou à un modèle schématique sémantique intériorisé (contrecarré par le gestaltisme phénoménologique des *formes sémantiques* de Cadiot & Visetti (2001, 2006)) invite à

regarder de près les régularités sémantiques dont les parémies sont porteuses dans leur singularité et à les généraliser.

À la différence de Kleiber, nous nous interrogerons non pas sur la compétence sémantique, mais sur la compétence formelle qu'il a évoquée et qui permet aussi de reconnaître un proverbe. Sur l'escorte de la littérature en la matière, notre recherche visera la description lexico-grammaticale d'une liste de parémies pour mettre en relief les régularités lexico-grammaticales les plus saillantes. Nous essayerons donc de faire ressortir des modèles lexico-grammaticaux de matrice proverbiale (et, en général, parémique) qui permettent, d'une part, la reconnaissance et l'extraction formelles des parémies dans le discours et, d'autre part, la 'production' de pseudo-parémies, détournements et autres séquences formulaires.

À côté de ce but, nous souhaitons réattribuer aux proverbes et aux parémies leur statut d'unité phraséologique. S'il vrai qu'elles relèvent de la parémiologie, il est vrai aussi que leur combinatoire peut faire l'objet d'étude au sein de la phraséologie, comme le remarque Permyakov déjà dans les années 1970 (1979 : 135). L'étroite division entre disciplines promues par certains parémiologues a marqué une séparation qui nuit, à notre avis, à l'une quant à l'autre. C'est pour rapprocher ces disciplines que nous transposerons des traitements propres à la phraséologie qui nous aideront à saisir les régularités formelles des parémies et les resituer dans le continuum phraséologique.

Nous nous servirons de l'établissement d'une liste de fréquence d'usage dans le discours écrit comme prétexte pour mettre en œuvre cette opération de rencontre interdisciplinaire. Malgré les parémies soient souvent menacés de disparition par les parémiologues eux-mêmes, la littérature parémiologique francophone ne dispose pas à l'heure actuelle d'une liste de fréquences qui mesure la vitalité parémique dans l'usage. Nous tirerons ainsi parti de la linguistique de corpus qui a ravivé la phraséologie. Plus précisément, elle a poussé le repérage automatique des séquences formulaires et, en général, le traitement statistique des données linguistiques. Malgré cela, la détection automatique et les études de fréquence des parémies basées sur corpus demeurent un terrain peu exploré, en milieu francophone en particulier. On peut comprendre le découragement et l'embarras de certains parémiologues face à une telle lourde tâche. Pourtant, à un moment où les corpus représentent des outils empiriques indispensables et que leurs tailles s'agrandissent au fur et à mesure, ce serait dommage de ne pas essayer cette voie. Ce sera donc pour mesurer la présence des parémies dans les textes que nous aurons besoin de décrire leurs régularités formelles. C'est par l'identification de ces régularités que nous pourrons les reconnaître plus aisément dans des corpus de français contemporain.

Plan de la thèse

Dans le premier chapitre, nous proposerons une réflexion théorique autour du proverbe en particulier, et des parémies en général, d'après deux perspectives qui, à notre connaissance, n'ont pas encore fait l'objet d'approfondissements. D'abord, nous essayerons de mieux décrire la relation entre quelques unités minimales descriptives en linguistique et le proverbe. Plus précisément, nous questionnerons le statut de ses unités comme, entre autres, l'*énoncé*, la *phrase* et la *structure*. Nous évaluerons leur utilisation par les parémiologues pour souligner leurs apports à l'encadrement des proverbes et des parémies comme faits linguistiques. Nous proposerons ainsi notre unité minimale : la séquence lexico-grammaticale. Par la suite, nous présenterons une *métaclassification* parémiologique. Elle mettra en évidence la quantité d'étiquettes attribuées aux séquences formulaires et leur variation dans l'usage par les spécialistes. Ce qui nous conduira à reconsidérer l'importance de la surface lexico-grammaticale et à expliciter les principes qui guideront notre classification pour dégager les régularités lexico-grammaticales de notre corpus parémique.

Le deuxième chapitre encadrera notre étude au sein de deux paradigmes de recherche : la *parémiologie empirique* et la *linguistique de corpus*. Nous détaillerons les traits saillants de chaque paradigme et aboutirons à la proposition finale d'un croisement entre les deux et concevoir ce que nous avons appelé : *parémiologie linguistique basée sur l'usage* dans le sillage de la *parémiologie linguistique* proposée par Conenna.

Le troisième chapitre montrera une revue critique de la partie de littérature parémiologique consacrée au traitement quantitatif des parémies, voir à l'établissement de listes de fréquence. Nous nous concentrerons sur les caractéristiques des listes de parémies, sur la nature des sources fouillées ainsi que sur les mesures et sur l'interprétation des fréquences. Nous avons essayé de tirer parti du nombre le plus élevé d'expériences pour proposer un cadre méthodologique fédérateur qui nous aidera pour notre étude.

Dans le quatrième chapitre, nous examinerons les pratiques liées à l'activité de repérage des parémies dans les textes. Nous distinguerons, d'une part, les études qui portent sur la reconnaissance des parémies dans des sources textuelles sur support papier et, d'autre part, celles qui approfondissent la détection (semi-)automatique. Ces dernières seront sous-distinguées d'après les requêtes informatiques sélectionnées par les parémiologues : les expressions régulières et les automates à états finis, notamment les grammaires locales. La

gamme des techniques illustrées complétera ainsi le cadre méthodologique que nous développerons dans le troisième chapitre.

Dans le cinquième chapitre, nous dévoilerons notre corpus parémique et motiverons notre choix. Par la suite, nous expliquerons l'étape d'annotation morphosyntaxique et de lemmatisation de notre corpus parémique à l'aide de l'annotateur automatique *TreeTagger*. Loin de nous avoir mis à l'abri d'un travail de révision manuelle, cette activité nous a permis de nous confronter directement avec les spécificités lexico-grammaticales des parémies. Ensuite, nous approfondirons notre démarche de classification lexico-grammaticale et quelques remarques sur des faits parémiques qui sont ressortis lors de l'analyse des séquences lexico-grammaticales.

Le sixième chapitre sera dédié à la description des corpus que nous interrogerons pour notre étude de fréquence et motivera notre choix par rapport au cadre méthodologique. Nous montrerons les programmes du logiciel *Unitex* qui nous serviront pour lancer nos recherches sur les corpus, mais aussi pour la modélisation de nos requêtes informatiques, c'est-à-dire des grammaires locales ou graphes syntaxiques. Pas à pas, nous dessinerons notre modèle de graphe et mettrons en avant la centralité de notre classification lexico-grammaticale pour sa constitution.

Pour conclure, le septième chapitre montrera une application de notre cadre fédérateur, de notre classification lexico-grammaticale et de nos modèles de grammaires locales à un échantillon de notre liste de parémies pour estimer leur fréquence dans les corpus Leipzig choisis ainsi que pour observer de près leur usage en discours.

Quelques principes-clés et quelques avertissements

Pour le déroulement de notre étude, nous ferons nôtre la suggestion méthodologique proposée par Kleiber de travailler dans le cadre d'une *linguistique cumulative* :

> « c'est-à-dire une linguistique qui ne réinvente pas la roue tous les matins » (dans Conenna 2010 : 67).

En ce sens, dans notre tentative de mettre en route une *parémiologie linguistique cumulative*, nous essayerons de tirer parti du nombre le plus élevé d'expériences qui ont abordé nos sujets de recherche.

D'autre part, nous nous inspirerons également de l'indication méthodologique (et professionnelle) suggérée par Benoît de Cornulier :

> « je dirais à un type jeune qui veut faire de la linguistique, qu'il faut vraiment qu'il approfondisse plusieurs langues très différentes les unes des autres » (dans Conenna 2010 : 54).

Toujours avec une attitude cumulative, nous accorderons une attention particulière à la littérature parémiologique francophone et, à même titre, nous thésauriserons les connaissances ainsi que les pratiques telles qu'elles sont exposées dans les littératures parémiologiques non francophones.

D'après nos propos translinguistiques et interdisciplinaires, il vaut mieux formuler quelques mots de précaution. Tout d'abord, une précision métaterminologique qui ne va pas sans conséquence. Comme nous croiserons plusieurs approches, il est évident que la notion de ce qu'on nommera par *proverbe* ou *parémie* pourra varier d'un chercheur à l'autre, d'une tradition parémiologique stricte à une approche phraséologique moins contraignante. Une étude comparée à proprement parler veut qu'il y ait un *tertium comparationis* objectif et partagé sur lequel on fonde la comparaison. Ce *tertium comparationis* n'existe pas (encore) en parémiologie parce qu'il correspondrait à une définition translinguistique et transculturelle sur le statut linguistique et sémiotique du proverbe et de la parémie. Sans rentrer ici dans les débats que nous évoquerons dans § 1, nous ferons abstraction de ce manque de convergence. Cela n'équivaut pas à dire que nous gommerons les nuances conceptuelles que les parémiologues et les phraséologues attribuent au *proverbe* et à la *parémie* dans leurs études. Notre abstraction s'appuiera sur le fait que tous les chercheurs mentionnés perçoivent le proverbe et la parémie comme 'un tout linguistique qui se tient à sa façon' – et cette référence conceptuelle générique est le seul *tertium comparationis* viable, pour l'instant. Une telle perception abstraite, soit-elle inspirée par des *a priori* théoriques ou par une expérience directe et personnelle, est partagée (nous croyons) par tous les parémiologues et nous suffira (en l'occurrence, cette perception doit plutôt nous suffire). Pour toute étude traitée dans les §§ 3-4, nous adopterons donc le (les) terme(s) que chaque chercheur a utilisé pour faire référence à ce 'tout qui se tient à sa façon', sans critiquer ou réfuter tel ou tel autre choix d'étiquetage. Dans les cas où les chercheurs utiliseraient les deux termes à la fois et de manière explicitement différente, nous le signalerons ouvertement dans notre analyse.

Quant à notre recherche, nous utiliserons *parémie* comme hyperonyme de *proverbe*, *dicton*, *aphorisme*, etc. De même, nous parlerons de *séquence formulaire* pour désigner toute

combinatoire linguistique préfabriquée et conventionnalisée par l'usage. L'affinité avec la *formulaic sequence* de Wray (2002 : 9) est étudiée de notre part, dans la mesure où toutes les parémies (ou, comme synonyme, toutes les *séquences parémiques*) font partie du langage formulaire. Par rapport à la définition de *formulaic sequence* de Wray, nous enlevons le côté cognitif qui sera plutôt à intégrer par la suite (§ Conclusions). Le recours aux étiquettes *séquence formulaire* et *séquence parémique*, d'une part, justifie le choix de *séquence lexico-grammaticale* (SLG) pour appeler notre unité minimale descriptive (§ 1). D'autre part, il s'applique bien au traitement informatique que nous décrirons au § 6.

Pour conclure, nous souhaitons encore préciser que, comme nous traverserons plusieurs langues et cultures, nous donnerons des traductions littérales en français des parémies. Cette démarche qu'on pourrait qualifier de simpliste (voire circonspecte) veut nous mettre à l'abri de tout constat (sans aucun doute) pertinent concernant l'équivalence interlinguistique des parémies. Nous connaissons l'envergure du débat traductologique en la matière (en guise d'exemples, nous citons Conenna 1995, 2011 ; Lavermicocca 2011 ; Quitout & Sevilla Muñoz 2009 ; Zouogbo 2008) et qui est comparable à celle du débat autour de la définition linguistique du proverbe. Il faut donc envisager nos traductions littérales[1] comme des aides à la compréhension des parémies en d'autres langues, non pas comme des suggestions d'équivalents ou de correspondants. Seulement dans les cas (rares) où la traduction correspond à un équivalent formel et sémantique en français (parfois indiquée dans les analyses comparées par les chercheurs mêmes), la traduction n'est pas signalée comme littérale. La visée traductologique mériterait une réflexion à part entière qui dépasse le cadre et les intentions de notre recherche.

[1] Les traductions littérales sont précédées par l'abréviation 'trad. litt.' dans le corps du texte et entre crochets.

CHAPITRE 1

LES PAREMIES.
UNITES DESCRIPTIVES ET CLASSIFICATIONS

Il y a trois grandes vulgates dans la littérature parémiologique :

i. la vulgate des parémiologues ;
ii. la vulgate du 'peuple' par l'intermédiaire explicite des parémiologues ;
iii. la vulgate du 'peuple du Web' (Marcon 2012).

(ii) et (iii) sont fonction de la compétence parémiologique des locuteurs natifs Δ et de leur *intérêt parémiologique* (Marcon 2012 : 131-132), à savoir de leur volonté et plaisir du partage des parémies ainsi que de toute expression conventionnalisée (ou en voie de conventionnalisation) du langage formulaire. Les vulgates (ii) et (iii) ne se penchent pas sur une réflexion métalinguistique sur la parémie. En revanche, la vulgate en (i) a longuement débattu sur la définition de *parémie* (en particulier, de *proverbe*) et sur leur statut linguistique. Autour de cette question fétiche, on peut distinguer :

- la *vulgate ontologique* qui veut la parémie en tant que concept actualisé sous plusieurs formes linguistiques formulaires ;
- la *vulgate logique* où la parémie est abordée avec les outils de la logique classique, notamment de la proposition ;
- la *vulgate catégorielle* pour laquelle la parémie est une catégorie linguistique à part entière, ayant ses propriétés morphosyntaxiques, sémantiques et discursives qui l'assemblent à d'autres catégories linguistiques (par exemple, le nom), d'après les propriétés mises en relief ou partagées avec telle ou telle autre catégorie linguistique ;
- la *vulgate figement* qui inclut les parémies parmi les expressions figées, c'est-à-dire parmi ces manifestations du langage formulaire à combinatoire (lexicale,

métrique, sémantique, syntaxique, etc.) préétablie et à variation (lexicale, métrique, sémantique, syntaxique, etc.) contrainte ;

- la *vulgate défaitiste* qui regroupe tous les parémiologues qui renoncent ou découragent toute tentative de définition.

Ce parcours incomplet veut témoigner la variété des angles d'attaque à l'égard de l'épineuse question : « Qu'est-ce qu'une parémie ? ». Or, nous souhaitons apporter notre contribution non vraiment pour donner une réponse, mais plutôt pour aborder la réflexion théorique d'après deux perspectives qui, à notre connaissance, n'ont pas encore fait l'objet d'approfondissements ponctuels.

Premièrement, nous essayons de mieux cerner la relation entre quelques unités minimales descriptives en linguistique et la parémie. Plus précisément, nous questionnons le statut, entre autres, de l'*énoncé*, de la *phrase* et de la *structure*, qui jouissent d'un consensus (apparent). Ces notions qu'on emploie comme unités descriptives fondamentales font souvent partie de l'outillage des parémiologues qui s'en servent pour saisir la nature linguistique du proverbe et, en général, des parémies. À l'aide de brèves reconstructions historiques et philologiques ainsi que du rappel de certaines critiques vibrantes autour de ces notions essentielles, nous mettrons en relief leurs points forts et leurs points faibles. En même temps, nous considérerons l'utilisation de ces unités de la part des parémiologues. Nous montrerons la cohérence théorique et descriptive de certains d'entre eux à l'égard de chacune de ces unités, mais aussi les enchevêtrements opérés par d'autres. Nous en soulignerons la pertinence et la fiabilité pour encadrer la parémie comme fait linguistique et proposerons, par la suite, notre unité minimale – la *séquence lexico-grammaticale* (§§ 1.1.5., 1.2.6.) – qui dépasse (voire, peut-être, inclut) les unités 'traditionnelles'.

Deuxièmement, nous proposerons une *métaclassification* parémiologique, en faisant nôtre la remarque de Maurice Gross :

> « […] l'approche taxonomique, c'est-à-dire la construction de classifications est une approche possible qui est a priori entièrement indépendante de la nature des concepts descriptifs employés. C'est au niveau de l'élaboration des critères de classement […] que peut éventuellement s'insérer une critique d'adéquation […] une telle critique ne pourrait, par définition, que remettre en question une classification *existante* […] » (Gross 1976 : 25).

Comme le dit Gross, l'activité de classification et sa remise en question relèvent d'un jugement sur les critères adoptés pour trier un ensemble de faits linguistiques et pour aboutir à une classification, non pas de l'activité de classification elle-même. D'une part, notre métaclassification se veut une reprise critique des classifications parémiologiques existantes. D'autre part, elle est une classification (susceptible de critiques successives, évidemment) qui dégage ses critères de tri à partir de ceux qui ont été adoptés par les parémiologues que nous citerons. Cette activité d'analyse et de synthèse en vue d'une classification globale nous servira à mettre en évidence :

- la quantité des étiquettes données aux séquences formulaires ;
- la polysémie que ces étiquettes acquièrent dans l'usage des spécialistes. Suivant leurs approches descriptives, les parémiologues déclenchent une *variation métaterminologique* qui résulte de l'absence d'un consensus sur des concepts et sur des notions qui pourraient représenter les plus petits dénominateurs communs parémiologiques. Ce qui finit par encourager, d'une part, la variation dans l'usage (en littérature parémiologique tout comme dans l'usage courant) et par créer, d'autre part, des mésententes et des débats théorico-méthodologiques peu conséquents au sein de la communauté parémiologique ;
- l'importance de ramener la parémie à l'évidence qu'elle fournit, c'est-à-dire à sa surface lexico-grammaticale, à sa description systématique sur la base d'un échantillon (pour autant que possible) exhaustif et à son usage, avant toutes autres spéculations (nécessaires, certes) dans une perspective sémantique, énonciative, textuelle, etc.

En ce sens, cette métaclassification nous permettra de proposer une classification lexico-grammaticale axée autant sur des critères déjà exploités que sur d'autres, encore ignorés par les parémiologues.

1.1. Les unités linguistiques

D'un côté, le souci du découpage en unités minimales est commun à toute l'histoire de la grammaire, y compris à l'histoire de la grammaire française (Colombat 1988 : 7). De

l'autre, les linguistiques cognitives basées sur l'usage font recours au processus cognitif du 'découpage en gros tronçons linguistiques' qu'est le *chunking*, c'est-à-dire :

> « *the process by which sequences of units that are used together cohere to form more complex units* » (Bybee 2010 : 7).

Dans ce sous-chapitre, il nous a donc paru raisonnable de réfléchir sur où situer la parémie par rapport aux unités minimales linguistiques ainsi que (et surtout) par rapport aux assemblages d'unités minimales qui sont considérés à leur tour comme des unités. En ce sens, nous avons fait référence aux notions évoquées par les parémiologues, notamment :

- la *phrase* et la *proposition* ;
- l'*énoncé*, l'*énonciation* et le *discours* ;
- la *structure* ;
- le *schéma* et la *Gestalt*.

Nous reconstruirons les grandes lignes de leur évolution et de leur usage en tant qu'unités linguistiques pour remettre en question le consensus et le débat qu'elles produisent en tant que fondements de la description et de la définition des parémies.

Dans notre revue, nous laisserons de côté toutes les unités minimales qui relèvent de l'oral : la *clause* et la *période* (Berrendonner 2002) ainsi que la plus récente *unité minimale de base* de nature syntactico-prosodique (Simon & Degand 2011). D'ailleurs, elles ne sont même pas mentionnées dans la littérature que nous avons consultée[2]. Nous ne traiterons pas non plus le *vers*, unité poétique avec laquelle la parémie se rapporte (surtout) en littérature, et le *texte*[3].

[2] Ces unités minimales de l'oral méritent en elles-mêmes une attention toute particulière qui va bien au-delà des propos de notre étude. Leur absence dans la littérature parémiologique est révélatrice, en tout cas, du fait que la plupart des analyses se concentrent sur les parémies à l'écrit, excepté les références au statut d'*énoncé*.
[3] Une référence, notamment à son autonomie textuelle, est en tout cas mentionnée dans les classifications que nous aborderons au § 1.2.

1.1.1. *Phrase / Proposition*

La littérature parémiologique dispose de nombreux exemples de description du proverbe par la *phrase*. Le 'proverbe ineffable', d'après les mots bien connus de Taylor :

> « The definition of a proverb is too difficult to repay the undertaking; […] An incommunicable quality tells us this sentence is proverbial and that one is not. Hence no definition will enable us to identify positively a sentence as proverbial » (Taylor 1962 [1931] : 3, c'est nous qui soulignons)

est souvent ramené à la « phrase insaisissable » (Béguelin 2000, Blanche-Benveniste 2002 : 20), même quand on souhaite éviter toute définition : la référence à la *sentence* par Taylor en est un exemple (Marcon 2011 : 107). L'intrusion de la notion de *phrase* traverse nombre de théories, d'approches et de modèles en sciences du langage. Cela est vrai notamment pour toutes les approches *transphrastiques* et *interphrastiques* (Béguelin 2002 : 86 ; Sarfati 2005 : 13) qui se fondent :

> « sur le postulat que les discours s'analysent en phrases, et que celles-ci sont les unités maximales de la combinatoire syntaxique » (Berrendonner 2002 : 23).

La grammaire s'est emparée de cette notion et :

> « relayée par la linguistique d'inspiration générativiste, s'y réfère comme à l'unité langagière par excellence, domaine de la morphosyntaxe et champ de l'analyse en constituants » (Béguelin 2002 : 85-86).

Pour une confirmation de ce phrase-centrisme de la syntaxe, il suffit de lire le sixième chapitre de la *Grammaire méthodique du français* de Riegel, Pellat et Rioul dans une réédition récente (2009) où l'approche générativiste est adoptée comme cadre pour toute l'analyse syntaxique proposée.

Au contraire, par une invitation à considérer plus qu'une seule syntaxe pour décrire les faits linguistiques (Blanche-Benveniste 2002 : 20), les travaux sur les discours oraux jouent souvent le *requiem* pour la notion de *phrase* :

> « Exit, donc, la phrase, avec remerciements pour les services rendus » (Berrendonner 2002 : 27),

surtout dans les cas où elle est retenue comme « bonne unité de calcul en grammaire » (Blanche-Benveniste 2002 : 8). Ils lui reconnaissent plutôt le statut d'« unité du savoir pratique » (Béguelin 2002 : 88) dont l'apprentissage et l'intériorisation sont la conséquence d'une insistance scolaire qui satisfait des exigences orthographiques et normatives (Béguelin 2000 : 49, 2002 : 88 ; Berrendonner 2002 : 27 ; Blanche-Benveniste 2002 : 7).

Le succès et la marque d'évidence en soi dont la phrase jouit (Béguelin 2000 : 49, 2002 : 85) sont à repérer dans le flou de sa définition et de son exploitation. Pour synthétiser les revues critiques proposées par Béguelin (2000) et par Berrendonner (2002), les critères d'autonomie (ou maximalité) syntaxique, d'unité logico-sémantique, de délimitation typographique (majuscule initiale et point final) et prosodique de début et fin ainsi que d'actualisation accordent à la phrase le caractère de *prototype* :

> « c'est-à-dire un assortiment de propriétés généralement vraies en même temps, mais dont aucune n'est strictement nécessaire, généralement caractéristiques, mais dont aucune n'est strictement équivalente. Cela lui confère de la *plasticité* et la rend apte à subsumer aussi bien les individus exemplaires qui vérifient tous ces critères […] que des spécimens marginaux qui satisfont l'un d'entre eux » (Berrendonner 2002 : 27, c'est nous qui soulignons).

Kleiber (2003) revient sur cet *adieu* à la *phrase-prototype* et sur les propriétés des unités de *clause* et de *période* que Berrendonner suggère comme unités de segmentation remplaçantes. Kleiber invite à mieux reconsidérer la phrase justement à la lumière de la sémantique du prototype. On pourrait ainsi concevoir la *phrase-prototype* centrale comme la résultante des propriétés typiques communes aux membres qui relèvent de la catégorie *phrase* (2003 : 18). Ce qui comporte une observation, d'une part, de tous les membres de la catégorie que les sujets parlants reconnaissent comme phrases (*ibid*.). D'autre part, l'identification du centre prototypique de la catégorie *phrase* doit passer par une réflexion soignée autour des propriétés typiques qui la caractérisent (*ibid*.). D'où la nécessité d'enrichir les critères approximatifs d'autonomie syntaxique, d'unité logico-sémantique, de délimitation typographique et prosodique ainsi que d'actualisation. Il faut dégager d'autres critères ou propriétés à partir de ces phrases que les sujets parlants découpent et signalent comme tels.

Pour Kleiber, un raisonnement pareil doit être valable non seulement pour la définition de l'unité *phrase*, mais aussi pour la définition de toute unité linguistique de segmentation (2003 : 22), y inclus la *clause* et la *période*, dont il met en évidence quelques limites[4].

Une autre unité est le pendant privilégié, voire l'unité constituante, de la phrase : c'est la *proposition*.

> « *Proposition* désigne en logique classique une suite de mots qui permet l'expression d'un jugement. La vieille *analyse logique* (noter l'épithète) faisait de la phrase simple une « proposition indépendante » et débitait la phrase composée en une « proposition principale » et une ou plusieurs « propositions subordonnées » » (Wilmet 2007 : 473).

La phrase, notamment la phrase canonique générativiste (ou *phrase simple*) du type :

> *groupe nominal* (GN) + *groupe verbal* (GV)

est identifiée avec le couple *sujet-prédicat* de la proposition logique : la phrase-proposition narre ainsi des états de certains *realia* (Béguelin 2000 : 55). En outre, la proposition sous-jacente sert à classer les phrases. Ce qui alimente cette « collusion entre […] syntaxe et logique » qui remonte à Port-Royal et à Beauzée au XVIII^{e} siècle (Marchello-Nizia 1979 : 47, Béguelin 2000 : 52). Cette collusion se perpétue encore aujourd'hui, comme le montre l'introduction au manuel de Gardes-Tamine (1990 : 11) qui est explicitement présenté comme « grammaire de la phrase » (1990 : 7)[5]. Sur l'identification entre *phrase* et *proposition* (et *période* en rhétorique, aussi), Le Goffic a peu de doutes :

> « L'histoire peut aider : le terme de *phrase* a dû sa fortune à ce qu'il réalise un compromis entre la « proposition » de la logique et la « période » de la rhétorique : de la « proposition » des logiciens, il retient l'idée d'une articulation centrale entre un sujet et un prédicat, cimentée par un acte de l'énonciateur (tel que, prototypiquement, une assertion) ; de la « période » de la rhétorique, le terme retient l'extension au-delà d'une proposition unique, la possibilité d'une

[4] On retrouvera cette approche prototypique, à savoir l'attention aux propriétés typiques observées pour la démarcation d'une catégorie/unité, dans la notion de *Gestalt* proposée par Lakoff (§ 1.1.4.) ainsi que dans notre *séquence lexico-grammaticale* (§ 1.1.5.). Dans les travaux sur les proverbes par Kleiber, c'est d'ailleurs l'analyse minutieuse de leurs propriétés typiques qui l'amènera à la proposition d'une catégorie sémantique-proverbes à part entière (cf. *infra*).

[5] Depuis, Gardes-Tamine a publié un autre manuel consacré à l'ordre des mots (2013). Il suffit de lire le sommaire pour se rendre compte du fait que cette collusion persiste. Le titre du chapitre 3 est en effet « La phrase minimale », mais le titre du premier paragraphe le dément et montre plutôt l'identification avec la proposition : « La proposition minimale : définition ».

certaine multiplicité, pour autant que celle-ci puisse se résorber dans l'unité d'une structure matrice. » (Le Goffic 2005 : 56).

À vrai dire, on soupçonne l'entrelacement de la *phrase* avec l'*énoncé* (§ 1.1.2.) et la *structure* (§1.1.3.), qu'il avoue, d'ailleurs :

« La phrase est donc, inséparablement, une réalité à la fois syntaxique (prédicative) et énonciative » (Le Goffic 2005 : 57).

Ce qui permet de satisfaire les chomskyiens, d'une part, et les saussuriens, de l'autre (*ibid*.).

Nous ne savons où l'on peut situer le proverbe et, en général, les parémies par rapport à la notion protéiforme ou par rapport au centre prototypique de la *phrase*, malgré son exploitation très courante en parémiologie. En ce sens, il suffit de jeter un coup d'œil aux classifications parémiologiques qui reposent sur la notion de *phrase* (§ 1.2.). La connivence logico-syntaxique entre phrase et proposition n'a fait que produire un élargissement du champ d'action des parémiologues. Nous donnons quelques exemples ci-dessous :

Le proverbe est une...
construction phrastique (Anscombre 2008a : 254)
forme sentencieuse (Anscombre 2003 : 160 ; Anscombre 2008a : 256 ; Anscombre 2008b : 18 ; Gouvard 1996 : 48)
phrase autonome (Anscombre 2008a : 254)
phrase figée (Conenna 1988, 1998a)[6]
phrase générique (à contenu implicatif) (Kleiber 2000 :49-50)
phrase générique (Kleiber 1989, 2000 : 41)
phrase générique typifiante a priori (Anscombre 2000 : 10)
phrase parémique (Anscombre 2008b ; 2013 : 102)[7]
phrase sentencieuse (Anscombre 2008a : 254-255)
signe-phrase (Kleiber 1989, 2000)

Tableau 1. Quelques étiquettes appliquées aux proverbes par le recours à la notion de *phrase* en parémiologie francophone.

6 Dans ce cas de figure, la notion de *phrase* ne correspond pas exactement aux nuances présentées dans ce paragraphe (§ 1.2.5.). Malgré cela, le proverbe est également conçu comme phrase.

7 Anscombre avoue qu'il préfère parler de *formes sentencieuses*, « le mot *phrase* étant suspect à plus d'un titre » (2008b : 18). Pourtant, il continue à envisager son argumentation autour de la notion de *phrase*. Le changement d'étiquette n'affecte pas vraiment le fond.

Au Tableau 1, on peut ajouter d'autres étiquettes collées aux parémies où les propriétés des verbes en déterminent leur classification phrastique (§ 1.2.5.). D'ailleurs, cette extension des propriétés qui relèvent de la rection verbale à la phrase est, pour Blanche-Benveniste, la conséquence d'un usage impropre qui montre l'absence de caractéristiques syntaxiques distinctives de la phrase en soi (Blanche-Benveniste 2002 : 20).

Ce foisonnement métaterminologique autour de la notion de phrase qui subsume, à son tour, la logique propositionnelle, est remis en question par Wilmet. Son approche critique et polémique à la grammaire lui permet de « se passer sans douleur » de la proposition comme unité descriptive fiable, et ce, à l'aide d'un proverbe. Par le recours à :

Qui vivra verra

Wilmet observe que celle qui serait la proposition principale en analyse logique (*verra*) se réduit à une partie grammaticalement discutable (2007 : 474, 496). Il redéfinit ainsi la phrase comme une somme de l'*énonciation* et de l'*énoncé* (2007 : 478) (§ 1.1.2.). Il distingue, d'une part, la *phrase matrice*, à savoir une partie 'suspendue' d'une *phrase unique complexe* et, d'autre part, la *sous-phrase*, à savoir la partie d'une phrase unique complexe qui fait qu'une phrase matrice s'achève en un tout. En l'occurrence, Δ *verra* est la phrase matrice, Δ étant le symbole qui indique l'enchâssement attendu par la sous-phrase *Qui vivra*. Malgré sa redéfinition critique où deux autres unités minimales (l'énoncé et l'énonciation) sont à la base de son raisonnement phrastique, le proverbe reste encore, pour Wilmet aussi, une phrase – qu'il définit, entre autres, comme séquence de mots (voir ci-dessous) – au statut énonciatif (§ 1.1.2.)[8].

Il faut souligner que, malgré l'orientation phrastico-propositionnelle, les étiquettes proposées par Kleiber pour décrire le proverbe sont les plus fidèles à l'étymologie et à l'histoire de la *phrase*. La « trinité dénomination-phrase-généricité » (Kleiber 2000 : 42), le nom-*name* (Kleiber 1989 : 233) qu'est le proverbe et qui débouche vers l'autonomisation d'une catégorie sémantique-proverbe en soi avec ses propriétés (proto)typiques (Kleiber 2010a), renvoient à une dimension lexicale[9]. Ce qui le rapproche de la notion de *phrase* telle

[8] Wilmet n'est pas le seul grammairien à faire recours aux proverbes pour mettre en relief des aspects grammaticaux. En continuité avec la tradition médiévale (Schulze-Busacker 2012), les proverbes et, en général, les parémies servent encore d'exemples pour montrer tel usage ou telle norme (par exemple, voir les chapitres consacrés aux références des groupes nominaux ou à la juxtaposition dans Riegel *et al*. (2009)).

[9] En passant, à ce propos, nous citons ce que Cram a affirmé sur la manière d'aborder le proverbe : « The proverb should be viewed as a lexical element with a quotational status. It is a lexical element in the sense that is a syntactic string […] reused as a single unit with frozen internal structure. Its quotational status derives from

qu'elle a paru dans l'histoire de la grammaire. Comme le rappellent Marchello-Nizia (1979 : 46) et Béguelin (2000 : 50-51), *phrasis* n'apparaît pas pour rendre compte de la syntaxe, mais du lexique. Son adoption en français (*frase*, *fraze*) s'applique pour indiquer toute tournure, locution et expression polylexicale. Comme le précise Béguelin, *phrasis* fonctionne en tant que catégorie lexicale (Béguelin 2000 : 51), et ce, jusqu'à la fin du XVII^{e} siècle[10]. C'est en ce sens qu'il faut donc envisager la *phras*éologie, à savoir comme une émanation disciplinaire (revue, augmentée, etc.) de toutes ces études sur *phrasis*.

Il est donc raisonnable de se demander si l'on peut revenir aux sources de *phrasis* pour décrire les parémies. On pourrait aller au-delà de la notion de *phrase* grammaticalisée et logique, voire scolaire, et garder la seule notion de *phrase graphique* comme une unité heuristique empirique pour toute exploitation typographique et orthographique, comme il en était à ses débuts (Béguelin 2002 : 88) et comme il en est maintenant pour le traitement automatique des langues (TAL). Autrement dit, il est raisonnable de se questionner sur le fait que les parémies en tant que *phrasēs* sont des combinatoires lexicales (outre que syntaxiques) et qu'elles ne sont pas non plus 'très très spéciales' dans la mesure où elles partagent des comportements d'agencement similaires aux autres unités phraséologiques. Au même titre, l'application de parcours méthodologiques communs à l'étude des unités phraséologiques deviendrait faisable (et souhaitable) pour parvenir à rendre compte de cette catégorie lexico-syntaxique qu'est *phrasis*-parémie comme « un grand mot en plusieurs tronçons » (Béguelin 2000 : 51).

1.1.2. *Énoncé* / *Énonciation* / *Discours*

Unité assez répandue en parémiologie, l'*énoncé* et l'*énonciation* se proposent comme deux piliers, à partir de Benveniste, pour introduire sur la scène linguistique le sujet qui parle (Sarfati 2005 : 11, 18). À la place de l'objectivité syntaxique, ces deux notions servent à prendre en considération la *subjectivité*, à savoir :

---

the fact that proverbial expressions are typically 'invoked' or 'cited' rather than straightforwardly asserted » (Cram 1983 : 54). Quoique l'approche diverge de celle de Kleiber, Cram aussi revient à la dimension lexicale et y ajoute la dimension situationnelle, plus proche de la vision du proverbe-énoncé (§ 1.1.2.).

[10] Le passage de groupement lexical à unité descriptive en syntaxe se concrétise par l'œuvre de deux pères Jésuites : en 1668, dans *Essay d'une parfaite grammaire de la langue française* rédigée par le Père Chiflet (Marchello-Nizia 1979 : 46) et en 1709 dans la *Grammaire françoyse sur un plan nouveau* par le Père Buffier (Marchello-Nizia, *ibid*. ; Béguelin 2000 : 51-52).

« l'unité psychique qui transcende la totalité des expériences vécues qu'elle assemble et qui assure la permanence de la conscience » (Sarfati 2005 : 18).

L'énoncé et l'énonciation se caractérisent parfois (et plus rarement) comme constituants de la phrase, parfois (et le plus souvent) comme cadres d'actualisation de la phrase grammaticale.

Pour le premier cas, nous rappelons la définition de Wilmet pour qui la *phrase* se configure par la rencontre d'une *énonciation* et d'un *énoncé*. Par quelques ajustements personnels de la notion d'*énonciation* telle qu'elle est proposée par Kerbrat-Orecchioni. Wilmet la définit comme la rencontre d'un *qui*, d'un *quand* et d'un *comment* (Wilmet 2007 : 479), c'est-à-dire, respectivement :

- d'une personne énonciative ou source personnelle qui permet de distinguer les formes de discours (direct, indirect, libre, rapporté) ;
- d'un repère énonciatif temporel ;
- d'une modalité (assertive, injonctive, etc.) (Wilmet 2007 : 480 et ss.).

L'*énoncé*, en revanche, est à considérer comme « le contenu de la phrase » (Wilmet 2007 : 494) qui résulte de l'interaction entre thème (sujet grammatical) et rhème et par la création d'une *prédication* entre eux (Wilmet 2007 : 495).

Pour le cas où l'énoncé et l'énonciation agissent comme cadres d'actualisation des phrases grammaticales, nous pouvons mentionner encore l'exemple du manuel de Gardes-Tamine qui décrit l'énoncé comme :

« un événement de parole concret et individuel […] un discours (ou […] partie de discours), tenu par une personne, qui est précédé et suivi d'un silence, et qui n'est pas descriptible ou organisé » (1990 : 8)

et encore, en ligne avec ce que nous avons mentionné de Benveniste :

« un énoncé est un **fragment de vécu**, enraciné dans une situation particulière. **Ce n'est pas une unité de langue**, abstraite, mais **de parole**, concrète » (*ibid.*).

L'énoncé reste en dehors de la syntaxe, sauf dans la mesure où un découpage syntaxique en permette une analyse par phrases grammaticales (*ibid*.). Ces phrases-segmentations objectives d'une unité subjective et concrète doivent également tenir compte de l'énonciation :

> « un acte de parole pris en charge par un sujet parlant […] dans une situation précise » (*ibid*.).

L'énonciation est l'incubateur de l'énoncé et lui fournit un ancrage pragmatique référentiel extralinguistique (1990 : 9). Autrement dit, énoncé et énonciation sont aux linguistes du discours (Benveniste, Culioli, Ducrot et Mainguenau, pour ne citer que les plus connus) ce que la phrase est à la syntaxe est aux grammairiens[11].

De la phrase grammaticalement normalisée et normée, on passe à la subjectivité ainsi qu'à la pléthore des approches pour analyser **le(s)** *discours*. Cette notion de discours ne fait pas non plus consensus (pour des revues sur la notion de *discours* : De Gioia & Marcon (à paraître), Sarfati 2005 : 9-15)[12]. C'est dans cette pléthore (parfois désabusée) qu'il faut encadrer tout étiquetage énonciatif et discursif du proverbe et des parémies. En gros, le flou définitionnel des notions de *discours*, d'*énoncé* et d'*énonciation* offre le même degré de plasticité que Berrendonner a imputé à la notion de *phrase* (§ 1.1.1.).

La souplesse de ces notions permet ainsi aux parémiologues de saisir les nombreuses facettes des parémies d'un coup. Ce qui peut entraîner un mélange et une réunion notionnels entre plusieurs niveaux d'analyse, notamment entre syntaxe, sémantique et analyse du discours, comme le montrent, d'ailleurs, les définitions de Gardes-Tamine ci-dessus. La rencontre de plusieurs niveaux d'analyse comporte souvent que certains parémiologues réajustent leur tir pour établir un cadre descriptif assorti. C'est le cas exemplaire d'Anscombre qui définit le concept [*proverbe*] par le biais de propriétés qui présument le recours aux notions de *discours*, à celle de personne énonciative collective *ON-locuteur*[13] :

> « Sera [proverbe] toute entité linguistique possédant certaines propriétés définitoires: […] a) […] des discours ON-sentencieux […] b) […] des discours autonomes, clos et minimaux […] » (Anscombre 2000 : 9, 13)

[11] Il est vrai aussi que la phrase acquiert une allure pragmatique, par exemple, dans la définition du *Bon Usage* de Grevisse où elle devient l'unité minimale de communication (Wilmet 2007 : 472).
[12] C'est pour cette raison que nous avons choisi de ne mentionner qu'à titre d'exemple les seules définitions de Gardes-Tamine. Ses définitions se rapprochent davantage de la lecture de Benveniste, père da la linguistique de l'énonciation moderne en milieu francophone.
[13] Il parle aussi de *phrase ON-sentencieuse* (Anscombre 2000, 2003), faisant ce lien entre syntaxe et discours tel qu'on le repère dans les définitions de Gardes-Tamine.

ainsi qu'à celle de *marqueurs médiatifs* (Anscombre 2006, 2011a) tel *comme le dit le proverbe* pour insérer et indiquer le proverbe-discours dans le discours tout-venant. Nous repérons également l'étiquette d'*énoncés sentencieux* comme hyperonyme de *parémies* (Anscombre 2006 ; Quitout 2002 ; Sevilla Muñoz 1993). L'étiquette d'*énoncé parémique* chez Anscombre est aussi employée sans aucune distinction avec celle de *phrase parémique* (Anscombre 2013). Ce qui témoigne de la coexistence des unités de *phrase* et d'*énoncé* dans l'argumentation du parémiologue.

Dans la lignée de la théorie de l'argumentation de la langue et de la théorie des stéréotypes (les deux sorties des recherches d'Anscombre), Gómez-Jordana Ferary analyse les *situations d'énonciation proverbiale*, à savoir les *occurrences proverbiales* dans son corpus textuel (Gómez-Jordana Ferary 2012 : 307-355).

Plus axé sur la prédication énonciative et concentré autant sur la forme et sur la *structure* (§ 1.1.3.) que sur la phrase et sur l'énoncé, il faut citer les travaux sur l'articulation sémantique et sur l'inférence des *énoncés proverbiaux* que Merji (2001, 2008) présente comme une sous-catégorie des *énoncés sentencieux* (Mejri 2008 : 169) ou « énoncés de la famille du proverbe » (Arnaud 1991 : 6).

Les linguistes de l'énonciation et du discours se sont eux-mêmes proposés de saisir le proverbe et les parémies. Pour Crépeau, le proverbe est un *énoncé de structure analogique* (Crépeau 1975 : 295), le processus cognitif de l'analogie faisant la rencontre à la fois de l'analyse du discours et de l'énonciation ainsi que de la notion de *structure*. Pour Mainguenau, le proverbe est un *énoncé non embrayé*, c'est-à-dire que l'introduction en discours se réalise sans que le locuteur-acteur de l'énonciation laisse des marques subjectives qui le relient à l'énoncé-proverbe (Paveau & Sarfati 2003 : 174-175). Plus récemment, Mainguenau a encore élaboré sur l'énonciation des proverbes et de tous les *énoncés sentencieux* (Mainguenau 2012 : 22, 27, 52) en termes d'*énonciation aphorisante* (*ivi*, 19-28), respectivement d'*aphorisation proverbiale* et d'*aphorisation sentencieuse* (*ivi*, 61-62). Nous synthétisons les aspects saillants de cette aphorisation par les mots de Mainguenau : comme d'autres « phrases sans texte » (*sic*, titre de son ouvrage), le proverbe est un énoncé détaché, « un fragment clos sur soi, facilement mémorisable […] qui renvoie l'image d'un monde stabilisé » (*ivi*, 62) et, en tant que « classe particulière des énoncés généralisants » (*ivi*, 61), « implique une sorte d''autorepérage' » (*ibid*.) en discours. En d'autres mots, l'aphorisation proverbiale s'empare de la notion de *phrase* et de la *proposition* logique sous-jacente (la généricité est, par exemple, une propriété observée sur le plan logico-phrastique) et contribue

(encore) à estomper les limites entre ces notions, déjà suffisamment floues de par elles-mêmes.

1.1.3. *Structure*[14]

La notion de *structure* est décidément celle qui caractérise de manière transversale les études parémiologiques et parémiographiques. Faudrait-il en conclure que le paradigme du structuralisme s'est emparé à jamais du proverbe et des parémies ? Ou y aurait-il une autre manière de percevoir la notion de *structure* qui a peu à voir avec les structuralismes en sciences du langage ?

Pour commencer, nous reprenons dans le Tableau 2 ci-dessous quelques-unes des qualifications de la notion de *structure* que l'on peut repérer dans la littérature francophone :

[14] Ce paragraphe est une réélaboration augmentée et détaillée d'une réflexion à peine ébauchée dans Marcon (2011 : 107-110).

***Structure* en parémiologie**
***structure actorielle – structure bi-membre – structure binaire – structure bipartite – structure conceptuelle – structure concessive – structure corrélative – structure de surface – structure énonciative – structure événementielle – structure formelle – structure hiérarchisée – structure immanente – structure implicative – structure inférentielle – structure lexico-grammaticale*[15] *– structure mélodique – structure métrique – structure monopartite – structure morphologique – structure morphosyntaxique – (macro-)structure narrative – structure parémique (prototypique) – structure phrastique – structure plurimembre – structure poétique – structure prédicative – structure prégnante – structure prosodique – structure prototypique – structure proverbiale – structure proverbioïde – structure proverb-like – structure quadripartite – structure rimique – structure rythmique – structure sémantique – structure sous-jacente – structure superficielle – structure syllabique – structure syntaxique – structure textuelle – structure tripartite – structure* **(suivie d'une proposition logique ou d'une séquence syntaxique et/ou lexicale)**

Tableau 2. Identités de *structure* dans la littérature parémiologique francophone au cours du XX^e^ siècle.

À partir de la deuxième moitié du XX^e^ siècle, la parémiologie francophone opère une 'déstructuration identitaire' de la notion de *structure*. Au premier abord, le domaine de la parémiologie éloigne l'unité épistémologique descriptive *structure* du paradigme qui porte son nom. Chaque parémiologue suggère à tour de rôle une nouvelle *identité de structure* qui réfère successivement à tel ou à tel autre aspect linguistique qu'on souhaite mettre en relief et (faire) reconnaître des proverbes et, en général, des parémies.

Toutefois, s'il est vrai que la vague du structuralisme a marqué l'évolution de la linguistique depuis la révolution saussurienne[16] et que la *structure* s'est imposée en raison de

[15] La *structure lexico-grammaticale* est également à repérer dans les travaux sur corpus de Sinclair. Pour mieux mettre en relief l'*idiom principle* (Sinclair 1991 : 109-110) en tant que 'principe de phraséologisation' de la langue, Sinclair fait recours à la notion de *structure* : « […] *to indicate a lexical item and its patterns and collocations (i.e.) „any privileges of occurrence of morphemes"* […] *lexical* […] *or grammatical* » (Sinclair 1991 : 104).

[16] Wilmet se limite à constater que la notion de *structure* et celle de *système* ouvrent la porte à la linguistique moderne et à son autonomisation méthodologique, sans s'attarder sur une définition de ces deux « concepts opératoires » (2007 : 22).

cet éclat paradigmatique (notamment, sur l'escorte du génértivisme chomskyien[17]), un parcours diachronique révèle que les racines de cette notion sont à repérer dans le passé.
Déjà depuis le XVI[e] siècle, *structure* est mentionné en grammaire. En 1531, dans sa grammaire *In linguam gallicam Isagôge*, Jacques Sylvius l'utilise pour étiqueter des cadres formels équivalents aux plans intralinguistique et interlinguistique (latin-français). De cette manière, il peut décrire les déclinaisons et la dérivation (Chevalier 2006 : 105-106). Chevalier précise :

> « Sylvius reste fidèle à la méthode des grammairiens latins qui consiste à définir les fonctions par la structure : c'est en différenciant les formes qu'on indique leur rôle par rapport au reste de la phrase » (2006 : 110).

C'est une 'méthode par la structure' qui hérite de la tradition médiévale et, encore plus loin, de l'Antiquité classique (2006 : 175) et qui restera bien vivante dans le paradigme linguistique du structuralisme contemporain. De la forme à la fonction par une description au moyen d'une structure, et ce, dans le cadre de la phrase (qui revient encore comme cadre idéal de l'analyse grammaticale) : on repère les germes de la grammaire générative de Chomsky ou de la méthode distributionnelle et transformationnelle du Lexique-Grammaire (Gross 1975).

Dans sa grammaire du latin *De causis linguae latina* qui date de 1540, Jules César Scaliger maintient la notion de *structura* qui indique tous les groupements de mots. Il oppose à celle-ci la notion de *dictio*, à savoir le mot isolé (Chevalier 2006 : 198). En termes de 'groupements', *structura* se rapproche de *phrasis* (§ 1.1.1.). Ce qui est confirmé par l'apparition des premières grammaires entièrement consacrées à la langue française. Jean Pillot garde la notion de *structure* et, dans sa *Gallicae linguae Institutio* de 1550, commence à appliquer l'analyse par groupements (au plus de deux éléments) dans le cadre de la phrase (2006 : 227). La solidarité entre les deux notions de *phrase* et de *structure* est ainsi confirmée par la grammaire. Le XVII[e] siècle et la *Grammaire générale de Port-Royal* consacrent les fonctions de chaque unité linguistique : c'est la réflexion sur les parties du discours qui l'emporte (2006 : 488). L'atomicité du signe linguistique – fondement de l'analyse syntaxique distributionnelle moderne basée sur corpus (§ 2.2.) – prend le dessus par rapport à

[17] Par exemple, Béguelin utilise la notion de *structure* pour faire référence aux *structures syntaxiques profondes* et *superficielles* de la grammaire générative chomskyienne (Béguelin 2000 : 89-90). De temps en temps, elle parle de *structure* pour argumenter autour de la *phrase*.

la vision ensembliste proposée par la *structure*-groupement linguistique. Ce n'est qu'au XVIII[e] siècle, quand l'abbé Noël-Antoine Pluche publie *La Mécanique des Langues, et l'Art de l'enseigner* en 1751, en plein esprit des Lumières et de l'*Encyclopédie*, que la grammaire se fait par la structure (2006 : 666). D'après la reconstruction historique de Jean-Claude Chevalier que nous avons suivie, c'est à Beauzée et à sa *Grammaire générale* de 1767 qu'on doit l'entrée officielle de la notion de *structure* dans l'outillage descriptif de la langue française :

> « ce mot *structure* n'est-il pas rigoureusement relatif au mécanisme des langues, et ne signifie-t-il pas la disposition artificielle des mots, autorisée dans chaque langue pour atteindre le but qu'on s'y propose, qui est l'énonciation de la pensée? » (2006 : 666 repris par Diderot & D'Alambert, vol. XVIII, p.1017)

côte à côte de la notion de *phrase*. De cette définition de Beauzée, nous dégageons une réflexion en ce qui concerne la nature même de la structure en tant qu'unité invoquée par les linguistes. Cette « disposition artificielle des mots » qu'est la structure sert à articuler les langues et à ordonner la pensée humaine. Elle existe au niveau cognitif, quoiqu'elle soit reconnue et saisie au moment des réalisations concrètes de la pensée humaine subjective : l'énonciation. Il faut donc considérer la formalisation et l'actualisation verbale de la pensée pour saisir la *structure*. Autrement dit, la *structure* existe quand il y a usage, quand il y a une surface à analyser qui se concrétise après la pensée. Par conséquent, comme le remarque Eco, la structure existe soit en fonction de ce qui est soit de ce qui n'est pas encore :

> « *È Struttura quella che non c'è ancora*. Se c'è, se l'ho individuata, ho tra le mani solo un momento mediano della catena che mi garantisce, al di sotto di questa, una struttura più elementare e onniesplicativa » (Eco 2008 [1968] : 323)[18].

Pour revenir aux identités attribuées à la notion de *structure* du Tableau 2, il est question de distinguer deux approches. D'une part, les parémiologues emploient la notion de *structure* pour décrire des relations et des corrélations dans l'ensemble des parémies. D'autre part, les parémiologues se dirigent vers *la* parémie, à savoir vers la structure-source qui fait la parémie. Autrement dit, soit les parémiologues procèdent par une méthode qui opère *par la*

[18] « *Est Structure ce qui n'est pas encore*. Si elle est là, si je l'ai cernée, j'ai sous la main seulement un moment médian de la chaîne qui me promet, au-dessous d'elle, une structure plus élémentaire et omni-explicative » (c'est nous qui traduisons).

structure comme un outil descriptif d'analyse, soit les parémiologues donnent un *sens absolu (ontologique) à la structure*. Il est raisonnable de croire que la totalité des parémiologues qui font référence à la *structure* se situe du côté méthodologique, comme le montrent la plupart des étiquettes dans le Tableau 2. L'identité de structure reconnue est ainsi un outil qui sert à mettre en évidence, d'après une perspective linguistique donnée :

> « un tutto formato di elementi solidali, tale che ciascuno dipenda dagli altri e non possa essere quello che è se non in virtù della sua relazione con gli altri » (Eco 2008 [1968] : 254)[19].

Au même titre, il est aussi raisonnable de croire que certains d'entre eux se laissent tenter par le fait de rendre absolue une identité de structure-outil opérationnel. Autrement dit, certains parémiologues essaient de pousser aux extrêmes une identité de structure pour qu'elle rende compte de la totalité proverbiale et parémique. Ce qui est manifesté (explicitement) par les étiquettes *structure proverbiale* et *structure parémique* ainsi que par *structure proverb-like*, *structure proverbioïde*, *structure immanente* et *structure sous-jacente*[20].

De ce bref parcours, il résulte que, comme pour les notions de *phrase*, *énoncé* et *énonciation*, celle de *structure* sert à autonomiser des groupements d'éléments perçus comme solidaires entre eux, et ce, d'après la perspective linguistique sélectionnée, voire privilégiée. Bien au-delà de toute approche structuraliste (*stricto sensu*), c'est la variation de cette perspective qui fait de la *structure* une unité prototype encore plus plastique (et moins saisissable) que la *phrase* et l'*énoncé*.

1.1.4. *Gestalt* / *Schéma*

Les notions de *Gestalt* et de *schéma* (qui résument aussi d'autres étiquettes que nous avons reconduites à ces notions) nous servent pour souligner la corrélation que les parémiologues établissent entre plusieurs perspectives linguistiques adoptées pour aborder les proverbes et, en général, les parémies. Plus précisément, *Gestalt* et *schéma* servent à uniformiser la perception et/ou l'observation de régularités et de répétitions qui ont trait aux parémies et qui les décrivent cognitivement ou formellement.

[19] « un tout formé d'éléments solidaires, de façon que chacun dépende des autres et ne puisse être ce qu'il est qu'en raison de sa relation avec les autres » (c'est nous qui traduisons).
[20] En milieu russophone, Permyakov rédige ses *Notes on Structural Paremiology* (Permyakov 1979) où la *structure* représente un outil descriptif appliqué à tous les niveaux de la description linguistique du proverbe (et, en général, des parémies). Dans ce cas, la notion de *structure* est pleinement insérée en linguistique structurale et sert à dévoiler et ordonner la nature à la source du proverbe.

Le moule ethnolinguistique (Paulhan)

Dans la littérature parémiologique francophone, une ébauche des notions de *Gestalt* et de *schéma* est à repérer au début du XX^{e} siècle. Paulhan (1993 [1925]) reconnaît au proverbe un caractère autoreproducteur de pensées, d'images et de combinatoires verbales contraintes sur les axes paradigmatique et syntagmatique. *L'Expérience* (ethnolinguistique) *du proverbe* (ou, comme il le dit, du « langage proverbial » (Paulhan 1993 [1925] : 13) à Madagascar en milieu malgache, introduit des observations que l'on peut considérer proto-générativistes, proto-transformationnelles et en faveur des études à venir sur le (omni-, quasi-, semi-) figement proverbial. Cela est vrai en particulier quand Paulhan reconnaît que le proverbe possède la propriété de devenir moule, et ce, pour fabriquer en série des objets de sa même famille :

> « le proverbe est à la fois moins et plus qu'un raisonnement ou une métaphore : il est l'un et l'autre à l'état figé […] Tout proverbe ainsi pouvait devenir un moule, un poncif susceptible de me donner, à quelques retouches près, des centaines de reproductions » (1993 [1925] : 34-35).

Paulhan évoque le processus de l'*analogie* que Legallois & François considèrent « comme processus productif fondamental » autant pour des usages nouveaux que pour des montages en série d'expressions idiomatiques :

> « les expressions idiomatiques servent souvent de patrons à des emplois nouveaux, ou alors s'inscrivent elles-mêmes dans des séries – des familles de constructions » (Legallois & François 2011 : 16).

D'autres réflexions de Paulhan, en revanche, sont plutôt proto-logiques, proto-contextualistes ainsi qu'en faveur des études sémantiques dynamiques autour du proverbe, comme le montre la citation suivante :

> « Il arrivait par la suite que le cadre abstrait, l'armature commune à toute une famille de proverbes se présentât d'abord à mon esprit : ce cadre ensuite se garnissait de mots » (1993 [1925] : 36).

Toutes ces remarques sur terrain témoignent, certes, la fascination de Paulhan à l'égard du « langage proverbial » perçu comme un code en soi (§ Greimas 1.2.5.). Surtout, ses considérations empiriques indiquent la complexité de l'interaction qui s'établit entre les éléments (abstraits, comme le « raisonnement » et la « métaphore », ainsi que concrets, comme les « mots ») qui composent les proverbes et, par conséquent, la complexité de l'interaction entre les niveaux qui se croisent et dialoguent dans et pour cette interaction. L'« armature commune à toute une famille de proverbes » dont il parle sera racontée à la fois formellement et logico-sémantiquement par d'autres parémiologues.

Gestalt linguistique (Lakoff)

Avant de détailler les quelques tentatives des parémiologues pour saisir l'« armature commune » des proverbes, nous souhaitons mieux éclairer ce qu'on entend par *Gestalt*, et ce, par un renvoi à l'étude de Lakoff (1977). Lakoff emprunte la notion de *Gestalt* à la psychologie et la personnalise pour élaborer une théorie des *Gestalts linguistiques* sur base empirique. Lakoff situe son analyse au sein de ce qu'il appelle *linguistique expérientielle*, c'est-à-dire la linguistique qui encadre les réalisations verbales dans la totalité de la vie physique, psychique et sociale de l'homme (1977 : 237). Ce qui heurte avec la vision réductrice du générativisme chomskyien que Lakoff veut ainsi contrecarrer. Il définit la *Gestalt* comme une *structure* qui organise la pensée, la perception et toutes les activités humaines (1977 : 246)[21]. Brièvement :

- une Gestalt est un tout analysable de façon holistique ou analysable en ses parties ;
- les propriétés reconnues pour une Gestalt comme tout peuvent être acquises par ses parties ;
- l'analyse des parties peut se faire par plusieurs points de vue, ce qui implique plusieurs analyses viables et fiables pour une seule Gestalt ;
- une Gestalt possède des propriétés prototypiques et non prototypiques ;
- comme il y a plusieurs points de vue pour analyser une Gestalt et ses parties, les propriétés des parties d'une Gestalt relèvent de plusieurs natures, tout comme leurs relations entre elles ;

[21] En quelque mesure, la notion de *structure* est évoquée ici dans l'acception mécaniste que Beauzée lui avait donnée au XVIII^{e} siècle (§ 1.1.3.).

- en raison des plusieurs analyses par plusieurs points de vue sur les parties, les Gestalts sont souvent intermodales (par exemple, la Gestalt perception visuelle influence la Gestalt linguistique) et peuvent interagir ou s'enchâsser entre elles, tout comme leurs propriétés (1977 : 246-247).

Cette synthèse par points (suivant, d'ailleurs, le style de présentation de Lakoff) restitue la vision d'une configuration commune à toute activité humaine, y compris la production langagière, qui peut se confronter avec la complexité, notamment celle des faits linguistiques. Chaque propriété relève, comme on le disait ci-dessus, d'un point de vue comme, entre autres, le point de vue grammatical, lexical, phonologique, etc. (1977 : 268). Chaque point de vue comporte la reconnaissance des propriétés prototypiques d'une Gestalt, à savoir celles que l'usage témoigne comme les plus fréquentes au sein de la même Gestalt (1977 : 249), et celles moins prototypiques. C'est par ses propriétés, notamment par celles prototypiques, que les locuteurs peuvent superposer partiellement des patrons (*partial pattern matching*) (1977 : 248) dans l'usage et produire, entre autres, des métaphores ou des expressions idiomatiques, y compris les parémies[22].

Gestalt appliquée à la sémantique des parémies (Visetti & Cadiot)

Ce détour par la *Gestalt linguistique* décrite par Lakoff aide à mieux comprendre l'étude de Visetti & Cadiot (2006)[23] qui s'intéressent à la genèse dynamique du sens proverbial. Ils contestent, d'abord, comme Lakoff, tous les schémas formels (non seulement ceux qui sont génératives). En tout cas, ils remettent en question :

> « la thèse d'un certain *schématisme* sémantique, de nature 'grammaticale' » (Visetti & Cadiot 2006 : 28)

[22] Le propos holistique de Lakoff consiste à dépasser l'approche générativiste et transformationnelle chomskyienne (1977 : 265), et ce, même au moyen d'une nouvelle perception visuelle de la surface des faits linguistiques. Plus précisément, Lakoff remplace la représentation sous la forme d'un arbre de dérivation syntaxique par une représentation sous la forme d'un réseau orienté qui décrit les relations entre parties d'après leurs propriétés. Cette distinction de représentation influencera la modélisation des langues en vue de traitements informatiques (§ 4.2.).

[23] On remarque que l'œuvre de Paulhan est ouvertement mentionnée parmi les références de leur ouvrage. Nous soulignons également que Lakoff n'est mentionné dans leur ouvrage que pour ses études successives sur la métaphore.

des linguistiques cognitives et de l'énonciation qui articulerait un modèle perceptif immanent des faits linguistiques. Par conséquent, ils suggèrent :

> « un retour critique aux écoles historiques de la Gestalt et en même temps à la philosophie phénoménologique […] [pour] un mode phénoménologique de théorisation, bien distinct des modes formels » (Visetti & Cadiot 2006 : 29, c'est nous qui soulignons).

Les auteurs sortent des modes formels pour postuler une théorie des *formes sémantiques*[24] qui décrit tant la relation perception-langage que la relation analogique entre formes de perception et formes sémantiques linguistiques[25]. Ils identifient trois formes : les *motifs*, les *profils* et les *thèmes*. Elles représentent trois phases de nature éminemment sémantique. Les *motifs* gèrent le processus de formalisation lexicale et grammaticale en tant que « germes de signification chaotiques et/ou instables » (2006 : 39). C'est la définition des *profils* en tant que phase de cristallisation (non pas de figement) intermédiaire qui rapproche les auteurs de la vision formelle et schématisante qu'ils ont rejetée :

> « Les profils renvoient aux dynamiques de stabilisation différentielle des lexèmes, qui s'interdéfinissent sur le fond de champ ou de domaines sémantiques, et corrélativement par détermination réciproque dans une syntagmatique (partiellement enregistrée, qu'il s'agisse de grammaire ou d'idiomaticité[26]) » (*ibid*.).

Pour finir, les *thèmes* agissent comme des chemins sémiotiques (linguistique et référentiel) qui supportent la modélisation des motifs et la stabilisation des profils (*ibid*.).

La synthèse de l'interaction de ces phases dans la dynamique proverbiale est à repérer dans une note en bas de page dans leur ouvrage sur les formes sémantiques de 2001. Cette note est décidément parlante en ce qui concerne le rapprochement avec les intuitions de Paulhan ainsi qu'avec la formalisation de la *Gestalt linguistique* par Lakoff. En outre, on

[24] Par *forme*, ils entendent une unité qui prend également en compte le côté morphologique des faits linguistiques, sans qu'elle soit identifiée seulement avec ceci. La forme doit en outre éviter toute réduction à un schéma, notamment à un schéma logique (Visetti & Cadiot 2006 : 30). La référence au terme *forme* est à motiver, malgré l'incohérence apparente, par le courant dénommé *psychologie de la forme* dont la notion de Gestalt en est le pilier.

[25] Ce propos est partagé par Lakoff, aussi, notamment par son encadrement dans une *linguistique expérientielle* (voir ci-dessus).

[26] Cet enregistrement partiel et cette dynamique de corrélation rappelle le *partial pattern matching* par propriétés prototypiques de Lakoff (voir ci-dessus).

constate que le formel rejeté par les auteurs, se fait promoteur d'innovations langagières par analogie. Les proverbes, ou leur :

> « gestalt linguistique […] porte un motif thématique transposable à une pluralité indéfinie de situations. Il y a bien, comme dans les paraboles, une thématique première à fonction allégorique : mais en même temps se trouve dans une armature, un moule, dont la transposabilité immédiate doit être perçue comme telle, puisqu'elle constitue justement le sens proverbial. Le proverbe est bien sûr une forme thématique déjà très avancée dans sa stabilisation [...]. Mais il promeut en même temps des motifs originaux, encore instables, qui condensent une analogie ouverte. La perception de cette transposabilité suspendue sans être effectuée correspond à la promotion de ces nouveaux motifs, qui sont plus particulièrement indexés sur certains syntagmes (notamment les groupes nominaux constituant l'*incipit*[27] [...]) ; le reste du matériel lexical et grammatical affichant des profils plus banalement attestés, donnés comme invariants pour ces transpositions mêmes » (Cadiot & Visetti 2001 : 160, c'est nous qui soulignons).

La transposabilité des formes sémantiques proverbiales passe ainsi par « certains syntagmes » et, donc, par certaines cooccurrences qui déclenchent de nouveaux germes de signification en attente de stabilisation, et ce, grâce aussi au « reste du matériel lexical et grammatical affichant des profils plus banalement attestés » dans l'usage. C'est l'usage qui donne une forme à l'instabilité sémantique et qui n'est pas forcément à interpréter comme le but ultime de l'appréhension d'un fait linguistique, mais plutôt comme un passage essentiel pour atteindre la compréhension de la parémie. D'ailleurs :

> « du point de vue cognitif, les formes schématiques, par leur abstraction, ne peuvent 'se donner' aux locuteurs que sous la forme d'exemplaires, donc de réalisations concrètes idéalisées » (Legallois & François 2011 : 13).

Contrairement à ce que défendent Visetti & Cadiot, nous suivrons cette perception de schématisation formelle de nature lexico-grammaticale. À notre avis, elle ne constitue ni une limitation à la transposabilité des formes sémantiques ni une négation de la *condensation* des qualités sémiotiques et expérientielles (Visetti & Cadiot 2006 : 35). La perception d'une

[27] C'est par le recours à ces parties initiales que les parémiologues ont essayé et essaient de repérer des occurrences des proverbes, de leurs variantes et de leurs détournements dans l'usage, notamment à l'écrit, sur corpus (§ 4.).

schématisation formelle est plutôt à voir comme *une des manifestations possibles* de cette condensation ou, pour le dire avec Lakoff, comme *un point de vue sur la Gestalt linguistique-proverbe*. D'ailleurs, son rejet comporterait, d'une part, un affaiblissement de la transposabilité des formes sémantiques (§ 7.2.) et, d'autre part, irait à l'encontre même de l'envergure dont la notion de *Gestalt* est porteuse de par sa nature.

Schémas formels **des parémies**

Du côté des études logico-formelles, on rencontre la reconnaissance d'un cadre holistique dans les contributions de beaucoup de parémiologues, quoique chacun d'entre eux porte un degré d'attention différent ou privilégié à l'égard de tel ou de tel autre point de vue linguistique.

Malgré son analyse facettée, Permyakov se concentre sur la perspective logique et ses *archinvariants logico-sémiotiques* sont justement 4 propositions logiques à l'origine de toute (re)production et de la classification des proverbes. De la lecture des *Jeux-Partis*, Buridant dégage des schèmes narratifs sémantiques et logiques qui deviennent des *traits de classification*. La logique de la proposition et la syntaxe de la phrase fondent la notion de *moule syntaxique* présentée par Achard-Bayle & Schneider (2010). Après une distinction entre proverbes hypotaxiques et proverbes parataxiques, ils parlent de *moule* pour faire référence à la binarité propositionnelle des proverbes et aux relations logiques qui se créent entre les deux propositions impliquées.

Entre phrase grammaticale, proposition logique, syntaxe et lexique, la notion de *moule proverbial* proposée par Gómez-Jordana Ferary (2012) sert à décrire les cadres de (re)production d'une liste de parémies qu'elle a établie (§ 3.2.11.).

Par une attention plus axée sur le lexique, Arnaud & Moon reconnaissent le *cliché proverbial* (1993 : 334) comme cadre lexico-grammatical qui (re)produit fréquemment des expressions sémantiquement corrélées au proverbe-matrice (§ 4.2.1.1.).

D'après la notion de *schéma phraséologique* proposée par Zuluaga et avec la même perspective lexicale et syntaxique, Anscombre (2011b) élabore celle de *matrice lexicale* pour décrire l'idiomaticité de certaines expressions linguistiques, y compris les proverbes. D'après Anscombre, la matrice lexicale est un « schéma comportant des unités linguistiques fixes » (2011 : 25), à savoir des unités grammaticales, « et des variables linguistiques » (*ibid.*), c'est-à-dire des unités lexicales. La matrice est plus ou moins productive et obéit aux contraintes propres au schéma ainsi qu'aux relations entre unités lexicales. Anscombre précise que :

> « les textes parémiques correspondent en fait à des matrices lexicales, les plus connues étant *Qui* [*GV*1] [*GV*2], [*øN*] *Qui* [*GV*1] [*GV*2], ou encore *A* [*GN*1] [*GN*2] » (2011 : 38).

Il ajoute que ces matrices sont en nombre limité d'après la classification taxinomique des structures proverbiales proposée par Gómez-Jordana Ferary (2003) et sur laquelle il s'appuie. Anscombre présuppose aussi que, sans aucune distinction typologique, les parémies partagent des matrices lexicales communes. Il constate encore que :

> « les relations qui interviennent dans ces matrices proviennent de choix culturels conventionnalisés […] [et] que ces contraintes pèsent sur les possibilités de manipulation des items produits » (2011 : 39).

Les matrices lexicales d'Anscombre et les moules proverbiaux de Gómez-Jordana Ferary sont déjà à repérer dans les travaux de Conenna sous le nom de *moules syntaxiques* (2000a, 2002). Avec une attention à la syntaxe et au lexique, elle identifie ces cadres lexico-grammaticaux de matrice proverbiale à combinatoire restreinte dans le sillage des études sur le figement. Ces moules sont significativement productifs dans le répertoire parémiologique français et sont dégagés à partir des *structures syntaxiques* décrites suivant la méthode du Lexique-Grammaire. Suivant la même méthode et les observations de Conenna, Mogorrón Huerta & Navarro Brotons reprennent le *moule proverbial* (2012). Sans l'encadrement d'une méthode et sans un travail systématique de description syntaxique, Schapira parle elle aussi d'un *modèle proverbial* (1999 : 95-99) pour faire référence au même fait. De façon pareille, dans son analyse des détournements proverbiaux, Barta l'appelle *schéma proverbial* (2005 : 140).

Outre les exemples en parémiologie francophone, la récurrence de schémas formels est mise en relief par des parémiologues qui travaillent en d'autres langues. En parémiologie anglophone, Dundes nomme ces cadres sous le nom de *proverb architectural formula* (1994 [1975] : 46) et spécifie qu'il doit y en avoir un nombre fini. Taylor défend l'idée de *frames* (1962 [1931] : 16) pour signifier l'existence de modèles de proverbialisation basés sur des proverbes existants, et ce, en raison de leur familiarité et de leur pertinence (1962 [1931] : 20). Nous soulignons que les *frames* ne peuvent être dégagés que par une comparaison

méthodique qui permet de remonter aux proverbes-sources de ces cadres[28]. En parémiologie hispanophone et dans le sillage de la leçon de Taylor, Arora (1998) fait un travail sur le *proverb patterning* de la partie initiale *El que nace* [trad. litt. : Celui qui naît] pour dégager des patrons de proverbialisation à partir d'occurrences réelles et qu'elle assemble en groupes de par les types de variations paradigmatiques et syntagmatiques[29].

1.1.5. Pour une *Gestalt parémique* : la *séquence lexico-grammaticale*

En guise de synthèse, on constate que les unités exploitées jusqu'à présent pour décrire et pour définir les proverbes et, en général, les parémies n'ont abouti qu'à l'appréhension d'une de ses facettes ou de quelques-uns de ses comportements linguistiques. Elles ont plutôt contribué à confondre le statut linguistique de *parémie*. Ce qu'on peut observer par les fluctuations et les superpositions métaterminologiques mises en œuvre par certains parémiologues. Nous délaisserons ces unités, et ce, :

- en raison de leur utilisation désabusée en parémiologie
- ainsi que (et surtout) à cause d'un manque de consensus, d'une part, en ce qui concerne leur même définition et, d'autre part, leur fiabilité en termes opérationnels et en vue de généralisations successives.

À l'exception près de quelques perplexités manifestées par certains et de quelques modulations dans l'étiquetage, il est étonnant d'observer que la quasi-totalité des parémiologues ait essayé de définir et de décrire les parémies par des unités minimales instables du point de vue notionnel. Par conséquent, bon nombre de parémiologues ont amplifié la confusion linguistique par le recours à des unités atteintes à leur tour par une confusion notionnelle.

Nous éviterons ainsi la *phrase* parce que la parémie, pour reprendre les mots de Berrendonner (§ 1.1.1.), n'est ni un « individu exemplaire » ni un « spécimen marginal » de

[28] Taylor ajoute : « no one has undertaken a study of this sort » (1962 [1931] : 16). Cela est vrai encore aujourd'hui pour les proverbes anglais.

[29] Au-delà des proverbes, la littérature en phraséologie foisonne aussi de schémas/cadres lexico-grammaticaux. C'est le cas : des *collocational frameworks* (Sinclair & Renouf 1991), des *quasi-segment*s (Habert *et al.* 1997 : 200) et des *concgrams* (Cheng *et al*. 2006) par la lecture des concordances en linguistique de corpus ; des *frames* en Grammaire des Patrons (Hunston & Francis 2000 : 25-27) ; des *collostructions* en Grammaire des Constructions (Stefanowitsch & Gries 2003 : 215) ; des *« schémas » figés* dans les études sur les expressions figées dans une partie de la francophonie (Lamiroy *et al.* 2010 : 80).

ce prototype unitaire. Comme la phrase, la parémie aussi est très souvent entrelacée, voire identifiée, avec la *proposition* logique et confondue avec les propriétés qui relèvent de ses constituants, notamment du verbe. De toutes les nuances de la notion de *phrase*, on ne retiendra que la *phrase typographique*, c'est-à-dire la suite de caractères qui commence par une majuscule et se termine par un point final, et ce, pour des raisons liées au traitement informatique que nous avons prévu (§§ 5.2, 6.). Nous laisserons aussi de côté les notions 'discursives' : l'*énoncé*, l'*énonciation* et le *discours* sont significativement subjectives et variables, d'après le linguiste interpellé, sans compter les enchevêtrements qu'elles entretiennent avec la notion de *phrase*. L'acception désormais généralisée de 'tout qui se tient' attribuée à la notion de *structure* pourrait nous convenir. Pourtant, sa déconstruction identitaire et sa resémantisation persistante observées au fil des siècles ainsi que dans la littérature parémiologique la plus récente nous conseillent de ne pas contribuer ultérieurement à cette 'pagaille pseudo-structurale'.

Nous approcherons la parémie comme une *Gestalt linguistique* au sens de Lakoff. Par l'adoption d'un point de vue à la fois lexical et grammatical, nous mettrons en relief la ***Gestalt lexico-grammaticale*** d'un échantillon de parémies (§ 5.1.). Cela veut dire que nous partirons de la surface lexico-grammaticale des parémies pour décrire les propriétés prototypiques qui résultent de l'interaction entre les constituants qui les composent. Sur l'escorte de certains *schémas formels* reconnus, comme le *moule syntaxique* de Conenna, le *cliché proverbial* d'Arnaud & Moon ou la *matrice lexicale* d'Anscombre, nous repartirons par une analyse en constituants, suivant l'approche distributionnelle harrisienne (Harris 1976). Comme le constate Maurice Gross :

> « il est vraisemblable que le premier critère de ressemblance (au-delà des répétition [*sic*] d'occurrences de mots) qui ait été découvert est celui que constituent les parties du discours » (1976 : 18-19).

Or, il est raisonnable de se demander si les parties du discours peuvent représenter des unités minimales descriptives fiables. Comme le montre le numéro 92 de *Langages* consacré à l'histoire des parties du discours (de l'Antiquité jusqu'à aujourd'hui, en passant par la *Grammaire générale et raisonnée* de Port-Royal), les parties du discours n'ont pas vraiment changé de peau, mais elles ont, certes, changé leur substance. Malgré de nombreuses critiques, souvent fondées sur l'identité entre *partie du discours* et *mot*, nous faisons nôtre la conclusion de Lagarde :

> « Envisagée dans sa fonctionalité [*sic*] épistémologique, l'antique théorie qui depuis des siècles modèle nos consciences grammaticales occidentales est celle qui permettrait d'obtenir des résultats tout à fait appréciables avec le minimum d'effort. Cela expliquerait pourquoi elle est finalement conservée par la plupart des linguistes, non seulement par ceux qui creusent le sillon traditionnel, mais aussi par ceux qui explorent des voies nouvelles » (1988 : 106).

En d'autres termes, les parties du discours fonctionnent bien et sans (trop de) peine[30]. Elles continuent à être employées dans les « voies nouvelles » que sont représentées par les *grammaires émergentes* (*usage-based*) : Grammaires des Patrons et Grammaire des Constructions (Legallois & François 2006), Analyse des Patrons sur Corpus (Hanks 2004a, 2004b), Théorie des Normes et des Exploitations, etc. Il est peut-être venu le temps de les accepter et de les considérer systématiquement pour une description exhaustive des parémies, comme l'a déjà fait Conenna (et d'autres, par la suite)[31].

Loin de proposer, pourtant, une Gestalt d'après la seule perspective syntaxique, nous suivrons l'exemple de ces linguistiques empiriques basées sur corpus. Nous prendrons également en compte le lexique. Par une approche contextualiste qui vient de John Rupert Firth (1957 [1951], 1968 [1957]), nous considérerons les unités lexicales comme unités porteuses d'un sens qui correspond (du moins) à un mot graphique et qui s'explicite par la cosélection des unités lexicales entre elles. Les débats sur les notions de *mot* et d'*unité lexicale* sont bien connus, notamment en sémantique (Anscombre 1994b), lexicologie (Polguère 2003) et en terminologie (L'Homme 2004). Comme nous venons de le dire, nous entendons par *unité lexicale* le *mot graphique*, à savoir une suite de caractères marquée par deux espaces en début et fin de suite, qui possède un *noyau sémantique* dans sa singularité. Par cela, nous n'entendons nullement ignorer l'existence des *unités polylexicales*, comme le soulignent les travaux sur la composition et le figement (G. Gross 1996). Toutefois, le niveau

[30] L'argument de la fonctionnalité épistémologique pourrait être insuffisant pour les détracteurs. On pourrait argumenter à juste titre que des grammaires et des théories linguistiques exploitant les notions de *phrase*, *énoncé* ou *structure*, satisfont les nécessités pour lesquelles elles ont été conçues. Comme nous espérons l'avoir montré ci-dessus, nous nous limitons simplement à constater ici que ces unités ne conviennent pas (ou qu'elles conviennent très peu) à un éclairage sur la nature linguistique, et plus précisément lexico-grammaticale, des parémies. De plus, ces unités se proposent comme un 'tout' pour appréhender le tout-parémie. Nous visons plutôt la constitution d'un tout-Gestalt-parémie par l'analyse des propriétés de ses constituants. Nous rejoignons donc les parémiologues, comme Kleiber, qui considèrent le proverbe/parémie comme une catégorie en soi, quoique ce dernier en défende le statut phrastique.

[31] On pourrait justement remarquer que les parties du discours, notamment dans la période de Port-Royal, résultent de l'adoption du cadre de la *proposition* logique pour la description et pour l'analyse linguistique. Cela n'équivaut pourtant pas au fait que nous acceptons tout court la proposition comme unité minimale. Nous acceptons les parties du discours en tant que constituants des chaînes écrites et parlées et comme unités épistémologiquement fiables.

de la composition (comme l'implique son nom, d'ailleurs) suit la différenciation d'unités simples à mettre ensemble pour constituer un tout cohérent et pour envisager une organisation hiérarchique du lexique. Autrement dit, la composition et le figement sont des opérations d'assemblage que nous effectuerons après toute opération de reconnaissance d'unités lexicales minimales autonomes. Leur assemblage s'appropriera, d'une part, des patrons syntaxiques codifiés par G. Gross et, d'autre part, de l'*idiom principle* sinclairien qui veut que :

> « […] the nature of the world around is reflected in the organization of language and contributes to the unrandomness. Things which occur physically together have a stronger chance of being mentioned together; also [...] the results of exercising a number of organizing features such as contrasts and series. [...] The principle of idiom is that a language user has available [...] a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments » (Sinclair 1991 : 110).

Des ensembles d'unités lexicales discrètes sont ressentis et exploités comme un choix linguistique unitaire, et ce, parce qu'ils reflètent l'organisation ensembliste du monde réel ainsi que nos processus de classification et catégorisation. Cette vision de la langue se rapproche de la vulgate *catégorielle* du proverbe qui est défendue notamment par Kleiber (2010a). Pour Kleiber, les proverbes-phrases dénomment et catégorisent des situations de vie récurrentes : ils sont donc des exemplaires centraux de l'*idiom principle* sinclairien. C'est d'ailleurs la vision de la parémie en tant que *catégorie lexico-grammaticale à part entière* que nous défendrons au cours de notre étude, à l'exception substantielle de l'encadrement phrastique. À ce propos, nous tendrons vers l'approche contextualiste sinclairienne dans la mesure où nous entendrons la parémie comme le résultat d'agencements lexico-syntaxiques récurrents de parties du discours et d'unités lexicales.

La considération de la polylexicalité sera l'étape suivante d'un processus de formalisation médié par les études sur le figement, sans que cela implique que nous regardons à la parémie comme une unité figée. Comme nous venons de l'expliquer, nous appréhenderons la *Gestalt lexico-grammaticale parémique* à l'aide d'une *unité de cosélection étendue des parties du discours et des unités lexicales*, à savoir la ***séquence lexico-grammaticale*** (SLG). Autrement dit, on retiendra à la base les agencements entre unités syntaxiques et lexicales discrètes de manière à intégrer, par la suite, les formalisations

suggérées par la syntaxe et par l'usage pour interpréter ses agencements à un niveau plus éloigné de la surface lexico-grammaticale.

Nos séquences lexico-grammaticales décrivent les parémies, mais ne sont pas à identifier *ipso facto* aux parémies. En d'autres termes, elles ne représentent pas LA parémie dans son intégralité. Elles n'ont donc pas la prétention de la définir. Elles constituent plutôt des agencements stratégiques conventionnalisés (par leur fréquence d'occurrence et par leur dispersion) que les locuteurs choisissent pour faire appel (par analogie ou explicitement) aux parémies ou pour parémiser des enchaînements de parties du discours et d'unités lexicales. Ce choix relève du principe cognitif du *chunking*. Ce principe essentiel en linguistique émergente se manifeste de manière évidente quand on se concentre sur toute expression préfabriquée et idiomatique (Bybee 2010 : 35). Il nous permet de représenter les parémies comme « *sequential connections between one word and the next* », sans oublier que « *such connections can have varying strength, depending on their frequency of co-occurrence* » (Bybee 2010 : 36).

L'interface lexique-syntaxe et la rencontre distribution-contexte nous permettront de réunir deux approches qui, séparément, ont produit et produisent, d'une part, des descriptions syntactico-centriques et abstraites (comme dans le cas des études de matrice harrisienne sur le figement) et, d'autre part, des recherches lexico-centriques sans généralisations théoriques (comme beaucoup d'études phraséologiques sur corpus le prouvent, à l'exception toute récente de la Théories des Normes et des Exploitations par Hanks). Notre perspective sur la *Gestalt linguistique parémique* sera ainsi lexico-syntactico-centrique. Bien évidemment, la perspective lexico-grammaticale n'est qu'une des perspectives possibles pour aborder la *Gestalt linguistique parémique* et pourra appuyer (nous l'espérons) les spéculations successives, suivant d'autres points de vue linguistiques.

La double perspective lexicale et syntaxique nous permettra aussi de suggérer une classification lexico-grammaticale des parémies conséquente (§ 1.2.5., 5.3.). De plus, la mise en relief des propriétés prototypiques formelles nous servira pour la modélisation des requêtes informatiques pour la reconnaissance automatique des parémies et d'autres séquences formulaires sur corpus (§ 6).

1.2. Métaclassification des classifications parémiologiques

Toute activité de classification présuppose un ensemble d'objets (réels ou abstraits) non ordonné et relativement fermé. Les parémies, un ensemble apparemment fermé d'objets

linguistiques échappent, pourtant, à toute activité de classification. Du moins, les critères de classification proposés n'appréhendent pas d'un seul coup toutes les facettes qui font de la parémie un objet multidimensionnel. En 1976, Gross remarque que :

> « la taxonomie a mauvaise réputation aujourd'hui dans le monde des linguistes » (1976 : 7).

On ne pourrait dire qu'il en est allé pareillement en parémiologie. Les parémiologues se sont toujours confrontés avec le « problème insoluble de classification » (Visetti & Cadiot 2006 : 317). Entre ontologie et vertige des listes parémiographiques, les parémies occupent une place dans plusieurs tentatives de rangement.

Dans les paragraphes qui suivent, nous nous concentrerons sur quelques-unes de ces tentatives proposées par les parémiologues. Nous présenterons ainsi une métaclassification parémiologique. Il s'agit d'une métaclassification incomplète et discutable. Elle essaie de mettre de l'ordre, pour autant que possible et d'après nos connaissances, par analogie et de manière diachronique (presque philologique), dans la littérature en la matière. Les classifications mentionnées concernent principalement les parémies françaises, quoique des expériences portant sur d'autres langues et sur des études comparées soient également mentionnées. Nous avons essayé de réunir ces classifications – soit qu'elles relèvent de l'intention des parémiologues, soit qu'elles découlent implicitement de leurs études – sous des macrocatégories. La distinction entre macrocatégories veut rendre compte du tiraillement méthodologique des parémiologues autour d'un (le plus souvent, de plusieurs) critère(s). Nous précisons, en tout cas, que le recours à une notion-clé sert parfois de pivot et de justification pour assurer la cohérence d'une analyse s'appuyant sur plusieurs critères.

Il s'avérera difficile, pour certains cas, de limiter une proposition de classification à une seule macrocatégorie. Autrement dit, les macrocatégories de cette métaclassification ne sont pas vraiment isolées entre elles, et ce, comme nous le rappelions au § 1.1., en raison des différentes natures et des différents statuts attribués aux parémies. Ce qui est, d'ailleurs, mis en évidence par le cas de Permyakov qui mériterait un traitement à part entière. D'après Permyakov, les classifications parémiologiques ont ignoré la nature polymorphe du proverbe qui demande :

> « a three-pronged approach, in which they are treated as phenomena of language, phenomena of thought and phenomena of folklore » (1979 : 9).

Sa classification adopte cette approche tripartite dont nous essayerons de rendre compte dans les diverses macrocatégories.

1.2.1. *Classifications typologiques (ou textuelles)*

Les enjeux des classifications typologiques (ou textuelles) suivent de près les tentatives de donner une définition de *proverbe* et de *parémie*, tout comme l'adoption (et le souhait d'une adoption sur une large échelle) de néologismes ou de resémantisations métaterminologiques en parémiologie.

Le caractère gnomique (à savoir la brièveté) a servi de condition nécessaire initiale, mais insuffisante, pour démarrer l'acheminement (sans fin) des classifications typologiques. Au début, les parémiologues se sont souvent inspirés du paradigme structuraliste et ont visé à isoler des traits proverbiaux à la manière d'une analyse sémantique componentielle. Greimas (1960)[32] amorce ce genre de démarche et met en relief les propriétés d'un code, de ce « système clos de signification » que représentent les proverbes et les dictons. Il les distingue ainsi d'après la présence ou l'absence d'une série de critères formels qui ont affaire à la parémie sous toutes ses facettes : syntaxique, sémantique, prosodique et lexicale. Il identifie, en tout cas, la *connotation*, à savoir le transfert métaphorique, comme le seul véritable critère distinctif entre les *proverbes* et les dictons. Nous remarquons que cette distinction proverbes/dictons sur la présence ou l'absence éventuelle de transfert métaphorique caractérisera les études postérieures de bien d'autres parémiologues, sans distinction de cadres théoriques et méthodologiques.

Quelques années plus tard, Rodegem (1972) présente une classification par traits des *locutions sentencieuses*, hyperonyme (initialement) choisi pour représenter toute formule sapientielle, et ce, sur la base du *Rythme* (R), de la *Norme* (N) et de la *Métaphore* (M). Le *proverbe* jouit, à son avis, de la co-présence de ces critères et se distingue, par exemple, du *dicton*, ce dernier ne possédant pas le trait M, quoiqu'il admette, en tout cas, qu'un dicton peut ambitionner à la proverbialisation par un usage métaphorique.

[32] Il est réducteur de notre part de restreindre le potentiel classificatoire de l'étude de Greimas à cette macrocatégorie. Sa contribution est souvent mentionnée par les parémiologues qui ont puisé et s'inspirent encore de son travail (plus ou moins consciemment). Pourtant, les efforts de Greimas de caractériser le *code parémique* autonome et fermé, ainsi que de caractériser deux genres textuels (*proverbes* et *dictons*), tout en ayant conscience de leur différence au cours de son argumentation, nous suggèrent de lui réserver une place d'honneur dans cette macrocatégorie.

Par la suite, l'étiquette grecque παροιμία a essayé de redéfinir les rapports entre le proverbe et ses 'semblables' sentencieux. *Parémie* se propose ainsi en tant qu'étiquette hyperonymique valable pour toute expression gnomique et formulaire sapientielle, sentencieuse et folklorique codifiée et faisant partie du patrimoine linguistico-culturel d'une langue donnée. Elle sert à la fois de porte d'entrée, ou plutôt de tête de la hiérarchie de ces expressions gnomiques et formulaires, ainsi que de contrepartie idéale par rapport à *unité phraséologique*, hyperonyme attaché (en gros) à toute expression formulaire non parémique. La distinction entre parémiologie et phraséologie est ainsi (apparemment) établie par ces deux métatermes[33]. Du moins, c'est la leçon qu'on dégage des *Notes On Structural Paremiology* de Permyakov, quoique l'auteur n'empêche pas *a priori* à la phraséologie au sens large de s'occuper de l'étude des parémies (1979 : 135). L'inverse, en tout cas, n'est pas prévu : les parémiologues sont confinés à l'étude des parémies, sans possibilité de se tourner vers les unités phraséologiques (*ibid*.).

Le terme *paremia* recouvre pour Permyakov des types de clichés qui sont :

> « folk sayings expressed in sentence form (e.g. proverbs, proverbial phrases, omens) as well as in the form of short strings of sentences representing an elementary scene or a simple dialogue » (1979 : 133).

La complexité des parémies et leurs anomalies structurelles et sémantiques par rapport aux unités phraséologiques, font que Permyakov décide de traiter les parémies :

> « as a separate class of language units, which I call the **paremiological level of language** » (1979 : 138; c'est l'auteur qui souligne)

où les proverbes se distinguent d'autres parémies pour leur structure linguistique externe phrastique fermée et pour leur structure interne linguistique synthétique.

En milieu francophone, c'est Rodegem qui redémarre la roue :

> « Quel vocable convient le mieux pour désigner l'ensemble des formules consacrées par l'usage, qu'on appelle généralement « proverbes » ? […] pour éviter les ambiguïtés, je réserverai le terme de parémie aux énoncés sentencieux pris globalement » (Rodegem 1984 : 121, c'est nous qui soulignons).

[33] Cette distinction, à notre avis, est justifiable, mais discutable (§ Introduction).

L'étirement métaterminologique est évident tout au long de son argumentation[34], alors que les annexes présentent un aperçu d'exemples de parémies, à savoir une taxinomie plate à un seul niveau, où l'on présente les proverbes, les maximes, les wellerismes, etc. l'un après l'autre.

Par des critères -filtres « pour ne laisser passer que les proverbes » (Arnaud 1991 : 7) comme dans une passoire linguistique, Arnaud classifie les « énoncés de la famille du proverbe » (§ 1.1.2.), tout en étant conscient que :

> « Les entités phraséologiques forment un fatras à première vue inextricable et rebelle au classement » (Arnaud 1991 : 7).

Le terme *parémie* laisse la place aux *énoncés* et aux *entités phraséologiques*. Ce qui estompe à nouveau les limites entre parémiologie et phraséologie.
Dans cette passoire linguistique, les unités phraséologiques coulent au travers des 5 filtres suivants :

1. la *lexicalité*, à savoir la préconstruction codifiée d'un proverbe et son partage par les locuteurs ;
2. l'*autonomie syntaxique* ou la possibilité de les énoncer sans ajouts et modifications ;
3. l'*autonomie textuelle* pour laquelle l'énonciation du proverbe ne relève pas du nombre des locuteurs ;
4. la *valeur de vérité générale*, c'est-à-dire que le contenu logico-sémantique du proverbe a une portée classificatoire des situations du monde réel ou discursif, qu'ils soient impliqués, évoqués et/ou créés ;
5. l'*anonymat* ou l''oubli' et la perte du premier locuteur absolu (Arnaud 1991 : 8-12).

Les proverbes franchissent tous ces critères de nature lexicale, logique et discursive, alors que les *locutions proverbiales* s'arrêtent, par exemple, au critère de l'autonomie syntaxique, car elles nécessitent d'autres éléments syntaxiques pour l'achèvement d'un énoncé. Les *énoncés phrastiques à valeur spécifique* comme :

[34] Il suffit de considérer l'utilisation de l'adjectif *proverbial* à la place de *parémique*, désormais accepté et répandu en littérature.

Il y a de l'orage dans l'air

ne possèdent pas une valeur de vérité générale. Dans ce cas de figure, cet énoncé s'applique à une situation nuageuse et dangereuse, qu'elle soit réelle (météorologique) ou perçue (psychique). Les *slogans* et les *aphorismes* sont rattachés à leurs locuteurs originaires, soient-ils des individus ou des groupes, ce qui ne leur permet de passer en proverbes. Nous soulignons que, par manque de familiarité requise par le critère de lexicalité, Arnaud écarte :

Oignez vilain, il vous poindra ; poignez vilain, il vous oindra

que la plupart des répertoires parémiographiques enregistrent comme *proverbe*. Il affirme :

> « […] peu de francophones, en dehors des spécialistes et des amateurs de la langue, le connaissent. Cela en justifie l'élimination dans la perspective synchronique et sélective qui est la nôtre » (Arnaud 1991 : 8).

Partageant l'approche d'Arnaud basée sur l'usage, il nous est, pourtant, difficile d'accepter une remise en question du statut d'un *proverbe* seulement par le recours à la familiarité. Celle-ci n'affecte pas la préconstruction codifiée de la suite lexicale en soi et repose plutôt sur des aléas extralinguistiques de nature sociale, économique et culturelle qui affectent toutes les études de familiarité[35].

Quelques années plus tard, dans son ouvrage sur les stéréotypes linguistiques en français, Schapira elle aussi rassemble des critères variés pour proposer sa classification des *formes brèves*. Entre autres[36], elle parle de *formes phrastiques figées* où elle inclut :

- les *apophtegmes* ;
- les *slogans* ;

[35] La fréquence sur corpus qui nous intéresse, ne peut non plus compromettre la nature linguistique du proverbe. La familiarité et la fréquence ne sont à interpréter que comme des indicateurs d'usage.

[36] Par souci d'exhaustivité, nous mentionnons aussi les autres catégories. Elle distingue les *genres libres* (aphorisme, maxime, pensée et sentence) en tant que genres littéraires et phrases signées qui agissent de manière autonome en discours et dont les origines sont à repérer dans des textes littéraires, mais aussi non littéraires (1999 : 48-50). Elle identifie aussi les *formules de politesse*, les *formules rituelles*, les *stéréotypes de circonstance* et les *truismes* (1999 : 51). Il nous paraît pertinent de souligner la juxtaposition de plusieurs métatermes qui relèvent de différentes branches linguistiques (*classe*, *énoncé*, *forme*, *formule*, *genre*, *phrase*). Ce qui témoigne de la difficulté de cerner tous les faits linguistiques dont la combinatoire dépasse l'unité lexicale simple.

- les *phrases de routine* ;
- les *énoncés parémiques* : « qui regroupent les classes des formules figées (proverbes, dictons et adages » (1999 : 50).

Dans ce cas, le terme *parémie* subit une mutation en faveur d'*énoncé parémique*. Cette variation métaterminologique restreint le champ d'attribution à des faits linguistiques relevant du discours et à un nombre réduit de combinatoires polylexicales stabilisées par l'usage, à savoir les *adages*, les *dictons* et les *proverbes*. Il en suit, donc, que le proverbe est une forme phrastique figée (ou plus simplement, *phrase figée*) à valeur discursive.

Dans ses études comparées français-espagnol et de manière différente par rapport à Schapira, Sevilla Muñoz[37] revient à plusieurs reprises sur la classification des *parémies*/*de las paremia* (1993, 2000, 2008). Pour elle, l'étiquette *paremia* « *es el archilexema del campo sapiencial* » (Sevilla Muñoz 2008 : 241) qui peut remettre de l'ordre dans la confusion métaterminologique autour du concept *refran* et, en général, des expressions sapientielles. En particulier, elle opte pour une classification pentapartite des parémies, où les proverbes et les dictons/*refranes* sont à insérer dans le regroupement des *parémies proprement dites*, à côté des :

- *parémies scientifiques*, telles que les théorèmes ou les axiomes ;
- *parémies comiques ou satiriques*, comme les wellerismes ;
- *parémies épiques*, à savoir les cris de guerre ou les devises ;
- *parémies publicitaires*, c'est-à-dire les slogans (Sevilla Muñoz 2008 : 242-244).

Nous constatons que le dégagement des genres textuels parémiques comporte parfois le recours au critère *domaine du savoir*[38]. En l'occurrence, Sevilla Muñoz identifie des *parémies scientifiques*. Encore, Rodegem fait référence à la norme, notamment à la *norme spécifique*, pour autonomiser les *adages juridiques* (1984 : 125). Toujours dans le domaine du droit, on assiste à une taxonomie approximative des différentes parémies à valeur juridique proposées par le juriste Cornu (2005) et auxquelles certains parémiologues se sont

[37] Nous remercions Julia Sevilla Muñoz pour la grande générosité et son estime.
[38] Le recours au *domaine* ouvre souvent la voie à des classifications thématiques (§ 1.2.4.).

attachés en vue d'une description linguistique (Asensio Sánchez 2008 ; Calvo Espiga 2008 ; Conenna *et al.* 2011 ; Gouvard 1999 ; Serrone 2013)[39].

Un cas à part entière est représenté par Anscombre[40], porteur du [*proverbe*]-concept et de la notion de *classe proverbiale* de nature ontologique (Anscombre 2000 : 9). Anscombre établit une classification des *phrases autonomes* (Anscombre 2008a, 2008b), plus précisément de la sous-classe des *phrases sentencieuses* (§ 1.1.1.). Il met au point une arborescence où les phrases sentencieuses se distinguent premièrement par la spécification (ou la non-spécification) d'auteurs et, ensuite, par les caractéristiques rythmiques et par le degré de métaphoricité (Anscombre 2008b : 22). La signature d'une phrase autonome, accompagnée d'un marqueur médiatif (ex. *comme le dit M. Dupont*), permet de regrouper les maximes, les sentences et les morales dans une seule sous-classe. L'absence de signature des phrases à *ON-locuteur* crée deux autres sous-classes :

1. les *phrases situationnelles* (ex. *Les jeux sont faits*) et
2. les *phrases parémiques*.

Cette sous-classe en prévoit d'autres encore, à savoir :

2a. les *tautologies* ;
2b. la sous-classe des phrases parémiques concernées par des schémas rythmiques « particuliers », d'où la double sous-classification entre :
    2bi. *[proverbe]* [41], concept qui correspond aux proverbes métaphoriques (ex. *Une hirondelle ne fait pas le printemps*) ;
    2bii. *[adage]* et *[dicton]*, c'est-à-dire phrases parémiques non métaphoriques concernant des aspects moraux (*[adage]*) et naturels (*[dicton]*).

Même si Anscombre précise que :

[39] La reconnaissance des parémies peut se réaliser également à un moment donné de l'évolution linguistique. Un exemple nous est fourni par Buridant qui élabore un schéma des parémies telles qu'elles sont employées au Moyen Âge (2011 : 254).
[40] Nous remercions Jean-Claude Anscombre pour ses remarques ponctuelles ainsi que pour sa générosité.
[41] Les crochets marquent la valeur de classe conceptuelle. Nous aurions pu situer les nombreux travaux d'Anscombre, sans aucun doute, dans la macrocatégorie classification logico-sémantique. Nous l'avons inséré dans cette macrocatégorie en raison des retombées évidentes de sa proposition en ce qui concerne la distinction entre genres textuels gnomiques. On pourrait dire qu'il s'agit d'une proposition de classification ontologico-textuelle.

« [sa] classification correspond à des propriétés linguistiques dont des marques sont éventuellement visibles en surface » (Anscombre 2008b :18),

où la *phrase* (malgré tout) reste l'horizon de sa classification[42], il ne touche aux aspects morphosyntaxiques et de surface qu'en 2011.

1.2.2. *Classifications fonctionnelles (ou discursives)*

D'après Meschonnic, le proverbe relève du discours et, sans exception, il affirme que les classifications formelles des proverbes sont vouées à l'échec (1976 : 419). Certains parémiologues ont, pourtant, essayé de proposer des classifications *ad hoc* ou à visée générique qui prennent en compte les proverbes en discours, notamment en contexte écrit[43].

Permyakov encore suggère une classification d'après les 7 fonctions pragmatiques que les parémies peuvent satisfaire (1979 : 141). C'est la fonction *modelling* (modélisation) à être concernée et à rapprocher les wellerismes, les anecdotes et les proverbes et à faire la différence avec d'autres parémies. Comme l'explique Permyakov :

« this function provides a verbal (or thought) model (scheme) of some real-life (or logical) situation » (1979 : 141),

ce qui ne va pas sans conséquence au moment du repérage même des proverbes en discours.

Après Permyakov et afin de décrire les énoncés proverbiaux médiévaux du *Recueil général des Jeux-Partis français*, Buridant (1976), par exemple, opérationnalise des classifications préexistantes et aboutit à une classification personnelle et hybride de ces énoncés (1976 : 405). Il s'appuie sur les *traits d'identification* de Rodegem (§ 4.1.1.), des traits qu'il estime « à un premier étage de caractères propres » (1976 :393), intègre des *traits de spécification* repris par Greimas « à un second étage » ou « signes particuliers » des proverbes (*ibid*.), tels les temps verbaux et les structures rythmiques binaire et ternaire (et à l'exception des traits archaïques) et également inspiré de Paulhan (1993 [1925]), il ajoute des

[42] Dans sa contribution de 2013, il propose encore la notion de *forme sentencieuse* (§ 1.1.1.) et parle encore de *phrase* (cette fois-ci) *parémique descriptives* (2013 : 109-110), comme les tautologies, ou *prescriptives* (2013: 108-109), comme les adages.

[43] Comme le souhaitait déjà Greimas (1960), malheureusement, et malgré les progrès actuels autour du français parlé, des études systématiques consacrées à l'usage des proverbes français à l'oral manquent dans le panorama parémiologique.

traits de classification, c'est-à-dire des cadres sémantico-sémiotiques récurrents, comme l'implication.

Un autre exemple de classification en contexte narratif est à attribuer à Schulze-Busacker (1985). À partir d'un corpus d' œuvres en vers en ancien et moyen français, elle dégage une classification tripartite et sur trois niveaux (1985 : 25-35, 167-177). Au premier niveau, elle identifie trois macrocatégories de proverbes qui témoignent de trois emplois discursifs précis dans son corpus : les *proverbes intégrés*, les *proverbes cités* et les *proverbes exploités*. Au deuxième niveau, chaque macrocatégorie peut se réaliser dans (voire contribuer à réaliser) trois choix rhétoriques reconnaissables dans les textes (trois « formes de la narration » (1985 :169)) et qui constituent le pivot inchangé de toute la classification : le *discours direct*, le *récit* et la *digression*. C'est au troisième niveau que sa classification se raffine et intègre des critères formels morphosyntaxiques et lexicaux. À ce moment, comme elle l'affirme tout au long de son étude, sa classification présente des catégories douteuses, vu que l'identification des catégories et l'attribution d'occurrences proverbiales à ces catégories deviennent une affaire de jugement personnel, à savoir d'une rencontre de quelques repères encyclopédiques disponibles au moment de l'étude, ainsi que d'intuition.

1.2.3. *Classifications logico-sémantiques*

Lorsqu'en 1968 Permyakov présente sa classification universelle des proverbes, son universalité découle principalement de son approche logique au matériel parémiologique. La reconnaissance d'*archinvariants logico-sémiotiques*, à savoir de 4 propositions logiques-mères auxquelles tout l'acquis proverbial peut être ramené. On remarque que les 4 propositions logiques sont régies pour l'essentiel par une relation d'implication (que d'autres parémiologues comme Buridant (1976), Riegel (1986) et Kleiber (2000) discuteront en milieu francophone) et qui concerne des objets et des propriétés. Permyakov subdivise ces 4 propositions en 2 sections (1979 : 21-22). Une première section regroupe 2 propositions logiques de relation directe entre objets et propriétés :

I. $P(x) \rightarrow P(y)$: si un objet P possède une propriété x, alors l'objet P possède aussi une propriété y (ex. *Ce qui vient du diable retourne au diable, Tout homme est menteur*)[44] ;

[44] C'est nous qui choisissons les exemples pour les propositions logiques d'après la liste des formes canoniques de *DicAuPro* (Conenna *et al.* 2006).

II. *(P~Q) → [∃ (P) → ∃ (Q)]* : pour un objet P existe un objet Q en raison d'une relation entre ces objets (ex. *Après la pluie le beau temps*, *Pas de montagne sans vallée*).

Une deuxième section regroupe les 2 autres propositions logiques qui établissent des relations plus complexes entre objets et propriétés :

III. *(P→Q) → [P(x) → Q(x)]* : pour un objet Q qui dépend d'un objet P, la propriété x de l'objet P est aussi possédée par l'objet Q (ex. *Double jeûne, double morceau* ; *La goutte vient de la goutte*) ;
IV. *[P(x) ∧ Q(¬x)] → (P > Q)* : pour un objet P qui possède une propriété positive x qu'un objet Q ne possède pas, il en suit que l'objet P sera préféré à l'objet Q (ex. *Mieux vaut bonne attente que mauvaise hâte* ; *Un bon ami vaut mieux qu'un parent*).

De ces 4 ramifications logiques dérivent d'autres sous-catégories logico-sémiotiques plus spécifiques, à savoir 2 *types structuraux logico-sémiotiques* pour chaque invariant. Il s'agit de deux variantes logiques des 4 archinvariants qui sont équivalentes entre elles. À leur tour, ces types structuraux logico-sémiotiques ont au moins 2 *sous-types structuraux logico-sémiotiques* : la juxtaposition d'objets ou de propriétés qui s'opposent (ex. grand – petit) ; l'opposition entre objets et propriétés et leur manque (ex. parole – silence) (1979 : 22). La classification se raffine encore parce que tout sous-type se divise en *groupes logico-thématiques* (*ibid*.), mais à ce stade, à notre avis, les propositions logiques se lexicalisent et se rapprochent davantage des catégories thématiques qu'on reconnaît dans les recueils parémiographiques.

Loin des propositions logiques génératives de Permyakov, Kleiber attribue quand-même au proverbe le statut de « catégorie sémantique *en soi* » (Kleiber 2010a : 137), suivant idéalement la suggestion de Permyakov quant à la nécessité de traiter les parémies, en général, comme une classe linguistique à part entière. De par sa nature, le proverbe lui-même est une entité linguistique capable de jouer le rôle de catégorie dans une classification idéale des 'situations de la vie', un peu comme le proverbe-signe de Permyakov. Chaque proverbe-catégorie regroupe ainsi des occurrences linguistiques (et extralinguistiques) hétérogènes sous une image. En ce sens, il reprend la distinction entre proverbes métaphoriques et proverbes non métaphoriques, et esquisse ainsi des *sous-classes sémantiques* proverbiales par oppositions, comme les classes proverbiales représentant et/ou non représentant la situation à

laquelle elles font référence, ou les classes de proverbes à quantification universelle et/ou « partitive » (Kleiber 2010b).

1.2.4. *Classifications lexicographiques (ou thématiques)*

Dans une étude comparée de recueils parémiographiques espagnols et italiens, Alonso Pérez-Ávila (2008) distingue les techniques de rangement parémiographique en deux familles, suivant les deux approches traditionnelles lexicographiques. D'une part, elle sépare la *parémiographie sémasiologique* qui permet de classifier les parémies d'après l'ordre alphabétique et le mot-clé. D'autre part, elle regroupe sous la *parémiographie onomasiologique* toute tentative de classification ayant forme d'ontologie, à savoir de concepts génériques (ou complexes) et de thèmes en relation entre eux.

Les classifications thématiques représentent l'interface populaire préférée par les recueils parémiographiques (Meschonnic 1976 : 419). On a ainsi des collections de sagesse (souvent enrichies de maximes et d'aphorismes *incognito*) très génériques, par exemple, autour de l'amour, qui sont de préférence l'apanage de livrets à visée commerciale. C'est le type d'expérience de classification que Visetti & Cadiot mènent pour manifester leur contrariété aux classements proverbiaux. Pour montrer leur manque de fiabilité, ils essaient de classer une liste d'environ 80 proverbes d'après les 21 *topoï* et rubriques qu'ils identifient, intégrés par 16 autres complémentaires et/ou alternatifs (Visetti & Cadiot 2006 : 317-337). Les intitulés de ces rubriques peuvent être des séquences, d'où la classe des proverbes appelée :

Il ne faut pas se fier aux apparences

qui inclut, entre autres, les proverbes :

Une hirondelle ne fait pas le printemps
Tout ce qui brille n'est point or

ou des noms, comme pour la classe *Prudence/danger* qui sert à ranger les proverbes :

Il ne faut pas réveiller le chat qui dort
Il ne faut pas émouvoir les frelons

Il ne faut pas remuer l'ordure.

On ne peut s'empêcher de remarquer que la répétition du déontique *Il ne faut pas* est un indice formel, un « soutien de l'intuition » (Visetti & Cadiot 2006 : 318), dont ils se servent pour le regroupement (Visetti & Cadiot 2006 : 335) et auquel Arnaud avait déjà attribué le statut de marqueur de proverbialisation (Arnaud 1991 : 9). En guise de conclusion, il reconfirment leur propos, tout en affirmant que :

> « le genre proverbial, avec son jeu d'accentuations, de fusions et de dissociations entre aspects perceptifs, praxéologiques, évaluatifs, modaux et temporels, […] s'oppose au propos même d'une classification, pour autant que celle-ci devrait reposer sur des principes de discrimination stabilisés » (Visetti & Cadiot 2006 : 337).

Quant aux recueils parémiographiques à visée académique ou « pour les professionnels du proverbe », la classification thématique se peaufine suivant l'intuition sémantique et l'expérience directe du parémiologue, parfois au détriment d'une consultation rapide. À cet égard, nous mentionnons encore Permyakov, notamment la dernière catégorie de sa classification logico-sémiotique : les *groupes logico-thématiques* (1979 : 22-23). Nous avons écarté cette catégorie de l'approche logique que Permyakov revendique parce que son opérationnalisation satisfait davantage des critères sémantiques attachés aux unités lexicales qui composent les proverbes, plutôt qu'à la logique propositionnelle tout court. À ce stade et pour cette catégorie, le lexique (les étiquettes linguistiques des *realia*, des références aux objets (au sens large) du monde réel et abstrait) sert de conjonction idéale entre la catégorisation purement logique et la catégorisation syntactico-pragmatique. Pour appuyer notre choix, nous pourrions citer les exemples d'un groupe logico-thématique et ses proverbes correspondants suggérés par Permyakov. Il présente le groupe *Producer and His Goods* qui inclut, entre autres, les proverbes :

> *The shoemaker is the worst shod* [trad. litt. : Le cordonnier est le plus mal chaussé]
> *The tailor is without a shirt* [trad. litt. : Le couturier est sans maillot]

Il en suit que l'étiquette choisie pour le groupe identifié répond à la présence de deux noms de professions (*shoemaker*, *tailor*) et de leurs biens (implicite dans le nom de profession et dans le verbe *shod* pour le premier, explicite dans le nom *shirt* pour le deuxième). Les

groupes logico-thématiques font coexister des ensembles lexicaux (que d'autres nommeraient classes d'objets ou *lexical sets*) qui sont les matériaux à mettre en proverbe d'après un ou plusieurs archinvariants logiques [45]. Pourtant, Permyakov défend leur nature logico-sémiotique qui n'est pas ancrée aux *realia* et que ces groupes constituent une classification ontologique. Il affirme :

> « The thematic groups […] encompass the logical meaning of all the pairs of paremiological realia and fit in very neatly with the logico-semiotic classification of proverbial sayings » (1979 : 157).

Il ajoute que si les archinvariants logico-sémiotiques aident à classifier les relations entre les objets réels que les proverbes évoquent :

> « the thematic classification characterises the logical substance of the content by naming these objects » (1979 : 158).

À notre avis, c'est justement cette activité de dénomination qui fait de ces groupes logico-thématiques des ensembles (motivés, raisonnés et limités) d'unités lexicales. C'est en ce sens que cette partie de l'activité de classification des proverbes par Permyakov ressemble (et s'assemble) davantage aux approches en parémiographie onomasiologique. Une classification thématique qui ressemble à celle de Permyakov (et ce, sans aucun étonnement scientifique[46]) est aussi proposée par le parémiologue finnois Matti Kuusi et implémentée par Outi Lauhakangas dans la base de données *Matti Kuusi International Database of Proverbs*[47] (Mieder 2004 : 16-20).

Or, comme les expériences de Permyakov et Kuusi suggèrent, les aléas des thèmes d'un proverbe sont intrinsèquement liés non seulement au degré de subjectivité des parémiographes (Alonso Pérez-Ávila 2008 : 450), mais aussi au niveau de sens qu'ils interpellent. Le rappel du *sens compositionnel* et/ou du *sens formulaire* qui constituent le *sens proverbial* (Tamba 2000) dépend de son usage en contexte. Les recueils parémiographiques ne fournissent pas de distinctions entre ces deux niveaux de sens, vu qu'ils mentionnent

[45] D'autres exemples de groupes logico-thématiques sont, par exemple, '*Big and small*', '*The old and the new*', '*Action and reaction*' (1979 : 157).
[46] Dans la préface à l'édition anglaise de son ouvrage que nous avons consultée pour notre analyse, Permyakov remercie ouvertement Matti Kuusi pour ses remarques.
[47] La base de données est disponible pour une consultation gratuite à l'adresse : http://lauhakan.home.cern.ch/lauhakan/int/cerpint.html (dernière consultation : 20/10/2013).

rarement des contextes d'usage. Quand ils les mentionnent, comme dans le cas du *Refranero Multilingüe*[48], les contextes (en l'occurrence, souvent littéraires) peuvent faire référence autant au sens compositionnel qu'au sens formulaire. Ce qui témoigne des difficultés encore rencontrées quand on souhaite proposer une classification fiable en parémiographie sémasiologique.

1.2.5. *Classifications lexico-grammaticales*[49]

Quant aux classifications lexico-syntaxiques, il faut, d'abord, opérer une distinction. Certains parémiologues préfèrent adopter une vue sur le tout de l'objet-proverbe et, en général, sur le tout de l'objet-parémie. Autrement dit, les classifications de ces parémiologues ne vont pas en deçà de la *phrase* ou, si elles prennent en compte l'en deçà de la phrase, elles le projettent sur le tout de la phrase pour la qualifier (voir ce que dit Blanche-Benveniste à ce propos, § 1.1.1.). En revanche, d'autres parémiologues se concentrent aussi bien sur le tout que sur les composantes du tout parémique. Pour ces cas, les classifications proposées partent toujours du proverbe-phrase, mais pour décrire ce qui le compose et les relations entre ses parties.

Greimas (1960) met en relief les propriétés syntaxiques (mais aussi lexicales, sémantiques et prosodiques) du « système clos de signification » que représentent les proverbes et les dictons, sans esquisser une classification explicite, à l'exception près de l'identification de trois 'unités syntaxiques' qui décriraient ces parémies : la phrase, la proposition à verbe explicite et la proposition averbale.

Permyakov distingue les proverbes d'autres parémies en raison de deux critères : la structure linguistique externe phrastique fermée et la structure interne linguistique synthétique. La *phrase* est le premier paramètre formel (externe, donc) de discrimination. Il précise, pourtant, que la phrase fermée est un tout qui ne laisse pas de place à d'autres composantes. La 'fermeture' de la phrase proverbiale est en effet expliquée comme une fixité autonome en discours, une forme achevée de clichéisation qui ne subit pas d'autres influences par le co(n)texte où elle est introduite :

[48] La base de données est disponible pour une consultation gratuite à l'adresse : http://cvc.cervantes.es/lengua/refranero/Default.aspx (dernière consultation : 20/10/2013).
[49] Ce paragraphe est une reprise revue et augmentée que nous avons tirée de Marcon (à paraître).

« Sentences that are full clichés, i.e. those consisting only of stable elements and not subject to changes or additions in speech, are called closed » (1979 : 10)[50].

Les proverbes-phrases fermées sont autant des *phrases simples* que des *phrases composées* (1979 : 14). Pourtant, la sémantique et la pragmatique l'emportent au fur et à mesure de sa classification linguistique. À un certain passage, la sémantique entretient des relations plus étroites avec la syntaxe. Permyakov affirme que les proverbes sont à la fois des *phrases fermées génériques*, à savoir des phrases qui relatent des faits réguliers et des coutumes[51], et des *phrases fermées particulières*, c'est-à-dire des phrases qui décrivent un fait individuel ou exceptionnel (1979 : 11). Encore, la distinction sémantique s'appuie sur des critères syntaxiques, notamment la personne et les modes des temps verbaux de ces phrases. À un autre niveau, la sémantique lexicale (ou plutôt la somme sémantique des unités lexicales) devient le paramètre central de l'argumentation de Permyakov, sans plus de liens avec la syntaxe. Les proverbes à proprement parler sont toujours entrelacés avec une *image*[52], c'est-à-dire que les 'vrais' proverbes ont un sens non-compositionnel[53] (1979 : 12). Au-delà de la sémantique lexicale, la sémantique du prédicat intervient pour créer une autre distinction entre les proverbes qui présentent deux prédicats sémantiquement opposés entre eux et ceux qui n'ont pas cette opposition (1979 : 15). Il fait une distinction, donc, entre les proverbes sémantiquement binaires[54] et les autres. La polarité de la phrase (entre sémantique et pragmatique) est aussi un autre critère linguistique de description et de distinction : il y a ainsi des phrases-proverbes à l'*affirmative* et d'autres à la *négative* (1979 : 14). Chacune de ces polarités peut se réaliser en 3 autres modalités : modalité *déclarative*, modalité *impérative* et modalité *interrogative* (*ibid*.). Comme on peut observer, le souci d'exhaustivité qui motive le travail de classification linguistique de Permyakov opère une sorte de syncrétisme entre les branches de la linguistique (dont on a des exemples plus récents que Permyakov a appliqués

[50] À l'inverse, les *phrases ouvertes* permettent la variation et la modification de certaines composantes. Les phrases proverbiales sont, pour Permyakov, des phrases ouvertes.
[51] La généricité et l'autonomie des proverbes en tant que phrases sont largement discutées, en milieu francophone, par Anscombre et par Kleiber.
[52] En milieu francophone, Conenna & Kleiber (2002).
[53] Pour Permyakov, les proverbes à sens compositionnel correspondent aux aphorismes (1979 : 13). L'opacité sémantique est l'argument central de toutes les études sur le figement des proverbes, malgré les quelques divergences sur la nature du figement proverbial (entre autres, Anscombre (2003, 2005) et Conenna (1988, 1998)).
[54] Sur la binarité sémantique comme trait distinctif des proverbes : en milieu anglophone, Milner (1969) ; en milieu francophone, Anscombre (2000). Nous rappelons que sur la binarité sémantique et sur la structure quadripartite du proverbe, Milner fonde une taxonomie sémantique des proverbes ancrée sur la sémantique lexicale de leurs constituants.

aux proverbes et aux parémies) dans une perspective syntaxique traditionnelle et biaisée par l'unité *phrase-proposition*.

Ce syncrétisme débouche sur deux constats significatifs de Permyakov lui-même. Le premier concerne le nombre de types grammaticaux de proverbes qu'on peut dégager de sa classification, à savoir un total de 192 types syntaxiques, et ce, sans tenir compte d'autres sous-distinctions possibles (1979 : 15). Le deuxième est le plus éclairant sur la validité de son approche. Permyakov écrit :

> « **In practice, proverbs and proverbial phrases are represented by all possible types of sentences** » (1979 : 17).

Ce qui équivaut au fait que l'unité phrase-proposition peut classifier le niveau parémiologique de la langue, quoiqu'elle permette de mettre au point une taxinomie rigoureuse, comme Permyakov le précise quelques lignes après (*ibid*.).

Dans une perspective phrastico-propositionnelle, Gómez-Jordana Ferary (2003) s'inspire des travaux de Conenna (sans pour autant se situer dans le cadre du Lexique-Grammaire (cf. *infra*)) pour classifier les structures syntaxiques des proverbes espagnols et français. En ce qui concerne les proverbes français, elle identifie 13 structures communes de diverses natures linguistiques et superposables entre elles : *phrases*, *formes* ou éléments récurrents en début de proverbe (Gómez-Jordana Ferary 2003 : 72). Pour chaque structure-source, elle spécifie des structures syntaxiques correspondantes. Par exemple, la *structure* « phrase nominale » [sic] décrit les proverbes qui présentent, entre autres, les structures syntaxiques :

À N N
À petit saint petite offrande

ou :

Tel N Tel N
Tel père, tel fils (Gómez-Jordana Ferary 2003 : 76-77).

Quelques années plus tard (Gómez-Jordana Ferary 2012), encore inspirée par les études de Conenna, elle identifie ceux qu'elle appelle *moules syntaxiques* pour les proverbes français et

espagnols (2012 : 75-145). Elle énumère ainsi une liste de 7 moules phrastiques et propositionnelles, parfois avec des composantes pleinement lexicalisées :

« - Phrase averbale binaire à article zéro en position frontale : *Jeu de mains, jeu de vilains* ;
- Phrase canonique à article zéro ou à article défini singulier LE-EL : *L'enfer est pavé de bonnes intentions* ;
- Proposition relative en QUI/QUIEN sans antécédent : *Qui aime bien châtie bien* ;
- Phrase impersonnelle injonctive/impersonnelle comparative : *Il faut battre le fer quand il est chaud* ;
- Phrase comparative averbale binaire : *Tel père, tel fils* ;
- Proposition relative restrictive avec antécédent précédé de l'article zéro : *Femme qui prend, se vend ; femme qui donne s'abandonne* ;
- Phrase juxtaposée ou coordonnée avec verbe à l'impératif : *Dis-moi qui tu hantes et je te dirai qui tu es* » (2012 : 139-140).

Or, les structures et les moules sont le résultat d'une relecture de quelques études parémiologiques (plus en milieu francophone qu'en milieu hispanophone) confrontées à une liste de parémies qu'elle a établie (pour les détails sur sa liste, § 3.2.11.). Elle ne présente pas une classification lexico-syntaxique conséquente[55]. Par exemple, elle isole :

« *À* N, N
Aux innocents, les mains pleines
À cœur vaillant, rien d'impossible
À père avare, enfant prodigue »

mais elle ne s'intéresse ni à une description des N (groupes nominaux) ni à donner une liste de toutes les séquences *À N, N* qui appartiennent au « groupe de proverbes prépositionnels en Prédicat-Sujet » (2012 : 104), déjà « phrase nominale » en 2003. Elle ne traite pas non plus la relation entre les surfaces lexico-grammaticales des parémies et leurs moules[56].

Une classification précédente à celle de Gómez-Jordana Ferary, c'est-à-dire celle de Mejri (1997), se situe toujours dans une perspective phrastico-propositionnelle enrichie de

[55] Par souci de clarté, nous précisons que la classification des parémies françaises et espagnoles n'est pas le but ultime de son étude de 2012. Malgré cela, l'organisation des parémies sur base syntaxique constitue le passage initial de toutes ses observations successives sur les aspects sémantiques et discursifs.
[56] Par exemple, pour le cas de *À père avare, enfant prodigue*, faudrait-il le classer sous le moule « phrase comparative averbale binaire » ? Et surtout, où sont finis les « proverbes prépositionnels » qui ont été reconnus ?

remarques syntaxiques et lexicales. Sur la base de la productivité interne au répertoire parémiologique français, il identifie 9 classes (très semblables aux moules de Conenna et de Gómez-Jordana Férary) de par la récurrence de certains éléments lexicaux et syntaxiques :

- Structures binaires avec groupes nominaux (*GN*) (ex. *Tel arbre, tel fruit*) ou prépositionnels (*GP*) (ex. *À bon chat, bon rat*)
- Phrases binaires du type *GN (ne pas) être GN* et une sous-distinction entre :
    - *GN* sans déterminants et ellipse nominale (ex. *Femme avisée est toujours modérée*)
    - *GN* sans déterminants sans ellipse nominale (ex. *Partir, c'est un peu mourir*)
    - *GN* avec déterminants (ex. *L'usage est le tyran des langues*)
- Pronom *QUI* suivi de deux groupes verbaux *GV* (ex. *Qui a bu boira*)
- *GN FAIRE GN* (ex. *L'union fait la force*)
- Déontique initial *IL (NE) FAUT (PAS) suivi de l'infinitif* (ex. *Il ne faut pas jeter l'huile sur le feu*)
- *ON NE PEUT PAS suivi de l'infinitif* (ex. *On ne peut pas être et avoir été*)
- *VALOIR* (ex. *Mieux vaut tard que jamais*)
- *Forme impérative* (ex. *Ne remets au lendemain ce que tu peux faire le même jour*)
- *Subordonnée circonstancielle suivie d'une principale* (ex. *Quand le chat n'est pas là, les souris dansent*)

Comme on peut le constater, les classes décrivent parfois la totalité, parfois certains éléments linguistiques récurrents, parfois une caractéristique syntaxique des parémies. Qui plus est, cette classification fait coexister les unités *phrase-proposition* et *structure*. En même temps, elle continue à garder certains acquis de la littérature parémiologique, comme la binarité syntaxique.

Le souci microdescriptif des parémies et l'élaboration de classifications à partir de la description de leurs constituants, relèvent, pour le français et pour l'italien, des études de Conenna (1988 et ss.) dans le cadre du Lexique-Grammaire de M. Gross[57] et pour l'italien seulement, de l'analyse de Bessi au sein du projet *Atlante Paremiologico Italiano* dirigé par Franceschi (2004). Comme dans le cas de Mejri, ces classifications réunissent les unités *structure* et *phrase*. Toutefois, il vaut mieux détailler et distinguer.

[57] Nous signalons aussi des études en espagnol (Blanco *et al.* 1995), en grec (Tsaknaki 2006) et en portugais (Chacoto 2007). En dehors du Lexique-Grammaire, nous mentionnons Paisii (1973) pour les proverbes français.

Bessi fait recours à la notion traditionnelle de *phrase-proposition*. Elle décide *a priori* de privilégier les *proverbes monophrastiques* (ou plutôt monopropositionnels) italiens de l'*Atlante* et d'identifier ceux qu'elle appelle des *modules de macrostructure* (*moduli di macrostruttura*). Elle présente ainsi une classification par ordre des composantes internes aux proverbes monophrastiques, ainsi qu'une classification sur base quantitative de ces modules pour montrer leur productivité dans son échantillon (2004 : 135-246). Bessi utilise à la fois les fonctions logiques (S = sujet, D = complément direct, I = complément indirect en relation avec la valence du verbe, C = complément indirect circonstanciel[58]) et les parties du discours (d = déterminant, N = nom, V = verbe, etc.). Ainsi, dans le module de macrostructure SVD (le module le plus productif, décrivant 525 proverbes qui correspondent à 35% de son échantillon), elle décrit les proverbes comme suit :

dN V dN
L'abito non fa il monaco[59]

Il ne serait pas trop compliqué de rapprocher la description de Bessi aux travaux de Conenna. Ces derniers, pourtant, s'appuient sur la classification syntaxique des *phrases figées* établie par M. Gross (1982). Après avoir reconnu le caractère de « figement spécial » du proverbe, Conenna range les régularités des répertoires parémiologiques français et italiens par des *classes syntaxiques* (1988, 2004) qui correspondent à des *structures syntaxiques*. La *structure* sert donc d'outil métalinguistique classificatoire, alors que la *phrase* fonctionne comme cadre descriptif du proverbe, quoiqu'elle ne soit pas à identifier avec celle de la grammaire traditionnelle. Elle est plutôt à ramener à l'approche distributionnelle et transformationnelle de Z. S. Harris et, plus précisément, à la distribution d'*arguments* autour d'un *opérateur* (notamment le verbe) (Harris 1976) ainsi qu'à l'idée de Maurice Gross pour qui la *phrase simple* est « l'unité significative du lexique » (1986a : 299)[60]. Conenna dégage aussi des *moules syntaxiques* (1998b, 2000, 2002). En outre, elle essaie de systématiser le répertoire parémiographique par un expédient à la fois empirique et lexicographique :

[58] Comme elle le dit, C joue le rôle de « 'cornice' […] una premessa spazio-temporale, che fa le veci di una proposizione subordinata introduttiva » (Bessi 2004 : 67). D'après une lecture des modules-catégories et de leurs occurrences, nous avons remarqué que la distinction entre compléments indirects I et C entraîne une distinction entre proverbes et dictons météorologiques, ceux-ci relevant de tout module contenant un C. Une description syntaxique fine a donc mis en relief une propriété linguistique pour mieux opérer une 'classification textuelle' (§ 1.2.1.).
[59] La négation n'est pas décrite.
[60] En ce sens, la *phrase simple* de Gross se rapproche de l'ancienne *phrasis* (§ 1.1.1.).

l'identification d'une ou de plusieurs unités lexicales récurrentes en début de proverbe. On assiste à la description systématique – la systématicité étant le véritable point fort en vue du traitement automatisé – de la classe des proverbes qui commencent par *Qui* (1988, 1998) ou *Il faut* (2000), et à la reconnaissance, par exemple, du moule syntaxique :

Qui a N a N

pour les proverbes :

Qui a femme a noise
Qui a bon voisin a bon matin

où N représente n'importe quel groupe nominal[61].

Mogorrón Huerta et Navarro Brotons (2012) eux aussi ont adopté le cadre du Lexique-Grammaire et réalisé une étude comparée français-espagnol. Ils décrivent les structures syntaxiques de 110 proverbes français qui commencent par la préposition *À/A* (Mogorrón Huerta et Navarro Brotons 2012 : 462-464) et dégagent 7 structures-moules proverbiaux syntaxiques qu'ils répartissent dans deux macro-catégories implicitement phrastiques : les proverbes nominaux (ou averbaux) et les proverbes verbaux[62]. Pour les proverbes nominaux, ils observent que la structure syntaxique la plus fréquente (44 proverbes) correspond à :

À N_1 N_2
À bon chat bon rat
À père avare, enfant prodigue

où tout N_n représente un *GN*. Pour les proverbes verbaux, ils constatent que 36 proverbes de leur échantillon sont décrits par :

À N_1 N_0 V N_2
Au pays des aveugles les borgnes sont rois

[61] La classe *Quand*/*Quando* est analysée de manière comparée français-italien par Lacavalla (2007). Pour une étude originale sur la classe *Qui* et qui intègre la métrique de Benoît de Cornulier : D'Andrea (2007, 2008).
[62] On a l'impression que les auteurs superposent les notions de *phrase*, *structure* et *proverbe*.

Quoique précieuse et nécessaire, surtout pour les renseignements sur leur productivité interne, la description syntaxique générique délaisse le lexique, notamment les actualisations d'unités lexicales spécifiques et leur fréquence[63].

1.2.6. *Pour une classification lexico-grammaticale (autre) des proverbes français*

En gros, nous pourrions réorganiser nos macrocatégories en deux groupes de par la perception et le traitement réservés aux parémies. Les macrocatégories se répartissent en fonction de :

- la parémie mise en relation avec d'autres *textes* ;
- la parémie et son identité autonome, c'est-à-dire un tout linguistique qui se tient de par elle-même.

Dans le premier groupe, nous avons toutes les classifications typologiques qui sous-entendent une perspective textuelle vis-à-vis du tout linguistique parémique. C'est dans cette macrocatégorie qu'on identifie tous les néologismes qui se répandent pour étiqueter les parémies. Au premier groupe appartiennent aussi les classifications fonctionnelles qui s'appuient sur des corpus textuels pour motiver les emplois parémiques en co(n)texte. Il est évident que la liste des fonctions et la distinction des parémies suivant ces fonctions dépendent de la nature des textes consultés.

En revanche, le deuxième groupe réunit les trois autres macrocatégories de notre métaclassification. Les classifications logico-sémantiques interpellent l'autonomie parémique isolée d'un (con)texte. Dans ce cas, il est souvent lié à la *phrase* et à la *proposition* logique. Sur cette dernière ainsi que sur la sémantique lexicale et donc sur l'*unité lexicale* à elle seule se fondent plutôt les aléas des classifications lexicographiques, outre que sur le tri alphabétique. Pour finir, les classifications lexico-syntaxiques font recours encore à la *phrase* et à la *proposition* tout comme à la *structure*, et ce, pour mettre en évidence des ressemblances entre constituants lexicaux et syntaxiques à l'aide de classes formelles.

[63] La prise en compte du lexique aurait mieux décrit, à notre avis, les « 'moules proverbiaux' qui aident à la reconnaissance des proverbes et [à] mettre en relief leurs propriétés lexicales » (Mogorrón Huerta & Navarro Brotons 2012 : 467).

D'après ce que nous avons explicité au § 1.1.5., nous proposerons un autre modèle de classification lexico-grammaticale. Nous estimons que c'est par une description exhaustive et systématique de la surface lexico-grammaticale d'un échantillon de parémies qu'on peut mettre en place et motiver toute autre sorte de classification. Bien évidemment, par rapport aux tentatives que nous avons présentées, nous utiliserons notre *séquence lexico-grammaticale* (SLG) pour classifier environ 1.800 proverbes (§ 5.1.).

La conjonction des approches distributionnelle et contextualiste ainsi que certains principes généraux qui façonnent les études en linguistique émergente nous serviront à établir notre modèle de classification lexico-grammaticale. Pour chaque parémie, nous examinerons par étapes successives la cooccurrence et la récurrence des parties du discours et des unités lexicales. Nous nous appuierons sur la fréquence (désormais *f*) des parties du discours et des unités lexicales ainsi que sur leur solidarité, à savoir sur leurs relations de co-sélection grammaticales (*colligations*, Sinclair 2004 : 32) et lexicales (*collocations*, Sinclair 2004 : 28) (§ 2.2.6.).

Nous adopterons une analyse lexico-grammaticale détaillée pour l'établissement d'une classification empirique plutôt que sur les généralisations syntaxiques et sur les classifications que nous avons mentionnées. Il s'agira d'une classification partielle et biaisée par notre corpus parémique, mais elle nous mettra à l'abri de raccourcis descriptifs ancrés dans la littérature introspective.

L'analyse lexico-grammaticale permettra de concevoir un modèle de classification à double articulation lexicale et syntaxique. Chaque classe relevant des niveaux les plus proches de la description lexico-grammaticale de surface complète, héritera des propriétés lexicales et grammaticales des niveaux les plus éloignés de la description complète. En ligne générale, les niveaux de classification et leur numérotation croissante correspondront au nombre de constituants analysés pour chaque parémie. Cela veut dire que le niveau 1 de la classification concernera la description lexico-grammaticale du constituant en début de parémie, le niveau 2 intéressera la description lexico-grammaticale de la suite des deux premiers constituants de la parémie et ainsi de suite, jusqu'à épuisement de la description. À nouveau en règle générale, une classe lexico-grammaticale correspondra à une SLG dont *f* dans notre corpus parémique sera supérieure à (désormais >) 1. Autrement dit, une SLG

acquerra le statut de *classe lexico-grammaticale* du niveau *n* si et seulement si elle sera en mesure de décrire du moins 2 parémies de notre corpus[64].

Par conséquent, nous n'accorderons pas d'importance à la présence ou à l'absence d'une partie du discours et/ou d'une unité lexicale pour dégager des classes. En revanche, nous attribuerons une attention particulière au principe de la *linéarité*. Ce principe, accepté depuis Saussure, répond, d'une part, aux exigences de modélisation de nos requêtes informatiques qui sont unidirectionnelles et, au plus, récursives et, d'autre part, est lié aux exigences d'une classification lexicographique par tri initial (alphabétique, syntaxique, etc.). Le fait de commencer la description lexico-grammaticale des parémies par leur début satisfait en outre leur nature d'unités morphosyntaxiques et sémantiques – que l'usage réitère, coupe, démantèle ou reconstruit – et reproduit le processus cognitif du *chunking*. Il nous a semblé pertinent de considérer tous les constituants élémentaires – qu'ils soient considérés comme parties du discours ou unités lexicales – et de décrire l'enchaînement des SLG parémiques sans donner la priorité à une partie du discours ou à une unité lexicale comme nœud descriptif, à la manière de la Grammaire des Patrons (Hunston & Francis 2000 : 37)[65]. Ce qui facilitera une meilleure évaluation de la solidarité entre ses constituants et leurs fréquences de cooccurrence à l'intérieur du corpus parémique.

L'interface lexique-grammaire analysée sur la base du critère quantitatif permettra aussi d'identifier et de mesurer la productivité des ***séquences lexico-grammaticales actualisées*** (SLG-a) de matrice proverbiale autant dans notre corpus parémique que dans d'autres corpus de français contemporain (§ 6). Nous distinguerons les ***séquences lexico-grammaticales actualisées partielles*** (SLG-ap), c'est-à-dire les SLG qui décrivent une partie d'un nombre $n > 1$ de parémies, et les ***séquences lexico-grammaticales actualisées complètes*** (SLG-ac), à savoir toute SLG qui décrit la totalité d'un nombre $n > 1$ de parémies de notre corpus. Par cette distinction, nous pourrons prendre en compte quelques formes d'analogie créative et productive de matrice parémique. Nous décrirons par la suite toutes les étapes de notre démarche de classification (§ 5.3.).

[64] D'après Sinclair, « a language pattern – however defined – has to occur a minimum of twice » (2004 : 28). Il est évident que nous garderons aussi toutes les SLG dont $f = 1$ en vue de la modélisation des requêtes informatiques. Ces SLG non productives seront classées sous leurs SLG-sources et classes respectives.

[65] Nous précisons que la reconnaissance des patrons par Hunston & Francis privilégie les cooccurrents à droite, surtout pour décrire les complémentations verbales. En tout cas, l'observation contextuelle n'exclut pas les co-occurrents à gauche.

1.3. En guise de brève conclusion : la forme et l'usage avant toute chose

Pour conclure ce chapitre, nos propositions de *séquence lexico-grammaticale* comme unité minimale descriptive pour saisir la *Gestalt linguistique parémique* et notre modèle de classification lexico-grammaticale suivent un rejet momentané d'une bonne partie de l'acquis parémiologique que nous venons de décrire. En d'autres termes, nous suspendons tout jugement sur la pertinence et sur la validité des études fondées sur des unités descriptives peu fiables et sur des aspects non formels. Quoique ces études généralisent des comportements et des propriétés linguistiques attachés aux parémies, la plupart de leurs conclusions ne se basent pas sur l'évidence formelle de la parémie, mais sur une présomption de nature introspective. Par conséquent, nous retiendrons seulement les expériences et les analyses menées suivant une approche empirique qui ont creusé les parémies pour faire ressortir des régularités, des ressemblances et des particularités. En ce sens, notre recherche se veut une étude en *parémiologie empirique* (Grzybek 2009, § 2.1.).

Par rapport aux vulgates des parémiologues autour de la définition de *proverbe* et de *parémie*, nous défendrons la *catégorielle*. La parémie est, pour nous, une catégorie linguistique à part entière. Nous nous occuperons de décrire de manière exhaustive ses propriétés morphosyntaxiques et lexicales. Autrement dit, nous envisageons une ***grammaire lexico-grammaticale parémique***. D'autres propriétés de la *catégorie parémie* pourront être mises en relief par l'adoption d'autres points de vue et d'autres approches, mais sans ignorer l'évidence de la forme (ou plutôt des formes) autant au sein du répertoire parémiologique que dans l'usage. Outre que parémiologie empirique, donc, notre démarche nous introduit dans le filon de la *parémiologie linguistique* (Conenna 2000). Plus précisément et de par notre approche, nous suggérons l'étiquette de ***parémiologie linguistique basée sur l'usage***, c'est-à-dire une parémiologie qui généralise des tendances et des comportements linguistiques sur l'escorte de données linguistiques formelles et d'usage authentique (§ 2.).

CHAPITRE 2

PAREMIOLOGIE LINGUISTIQUE BASEE SUR L'USAGE

Le présent chapitre veut encadrer notre étude au sein de deux paradigmes de recherche :

1. l'un en parémiologie, c'est-à-dire la *parémiologie empirique* (§ 2.1) ;
2. l'autre en sciences du langage, à savoir la *linguistique de corpus* (§ 2.2).

Nous présenterons les traits saillants de chaque paradigme pour aboutir à la proposition finale d'un croisement entre les deux et concevoir ce que nous avons appelé *:* ***parémiologie linguistique basée sur l'usage*** (§ 2.3) dans le sillage de la *parémiologie linguistique* de Conenna (§ 2.1).

2.1. Parémiologie empirique

Ce sous-chapitre introduit les notions-clés qui délimitent le périmètre de la *parémiologie empirique*. Après une brève définition, nous nous concentrerons sur quelques aspects théoriques et terminologiques qui nous serviront pour mieux établir l'étendue de notre étude portant sur l'usage des parémies (§ 7). Nous présenterons également le recensement systématisé des recherches qu'il est possible d'effectuer en parémiologie empirique. Ce qui nous permettra de mieux situer notre travail et de proposer une nouvelle intersection entre la parémiologie empirique et notre approche linguistique.

2.1.1. *Essor de la parémiologie empirique*

C'est pour organiser une revue des études empiriques sur les proverbes que Grzybek & Chlosta (1993) introduisent l'étiquette de *empirischen sprichwortforschung* (*parémiologie empirique*). Peu de temps après, ils y reviennent dessus et, tout en souhaitant une rencontre (probablement inspirée de l'ouverture méthodologique de Permyakov[66]) entre l'approche folklorique et l'approche empirique aux proverbes, ils opèrent une distinction entre *empirical paremiology* et *empirical paremiography* (*parémiographie empirique*) (Chlosta & Grzybek 1995). Les deux parémiologues esquissent ainsi une définition de ces branches de la parémiologie :

> « Both empirical paremiology and empirical paremiography are characterized by the attempt to study contemporary proverb usage, to work with or to provide authentic material » (Chlosta & Grzybek 1995 : 82, c'est nous qui soulignons).

L'usage du proverbe dans un environnement sociolinguistique d'occurrence authentique constitue le but principal de la parémiologie et de la parémiographie abordées avec une approche empirique. Ce qui implique une attention particulière à la pluralité formelle de chaque occurrence :

> « […] empirical paremiology by no means denies proverbial varieties, but, on the contrary, attempts to structure the 'infinite' number of collected proverbs (and their variants) […] » (*ivi*, 68).

Autant la forme que la quantité ('infinie') de parémies : c'est cette variété que vise la parémiologie empirique. Ils ajoutent :

> « [...] the quantitative results obtained by empirical paremiology are only an intermediate objective; first and foremost, these results should be understood as a basis, which subsequent (paremiological, philological, linguistic, etc.) studies can take for their starting point » (*ibid*., c'est nous qui soulignons).

[66] Grzybek (1994) a ouvertement apprécié et défendu la richesse du traitement sémiotique que Permyakov a réservé aux parémies, tout en admettant qu'il laisse des questions à résoudre.

La parémiologie empirique ne suffit pas à elle-même ou, du moins, son autonomie disciplinaire et ses conclusions sont relatives et temporaires. Nous soulignons cette précarité des résultats quantitatifs (mais aussi qualitatifs) dans la mesure où ils sont fonction des observables utilisés ainsi que de la méthodologie adoptée pour la description et pour l'analyse des observables. D'ailleurs, cette précarité qui freine toute généralisation caractérise l'approche empirique en soi, et ce, même en dehors de la parémiologie. En linguistique, par exemple, le recours aux corpus assure des résultats à moyen terme et fortement entrelacés avec les observables. Rien n'empêche que ces résultats deviennent un repère à exploiter ou à réfuter dans le cadre d'études successives (§ 3.2.5.). En ce sens, il nous intéresse de remarquer au passage la distinction nette qui est opérée (plus ou moins explicitement) par Chlosta & Grzybek entre parémiologie et linguistique. Conenna neutralisera cet écart par la proposition d'une *parémiologie linguistique* :

> « Il est donc important de créer un domaine strictement linguistique dans les études sur les proverbes : en quelque sorte, une *parémiologie linguistique* ayant sa place à côté d'autres domaines reliés à l'anthropologie, à l'ethnographie, à la dialectologie, à la littérature, etc. » (Conenna 2000a : 28)[67].

Pour revenir à la proposition de Chlosta & Grzybek, leur première vue sur la parémiologie empirique est, pourtant, ancrée dans la notion de *familiarité*, c'est-à-dire au degré de connaissance qu'un échantillon donné d'une population possède vis-à-vis du répertoire parémiologique de telle ou telle autre langue. Par des enquêtes socio-démologiques et par des analyses statistiques, il faut enfermer les parémies contemporaines dans des listes qui permettent, par la suite, la mise à jour des recueils parémiographiques. En ce sens, on court le risque d'enregistrer une 'grande mortalité parémique', souvent mise en avant par les détracteurs de la parémiologie ainsi que par les parémiologues eux-mêmes, mais repoussée par d'autres (§ 3.2.1.). Malgré la disparition éventuelle des parémies du répertoire contemporain d'une langue (au moment de l'étude empirique, ajoutons-nous), Chlosta & Grzybek précisent que :

[67] Plus récemment, De Gioia (à paraître) a défendu l'idée d'englober la dialectologie au sein de la parémiologie linguistique.

> « […] none would ever eliminate them from the traditional proverb treasury – they remain important witnesses of a culture's proverbial richness, though witnesses of days passed by… » (1995 : 68).

Bref : le temps passe, les parémies restent, même quand elles ne sont plus courantes dans l'usage[68]. L'usage, soit-il estimé par des questionnaires ou par des occurrences dans des textes, n'est pas le tyran des parémies. Il est seulement à interpréter comme un indicateur de connaissance ou de récurrence.

Ces prolégomènes à la parémiologie empirique de Chlosta & Grzybek s'accompagnent à l'établissement d'une métaterminologie ponctuelle. C'est le cas, par exemple, de *corpus expérimental* (*experimental corpus*) que les parémiologues utilisent pour faire référence à la liste de parémies sélectionnées qu'on soumet à un échantillon d'une population donnée en vue d'une étude de familiarité (Chlosta & Grzybek 1995 : 69). Le corpus expérimental devrait, pour eux, contenir un nombre bien délimité de parémies potentiellement connues. On est bien loin ici de la notion de *corpus* en linguistique de corpus (§ 2.2.2.).

2.1.2. *Systématisation de la parémiologie empirique*

Quelques années plus tard, Grzybek (2009) et Grzybek & Chlosta (2009) élaborent davantage leur métaterminologie et proposent une nouvelle systématisation des études possibles et des méthodologies que l'on peut envisager en parémiologie empirique. S'inspirant de la distinction introduite par Čermak entre « **knowledge** of proverbs » et « actual **usage** of proverbs » (2007 [1998] : 569), ils séparent les notions de *familiarité* et de *fréquence* d'usage pour caractériser deux parcours de recherche distincts :

I. le parcours des études de familiarité « *knowledge-based* » et « *subject-dependent* » (Grzybek 2009 : 215-216 ; Grzybek & Chlosta 2009 : 96) ;
II. le parcours des études de fréquence « *text-based* » (*ibid*.).

Par cette distinction, il en suit donc une spécialisation métaterminologique :

[68] À ce propos, voir le choix opposé fait par Arnaud (§ 1.2.1.).

1. la notion et le terme de *familiarité* s'appliquent dans le cas d'études qui interpellent un échantillon donné d'une population de sujets à laquelle on s'intéresse pour déterminer leur degré de *connaissance* du répertoire parémiologique (ou d'une partie de ce répertoire). Au contraire,
2. la notion et le terme de *fréquence* sont à employer si et seulement si des études visent les occurrences d'éléments du répertoire parémiologique d'une langue dans des *textes*[69].

Cette distinction entraîne une réorganisation interne des approches. D'une part, les études de familiarité basées sur la connaissance par des sujets peuvent consister en :

Ia. la rédaction spontanée d'un nombre *p* de parémies de la part d'un nombre *s* de sujets, *p* relevant de la connaissance des sujets et d'aspects extralinguistiques ;

Ib. l'évaluation introspective de la part d'un nombre *s* de sujets quant à la connaissance ou à la méconnaissance d'une liste préétablie d'un nombre *p* de parémies. À ce propos, les auteurs qualifient cette technique de *full text presentation* (Grzybek 2009 : 217 ; Grzybek & Chlosta 2009 : 98) ;

Ic. l'évaluation introspective de la part d'un nombre *s* de sujets quant au degré de connaissance quantifié par une plage de valeurs chiffrées et concernant une liste préétablie d'un nombre *p* de parémies. Ce que les chercheurs nomment *full text rating* (*ibid*.) ;

Id. la reconnaissance de la part d'un nombre *s* de sujets d'un nombre *p* de parémies par la soumission de leur partie initiale. La reconnaissance et l'achèvement formel des parémies que comporte la *partial text presentation* (*ibid*.) sont ainsi des indicateurs de connaissance.

D'autre part, les études de fréquence d'usage dans des textes se distinguent en :

[69] Dans une revue sur la notion de fréquence en linguistique basée sur l'usage et, en général, en sciences du langage, Loiseau distingue entre *fréquence émique* (ou *fréquence intuitive*) et *fréquence étique* (ou *fréquence mesurée*). La première fait référence à la fréquence établie sur un jugement introspectif, alors que la deuxième résulte de l'application d'une méthode qui mesure un tel fait linguistique (Loiseau 2011 : 65). La fréquence émique correspond à la notion de familiarité, tandis que la fréquence étique à la notion de fréquence en parémiologie empirique. Pour éviter de poursuivre la confusion qui s'est produite en parémiologie, nous nous tiendrons à la distinction proposée par Grzybek & Chlosta, tout en étant conscient qu'une distinction pareille est également faite au sein des approches en linguistique basée sur l'usage.

IIa. études longitudinales d'occurrence des parémies pour un nombre *s* de sujets suivis au cours de leurs activités quotidiennes (Grzybek 2009 : 216 ; Grzybek & Chlosta 2009 : 97) ;

IIb. études des occurrences des parémies dans un nombre *r* de recueils parémiographiques (*ibid*.) ;

IIc. études des occurrences des parémies dans les médias, notamment dans la presse, ainsi que sur le Web (*ibid*.) ;

IId. études des occurrences des parémies basées sur *corpus* (dans l'acception du terme en linguistique de corpus (§ 2.2.2.)) (Grzybek 2009 : 217 ; Grzybek & Chlosta 2009 : 97).

Nous partageons toutes les réserves de Grzybek & Chlosta à l'égard de Ia – Id et de IIa – II d (*ibid*.). Nous soulignons aussi que ces réserves s'atténuent (mais ne disparaissent pas) à l'égard des études de fréquence sur corpus. Ils précisent que :

> « […] such analysis has the advantage of electronic search and retrieval strategies, provided one previously defines what one wants to search » (Grzybek & Chlosta 2009 : 97).

Les études sur corpus nécessitent qu'au préalable les parémiologues établissent le *comment* et surtout le *quoi parémique* de leur interrogation. L'avantage de l'interrogation électronique est, certes, évident au moment même de l'interrogation (le dépouillement électronique assure une vitesse décidément plus élevée d'un dépouillement manuel), non pas en termes de préparation du *comment* (§ 3-4) et du *quoi* (§ 5). L'avantage de l'interrogation électronique demande également une réflexion approfondie sur l'objet à interroger, c'est-à-dire sur la composition du corpus, tout comme sur l'interprétation quantitative et qualitative des résultats (§§ 2.2.4., 2.2.5., 2.2.6.).

En tout cas, la typologisation de Grzybek & Chlosta sert à mieux éclairer la nature des résultats qu'on peut obtenir en parémiologie empirique. On verra que les parémiologues essaient très souvent de faire rencontrer les notions de *fréquence* et de *familiarité*, et ce, pour cerner le *minimum parémiologique* (par émulation de Permyakov) d'une langue ou pour mesurer la *contemporanéité* (§ 3.2.2.) ou la *popularité* (§ 3.2.12.) d'un répertoire parémiographique.

2.2. Linguistique de corpus

Disposer et organiser des observables afin de décrire un fait linguistique : c'est la linguistique de corpus. Elle connaît ses débuts en plein générativisme chomskyien dans les années 1960-70, quand :

> « many early corpus linguists almost felt as if they had to work in secret cells »
> (Lindquist 2009 : 9)

pour éviter les critiques des structuralistes, au point que l'association anglophone la plus connue en linguistique de corpus (l'*International Computer Archive of Modern and Medieval English – ICAME*) voit le jour autour de la table de cuisine de Stig Johansson, un des pionniers de cette nouvelle approche aux langues (*ibid*.). Pour attendre son véritable essor, il faudra attendre la parution du livre *Corpus, Concordance, Collocation* de John McArthur Sinclair (1991) qui représente un fondement incontournable de la linguistique de corpus contemporaine.

Malgré sa diffusion à l'échelle mondiale, quelques questions se posent sur son compte :

1. faut-il considérer la linguistique de corpus seulement comme une méthodologie ? A-t-elle aussi le statut de branche de la linguistique ? (§ 2.2.1.)
2. Qu'est-ce que c'est qu'un corpus ? (§ 2.2.2.) Comment le concevoir ? (§§ 2.2.4. et 2.2.5.)
3. Quelle est la place des hypothèses en linguistique de corpus ? (§ 2.2.3.) Comment et pourquoi exploiter un corpus ? (§ 2.2.6.)

2.2.1. *Linguistique de corpus : entre méthodologie et théorie*

C'est pendant les années 1980 que la linguistique de corpus profite des progrès informatiques et des premiers pas en matière de traitement textuel automatisé pour se présenter comme une voie empirique alternative à l'introspection. Comme le fait entrevoir le titre d'une contribution de Fillmore (1992), la *linguistique de fauteuil* (*armchair linguistics*) s'est transformée en *linguistique de fauteuil assistée par l'ordinateur* (*computer-aided armchair linguistics*). Ainsi, les linguistes assis à leurs bureaux vérifient des hypothèses par

le support d'exemples réels, que ces hypothèses précèdent l'interrogation des textes ou qu'elles dérivent de l'exploration de ces derniers (§ 2.2.3.). L'expérience devance la théorie dépourvue d'ancrages empiriques et munie seulement de l'intuition.

La question de trancher sur l'identité de la linguistique de corpus comme méthodologie ou comme branche de la linguistique est loin de trouver une véritable solution (McEnery *et al*. 2006 : 7-8). Dans un chapitre consacré aux courants philosophiques et aux courants linguistiques qui ont contribué à l'établissement et à l'évolution de la linguistique de corpus, Stubbs (2007) offre un exemple de justification théorique. Par sa contribution, l'auteur comble le vide d'une trentaine d'années qui ont fait de la linguistique de corpus un terrain à exploiter sans contraintes d'aucune sorte.

Certes, l'impact de la linguistique de corpus et la révolution paradigmatique qu'elle a déclenchée dans les recherches linguistiques actuelles sont sans aucun doute évidents, et ce, par le nombre élevé de linguistes qui font recours à ces observables. L'approche par corpus se veut un moyen pour des formalisations au plan théorique (comme l'a fait récemment Hanks avec sa *Théorie des Normes et des Exploitations*), d'où il nous paraît raisonnable l'identification de la linguistique de corpus avec une méthodologie souple[70] et catalysatrice de nouvelles théorisations (§ 2.2.3.) ou de redéfinition des frontières d'un ou de plusieurs domaines (§ 2.3).

2.2.2. *Corpus*

Il n'y a pas une seule définition de *corpus*. Beaucoup de linguistes ont essayé d'en donner une définition satisfaisante. Ci-dessous, nous parcourons quelques-unes de ces définitions afin de circonscrire l'usage du terme *corpus*.

> Sinclair (1991 : 171) : « a collection of naturally-occurring language text, chosen to characterize a state or variety of language ».

> Francis (1992 : 7) : « a collection of texts assumed to be representative of a given language, dialect, or other subset of language, to be used for linguistic analysis ».

[70] À l'égard de cette souplesse, il suffit de considérer le pluriel « *Les linguistiques de corpus* » employé comme titre du manuel de Habert *et al.* (1997).

Leech (1992 : 116) : « […] computer corpora are rarely haphazard collections of textual material : they are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) representative of some language or text type ».

Sinclair (1995 : 17) : « a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language »[71].

Kennedy (1998 : 3-4): « the definition of a corpus as a collection of texts in electronic database can beg many questions for there are many different kinds of corpora. […] A corpus constitutes an empirical basis not only for identifying the elements and structural patterns which make up the systems we use in a language, but also for mapping out our use of these systems ».

Pearson (1998 : 43) : « there appears to be a consensus that a corpus is an artefact; it is selected, chosen or assembled according to explicit criteria. It is stored in electronic form. It consists of pieces of naturally occurring language ».

Rastier (2004) : « Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications ».

Condamines (2005 : 18) : « Il est désormais acquis qu'un corpus n'est pas un ensemble de données langagières en vrac, mais des données (en l'occurrence textuelles) qu'on décide de regrouper pour une étude particulière […]. Le corpus est ainsi à distinguer de la base de données textuelles [...] »

Sinclair (2005) : « A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as

[71] Dans leur manuel sur les linguistiques de corpus, Habert *et al.* (1997 : 6-7) font recours à cette définition.

possible, a language or language variety as a source of data for linguistic research ».

McEnery *et al.* (2006 : 5) : « There are many ways to define a corpus [...], but there is an increasing consensus that a corpus is a collection of (1) machine-readable (2) authentic texts [...] which is (3) sampled to be (4) representative of a particular language or language variety. While all scholars agree upon the first two qualities, there are differing opinions regarding what can be counted as representative ».

Baker *et al.* (2006 : 48) : « […] a corpus is a collection of texts (a 'body' of language) stored in an electronic database. Corpora are usually large bodies of machine-readable texts containing thousands or millions of words ».

Le terme *corpus* garde une aura de textualité qui est reconductible aux parties qui le composent. Au fur et à mesure, il faut admettre, toutefois, que les linguistes ont convergé vers des traits définitoires de plus en plus distinctifs : le *corpus* est (i) une collection (ii) de textes (iii) au format électronique (iv) en raison d'un choix (v) motivé par des critères. C'est autour de (iv) et (v), c'est-à-dire d'un choix raisonné, que la linguistique de corpus comme méthodologie souple (§ 2.1.1.) abandonne le singulier pour le pluriel, à savoir pour **les** linguistique**s** de corpus (Habert *et al*. 1997). Les logiques de ce choix peuvent être guidées par les finalités d'une étude linguistique, par les aspects à enquêter de tel ou tel autre fait ou objet linguistique ou tout simplement par l'intérêt de disposer d'observables linguistiques.

2.2.3. *Hypothèses et observables*

Tognini-Bonelli (2001) suggère une macrotypologie des études qu'il est possible d'envisager en linguistique de corpus. Elle distingue trois types d'études d'après la relation que les linguistes établissent entre corpus et questions/hypothèses de recherche. On peut :

1. exploiter un corpus comme un moyen pour confirmer ou pour réfuter une hypothèse conçue avant la constitution et l'interrogation d'un corpus (*corpus-based*) ;

2. se laisser emporter par les observables pour dégager et pour ramasser des indices afin de construire des hypothèses (*corpus-driven*) ;
3. utiliser un corpus pour l'apprentissage d'une langue, surtout de son lexique (*data driven*).

Dans le premier cas, elle parle d'*études basées sur corpus*, alors que dans le deuxième cas, elle parle d'*études guidées par corpus*. Dans le troisième cas, qui est le cas le plus proche de la didactique des langues, elle parle d'*apprentissage guidé par les données*.

Il arrive très souvent d'opposer nettement les études basées sur corpus et les études guidées par corpus. Les linguistes qui se penchent sur le corpus pour vérifier la fiabilité d'une hypothèse au sein d'une théorie linguistique, envisagent le corpus comme un appui pour la confirmation ou (plus rarement) pour la remise en question d'une théorie linguistique. En revanche, les linguistes qui se servent des corpus pour faire ressortir et interpréter des comportements linguistiques, considèrent le corpus comme un outil de fouille et de découverte pour la création ou pour l'esquisse de 'nouvelles' théories linguistiques. Autrement dit, soit l'hypothèse précède, soit elle suit les observables. Cette distinction est en tout cas assez fictive et abstraite. La pratique montre souvent leur rencontre fusionnelle et féconde : les observables améliorent ou réfutent des hypothèses et, souvent, en déclenchent d'autres.

2.2.4. *Critères de constitution d'un corpus*

Pour veiller à leur choix, les linguistes de corpus font recours à différents critères. Ci-dessous, nous en présentons quelques-uns parmi les plus utilisés et auxquels nous ferons référence pour le choix et la composition de nos corpus (§ 6).

2.2.4.1. Taille

De préférence, des linguistes se penchant sur des études ciblées n'utiliseront pas les mêmes corpus que des linguistes qui souhaitent 'prédire' des comportements linguistiques récurrents. Une distinction s'impose donc entre ***corpus échantillonnés*** (*sample corpora*) de taille restreinte (normalement, quelques millions d'occurrences), représentatifs et balancés par rapport à une variété de langue cernée (McEnery & Hardie 2012 : 8-9), d'une part, et les

corpus moniteurs (***monitor corpora***) ayant une taille d'envergure (au-delà des quelques centaines de millions d'occurrences) et constamment enrichis avec des textes variés (McEnery & Hardie 2012 : 6-7). En raison de leur augmentation durable, les corpus moniteurs s'appellent aussi *corpus ouverts ou dynamiques* (§ 2.2.4.2.).

Dans les deux cas se pose le problème de la généralisation des résultats obtenus. Autrement dit, quelle quantité d'observables permet de valider les résultats d'une étude sur corpus ? La réponse dépend autant du progrès des technologies en matière de stockage des données que de l'ouverture du corpus.

Si au début de l'approche par de corpus, une taille raisonnable était estimée aux alentours d'un million d'occurrences, à l'heure actuelle une taille satisfaisante pour une première ébauche d'étude est à estimer aux alentours d'une centaine de millions, jusqu'au cas limite des corpus collectés à l'aide des données textuelles sur le Web, comme le corpus *frWaC* (Baroni *et al*. 2009) d'environ 1 milliard et 600 millions d'occurrences, ou *Google Books Ngram* qui atteint une taille (toujours agrandissante) de quelques milliards d'occurrences (Michel *et al*. 2011).

Nous partageons l'avis de Sinclair quand il suggère qu'un corpus devrait augmenter de taille au fur et à mesure (1991 : 19). Pourtant, en généralisant un critère adopté par Pearson (1998 : 59), la quantité d'observables d'un corpus doit être le pendant d'un besoin qualitatif.

2.2.4.2. Temps

Le critère du temps peut être abordé sous deux points de vue. On peut s'intéresser au temps :

1. comme *contexte chronologique* de production des textes ou
2. comme *délai* qu'on se donne pour la collection des textes.

Pour le premier cas, sont ciblées la diachronie et la synchronie d'une langue ou d'une variété linguistique représentées par les textes choisis. Supposons, par exemple, qu'on s'attache à la variation et au changement d'agencements syntaxiques. comme la permutation entre l'adverbe *mieux* et le verbe *valoir* (Marcon 2012). Pour cette étude, le ***corpus*** doit être ***diachronique***, à savoir composé de textes dont la production se situe sur plusieurs points de

la ligne du temps qui appartiennent au passé, et ce, par rapport au moment de la collection du corpus. Autrement dit, le corpus rassemble les textes rédigés dans des siècles (anciens ou récents) le plus souvent éloignés, ou encore dans des décennies plus récentes relativement au moment de la collection du corpus afin d'apprécier des changements linguistiques tout au long du temps (McEnery & Hardie 2012 : 94-120). Par contre, supposons de décrire l'écriture des SMS belges francophones en 2004 (Fairon & Klein 2006) et que l'opération de collection se réalise approximativement à la même période. Dans ce cas, le ***corpus*** s'appelle ***synchronique***, c'est-à-dire qu'il se compose de textes produits et collectés à un point précis de la ligne du temps (Gries 2009 : 10), et servant moins à l'analyse de changements linguistiques, mais plutôt à la récurrence ainsi qu'à la régularité des agencements entre mots.

Pour le deuxième cas, prenons encore l'exemple des SMS. Dans le cadre du projet SMS4SCIENCE[72], la collection du corpus de SMS a poursuivi en Belgique, mais aussi en France, en Suisse et au Québec. Le corpus peut se dire ainsi ***ouvert ou dynamique*** parce que la collection continue au fur et à mesure. Il pourrait éventuellement devenir un corpus moniteur (§ 2.2.4.1.) de la variété 'langage SMS francophone', si l'on prévoyait des ajouts constants. Le corpus *Google Books Ngram* mentionné au § 2.2.4.1. est un autre exemple de corpus ouvert. Au contraire, l'œuvre complète d'un écrivain décédé est un exemple de corpus ***fermé ou statique*** (sauf si l'on découvre des épreuves de romans posthumes…). Les corpus statiques ont donc une taille fixe (parfois restreinte, aussi) et correspondent souvent à des corpus échantillonnés (§ 2.2.4.1.) (Gries 2009 : 11).

Dans les études sur l'acquisition d'une langue (notamment étrangère), la variable *temps* permet aussi de distinguer les ***corpus longitudinaux***, c'est-à-dire des collections de textes qui tracent le parcours d'apprentissage de la langue d'un ou de plusieurs individus au fur et à mesure que le temps passe (Granger 2002 : 11).

2.2.4.3. Médium

La linguistique de corpus anglophone s'est davantage distinguée pour la collection de ***corpus écrits*** : le tout premier *BROWN Corpus* dans les années 1960 (Hunston 2002 : 15), *The Bank of English* (BoE) débutant dans les années 1980 comme *COBUILD Corpus* par la volonté de Sinclair ainsi que le *Corpus of Contemporary American English* (COCA)[73] en sont

[72] La présentation du projet est disponible à l'adresse suivante : http://www.sms4science.org (date de consultation : 10/11/2013).
[73] Le corpus est consultable gratuitement ici : http://corpus.byu.edu/coca/ (date de consultation : 10/11/2013).

des exemples d'envergure (McEnery & Hardie 2012 : 6-7). Par contre, la linguistique de corpus francophone s'est plutôt concentrée sur l'oral, comme le montrent l'*Inventaire des* ***corpus oraux*** (en ligne en 2005-2006) et la toute récente *Base de données des corpus oraux de français hors de France* mis au point par la *Délégation générale à la langue française et aux langues de France*[74], tout compte tenu de l'exception notable qu'est *Frantext* en évolution constante[75]. Malgré les soucis techniques liés, entre autres, aux protocoles de transcription, la collection de textes oraux continue à susciter l'intérêt des linguistes français (Bruxelles *et al.* 2009). La naissance (tardive par rapport au monde anglophone) du *Consortium Corpus Écrits* de l'*Institut de Linguistique Française* en 2013 en est une preuve[76], quoique des centres de recherche se soient engagés dans la collecte de corpus écrits avant sa création (entre autres, le LIDILEM de Grenoble et le corpus *Scientext*[77]).

En outre, l'oralité a subi une forte stigmatisation quant à son statut social face à l'écriture. La dichotomie *écriture / formalité* vs. *oralité / informalité* a été déjà remise en cause (Koch & Oesterreicher 2001) et elle est décidément bouleversée par le ***Web*** 2.0, notamment par l'écrit numérique de la messagerie instantanée, des blogs, des réseaux sociaux, des wikis (§ 2.2.5.2.), mais aussi par les interactions mobiles, notamment par l'écriture de SMS et par d'autres apps.

Il faut en tout cas souligner que le défi le plus récent en linguistique de corpus est représenté par la multimodalité : les ***corpus multimodaux*** qui intègrent image et parole, par exemple, commencent à faire leur parution, notamment pour l'analyse simultanée de la gestuelle et de la parole (McEnery & Hardie 2012 : 5).

[74] L'inventaire et la base de données sont consultables à l'adresse suivante : http://www.dglflf.culture.gouv.fr/ (> *Études et recherches* ; date de consultation : 10/11/2013).
[75] La documentation concernant *Frantext* est disponible à l'adresse : http://www.frantext.fr/ (date de consultation : 10/11/2013).
[76] La présentation du Consortium et les activités de ses groupes de travail sont disponibles sur le site : http://corpusecrits.corpus-ir.fr/ (date de consultation : 10/11/2013). Un projet pareil d'envergure européenne est représenté par le *Expert Advisory Group on Language Engineering Standards* (EAGLES) et par son évolution en *International Standards for Language Engineering* (ISLE). Ces projets ont élaboré des normes et des standards pour l'utilisation des ressources linguistiques en traitement automatique du langage, y compris pour la définition, la typologisation et la composition des corpus qui datent du 1996.
[77] Le corpus *Scientext* et sa présentation sont à repérer à l'adresse : http://scientext.msh-alpes.fr/ (date de consultation : 10/11/2013).

2.2.4.4. Langue

Très souvent les corpus sont ***monolingues***, mais ils peuvent être aussi ***bilingues*** ou ***multilingues***. La présence de plusieurs langues peut être gérée de deux façons :

1. soit par la sélection de textes relevant d'un même genre textuel et d'un même sujet pour chaque langue ;
2. soit par la sélection de textes dans une langue source et de leurs traductions respectives dans une (ou plusieurs) langue(s) cible(s).

En général, les premiers sont appelés ***corpus comparables*** alors que les deuxièmes ***corpus parallèles*** (McEnery & Hardie 2012 : 18-21), ces derniers nommés ***corpus alignés*** (L'Homme 2004 : 131-134) parce que les segments textuels (normalement, des phrases graphiques ou des paragraphes) des textes en langue source sont mis en relation avec leurs correspondants en langue source. Il est évident que la constitution de corpus bi- ou multilingue représente un enjeu majeur pour les études en traductologie, auxquelles Laviosa (2002) a donné la contribution la plus complète, notamment en ce qui concerne la redéfinition des universaux de traduction et l'enseignement de la traduction à l'aide de corpus bilingues. Dans le cadre de la Traduction Assistée par Odinateur (TAO) et de la Traduction Automatique (TA), les mémoires de traduction nécessitent également la constitution de corpus multiingues alignés (Baker & Saldanha 2008 : 48, 163). La collection de corpus bi- et multilingues, surtout satisfait encore l'intérêt des lexicographes et des terminographes qui souhaitent extraire automatiquement des lexiques et des terminologies multilingues (L'Homme 2004 : 207-210).

Nous verrons par la suite (§ 6.1.2.) qu'on peut dépasser cette typologie par le non-respect de la séparation entre langues comme critère de collection des corpus multilingues. Nous montrerons donc que, lorsque le but de l'étude réfléchit sur des sujets autres que la comparaison entre langues, il est possible de réunir des textes appartenant à plusieurs langues dans un seul corpus.

2.2.4.5. Discours

La distinction entre *langue générale* et *langue de spécialité* ou *spécialisée* (Cabré 1998 ; Lerat 1995) et, plus récemment, celle entre *discours général* et *discours de spécialité* ou *spécialisée* (Charaudeau et Maingueneau (2002), s.v. 'spécialité (discours de – / langue de –)').

Étant donné cette distinction, en linguistique de corpus il est possible de différencier les ***corpus généraux*** et les ***corpus spécialisés*** (Bowker & Pearson 2002 : 11-12). Les premiers représentent une langue en un nombre *n* de discours, pour autant que possible, différents entre eux. D'habitude, les corpus généraux sont aussi des corpus moniteurs (§ 2.2.4.1.). En revanche, les deuxièmes servent à étudier des faits linguistiques dans des textes qui relèvent d'un seul discours. Ce discours peut décrire une variété linguistique, une communauté de professionnels et de spécialistes ou une communauté démographique.

Parmi les corpus spécialisés qui, entre autres, sont très utilisés en terminologie textuelle (Bourigault & Slodzian 1999, L'Homme 2004 : 123-129), on peut inclure aussi les ***corpus d'apprenants*** (Granger 2002), c'est-à-dire les collections de textes élaborés par les apprenants d'une langue étrangère, ainsi que les ***corpus d'enseignants*** (O'Keeffe *et al.* 2007 : 220-221), à savoir des corpus longitudinaux (§ 2.2.4.2.) qui enregistrent les interactions enseignants-élèves et/ou enseignants-enseignants.

Cependant, la nature des uns et des autres s'appuie aussi sur les *genres* textuels qui forment les corpus (§ 2.2.4.6.).

2.2.4.6. Genre

Dès son essor en littérature, la notion de *genre* a intéressé plusieurs branches de la linguistique. Les linguistes Kerbrat-Orecchioni et Traverso soulignent que :

> « on peut difficilement décrire une interaction quelconque sans prendre en compte le genre dont elle relève, les genres étant définis comme *des catégories abstraites qui regroupent, sur la base d'un certain nombre de critères, des unités empiriques se présentant sous forme de "textes" ou de "discours"*. » (2003).

On constate ainsi que la notion de *genre* caractérise toute unité de n'importe quel médium, son but étant la classification et la catégorisation d'ensembles plus ou moins homogènes de productions linguistiques achevées. Du côté de la terminologie textuelle (Bourigault & Slodzian 1999) qui fait des corpus son point de départ, la notion de *genre* permet de comprendre la situation de production de tel ou tel autre texte pour mieux gérer la variation terminologique (Aussenac-Gilles & Condamines 2009) ainsi que pour l'établissement de ressources terminographiques adaptées (Aussenac-Gilles *et al*. 2002). Du côté de l'analyse des discours oraux, Kerbrat-Orecchioni & Traverso nous rappellent :

> « qu'il existe deux sortes de genres […] : (1) G1 : catégories de textes plus ou moins institutionnalisées dans une société donnée. […] (2) G2 : « types » plus abstraits de discours caractérisés par certains traits de nature rhétorico-pragmatique, ou relevant de leur organisation discursive. Ainsi un guide touristique serait-il un « genre » constitué de différents « types », les genres typologiquement purs étant en tout état de cause rares, voire inexistants » (Kerbrat-Orecchioni & Traverso 2004 : 41-42).

Par cet aperçu, il paraît évident que la notion de *genre* fait elle aussi l'objet de tiraillements au sein de la linguistique. La distinction G1/G2 proposée par Kerbrat-Orecchioni & Traverso est fusionnée par Aussenac-Gilles & Condamines, pour qui la notion de *genre* recouvre à la fois celle de catégorie textuelle codifiée et celle de texte/discours (écrit, oral, etc.) avec une macro-organisation (comme la progression thématique) et une micro-organisation (comme les choix lexicaux et syntaxiques) internes caractérisantes. C'est à cette définition inclusive de *genre* que nous ferons référence dans notre étude, tout en sachant qu'elle pourrait rencontrer des détracteurs.

2.2.5. *Constitution d'un corpus*

Au fur et à mesure de leur exploitation en linguistique, les corpus sont de plus en plus simples à repérer. Cela dit, il est parfois difficile de repérer des bonnes sources.

2.2.5.1. Corpus prêts à l'emploi

Entre autres, le répertoire de David Lee[78], la liste de diffusion *Corpora List*[79] ainsi que les initiatives mentionnées *infra*, sont des repères pour la recherche de corpus prêts à l'emploi. Parfois des restrictions de confidentialité, commerciales ou liées aux droits d'auteur empêchent l'accès direct aux données. Pour pallier ces restrictions, au fur et à mesure les centres de recherche académiques mettent à la disposition des corpus pour leur consultation ou leur téléchargement gratuits. C'est le cas, entre autres, du *Centre National de Ressources Textuelles et Lexicales*[80], le site REDAC du laboratoire CLEE-ERSS de l'Université de Toulouse II-Le Mirail[81] ou la *Leipzig Corpora Collection* mise au point par le Département d'Informatique de l'Université de Leipzig (§ 6.1.1.).

2.2.5.2. Le Web comme corpus

Pour éviter ces restrictions, le Web est une solution (presque) évidente. Le grand réseau virtuel couvre la (plupart de la) planète, et se diversifie sans cesse. Des textes sont disponibles à presque chaque page. Peut-on vraiment considérer le Web comme corpus ? D'un côté, Sinclair affirme que :

> « *The World WideWeb is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective*» (2005).

D'un autre, après avoir défini le Web comme « *immense, free and available by mouse-click* », caractérisé par « *hundreds of billions of words of text* », Kilgarriff & Grefenstette tranchent nettement sur l'affaire :

> « *The answer to the question "is the web a corpus ?" is yes* » (Kilgarriff & Grefenstette 2003 : 1 -2).

[78] Le répertoire se trouve à l'adresse : http://www.uow.edu.au/~dlee/CBLLinks.htm (date de consultation : 10/11/2013).
[79] Les archives sont à l'adresse : http://www.hit.uib.no/corpora/ (date de consultation : 10/11/2013).
[80] Les ressources du CNRTL sont disponibles sur le site : http://cnrtl.fr/ (date de consultation : 10/11/2013).
[81] Les corpus développés par le CLEE-ERSS sont à repérer : http://redac.univ-tlse2.fr/ (date de consultation : 10/11/2013).

Les positions de Sinclair et de Kilgarriff & Grefenstette exemplifient les points de vue opposés sur le Web conçu comme corpus. D'une part, nous partageons l'avis de Sinclair dans la mesure où il affirme à juste titre que le Web évolue et se diversifie indépendamment d'une architecture linguistique préétablie. L'éclat du *Web 2.0* (O'Reilly 2005), notamment des réseaux sociaux, des wikis, des blogs, tout comme de l'audiovisuel met au centre le partage de données qui relèvent de plusieurs codes sémiotiques (verbal, visuel, auditif, etc.) sans contraintes linguistiques. D'autre part, on ne peut nier l'appui que la consultation de masses de données textuelles apportent à la mise au point d'hypothèses linguistiques. C'est grâce à la présence de masses textuelles en ligne que nous pouvons reconduire l'existence d'une interface comme *WebCorp* (Renouf *et al.* 2007). *WebCorp* exploite l'indexation des pages Web faite par des moteurs de recherche (*Bing*, *Google* etc.) pour appliquer des critères linguistiques à un ensemble de données confus.

2.2.5.3. Le Web pour les corpus

Le Web peut représenter une source d'où puiser des textes. Des logiciels (*webcrawler*s) facilitent l'aspiration de données linguistiques (texte brut) des sites Web. Dans ce cas, le Web est considéré comme une matière première à retravailler pour un corpus-produit final. C'est dans cette vision que s'inscrivent des logiciels comme *GlossaNet 2* (Fairon *et al.* 2008)[82] et *BootCaT* (Baroni & Bernardini 2004) intégré au *Sketch Engine* (Kilgarriff *et al.* 2004) qui cernent, respectivement, a priori et a posteriori les sites Web à exploiter. *GlossaNet* interroge une liste fermée de flux RSS des sites Web (notamment, de presse en ligne, mais personnalisable) pour collecter les textes bruts. Le logiciel renvoie donc les résultats d'une recherche qu'un utilisateur introduit sous forme d'expression régulière ou de graphe (§ 4). *WebBootCat* fait recours aux moteurs de recherche pour repérer les sites les plus pertinents aux *seeds* (suites de mots-clés) qu'un utilisateur soumet. Les résultats proposés sont par la suite filtrés manuellement par l'utilisateur, d'après ses finalités et ses besoins.

[82] Le logiciel se trouve à l'adresse : http://glossa.fltr.ucl.ac.be/ (date de consultation : 10/11/2013). *GlossaNet 3* sera bientôt mis en ligne.

2.2.5.4. Étiquetage morphosyntaxique et lemmatisation

Par *annotation*, on indique toutes les opérations (manuelles et/ou automatiques) d'enrichissement d'un corpus par d'autres informations, souvent de nature linguistique. Comme le précisent (Habert *et al*. 1997 : 15-36 ; 169-174), par étiquetage on entend l'opération d'attribution (manuelle ou, le plus souvent, automatique) de la partie du discours et du lemme à chaque mot graphique.

Les étiqueteurs adoptent des *jeux d'étiquettes*, c'est-à-dire des ensembles de catégories à attribuer à chaque mot graphique. Ces catégories relèvent du degré de précision que l'on envisage pour une description ou analyse linguistique. Il peut donc y avoir des jeux d'étiquettes minutieux, comme ceux morphosyntaxiques des dictionnaires électroniques au format DELA (Courtois 1994-1995), ou moins spécifiques, comme ceux de l'étiqueteur automatique *TreeTagger* (Schmid 1994) (§ 5). Il faut être conscient des jeux d'étiquettes utilisés et des perspectives linguistiques (et informatiques) qui ont guidé leur conception. De même, surtout quand on emploie des étiqueteurs automatiques, il faut veiller à la correction éventuelle d'annotations erronées ou non pertinentes par rapport à l'approche linguistique adoptée.

Dans des ***corpus annotés*** qui s'opposent ainsi aux ***corpus non annotés*** (*raw corpora*) (Gries 2009 : 9), il est possible de cibler des séquences de mots graphiques non seulement d'une forme de surface exacte, mais aussi des séquences plus abstraites (*motifs* ou *patrons*) portant sur des informations linguistiques qui ont préalablement enrichi le corpus.

2.2.5.5. Numérisation et documentation

Dans les cas où l'on décide de ne pas exploiter un corpus prêt à l'emploi ou le Web comme ou pour les corpus et qu'on se tourne vers des textes au format papier, ces derniers nécessitent de transcriptions, d'opérations de reconnaissance optique, etc. pour en permettre leur conversion au format électronique. L'importance de la numérisation repose non seulement sur la notion même de *corpus*, mais aussi sur l'application d'une série de traitements textuels automatiques comme, entre autres, les concordances (§ 2.2.6.).

Une fois le corpus compilé, il est aussi indispensable de fournir une documentation détaillée sur sa composition. La transparence à l'égard de la composition ainsi que de toute

sorte de (pré)traitement (Sinclair 1991 : 13) sert à mieux évaluer les résultats d'une étude basée sur corpus ainsi qu'à favoriser la comparaison avec d'autres études sur le même sujet.

2.2.6. *Approche phraséologique sur corpus : concordances, cooccurrences*

D'après Gries, il y a trois méthodes pour aborder les données issues d'un corpus : i) se concentrer sur les listes de fréquence des mots ; ii) observer les cooccurrences lexicales ; iii) focaliser l'attention sur les cooccurrences lexico-grammaticales à l'aide des concordances (Gries 2009 : 12-18). Dans les paragraphes suivants, nous approfondirons respectivement iii) et ii).

L'établissement des concordances remonte déjà au XIII^{e} siècle pour faciliter les études bibliques, notamment leur exégèse. Peu à peu, elles se sont répandues à d'autres domaines comme forme d'édition du texte. Quelques siècles plus tard, le développement informatique a vulgarisé la technique *Key Words In Context* (KWIC) supportée par l'ordinateur. C'est l'ordinateur qui crée une série verticale de lignes de texte où un mot (*nœud*) est situé au milieu dans une fenêtre donnée de lecture à droite et à gauche (Hunston 2002 : 39). Tout en se démontrant un des outils privilégiés d'exploration des corpus, la longueur d'une concordance ne répond pas vraiment à des normes. Elle peut correspondre à une *phrase graphique*, la dépasser ou y rester en deçà.

La lecture des concordances varie aussi, et ce, d'après les faits et les objets linguistiques auxquels on s'intéresse. Outre le tri alphabétique des mots à droite ou à gauche d'un nœud, les concordances facilitent l'observation de la *typicité* (*typicality*) d'un comportement linguistique (Hunston 2002 : 42). En l'occurrence, il est possible de constater le voisinage entre mots graphiques ainsi que de calculer leur fréquence d'occurrence. Cette approche contextualiste de lecture a l'avantage de mettre en relief comment les mots interagissent entre eux. Ce qui permet de décrire à la fois les comportements centraux d'une unité lexicale ainsi que ces agencements syntaxiques préférés. Autrement dit, la lecture des concordances révèle des combinatoires lexico-grammaticales privilégiées.

Ce n'est donc pas par hasard si l'impulsion majeure à examiner la combinatoire des mots et, par conséquent, au développement des études en phraséologie, vient de l'observation directe et des méthodes statistiques qu'ont offertes les corpus. Sinclair l'explicite :

> « Phraseology is the ideal point of contact between a corpus and a description, because it accepts surface phenomena, and this, initially, is what a corpus provides [...] » (Sinclair 2008 : xvi).

Sinclair explique également que l'approche phraséologique représente la manière la plus holistique d'aborder le lexique (2008 : xv). Il souligne :

> « One of the great strengths of a phraseological approach is the preservation of the integrity of text for much longer than alternative approaches to description, and in turn this entails the preservation of meaning. Eventually the analytic process of abstraction, generalisation and formalisation will replace the text with an intralinguistic representation, but the specific details of the text will be able to exert maximum effect on the description » (Sinclair 2008 : xvii - xviii).

L'approche phraséologique sur corpus garde au maximum les *cooccurrences* entre les mots. La démarche de formalisation ne devient qu'une étape subséquente pour une appréhension globale des données.

Dans sa contribution *The Search for Units of Meaning* (Sinclair 2004 [1996]), Sinclair propose une taxinomie (quoiqu'il ne la présente pas explicitement comme telle) des cooccurrences qu'on peut repérer dans des corpus. Les *extended units of meaning* (*unités étendues de sens*) constituent la 'nouvelle devise' de la description linguistique (Tognini-Bonelli 2001 : 18) et représentent une référence incontournable pour les études phraséologiques basées sur l'usage. La taxinomie prévoit :

- la *collocation* ou la cooccurrence récurrente de mots (graphiques) entre eux ;
- la *colligation*, c'est-à-dire la cooccurrence récurrente entre mots graphiques en tant qu'unités qui relèvent de telle ou telle partie du discours ;
- la *préférence sémantique*, à savoir la cooccurrence récurrente entre mots graphiques en raison du partage et/ou de la création partagée d'un trait sémantique ;
- la *prosodie sémantique* ou la portée communicative et pragmatique d'une cooccurrence récurrente, cette portée n'étant pas nécessairement corrélée avec la préférence sémantique, mais plutôt avec l'intention du locuteur/énonciateur.

Cette taxinomie explore et s'appuie sur les observables-textes, comme le dit le titre de l'ouvrage d'où nous l'avons tirée (*Trust the text*), pour décrire le continuum qui part de la surface lexico-grammaticale et va vers l'intention sémantico-pragmatique du locuteur/énonciateur (ou, bien entendu, le chemin à l'inverse). Le lexico-centrisme de cette taxinomie sinclairienne peut se rapprocher en partie de l'approche harrisienne, comme le reconnaît, d'ailleurs, Sinclair lui-même (2004 [1996] : 25), et nous osons également le rapprocher de cette continuation de l'approche harrisienne (quoique davantage syntactico-centrique) qu'est représentée par le Lexique-Grammaire de Maurice Gross. D'où, d'ailleurs, notre proposition d'une rencontre entre les deux approches.

En milieu francophone (et d'autres, aussi), on a maintenant tendance à emprunter la taxinomie sinclairienne. Avant son utilisation 'systématique', Habert *et al.* fait référence simplement aux *cooccurrences* en général et aux *collocations*, mais ils amplifient cette liste du côté de la surface lexico-grammaticale. Ils parlent ainsi de *segments répétés*, à savoir :

> « toute suite d'unités textuelles reproduite sans variation à plusieurs endroits d'un corpus. Le nombre des unités qui composent le segment est sa *longueur* » (1997 : 200)[83].

et les *quasi-segments* (*répétés*), c'est-à-dire des suites d'unités reprises partiellement et qui englobent, entre autres, la notion de *motif* par Longrée & Mellet :

> « De manière strictement formelle, un motif se définit par l'association récurrente de *n* éléments du texte muni de sa structure linéaire […], laquelle donne une pertinence aux relations de successivité et de contiguïté […]. Ainsi, si le texte est formé d'un certain nombre d'occurrences des éléments A, B, C, D, E, un motif pourra être la micro-structure[84] récurrente ACD ou bien encore AA, etc., sans que l'on préjuge ici de la nature des éléments A, B, C, D, E en question » (2013 : 66, ce sont les auteurs qui soulignent).

2.3. Pour une parémiologie linguistique basée sur l'usage

Grâce à l'encadrement de Grzybek & Chlosta, nous pouvons qualifier notre étude comme un exemple de recherche en parémiologie empirique. Plus précisément, leur

[83] Les segments répétés sont souvent appelés *clusters* ou *lexical bundles* en milieu anglophone.
[84] Nous avons souligné le rapprochement entre la phraséologie et la notion de *structure* au § 1.

systématisation typologique nous a permis de faire de l'ordre dans la littérature parémiologique et de mettre de côté toutes les études qui ont visé la *familiarité* des parémies, et ce, malgré le fait qu'elles emploient la notion de *fréquence* (de manière impropre par rapport à ce que nous avons expliqué au § 2.1). Nous garderons la distinction familiarité/fréquence et concentrerons seulement sur la fréquence [85]. L'attention à la fréquence des parémies, notamment sur corpus, pourra, premièrement, combler un vide actuel en parémiologie française : le manque d'une liste de fréquence d'occurrence des parémies françaises [86]. Deuxièmement, le fait de cibler la seule fréquence fournira des données quantitatives (réfutables, fiables, incomplètes, satisfaisantes, etc.) qui pourront contribuer à mieux comprendre la relation qui s'établit entre familiarité et fréquence, notamment dans le cadre d'un minimum parémiologique français (Zouogbo 2011) ou de la contemporanéité/popularité des parémies.

Compte tenu de notre approche empirique, nous revendiquons une attention particulière au « strictement linguistique ». Pour cette raison, notre étude relève de la *parémiologie empirique* qu'on peut qualifier de *linguistique*, dans le sillage de la *parémiologie linguistique* cernée par Conenna (§ 2.1). Or, il nous faut encore préciser que l'approche qui caractérise notre description lexico-grammaticale et notre attention systématique aux observables comme on le fait en linguistique de corpus (§ 2.2.), nous permet de tracer un parcours en parémiologie qui est à la fois empirique, linguistique et basée sur la preuve de textes authentiques. C'est pour cela qu'on peut véritablement définir notre recherche comme un exemple de ***parémiologie linguistique basée sur l'usage***. Ce sous-domaine hybride se situe ainsi à la croisée de la parémiologie empirique, d'une part, et de la parémiologie linguistique, de l'autre.

Une fois de plus, donc, la linguistique de corpus conçue comme méthodologie invite à une nouvelle réflexion théorique (§ 2.2.1.), en l'occurrence, en parémiologie. Nous serons supporté par les faits observés dans des *corpus* (§ 2.2.2.). Les critères de constitution et les différentes typologies de corpus que nous avons présentés aux § 2.2.4. et 2.2.5., nous serviront de repère, d'une part, pour mieux encadrer les expériences sur corpus par d'autres parémiologues (§ 3). D'autre part, ils nous aideront à choisir et concevoir nos corpus (§ 6.1.). Plus précisément, nous viserons :

[85] Nous mettrons en évidence les mérites et les failles de certaines études de fréquence d'occurrence des parémies sur corpus dans notre recensement au § 3.
[86] Excepté le cas tout particulier et isolé d'Arnaud & Moon (1993) que nous décrirons au § 3., nous n'avons pas repéré une liste complète et basée sur corpus pour les proverbes et, en général, pour les parémies en français.

- la *taille* : sur l'escorte des conclusions qui découleront des expériences d'autres parémiologues, nous essayerons d'identifier et de constituer le(s) corpus dont le nombre d'occurrences soi(en)t le plus approprié par rapport à la finalité ainsi qu'aux limites informatiques des logiciels et de l'ordinateur que nous allons exploiter. En même temps, la taille des corpus utilisés ou conçus par d'autres parémiologues nous servira pour mieux établir et interpréter les seuils de fréquence d'occurrences (§ 3.3.3.) ;
- le *temps* : vu que nos requêtes informatiques ne modéliseront que des proverbes d'une variété de français contemporain, nous éviterons des corpus diachroniques. Nous nous concentrerons au plus sur des corpus qui visent la micro-diachronie, à savoir une période de temps relativement circonscrite, et des corpus fermés[87] ;
- le *médium* : malgré le réservoir à corpus oraux pour le français, nous avons décidé de nous concentrer sur le médium le plus étudié par les expériences que nous analyserons au § 3. D'ailleurs, le choix du médium influence la modélisation des requêtes informatiques (§ 4.) ;
- la *langue* : il est évident que notre (ou nos) corpus privilégiera (ou privilégieront) le français. Malgré cela, nous montrerons que, sans avoir l'obligation de qualifier un corpus de bilingue et dépassant donc la typologie normalement proposée en linguistique de corpus, il est possible de fusionner (du moins) deux langues. Ce qui sera motivé, lors de la constitution d'un corpus, par la centralité qu'acquerra un autre critère (celui du discours), sans que cela préjuge la finalité de notre étude ;
- le *discours* : nous essayerons de choisir et de constituer un (ou des) corpus qui représente(nt) un (ou plusieurs) discours spécialisé(s) ou plutôt plusieurs discours. Nous veillerons à faire un choix à la fois :
    - qui facilite en principe une comparaison avec les résultats relevés dans la littérature disponible en la matière et
    - qui s'inspire des points que nous estimons comme forts et faibles des études auxquelles nous ferons référence.
- les *genres* : c'est du choix concernant le discours qui dépendra la liste des genres textuels à prévoir. Quoiqu'il arrive, nous nous tiendrons aux étiquettes qui codifieront

[87] Nous avons déjà réalisé des études sur corpus dynamiques à l'aide de *GlossaNet 2* (Marcon 2011, 2013). Nous avons décidé de privilégier, pour l'instant, les résultats quantitatifs obtenus par des corpus fermés. Nous reviendrons sur les résultats de nos études précédentes pour effectuer des comparaisons (§ 7).

ces textes, sans remettre en question leur dénomination sur la base de leur structuration interne.

En raison de cette forte caractérisation de la notion de *corpus*, nous préférons rejeter le terme *corpus expérimental* suggéré par Grzybek & Chlosta (§ 2.1.). Pour éviter une polysémie risquée, nous indiquerons l'ensemble des parémies à décrire et à rechercher par le terme ***liste***. Nous indiquerons les critères pour la mise au point de cette liste au § 3.3.1.

Après avoir établi notre (ou nos) corpus et notre liste, il nous faudra également définir comment aborder la lecture des concordances. Plus précisément, nous déciderons si la *phrase graphique* pourra représenter une longueur textuelle suffisante pour aborder les occurrences des proverbes et si elle nous servira pour mieux comprendre la nature linguistique du proverbe et de la parémie.

Il est aussi évident que notre *séquence lexico-grammaticale* (§ 1.1.5.), est ouvertement débitrice de l'approche phraséologique basée sur corpus et de la taxinomie sinclairienne (§ 2.2.6.) dans la mesure où elle rassemble la *collocation* et la *colligation* en une seule unité formelle. Différemment de l'approche contextualiste sinclairienne et de manière tout à fait complémentaire à celle-ci, nous attribuons à la séquence lexico-grammaticale la capacité de classifier et d'ordonner le répertoire parémiologique français du point de vue lexico-grammatical (§ 5.3.)[88].

[88] À ce propos, on peut objecter que le terme *patron* qui caractérise les *unités étendues de sens* sinclairiennes, nous mettrait à l'abri de tout néologisme. Cela n'est vrai qu'en partie. La notion de *patron* permet de réunir nos besoins linguistiques et notre approche informatique d'un seul coup. Cette réunion, pourtant, crée de l'ambiguïté, voire une superposition qu'on peut aisément observer dans la littérature ainsi que dans les ouvrages de lexicographie spécialisée. Par exemple, le glossaire en linguistique de corpus par Baker *et al.* (2006), s.v. '*pattern*' écrit : « see **regular expression** » (§ 4.2.1.). Outre l'ambiguïté, le *patron linguistique* tout court fait de plus en plus l'objet d'une variation au sein de la communauté linguistique. D'après le parcours critique de Legallois & François, Saussure parlait déjà de *patrons réguliers* pour souligner la schématicité de la *parole* (2011 : 12). De Ferdinand de Saussure, la notion de *pattern* a eu son évolution en milieu anglophone, aussi (Hanks 2008) et constitue maintenant une unité constitutive de la Grammaire des Patrons (le patron ayant ici une identité linguistique à part entière) et de la Grammaire des Constructions (Legallois & François 2006). Malgré cette variation, un trait commun de ces interprétations du *patron* tout comme de la notion de *segment* que nous avons mentionnée (§ 2.2.6.), consiste en la reconnaissance d'une suite ou *séquence* continue d'éléments grammaticaux et lexicaux. D'où notre *séquence lexico-grammaticale* qui veut décrire les parémies, notamment (i) leur totalité et (ii) leur linéarité lexico-grammaticale. Toutefois, (i) et (ii) ne sont pas toujours des caractéristiques attribuées au *patron*. (i) et (ii) sont plutôt à repérer dans les schémas/cadres lexico-grammaticaux phraséologiques que nous avons signalés dans la note 25 au § 1.1.4. De plus, s'il est vrai que les patrons aident à organiser, par exemple, la microstructure de l'entrée d'un dictionnaire, il est vrai aussi qu'ils ne sont pas utilisés pour classifier le lexique de la même façon que nous l'envisageons dans notre étude.

CHAPITRE 3

FREQUENCE ET PAREMIES.
UN CADRE METHODOLOGIQUE FEDERATEUR

Dans ce chapitre, nous présenterons quelques études parémiologiques qui ont délibérément visé l'établissement d'une liste de fréquence des parémies dans des textes, soient-ils des corpus (§ 2), des recueils parémiographiques ou d'autres sources. Pareillement, nous analyserons des études parémiologiques qui s'appuient sur des données quantitatives obtenues par l'interrogation de corpus, quoiqu'elles motivent des conclusions de nature qualitative. Pour certains cas, nous avons également considéré des études phraséologiques où le questionnement sur la fréquence des parémies intègre un cadre de réflexion et de description qui intéresse les faits linguistiques les plus divers. Suivant la distinction que nous avons expliquée aux § 2.1 et § 2.3., nous avons exclu toutes les études de *familiarité*, excepté une seule étude pour son intérêt métaterminologique (§ 3.1.3.).

Au fil de notre revue, nous nous focaliserons sur :

1. l'établissement (éventuellement préalable) des *listes* de parémies ;
2. les *sources* utilisées et leurs caractéristiques ;
3. la *fréquence* d'usage des parémies dans un corpus, ses procédés de calcul, l'établissement des seuils et des plages de valeurs ainsi que leur interprétation.

Nous avons essayé de tirer parti du nombre le plus élevé d'expériences qui ont abordé la fréquence des parémies et leur usage dans des textes écrits, et ce, au fur et à mesure qu'elles ont été publiées à partir de la deuxième moitié du XX^{e} siècle. Nous avons estimé d'accorder une attention particulière à la littérature parémiologique francophone et, à même titre, de thésauriser les connaissances ainsi que les pratiques telles qu'elles sont exposées dans les

littératures parémiologiques non francophones[89]. Comme on le verra au § 4, nous avons distingué les fréquences d'usage calculées d'après le dépouillement manuel de sources au format papier, d'une part, et celles obtenues par l'interrogation de sources au format électronique, d'autre part. L'ordre chronologique de présentation montrera que la fréquence d'usage – qui est aussi influencée par les techniques de reconnaissance automatique choisies et disponibles – pose aux parémiologues les mêmes interrogatifs et entraîne des réponses similaires à chaque période. Au bout de notre parcours, nous élaborerons une synthèse critique en vue d'harmoniser le cadre méthodologique et de proposer des critères communs pour mener des études de fréquence d'usage sur corpus (§ 3.3.). Sur cette harmonisation méthodologique, nous fonderons le restant de notre recherche.

Avant de commencer ce parcours critique, il vaut mieux rappeler quelques avertissements que nous avons précisés dans l'Introduction. Comme nous croiserons plusieurs approches, il est évident que la notion de ce qu'on nommera par *proverbe* ou *parémie* variera. Nous utiliserons les deux termes tout en respectant les acceptions d'un parémiologue à l'autre. Vu le débat traductologique en parémiologie, nous ne donnerons que des traductions littérales en français des proverbes et des parémies en d'autres langues. Il faut considérer ces traductions[90] comme des aides à la compréhension, sauf dans les cas où les parémiologues en suggèreraient un équivalent formel et sémantique.

[89] Pour autant que possible, nous avons essayé d'atteindre une exhaustivité maximale, tout en admettant que certaines périodes et langues sont inévitablement restées à côté de notre analyse. De même, dans la plupart des études présentées, nous avons accédé directement aux sources primaires. Pour les rares cas où la consultation directe des sources s'est avérée impossible, nous avons cité des sources secondaires qui font référence à ces études.

[90] Les traductions littérales sont précédées par l'abréviation 'trad. litt.' dans le corps du texte et entre crochets.

3.1. Fréquence et format papier

Les paragraphes suivants exposent 4 études de fréquence (ou du moins de traitement quantitatif) des proverbes et, en général, des parémies, ainsi que des traits formels qui les caractérisent et/ou des fonctions discursives attachées à leur usage au discours écrit au format papier. Nous aurions pu mentionner des nombreux cas portant surtout sur des œuvres littéraires françaises[91]. Comme notre recherche concerne des corpus au format électronique (§ 3.2.), nous avons opéré une sélection de ces études 'sur papier'. Nous nous sommes donc concentré sur les travaux de Kuusi (1953) (§ 3.1.1.), Buridant (1976) (§ 3.1.2.), Rodegem (1984) (§ 3.1.3.) et Schulze-Busacker (1985) (§ 3.1.4.). Ces études fournissent des éléments d'ancrage avec la tradition parémiologique ainsi que des éléments novateurs et d'ouverture au traitement informatique des données linguistiques. Leurs réflexions, leurs suggestions et leurs techniques semblent confirmer que certains questionnements théoriques et méthodologiques, tout comme certaines astuces empiriques, se perpétuent d'un parémiologue à l'autre, plus ou moins sciemment.

3.1.1. *Kuusi* (1953)

C'est en dehors de la parémiologie française que nous avons repéré une toute première expérience. Il s'agit d'une étude par Matti Kuusi sur la *popularity* (*popularité*) (terme repris par la suite par Mieder (§ 3.2.2.) et par Grzybek (§ 3.2.12.)) de certains proverbes du répertoire parémiologique finnois. L'attention au caractère (quantitativement) partagé est, de par le chercheur, une *conditio sine qua non* du proverbe en soi :

> « Proverbs are common sayings among the people ; commonness is their state of being » (1998 [1953] : 24, c'est nous qui soulignons).

Toutefois, comme :

> « Proverbs do not […] come accompanied by records of their commonness, as do the plants of a school botany » (*ibid*.)

[91] Pour un relevé des articles en français sur des analyses (entre autres) linguistiques et littéraires de 1800 à 2008 : Gutiérrez Sánchez (2008).

il essaie de combler ce vide quantitatif, et ce, au fur et à mesure des siècles, notamment pour la période qui va du début du XIX^{e} jusqu'au début du XX^{e} siècle. Au fond de cette quantification se trouve une raison psycho-sociale :

> « if we were able to discern the strengths that have lifted […] to the peak of proverbial popularity […], we should know much about the entire direction of the development or the Finnish psyche, general taste and style of speech and cultural history over a period of centuries » (1998 [1953] : 28-29).

Ce genre d'ambition à l'égard de la mesure culturelle et éthique d'un peuple par les proverbes sera partagée aussi par d'autres parémiologues (voir Lau au § 3.2.6.)[92].

Liste

Kuusi démarre son enquête sur la popularité sans une liste de proverbes.

Sources

L'établissement de la popularité des proverbes se fait autour de leur répétition dans des manuscrits, des recueils et des collections parémiographiques, et ce, autour de deux dates. Le choix des dates est motivé par un fait historique : l'incendie de la ville de Turku en 1827 (1998 [1953] : 25) qui a provoqué la perte d'une grande partie du patrimoine écrit ancien finnois. Cette année représente ainsi l'*année-pivot*[93] pour estimer la première vingtaine de proverbes à retenir comme les plus populaires. De manière spéculaire, c'est l'année 1930 qui est choisie comme *année-contraste*[94]. Plus précisément, pour les proverbes avant 1827, Kuusi se concentre sur les « *primary sources* » (*ibid*.) dont nous ne disposons pas de précisions supplémentaires. Les autres proverbes sont tirés d'œuvres ainsi que d'une cinquantaine de collections parémiographiques venant de différentes régions finnoises et recueillies avant l'année-contraste 1930 (*ibid*.).

[92] La quantification de la culture d'un peuple de par les mots et les textes est une tendance très à la mode et relancée par la *culturomics* (culturomique) grâce au développement de logiciels ad hoc, comme *Google Ngram Viewer* (Michel *et al*. 2011). Pour une application de *Google Ngram Viewer* en parémiologie : Marcon (2012).
[93] C'est nous qui utilisons cette étiquette.
[94] C'est nous qui utilisons cette étiquette.

De toute façon, Kuusi est bien conscient des limites de sa recherche, surtout en ce qui concerne le calcul de la fréquence des proverbes avant 1827. Outre le nombre (inévitablement) réduit des sources consultables, il prend en compte aussi la contrainte socio-éducative, à savoir le fait que le goût et les préférences des copistes ont joué un rôle d'inclusion et d'exclusion de certains proverbes (1998 [1953] : 34-35). Au contraire, Kuusi défend le choix des collections consultées pour le calcul de la fréquence avant 1930, leur but étant un enregistrement exhaustif de toute occurrence proverbiale (*ibid*.).

Fréquence

La popularité calculée n'est qu'une *fréquence parémiographique*[95], non pas une fréquence d'usage[96]. Nous reprenons ci-dessous les cas des trois premiers proverbes (sur un total de 20) tirés de la liste de fréquence pour les sources avant 1827 et les trois premiers proverbes (toujours sur un total de 20) de la liste pour les sources avant 1930 avec leurs fréquences respectives dans les deux périodes prises en examen[97] :

[95] C'est nous qui utilisons cette étiquette pour décrire la fréquence calculée par la seule consultation de ressources parémiographiques.
[96] Les aspects statistiques seront traités de la même façon lors de la mise au point du recueil *Proverbia septentrionalia*.
[97] Pour les traductions littérales, nous nous appuyons sur les traductions littérales anglaises telles qu'elles sont proposées par McKenna, traducteur du finnois vers l'anglais de la contribution de Kuusi. McKenna suggère aussi des proverbes équivalents anglais.

Rang	Proverbe [traduction littérale]	*f* lexicographique (1827)	*f* lexicographique (1930)
1)	*Sanasta sana tulevi,* *kipinästä maa kytevi* [D'un mot vient un mot, d'une étincelle, la terre prend feu]	23	0
2)	*Kaunis kakku päältä nähden, vaan on sirkkoja sisällä, akanoita alla kuoren* [Le gâteau est bien au-dessus, mais il y a des grillons à l'intérieur, de l'ivraie sous la croûte]	22	8
3)	*Sanasta miestä sarvesta härkää* [Prends un homme par sa parole, prends un taureau par ses cornes]	21	14

Tableau 3.Fréquence parémiographique des 3 premiers proverbes les plus fréquents dans les recueils parémiographiques avant 1827 d'après Kuusi (1998 [1953]).

Rang	Proverbe [traduction littérale]	*f* brute (1827)	*f* brute (1930)
1)	*Lisäna rikka rokassa hāmāhäkki taikinassa* [Un peu de mauvaises herbes dans la soupe, araignées dans la pâte]	10	23
2)	*Mies tulee räkänenāstä,* *vaan ei tyhjän naurajasta* [Un homme vient d'un pleurnicheur, non pas d'un railleur frivole]	10	23
3)	*Ei haukku haavaa tee,* *jos ei koira purra saa* [Les aboiements ne font pas de blessures, si le chien ne mord pas]	6	19

Tableau 4. Fréquence parémiographique des 3 premiers proverbes les plus fréquents dans les recueils parémiographiques finnois entre 1827 et 1930 d'après Kuusi (1998 [1953]).

De ces données quantitatives, Kuusi dégage des tendances qualitatives et culturelles, comme le changement des images proverbiales dû à une mutation des coutumes et des habitudes sociales (1998 [1953] : 31-32) ainsi que la propension à l'humour et à une attitude

moins conservatrice et contraignante, mais plutôt conciliante et bienveillante sur les hommes et ses actions (1998 [1953] : 32-33). À côté de ces remarques, Kuusi constate aussi des tendances linguistiques formelles, à savoir l'éloignement du mètre qui relève de la tradition épique finnoise (1998 [1953] : 29), le sous-emploi de l'allitération (*ibid*.) et, au niveau syntaxique, la composition phrastique de plus en plus élaborée (1998 [1953] : 30-31).

Du point de vue purement quantitatif, nous soulignons que les 40 proverbes dépassent largement le seuil de fréquence (désormais *f*) de 10 occurrences. Il n'y a pas vraiment d'écarts significatifs entre les proverbes aux premiers rangs (ceux des Tableaux 1-2) et les autres. Ce qui implique que le total des sources dépouillées est réduit. Nous remarquons encore que la moitié des 40 proverbes (10 pour la première liste, 11 pour la deuxième) ont $f =$ 14 et 15. Ce regroupement autour d'une ou de plusieurs valeurs sera également observée pour d'autres études que nous analyserons dans les pages qui suivent.

Pour finir avec une réflexion méthodologique de portée générale, nous citons les mots de Kuusi en conclusion de son enquête :

> « It is […] my belief that the analysis of individual proverbs can provide only an incomplete explanation of why their popularity increases or declines. With the aid of general comparisons, […] one can track really significant developmental tendencies and regularities » (1998 [1953] : 37-38).

Une vue d'ensemble des proverbes, sans trop de biais au départ, permettrait de mieux comprendre les tendances d'évolution (en macrodiachronie) des proverbes ainsi que leurs régularités en termes quantitatifs et qualitatifs, et ce, grâce à la consultation de textes. On pourrait repérer dans ces mots les prolégomènes à une parémiologie empirique basée sur les observables plutôt que sur l'introspection.

3.1.2. *Buridant* (1976)

Parmi les études francophones visant les « traces proverbiales » (J. Cerquiglini & B. Cerquiglini 1976 : 374), celle de Buridant constitue un point de départ idéal. Son travail représente un exemple de fusion entre tradition et modernité en ce qui concerne le traitement linguistique des proverbes. Il s'agit, à notre avis, d'une étude-manifeste qui touche la plupart des aspects théoriques et méthodologiques de notre recherche.

Liste

Le questionnement de Buridant concerne directement le proverbe en tant qu'objet d'étude, notamment sa définition et son repérage :

> « Que faut-il, en effet, considérer comme proverbe dans ce corpus ? Quelles en sont les limites ? » (1976 : 378).

Buridant se montre critique à l'égard de tous ces parémiologues qui, au moment des dépouillements textuels en quête de proverbes, ont décidé d'attribuer le statut de proverbe à toute expression recensée dans des recueils parémiographiques. Plus précisément, il conteste tous les cas où ces recueils ne sauraient pas confirmer le statut de certaines expressions pour lesquelles l'arbitraire s'impose. Buridant en conclut que :

> « Le recours exclusif à l'autorité des recueils, pour l'identification des proverbes, n'est donc pas sans risque : quelle que soit l'étendue de sa culture parémiologique, nul chercheur ne peut prétendre, même pour une période relativement restreinte, connaître le corpus complet des proverbes qui y sont enregistrés » (1976 : 379).

Et quelques lignes après :

> « Le seul recours aux références externes est donc une chausse-trape […] » (1976 : 388).

Aucune liste de proverbes préalable et aucune connaissance – pour vaste qu'elle soit, en diachronie et en synchronie – des proverbes ne peuvent mettre à l'abri de fautes et de défaillances. La reconnaissance des proverbes dans les textes est donc à envisager comme un défi qui ne donne pas tous les résultats auxquels on pourrait s'attendre. Certes, Buridant fait référence ici à la tâche de reconnaissance de proverbes à l'époque du moyen français. Rien n'empêche que cette remarque soit aussi valable pour la détection des proverbes en français contemporain, notamment en ce qui concerne la fiabilité et l'exhaustivité des recueils parémiographiques (parfois incomplets, parfois datés). Par conséquent :

> « Il est donc nécessaire de retenir [...] des critères internes, c'est-à-dire des éléments de définition formalisables qui en permettent le repérage » (1976 : 390).

et qu'on identifie dans :

> « […] des études […] qui se sont efforcées de déceler les éléments constitutifs de la locution sentencieuse » (*ibid*.).

Buridant envisage, certes, de compléter les recueils parémiographiques par l'établissement de critères linguistiques qui éclairent le statut du proverbe. Pourtant, il finit par être pris au piège de cette tautologie (inévitable) que comporte la reconnaissance des proverbes dans des textes : le recours aux *références externes*[98] que lui-même a désapprouvé. Or, ces références ne sont pas des recueils parémiographiques, mais les travaux de collègues parémiologues (notamment Rodegem) et sémiologues (à savoir Greimas). Pourtant, il est aussi risqué de s'appuyer sur l'*autorité des recueils*[99] que sur l'*autorité des parémiologues*[100] eux-mêmes, et ce, malgré leurs connaissances et leurs esprits. Il semble évident qu'on ne peut éviter un repère externe – une ou plusieurs sources estimées comme fiables, récentes, pertinentes, etc. – pour faire face à la reconnaissance des proverbes. Malgré leurs faiblesses, les références externes restent indispensables pour faciliter cette entreprise.

Sources

L'étude de Buridant se concentre sur les textes du Moyen Âge, notamment sur le *Recueil général des Jeux-Partis français* par Langfors, Jeanroy et Brandin. À cet égard, il suffit de mentionner le tout début de sa contribution, où Buridant précise qu'on peut comprendre :

> « ces jeux-partis comme un corpus, c'est-à-dire un ensemble fini d'énoncés homogènes qu'on peut soumettre à une analyse » (1976 : 377)

et que :

> « l'homogénéité de ce corpus est assurée par la technique et le sujet des jeux-partis » (*ibid*.).

[98] C'est nous qui utilisons cette étiquette.
[99] C'est nous qui utilisons cette étiquette.
[100] C'est nous qui utilisons cette étiquette.

Le linguiste contemporain a l'impression de lire un travail moderne de linguistique de corpus où l'on définit son corpus par rapport à son homogénéité interne et aux spécificités du genre textuel qui le compose, quoiqu'il n'en soit pas encore le cas.

Fréquence

Au bout du dépouillement du *Recueil*, les traits établis par Buridant pour la reconnaissance des proverbes (§ 4.1.1.) n'apportent qu'une contribution quantitative exiguë. D'après les 107 proverbes reconnus, seulement 12 relèvent d'une reconnaissance à l'aide de critères linguistiques, alors que 84 sont détectés moyennant l'autorité des références externes (1976 : 405-406).

L'étude devient véritablement quantitative quand Buridant corrèle la fréquence d'usage de chaque proverbe par rapport à leur position et à leur signalement dans le *Recueil*. Plus précisément, il répartit les occurrences de chaque proverbe de par leur position en début, fin ou dans le corps de chaque strophe ainsi qu'en fonction de la présence ou de l'absence de « formules d'introduction » (1976 : 406)[101]. Il est intéressant de constater que les proverbes privilégient autant le corps de la strophe (45 sur 107) et la fin de la strophe (44 sur 107). Le plus souvent, il ne sont pas signalés par une formule d'introduction (75 sur 107). Il en ressort donc que :

> « les partenaires des jeux-partis utilisent donc des proverbes […] comme point d'accrochage de l'argumentation à des places qui tendent à devenir des places fixes » (*ibid.*).

Sans vouloir contraindre la position des proverbes à tout prix, nous souhaitons souligner que d'autres études parémiologiques basées sur des textes non médiévaux (Marcon 2011) observent un positionnement similaire, à savoir dans le corps ou en fin de textes. On pourrait même dire que le positionnement des proverbes dans les textes pourrait constituer un trait pour la reconnaissance des proverbes (§ 4.).

Quelques mots encore sur les formules d'introduction des proverbes. Les résultats de Buridant en diachronie semblent remettre en question un des arguments défendus par Anscombre en français contemporain (Anscombre 2000, 2011b). Il est intéressant de constater que, en l'occurrence, ces formules pointent des proverbes et contribuent à leur

[101] Quelques parémiologues les ont rebaptisées par la suite *formules* (Conenna 2004), *proverbial introducers* (introducteur proverbial, (Čermák 2007 [1998] : 575)) ou *marqueurs médiatifs* (Anscombre 2011)).

reconnaissance, quoiqu'elles ne soient pas vraiment nécessaires pour introduire les proverbes dans le discours (voir Conenna § 4.2.2.1.).

3.1.3. *Rodegem* (1984)

« La parémie peut être mise en chiffres » (1984 : 122).

Cette affirmation de Rodegem suit de quelques lignes sa proposition à l'égard du terme *parémie* – terme portemanteau des « énoncés sentencieux pris globalement » (1984 : 121) – et précède de quelques lignes sa définition « comme sentence lapidaire normative » d'origine orale (1984 : 122). En même temps, Rodegem crée le néologisme *parémiométrie* (*ibid*.) pour indiquer l'étude quantitative menée sur ses échantillons de parémies.

La contribution de Rodegem est ainsi marquée par l'intention de médiation entre plusieurs métatermes parémiologiques en vue d'une compréhension globale des traits définitoires quantifiables et quantifiés qui servent à décrire chaque parémie. Son étude peut être considérée comme un préambule métaterminologique et méthodologique pour l'étude de la fréquence d'occurrence des proverbes français dans des corpus. À notre connaissance, il s'agit de la première étude française parémiologique qui généralise les propriétés propres aux parémies (et, entre autres, les proverbes) sur base quantitative. Nous mentionnons cette étude non pas comme une véritable étude de fréquence, mais, après l'étude de Buridant (§ 3.1.2.), plutôt comme un tournant quantitatif pour la parémiologie francophone du XX^{e} siècle. Elle représente une preuve de la prise de conscience croissante quant à la possibilité et à la nécessité de mettre en chiffres toutes les parémies.

Liste

Son travail ethnolinguistique pionnier ambitionne à comparer « cent proverbes [*sic*] français » (*ibid*.) tirés de deux recueils parémiographiques (*Le Littré* et *Larousse*) avec « cent formules extraites d'un corpus de quatre mille énoncés sentencieux provenant du Burundi » (*ibid*.). On ignore, pourtant, les listes des parémies sélectionnées. On ne dispose non plus des critères qui ont guidé l'auteur à privilégier telle ou telle autre parémie en français et en rundi.

Sources

À notre connaissance, aucun corpus n'est dépouillé.

3.1.4. *Schulze-Busacker* (1985)

L'ouvrage de Schulze-Busacker est axé sur le dépouillement manuel de textes littéraires narratifs en ancien et moyen français. La réflexion sur les critères de reconnaissance des proverbes (que nous analyserons séparément au § 4) rapproche cette étude des questionnements récents sur la détection des proverbes. En même temps, à notre connaissance, il s'agit de la première étude en parémiologie française qui montre une description quantitative ponctuelle. C'est à Schulze-Busacker qu'on doit la première liste de fréquence des proverbes français sur des critères statistiques descriptifs, quoique son travail porte sur une variété diachronique très éloignée par rapport à celle que nous ciblons dans notre recherche.

Liste

Schulze-Busacker est dépourvue d'une liste de proverbes de départ. Malgré cela, la consultation systématique du métarecueil parémiographique de Morawski de 1925 pour la vérification du statut proverbial des formes reconnues, fait qu'elle dispose d'une *liste de référence*[102], voire liste d'inclusion et d'exclusion. Ce recueil rassemble 2.500 proverbes. Il est une synthèse de 29 recueils parémiographiques inclus dans 22 manuscrits français qui datent du XIII^{e} au XV^{e} siècle (1985 : 17). Cette exhaustivité a prôné Schulze-Busacker à privilégier le recueil de Morawski parmi d'autres ressources qu'elle a, en tout cas, consultées pour des contre-vérifications (1985 : 17-18). La référence constante à Morawski, notamment « le rapprochement de l'énoncé à un ou plusieurs proverbes recueillis » (1985 : 18), d'une part, constitue un appui irréfutable et, d'autre part, fonde certains jugements subjectifs inévitables, comme le présuppose le « rapprochement » dans la citation que nous avons reprise.

[102] C'est nous qui utilisons cette étiquette.

Sources

Schulze-Busacker sélectionne 112 textes littéraires narratifs, notamment des romans et des contes courtois, rédigés au Moyen Âge (1160-1300) (1985 : 13-14). Les œuvres de son corpus totalisent environ 665.000 vers (1985 : 20) et excluent les textes en prose parce qu'ils « contiennent beaucoup moins d'éléments proverbiaux que les œuvres en vers » (1985 : 14). Cette remarque apparemment générique et, pourtant, vérifiée sur la base de son expérience, nous conduit à reconsidérer l'importance que recouvre le rythme en diachronie pour la reconnaissance, outre que pour l'usage, des proverbes[103].

Fréquence

Dans son corpus de vers, Schulze-Busacker reconnaît 1.939 occurrences qu'elle peut « rattacher d'une manière plus ou moins directe à 649 proverbes » (1985 : 20), à savoir « une moyenne de deux à trois occurrences par mille vers » (*ibid*.). Au premier abord, ce bref rapport statistique descriptif témoigne une vivacité proverbiale dans les textes versifiés et un usage courant à l'écrit au Moyen Âge. Toutefois, cette moyenne arithmétique est mieux encadrée dans son analyse des fréquences pour chaque proverbe identifié. Seulement 21 proverbes possèdent $f > 10$, dont 5 dépassent le seuil des 20 occurrences, comme nous reproduisons dans le Tableau 5 (1985 : 22) :

Rang	**Proverbe [orthographe contemporaine]**	***f* brute**
1)	*Mius vaut mourir a joe* *que vivre a onte* [Mieux vaut mourir à joie que vivre à honte]	42
2)	*Fame veult touz jours faire* *ce que len lui vee* [Faim/renommée veut toujours faire ce qu'on lui interdit]	25

[103] En ce qui concerne les textes versifiés, le respect de la syllabation et des accents jouent un rôle central pour garder la métrique privilégiée et les effets acoustiques désirés.

3)	*Amor veint tute rien* [Amour vainc tout et rien]	23
4)	*Qui croit et aimme fole femme /* *Il gaste avoir et cors et ame* [Qui croit et aime folle femme / Il va gâcher et corps et âme]	22
5)	*Teus cuide gaingnier qui pert* [C'est celui qui gagne qui perd]	20

Tableau 5. Fréquence des 5 premiers proverbes français les plus fréquents d'après Schulze-Busacker (1985).

Quant aux *f* des 628 autres proverbes :

- $f = 1$ pour 273 proverbes ;
- $f = 2$ pour 140 proverbes ;
- $3 \leq f \geq 5$ pour 170 proverbes ;
- $6 \leq f \geq 10$ pour 45 proverbes (*ibid*.).

Si nous traduisons ces *f* en pourcentages, environ 10% des proverbes (66) ont $f > 5$ alors que le restant 90% restent en deçà de ce seuil. D'une part, cet écart suggère une préférence d'usage accordée aux proverbes utilisés « avec une certaine insistance », et ce, malgré « les nombreux proverbes médiévaux conservés » (1985 : 24). La parémiologue justifie cette répétition comme fonction des choix stylistiques partagés par les auteurs ainsi que des schémas narratifs. D'autre part, avec un certain recul sur les données statistiques, nous observons qu'en général surtout si l'on prend en compte la variété des proverbes, dont $1 \leq f \geq 2$ – qui représentent plus de la moitié des proverbes reconnus (413 sur 649) – , les séquences formulaires à l'époque jouissent d'une vitalité remarquable dans les textes littéraires[104].

Certes, « le rapprochement de l'énoncé à un ou plusieurs proverbes recueillis » (1985 : 18) dans le métarecueil de Morawski met en évidence la difficulté de catégorisation des proverbes reconnus au moment du dépouillement et, par conséquent, du calcul de leur fréquence. Il est, d'ailleurs, possible de réinterpréter le nombre total d'occurrences des

[104] Cette vitalité suit, à notre avis, le plaisir créatif attaché à la séquence formulaire et à son immédiateté qui se manifestent, malgré *f* très basses, sous plusieurs formes. Ces formes agissent comme des *pré-textes*, dans le sens où elles sont des textes qui précèdent (en cotexte ou en dehors de ce cotexte) les œuvres où elles sont attestées, et qu'elles représentent des textes qui motivent et appuient la narration cotextuelle.

proverbes en relation avec 3 macrocatégories d'usage des proverbes en co(n)texte que Schulze-Busacker a élaboré :

- le *proverbe cité*, c'est-à-dire le proverbe repris tel quel (1985 : 27-30 ; 168) ;
- le *proverbe intégré*, à savoir le proverbe qui a subi des modifications formelles pour s'adapter au co(n)texte d'occurrence (1985 : 25-27 ; 167-168) ;
- le *proverbe exploité* ou le proverbe employé directement ou indirectement pour créer un texte ou une partie d'un texte, tout en disposant de certains procédés formels utilisés pour la citation ou pour l'intégration des proverbes (1985 : 30-35 ; 168).

D'après cette tripartition, sur le total des 1.939 occurrences, seulement 444 correspondent à des proverbes cités et 139 à des proverbes exploités, alors que 1.356 renvoient à des proverbes intégrés (1985 : 25). Ces chiffres nous aident à mieux justifier la précaution qui se cache derrière le mot *rapprochement* et le degré de subjectivité qui a caractérisé la 'détection à vue' des proverbes.

Schulze-Busacker raffine ultérieurement la répartition des occurrences et enchevêtre ces 3 macrocatégories avec 3 situations d'énonciation : le *discours direct*, le *récit* et la *digression* (1985 : 36). Il en suit que les proverbes privilégient le discours direct (1.176 occurrences) par rapport au récit (540 occurrences) et à la figure de style de la digression (223 occurrences) (*ibid*.).

Si l'on insère un troisième niveau d'affinement des fréquences qui décrit les procédés que Schulze-Busacker attache à chaque macrocatégorie (*ibid*.), il en ressort, de manière transversale aux trois situations d'énonciation, que :

- presque la moitié des proverbes cités sont isolés de leur co(n)texte d'occurrence (227 sur 444 occurrences) ;
- un peu moins de la moitié des proverbes intégrés sont reconnus grâce au maintien de leur noyau proverbial (§ 4.3.) (662 sur 1.356 occurrences) ;
- la plupart des proverbes exploités sont des proverbes en série (1985 : 32-34), à savoir des suites de proverbes qui servent à développer le texte, notamment des digressions de la part des auteurs (102 sur 139 occurrences).

Vu tous les obstacles qui découlent de l'analyse d'une variété diachronique dépourvue d'une véritable norme parémiographique, l'étude de Schulze-Busacker constitue sans aucun doute la preuve par excellence de faisabilité d'une étude de fréquence d'usage des proverbes à l'écrit.

3.2. Fréquence et format électronique

Dans ce sous-chapitre, nous analyserons 19 études de fréquence et de traitement quantitatif des parémies ainsi que de leurs traits formels et/ou de leurs processus de variation lors de l'usage à l'écrit dans des sources au format électronique. Nous nous attarderons sur une sélection que l'on peut séparer en :

- études intralinguistiques et
- études comparées.

Pour les premières, nous nous focaliserons sur 4 études concernant les parémies anglaises (§§ 3.2.1, 3.2.2., 3.2.4., 3.2.5.), 3 qui visent les parémies allemandes (§§ 3.2.9, 3.2.12.), 2 respectivement pour les parémies espagnoles (§§ 3.2.6., 3.2.15.) et tchèques (§ 3.2.8.), ainsi qu'une étude respectivement pour les parémies en bulgare (§ 3.2.13.), en estonien (§ 3.2.7.), en français (§ 3.2.10.) et en polonais (§ 3.2.14.). Quant aux analyses comparées, nous en traiterons 3 pour le couple espagnol-français (§ 3.2.11., 3.2.16.) et une seule pour le couple anglais-français (§ 3.2.3.). D'orès et déjà, précisons que la contribution majeure quant aux approches empiriques basées sur corpus arrive des littératures parémiologiques et phraséologiques anglophone et hispanophone.

Comme pour les études sur papier, les études sur des sources électroniques seront présentées par ordre chronologique pour mettre en relief les évolutions et les enchâssements croisés méthodologiques entre parémiologues et langues.

3.2.1. *Norrick* (1985)

Dans son ouvrage consacré aux proverbes anglais, Norrick (1985) considère les proverbes comme des textes à part entière et souhaite analyser leur sémantique et leur valeur en contexte. La fréquence d'usage ne constitue pas une finalité explicite de son étude.

Pourtant, Norrick glisse des informations quantitatives qu'on peut extrapoler et offre des suggestions pour des études quantitatives et qualitatives.

Liste

Au début de sa recherche, Norrick établit une liste de 200 premiers proverbes, phrases proverbiales et autres formules stéréotypées qu'il repère sous la lettre F de l'*Oxford Dictionary of English Proverbs* de Wilson de 1970. Plus précisément, ces 200 proverbes correspondent à 172 formes canoniques et 28 variantes (1985 : 7-8). Sans prendre en considération des études de familiarité *a priori* ou *a posteriori*, Norrick échantillonne ses formes sur la base d'une ressource lexicographique, entre autres, dit-il :

> « to eliminate idiosyncratic experience and personal preference from the identification process » (1985 : 8).

Comme il doit démontrer la validité des propriétés du *proverb*, de ses définitions ainsi que de sa classification typologique par rapport à d'autres genres textuels gnomiques et folkloriques (1985 : 67-79), Norrick évite le biais initial d'un jugement personnel ou confié à la connaissance des locuteurs natifs. Norrick fait confiance – peu ou prou – à une norme descriptive lexicographique dont la codification est assez contemporaine au déroulement de sa recherche. Ce qui l'écarte de certains parémiologues (§ 3.2.3.) et le rapproche d'autres (§ 3.2.4.).

Sources

Tout d'abord, Norrick décide d'observer les proverbes anglais à l'oral en raison du fait que :

> « given that the proverb is essentially conversational, proverbs in conversation provide a natural point of departure for a study of proverbs in texts generally » (1985 : 12).

Ce propos doit, pourtant, faire face à l'épreuve des corpus disponibles au moment de son étude. Il repère à l'oral une seule occurrence du proverbe :

the proof of the pudding is in the eating [trad. litt. : la preuve du pudding c'est dans le manger] (1985 : 13)

ainsi qu'un seul détournement du proverbe :

A bird in the hand is worth two in the bush [trad. litt. : Un oiseau dans la main vaut mieux que deux dans le buisson]

dans le *Corpus of English Conversation* établi par Svartvik & Quirk en 1980. Les mêmes résultats décevants apparaissent lors de l'exploitation du corpus de transcriptions de conversations mis au point par Crystal & Davy en 1975 (1985 : 6)

Norrick décide ainsi de se pencher sur l'écrit. Dans ce cas, aussi, il constate les difficultés que la fiction narrative littéraire, tout comme les pièces théâtrales – quoiqu'en mesure inférieure, à son avis, en raison de leur nature dialogique – posent en vue de son analyse sémantique des proverbes. Malgré cela, il opte pour une revue des proverbes dans le *Shakespeare Index* établi par Tilley en 1950, plus précisément dans 4 pièces de William Shakespeare : « in order to ensure an ample and sufficiently varied corpus » (1985 : 19). De manière assez étonnante par rapport aux études successives d'autres chercheurs, il préfère ignorer la presse, et ce, sur la base d'expériences directes et indirectes peu prometteuses.

Il faut en tout cas souligner que Norrick a (peut-être) pressenti les changements que l'exploitation des premiers corpus est en train d'apporter, notamment en lexicographie et en syntaxe. C'est de cette façon que nous avons interprété l'ouverture du chapitre qui clôture son ouvrage :

« One natural point of departure for future work on proverbs is thus an extension of the corpus treated here. Proverbs in naturally occurring conversation, in plays and other literature, in the press etc. […] should be collected, classified and compared » (1985 : 169, c'est nous qui soulignons).

Cette affirmation représente le préambule de toutes les analyses successives des proverbes sur corpus. Elle révèle également une sensibilité quant à la répartition que les proverbes peuvent montrer de par le genre textuel. Norrick insiste sur la nécessité d'amasser des données linguistiques, mais aussi de prévoir une analyse systématique et une formalisation de ces mêmes données, ainsi que de se pencher sur leur interprétation. Il indique explicitement la

voie de la collection des données linguistiques réelles comme la meilleure voie pour traiter le proverbe et pour remettre en question son statut à l'ère contemporaine.

Fréquence

Pour revenir aux données quantitatives et à leur relation avec le genre textuel interrogé, Norrick suggère l'exploitation d'œuvres théâtrales en vue d'atteindre une quantité suffisante d'exemples. En guise de contre-exemple, Norrick affirme avoir repéré seulement 3 citations, dont 2 proverbes, à l'issue du dépouillement manuel randomisé de copies du journal *International Herald Tribune* pendant 9 mois (1985 : 7). Il en va de même pour une de ses étudiantes qui ne collecte que 2 autres proverbes (qui plus est, détournés) dans 5 copies choisies au hasard de la revue *Time* (*ibid*.).

Malgré tout, Norrick ne renonce pas à leur analyse (1985 : 22-24) et se concentre sur 4 pièces de Shakespeare. Il y reconnaît 20 proverbes, y compris des variantes et des allusions (Norrick 1985 : 19). Grâce aux exemples illustrés (1985 : 19-22), Norrick arrive à des conclusions, d'une part, stylistiques sur l'écriture de Shakespeare et, d'autre part, sémantiques sur l'usage des proverbes en tant que *templates* (1985 : 21) de la sagesse populaire introduits dans la conversation des personnages et finalisés au déroulement de la pièce. Ces *templates* s'insèrent dans un discours et encadrent son déroulement, tout en véhiculant parfois un jugement de valeur (1985 : 24).

Au bout de l'observation d'un nombre très restreint d'occurrences, on s'attendrait à une attitude défaitiste sur le destin du proverbe et sur son emploi en discours. Au contraire, aux parémiologues qui prophétisent la mort du proverbe, il répond avec empirisme :

> « we should reserve judgement on the fate of the proverb until more statistics are available » (1985 : 169).

3.2.2. *Mieder* (1993)

À propos de dépouillement textuel, en 1993, Mieder consacre un article au proverbe anglais :

The apple doesn't fall far from the tree [La pomme ne tombe pas loin de l'arbre]

et à son historique en diatopie, de l'Orient à l'Occident. C'est à l'allemand que Mieder attribue l'origine de ce proverbe, à savoir :

Der Apfel fällt nicht weit vom Stamm (1995 [1993] : 224)

et décrit son parcours d'acquisition dans le répertoire parémiographique anglophone étatsunien.

Liste

En 1993, la recherche de Mieder cible un seul proverbe. Il n'y a donc pas de liste.

Sources

Le parémiologue veut examiner la réception et l'usage en discours du proverbe anglais. Pour ce faire, Mieder entreprend la consultation de différentes bases de données textuelles numérisées. Outre le corpus des fables des frères Grimm en version électronique, Mieder lance des recherches dans la base de données en ligne *LEXIS*/*NEXIS* (1995 [1993] : 224) (§ 3.2.6.). Il décrit ce « giant database system » (*ibid*.) comme une collection ouverte de textes de presse, rapports, transcriptions d'émissions télévisées et d'autres matériaux rassemblés depuis la fin des années 1970.

Fréquence

Les performances de ses techniques d'interrogation trop génériques (§ 4.2.1.2.) comportent du bruit : sur un total de 232 occurrences repérées, 123 ne sont pas pertinentes (*ibid*.). Les 109 occurrences restantes montrent un regain progressif de la fréquence d'occurrence brute du proverbe en microdiachronie : d'une seule occurrence en 1981 jusqu'à 28 occurrences en 1992. Cette progression positive de la fréquence est justement interprétée par Mieder comme une conséquence directe de la croissance postérieure des documents dans

la base de données (*ibid*.). Malgré cela, Mieder ne cache pas son enthousiasme[105] et affirme que :

> « the 109 references from 1981 to 1992 have never before been registered by proverb scholars, and they most certainly establish the modern American currency of this proverb » (*ibid*.)

qu'il complète peu loin avec les mots suivants :

> « Database searching for particular proverbs is truly revolutionizing paremiography as it has been known thus far » (1995 [1993] : 244).

Ce constat programmatique pour avantager l'étude de la *contemporanéité* (*currency*) proverbiale n'a plus lieu dans sa contribution rédigée l'année suivante, en 1994, sur la relation entre minimum parémiologique et littératie culturelle. La fréquence synchronique d'usage des proverbes constitue un des soucis soulevés par cette contribution. Pourtant, l'usage du mot *frequency* s'accompagne presque toujours du mot *familiarity*, ainsi que des mots *traditionality* et *currency*, au point que leur entrelacement notionnel en est indissociable dans son argumentation. En guise d'exemple, on peut citer le passage suivant :

> « […] two extremely important questions […] go beyond purely linguistic aspects of proverbial texts. The one deals with the diachronic problem of traditionality, i.e. the fact that any text to qualify as a proverb must have (or have had) some currency for a period of time. Related to this is the synchronic question of frequency of occurrence or familiarity of a given text at a certain time. None of the dozens of proverb definitions can answer these questions, and yet any proverb must "prove" a certain traditionality and frequency in order to be considered verbal folklore » (1995 [1994] : 13, c'est nous qui soulignons).

Il ajoute quelques lignes après :

> « These questions are not new, but they need to be addressed in a more scientific fashion using modern means of statistical research » (*ibid*.).

[105] En note, il remercie ses collègues comme suit : « I would like to thank my colleagues […] at the University of Vermont for introducing me to the new world of databases. They know how excited I was when I learned about this 'marvelous' technology » (1995 [1994] : 270, c'est nous qui soulignons).

Quand Mieder écrit ces mots, la linguistique de corpus est en plein essor – *Corpus, Concordance, Collocation* de Sinclair a déjà fait sa parution –, mais le dépouillement de grands corpus n'est pas encore répandu chez les linguistes et balbutie parmi certains parémiologues (§ Arnaud et Moon 3.2.3.). Mieder est donc fort conscient de l'apport des études quantitatives (1995 [1994] : 25). Malgré cela, toutes les études mentionnées dans sa contribution depuis les années 1930 (Albig, Bain, Levin, Tillhagen, Permyakov, Kuusi, Marzolf, Hirsch etc.) soulignent que, lorsqu'il mentionne le terme *frequency*, il fait plutôt référence au nombre de fois que tel ou tel autre proverbe est reconnu, signalé ou rédigé par des locuteurs lors d'expériences sur terrain de toute sorte. Pourtant, la distinction entre fréquence et familiarité a l'air d'être plus nette quand il décrit son expérience de parémiographe pour l'anglais :

> « How else was I to come up with these 1200 texts [proverbs] but to go to some of the historical English and Anglo-American proverb collections and letting my scholarly knowledge of proverbs together with my subjective feeling be the guide to decide whether any given text had enough currency (frequency, traditionality, familiarity, etc.) to be included. […] I stuck out my proverbial neck at times and marked some proverbs in the notes as being particularly "popular", but I remember a certain scholarly unease since I was not really basing this judgment on demographic research » (1995 [1994] : 26, c'est nous qui soulignons).

Ce passage nous éclaire sur le fait que la fréquence et la familiarité sont à leur tour, pour Mieder, fonction de la contemporanéité (*currency*) et de la popularité des proverbes[106]. Ensuite, la conscience de donner un jugement de popularité proverbiale sur la base de recherches démographiques, non pas d'une intuition personnelle, confirme que, dans son raisonnement, Mieder privilégie la compétence parémiologique telle qu'elle est exprimée par des locuteurs natifs plutôt que la fréquence d'usage observée dans des textes. Il semble ainsi oublier son enthousiasme (du moins, en 1994) à l'égard du dépouillement des textes électroniques.

[106] Une dizaine d'années plus tard, cette corrélation entre fréquence et familiarité sera prouvée de façon algébrique par Grzybek (§ 3.2.12.).

3.2.3. *Arnaud & Moon* (1993)

L'étude d'Arnaud & Moon (1993) représente la première tentative (et peut-être la seule, à notre connaissance) de calcul de la fréquence d'occurrence des proverbes dans une optique comparée anglais-français. Cette étude pionnière a le mérite de suggérer des pistes et des observations méthodologiques modernes, dont nous nous inspirerons. Par ailleurs, elle présente des résultats peu connus (ou méconnus) sur la fréquence et sur l'usage des proverbes anglais et français[107].

Liste

Arnaud & Moon se concentrent sur la recherche de leur « ensemble prédéfini de proverbes » (1993 : 325), à savoir leur :

> « liste de 240 proverbes [anglais] précédemment rassemblés pour une enquête de familiarité » (*ibid*.).

Dans la démarche suivie, on constate un aspect méthodologique novateur : la corrélation entre familiarité et fréquence. Plus précisément, on remarque que le calcul de la fréquence des proverbes suit ou dépend d'une liste de proverbes collectés et/ou ordonnés selon leur degré de connaissance exprimé par un échantillon de personnes-locuteurs natifs.

Sources

Plutôt que sur les techniques d'interrogation (que nous décrirons au § 4.2.1.1.), l'attention d'Arnaud & Moon se concentre sur les corpus eux-mêmes ainsi que sur leur composition. Les auteurs se focalisent sur leurs sources pour mieux cerner la portée de leur étude de fréquence, à savoir pour mettre en relation la présence d'un proverbe avec un ou plusieurs genres textuels. Pour les proverbes anglais, Moon peut s'appuyer sur l'*Oxford Hector Pilot Corpus*, un corpus d'environ 18 millions d'occurrences et composé, pour l'essentiel, d'écrit, notamment de textes tirés de la presse britannique, ainsi que (par ordre

[107] Lors de notre dépouillement de la littérature parémiologique et phraséologique, nous avons constaté que cette étude est rarement mentionnée.

décroissant) par des textes non fictionnels, de fiction, des transcriptions de l'oral et par d'autres textes. En revanche, pour les proverbes français, Arnaud ne peut tirer parti d'aucun corpus informatisé. À la suite d'une étude pilote menée dans la base de données *Frantext* qui inclut des textes littéraires, Arnaud observe la quasi-absence des proverbes que lui-même avait estimés comme familiers (Arnaud 1992). Ainsi :

> « il fût décidé de renoncer à consulter *Frantext* » (Arnaud & Moon 1993 : 324)

en faveur d'une collection de 374 occurrences de proverbes mise au point par le jugement d'Arnaud et de par leur hasard d'attestation et de perception de la part de l'auteur. La période de collecte devrait couvrir environ 8 mois au début des années 1990[108]. Les occurrences des proverbes relèvent principalement des médias (298 occurrences), mais aussi de conversations orales et d'autres genres textuels écrits.

Fréquence

L'asymétrie entre le corpus anglais et la collection manuelle pour le français doit nous mettre en garde sur toute interprétation des résultats. En tout cas, par rapport aux 374 occurrences des proverbes français, Moon en repère 603 pour les proverbes anglais.

Au moment de la présentation des fréquences brutes d'occurrence des proverbes anglais et français, les auteurs reprennent seulement les proverbes dont $f \geq 4$. En tête de liste en anglais, on rencontre (1993 : 326) :

Rang	Proverbe [traduction littérale]	*f* brute
1)	*It's the last straw* *that breaks the camel's back* [C'est la dernière paille qui casse le dos du chameau]	40
2)	*Enough is enough* [Trop, c'est trop]	18

[108] Nous avons reconstruit la période de collecte en mois de par la précision apportée par Arnaud : « […] le nombre d'occurrences collectés […] s'élève à 374 occurrences, ce qui permet d'estimer aux environs de 1,5 le nombre d'occurrences proverbiales rencontrées quotidiennement » (1993 : 325).

3)	*First come, first served* [Venu en premier, servi en premier]	15
4)	*Every cloud has a silver lining* [Chaque nuage a une lueur d'espoir]	14
	Small is beautiful [Petit est beau]	14
	The chain is no stronger *than its weakest link* [La chaîne n'est pas plus forte que son maillon le plus faible]	14
7)	*Live and let live* [Vivre et laisser vivre]	12
8)	*A leopard does not change its spots* [Un léopard ne change pas ses taches]	11[109]
9)	*A drowning man will clutch at a straw* [Un homme qui se noie se cramponnera à une paille]	10
	You can't have your cake and eat it [Tu ne peux pas avoir ton gâteau et le manger]	10

Tableau 6. Fréquence des 10 premiers proverbes anglais les plus fréquents d'après Moon.

Le tout premier proverbe :

It's the last straw that breaks the camel's back

a $f = 40$, quoique seul 5 de ses occurrences respectent la forme canonique. En effet, les 35 autres occurrences correspondent à une réduction du proverbe aux deux groupes nominaux *last straw* et *camel's back* (1993 : 328). Il s'avère que la réduction à « des unités syntaxiques de niveau inférieur » (1993 : 329) concerne la plupart des occurrences de tous les proverbes anglais dont $f > 10$. Seule exception, le proverbe au deuxième rang :

Enough is enough [Trop c'est trop]

[109] Dans la liste proposée par Moon, ce proverbe aurait 10 occurrences, alors que les deux qui suivent en compteraient 11. Comme la liste est évidemment triée par ordre de fréquence décroissante, nous croyons qu'il s'agit d'une faute de transcription.

pour lequel les 18 occurrences reflètent la forme canonique (de par sa nature brève) (*ibid*.). Moon ne signale pas non plus des cas de variantes proverbiales par adjonction, par exemple. Ce qui suggère une tendance générale au rétrécissement formulaire, du moins en anglais.

On remarque l'écart entre les occurrences du proverbe au premier rang (40) et de celui au deuxième rang (18). Tous les autres proverbes ne dépassent pas le seuil de $f = 15$. Plus précisément :

- 8 proverbes ont $15 \leq f \geq 10$;
- 10 proverbes ont $8 \leq f \geq 6$;
- 18 proverbes (et 2 variantes corrélées à 2 proverbes) ont $f = 5$;
- et 20 proverbes ont $f = 4$ (1993 : 326-327).

Il s'en suit que, sur un total de 240 proverbes familiers, 182 ont $3 \leq f \geq 0$. Autrement dit, environ 76% des proverbes dans la liste de départ de Moon comptent de 0 à 3 occurrences. Premièrement, ce résultat interpelle la nature même des proverbes. Par cela, nous faisons référence au fait que les proverbes sont des dénominations-catégories de et pour diverses situations de vie. Par conséquent, un usage approprié des proverbes est en rapport étroit avec une situation discursive qui décrit ou s'adapte à une situation de vie donnée dont l''étiquetage' satisfait la forme et le contenu de tel ou tel autre proverbe[110]. La fréquence basse d'un nombre élevé de proverbes n'équivaut pas tout court à leur disparition.
Deuxièmement, la fréquence basse de la plupart des proverbes implique que la familiarité est une mesure non directement proportionnelle à la fréquence. La familiarité élevée d'un nombre *n* de proverbes peut correspondre comme ne peut correspondre à une fréquence élevée (§§ 3.2.9., 3.2.12.).

Quant au français, nous sommes contraint par le manque d'une liste de départ des proverbes à rechercher, outre que pour le manque d'un corpus. Le Tableau 7 ci-dessous reprend les 10 proverbes les plus fréquents (1993 : 328) :

[110] Un argument pareil est aussi défendu par les auteurs.

Rang	Proverbe	*f* brute
1)	*Une fois n'est pas coutume*	20
2)	*Trop, c'est trop*	12
	Un train peut en cacher un autre	12
4)	*À tout seigner tout honneur*	8
	On ne peut pas avoir le beurre et l'argent du beurre	8
	On ne tire pas sur une ambulance	8
7)	*L'enfer est pavé de bonnes intentions*	7
8)	*Les mêmes causes produisent les mêmes effets*	6
	Une hirondelle ne fait pas le printemps	6
10)[111]	*Chassez le naturel, il revient au galop*	5

Tableau 7. Fréquence des 10 premiers proverbes français les plus fréquents d'après Arnaud.

Nous nous limitons à observer que le premier proverbe français en tête :

Une fois n'est pas coutume

totalise 20 occurrences. Au deuxième rang *ex aequo* :

Trop, c'est trop
Un train peut en cacher un autre

comptent 12 occurrences respectivement. Comme pour l'anglais, le proverbe au premier rang s'écarte des autres. Arnaud observe, en outre, que *Trop, c'est trop* occupe le même rang que son équivalent anglais *Enough is enough* et toujours employé sous sa forme canonique (1993 : 329). La liste de fréquence des proverbes prévoit encore :

- 6 proverbes dont 8 (ex. *On ne tire pas sur une ambulance*) $\leq f \geq$ 6 (ex. *Une hirondelle ne fait pas le printemps*) ;

[111] Au même rang avec la même fréquence : *Il n'y a pas de fumée sans feu* ; *Les absents ont toujours tort* ; *On n'est jamais si bien servi que par soi-même* ; *Un malheur n'arrive jamais seul* ; *Vérité en deçà des Pyrénées, erreur au delà*.

- 6 proverbes dont *f* = 5 (ex. *Il n'y a pas de fumée sans feu*, *Vérité en deçà des Pyrénées, erreur au-delà*) ;
- 7 proverbes dont *f* = 4 (ex. *Tous les chemins mènent à Rome*, *On ne change pas une équipe qui gagne*) (1993 : 328).

Il est intéressant de s'attarder sur le commentaire qu'Arnaud fournit pour expliquer la fréquence du proverbe :

Vérité en deçà des Pyrénées, erreur au-delà

qu'il attribue à « l'imitation des journalistes les uns par les autres » (1993 : 330), à savoir à un « effet de mode » (1993 : 332). D'une part, cette lecture confirme l'apport significatif des médias en termes d'occurrences. D'autre part, elle sert d'argument pour défendre le manque de familiarité de ce proverbe dans son étude (Arnaud 1992). Si l'on reprend cette étude de familiarité conduite sur une population de 162 informateurs entre 18 et 23 ans, on remarque que *Vérité en deçà des Pyrénées, erreur au-delà* est au 359^{e} rang (sur un total de 401) de son classement (indice de familiarité = environ 15/100) (Arnaud 1992 : 225). Nous avons donc un cas inverse par rapport à la relation familiarité-fréquence observée pour les proverbes anglais, mais qui confirme le même propos. Un indice de familiarité bas n'implique pas directement une fréquence basse. La basse familiarité d'un nombre *n* de proverbes peut correspondre comme ne peut correspondre à une basse fréquence d'usage (§§ 3.2.9., 3.2.12.).

Au-delà du classement des proverbes par fréquence brute, Arnaud & Moon répartissent les occurrences d'après leur positionnement dans le discours. Autant en anglais qu'en français, la quasi-totalité des proverbes est incorporée dans un texte (93% en anglais, 73% en français). Ils préfèrent le corps central du texte (85% en anglais, 76% en français) comme « argument discursif » (1993 : 331), plutôt que la position initiale ou finale, comme clôture élégante et efficace (*ibid*.). Ces deux derniers cas sont, pourtant, privilégiés dans les journaux télévisés et les articles : les médias utilisent les proverbes comme pré(-)textes ou, d'après les auteurs, comme transitions (1993 : 332).

Une autre manière de réinterpréter les fréquences est offerte par une répartition des données quantitatives en rapport avec les processus de variation. Nous opérons une synthèse de ces répartitions dans le Tableau 8 (1993 : 332-337) :

	PROCESSUS DE VARIATION	ANGLAIS	FRANÇAIS
1	**Forme canonique (Ø variations)**	135/603	188/374
2	**Substitution d'un élément lexical**	56/603	55/374
3	**Substitution de plus d'un élément lexical**	16/603	9/374
4	**Ajout d'éléments lexicaux**	33/603	11/374
5	**Troncation**	16/603	18/374
6	**Réduction à niveau inférieur**		
	Groupe adjectival	29/603	0
	Groupe nominal	81/603	9/374
	Groupe verbal	139/603	12/374
7	**Cas complexes inclassables**	95/603	52/374

Tableau 8. Répartition des occurrences des proverbes anglais et français d'après les 7 processus de variation identifiés par Arnaud & Moon (1993).

Or, les comptes ne sont (apparemment) pas bons. Les sommes des occurrences pour les deux langues restent en dessous des totaux déclarés (600 pour l'anglais, 354 pour le français). Il se peut que, pour certains cas, les processus interviennent simultanément. Tout en faisant abstraction des chiffres pour quelques instants, la répartition révèle de manière frappante comment la réduction syntaxique affecte massivement les proverbes anglais, alors que la substitution lexicale l'emporte en français. On dirait que l'anglais tend à parcelliser les proverbes en discours sous forme de collocations (plutôt verbales) de matrice proverbiale. En français, il paraît que les proverbes continuent à garder leur combinatoire syntagmatique de départ et que la variation intéresse l'axe paradigmatique. Ces résultats sont certes à tenir en compte au moment de la modélisation de nos requêtes informatiques (§ 6.3.).

Si l'on considère maintenant la fréquence des proverbes répartie par genres textuels, il faut constater qu'en anglais, la « pauvreté en proverbes des textes non fictionnels » (1993 : 326) s'oppose à leur présence massive (419 sur 603 occurrences) dans les textes de presse. En français, d'une part, il faut prendre en considération la conclusion logique qui découle de la consultation de *Frantext*, c'est-à-dire que « le proverbe [français] tend à être absent de la littérature sérieuse » (1993 : 326) et, d'autre part, que les médias se font les principaux porteurs de proverbes. En guise de conclusion, on pourrait oser synthétiser les résultats d'Arnaud & Moon comme suit : les proverbes préfèrent la presse et les médias, alors qu'ils

négligent la littérature, à savoir la fiction et la littérature scientifique, tout comme les conversations à l'oral. Certes, le déséquilibre des sources anglaises et françaises ne nous permet pas d'attribuer une confiance absolue à l'égard de ces conclusions qui, d'ailleurs, demeurent des piliers incontournables en parémiographie anglaise et française.

3.2.4. *Lau* (1996)[112]

Aux États-Unis, en 1996, Kimberly Lau publie sur la revue *Proverbium* une étude de fréquence d'usage des proverbes basée sur un corpus d'anglais américain. Son étude se concentre sur les 10 proverbes les plus fréquents afin d'extrapoler les valeurs et les comportements les plus en vogue aux États-Unis dans la période 1970-1990 (§ 3.1.1.). Confier l'expression d'une seule culture nationale aux proverbes les plus fréquents est une entreprise quelque peu ambitieuse et risquée, comme elle-même l'avoue, et ce, en raison de la genèse commune – ou de la traduction/adaptation – de certains proverbes dans plusieurs langues/cultures. De façon pareille, la présence de proverbes antinomiques dans un même répertoire parémiographique complique la tâche (2003 [1996] : 231-232). Malgré ces défaillances, Lau est persuadée que :

> « The very proverbs which people choose to use with the greatest frequency imply certain understandings about the culture in which they are found » (2003 [1996] : 232).

En d'autres termes, s'il est vrai que Lau se focalise sur les coutumes et sur les comportements transmis par les proverbes, elle établit en même temps une liste de fréquence exhaustive qu'elle reprend dans les Annexes B de sa contribution (2003 [1996] : 249-252). Sans détailler son analyse sociolinguistique, nous sommes convaincu que la haute fréquence d'usage des proverbes est certainement un indice significatif pour mieux comprendre la culture d'une nation, mais qu'elle est plutôt un indice de certains mécanismes productifs phraséologiques.

Liste

Pour « practical considerations » (2003 [1996] : 233), Lau définit un premier seuil de 315 proverbes qui reflète le seuil fixé par l'expérience d'établissement du minimum

[112] Nous remercions Kimberly Lau pour son support et pour son enthousiasme.

parémiologique russe de Permyakov (1997 [1982]). Elle concentre son attention sur les « *self-contained proverbs* » (*ibid*.), c'est-à-dire rien que sur les proverbes, non pas sur d'autres expressions idiomatiques ou folkloriques. Ses 315 proverbes et phrases proverbiales ressortent du dépouillement du recueil *Modern Proverbs and Proverbial Sayings* de Whiting (1989) qu'elle élève au rang de recueil parémiographique essentiel[113]. Plus précisément, les proverbes ayant le nombre le plus élevé de citations mentionnées dans le même recueil sont insérés dans cette première liste (*ibid*.). Elle compare ainsi sa première liste avec 3 autres recueils. Ce croisement de ressources parémiographiques synchronique résulte en une liste finale composée de 188 proverbes (*ibid*.).

Sources

Lau n'utilise pas vraiment un corpus, mais une base de données textuelle : la base *LEXIS/NEXIS ALLNWS* (comme Mieder (§ 3.2.2.)). Au moment de l'interrogation par Lau, cette base recense environ 3.500 textes, dont à peu près 2.300 textes relèvent de la presse étatsunienne ainsi que d'autres pays[114] (2003 [1996] : 232).

Fréquence

Comme le moteur de recherche de la base *LEXIS/NEXIS ALLNWS* l'a empêché de saisir certains mots à haute fréquence (§ 4.2.1.3.), Lau ne peut qu'approcher le calcul de la fréquence des proverbes contenant ces mots. Plus précisément, pour tout proverbe contenant un ou plusieurs de ces mots et dont *f* brute > 1.000, elle a lu seulement les 100 premières occurrences pour évaluer la précision moyenne des résultats renvoyés par la base. Par une

[113] Pour comprendre le choix opéré par Lau, on peut s'appuyer sur ce que Mieder affirme à propos de la renommée étatsunienne de Whiting qu'il appelle « the American paremiographer par excellence » (2004 : 122). Mieder nous offre quelques détails sur son dictionnaire parémiographique historique : « […] individual proverbs and proverbial expressions are arranged alphabetically according to key words. For each proverb the editors supply historical references from Middle Ages on, often including the earlier classical and/or biblical references. At the end of such historical monographs on individual proverbs, cross-references to other proverb collections […] are cited as well » (*ibid*.). Sur le prestige de Whiting et sur le nombre des références citées pour chaque proverbe, Lau peut établir la première liste de 315 proverbes à filtrer successivement. Au passage, on constate que la notion de *historical monograph* pour chaque proverbe mentionnée par Mieder, ressemble à celle de *historique du proverbe* proposée par Conenna (2002) pour le français. En ce sens, nous observons que Mieder précise que : « Even though this methodology for major historical proverb collections has been long established, it is being followed by more or less exclusively only in the Anglo-American world […] » (*ibid*.). Bien évidemment, nous pouvons *a posteriori* rectifier ce constat et inclure de plein droit l'originalité du projet parémiographique francophone *DicAuPro* (Conenna *et al*. 2006).

[114] La seule origine non-étatsunienne d'une partie des textes serait un argument suffisant pour remettre en question les résultats obtenus.

simple proportion mathématique, Lau a appliqué cette moyenne d'occurrences exactes au nombre total d'occurrences repérées « to approximate the number of 'true' results for a given proverb » (*ibid*.). Ce qui finit par compromettre la fiabilité des fréquences finales.

De son interrogation, Lau obtient ainsi les 10 proverbes les plus fréquents suivants (2003 [1996] : 234) :

Rang	Parémie [traduction littérale]	*f* brute
1)	*Enough is enough* [Trop, c'est trop]	15.808
2)	*Time will tell* [Le temps le dira]	14.226
3)	*First come, first served* [Venu en premier, servi en premier]	13.050
4)	*Forgive and forget* [Pardonne et oublie]	5.097
5)	*Time is money* [Le temps, c'est de l'argent]	3.770
6)	*History repeats itself* [L'histoire se répète]	3.713
7)	*Time flies* [Le temps vole]	3.673
8)	*Better late than never* [Mieux vaut tard que jamais]	3.493
9)	*Out of sight, out of mind* [Loin du regard, loin de l'esprit]	2.902
10)	*Boys will be boys* [Les garçons seront des garçons]	2.103

Tableau 9. Fréquence des 10 premiers proverbes anglais les plus fréquents d'après Lau (1996).

La fréquence d'au moins 4 de ces 10 proverbes (*Enough is enough*, *Forgive and forget*, *Time is money* et *Boys will be boys*) se rapproche de la fréquence réelle, mais ne correspond pas vraiment à leur nombre réel d'occurrences. De toute façon, compte tenu de l'approximation, c'est le nombre très élevé d'occurrences ce qui nous étonne de cette liste. Le dépassement du

seuil de 1.000 occurrences n'est observé que dans cette étude et semble contredire toute présomption de rareté ou de disparition des proverbes dans le discours écrit. En général :

- 23 sur 188 proverbes ont $f > 1.000$,
- et 55 sur 188 proverbes ont $100 > f < 1.000$.

Du moins, ces résultats servent à prouver la validité du constat de Norrick sur la nécessité de vérifier la présence et l'absence des proverbes dans des corpus de taille critique (§ 3.2.1.).

Il est également surprenant de découvrir dans l'Annexe B que 17 sur 188 proverbes et phrases proverbiales communs à 4 recueils parémiographiques ont $f = 0$. De cette absence, on peut en tirer deux leçons :

a) aucun recueil parémiographique ne peut vraiment mettre à l'abri d'un échec lors de l'interrogation d'un corpus et
b) ceux qui se méfient des recueils parémiographiques ont raison de le faire, surtout quand ces recueils incluent des proverbes qui ne sont (peut-être) plus tous courants.

Pourtant, pour une étude de fréquence, il vaut décidément mieux s'appuyer sur un recueil exhaustif et daté plutôt que de s'appuyer sur une liste censée représenter les proverbes contemporains, et ce, sur la base de la compétence parémiologique plus ou moins développée du parémiologue et/ou de la connaissance des locuteurs natifs d'une langue. Il faudrait valider une telle contemporanéité par des données quantitatives. Cette évidence appartient aux résultats des enquêtes sociolinguistiques sur la familiarité des proverbes, aussi, mais elle ne peut non plus ignorer l'évidence offerte par la fréquence dans des corpus écrits ou oraux, malgré tous leurs défauts.

3.2.5. *Moon* (1998)

À la suite de son étude contrastive anglais-français avec Arnaud, dans sa monographie de 1998, Moon s'attache, en général, aux expressions figées et idiomatiques (*Fixed Expressions and Idioms* – FEI) anglaises, y inclus les proverbes.

Liste

Dans son ouvrage, Moon mentionne explicitement une base de données des expressions figées et idiomatiques où elle a aussi des proverbes. Il nous semble que la plupart des proverbes correspondent à ceux utilisés pour l'étude avec Arnaud (1993). Au moment de la présentation de la répartition des expressions de sa base de données, on apprend que les proverbes sont environ 274 proverbes (1998 : 63)[115], ce qui fait partiellement écho au total des 240 proverbes qu'elle a mentionné dans son étude avec Arnaud.

Sources

Comme dans Arnaud & Moon (1993), sa recherche porte sur les résultats obtenus grâce à l'*Oxford Hector Pilot Corpus* (OHPC) et ses 18 millions d'occurrences. À cet égard, Moon approfondit le problème de la taille du corpus. Son avis est tranchant quant au bien-fondé de son choix :

> « It became clear in the course of the present study that OHPC was too small to give conclusive information concerning transformations, inflection potential, and variations […] » (1998 : 48).

Cette remarque sur les résultats exhaustifs attendus concerne plutôt l'impossibilité de suivre le potentiel créatif de la surface des expressions figées et idiomatiques. Elle ne mentionne pas explicitement, comme on peut lire, la fréquence d'usage. On constate, en tout cas, que Moon se questionne sur la taille du corpus parce qu'au bout du compte, elle influence la fréquence repérée des expressions figées et idiomatiques. Aucune règle n'est vraiment précisée : il n'y a pas une taille de corpus à satisfaire pour mener à bien une recherche sur les expressions figées et idiomatiques. Il suffit de s'assurer, en gros, que le corpus soit le plus grand possible : *the larger*, *the better*.

Fréquence

Il vaut mieux expliquer d'abord que du questionnement sur son corpus suit la précaution de Moon sur la validité des résultats acquis :

[115] Moon indique que la macrocatégorie *formulae* (formules) se compose de 1.443 expressions, dont 19% sont des proverbes (1998 : 62-63). Nous avons calculé le nombre de 274 à partir de ces données.

> « I am *not* claiming […] that figures and statistics concerning events observed […] are universal truths. However, I *am* claiming that figures and statistics can be regarded as reasonable benchmarks, which may then be tested against or compared with other corpora […] » (1998 : 49 ; c'est l'auteure qui souligne).

C'est ce principe qui marque toute recherche empirique sur corpus. Les fréquences calculées sur corpus ne révèlent pas le véritable usage quantitatif d'un fait linguistique dans une langue donnée, et cela est vrai aussi – on dirait même *surtout* – pour les proverbes. Il n'en reste pas moins vrai que ses fréquences sont exploitables et, comme Moon l'affirme, elles représentent des bancs d'essai (*benchmarks*) pour des recherches similaires sur d'autres corpus. Ces recherches peuvent ainsi les confirmer ou les réfuter ou devenir un complément d'enquête pour les études de familiarité.

Malgré cette précaution sur les chiffres, Moon est persuadée que ses résultats seront probablement identiques aux résultats obtenus par l'interrogation d'autres corpus :

> « I am also claiming that, in spite of the shortcomings of OHPC, gross tendencies observed in it are likely to be observed too in other corpora […] albeit with different distributions […] » (1998 : 49).

Ces tendances quantitatives brutes observées seront également observées lors d'une recherche similaire sur des corpus comparables d'anglais britannique. L'analyse quantitative fait ainsi la rencontre de son contrepoint dialectique : l'analyse qualitative, notamment la distribution quantitative calculée d'après un critère qualitatif. Comme on l'avait indiqué au § 3.2.3., l'OHCP est composé, pour l'essentiel, de textes tirés de la presse britannique, quotidiens (59.5%) et magazines (6.6%) (1998 : 48) et sous-représente d'autres genres écrits ainsi que l'oral (3.1%) (*ibid*.). En gros, les fréquences sont aussi fonction des genres dont le corpus est composé.

Pour présenter ses résultats, Moon se concentre uniquement sur les fréquences brutes (au plus normalisées par million d'occurrences) et laisse ainsi de côté d'autres mesures statistiques, tels le *t-score* et l'information mutuelle (*MI*) (1998 : 58). Son calcul inclut les variations et les adaptations en discours des expressions, hormis certaines transformations et réductions syntaxiques (1998 : 59). Ce choix, peu motivé du point de vue linguistique, est suivi par un autre motivé par la statistique probabiliste. Moon fixe le seuil pour que *f* d'une

expression soit statistiquement significative à 5 occurrences[116] (1998 : 57). De même, Moon décide de représenter la distribution des fréquences par 9 plages de valeurs (1998 : 59), où « the dividing-lines are **arbitrary** » (*ibid*., c'est nous qui soulignons). D'après la significativité statistique qu'elle y confie, on résume les 9 plages comme suit :

- *f* non significative ($f \leq 4$) qui correspond aux plages I et II ;
- *f* basse ($5 \geq f \leq 35$) pour les plages III (de 5 à 17 occurrences brutes) et IV (de 18 à 35, à savoir entre 1 et 2 occurrences normalisées par million) ;
- *f* moyenne ($36 \geq f \leq 899$) de la plage V à la plage VII (de 2 à 50 occurrences normalisées par million) ;
- *f* élevée ($f \geq 900$) pour les dernières bandes VIII et IX (de 50 à plus de 100 occurrences normalisées par million).

La macrocatégorie *formulae* (formules)[117] recouvre 21% des expressions figées et idiomatiques recherchées sur corpus (1998 : 61) et inclut aussi 274 proverbes, dont 59% ont un sens métaphorique (1998 : 63). La quasi-totalité des proverbes analysés a *f* non significative, c'est-à-dire que la plupart d'entre eux ont *f* brute entre 1 et 4 occurrences. Ils sont suivis par les proverbes dont $f = 0$ et par les très rares proverbes qui ont $5 \geq f \leq 35$ et $36 \geq f \leq 89$ (*ibid*.). D'où la conclusion que :

> « their occurrence or non-occurrence in corpora of this size is almost entirely a matter of chance » (*ibid*.).

Seulement quelques rares proverbes anglais atteignent une fréquence basse ou moyenne, mais Moon ne précise pas de quels proverbes il s'agit et s'il existe une corrélation entre leur fréquence et d'autres études de familiarité, au moins de ne s'appuyer sur la liste de proverbes proposée dans Arnaud & Moon (1993).

---

[116] En statistique probabiliste appliquée aux études en sciences sociales (y compris la linguistique), le seuil de 5% est d'habitude considéré comme le seuil pour qu'une hypothèse à tester (par exemple, la cooccurrence entre deux mots) soit acceptée. Implicitement, nous croyons qu'elle suppose que si, en l'occurrence, un proverbe a $f > 5$, alors sa présence dans le corpus n'est pas à attribuer au hasard et tel proverbe acquiert une importance particulière parmi les autres ainsi que dans le corpus même. Notre spéculation peut s'appuyer sur son observation que nous citons à la page suivante, quand elle commente la fréquence des proverbes dans son corpus (1998 : 63). Outre cette mention, elle n'explicite nulle part ce que nous venons d'expliquer.

[117] Outre les proverbes, cette macrocatégorie prévoit : 1010 *simple formulae* comme *you know* ; 29 *sayings* comme *an eye for an eye* (pour nous, un proverbe tronqué/réduit) ; 130 *similes* comme *as good as gold*. C'est nous qui avons dégagé la répartition exacte d'après le total de 1.443 FEI déclaré par Moon (1998 : 62) et le tableau qui reprend les pourcentages de chaque composant de la macrocatégorie *formulae* (1998 : 63).

D'après ces chiffres, toutefois, la répétition de cette expérience avec des corpus comparables produirait les mêmes résultats, du moins si l'on se tient à cette remarque de portée générale sur les fréquences des expressions figées et idiomatiques :

> « It is reasonably unlikely that FEIs with frequencies of 2 per million and above would fail to occur at all in comparable corpora. Similarly, it would be surprising if FEIs that do not occur in OHPC, or that occur with frequencies no better than chance, were then found to be highly frequent in comparable corpora » (1998 : 49).

À première vue, de telles conclusions auraient de quoi décourager d'autres démarches similaires. En revanche, on peut tout simplement émettre l'hypothèse que les proverbes et, en général, les séquences formulaires nécessitent d'un traitement et d'une évaluation de leurs fréquences d'occurrence qui leur sont propres, sans prendre en compte[118] et les paramètres lexicométriques traditionnels. Autrement dit, les plages arbitraires de fréquence identifiées par Moon présupposent que l'ensemble des expressions figées et idiomatiques soit assez homogène pour entreprendre une analyse quantitative qui se fonde sur des répartitions et sur des principes communs. D'ailleurs, Moon elle-même reconnaît que les expressions figées et idiomatiques sont composées du moins de deux mots graphiques et qu'il serait plus appropriées de les considérer comme un tout, à savoir comme des unités lexicales (1998 : 57), ce qui aurait impliqué – ou implique, en tout cas – une adaptation pondérée des mesures de la fréquence de chaque catégorie ou macrocatégorie d'expressions. Il en ressort que tout fait linguistique formulaire, y compris les proverbes, ont besoin de mesures quantitatives différentes par rapport à d'autres combinatoires linguistiques idiomatiques, notamment par rapport à toute combinatoire qui remplit une fonction discursive non évaluative et qui n'hérite pas d'une référence à une tradition culturelle et sociale. Cela est vrai pour le cas de l'anglais, mais il est raisonnable de pouvoir étendre cette considération à d'autres langues indo-européennes (§ 3.2.6.).

Un regard à la distribution des proverbes par genres textuels, notamment des 702 occurrences de 208 proverbes (1998 : 69), nous montre comment le calcul de la fréquence sur corpus ne s'écarte pas de la nature des textes qui le composent. Les occurrences des

[118] D'après les pourcentages de Moon (1998 : 63), seulement 43 formules simples atteignent une *f* brute entre 180 et 899 occurrences par million, alors que la *f* brute de 664 formules simples oscille entre 0 et 17 occurrences. À la différence des proverbes et des formules simples, quelques rares similitudes et citations ne dépassent pas le seuil des 35 occurrences maximales. Pareillement aux autres sous-catégories, la plus grande partie des similitudes (115 sur 130) et des citations (16 sur 29) ont une *f* brute entre 0 et 4 occurrences.

proverbes reflètent de manière quasi spéculaire la hiérarchie de la composition par genre de l'OHCP. Plus précisément, 71% des proverbes sont à repérer dans la presse britannique (environ 66% de l'OHCP). Les écrits de fiction et de non-fiction contiennent les mêmes pourcentages d'occurrence des proverbes (12% face à 11% de textes de fiction et face à 18% de textes de non-fiction de l'OHCP). L'oral inclut 3% des occurrences des proverbes. Si la presse prend ainsi le dessus, elle l'a, dans ce cas, par la composition de l'OHCP.

3.2.6. *Corpas Pastor* (1998)

En 1998, Corpas Pastor entreprend une étude sur la fréquence des parémies espagnoles sur corpus. Tout d'abord, il faut préciser qu'elle utilise l'hyperonyme *paremia* (parémie) pour dépasser les incertitudes et les disputes académiques sur les classifications typologiques (2003 [1998] : 89). Pourtant, elle qualifie le *réfran* (proverbe) comme « la paremia por excelencia » (2003 [1998] : 90) en raison du fait que le *réfran* remplit tous les critères de reconnaissance d'un proverbe énumérés par Arnaud (1991). On remarque ici que la réflexion interlinguistique sur le proverbe fertilise mutuellement le champ de la parémiologie. En outre, sans aucune hésitation, Corpas Pastor inclut les parémies dans le *continuum* phraséologique. Plus précisément, elle affirme que les parémies sont une :

> « clase de unidades fraseológica » qui « engloba, a su vez, distintos subtipos que no resultan fáciles de delimitar » (2003 [1998] : 89).

« Las parientes pobres » (2003 [1998] : 78) de la recherche en phraséologie récupèrent ainsi leur place dans cette étude basée sur corpus qui veut (à juste titre) remettre en question la soi-disant disparition des parémies.

Liste

Corpas Pastor démarre son étude d'une liste de départ composée de 100 parémies sélectionnées au hasard, dont la plupart sont reconnues comme des *refranes* (2003 [1998] : 90). Aucun lien n'est fait avec des études défaitistes sur le sort des parémies espagnoles. De même, la liste est établie sans tenir compte de la compétence parémiologique des locuteurs espagnols et d'études précédentes sur ce sujet, comme celle de Sevilla Muñoz & Diaz (1997).

Sources

Corpas Pastor fait siens les soucis et les considérations finales de Norrick (§ 3.2.1.) pour quantifier ainsi que pour qualifier le véritable usage des parémies espagnoles dans le *corpus Vox-Bibliograf* (CVB), un corpus de référence d'espagnol péninsulaire d'environ 10 millions d'occurrences (2003 [1998] : 86-87). Ce corpus équilibré entre discours écrit et discours oral se compose pour la plupart de textes échantillonnés qui datent de 1950 au début des années 1990 (*ibid*.). Le sous-corpus du discours écrit est essentiellement composé par des textes de non-fiction (35%) et de fiction (35%), ainsi que par des textes de presse (25%). Le sous-corpus du discours oral inclut des conversations ainsi que des dialogues, des entrevues et des monologues médiatisés (radio et télévision), tous représentés en pourcentages identiques (*ibid*.).

Outre le CVB, Corpas Pastor fait référence à un *corpus de citas* (CC), à savoir un corpus d'à peu près 2.000 citations qu'elle-même a recueillies dans la presse ou dans des textes de fiction au cours de 6 ans dans les années 1990 (2003 [1998] : 88).

Fréquence

Corpas Pastor repère 79 sur les 100 parémies espagnoles de sa liste, ce qui veut dire qu'environ 20% des parémies ont $f = 0$ (2003 [1998] : 90). Les 79 parémies dont $f \geq 1$ totalisent 166 occurrences, à savoir une moyenne brute (*type/token ratio*) égale à 2,10 (*ibid*.). 124 de ces 166 occurrences sont attestées dans le sous-corpus du discours écrit et, de manière similaire aux résultats d'Arnaud & Moon (1993) et de Moon (1998), la plupart d'entre elles intéressent la presse écrite (35.54%, soit 59 occurrences) (*ibid*.).

À partir des annexes de son étude, nous rétablissons la liste triée par ordre décroissant de fréquence brute des 10 parémies les plus fréquentes dans le CVB (2003 [1998] : 101-107) :

Rang	Parémie [traduction littérale]	*f* brute
1)	*La vida es sueño* [La vie est un rêve]	12
2)	*El tiempo es oro* [Le temps, c'est de l'or]	10
3)	*Un día es un día* [Un jour est un jour]	8
4)	*A río revuelto, ganancia de pescadores* [À fleuve troublé, gain de pêcheurs]	6
	De tal palo, tal astilla [De telle branche, telle écharde]	6
	Del dicho al hecho va mucho trecho [Du dit au fait il y a beaucoup de distance]	6
	El hombre es un animal político [L'homme est un animal politique]	6
8)	*A vivir, que son días* [À vivre, il y a des jours]	5
9)	*Agua pasada no mueve molino* [Eau passée ne bouge pas moulin]	4
	Si te he visto no me acuerdo [Si je t'ai vu je ne me souviens pas]	4

Tableau 10. Fréquence des 10 premiers proverbes espagnols les plus fréquents d'après Corpas Pastor (1998).

Les parémies dépassent à peine le seuil de $f = 10$. Précisons encore que la parémie en tête de liste n'est pas un proverbe en soi, mais le titre d'une œuvre homonyme et une partie d'un extrait (proverbialisé) de cette œuvre rédigée par Caldéron de la Barca au XVII^{e} siècle. Le premier proverbe apparaît, en revanche, au deuxième rang (*El tiempo es oro*).

Dans le CVB, Corpas Pastor repère respectivement 21.08% (35 occurrences) dans les textes de fiction et 18.07% (30 occurrences) dans les textes de non-fiction. Par rapport au nombre exigu de parémies identifiées par Arnaud & Moon (1993) et Moon (1998), le sous-corpus oral du CVB délivre 25.3% (42 occurrences) des résultats totaux.

Outre la répartition en genres, Corpas Pastor indique aussi les pourcentages qui relèvent de la fréquence d'occurrence des parémies selon les processus de variation reconnus. Elle souligne, d'abord, que, sur un total de 166 occurrences, environ 40% témoignent au moins une variation formelle (2003 [1998] : 91), à savoir environ 66 occurrences. De ces occurrences parémiques modifiées, elle détaille les pourcentages pour chaque processus de variation. Nous synthétisons ses résultats (2003 [1998] : 92-94) dans le Tableau 11 où l'on remarque que le processus le plus rencontré dans le CVB est la réduction :

PROCESSUS DE VARIATION	% (SUR ENVIRON 66 OCCURRENCES PAREMIQUES)
Réduction d'un ou de plusieurs éléments formels	54.84%
Substitution d'un ou de plusieurs éléments formels	16.38%
Variation morphosyntaxique	7.15%
Adjonction d'éléments formels dans la parémie	5.61%
Enchevêtrement de processus	14.84%

Tableau 11. Répartition des occurrences des proverbes espagnols d'après les 5 processus de variation identifiés par Corpas Pastor (1998)[119].

Par la suite, Corpas Pastor compare les résultats obtenus sur CVB avec son *corpus de citas* (CC). Après la recherche des 100 parémies dans son CC, elle identifie 76 parémies et 103 occurrences, ce qui équivaut à une moyenne brute (*type/token ratio*) de 1,36 (2003 [1998] : 91). Il en suit qu'environ 25% des parémies de sa liste de départ ont $f = 0$ dans le CC.

Lors de la comparaison intralinguistique de ces données quantitatives, Corpas Pastor constate, d'une part, les avantages qui découlent du choix d'un corpus de référence – même si elle ne les mentionne pas ouvertement – et d'autre part, elle souligne que :

> « Los datos […] ponen de manifiesto la alta frecuencia de uso de las paremias en el discurso » (2003 [1998] : 91).

Certes, les données quantitatives prouvent l'usage des parémies tant à l'écrit qu'à l'oral. Toutefois, il reste à comprendre de quelle manière interpréter ces mêmes données et surtout

[119] La somme des pourcentages partiaux atteint 98.82%. Il peut y avoir des calculs erronés de la part de l'auteure.

de quelle façon justifier que la fréquence d'usage des parémies espagnoles (d'ailleurs, sélectionnées au hasard) est élevée (*alta*). Comme pour Moon (§ 3.2.5.), mais à l'inverse, la mesure statistique des parémies remet en question le traitement lexicométrique usuel. Les seuils et les plages de valeurs sont restés (et restent) assez arbitraires et le peaufinement des résultats est souvent fonction des connaissances personnelles des auteurs. La justification sur la valeur discursive des parémies, à savoir que :

> « La frecuencia de uso de las paremias no se explica por una simple cuestión de estética ; obedece más bien a necesidades funcionales y de construcción del texto » (2003 [1998] : 95).

ne peut non plus suffire à motiver « la alta frecuencia ». De même, les $f \geq 5$ (seuil minimal de *f* basse pour Moon) dans CVB de certaines parémies comme :

> *De tal palo, tal astilla* [trad. litt. : De telle branche, telle écharde][120] (6 occurrences) (2003 [1998] : 102)

ou :

> *El tiempo es oro* [Le temps, c'est de l'or][121] (10 occurrences) (2003 [1998] : 104)

ne sont pas vraiment mises en relation ni avec leur $f = 0$ dans CC, ni avec des tests de familiarité a posteriori pour essayer de repérer d'autres ancrages.

3.2.7. *Järv* (1999)

La recherche de Risto Järv sur l'usage des proverbes estoniens dans la presse vise plutôt l'observation de leur contexte d'usage que leur fréquence (1999 : 81).

Liste

Järv ne part ni d'une liste préétablie de proverbes à rechercher. Il s'appuie seulement sur des introducteurs des proverbes en discours (§ 4.2.1.7.). Il est conscient qu'une liste

[120] Correspondant français suggéré par *El Refranero Multilingüe* : *Tel père, tel fils*.
[121] Correspondant français suggéré par *El Refranero Multilingüe* : *Le temps, c'est de l'argent*.

préétablie de proverbes à rechercher, et tirée d'un recueil spécialisé aurait apporté des résultats plus fiables, et ce, malgré la difficulté des variantes proverbiales à surmonter (1999 : 82).

Sources

Malgré les intentions qualitatives, Järv présente également un aperçu quantitatif basé sur le dépouillement de deux corpus de presse. Plus précisément, il a analysé la fréquence d'occurrence des proverbes estoniens dans les articles des journaux en ligne *Postimees* et *Eesti Päevaleht* pour la période entre octobre 1995 et novembre 1997 (1999 : 101).

Fréquence

Compte tenu des techniques assez limitées d'interrogation du corpus, ses résultats indiquent qu'environ 300 articles contiennent au moins un proverbe (1999 : 82) et que le mot-clé *vanasõna* (proverbe) possède la fréquence d'occurrence la plus élevée dans son corpus (154 sur 231 occurrences totales pour les quatre mots-clés recherchés) (1999 : 81).

Compte tenu de la taille réduite de son corpus, Järv évalue la répartition d'occurrences des proverbes dans les différentes rubriques des deux journaux. Au total, les rubriques consacrées à la politique et aux lettres des lecteurs contribuent presque dans la même mesure (19% et 18.6% respectivement), suivie par les rubriques de culture, de sport et *people* (1999 : 88-89). Il déclare que les résultats des rubriques en tête de liste ne sont pas vraiment étonnants. À son avis, les politiciens tendent à utiliser des proverbes pour que les gens ne les oublient pas et leur fassent confiance, et que les lettres des lecteurs contiennent souvent des protestes moralisantes et didactiques (1999 : 89-90). Mises à part ces constatations, Järv signale aussi que la rubrique des lettres des lecteurs du journal *Postimees* contribue massivement (33 occurrences) par rapport à la même rubrique du *Eesti Päevaleht* (10 occurrences) (1999 : 89). Pourtant, il passe sous silence trois répartitions à peu près identiques entre les deux journaux. Les rubriques politique, culture et *people* ont presque le même nombre d'occurrences de proverbes dans chaque journal. Autrement dit, ces trois rubriques-contextes sont intéressées par l'usage des proverbes de presque la même mesure.

Järv rentre également dans les détails de la répartition par auteurs et quoiqu'avec précaution, il affirme que les hommes agrémentent leurs textes de proverbes plus que les femmes (1999 : 92).

Après cette observation sociolinguistique, il donne aussi la liste des proverbes repérés par les marqueurs et dont la fréquence est la plus élevée. On reprend ci-dessous la liste triée par ordre décroissant (1999 : 98-99) :

Rang	**Parémie** **[traduction littérale]**	***f* brute**
1)	*Üheksa korda mõõda, üks kord lõika* [Mesurer neuf fois, couper une fois]	7
2)	*Väiksed vargad ripuvad võllas,* *suured sõidavad tõllas* [Les petits voleurs suspendus à la potence, les grands guident dans un char]	5
3)	*Parem hilja kui mitte kunagi* [Mieux vaut tard que jamais]	4
4)	*Pada sõimab katelt,* *ühed mustad mõlemad* [La casserole se moque de la bouilloire, toutes les deux sont noires]	3
	Parem pool muna kui tühi koor [Mieux vaut la moitié d'un œuf qu'une coquille vide]	3
	Ära enne vana kaevu täis aja, kui uus valmis ei ole [Ne remplit pas le vieux puits si le nouveau n'est pas prêt]	3
	Pea tehtud pilla-palla, *kaua tehtud kaunikene* [Ce qui est fait en vitesse est de la pagaille, ce qui est fait doucement est bien]	
	Teise silmas näed pindu, *enda silmas ei näe palki* [La paille qui est dans l'œil de l'autre et ne pas regarder la poutre dans son œil]	3
	Tasa sõidad, kaugemale jõuad [Le plus lent vous conduisez, davantage vous obtenez]	3

Pärast riidu ei ole tarvis rusikad näidata [Après la querelle il n'est pas nécessaire de montrer les poings]	3

Tableau 12. Fréquence des 10 premiers proverbes estoniens les plus fréquents d'après Järv (1999).

Järv observe tout de suite que les proverbes qu'il estime être parmi les plus connus sont exclus de cette liste, et ce, parce que la proverbialité d'un énoncé doit être marquée, à son avis, surtout pour des proverbes peu connus (1999 : 99). Nous ne sommes pas spécialiste de l'estonien pour pouvoir remettre en question un tel constat, mais il nous est difficile de croire qu'au moins 2 proverbes de cette liste – *Parem hilja kui mitte kunagi* et *Parem pool muna kui tühi koor*[122] – ne soient pas des proverbes acquis par des locuteurs natifs estoniens[123]. De toute façon, cette observation de Järv permet de remarquer encore une fois comment toute étude de fréquence des proverbes finit par mettre ses résultats en relation avec la familiarité, et que fréquence et familiarité sont nécessairement en rapport l'une avec l'autre.

Nous remarquons, pour finir, que les fréquences d'occurrence de cette liste de proverbes 'marqués' effleurent à peine le seuil de la dizaine. Ces performances sont décidément en ligne avec la plupart des études mentionnées dans notre revue de la littérature. Encore comme pour d'autres études, nous constatons que les proverbes tendent à se regrouper autour d'une même valeur de fréquence qui correspond, en l'occurrence, à 3.

3.2.8. *Čermák* (1998, 2003)

Čermák essaie d'établir un lien entre linguistique de corpus, fréquence et *minimum parémiologique*. Depuis son introduction par Permyakov, toutes les études successives en plusieurs langues n'ont en effet visé que la familiarité des proverbes pour définir un ensemble minimal de 'proverbes à porter' par chaque locuteur natif. Les contributions de Čermák ont le mérite de faire en sorte que la fréquence (à l'écrit) participe en tant que paramètre quantitatif pour l'établissement du minimum parémiologique, quoique de manière différente. D'une

[122] Pour ces deux proverbes, on constate au passage un schéma/cadre lexico-grammatical discontinu commun : *Parem… kui… .*

[123] Une recherche de la forme canonique de ces 2 proverbes dans le corpus de référence estonien *EstonianRC* (249.745.108 occurrences) consultable grâce à l'interface *Sketch Engine* (Kilgarriff *et al*. 2004) nous donne les fréquences brutes suivantes : 119 occurrences pour *Parem hilja kui mitte kunagi* et 49 occurrences pour *Parem pool muna kui tühi koor*. Cet usage, sans faire d'ailleurs recours à aucun marqueur de proverbialité, ne suggérerait pas vraiment un manque de connaissance chez les locuteurs estoniens.

part, en 1998, la référence au minimum parémiologique sert de prétexte pour remettre en question la relation entre familiarité et fréquence. D'autre part, en 2003, le minimum parémiologique est identifié tout court avec la fréquence d'usage sur corpus.

Liste

Les deux études de 1998 et 2003 divergent pour les listes de départ des proverbes.
Pour la première, Čermák utilise une liste de proverbes tchèques familiers mise au point par Franz Schindler en 1993. Cette liste inclut 99 proverbes triés par indice de connaissance (la valeur la plus basse retenue étant 78.5%).
Pour la deuxième, Čermák part d'une base de données d'unités phraséologiques collectées durant les 15 ans qui précèdent son étude. Il décide de calculer la fréquence de 241 proverbes tchèques[124] disponibles dans sa base à la façon de Moon (§ 3.2.5.). Les critères de sélection sont – tautologiquement, à un premier abord – la fréquence et la contemporanéité (2007 [2003] : 599). À la place de la familiarité comme biais de constitution de la liste de sa première étude, Čermák choisit l''actualité' des proverbes, sans donner aucune explication à cet égard[125].

Sources

Čermák contribue à la création et à la croissance progressive du *Czech National Corpus* (CNC)[126]. Pour ses recherches, Čermák tire parti de la composante SYN2000 qui compte environ 23 millions d'occurrences en 1998 (2007 [1998] : 570) et 100 millions d'occurrences en 2003. Pour cette dernière mise à jour, Čermák spécifie que SYN2000 se compose d'écrit (période entre 1990-1999) et se divise en : 60% de textes de presse, 15% de textes de fiction et 25% de textes de non-fiction (dont 14% de textes en sciences humaines et sociales et 11% en sciences non sociales) (2007 [2003] : 604).

Il souligne également que SYN2000 est un corpus équilibré, et on apprend du site qu'il est aussi lemmatisé et annoté morphosyntaxiquement.

[124] Au même endroit (2007 [2003] : 599), il parle d'une liste de 243 proverbes et, quelques lignes après, de 241 proverbes. De la lecture intégrale de sa contribution, nous avons retenu ce dernier chiffre.
[125] Nous précisons que Čermák lui-même arrive à la conclusion que ces comparaisons ne sont pas faisables à cause de divergences méthodologiques importantes (2007 [2003] : 605).
[126] Pour des renseignements supplémentaires, on peut visiter le site : http://ucnk.ff.cuni.cz/english/struktura.php (date de consultation : 11/11/2013).

Malgré tout, un corpus de 100 millions d'occurrences, pour le chercheur, n'est pas encore un corpus satisfaisant (2007 [2003] : 605).

Fréquence

En 2003, Čermák repère un total de 2.776 occurrences (2007 [2003] : 598-599), à savoir une moyenne de 12 occurrences par proverbe. Il estime ainsi que 0.128% du SYN2000 est composé de proverbes[127] et qu'on peut reconnaître un proverbe tous les 36.000 mots graphiques du SYN2000 (2007 [2003] : 599). Il souligne que « much more data is needed » (2007 [2003] : 606) et répète exactement ce que Norrick avait affirmé une vingtaine d'années auparavant (§ 3.2.1.). En l'occurrence, le constat empirique est prouvé par les résultats obtenus de son étude de 1998. Dans le (sous-)corpus d'environ 23 millions d'occurrences, en effet, le chercheur repère un total de 284 occurrences proverbiales, à savoir un proverbe tous les 80.000 mots graphiques (2007 [1998] : 577). L'augmentation de la taille du corpus a effectivement contribué à une augmentation des occurrences moyennes[128].

À propos des seuils de fréquence, citons maintenant ce que Čermák dit sur l'étendue du minimum parémiologique et des paramètres quantitatifs à adopter :

> « No really convincing criteria have been offered for the **size of a paremiological minimum**. […] its selection, somewhere on the decreasing frequency scale, is rather an arbitrary one. Thus, […] only those proverbs are listed whose frequency is 10 or higher » (2007 [2003] : 606).

Vu la difficulté à dégager des critères autres que la fréquence décroissante, Čermák revient à l'aléa d'un seuil minimum qu'il fixe à $f = 10$ pour qu'un proverbe rentre dans un minimum parémiologique tchèque potentiel[129]. Il double ainsi le seuil de 5 établi par Moon (§ 3.2.5.) et obtient une liste de 100 proverbes.

Apparemment, Čermák ne définit pas d'autres seuils de fréquence. Nous remarquons, toutefois, qu'au moment de la présentation des fréquences, il préfère introduire seulement les

[127] Ce calcul dérive de la longueur moyenne des proverbes tchèques que Čermák estime à 4.6 mots graphiques.
[128] La référence à la taille du corpus pour relativiser (à juste titre) ses résultats cache aussi, à notre avis, un préjugé théorico-méthodologique diffusé : les proverbes les plus fréquents doivent coïncider nécessairement avec les proverbes les plus familiers (§§ 3.2.3.).
[129] À notre avis, plutôt que de fixer un seuil minimal pour *f*, il serait question de se demander quelle place occupe *f* dans corpus en rapport avec les critères qualitatifs choisis pour des études de familiarité. La compétence parémiologique des locuteurs natifs doit faire partie de toute considération d'où la nécessité de prendre en compte familiarité et fréquence pour qu'un minimum parémiologique soit établi (Zouogbo 2011 ; Marcon 2013). Par conséquent, le corpus ne peut et ne doit pas remplacer les enquêtes sociolinguistiques, mais il peut et doit constituer leur support à la fois complémentaire et indépendant.

proverbes dont $f \geq 5$ (2007 [2003] : 605), de façon tout à fait similaire à Moon. Un regard à ces fréquences lui permet d'élaborer une corrélation qu'il met d'ailleurs en gras dans son texte et qui peut s'appliquer aussi aux résultats de Moon et de Corpas Pastor (§ 3.2.6.) :

> « **proverb frequency is inversely proportional to their number in the rank** » (2007 [2003] : 605).

En gros, la plupart des proverbes tchèques présentent globalement une *f* basse et au fur et à mesure que les fréquences brutes respectives diminuent, le nombre de proverbes augmente de façon conséquente. Plus précisément, selon notre interprétation des données fournies par Čermák (*ibid.*)[130] :

- 38% des proverbes tchèques analysés (91 sur 241) ont $0 \leq f > 5$,
- 43% (104 sur 241) possèdent $5 \leq f \geq 15$ – les pics représentés par 16 proverbes à $f = 5$ et 16 autres proverbes à $f = 7$ –,
- alors que 19% (46 sur 241) dépassent $f > 15$.

Čermák spécifie encore que :

> « this correlation does not represent a smooth curve, as there seems to be a prominent break around the frequency band of 12 » (*ibid.*).

En revanche, nous sommes de l'avis que la véritable rupture se réalise déjà à partir des 7 proverbes ayant $f = 15$, d'où notre choix de présentation des données en plages de fréquences[131]. Comme pour Moon (§ 3.2.5.) et pour Corpas Pastor (§ 3.2.6.), les chiffres confirment que l'estimation de la fréquence des proverbes ne peut être assimilée à celle de toute autre unité lexicale simple ou composée, ce qui implique l'établissement de mesures, du moins de seuils adaptés pour l'interprétation des données quantitatives.

[130] Par un tri décroissant, Čermák énumère les *f* et leur nombre de proverbes respectif dans une liste à plat, sans aucun regroupement par bandes de fréquence et pourcentages. C'est nous qui avons calculé ces pourcentages sur la base de ses données brutes.

[131] Les 7 proverbes ayant $f = 15$ sont suivis par 7 autres proverbes dont $f = 14$, par 4 proverbes de $f = 13$ et par 10 proverbes qui ont $f = 12$. Čermák choisit ainsi la dizaine de proverbes comme *prominent break*. En revanche, nous estimons que les 7 proverbes dont $f = 15$ constituent la véritable rupture dans une liste où seul 4 autres proverbes ont $f = 17$. D'ailleurs, à toute autre $f \geq 16$ ne correspondent pas plus de 3 proverbes.

Ci-dessous, nous faisons une synthèse des listes de fréquence que Čermák a rédigées sur la base de ses deux études sur corpus. Nous présentons seulement les 10 proverbes les plus fréquents qui comptabilisent à eux seuls : 86 sur 284 occurrences dans (2007 [1998] : 579) et 585 sur 2.776 occurrences dans (2007 [2003] : 606-607) :

Rang	Proverbe [traduction littérale]	*f* brute
1)	*Pozdě bycha honit* [Ce serait tard de continuer]	11
	Lepší vrabec v hrsti než holub na střeše [Mieux vaut moineau en main que pigeon sur toit]	11
	Není šprochu, aby na něm nebylo pravdy trochu [Pas de Sproch, s'il n'y avait pas un peu de vérité]	11
4)	*Vlk se nažral a koza zůstala celá* [Le loup mange sa nourriture et la chèvre reste entière]	10
5)	*Boží mlýny melou pomalu, ale jistě* [Les moulins de Dieu moulent lentement, mais sûrement]	9
	Pod svícnem bývá tma [Sous le chandelier il est sombre]	9
7)	*Ráno moudřejší večera* [Matin plus sage que soir]	7
8)	*Bližší košile než kabát* [Plus loin de chemise que de manteau]	6
	Kdo jinému jámu kopá, sám do ní padá [Qui creuse une fosse, lui-même tombera]	6
	Kdo chce psa bít, hůl si najde [Qui veut battre un chien, trouve le bâton]	6

Tableau 13. Fréquence des 10 premiers proverbes tchèques les plus fréquents d'après Čermák (1998)[132].

[132] Nous nous sommes arrêté aux 3 premiers proverbes triés par ordre alphabétique dont *f* = 6. Les 5 autres proverbes mentionnés dans cette liste et dont *f* = 6 ne sont pas présents dans la liste des 10 proverbes les plus fréquents du Tableau 11.

Rang	Proverbe [traduction littérale]	*f* brute
1)	*Účel světí prostředky* [La fin justifie les moyens]	89
2)	*Nic není zadarmo* [Rien n'est gratuit]	88
3)	*Oko za oko, zub za zub* [Œil pour œil, dent pour dent]	76
4)	*Mnoho povyku pro nic* [Tant de bruit pour rien]	71
5)	*Pravda vítězí* [La vérité l'emporte]	50
6)	*Vlk se nažral a koza zůstala celá* [Le loup mange sa nourriture et la chèvre reste entière]	48
7)	*Naděje umírá poslední* [L'espoir meurt en dernier]	44
8)	*Všechno zlé je k něčemu dobré* [Chaque nuage a une lueur d'espoir]	40
	Za málo peněz málo mugik [Pour peu d'argent peu de musique]	40
10)	*Boží mlýny melou pomalu, ale jistě* [Les moulins de Dieu moulent lentement, mais sûrement]	39
	Stará láska nerezaví [Le vieil amour ne meurt jamais]	39

Tableau 14. Fréquence des 10 premiers proverbes tchèques les plus fréquents d'après Čermák (2003).

Nous remarquons que rien que 2 proverbes (*Vlk se nažral a koza zůstala celá* et *Boží mlýny melou pomalu, ale jistě*) sont communs aux deux listes. La taille du corpus (sans oublier les deux listes de départ) fait la différence.

Un regard aux chiffres nous montre également que la variation de taille impose des écarts de fréquence significatifs que nous avons aussi constaté pour d'autres études mentionnées jusqu'à présent. Dans le cas de la liste établie en 1998, par exemple, les 10

proverbes en tête occupent des rangs différents en fonction d'une ou maximum deux occurrences. En revanche, les 10 proverbes en tête de liste en 2003, notamment les proverbes des rangs les plus élevés, agissent comme des *outliers*, à savoir des données ayant des valeurs extrêmes et externes par rapport au reste de la distribution de proverbes. Les proverbes des rangs les plus élevés s'écartent entre eux d'une dizaine d'occurrences (ex. rangs 1 et 2 vs. rangs 3 et 4) et d'une vingtaine-trentaine d'occurrences des restants.

Il paraît évident que l'établissement de seuils de fréquence dépend de la taille du corpus qu'on interroge. En outre, les deux expériences de Čermák nous suggèrent que la taille idéale d'un corpus pour des spéculations raisonnables sur l'usage des proverbes doit s'attester dans les alentours de la centaine de millions d'occurrences. Le regroupement des fréquences des proverbes autour de valeurs très similaires dans un corpus de taille inférieure ne permet pas de dégager des préférences d'usage, compte tenu des aléas de composition du corpus même (genres textuels, sujets, etc.).

À côté de ces remarques, le regroupement de la plupart des fréquences des proverbes autour de valeurs (relativement) basses nous indique deux parcours de réflexion. D'une part, le regroupement reflète la corrélation qu'on observe entre formes (*types*) et occurrences (*tokens*) dans un corpus. Ce qui veut dire que le taux d'augmentation des occurrences dans un corpus est normalement plus élevé que le taux d'augmentation des formes (Habert *et al*. 1997 : 197-198). D'autre part, les regroupements de proverbes et les proverbes-*outliers* nous invitent à prendre en considération ces valeurs comme des indications pour l'établissement de seuils de fréquence. Autrement dit, la distribution des proverbes dans un corpus semble suggérer que les trois mesures de tendance centrale, comme la *médiane*, la *mode*[133] et la *moyenne*, accompagnées éventuellement de quelques mesures de dispersion[134], peuvent constituer de premiers repères statistiques pour caractériser différents seuils de fréquence dans une même distribution de proverbes.

Pour revenir à Čermák, il se concentre sur les 10 proverbes les plus fréquents de la liste de 2003 pour présenter la répartition de leurs occurrences dans les genres textuels du CNC (2007 [2003] : 604, 606)[135]. Il est intéressant de se pencher non seulement sur les données quantitatives observées pour cet échantillon, mais surtout sur les considérations que Čermák nous offre. D'après ses moyennes et ses pourcentages, la plupart des occurrences des 10 proverbes sont attestées dans les textes de presse (64.4%).

[133] Pourvu qu'on ignore les proverbes dont $f = 0$, évidemment.
[134] Par exemple les *quartiles*, la *variation standard* et le *coefficient de variation*.
[135] Pour les 99 proverbes de son étude de 1998, il décrit plutôt sur les usages qu'il appelle prototypiques et non-prototypiques en contexte (2007 [1998] : 572-576).

Sous-corpus CNC-SYN2000	% sur total CNC-SYN2000	% occurrences proverbiales sur total CNC-SYN2000
Fiction	15	19.8
Presse	60	64.4
Sciences sociales	14	13.7
Sciences non sociales	11	2

Tableau 15. Répartition des occurrences des proverbes tchèques d'après les 5 processus de variation identifiés par Čermák (2003)[136].

Pourtant, Čermák affirme que la contribution majeure en termes absolus est à attribuer aux textes de fiction (19.8%) du CNC (2007 [2003] : 604). Plus précisément, Čermák estime que l'écart entre le pourcentage d'occurrence de ces 10 proverbes dans les textes de fiction (19.8%), et le pourcentage des textes littéraires dans le corpus (15%) est statistiquement significatif par rapport à l'écart entre le pourcentage d'occurrence dans les textes de presse (64.4%) et le pourcentage des textes de presse du CNC-SYN2000 (60%). Compte tenu du déséquilibre du corpus en faveur des textes de presse[137], on comprend qu'on aurait prévu une meilleure performance du sous-corpus des textes de presse par rapport au sous-corpus des textes de fiction. Nous croyons, pourtant, qu'il est possible de nuancer la performance du sous-corpus littéraire avec un peu de recul. D'une part, il faut considérer les performances de deux autres sous-corpus de textes en sciences humaines et sociales et en sciences non sociales. D'autre part, il faut réfléchir sur le choix de présentation des données quantitatives des proverbes examinés.

Quant aux performances des autres sous-corpus, notamment du sous-corpus de textes en sciences humaines et sociales, il n'existe pas un véritable écart entre le pourcentage d'occurrence des 10 proverbes (13.7%) et le pourcentage que ce sous-corpus représente dans le CNC (14%), ce que Čermák commente comme « somewhat surprisingly » (*ibid*.). Au contraire, rien que 2% des occurrences de proverbes sont la contrepartie de 11% de textes que

[136] La somme des pourcentages de la colonne des occurrences proverbiales atteint 99.9%. Les arrondissements des pourcentages sont présentés par Čermák. C'est nous, par contre, qui avons créé ce tableau pour mieux visualiser les données quantitatives de son étude.
[137] Ce qui remet en question aussi la qualification de corpus représentatif attribuée au CNC-SYN2000 (2007 [2003] : 598).

le sous-corpus en sciences non sociales occupe dans le CNC. Il est intéressant de souligner comment les genres textuels que Moon et Corpas Pastor ont étiquetés comme textes de non-fiction (§§ 3.2.5.-3.2.6.) influencent différemment les fréquences des proverbes. L'expérience de Čermák montre que la distinction entre sciences sociales/humaines et sciences non sociales influence la fréquence enregistrée pour les proverbes[138].

En ce qui concerne le choix de présentation des données quantitatives, l'auteur oublie de mettre sous la loupe les écarts de fréquence les plus évidents, c'est-à-dire ceux qui intéressent chaque proverbe dans chaque sous-corpus. Par exemple, par rapport aux sous-corpus de textes de fiction et textes de presse, les occurrences des deux proverbes les plus fréquents :

Účel světí prostředky
Nic není zadarmo

privilégient de manière frappante le sous-corpus de textes de presse (73.3% et 81% respectivement) et délaissent les textes de fiction (16.3% et 5.9% respectivement). Au contraire, les occurrences du proverbe :

Naděje umírá poslední

se répartissent de façon décidément plus équilibrée dans les deux sous-corpus (36% en fiction et 39.5% dans les textes de presse). Peut-être par souci de concision, Čermák passe sous silence ces écarts qui soulèvent plusieurs interrogatifs sur son étude ainsi que sur la nôtre[139]. D'abord, le *degré d'idiomaticité* (Schmale 2013 : 35-36) et l'*ancrage pragmatique* (Schmale 2013 : 36-37) de chaque proverbe, c'est-à-dire l'analyse de la portée sémantique, référentielle et situationnelle (sociale et stylistique) dans le discours (en l'occurrence, écrit) et de l'emploi pertinent (cohérent ou volontairement humoristique) du proverbe en discours. Par la suite, en

[138] Il est vrai aussi que la répartition des disciplines pour ces deux sous-corpus et le choix des genres textuels ont affecté le calcul de la fréquence. D'une part, nous croyons que l'attribution du domaine de l'économie au sous-corpus des sciences non sociales (2007 [2003] : 604) est tout à fait discutable, alors que sa place naturelle demeurerait davantage parmi les sciences sociales. D'autre part, Čermák lui-même avoue un excès d'aisance lors du choix des textes, notamment des journaux sur la chasse et sur le bricolage, comme partie du sous-corpus en sciences non sociales (*ibid*.). C'est en gros à ces textes que Čermák attribue ce 2% d'occurrences (ou, plus précisément, de 6 proverbes sur 10) dans le sous-corpus en sciences non sociales, et ce, en raison de : « commentaries used and a loose type of style » (*ibid*.).

[139] Čermák a analysé les fonctions et les usages des proverbes sur corpus, mais par généralisation de certaines tendances (comme le veut, d'ailleurs, la linguistique de corpus), sans se concentrer sur des études de cas spécifiques.

(macro- et micro-)diachronie, les écarts entre sous-corpus invitent à s'intéresser à l'*historique du proverbe* (Conenna 2002) autant en d'une perspective formelle (morphosyntaxique et lexique), mais aussi pour ce qui concerne la *proverbialisation* (Schapira 2000) de certaines séquences formulaires. À ce propos, on peut prendre encore le cas de *Účel světí prostředky* [La fin justifie les moyens] qui, au départ aphorisme, finit par être éloigné de son auteur/énonciateur (Nicolas Machiavel) et arraché de son genre textuel d'origine (« *De Principatibus* », traité en sciences sociales du XVI^{e} siècle) pour être réutilisé dans d'autres genres textuels (notamment des articles de presse).

Pour conclure, l'originalité des études de Čermák nous invite à ne pas sous-estimer la présentation et l'interprétation des résultats quantitatifs et à faire recours à d'autres mesures en statistique descriptive, outre la seule moyenne arithmétique et les pourcentages. Quant à l'établissement du corpus, les écarts concernant la fréquence de chaque proverbe dans les sous-corpus nous stimulent à bien tenir compte de la composition de notre corpus.

3.2.9. *Ďurčo* (2005, 2006)

D'après ce que nous avons appris par la contribution de Grzybek (2009 : 219-220) et de Zouogbo (2011 : 101-102, 104), Peter Ďurčo s'intéresse à l'étude de fréquence des proverbes allemands, et ce, en vue de la constitution d'un minimum parémiologique allemand.

Liste

Pour son étude de 2005, le chercheur fait recours à une liste de 151 proverbes allemands préalablement indiqués comme familiers. Plus précisément, Ďurčo a utilisé la liste de proverbes et leurs indices de familiarité disponibles dans une étude pilote menée par Grzybek (1991) ainsi que d'autres proverbes familiers qu'il obtient par d'autres enquêtes (Grzybek 2009 : 229). L'estimation de la familiarité précède celle de la fréquence, les deux entrelacées.

Ce qui est confirmé par l'étude de 2006 où l'établissement de la liste de 385 parémies (Zouogbo 2011 : 101) résulte d'un filtrage qui part de la consultation des recueils parémiographiques et passe par la suite à travers le jugement des spécialistes (non seulement des parémiologues). De par leur compétence en matière et leur connaissance des parémies, ils

produisent une liste à soumettre à une analyse de fréquence sur corpus. Ďurčo enlève ces parémies dont $f = 0$ et enrichit sa liste avec d'autres parémies éventuellement reconnues dans le corpus. C'est donc par l'ensemble norme parémiographique-spécialistes-corpus qu'on débouche sur la liste à soumettre aux sujets interpellés pour l'étude de familiarité. Contrairement à ce qu'il a effectué en 2005, Ďurčo démarre cette fois-ci d'une mesure de fréquence à la fois parémiographique et sur corpus nuancée par le jugement de familiarité des spécialistes pour aboutir à la mesure de familiarité de la part des non-spécialistes.

Sources

Pour le calcul de la fréquence, Ďurčo exploite la base de données *Mannheim Cosmas II* gérée par l'*Institut für Deutsche Sprache* qui, à présent, inclut un corpus (*DeReKo-Korpus*) d'environ 5 milliards de mots, à savoir 5 millions de pages de livres numérisés[140]. Si l'on se tient à Zouogbo (2011 : 101), la taille du corpus au moment de la consultation par Ďurčo en 2006 tourne autour de 1,7 milliard d'occurrences.

Fréquence

Pour l'étude de 2005 menée par Ďurčo, Grzybek élabore un graphique synthétique (Figure 1) où la courbe représente les pourcentages des *f* (où *f* maximale correspond au pourcentage le plus élevé (100%)) et la ligne en haut montre l'indice de familiarité (entre 86.57% et 100% par rapport aux résultats obtenus par l'échantillon de locuteurs interrogés) de chaque proverbe.

[140] Pour d'autres renseignements : http://www.ids-mannheim.de/cosmas2/ (date de consultation : 14/11/2013).

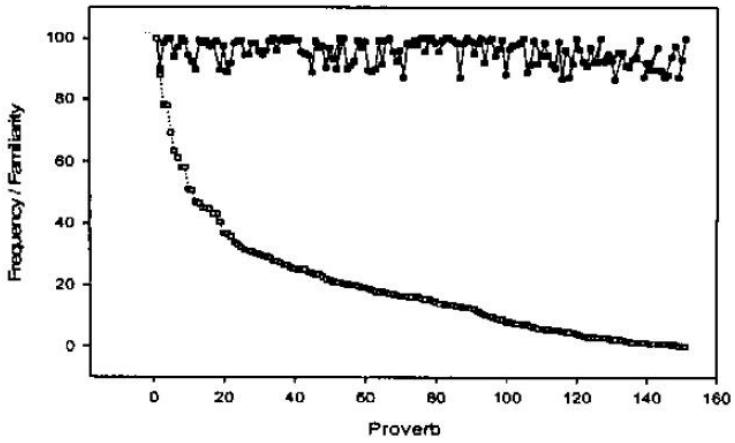


Figure 1. Graphique de corrélation entre le nombre de proverbes allemands examinés et leur fréquence/familiarité respectives (en %) d'après Ďurčo (2005) (tiré de Grzybek 2009 : 229).

En ce qui concerne *f* et de façon très générale, ce graphique prouve que, comme pour d'autres études mentionnées, la plupart des proverbes (on dirait environ 110 sur 151 proverbes) ont *f* < à ce 20%, à savoir *f* non élevée[141]. Les regroupements des proverbes autour de certaines valeurs ainsi que les ruptures de ces regroupements et la présence de proverbes-*outliers* (§ 3.2.8.) sont ainsi confirmés en allemand, aussi.

Quant à la corrélation fréquence-familiarité, Ďurčo conclut qu'elle n'existe pas, quoique Grzybek souligne que ce manque de corrélation est à interpréter par le fait que :

> « analyzing familiar proverbs only, it will hardly be possible to find any reliable insight into the FAM-FRQ relation » (*ibid*.).

En gros, on ne peut mieux comprendre les rapports entre fréquence d'usage et familiarité qu'à partir d'un échantillon de proverbes qui n'est pas biaisé au départ par l'une de ces variables. Comme l'ajoute encore Grzybek :

> « The only conclusion to be drawn from Ďurčo's study – and this conclusion is valuable indeed –, is the fact that highly familiar proverbs may, but need not occur frequently » (*ibid*.).

[141] Malheureusement, nous ignorons les *f* brutes.

En d'autres termes, la familiarité d'un proverbe ne garantit pas a priori *f* élevée en contrepartie, comme nous le rappelions pour commenter la recherche d'Arnaud & Moon (§ 3.2.3.). Une telle constatation ouvre la voie – elle donne plutôt le droit d'existence et d'étude – aux proverbes que certains parémiologues ignorent volontairement parce qu'ils ne sont pas 'contemporains',ce jugement étant fondé uniquement sur des études de familiarité.

Dans l'étude de 2006, la séparation entre *fréquence* et *familiarité* se concrétise par le moyen d'une hiérarchisation méthodologique en vue du but ultime d'un minimum parémiologique. Au premier abord, la familiarité apparaît dans une position subjuguée et asservie à la fréquence. Les filtres méthodologiques appliqués par Ďurčo veulent que la fréquence parémiographique précède la familiarité estimée par les spécialistes, tout comme la fréquence d'occurrence sur corpus précède le jugement de familiarité par un échantillon de la population germanophone. Pourtant, en nous inspirant de la sagesse proverbiale d'origine évangélique, *les premiers sont* (dès maintenant) *les derniers* : c'est la familiarité-dernière qui juge de la pertinence des résultats obtenus par la fréquence-première. Autrement dit, pour reprendre la distinction faite par Grzybek & Chlosta en parémiologie empirique (§ 2.1.2.), la connaissance gagne sur les textes. Ce choix méthodologique est, certes, motivé par l'établissement d'un minimum parémiologique qui se veut, de par sa nature, orienté à la familiarité. Malgré cela, il est vrai aussi que ce choix ne donne pas le même statut épistémologique aux notions de fréquence et de familiarité et qu'il n'y a pas vraiment un « compromis », comme le dit Zouogbo (2011 : 101), entre connaissance, compétence et occurrence d'usage. Il y a plutôt, à notre avis, une 'négociation' où les jugements (compétents ou non) de familiarité s'emparent d'une primauté décisionnelle au détriment des attestations de la fréquence sur corpus.

3.2.10. *Anderson* (2006)

Dans son analyse sur la phraséologie du français administratif basée sur corpus, Anderson jette un coup d'œil, entre autres, à l'usage des locutions idiomatiques ainsi que des proverbes. Nous avons ainsi un aperçu de l'usage des proverbes dans le domaine juridico-administratif.

Liste

L'intérêt d'Anderson pour la recherche des locutions idiomatiques et des proverbes est motivé par le constat que Rey & Chantreau expriment dans la préface de leur édition du *Dictionnaire des expressions et locutions* en 1993 :

> « aucun discours ou presque ne peut faire l'économie des locutions, lieux communs éculés ou produits plaisants de l'imagination populaire » (1993 : xiii).

Sur l'escorte de cette déclaration, Anderson décide d'utiliser ce dictionnaire pour établir la liste des locutions et des proverbes à rechercher dans ses deux sous-corpus. D'une part, elle motive ce choix par le fait que Rey & Chantreau affirment avoir collationné à la fois des recueils précédents et des données d'usage (y inclus la fréquence, paraît-il) ressorties de collections textuelles littéraires et journalistiques (2006 : 128). D'autre part, Anderson justifie la préférence pour ce dictionnaire en raison de sa taille (il inclut 11.647 expressions), sa généricité et son autorité (2006 : 129).

Sources

Pour son enquête, Anderson crée un corpus spécialisé (nommé FRADCO) qui se compose de deux sous-corpus :

- le corpus FRNACO (1.056.164 occurrences) qui est censé représenter le français administratif hexagonal et inclut des communiqués de presse, des discours officiels, des projets de lois, des rapports et des textes fondamentaux de France ;
- le corpus FREUCO (1.065.389 occurrences) qui doit rendre compte du français administratif de l'Union européenne et inclut, lui aussi, des communiqués de presse, des discours officiels ainsi que des plans d'action, des rapports, des traités et des documents officiels des travaux menés par chaque institution européenne (2006 : 225-228).

Tous les documents sont téléchargés des sites institutionnels manuellement. Il n'y a donc pas d'aspiration automatique de données linguistiques de sites Web (§§ 3.2.13.-14.).

On constate l'équilibre entre les deux sous-corpus aussi bien sur le plan quantitatif (ils ont presque la même taille en nombre d'occurrences) que sur le plan textuel (ils contiennent presque les mêmes genres textuels). Certes, les tailles sont réduites, mais, quand on s'intéresse à des faits linguistiques d'un discours spécialisé, il arrive très souvent de traiter des corpus d'une taille pareille. De par leur nature, on remarque aussi que certains genres textuels, comme les rapports et les traités, tendent à contribuer de manière plus significative d'autres genres du point de vue quantitatif. Par exemple, les 10 rapports officiels de FRNACO comptent plus d'occurrences que les 330 communiqués de presse, respectivement 195.596 face à 173.546 occurrences (2006 : 227).

Fréquence

Si l'on se concentre maintenant sur la fréquence des locutions en général, Anderson repère des occurrences seulement pour 633 sur un total de 11.647 locutions. Cela veut dire que 94.57% de locutions ont $f = 0$ (2006 : 131).

En ce qui concerne les proverbes, toutes les formes canoniques du *Dictionnaire* de Rey & Chantreau font partie de ce pourcentage. Autrement dit, aucun proverbe n'a été repéré dans les deux sous-corpus (2006 : 135). Le silence du français administratif est interprété, d'une part, en fonction de la taille réduite des deux sous-corpus et, d'autre part, à cause du manque d'une annotation morphosyntaxique des corpus mêmes (*ibid*.). Cette dernière affirmation souligne l'importance de prendre en compte la facette syntaxique, outre la facette lexicale, au moment de la détection sur corpus. En général, ce silence suggère de porter attention à la composition du corpus, notamment au critère du discours.

3.2.11. *Gómez-Jordana Ferary* (2006)[142]

La thèse de doctorat[143] de Gómez-Jordana Ferary poursuit le *leitmotiv* le plus connu en littérature parémiologique et ambitionne à proposer une définition linguistique du proverbe à partir de l'usage observé à l'écrit et à l'oral. Dans une perspective comparée

[142] Nous remercions Sonia Gómez-Jordana Ferary pour sa disponibilité et pour sa générosité.
[143] La thèse a fait l'objet d'une publication en 2012. Notre revue fera également référence aux données que nous avons consultées à la *Biblioteca de la Universidad Complutense de Madrid*, notamment aux annexes (liste des proverbes et contextes d'occurrence) qui ne sont pas disponibles dans la publication.

français-espagnol, son étude approfondit les aspects sémantiques et discursifs, autant en synchronie qu'en diachronie. Pourtant, pour atteindre son but, elle fait un détour en syntaxe en vue d'une énumération des *moules proverbiaux* (§ 1.2.5.). Ce qui la conduit, au préalable, à s'interroger sur les critères d'établissement d'une liste de proverbes et à quantifier les occurrences d'usage.

Liste

Comme nous venons de l'expliquer, sa recherche vise la définition linguistique d'après l'usage courant en contexte. Dès le départ, donc, il en suit que la désuétude proverbiale ne peut faire partie de son étude. Un jugement de contemporanéité « hors contexte » (2012 : 22) agit ainsi de filtre pour la mise au point de la liste des proverbes[144] français et espagnols. Plus précisément, ce filtre consiste en une enquête de familiarité sociolinguistique franco-espagnole auprès d'une centaine de personnes (de 13 à 86 ans) qui appartiennent à différentes couches sociales et qui proviennent de quelques régions hexagonales et ibériques (note 11, *ibid*.)[145]. Ces personnes ont transcrit pendant un mois leur répertoire parémiologique à elles, sans enrichir leurs listes de proverbes d'après la consultation de recueils parémiographiques[146]. Gómez-Jordana Ferary commente :

> « Grâce à cette enquête, nous avons obtenu une liste de parémies que nous pouvions considérer comme actuelles et nous ne sommes pas tombée dans le piège de nous baser uniquement sur celles qui sont proposées par les dictionnaires, où dans la plupart des cas, se glissent des formes archaïques et divers genres de locutions » (*ibid*.).

Nous partageons son avis quant au mélange linguistique souvent anachronique que les recueils parémiographiques proposent (§ 3.2.4.). En même temps, nous nous interrogeons sur le biais que pourrait constituer son échantillon de personnes. Non seulement ils auraient pu fausser l'enquête à tout moment[147], mais l'étendue même de la liste obtenue est forcément le

[144] Par souci de clarté, nous remarquons que la chercheuse préfère les termes *proverbes* et *formules*. De temps en temps, elle parle de *phrases* ou de *parémies*.
[145] La période de l'enquête et le nombre de personnes françaises et espagnoles ne sont pas précisées.
[146] C'est le type d'étude de familiarité *Ia* décrit par Grzybek & Chlosta et que nous avons mentionné au § 2.1.2.
[147] Entre autres : qui garantit à 100% à la chercheuse qu'elles n'ont pas jeté un coup d'œil aux proverbes dans des recueils ? Combien de temps ont-elles vraiment consacré à cette tâche de collecte introspective ? Se sont-elles confrontées avec d'autres personnes ? Et encore, comme le soulignent Grzybek & Chlosta (2009 : 97) pour ce type d'étude de familiarité, la récupération en mémoire d'un proverbe dépend de la situation extralinguistique

résultat d'une compétence parémique subjective, qui plus est, limitée[148]. Et encore, le silence de la mémoire vis-à-vis de certains proverbes lors d'une collecte passive n'implique pas que ces proverbes sont inconnus aux personnes enquêtées et, par conséquent, non actuels. Autrement dit, les listes des proverbes reçues peuvent cacher d'autres proverbes que les personnes connaissent et reconnaîtraient en contexte, mais qu'elles ont tout simplement oublié ou égaré dans leur mémoire à long terme. En outre, comme on l'a vu et on le verra, l''actualité' de tel ou tel autre proverbe (ou telle ou telle autre parémie) ne peut se passer d'une prise en compte (même partielle) de leur production de la part des locuteurs, aussi bien à l'écrit qu'à l'oral.

De plus, Gómez-Jordana Ferary a nettoyé la liste obtenue sur la base de critères linguistiques, comme l'infinitif du verbe qui peut être conjugué dans des locutions (2012 : 77). Un jugement linguistique *a priori* influe sur l'acheminement *a posteriori* vers une définition linguistique du proverbe. Nous comprenons que des personnes non expertes peuvent percevoir des proverbes (et, en général, des parémies) de manière impropre par rapport à ce que la littérature parémiologique suggère. Pourtant, d'une part, cela aurait pu représenter une occasion pour remettre en question certains acquis parémiologiques. D'autre part, l'introduction d'un 'filtre linguistique' expert (§ 3.2.15.) nous confirme dans le propos que le recours à des non experts pour l'établissement d'une liste de proverbes ne met pas à l'abri d'autres biais, peut-être plus délicats et complexes que la désuétude et l'anachronisme des recueils parémiographiques. Ainsi, de manière presque contradictoire, Gómez-Jordana Ferary ajoute que :

> « Nous avons ajouté à notre corpus quelques formules provenant de *Diccionario de refranes* de Campos et Barella (1993) et de *Le Dictionnaire des proverbes et dictons de France* de Dournon » (*ibid*.).

Le souci de l'actualité authentique se perd au milieu des recueils parémiographiques. De plus, c'est la chercheuse elle-même qui sélectionne certaines parémies, sans donner des explications sur les critères qui guident et motivent ce choix tout à fait curieux par rapport à son intention initiale. À ce propos, elle affirme encore que seulement les proverbes tirés de ces deux recueils (ainsi que d'autres, d'ailleurs) sont mentionnés tout au long de son étude

(vécue ou imaginée) auquel un proverbe peut faire référence. Combien de situations ces personnes peuvent vivre ou imaginer ? Attachent-elles un proverbe à ces situations ?

[148] Ce qui est souligné par Grzybek & Chlosta (2009 : 97). En confirmation de cela, Gómez-Jordana Ferary précise qu'en moyenne, elle a reçu 20 proverbes de chaque Français et 30 de chaque Espagnol (*ibid*.).

(Gómez-Jordana Ferary 2012 : 23). Ce qui contredit toujours l'élan vers les non-experts et la préoccupation d'une étude biaisée par la désuétude parémiographique.

De cette façon, elle obtient une liste de 700 proverbes pour chaque langue (*ibid*.). D'après notre consultation des annexes de sa thèse, cette liste inclut aussi des variantes morphosyntaxiques, comme :

L'oisiveté est mère de tous les vices
L'oisiveté est <u>la</u> mère de tous les vices

des variantes lexicales :

Qui ne risque rien n'a rien
Qui ne <u>tente</u> rien n'a rien

ainsi que des variantes où l'on observe une permutation interne :

Mieux vaut un petit chez soi qu'un grand chez les autres
<u>Un petit chez soi vaut mieux</u> qu'un grand chez les autres

Dans cette liste, aucune distinction n'est faite entre les résultats obtenus de l'enquête sociolinguistique et les ajouts tirés des recueils parémiographiques, quoique la lecture intégrale de la liste donne quelques indications à cet égard[149].

Sources

Par la suite, Gómez-Jordana Ferary recherche ses 700 proverbes en français et en espagnol dans des corpus écrits et oraux. Parallèlement, elle distingue (Gómez-Jordana Ferary 2012 : 23) :

- les textes littéraires : *Frantext* pour le français et le *CREA* de la Royale Académie de la Langue pour l'espagnol ; pour les deux, elle se tient à la

[149] On repère : « *Grande gueule petit bras* » suivi de la précision « (proverbe du monde du volley-ball) ».

période de 1970 à 2000, à peu près ; la recherche sur textes électroniques est aussi intégrée par la lecture de romans, essais et bandes dessinées ;

- les textes de presse : *Le Monde sur CD-Rom* pour le français et *El País* pour l'espagnol de 1999 à 2006 à peu près ; elle les choisit parce qu'ils sont deux journaux « de référence » (*ibid*.) ;
- les textes de la communication virtuelle : SMS et courriers électroniques, dont la provenance n'est pas précisée ;
- les transcriptions orales : de conversations à programmes télévisées ou radiophoniques, sans aucune précision de temps de collecte et de sources.

Fréquence

Il serait incorrect de parler de fréquence dans le cas de l'étude de Gómez-Jordana Ferary. Elle ne vise pas à établir une liste de fréquence des proverbes espagnols et français, mais repère seulement les occurrences de ces proverbes, sans qu'il y ait un détail quantitatif pour chaque proverbe. Il est, en tout cas, intéressant de dégager une réflexion justement sur l'actualité et sur la contemporanéité des proverbes, et ce, sur la base du nombre total d'occurrences repérées. Gómez-Jordana Ferary dit :

> « Le résultat global est d'environ huit cents occurrences, réparties à peu près de la même façon en français et en espagnol » (2012 : 24).

Par un arrondissement de 800 occurrences totales et donc d'environ 400 occurrences par langue, il en suit qu'en français, sur un total de 700 « proverbes contemporains » recherchés, on obtient 0,57 occurrences par proverbe en moyenne. Nous savons très bien que la moyenne est un indicateur fautif et très sensible aux données aberrantes, mais elle révèle que l'usage de ces proverbes qualifiés comme contemporains est faible et qu'un nombre *n* assez élevé de ses proverbes doit avoir $f = 0$. Nous avons du mal à partager la conclusion de Gómez-Jordana Ferary quant au fait que :

> « aussi bien en France qu'en Espagne, les proverbes sont toujours en vigueur » (*ibid*.),

non pas dans l'absolu, mais d'après ses résultats qui, d'ailleurs, ne s'alignent que partiellement avec ceux que nous avons présentés dans les autres études de notre revue.

3.2.12. *Grzybek* (2009)

En raison de ses intérêts en linguistique quantitative, Grzybek décide de tester son modèle algébrique non linéaire pour décrire la corrélation entre fréquence et familiarité pour certains proverbes allemands. Cette corrélation qu'il exprime comme suit :

FAM = *f*(FRQ)[150]

va sous le nom de *popularity* (popularité) (2009 : 224)[151].

Liste

Il part d'une liste de 20 proverbes allemands qu'il a examinés dans le cadre d'une de ses études de familiarité et d'usage 'présumé et déclaré' (à savoir sur base purement introspective) menée en 1983 (2009 : 220). Dans le cadre de la contribution, il effectue une nouvelle étude de familiarité sur la même liste de proverbes.

Sources

Outre la familiarité, Grzybek extrapole *f* brute de ces 20 proverbes 'doublement' familiers de par l'interrogation de la base de données *Mannheim COSMAS II* (§ 3.2.9.).

[150] On lit : la familiarité est fonction de la fréquence. Pour revenir à la distinction entre *fréquence émique* et *fréquence étique* en linguistique basée sur l'usage (voir § 2.1.2.), Loiseau précise que : « la fréquence émique peut diverger de façon systématique de la fréquence étique, particulièrement quand des représentations linguistiques sont en jeux [mais] *ce sont les mesures quantifiées (étiques) qui sont des approximations de mesures intuitives (émiques)*, et non l'inverse. Les objets pertinents sont en effet les fréquences intuitives, les fréquences quantifiées n'ont de statut et d'intérêt que comme approximation des fréquences perçues par les locuteurs. » (Loiseau 2011 : 67-68, c'est l'auteur qui souligne). Si l'on transpose cette dynamique en parémiologie empirique, cela implique qu'est normale la divergence entre les résultats d'une étude de familiarité et d'une étude de fréquence qui portent sur les mêmes proverbes-objets linguistiques. Il faut encore ajouter que la mesure de la fréquence reflète à peu près la familiarité et la perception de la fréquence de ces mêmes proverbes. Cette vision renverse donc la corrélation établie par Grzybek et on aurait plutôt : FRQ = *f*(FAM). Le défi de la corrélation entre familiarité et fréquence (qui relève en partie du paradoxe de l'œuf et de la poule) ne peut être résolu qu'avec des données quantitatives cohérentes et bien encadrées du point de vue méthodologique. Ce qui n'est pas le cas pour le répertoire parémiologique français, mais aussi d'autres langues (§ Conclusions).

[151] Il a effectué un test pareil pour les proverbes anglais étatsuniens (Grzybek & Chlosta 2009).

Fréquence

En ligne générale, on observe que le seul proverbe :

Wer Brot hat, dem gibt man Brot [trad. litt. :Qui a du pain, qu'il donne du pain]

a $f = 0$, alors que 9 autres proverbes ont $8 \geq f \leq 42$. En tête de liste, Grzybek identifie les 10 proverbes du Tableau 16 :

Rang	**Proverbe [traduction littérale]**	***f* brute**
1)	*Sicher ist sicher* [Assurer est assurer]	475
2)	*Wenn einer eine Reise tut, so kann er was erzählen* [Si on planifie un voyage, alors on peut en parler]	320
3)	*Zeit ist Geld* [Le temps est d'or]	303
4)	*Ein unglück kommt selten allein* [Un déplaisir vient rarement seul]	197
5)	*Morgenstund hat Gold in Mund* [Le matin a de l'or en bouche]	120
	Nachher ist man immer klüger [Par l'après on est toujours sage]	120
7)	*Geteiltes Leid ist halbes Leid* [Une peine commune est une peine à moitié]	97
8)	*Über Geld spricht man nicht* [De l'argent personne ne parle]	73
9)	*Spare in der Zeit, dann hast du in der Not* [Gagne du temps, quand tu es en difficulté]	67

10)	*Schadenfreude ist den reinste Freude*	
	[La *Schadenfreude* est	47
	la joie la plus parfaite]	

Tableau 16. Fréquence des 10 premiers proverbes allemands les plus fréquents d'après Grzybek (2009).

On observe que *f* atteint l'ordre des centaines parmi les 5 premiers proverbes et que les 5 autres restants ne dépassent pas le seuil des 100 occurrences. Les fréquences un peu plus importantes (en termes quantitatifs) que ceux d'autres études mentionnées sont certainement influencés par la taille du corpus. À ce propos et comme nous l'avons constaté, cette liste prouve aussi qu'un écart significatif se produit entre les proverbes en tête de liste et les autres qui suivent. D'étude en étude, on comprend que cet écart s'accroit de manière directement proportionnelle à l'augmentation de la taille du corpus : plus le corpus est grand, plus l'écart entre les proverbes en tête de liste et les restants se fait important.

3.2.13. *Hrisztova-Gotthardt & Gotthardt* (2011)

La contribution de Hrisztova-Gotthardt & Gotthardt porte sur l'estimation de la fréquence de proverbes bulgares à partir de la consultation du Web.

Liste

Les auteurs tirent parti d'une liste de 2.301 proverbes (2011 : 254) éventuellement hiérarchisés en lemmes et variantes, le choix des lemmes reposant sur *f* dans un corpus qui rassemble 8 ans de presse bulgare (journal *Standart*, 2011 : 256). C'est une option originale par rapport aux études analysées jusqu'à présent. Précisons encore qu'il s'agit d'une liste s'appuyant sur deux recueils parémiographiques bulgares (2011 : 250).

Sources

Les auteurs décident d'explorer le Web comme corpus (§ 2.2.5.2.) par l'intermédiaire de deux moteurs de recherche : *Bing* et *Google* (2011 : 251-252).

Fréquence

À l'issue de l'interrogation des moteurs de recherche (§ 4.2.1.14.), les chercheurs présentent les 10 parémies les plus 'fréquentes' :

Proverb (Variants)	Translation	Individual hits	Total hits	% (of total hits)	Source
Който търси, намира. (Търсете и ще намерите.)	He who seeks, will find. (Seek and you shall find.)	439+268	707	1,30%	Google
Всяко зло за добро.	No evil without good.	637	637	1,18%	Bing
По-добре късно, отколкото никога.	Better late than never.	594	594	1,10%	Bing
Глас народен - глас божи. (Глас народен - глас божий.)	The voice of the people [is] the voice of God. (The voice of the people [is] the voice of God almighty.)	426+150	576	1,06%	Google
Една птичка пролет не прави. (Една лястовица пролет не прави.)	One bird does not make spring. (One swallow does not make spring.)	455+103	558	1,03%	Google
Нищо ново под слънцето. (Няма нищо ново под слънцето.)	Nothing new under the sun. (There is nothing new under the sun.)	318+212	530	0,98%	Bing

Proverb (Variants)	Translation	Individual hits	Total hits	% (of total hits)	Source
Съединението прави силата.	Union makes strength.	484	484	0,89%	Bing
Бързай бавно.	Hurry slowly.	478	478	0,88%	Google
Старост - нерадост.	Old age [brings] no joy.	465	465	0,86%	Google
Апетитът идва с яденето.	Appetite comes with eating.	461	461	0,85%	Google

Figure 2. Tableau des occurrences des 10 proverbes bulgares les plus 'fréquents' sur le Web repris par Hrisztova-Gotthardt & Gotthardt (2011 : 259-260).

Hrisztova-Gotthardt & Gotthardt concentrent leur attention : (i) sur la recherche de formes exactes (variantes incluses) (§ 4.2.1.14.), ignorant les processus de variation des parémies en discours ainsi que l'insertion des attestations parémiques non canoniques dans le calcul de f ; (ii) sur le Web dont ils sont conscients de la dynamicité (2011 : 251) et par conséquent de l'impossibilité de réfléchir sur de véritables fréquences parémiques.

Il faut préciser que les auteurs comparent la liste des proverbes ci-dessus avec une autre liste de *f* obtenue après l'interrogation du corpus de presse bulgare mentionné plus haut. La comparaison ne se fait pas par le nombre d'occurrences, mais par le rang. Ce qui les surprend est que les 10 premiers proverbes correspondent dans les deux listes (2011 : 260).

De plus, ils estiment que, comme 65 proverbes couvrent la moitié des attestations repérées sur le Web, il serait probable qu'ils représentent les proverbes bulgares modernes les plus 'fréquents'. À ce propos, nous avons quelques doutes. Nous constatons en effet que les 10 premiers proverbes sont communs au répertoire francophone (mais aussi italophone). En gros, il s'agit de proverbes appartenant à une même matrice européenne, voir latine. Il est donc possible que des traductions sur le Web aient altéré les résultats obtenus. D'où notre souhait que les auteurs entreprennent des recherches ultérieures dans le corpus moniteur bulgare, comme ils l'annoncent (*ibid*.).

3.2.14. *Rozumko* (2012)

Dans un ouvrage consacré à l'anglicisation des lexiques européens, Rozumko considère l'accueil de certains proverbes anglais et leur traduction/adaptation à l'intérieur des recueils parémiographiques polonais. En guise d'appui aux observations qualitatives, Rozumko décide de compléter son étude par un détail de nature quantitative : la fréquence dans des corpus polonais.

Liste

Suite à une analyse des recueils parémiographiques polonais, elle définit une liste de 10 emprunts parémiques du répertoire parémiographique anglophone que nous reprenons dans le tableau ci-dessous (2012 : 266) :

1	*Mój dom to moja twierdza* (A man's home is his castle)
2	*Nie oceniaj książki po okładce* (You can't tell a book by its cover)
3	*Do tabga trzeba dwojga* (It takes two to tango)
4	*Mężczyźni wolą blondynki* (Gentlmen prefer blondes)
5	*Fakty są uparte* (Facts are stubborn things)
6	*Fakty mówia [same] za siebie* (Facts speak for themselves)
7	*Liczby nie kłamią* (Figures don't lie)

8	*Nie istnienje coś takiego jak darmowy obiad* (There's no such a thing as a free lunch)
9	*Kobieta bez mężczyźni jest jak ryba bez roweru/* *Mężczyźna kobiecie jest potrzebny jak rybie rower* (A woman without a man is like a fish without a bicycle)
10	*Morderca/przestępca/sprawca zawsze wraca na miejsce zbrodni* (The murderer/criminal returns to the scene of crime)

Tableau 17. Liste des 10 emprunts du répertoire parémiographique anglophone et leur traduction/adaptation en polonais d'après Rozumko (2012).

Il s'agit d'emprunts parémiques contemporains. En effet, le dépouillement concerne les recueils polonais à partir de la fin du XIX^{e} siècle (2012 : 263). Nous constatons également que Rozumko prend en compte des variantes formelles, comme dans les cas des proverbes 6, 9 et 10.

Sources

Avant d'entamer la description ponctuelle du comportement de chaque emprunt parémique, Rozumko présente quelques précisions sur les 3 corpus qu'elle a interrogés (et que nous compléterons par d'autres données repérées sur leur composition). Elle travaille avec :

- le *Polskiego Wydawnictwa Naukowego* (PWN) *Corpus*, notamment sa version en ligne : un corpus d'environ 40 millions d'occurrences et composé surtout d'écrit papier numérisé (livres, éditions, brochures), mais aussi d'oral (transcriptions) et de matériel écrit aspiré par des sites Web[152] ;
- le *IPI PAN Corpus* : un corpus consacré à l'écrit polonais – papier numérisé, électronique et Web – de taille supérieure à 100 millions d'occurrences (à savoir 250 millions de segments[153]) et annoté morphosyntaxiquement au format XML[154] ;

[152] Le corpus est consultable à l'adresse : http://korpus.pwn.pl/index_en.php (date de consultation : 11/11/2013).
[153] Dans sa contribution, Rozumko parle de 250 millions d'occurrences, alors que la documentation du corpus précise ce que nous venons de rédiger.
[154] D'autres renseignements sont disponibles à l'adresse : http ://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/book_en.pdf (date de consultation : 11/11/2013).

- le *Polish Web Corpus* (Radziszewski *et al*. 2011) : un corpus entièrement composé de matériel écrit aspiré sur le Web, sans aucune contrainte, pour un total de 128.185.119 occurrences [155] , annoté morphosyntaxiquement et lemmatisé grâce au logiciel en ligne *Sketch Engine* (§ 2.2.).

La variété en taille, médium, prétraitement informatique et périodes de temps couvertes par les corpus offre un éventail diversifié pour le repérage des occurrences.

Nous soulignons qu'il s'agit de la première étude qui exploite l'écrit du Web comme source pour la constitution automatique d'un corpus, comme l'est le *Polish Web Corpus*[156]. Pourtant, à ce trio de corpus fermés, Rozumko décide également de consulter le Web par le moteur de recherche *Google* (Web comme corpus, § 2.2.5.2.). Elle juxtapose ainsi les *f* brutes calculées dans les corpus et le nombre d'attestations repérées par *Google*.

Fréquence

Dans la figure ci-dessous, nous reprenons le tableau mis au point par Rozumko avec l'indication des occurrences repérées pour chaque emprunt parémique en polonais :

[155] Rozumko indique le chiffre de 103.028.410 millions de mots, alors que la documentation du corpus précise ce que nous avons repris dans le texte.
[156] Dans le cas d'Anderson (§ 3.2.9.), les textes sont téléchargés manuellement du Web et les sources du téléchargement sont ciblées. Rozumko utilise, en revanche, de l'écrit aspiré du Web sans aucune limitation de sources (URL).

Table 1. Frequency of selected English proverbs in Polish

Proverbs	Frequency in the *PWN* corpus (40 million words)	Frequency in the *IPI PAN* corpus (250 million words)	The Polish web corpus (103,028,410 words)	Google search (20 July 2011)
(1) *Mój dom to moja twierdza* (Eng. *A man's home is his castle*)	3	10	8	88,500
(2) *Nie oceniaj książki po okładce* (Eng. *You can't tell a book by its cover*)	–	–	2	217,000
(3) *Do tanga trzeba dwojga* (Eng. *It takes two to tango*)	17	15	23	570,000
(4) *Mężczyźni wolą blondynki* (Eng. *Gentlemen prefer blondes*)	9	5	2	800,000
(5) *Fakty są uparte* (Eng. *Facts are stubborn things*)	1	–	–	3,110
(6) *Fakty mówią [same] za siebie* (Eng. *Facts speak for themselves*)	3	17	25	465,000
(7) *Liczby nie kłamią* (Eng. *Figures don't lie*)	2	11	3	984,000
(8) *Nie istnieje coś takiego jak darmowy obiad* (Eng. *There's no such thing as a free lunch*)	–	2	10	97,300
(9) *Kobieta bez mężczyzny jest jak ryba bez roweru/Mężczyzna kobiecie jest potrzebny jak rybie rower* (Eng. *A woman without a man is like a fish without a bicycle*)	1	–	–	69,206
(10) *Morderca/przestępca/sprawca zawsze wraca na miejsce zbrodni* (Eng. *The murderer/criminal returns to the scene of the crime*)	–	–	1	4,725

Figure 3. Tableau des occurrences des 10 emprunts parémiques en polonais repris par Rozumko (2012 : 266).

On ne va ni interpréter le nombre d'attestations de *Google* comme un indice d'immortalité de ces 10 proverbes en polonais ni leur donner le statut de fréquence, au moins que l'on cède à la tentation de considérer vraiment le Web comme corpus (§§ 3.2.13.-16.). En tout cas, il est intéressant de constater que le Web n'est jamais silencieux à l'égard de ces emprunts, malgré toutes les précautions qu'il faut prendre sur la fiabilité de ces résultats.

Mis à part le Web, ce tableau confirme l'appel de Norrick à consulter des corpus de taille critique avant de décréter le décès des proverbes. Le *Polish Web Corpus* est le corpus qui renvoie le nombre le plus élevé d'occurrences totales. Pourtant, l'oscillation des *f* brutes

de chaque proverbe dans les corpus témoigne que la taille d'un corpus n'est pas tout ce qu'il faut. Par exemple, les fréquences de :

Do tanga trzeba dwojga [trad. litt. : Pour le tango il faut deux]

et de :

Liczby nie kłamią [trad. litt. : Chiffres ne mentent pas]

sont éloquentes à ce propos. Pour le premier (proverbe dispersé assez uniformément dans les trois corpus par rapport aux autres), on constate la performance du *PWN Corpus*, en l'occurrence fortement influencée, à notre avis, par sa composante livresque et par les transcriptions orales. Pour le deuxième, il est intéressant d'observer que *f* enregistrée dans *IPI PAN Corpus* surclasse celle enregistrée dans le *Polish Web Corpus*, et ce, malgré leur taille tout à fait comparable. Ces résultats divergents entre corpus pour une même liste de proverbes nous recommandent encore que non seulement la taille, mais aussi (et surtout) les genres textuels et les discours influencent les fréquences des proverbes.

3.2.15. *Barani* (2012)

Dans sa thèse de doctorat soutenue à l'Université de Salamanca, Barani établit une liste de fréquence des parémies espagnoles sur corpus. Plus précisément, l'établissement de cette liste vise la constitution de matériaux didactiques pour l'enseignement de l'*espagnol langue étrangère* (ELE) aux perses (2012 : 1)[157].

Liste

Pour ce propos, elle estime de 'pré-juger' la fréquence de milliers de parémies dont les recueils parémiographiques espagnols sont remplis. D'abord, par une forme de respect à l'égard de certains parémiographes espagnols, Barani transcrit l'intégralité des parémies

[157] Ailleurs et sur la base de la fréquence sur corpus en espagnol, Barani établit également un glossaire bilingue espagnol-farsi de 50 proverbes (Barani 2007).

recueillies dans les œuvres de cinq d'entre eux[158] (2012 : 38). Par la suite, elle applique un 'filtre natif', à savoir la compétence parémiologique d'un seul locuteur espagnol :

> « para decidir cuáles podrían considerarse más o menos "frecuentes", "conocidas" o "normales" ». (*ibid*.)

La recherche de la fréquence des parémies sur corpus est ainsi biaisée par leur familiarité telle qu'elle est ressentie par un seul locuteur natif espagnol. En outre, la suite des adjectifs « *frecuentes* » (fréquentes), « *conocidas* » (connues) et « *normales* » (normales), tous entre guillemets, confirme l'intention, d'une part, d'un jugement subjectif (quelle normalité pour une parémie ?) et met en relief, d'autre part, des imprécisions théoriques (la connaissance d'une parémie n'est pas – nécessairement – liée à sa fréquence, comme nous l'avons constaté dans notre revue) et méthodologiques (faire appel à l'introspection d'un seul locuteur pour la quantification d'un usage linguistique).

De plus, Barani a appliqué un 'filtre théorique' basé sur la littérature en matières parémiographique et phraséologique (notamment, les apports de Corpas Pastor), et ce, pour distinguer les parémies de tout autre expression idiomatique et unité phraséologique[159]. Ces deux filtres lui ont permis de créer, en tout cas, une « *lista provisional* » (liste provisoire) de 397 parémies.

Par la suite, cette liste est soumise à une étude de fréquence parémiographique dans 12 recueils parémiographiques (2012 : 40). Un 'filtre parémiographique' (à la façon de Lau, § 3.2.4.) est ainsi introduit pour mettre au point une « *lista maestra* » (liste principale) de :

> « expresiones que aparecen un mínimo de 3 veces en la totalidad de las 12 obras » (*ibid*.)

Le seuil $f = 3$ pour la création de cette liste principale est fixé de façon arbitraire. Nous en déduisons seulement que chacune des 397 parémies de la liste provisoire doit apparaître du moins dans 3 des recueils sélectionnés, à savoir 25% du total des recueils. Aucune autre motivation statistique de ce seuil n'est donnée. De même, aucun critère de sélection sur lequel appuyer le choix des recueils parémiographiques n'est mentionné. En outre, les 12 recueils choisis incluent 4 des 5 recueils utilisés pour l'établissement de la liste provisoire.

[158] À cet égard, l'auteure ne motive pas les critères qui l'ont conduite à choisir ces 5 recueils parémiographiques.
[159] L'auteure admet elle-même ces limitations, quoiqu'elle affirme que sa démarche peut satisfaire son but de recherche et donc le développement de son étude (2012 : 39-41).

Aucune raison pour la réintroduction de ces recueils n'est présentée par Barani et ce choix réfute, d'une part, le calcul de la fréquence parémiographique et, d'autre part, remet en question le seuil minimal $f = 3$[160]. Malgré ces aspects, Barani obtient une « *lista maestra* » de 277 parémies.

Elle se penche aussi sur la question délicate de la forme canonique et des variantes proverbiales qu'elle résout empiriquement, faute de références théoriques (et, ajoutons, philologiques), en faisant recours à celles que nous synthétisons comme :

- l'*autorité de la norme*, à savoir les formes enregistrées dans les recueils parémiographiques (choix – parfois tautologique – partagé par plusieurs parémiographes mentionnés dans le présent chapitre) ;
- l'*autorité de l'usage*, à savoir les formes repérées sur corpus ainsi que sur le Web par un moteur de recherche (en l'occurrence, *Google*) (2012 : 41).

Barani affirme privilégier l'autorité de l'usage, quoiqu'elle admette que trancher n'est pas toujours facile (2012 : 42). À la fois, donc, Barani mène une étude basée sur corpus, vu qu'elle vérifie une norme parémiographique de par son usage à l'écrit, et une étude guidée par le corpus, parce qu'outre à confirmer la norme parémiographique, elle enregistre d'autres usages pour adapter, voire enrichir, la norme parémiographique. En quelque sorte, l'usage établit une normalisation qui correspond à ou qui se distingue de la norme recherchée.

Sources

L'adaptation de la norme parémiographique à l'usage ainsi que le calcul de la fréquence reposent sur l'interrogation d'un corpus de presse espagnole (choix souvent partagé par les parémiologues-détecteurs de proverbes). Plus précisément, il s'agit des textes sur CD-Rom de l'édition madrilène du quotidien *El Pais* pour la période 2000-2001 (2012 : 5-6) pour un total estimé de 40 millions d'occurrences[161].

Barani motive le choix de ce corpus comme suit :

[160] La liste principale peut en effet compter des parémies qui apparaissent dans au moins 3 des 4 recueils réintroduits. Pour garder une cohérence méthodologique, on aurait pu procéder au calcul de la fréquence parémiographique : soit par une vérification croisée de la présence des parémies dans des recueils parémiographiques non employés pour l'établissement de la liste provisoire ; soit par une vérification interne de la présence des parémies aux 5 recueils parémiographiques employés pour l'établissement de la liste provisoire.
[161] Barani précise que ce chiffre est calculé sur une estime personnelle parce que l'implémentation du corpus et de son logiciel d'interrogation ne suivent pas les exigences propres aux recherches linguistiques (2012 : 7-8).

« En su gran mayoría son textos también de una variedad estilísticamente neutra y estándar, ni muy formal ni muy coloquial, con temas de actualidad y de muy distintos tipos. Se podría argumentar, además, que este periódico es el más representativo y prestigioso en el ámbito de España, y un referente a escala mundial » (2012 : 6).

La variété linguistique d'espagnol standard et peu formel, ainsi que le traitement de plusieurs discours et le prestige à l'échelle internationale sont donc les arguments qui assurent, d'une part, le calcul de la fréquence des parémies et, d'autre part, la satisfaction du but pédagogique de son étude. Quoique les textes de presse prévoient une sous-catégorisation en plusieurs genres textuels (2012 : 7), on peut affirmer qu'elle restreint son attention au seul macro-genre textuel 'article de presse'.

Fréquence

Barani présente ses résultats par un tri décroissant des *f* brutes pour chaque parémie de sa liste principale, sans aucune répartition en plages (2012 : 228-234). Seulement la parémie la plus fréquente, à savoir :

A la tercera va la vencida [trad. litt. : À la troisième va la victoire]

a $f = 41$, alors que toutes les autres restent en dessous de la quarantaine. Plus précisément :

- du rang 2 de la liste de fréquence, à savoir de *Ojo por ojo, diente por diente* [Oeil pour oeil, dent pour dent] au rang 11, où l'on a *Todos los caminos llevan a Roma* [Tous les chemins mènent à Rome] : $31 \leq f \geq 20$;
- du rang 12 pour *La ley del embudo ; para mí lo ancho y para ti lo agudo* [trad. litt. : La loi de l'entonnoir ; pour moi je l'enlargis et pour toi je le retrécis] au rang 32 de *Zapatero, a tus zapatos* [trad. litt. : Cordonnier, à tes chaussures] : $20 < f \geq 10$;
- du rang 33 pour *Cada cosa a su tiempo* [Chaque chose a son temps] au rang 70 pour *Sobre gustos no hay nada escrito* [trad.litt. : Sur les goûts il n'y a rien d'écrit] : $10 < f \geq 5$;
- du rang 71 au rang 140 : $5 < f \geq 2$;

- et du rang 141 au rang 180 : $f = 1$[162].

On en conclut donc que 97 parémies, soit environ 35% des parémies de la liste principale, ont $f = 0$. Barani ne rentre pas dans les détails de ce silence qu'elle attribue au fait que certaines parémies ont davantage de probabilités d'usage à l'oral qu'à l'écrit (2012 : 52). 110 parémies comptent $5 < f \geq 1$, c'est-à-dire environ 60% des parémies repérées dans le corpus. En général, Barani attribue ce nombre élevé de parémies à fréquence basse à la taille trop exiguë du corpus (*ibid*.). Comme pour d'autres cas, nous remarquons que les proverbes tendent à se regrouper autour de certaines valeurs basses et que certains se comportent comme à la façon d'*outliers*, surtout quand on se rapproche de la tête de liste.

Il faut souligner, en outre, que toutes les fréquences relèvent non seulement du simple repérage sur corpus, mais aussi d'opérations de désambiguïsation et de nettoyage menées par l'auteure. À côté de l'effacement d'occurrences à valeur non parémique (2012 : 46-48), elle a également effacé toutes les occurrences où les parémies sont utilisées en tant que :

> « nombres de películas, programas de televisión, obras de teatro, grupos y espectáculos musicales, relatos, ensayos » (2012 : 47),

à savoir tous les emplois non discursifs ou, en d'autres mots, tous les emplois des parémies comme 'étiquettes'. Nous soulevons quelques perplexités à l'égard de ce dernier choix. Il méconnaît (et sous-estime) non seulement des occurrences attestées, mais aussi une des fonctions discursives d'une partie des parémies. On fait référence ici à la lecture sémiotique de dénomination que Kleiber (1989, 2000, 2010a) a attribué aux proverbes-catégories des situations de vie, et qui est reconnue pour d'autres langues[163].

Mis à part ce choix, Barani présente une répartition des occurrences totales par rubriques. Sur un total de 1.059 occurrences, la plupart des attestations se divisent entre les rubriques *Única* (219 occurrences), *Opinión* (176 occurrences), *Cultura* (137 occurrences),

[162] C'est nous qui avons réparti les parémies en bandes de fréquence.

[163] À propos des proverbes tchèques, Čermák les dénote à la fois comme « name for a dynamic extralinguistic complex, made up of a […] complex and changing set of denotata (with their referents being stable neither in identity nor in number, sometimes) » (2007 [1998] : 572) et comme « a story, which is either viewed in its progress or as being concluded » (*ibid*.). En parémiologie russe, Permyakov ne parle pas de nom ou d'histoire, mais reconnaît, entre autres, aux proverbes une propriété de dénomination de grande envergure. Il définit les proverbes comme : « signs of situations or of a certain type of relationships between objects. Being signs, they must possess (and do indeed possess) special semiotic properties characteristic of all signs » (1979 : 20). Kleiber arrive à la même conclusion quelques années plus tard en suggérant pour les proverbes un traitement catégoriel sémantique en soi (2010a).

España (122 occurrences) et *Deportes* (115 occurrences) (2012 : 53). Malheureusement, l'auteure arrête à ce stade la description en raison de la visée de son étude[164]. De toute façon, on constate que les parémies tendent à apparaître à chaque fois où l'on donne un avis sur un fait ou dans les cas où l'on traite des évènements culturels et sportifs.

À côté de cette répartition textuelle, Barani élabore la répartition des fréquences de chaque parémie en rapport avec les processus de variation repérés (2012 : 238-243). 577 occurrences renvoient aux formes expressément visées par la recherche, alors que 482 occurrences sont les variantes reconnues. Nous reproduisons ci-dessous le tableau avec la répartition des occurrences proverbiales et leurs pourcentages respectifs par rapport aux processus de variation :

TIPO DE MODIFICACIÓN	NÚMERO TOTAL DE CASOS	PORCENTAJE DEL TOTAL DE CASOS MODIFICADOS
ADICIÓN	27	5.6%
REDUCCIÓN	102	21.2%
SUSTITUCIÓN	75	15.6%
MOD GRAMATICAL	81	16.8%
MOD COMBINADA	72	14.9%
MOD COMPLEJA	125	25.9%
TOTAL	482	100%

Figure 4 Tableau de répartition des occurrences des proverbes espagnols par processus de variation reconnu avec leurs pourcentages respectifs (Barani 2012 : 244).

La modification complexe[165], la réduction et les variantes morphosyntaxiques intéressent environ la moitié des variantes. Plus rarement, en revanche, les parémies font l'objet de véritables ajouts créatifs. Ces indications, pour autant que partielles qu'elles soient, nous conseillent des pistes à privilégier lors de la création de nos requêtes informatiques.

[164] Il n'y a pas une description, même générique, des contenus de chaque rubrique.

[165] Par « *modificación compleja* », Barani entend toute variation qui ne peut se résoudre par la prise en compte d'une variation formelle (comme l'adjonction, la réduction, la substitution, la variation morphosyntaxique ou une combinaison entre ces processus de variation de surface). La modification complexe demande l'application d'un schéma d'interprétation logique de la parémie aux contextes où les variantes de cette parémie sont repérées (2012 : 97-106).

3.2.16. *Mogorrón Huerta & Navarro Brotons (2012), Navarro Brotons (2013)*

Il faut préciser que la contribution de Mogorrón Huerta & Navarro Brotons (2012) que nous examinerons, est à entendre sans solution de continuité avec une thèse ponctuelle en parémiologie contrastive espagnol-français dirigée par Mogorrón Huerta et soutenue par Navarro Brotons à l'Université d'Alicante en 2013. Il ne s'agit pas vraiment d'une étude de fréquence des proverbes, mais d'une recherche lexicographique et syntaxique dans le cadre de la méthode du Lexique-Grammaire. Elle nous intéresse dans la mesure où certains passages concernant l'établissement de la liste de proverbes à décrire font explicitement recours en même temps aux notions de *fréquence* et de *corpus*[166].

Liste/Sources/Fréquence

Navarro Brotons circonscrit un échantillon de proverbes espagnols et français sur la base d'une équivalence interlinguistique lexico-grammaticale. Elle choisit :

> « les proverbes espagnols et français dont la structure syntaxique commence par *a*/*à*, *más vale*/*mieux vaut*, *no*/*ne*, *quien*/*qui* » (2012 : 453)[167].

Il est intéressant de se concentrer quelques instants sur les motivations de cette sélection. À côté de la simple constatation que le manque d'une telle sélection aurait comporté une charge massive de travail en termes de collecte et description, la raison qui justifie cette restriction réside dans le fait que :

> « ces quatres structures parémiques sont parmi les plus productives des deux langues étudiées et présentent des milliers de formes » (2012 : 454).

En amont, il y a un biais de fréquence concernant le mot ou les mots initiaux des proverbes. Cette fréquence est interne aux répertoires espagnol et français, et se fonde sur un constat de Conenna (2000). De toute façon, c'est sur ce critère linguistique qu'elle établit sa liste à

[166] Dans notre présentation, nous croiserons et intégrerons les données de la contribution de 2012 avec celles de la thèse de 2013.
[167] Dans le troisième chapitre de sa thèse, on apprend que sous les parémies qui commencent par *quien*/*qui*, Navarro Brotons inclut également les parémies qui commencent par *a quien*/*à qui* et *lo que*/*ce qui* (2013 : 143-144).

laquelle elle attribue le statut de *corpus*, suivant la définition des normes EAGLES (2013 : 82).

La première étape de sélection a démarré par le dépouillement des recueils parémiographiques espagnols et français. Pour le français, Navarro Brotons s'appuie au départ sur le *Dictionnaire de proverbes et dictons* des éditions *Le Robert*, sur le dictionnaire bilingue français/espagnol *Selección de refranes y sentencias* de Cantera et de Vicente ainsi que sur le dictionnaire multilingue *Dictionary of European Proverbs* de Strauss (2013 : 83). Les motivations de ce choix initial se résument en deux raisons de convenance : d'une part, le fait que ces dictionnaires disposent d'un grand nombre de parémies qui satisfont le critère lexico-syntaxique ; d'autre part, le fait que les parémies sont présentées par ordre alphabétique (*ibid*.). Il n'y a donc pas vraiment de motivations philologiques ou, en tout cas, ayant trait à un aspect parémiologique à proprement parler.

Par la suite, elle intègre des variantes et des définitions venant du *Dictionnaire des proverbes et des dictons de France* par Dournon pour les éditions *Larousse* ainsi que du *Refranero Multilingüe* (2013 : 83-84). De ce recensement, elle obtient deux listes d'environ 2.700 parémies françaises et 7.300 parémies espagnoles.

Vu l'étendue des listes, la deuxième étape – et c'est à ce stade que la sélection vire vers la contemporanéité et l''actualité' des proverbes tant évoquée par certains parémiologues – consiste :

> « à sélectionner les formes les plus utilisées actuellement de façon à ne travailler qu'avec des proverbes usités » (2012 : 454).

Au biais distributionnel de nature formelle s'ajoute ainsi le biais de l''actualité' des proverbes. Pour appliquer ce deuxième filtre et parvenir à deux listes de parémies, la tentation du Web tout-venant comme témoin de la contemporanéité parémique l'emporte :

> « nous avons réalisé une recherche sur le moteur de recherche *Google* dans le but de sélectionner les proverbes qui apparaissent le plus fréquemment sur la Toile » (2012 : 454).

Cet indice approximatif est en partie mitigé par la précaution des auteurs quant à la consultation a posteriori de corpus textuels (*ibid*.). Il est dommage que cette consultation n'ait pas fait l'objet d'un travail *a priori*, peut-être juxtaposé à la recherche sur le Web (§§ 3.2.13.-14.), les corpus espagnols étant largement disponibles. En outre, on n'a accès ni à la liste

provisoire des parémies dépouillées dans les recueils parémiographiques ni au nombre d'attestations renvoyées par *Google*. C'est en effet l'établissement de seuils d'attestation des parémies sur le Web qui décide de leur usage courant et de leur description syntaxique. Navarro Brotons définit ainsi 5 plages pour répartir les résultats obtenus :

> « 1. No aparece: 0 resultados.
> 2. Muy escasa: resultados que oscilan entre 1 y 100.
> 3. Escasa: los resultados comprendidos entre 100 y 1.000.
> 4. Normal: resultados entre 1.001 y 10.000.
> 5. Masiva: resultados a partir de 10.001 » (2013 : 85).

Toutes les parémies dont $f = 0$ et $1 \geq f < 100$ sont mises de côté et le seuil de la centaine d'attestations suffit pour satisfaire le critère de contemporanéité. Tous comptes faits, Navarro Brotons réduit sensiblement ces deux listes provisoires : d'environ 2.703 à 889 parémies espagnoles ; d'à peu près 1.899 à 697 parémies françaises (*ibid*.). Quelque 67% des parémies de la liste provisoire espagnole et 63% des parémies de la liste provisoire française sont effacées. D'une part, ces chiffres servent certainement d'indicateur de l'exhaustivité du recensement de Navarro Brotons et de l'envergure des recueils consultés. D'autre part, ces résultats pointent encore le doigt contre les dictionnaires qui enregistrent des quantités de sagesse parémique oubliée et rétro[168]. En tout cas, il est raisonnable de se demander si une description syntaxique doit seulement concerner des parémies 'actuelles' alors que la description syntaxique de parémies 'moins actuelles' (comme celle de la plage 2) aurait confirmé ou réfuté leur 'actualité' par la répétition de leur identité structurelle syntaxique. Pour notre propos, il est en outre dommage de ne pas avoir l'occasion de consulter la liste des parémies non repérées et/ou peu usitées sur le Web pour effectuer des comparaisons. De toute façon, on apprend que les parémies des listes finales se répartissent comme suit :

[168] C'est une des intentions de l'étude de Navarro Brotons que de mettre l'accent sur la nécessité de recueils parémiographiques à la page. Sa thèse s'inscrit dans le cadre d'un projet de recherche sur les constructions verbales figées les plus fréquentes en espagnol, dirigé par Mogorrón et financé par le Ministère de l'Éducation et de la Science espagnol (2012 : 468).

Paremias españolas

Estructuras	Número de paremias tras el primer vaciado de diccionarios	Número de paremias tras la criba en Internet
A	716	292 (265 + 27)[91]
MÁS	256	98
NO	760	208
QUIEN	971	292 (271 + 21)[92]
TOTAL	2.703	889

Tabla 12. Paremias españolas

Paremias francesas

Estructuras	Número de paremias tras el primer vaciado de diccionarios	Número de paremias tras la criba en Internet
À	500	159 (153 + 6)[93]
MIEUX	162	82
NE	656	243
QUI	581	213 (190 + 23)[94]
TOTAL	1899	697

Tabla 13. Paremias francesas

Figure 5. Répartition des listes provisoires et finales des parémies espagnoles et françaises par mot initial (Navarro Brotons 2013 : 85).

Il faut encore préciser que la fréquence est aussi invoquée par les auteurs au moment où ils doivent trier les variantes proverbiales concernées, en l'occurrence, par la présence ou par l'absence de déterminants. La fréquence devient ainsi un pivot discriminant qui autorise un jugement de justesse de telle ou telle autre variante :

> « de façon à ne travailler qu'avec des variantes fréquentes et pouvoir éliminer des variantes dues à un usage incorrect » (2012 : 462).

À titre d'exemple, les auteurs mentionnent les variantes morphosyntaxiques espagnoles suivantes :

Al enemigo que huye, puente de plata
A enemigo que huye, puente de plata

D'après leur consultation :

> « la forme *Al enemigo que huye, puente de plata* apparaît plus de 20 000 fois sur Internet tandis que la forme *A enemigo que huye, puente de plata* apparaît plus de 120 000 fois » (*ibid*.).

Par conséquent, pour la description syntaxique et pour le calcul de la productivité interne de la structure syntaxique : *À* N_1 N_2, c'est la variante *A enemigo que huye, puente de plata* qui est retenue (2012 : 460), non pas *Al enemigo que huye, puente de plata*, et ce, pour l'exigüité (20.000 !) d'attestations repérées par *Google*. Les chiffres motivent des choix qualitatifs. Et pourtant, on emploie des chiffres et la 'synchronie' entre une interrogation du Web et le moment du développement d'une enquête linguistico-parémiologique pour juger de la justesse d'usage d'une variante morphosyntaxique par rapport à une autre. On ignore ainsi la dynamique de l'usage proverbial en diachronie, soit-elle macrodiachronie (au fil des siècles) (Conenna *et al.* 2006) ou microdiachronie (au fil des années) (Marcon 2011, 2012). On empêche l'usage concurrentiel des variantes morphosyntaxiques pour une seule surface proverbiale en vue d'une description syntaxique monoréférentielle, et ce, même en contradiction avec le principe de l'exhaustivité descriptive propre à la méthode du Lexique-Grammaire. Au fond de ces choix se trouve un nombre d'attestations qui est un indice chiffré basé sur une archive virtuelle qui contient du matériel écrit (et bien d'autres matériaux). En effet, on fait référence à une fréquence qui n'est pas vraiment telle, mais un chiffre qui présume une fréquence et décide improprement, en l'occurrence, du sort syntaxique des répertoires parémiologiques espagnol et français.

Or, s'il est vrai que les considérations sur la fréquence et sur la productivité interne sont subordonnées à la visée descriptive d'ordre syntaxique et d'ordre lexicographique, elles représentent quand même le point de départ de toute l'analyse et orientent ses résultats. Quoiqu'elle passe en arrière-plan, la fréquence rentre forcément en ligne de compte à chaque fois qu'elle est imbriquée dans des études qualitatives et qu'elle concerne la constitution d'une liste de parémies à analyser, comme dans ce cas. Même après tous ces passages quantitatifs que nous venons de présenter, Navarro Brotons précise que la description syntaxique est fonction de la productivité d'une structure syntaxique à l'intérieur des listes finales des parémies. Ce qui veut dire que les parémies font partie des tables du Lexique-Grammaire si et seulement si la même structure syntaxique est identifiée au moins 2 fois dans la liste finale des parémies (2013 : 163, 227, 283, 335). On peut interpréter ce dernier souci quantitatif par la création des classes d'équivalence syntaxique des parémies. Pourtant, il est évident que tous les comptes sur l'actualité des parémies en amont biaisent significativement les résultats obtenus.

3.3. Proposition d'un cadre méthodologique fédérateur pour les études de fréquence

Au terme de notre analyse critique, nous essayons de tirer parti des points forts et des points faibles des expériences d'autres parémiologues pour proposer un cadre méthodologique fédérateur ainsi qu'une terminologie conséquente. Bien entendu, notre proposition se veut une tentative inclusive, mais aussi non exhaustive et non impérative qui cherche à rendre compte des études que nous avons mentionnées (et de la nôtre). Notre souhait est que ce cadre enrichisse la typologie des études de fréquence élaborée par Grzybek & Chlosta en parémiologie empirique (§ 2.1.2.). Nous espérons qu'il puisse favoriser la compréhension mutuelle entre parémiologues et une meilleure interopérabilité. En tout cas, ce cadre méthodologique constituera l'arrière-plan idéal des choix que nous effectuerons par la suite.

3.3.1. *Liste*

Plutôt que *corpus expérimental* (§ 2.1.1.), le terme *liste* indiquera l'ensemble des proverbes et, en général, des parémies qui font l'objet de l'étude de fréquence, comme nous l'avons déjà anticipé (§ 2.3.). Faute d'un repère méthodologique commun à cet égard, à partir des expériences creusées dans les paragraphes précédents, nous dégageons la macrotypologie qui suit :

- ***liste zéro*** (***l0***), pour faire référence à une liste *l* vide et au recours à :
  - des introducteurs et des séquences formulaires qui signalent des parémies en discours ;
  - des unités lexicales qui appartiennent aux parémies.

  Les introducteurs et les unités lexicales sont choisis :
  - ◊ au hasard ;
  - ◊ d'après une (ou plusieurs) expérience(s) disponible(s) dans la littérature ;
  - ◊ suivant une étude préalable ou juxtaposée du même parémiologue.

  Rien n'empêche que *l0* soit accompagnée à une des deux listes que nous présenterons ci-dessous. Celles-ci seront employées comme des ***listes de***

référence, c'est-à-dire comme des listes de vérification, inclusion ou exclusion des résultats obtenus.

- ***liste cible*** (***lc***) pour indiquer une liste *l* dont le nombre de parémies $p \geq 1$ et dont les parémies appartiennent à :
  - d'autres études de fréquence, à savoir *l* reprend l'intégralité ou une partie d'une (ou plusieurs) liste(s) de parémies établie(s) par d'autres parémiologues ;
  - des études de familiarité, c'est-à-dire que *l* inclut l'intégralité ou une partie d'une (ou plusieurs) liste(s) de parémies établie(s) par le parémiologue lui-même et/ou par d'autres parémiologues ;
  - d'autres expériences disponibles dans la littérature, qu'elles soient assez contemporaines ou éloignées dans le temps.
- ***liste parémiographique*** (***lp***) pour signifier une liste *l* dont le nombre de parémies $p \geq 1$ tirées d'un (ou plusieurs) recueil(s) parémiographique(s)[169]. Le(s) recueil(s) parémiographique(s) peu(ven)t être sélectionné(s) sur la base d'un (ou plusieurs) critère(s), notamment :
  - ◊ au hasard ;
  - ◊ d'après leur(s) public(s) de référence (spécialistes et/ou non spécialistes) ;
  - ◊ selon un critère linguistique (ex. classement syntaxique, description d'une variété diatopique, description comparée entre langues, etc.) ;
  - ◊ suivant un critère philologique, c'est-à-dire reconstruire l'*historique du proverbe* (Conenna 2002) et, en général, d'une parémie ;
  - ◊ en tenant compte d'autres critères extralinguistiques (ex. données des ventes, distribution sur l'échelle locale, nationale ou internationale, disponibilité/gratuité de la consultation, voisinage/éloignement temporel de la parution par rapport au moment de l'étude, année-pivot et année-contraste d'un recueil ou d'un ensemble de recueils parémiographiques).

  Une fois le(s) recueil(s) choisi(s), ***lp*** sera :

[169] Par *recueil parémiographique* on entend tout dictionnaire, glossaire et/ou toute collection de parémies conçus suivant une approche parémiographique sémasiologique ou onomasiologique.

- ***intégrale*** si elle comprend toutes les formes canoniques (et/ou toutes les variantes éventuellement mentionnées) du(des) recueil(s) parémiographique(s) ;
- ***échantillonnée*** dans le cas où elle contiendrait une partie des formes canoniques (et/ou des variantes éventuellement mentionnées) du(des) recueil(s) parémiographique(s). L'échantillonnage peut s'effectuer :
  - ◊ au hasard ;
  - ◊ par tri alphabétique (ex. parémies qui commencent par une lettre alphabétique) ;
  - ◊ par tri de fréquence concernant le(s) premier(s) mots graphiques des parémies ;
  - ◊ par mots-clés thématiques indexés ;
  - ◊ par tri de fréquence concernant les sources mentionnées pour chaque parémie.

Au cas où le parémiologue prévoirait un processus de filtrage pour aboutir à une de ces listes, nous suggérons le terme ***liste bêta*** (***lβ***) pour parler d'une liste *l* ayant un nombre de parémies $p \geq 1$ et qui doit être soumise à une ***référence*** :

- ***linguistique*** (ex. exclusion/inclusion de parémies qui commencent par un mot graphique, exclusion/inclusion de parémies qui présentent un fait morphosyntaxique, sémantique ou autre, etc.) ;
- ***extralinguistique*** (ex. exclusion/inclusion de par le jugement donné par des spécialistes, exclusion/inclusion de par une étude de familiarité, exclusion/inclusion de par la mise à jour d'un recueil parémiographique ou de par le recours à des recueils parémiographiques non consultés, etc.).

À l'issue du processus de filtrage, on peut ainsi déboucher sur une liste *l* ayant un nombre de parémies $p \geq 1$ supérieur, inférieur ou (éventuellement) égal à *lβ* et qui peut correspondre à *l0*, *lc* ou *lp*.

En fonction de(s) critère(s) d'établissement d'une liste et de(s) processus (éventuels) de filtrage choisis, il est envisageable de concevoir et d'exploiter une liste de la même façon qu'on conçoit et qu'on exploite un corpus. Seulement dans ce cas de figure, on parlera de

corpus parémique pour faire référence à une *l0*, *lc* ou *lp* qui possèdent les critères (i)-(v) que nous avons identifiés pour définir un corpus (§ 2.2.2.) et dont les traitements d'établissement et d'utilisation correspondent à ceux réservés aux corpus.

3.3.2. *Sources*

Nous n'allons pas vraiment remettre en question la typologie des corpus approfondie au § 2.2. Nous nous concentrerons plutôt sur les choix que les parémiologues ont opérés dans leurs recherches afin de les systématiser. D'abord, il faut distinguer :

A. la consultation d'un corpus ;
B. la consultation d'une base de données textuelle ;
C. la consultation du Web comme corpus.

Pour A, selon les critères présentés au § 2.2. :

- *Taille*

  Au fur et à mesure, le nombre d'occurrences des corpus a augmenté et, par conséquent, la fréquence des parémies. D'après les études mentionnées, on peut distinguer :

  - corpus de taille ≤ environ 1 million d'occurrences, ce qui est notamment le cas des corpus spécialisés en raison du discours traité ou du genre textuel ciblé ;
  - corpus d'une taille entre environ 10 millions et 40-50 millions d'occurrences, c'est-à-dire les corpus spécialisés de presse et les tout premiers corpus généraux ;
  - corpus de taille qui compte plus ou moins 100 millions d'occurrences, à savoir les corpus généraux qui commencent à tirer parti des procédures de numérisation du format papier et de la vulgarisation du Web pour les corpus ;
  - corpus de taille ≥ 1 milliard d'occurrences, notamment les corpus dynamiques et les corpus qui résultent de l'aspiration massive de données textuelles du Web.

- *Temps*

  Il n'y a pas vraiment de corpus synchroniques qui sont interrogés. Il y a plutôt manière de séparer :

  - des sources textuelles (et donc elles ne sont pas de *corpus* au sens que nous avons spécifié) macrodiachroniques au format papier, c'est-à-dire des collections de sources qui traversent les siècles et qui sont très rarement utilisées ;
  - des *corpus microdiachroniques*, à savoir des corpus qui couvrent une décennie ou, au plus, quelques décennies d'un même siècle. En ce sens, on pourrait sous-distinguer :
    - ◊ des *corpus strictement microdiachroniques* ou ces corpus qui rassemblent des textes qui relèvent au maximum d'une seule décennie et se rapprochent davantage d'un corpus synchronique ;
    - ◊ des *corpus approximativement microdiachroniques*, voir les corpus dont les textes réunis dépassent une seule décennie et restent en deçà d'un siècle.

- *Médium*

  La plupart des études exploitent des corpus de l'écrit. Très rarement, l'interrogation concerne l'oral qui est, d'ailleurs, souvent confondu à l'écrit dans un même corpus. Ce qui est le cas des tout premiers corpus généraux. Plus récemment, l'introduction du Web pour les corpus invite à prendre en considération les caractéristiques de l'écrit 2.0 et, plus en général, de l'écrit numérisé[170].

- *Langue*

  Les corpus monolingues et les corpus comparables sont les deux types de corpus intéressés par les recherches examinées. Aucun cas d'exploitation de corpus parallèles n'est présent en littérature. Il en va de même pour des corpus qui mélangent deux (ou plusieurs) langues en un seul corpus.

- *Discours*

  La variété des discours représentée dans un seul corpus est décidément

[170] Nous nous sommes déjà confronté avec les particularités de l'écriture sur le Web, notamment en ce qui concerne la presse en ligne et les commentaires des lecteurs aux articles (Marcon 2011) ainsi qu'avec les réseaux sociaux (Marcon 2012).

préférée aux discours spécialisés. Non seulement le choix de ces derniers affecte sensiblement la taille du corpus, mais aussi il restreint, de par leur nature, le public ciblé. La variété des discours cible souvent un public hétérogène, et ce, par le recours à la presse et à toutes ses rubriques spécialisées. Moins souvent, les discours littéraires, des sciences humaines et/ou des sciences sociales font l'objet d'une enquête ponctuelle. Ils sont plutôt juxtaposés l'un à l'autre.

- *Genre*

  De manière conséquente au discours, on peut affirmer que macrogenre textuel *article de presse* et ses déclinaisons l'emportent, par exemple, sur les *articles scientifiques*, les *essais*, les *rapports*, les *traités* ou les *courriers électroniques*.

La comparaison entre corpus est également possible, mais elle intéresse plutôt les fréquences, non pas la composition des corpus.
Nous soulignons encore qu'aucune étude n'a exploité un corpus étiqueté.

Quant à B, il y a, certes, moyen de rapprocher la consultation des bases de données textuelles à ce que nous avons expliqué pour A. Il ne faut pas oublier, en tout cas, qu'une archive n'est pas conçue de la même façon et pour les mêmes finalités d'exploitation d'un corpus. Son interrogation diffère et les traitements informatiques envisageables pour une base de données sont limités par rapport à ceux d'un corpus.

En ce qui concerne C, nous avons des fortes perplexités à l'égard de l'identité Web = corpus. On peut interroger le Web *comme s'il s'agit d'*un corpus, mais l'exploitation tout court de ses résultats (autant pour des descriptions quantitatives que qualitatives) doit être mitigée par des jugements de pertinence successifs. Les résultats de l'exploitation du Web comme corpus sont au plus des suggestions ou des indications d'orientation pour le déroulement d'une étude de fréquence (à n'importe quelle étape). Ils ne sont ni les termes ultimes de l'étude ni leur seule référence. Autant vaut considérer, donc, le Web (et les bases de données textuelles) comme des réservoirs où puiser les textes qui constitueront un corpus et déclarer tout de suite les critères qui guident leur constitution.

3.3.3. *Fréquence*

Avant de proposer une réorganisation des recherches analysées, nous souhaitons préciser que :

> « […] quand on quantifie des faits linguistiques, […], on interprète, par des approximations quantitatives, une dimension déjà présente dans le fonctionnement de l'objet » (Loiseau 2011 : 67).

Dans notre recherche, la répétition dans des corpus soulignera la dimension lexico-grammaticale qui est en œuvre dans le fonctionnement de l'objet linguistique-proverbe (et, en général, de l'objet linguistique-parémie). C'est cette dimension que nous considérerons comme étant à la base de la (re)production des proverbes et des parémies ainsi que de séquences formulaires détournées et/ou nouvelles. On est bien conscient que :

> « […] les fréquences étiques (mesurées empiriquement) ne peuvent être que des indices reflétant une conjonction complexe de causes historiques diverses, interprétables seulement en prenant en compte la diversité des phénomènes variationnels et textuels, et relevant d'une herméneutique […] » (Loiseau 2011 : 75).

D'ailleurs, la présente proposition de cadre méthodologique ainsi que les points théoriques (§ 1.) et paradigmatiques (§ 2.) de départ, se veulent un parcours herméneutique possible pour réfléchir sur les fréquences des proverbes et des parémies issues de l'interrogation d'un corpus. Nous avons déjà considéré la « diversité des phénomènes variationnels et textuels », voire les « causes historiques », dans la mesure où nous avons essayé de mieux cerner les choix qui caractérisent la mise au point de la *liste* et la constitution du *corpus*. Nous enrichissons ce parcours herméneutique par d'autres considérations que l'ensemble des études examinées nous suggère sur la récurrence des parémies en elle-même dans différents corpus et en plusieurs langues.

Tendances

En ligne générale, au moment de l'interrogation d'un corpus général, on peut s'attendre que :

1) une partie des proverbes (ou des parémies) de la liste a $f = 0$ et que, par rapport au total, le pourcentage de ceux (ou de celles) ayant $f = 0$ est d'autant plus élevé que la taille du corpus est restreinte et, à l'inverse, d'autant moins élevé que la taille du corpus devient de plus en plus importante. $f = 0$ n'implique pas le manque de connaissance ou la disparition de ces proverbes (ou parémies). Il est plutôt raisonnable de croire que cette absence est à corréler au fait qu'il n'y a pas d'*ancrage pragmatique* (Schmale 2013 : 36-37) dans les textes du corpus : l'usage des proverbes (ou des parémies) est inapproprié et ne sert pas à catégoriser (par jugement évaluatif, par référence directe ou par analogie) les situations extralinguistiques sur lesquelles portent les textes du corpus ;
2) la distribution des f des proverbes (ou des parémies) est assez homogène, c'est-à-dire que la quasi-totalité des parémies se regroupe autour des mêmes valeurs chiffrées ;
3) indépendamment de la langue, au moins 50% des proverbes (ou des parémies) d'une liste ont des f qui privilégient certaines valeurs chiffrées et que ces valeurs sont comprises entre :
    a. $1 \leq f \geq 5$ pour les corpus d'une taille qui va environ de 10 jusqu'à 50 millions d'occurrences ;
    b. $1 \leq f \geq 10$ pour les corpus dont la taille se situe autour de la centaine de millions d'occurrences.

    Ce pourcentage tend à augmenter en fonction de l'augmentation de la taille du corpus. Excepté les proverbes (et, en général, les parémies) dont $f = 0$, ces plages de valeurs se construisent autour de la mode de la distribution, c'est-à-dire dans une fenêtre (indicative) de ± 2 pour a. et de ± 5 pour b. par rapport à la valeur la plus représentée dans la distribution.
4) lorsqu'on interroge des corpus dont la taille commence à s'élever environ à 10 millions d'occurrences et dépasse cette limite, certaines parémies tendent à s'écarter du reste de la distribution ; on a donc des proverbes- (ou parémies-) *outliers* ;

5) le nombre des proverbes- (ou parémies-) *outliers* et les écarts qui se produisent entre ceux-ci et le restant de la distribution, tendent à augmenter quand la taille du corpus augmente, elle aussi.

Il faut en conclure que, dans un corpus général, les *f* des proverbes et des parémies qui se situent dans les plages de valeurs mentionnées au 3) sont à interpréter comme attendues et normales. Autrement dit, il ne s'agit pas de *f* basses ou non significatives, mais plutôt de *f* pertinentes et prévues pour des séquences formulaires. Il faudrait renverser ou, du moins, ajuster l'optique traditionnelle lexicométrique où la normalité et la significativité d'une distribution sont habituellement confiées aux *f* attestées et/ou attendues ayant des valeurs élevées. En ce sens, on peut proposer l'adoption d'une ***optique parémiométrique*** (suivant la suggestion terminologique de Rodegem, § 3.1.3.) qui remet en question certains acquis et certaines tendances observées en lexicométrie traditionnelle.

Nous suggérons également de regarder avec quelques précautions ces proverbes- ou parémies-*outliers* mentionnés aux 4) - 5) parce que leur *f* :

- ou relève d'un effet de mode (comme l'a observé Arnaud, § 3.2.3.) et on aurait un(e) ***proverbe/parémie en vogue*** ou ***à la mode*** ;
- ou signale la présence du nombre *n* de situations extralinguistiques décrites, commentées ou comparées à la situation prototypique dénommée par un proverbe ou une parémie. Dans ce cas de figure, ces **parémies-*outliers*** joueraient le rôle de pointeurs d'une dynamique situationnelle récurrente, d'où leur fréquence ;
- ou indique les deux en même temps.

Toujours en ligne générale, au moment de l'interrogation d'un corpus spécialisé, il faut s'attendre à que la quasi-totalité des proverbes ou des parémies recherchés ait $f = 0$. Sauf pour les cas où l'on aurait un intérêt scientifique spécifique sur des parémies dont l'usage est lié à un domaine (par exemple, les adages juridiques) ou que le discours spécialisé touche de près la parémiologie (§ 6.1.2.), il paraît évident que les corpus spécialisés sont déconseillés pour effectuer une étude de fréquence de portée générale.

Présentation des données quantitatives

Quant à la présentation de la fréquence, jusqu'à présent la grande majorité des parémiologues a fait recours à la *fréquence brute* d'occurrence, à l'exception de Moon qui est la seule à introduire la *fréquence normalisée* (§ 3.2.5.). De temps en temps, on a traduit ces mesures de la fréquence en *pourcentages*. Les regroupements en *plages* sont envisageables, mais elles sont (souvent) arbitraires qui servent à justifier une vue synthétique sur l'ensemble de la distribution.

On a fait recours à la *moyenne arithmétique* normalement pour définir la relation :

- entre la longueur des proverbes (ou parémies) de la liste et en établir une longueur moyenne ;
- entre la longueur moyenne des proverbes (ou parémies) et le nombre d'occurrences totales du corpus interrogé pour définir l'intervalle de parution (en occurrences) d'un proverbe (ou parémie) dans un corpus ;
- entre le nombre *p* de proverbes (ou parémies) et le nombre *o* total d'occurrences repérées (*type*/*token ratio*).

Pour le premier cas, le calcul concerne la même catégorie dont les membres ont (à peu près) les mêmes caractéristiques. Au contraire, les deux autres suggèrent une approximation assez douteuse, vu que l'on compare deux catégories, de fait, difficiles à comparer. Les occurrences d'un corpus coïncident d'habitude avec les mots graphiques simples (et parfois, avec quelques unités polylexicales préalablement codifiées, dans le cas de logiciels ou programmes *language-dependent*) et leur calcul est influencé par la norme de dépouillement automatique, c'est-à-dire par les grammaires (informatiques) de segmentation et de tokenisation du corpus (Habert *et al*. 1997 : 194-197). Au moins d'avoir une norme de dépouillement où les proverbes (et les parémies) sont prévus[171], la définition d'un intervalle moyen de parution d'un proverbe (ou parémie) et le *type/token ratio* parémique ne sont que des mesures fort approximatives et très peu fiables[172].

[171] Ce qui n'est pas le cas pour la quasi-totalité des logiciels, avec la seule exception du traitement de certains proverbes grecs sous *Unitex* (§ 4.2.2.2.).
[172] En général, on évite la moyenne arithmétique parce qu'elle est sensible aux données aberrantes. En l'occurrence, elle est sensible aux parémies-*outliers*.

La présentation de *f* des parémies concerne aussi trois types de répartition de la distribution d'après :

- les genres textuels qui composent le corpus ;
- les sous-corpus d'un corpus général ;
- les processus de variation.

En ce qui concerne les genres textuels et les sous-corpus, on fera référence encore aux fréquences brute et normalisée ainsi qu'aux pourcentages, pourvu que la répartition se base sur une documentation éclairée du corpus et sur des catégories discrètes.
Quant aux processus de variation, compte tenu des différentes étiquettes qu'on leur attribue[173], nous pouvons quand même identifier des macrotypes de variation qui sont communs à plusieurs langues. On distinguera :

1. l'**adjonction** à n'importe quel endroit de la séquence-parémie et de tout type d'unité ;
2. la **réduction**, qui inclut la troncation, l'allusion ou la reprise d'un ensemble d'unités qui relèvent de la séquence-parémie ;
3. la **permutation** de l'ordre des unités de la séquence-parémie ;
4. la **substitution paradigmatique**, c'est-à-dire le remplacement d'une (des) partie(s) du discours et unité(s) lexicale(s) relevant de la même (des mêmes) *classe(s) d'objets* (G. Gross & Clas 1997, Le Pesant & Mathieu-Colas 1998), *colligation* ou *préférence(s) sémantique(s)* (§ 2.2.6.).

Lors du calcul des occurrences et de leur répartition, on distinguera :

- la ***fréquence d'usage créatif***, c'est-à-dire le nombre d'occurrences qui témoignent du recours (i) soit à un de ces 4 processus de variation, (ii) soit à leur enchevêtrement en une seule occurrence, (iii) soit à toute autre réutilisation par **analogie formelle et sémantique** avec les éléments qui composent la séquence-parémie ;

[173] Il y a des études qui visent le détournement/défigement des proverbes et des parémies, comme, entre autres, celles de Barta (2005, 2006) qui détaillent minutieusement toutes les possibilités de variation et d'enchevêtrement entre les processus.

- la ***fréquence d'usage standard***, à savoir le nombre d'occurrences qui reproduisent la forme exacte recherchée, qu'elle relève d'une norme parémiographique ou non.

De cette façon, on pourra mieux comprendre l'influence de l'*autorité de la norme* ou de l'*autorité de l'usage* (§ 3.2.15.) dans les corpus interrogés.
Pour ce qui en est de la variation morphosyntaxique, nous proposons que :

- dans les cas où elle permettrait une adaptation nécessaire pour l'insertion d'un proverbe (d'une parémie) et qu'elle est le seul processus de variation mis en œuvre, l'occurrence de ce proverbe (de cette parémie) est comptabilisée dans la fréquence d'usage standard ;
- dans les cas où la variation morphosyntaxique interviendrait pour favoriser l'insertion d'un proverbe (d'une parémie) et qu'elle interagit avec d'autres processus de variation, l'occurrence de ce proverbe (de cette parémie) est énumérée dans la fréquence d'usage créatif.

3.4. En guise de conclusion

Dans ce chapitre, nous avons illustré l'apport de quelques études de fréquence (ou, en général, de traitement quantitatif) en parémiologie et en phraséologie. Au cours de ce recensement critique, les trois fils rouges de la liste, des sources et du calcul et de la présentation des données de fréquence, nous ont permis de mettre au point une proposition de cadre méthodologique fédérateur. Ce cadre a essayé de synthétiser et d'organiser de façon systématique les interrogatifs et les réponses que ces études ont montrés à différentes périodes et pour plusieurs langues. Notre tentative d'harmonisation méthodologique veut établir des repères communs pour les études de fréquence en parémiologie empirique, mais surtout une palette de choix de référence pour notre recherche.

Pourtant, il est évident que les tendances des fréquences enregistrées par les travaux que nous avons examinés, d'une part, sont influencées par l'optique lexicométrique traditionnelle qui ne peut suffire à encadrer les études parémiométriques. D'autre part (et surtout), les données quantitatives sont fonction des techniques de reconnaissance ('à vue' et automatiques) choisies par les parémiologues parmi les techniques disponibles. Notre proposition d'un cadre méthodologique nécessite donc d'être complétée par une vue

d’ensemble critique sur les approches au repérage des proverbes (et des parémies) sur corpus (§ 4).

CHAPITRE 4

REPERAGE DES PAREMIES. UN EMPIRISME FEDERATEUR

Dans ce chapitre, nous focaliserons notre attention sur l'activité de repérage des parémies dans des textes. Pour mieux distinguer les démarches empiriques adoptées, nous séparerons les études en deux grandes sections :

4. les études qui portent sur la détection 'à vue', c'est-à-dire sur la reconnaissance des parémies dans des sources sur support papier (§ 4.1.) ;
5. les études qui envisagent la détection (semi-)automatique, à savoir la reconnaissance des parémies dans des sources au format électronique, et ce, à l'aide de requêtes informatiques (§ 4.2.).

Nous essayerons de mettre en relief les combinaisons des techniques adoptées ainsi que leurs ajustements (éventuels) au fur et à mesure. Ces techniques nous serviront pour compléter le cadre méthodologique que nous avons développé au § 3.3. ainsi que pour proposer un point de départ pour le repérage automatique des parémies dans des corpus (§ 4.3.).

4.1. Détection 'à vue'

Les paragraphes qui suivent résument quelques expériences qui concernent le dépouillement manuel de sources au format papier et la reconnaissance visuelle des parémies. Comme on le verra, les suggestions des parémiologues s'insèrent dans le sillage d'une approche structuraliste par traits dont l'opérationnalisation informatique est parfois complexe. La généricité de quelques-uns des traits suggérés représenterait un obstacle à la précision des

résultats. Malgré cela, nous avons choisi de considérer ces études parce que, par rapport à certaines pratiques plus récentes de détection automatique, elles constituent des contre-exemples qui témoignent de l'importance d'une réflexion linguistique ponctuelle et préalable ainsi que d'une observation empirique systématique.

4.1.1. *Buridant* (1976)

De la littérature parémiologique disponible au moment de son étude, Buridant dégage trois typologies de traits formels en vue de la détection des proverbes. D'abord, il reprend les *traits d'identification* (1976 : 391-393) de Rodegem (1972), notamment le rythme et deux critères sémantiques, à savoir la norme et la métaphore. Par ces traits, il peut opérer une distinction typologiques entre proverbes, adages, maximes et dictons. Ensuite, Buridant se concentre sur les *traits de spécification* (1976 : 393-395) du proverbe et qui « n'apparaissent que facultativement et isolément dans tel énoncé ». Ces traits :

- absence d'article,
- ordre des mots 'non conventionnel',
- pronom relatif *qui* autarcique,
- structure rythmique binaire,
- présent de l'indicatif et impératif de préférence,

sont repris de Greimas et représentent un aléa formel pour la reconnaissance des proverbes. Autrement dit, il s'agit de traits récurrents dans le répertoire parémiologique, mais qui ne sont pas forcément présents. Ce qui implique un jugement subjectif du parémiologue face à chaque séquence formulaire.
Pour finir, Buridant identifie des *traits de classification* (1976 : 395-398), à savoir :

> « […] des schèmes qui se dégagent à travers un certain nombre de proverbes et qui dessinent, pour ainsi dire, leur morphologie » (1976 : 395-396).

Un nombre *n* d'agencements séquentiels et formels récurrents décrivent et rassemblent les proverbes en groupes. Ces agencements résultent d'une combinatoire qui met en jeu un nombre fermé d'éléments linguistiques récurrents et, par conséquent, une sémantique/narration particulière (implication de convenance ou de conséquence, identité,

concaténation entre implication et identité, possession, préférence). C'est à ce stade que Buridant introduit la référence au travail de Paulhan de 1925, notamment à l'existence d'un *moule* (Paulhan 1993 [1925]) (re)producteur de séquences formulaires analogues (§ 1.1.4.).

Outre ces traits, Buridant remarque aussi que les proverbes peuvent être signalés par des « termes introducteurs en nombre restreint » et donc faciles à repérer dans les textes (1976 : 398-400), comme *sachiez* ou *oïr dire*.

Comme on le disait au tout début, l'étude de Buridant met le doigt sur les difficultés que la tâche de reconnaissance des proverbes comporte, et ce, parce qu'il faut tenir compte qu'on peut repérer en discours :

> « métaphores qui se résolvent en proverbes, proverbes qui se dissolvent en métaphores, jeu de comparaisons coulées dans des proverbes, jeu de variations sur des schèmes de proverbes […] » (1976 : 418).

Quoiqu'on ne dispose pas toujours de réponses satisfaisantes pour faire face à une telle variété, l'étude de Buridant a le mérite d'éveiller notre conscience sur l'étendue des variations morphologiques et sémantiques qui intéressent les proverbes.

4.1.2. *Rodegem* (1984)

Trois ordres de traits aident Rodegem à distinguer les parémies entre elles (1984 :133) :

{1} la *structure rythmique* (1984 : 122-123) : seul trait distinctif corrélé à la forme, la structure rythmique caractérise toute parémie (y inclus le proverbe), à l'exception près des locutions proverbiales et des wellerismes. À cet égard, Rodegem ne suggère aucune opérationnalisation, sauf les brèves remarques sur une *symétrie* prosodique des parémies qui s'appuie sur un effet de répétition régulière[174].

{2} la *structure sémantique* (1984 :123-124) : la référence au monde réel s'établit autant sur le plan d'une *structure analogique* que sur le plan de la *dénotation*. La première, trait distinctif propre aux proverbes (comme aux dictons et aux

[174] Pour des approfondissements sur le rythme des proverbes français, en particulier, cf. Anscombre (2000), D'Andrea (2005, 2008).

locutions proverbiales), fait que le sens exprimé par la composition logico-sémantique des éléments qui constituent la parémie, entretiennent « une relation d'équivalence » (1984 : 123) en rapport avec d'autres éléments du monde réel des locuteurs qui l'emploient. Autrement dit, les parémies – et les proverbes, en particulier – cachent, comme des codes chiffrés succincts, « du 'déjà-vu' » (*ibid*.), à savoir une expérience éprouvée, un fait vécu et une évidence. Pour Rodegem, « le contenu du message sentencieux est empirique » et donc « la parémie n'innove pas » (*ibid*.) : elle réécrit le réel sous forme d'analogies cryptées qui fait « appel au subconscient » (1984 : 126) et qu'il faut interpréter par le recours à des schémas logiques abstraits[175]. La *dénotation* relève plutôt des maximes, des slogans et des devises, qui font directement référence au mondé réel.

{3} la *norme* : trait « commun dénominateur des parémies » (1984 : 124) qui caractérise les proverbes dans la mesure où ils renvoient à une *norme générale*. Elle se fonde sur l'autorité qui vient du partage social de certaines règles et coutumes. Comme « la norme transparaît dans la forme » (1984 : 125) et privilégie l'injonction et l'indicatif comme ses modes verbaux[176], elle s'exprime comme un tout discursivo-textuel en tant que :

a) constat : la parémie est à l'indicatif et suit l'ordre non marqué sujet-verbe-objet (SVO) ;
b) avertissement : la parémie acquiert la forme d'une implication '*Si…alors…*' ;
c) dénonciation satirique ;
d) conseil : souvent modulé de manière positive ;
e) préférentielle : la parémie commence, par exemple, par '*Il vaut mieux…*' ;
f) dissuasion : souvent modulée de manière négative ;
g) obligation : la parémie commence, par exemple, par '*Il faut…*' ;
h) défense : l'impératif se manifeste dans la parémie (*ibid*.)[177].

[175] À cet égard, on reconduit cette remarque au § 1.
[176] À côté de la *norme générique*, Rodegem ajoute la *norme spécifique* des adages juridiques et des slogans et la *norme restreinte* des apophtegmes, des devises et des locutions proverbiales. Elles s'appliquent respectivement à un groupe limité et à un seul individu (1984 : 125).
[177] Pour les lettres a), b), e), g) et h), c'est nous qui dégageons les traits formels d'après les exemples présentés par Rodegem. Pour les lettres c) et f), nous respectons à la lettre ce que Rodegem affirme. Pour la lettre c), nous nous limitons à respecter la version de Rodegem, sans pourtant être en mesure de dégager un véritable trait

Comme le précise Rodegem lui-même, l'identification de ces traits constitue, dans son ensemble, une « analyse descriptive sommairement développée » (1984 : 127). Nous n'avons ni accès à une énumération ponctuelle de certaines régularités ni à des exemples. Cette analyse, toutefois, se veut un premier repère pour la mise en relation de la détection des parémies (et donc des proverbes) avec leur fréquence.

4.1.3. *Schulze-Busacker* (1985)

Schulze-Busacker s'appuie, d'abord, sur la littérature en parémiologie disponible au moment de son analyse. Elle accorde une attention particulière au travail consacré aux traits formels proverbiaux, d'où la centralité de l'autonomie formelle du proverbe, de la syntaxe archaïque, de la construction binaire syntaxique, mais aussi rythmique, ainsi que la proéminence du temps présent en tant que temps anhistorique (1985 : 16).

À côté des spéculations de la littérature, Schulze-Busacker juxtapose des astuces empiriques qui portent tant sur le lexique proverbial que sur le lexique paraproverbial. Plus précisément, elle se concentre sur des séquences répétées en début de proverbe, comme :

Miauz vaut [Mieux vaut]
Tel cuide [Celui qui]

ainsi que sur des séquences répétées à l'intérieur ou en fin de proverbe, comme :

vengier sa honte [venger sa honte] (*ibid*.).

Elle remarque également la répétition de séquences discontinues, comme :

Tel… tel…

mais sans élaborer sur ces fenêtres de cosélection (*ibid*.). On pourrait dire qu'il s'agit de constatations phraséologiques, tout comme d'une conscience des difficultés que comporte le fait de s'appuyer sur la seule matrice lexicale des proverbes au moment de la détection dans

formel. D'ailleurs, la satire relève davantage d'une interprétation sémantique et pragmatique. D'après les traits identifiés par Rodegem, elle serait donc à reconduire à la structure analogique.

des textes. Ce n'est pas par hasard qu'elle qualifie ces astuces comme « des éléments lexicaux qui peuvent mener sur la piste d'un proverbe » (*ibid*.). Schulze-Busacker envisage ainsi la reconnaissance d'un proverbe dans son intégralité, mais aussi (et surtout, comme on apprend de son classement conséquent des proverbes (§ 1.2.2.)) sous autres formes.

Quant au lexique paraproverbial, Schulze-Busacker se focalise sur « les formules introductives ou conclusives », notamment « le 'car' introduisant la subordonnée proverbiale » (*ibid*.). La reconnaissance d'un proverbe peut ainsi être signalée par un adverbe causal qui présuppose le déploiement d'une leçon morale ou d'une explication quelconque apportée par un proverbe. Par l'adoption de cet expédient, Schulze-Busacker prouve sa familiarité avec les mécanismes d'enchaînements des proverbes en discours, mais, malheureusement (pour nous), elle n'énumère pas de manière systématique toutes les occurrences de lexique paraproverbial dont elle s'est servie lors de son dépouillement.

Outre ces indices lexicaux, Schulze-Busacker en ajoute un autre avec toute précaution : la répétition de la même formule dans des contextes différents (*ibid*.), à savoir dans diverses situations d'énonciation. Elle souligne qu'il s'agit d'une façon peu fiable pour reconnaître des proverbes, parce qu'on peut tomber sur « une formule fixe ou […] une expression sentencieuse » (*ibid*.). Certes, la répétition dans plusieurs contextes d'occurrences n'est pas en soi un critère qui permet de proverbialiser une séquence formulaire : nous partageons le même avis que Schulze-Busacker à cet égard. Toutefois, nous avons également envie de remarquer que, du point de vue linguistique, pour elle, l'assimilation textuelle entre les proverbes et d'autres séquences formulaires n'est pas admissible, alors que certains traits formels peuvent les rapprocher les uns des autres.

Un dernier critère « décisif pour l'identification d'un énoncé comme proverbe médiéval » (*ibid*.) est la référence à la codification lexicographique de ces proverbes au Moyen Âge. Un proverbe médiéval est reconnu comme tel en raison de sa présence dans le métarecueil parémiographique établi par Morawski en 1925 (§ 3.2.4.).

Par rapport à Buridant et Rodegem, Schulze-Busacker réfléchit sur les procédés formels qui caractérisent les proverbes repérés et distinguent les sous-catégories de sa classification (§ 1.2.2.). A posteriori, ces procédés représentent des fondements empiriques pour une opérationnalisation informatique éventuelle. En ce qui concerne la macrocatégorie des *proverbes cités*, Schulze-Busacker sépare trois sous-catégories :

1. les *proverbes isolés*, des proverbes à la forme canonique écartés du suivi de la situation énonciative ;

2. les *proverbes marqués*, qui correspondent à des proverbes à la forme canonique signalés par des formules d'introduction ou de conclusion ;
3. les *proverbes développés*, qui impliquent un démantèlement des composantes formelles du proverbe avec le but de réinventer ses « éléments lexicaux et conceptuels […] pour en formuler un commentaire ajusté au contexte immédiat » (1985 : 29).

Lors d'une opérationnalisation informatique, les deux premières sous-catégories ne posent pas de véritables problèmes parce qu'elles reposent sur la forme canonique du proverbe et sur un ensemble (relativement) fermé d'introducteurs (§ 4.2.2.1.). En revanche, la troisième sous-catégorie échapperait, excepté dans les cas où les requêtes informatiques les prévoiraient ou garderaient une flexibilité suffisante pour garantir une reconnaissance du moins partielle.

La sous-catégorisation se confronte à la subjectivité et à la nécessité d'une prédiction informatique éventuelle au moment où elle intéresse la macrocatégorie *proverbes intégrés*. Pour ces proverbes, Schulze-Busacker différencie :

1. l'*expression proverbiale*, c'est-à-dire des variations libres de la forme canonique proverbiale qui touchent à la structure syntaxique et métrique et qui conservent globalement le concept et le lexique, sauf quelques remplacements d'unités lexicales rapprochées par le sens (1985 : 26-27) ;
2. le *noyau proverbial*, à savoir des variations qui portent sur des « éléments lexicaux et notionnels qui, isolément ou combinés, permettent de rattacher le passage à un modèle proverbial » ou sur un « moule syntaxique, sans utiliser un vocabulaire sémantiquement proche ou identique et une notion comparable » (1985 : 26).
3. L'*arrière-pensée proverbiale*, où la séquence formulaire examinée est rapprochée d'un proverbe parce qu'elle garde une « 'mentalité proverbiale' » (1985 : 25) en raison d'une ressemblance conceptuelle avec d'autres proverbes ou en raison d'une unité lexicale qui évoque d'autres proverbes (1985 : 25-26).

Remarquons que cette sous-catégorisation est essentiellement basée sur la variance du degré de fidélité aux composantes formelles du proverbe et à sa portée sémantique. Par conséquent, le degré d'interprétation et de subjectivité augmente en fonction de la présence ou de

l'absence de traits formels repérés en contexte. Ainsi, la détection automatique d'une arrière-pensée proverbiale sera la conséquence d'une interprétation des données linguistiques reconnues à l'aide de requêtes informatiques génériques fortement axées sur la cosélection (plus ou moins continue) d'unités lexicales. Ce qui engendrerait du bruit dans les résultats. Ce bruit pourrait être réduit lorsque l'interface lexique-syntaxe se fait plus explicite, qu'on accorde une importance particulière à la composante lexicale, comme dans le cas des expressions proverbiales, ou qu'on équilibre lexique et syntaxe et on s'appuie sur des cosélections lexicales et syntaxiques, comme le suggèrent les noyaux proverbiaux. À ce propos, nous tenons à rappeler que les notions mêmes de *noyau proverbial* et surtout de *moule* sont souvent évoquées dans la littérature en parémiologie (§ 1.1.4.). De toute façon, ces trois sous-catégories montrent que, même dans le cas d'une variété diachronique de français loin de celle contemporaine, les proverbes (et le langage formulaire) peuvent être reconnus grâce à la répétition/régularité de cosélection de certains traits formels.

Et la répétition/régularité de cosélection est en quelque sorte à la base de la sous-catégorisation qui concerne les *proverbes exploités*, quoiqu'elle réponde davantage à des exigences stylistiques et textuelles mises en avant par l'auteur. Schulze-Busacker sépare :

1. les *proverbes en série*, à savoir des enchaînements de proverbes qui profitent des proverbes cités et des proverbes intégrés, notamment des expressions proverbiales (1985 : 32-33) ;
2. les *proverbes en tant qu'élément constitutif* d'un texte, c'est-à-dire des amplifications d'un proverbe à une partie de texte par l'exploitation de son noyau proverbial ou par l'enchâssement d'autres proverbes dont il demeure le fil rouge thématique (1985 : 33) ;
3. les *proverbes sous-jacents à une œuvre entière* ou des proverbes qui démarrent ou inspirent la narration de textes dans leur intégralité, que ces proverbes soient explicités ou qu'ils restent implicites tout au long des textes (1985 : 30-32).

L'exploitation dépasse l'objet-proverbe. Son opérationnalisation informatique serait facilitée pour la reconnaissance des proverbes en série ou des proverbes sous-jacents explicites, même par une application à cascade de requêtes qui décrivent la forme canonique des proverbes et qui prennent en compte certaines modifications lexico-grammaticales. Au contraire, pour les proverbes en tant qu'éléments constitutifs et pour les proverbes sous-jacents implicites, il

devient crucial de comprendre quelle composante lexicale des proverbes est la plus significative ou la plus susceptible d'inspiration textuelle.

En général, l'empirisme de Schulze-Busacker nous a donné un éventail de possibilités à retenir en vue de la modélisation de nos requêtes informatiques, quoique sa réflexion porte sur une variété diachronique de français qui ne fait pas vraiment l'objet de notre recherche. En tout cas, ses remarques sur les actualisations et sur les exploitations des proverbes dans des textes sont similaires à celles que d'autres parémiologues ont observées par la suite, au moment de l'interrogation de sources textuelles électroniques.

4.2. Détection automatique

Nous aborderons par la suite les études qui se sont concentrées sur le repérage automatique des parémies dans des sources au format électronique. Par souci de clarté, nous avons distingué les expériences d'après la représentation choisie pour la modélisation des requêtes informatiques. D'une part, nous traiterons les études s'appuyant sur les expressions régulières (§ 4.2.1.) et, d'autre part, celles qui ont préféré les automates à états finis (§4.2.2.). Il en résultera que le recours (sauvage ou, du moins, simpliste) à la reconnaissance de patrons par expressions régulières dépassera en termes quantitatifs le recours aux automates à états finis. Ces derniers sont en effet restés jusqu'à présent contraints à un cadre théorique et méthodologique qui a conditionné leur utilisation pour la reconnaissance des parémies.

4.2.1. *Repérage par expressions régulières*

Par *expression régulière* on entend toute suite de caractères ou symboles encodés d'après une syntaxe qui permet le repérage de patrons textuels, soient-ils simples, comme un mot graphique, ou complexes et contraints, tel :

^*repér** [a-z]{1-3} *expressions régulières*$

qui permettrait la reconnaissance du titre de notre paragraphe où, en l'occurrence : ^ contraint le début de notre première chaîne de caractères à reconnaître ; * favorise la reconnaissance de toute chaîne de caractères contenant *repér* ; [a-z] indique la nature alphabétique de la chaîne

qui suit l'espace / / et {1-3} sa longueur en nombre de caractères ; $ contraint la fin de la chaîne à reconnaître par le mot graphique *régulières* (Fitzgerald 2012).

On verra par la suite que beaucoup de parémiologues ont fait recours aux expressions régulières. Qu'ils aient saisi la forme exacte (partielle ou complète) d'une parémie ou qu'ils aient recherché des introducteurs ou des fenêtres de cosélection lexicale, la plupart des parémiologues ont tiré parti de l'immédiateté des résultats obtenus par l'utilisation des expressions régulières, non sans souci.

On constatera que la modélisation des requêtes informatiques ainsi que les données renvoyées approfondiront des parcours de réflexion sur la nature linguistique des parémies que nous avons déjà entamés dans l'Introduction et au § 1. Ce qui témoigne encore de l'avantage d'une approche empirique.

4.2.1.1. Arnaud & Moon (1993)

Arnaud & Moon abordent les difficultés posées par l'interrogation d'un corpus en vue de la détection des proverbes (1993 : 324-326). Toutefois, leur réflexion ne se penche pas tellement sur les problématiques qui concernent la création des requêtes informatiques. Ils privilégient la recherche guidée par :

> « la co-occurrence de deux ou plusieurs mots lexicaux de chaque proverbe » (1993 : 325).

Par ce choix, ils accordent à la *collocation* une place centrale en parémiographie, et ce, en ligne avec les premiers ouvrages lexicographiques issus d'études sur corpus, comme celle de Sinclair (1991), dont Moon sera une collaboratrice fidèle. Malgré la centralité de la collocation, Arnaud & Moon délaissent le potentiel des schémas collocationnels récurrents, sans décrire à l'avance et systématiquement leur échantillon de proverbes d'après leurs cooccurrences lexico-grammaticales les plus répétées. À ce propos, nous soulignons une remarque d'Arnaud sur les occurrences du proverbe :

> *Un train peut en cacher un autre*.

Toutes ses 12 occurrences sont des variantes obtenues par substitution du nom *train* et par adaptation morphosyntaxique éventuelle des déterminants *un*. Les occurrences repérées exploitent donc la séquence :

Un(e) X peut en cacher un(e) autre

où la position de *X* n'est jamais remplie par l'unité lexicale *train*. Arnaud commente :

> « Sans doute peut-on voir dans sa fréquence la conséquence de son statut de cliché » (1993 : 329).

Une armature lexicale et grammaticale se présente de manière récurrente que les auteurs appellent *cliché proverbial* (§ 1.1.4.). Ils s'inspirent du travail contemporain de Grunig sur la publicité française qui parle d'un « squelette commun minimal » se reproduisant au fur et à mesure des exploitations proverbiales. Arnaud et Moon observent eux aussi l'apparition de ce squelette et l'appellent, comme nous le disions, cliché proverbial, à savoir :

> « un cadre avec une ou plusieurs cases vides où peuvent venir s'insérer des éléments lexicaux, et qui repose sur la forme canonique d'un proverbe » (1993 : 334).

Il n'est pas tout simplement question de cadre en soi et de cases vides. Une propriété du cadre est justement la fréquence de réemploi et d'exploitation du même cadre :

> « Tous les proverbes exploités ne produisent pas de cliché lexical : il faut pour cela que la quasi-totalité des occurrences observées soient exploitées » (*ibid*.).

La reconnaissance de ce cadre est ainsi corrélée à sa réutilisation créative plusieurs fois où l'invariance est représentée par la sémantique dont ce cadre est porteur. Pour Arnaud et Moon :

> « les clichés proverbiaux ne gardent souvent en discours qu'un lien sémantique tenu, voire inexistant, avec leur proverbe-support » (1993 : 335).

Nous nous permettons d'ajouter que c'est ce « lien sémantique tenu » qui confirme l'existence d'un continuum. Dans ce continuum, les parémies (souvent relégués à un pôle isolé) commencent à dialoguer et à établir des relations linguistiques avec les autres unités phraséologiques, voire à contribuer à leur création. Là où les liens sémantiques s'atténuent, les cadres se grammaticalisent et/ou des bribes se lexicalisent. Dans le premier cas, les cadres gardent un squelette lexico-syntaxique qui garantit la reproduction de la sémantique prédicative du proverbe-support initial :

> *Les X se suivent et (ne) se ressemblent (pas)* [proverbe-support : *Les jours se suivent et ne se ressemblent pas*]
> *Half (a(n)) X is better than none* [proverbe-support : *Half a loaf is better than none*]
> *X is in the eyes of the beholder* [proverbe-support : *Beauty is in the eyes of the beholder*][178] (1993 : 334-335).

Dans le deuxième cas, le proverbe est réduit à une collocation ou, plus génériquement, à une cooccurrence lexicale qui maintient de manière plus évidente (ou formellement) son lien sémantique avec le proverbe-support et qui se prête à tout réemploi discursif.
Ces processus d'érosion formulaire par l'usage déstabilisent la combinatoire initiale et longue du proverbe et ouvrent le chemin à une combinatoire nouvelle et brève. Pour les cadres grammaticalisés, les proverbes sont reconstruits du point de vue formel et subissent une extension sémantique, et ce, dans la mesure où le cadre prédicatif du proverbe-support le permet. Pour les bribes lexicales, on observe plutôt un ajustement dans de nouveaux cadres syntaxiques qui, à leur tour, pourront réutiliser et étendre la sémantique originaire du proverbe-support dont les bribes lexicales sont porteuses.
On pourrait en conclure que ces cadres grammaticalisés et ces bribes lexicalisées à partir de proverbes-supports constituent le nœud d'échange de la (tant évoquée et débattue) fixité proverbiale à la mobilité/variation phraséologique. C'est l'opérationnalisation informatique de quelques-uns de ces cadres et bribes que nous essayerons de mettre en place, et ce, moyennant une description lexico-grammaticale préalable (§ 5) basée sur notre corpus parémique (§ 3.3.1. et § 5).

Un dernier élément sur lequel Arnaud et Moon se penchent (*a posteriori*) et qui pourrait intéresser une détection automatique des proverbes sur corpus concerne la

[178] C'est nous qui indiquons les proverbes-support d'après les listes présentées de proverbes dans leur article.

signalisation des occurrences proverbiales. Comme d'autres parémiologues l'ont remarqué, les occurrences des proverbes peuvent faire l'objet d'une introduction en discours par une désignation explicite par de simples étiquettes nominales (ex. le mot *proverb*, *proverbe*) ou par des formules qui soulignent une attribution à un savoir partagé ou à un énonciateur inconnu (ex. *on dit*, *tout le monde sait que*) (1993 : 339).

4.2.1.2. Mieder (1994)

Très excité par la consultation d'une base de données riche telle *LEXIS/NEXIS* (§ 3.2.2.), Mieder prend son temps pour détailler la syntaxe de recherche à employer pour codifier le proverbe :

The apple doesn't fall far from the tree.

Suivant la syntaxe de recherche de la base de données, il crée l'expression régulière :

« *apple /within 3 words of/ fall /within 3 words of/ tree* » (1995 [1993] : 243).

Il s'agit d'un patron construit autour de mots-clés lexicaux du proverbe – et qu'il appelle « the triad » (*ibid*.) – ainsi que de ses variantes paradigmatiques (à savoir des synonymes) et de ses positions dans des fenêtres de recherche.

Il est encore intéressant de constater que ce patron lexical ressemble au cliché proverbial d'Arnaud & Moon (§ 4.2.1.2.) par le respect de la syntagmatique du proverbe. Pourtant, à la différence de ce dernier, le patron de Mieder décrit seulement les unités lexicales non grammaticales du proverbe, alors que les unités grammaticales sont (implicitement) soumises à la variation paradigmatique.

Comme nous l'avons précisé au § 3.2.2., malgré l'encadrement lexical, cette technique d'interrogation renvoie des résultats non pertinents et comporte du bruit : 123 occurrences (sur 232) ne concernent pas le proverbe recherché (*ibid*.).

4.2.1.3. Lau (1996)

Comme Mieder, Lau interroge la base de données *LEXIS*/*NEXIS* par un moteur de recherche de manière assez analogue à ce qu'on pourrait envisager avec un concordancier ou tout autre logiciel de traitement des données textuelles. Lau prend son temps pour décrire ses techniques d'interrogation. Elle est très claire :

> « Each proverb was searched as a complete phrase » (2003 [1996] : 233).

La forme canonique repérée dans les quatre recueils qu'elle a sélectionnés (§ 3.2.4.) fait l'objet de ses requêtes informatiques. Il en va de même dans les cas de deux ou plusieurs variantes morphosyntaxiques, toujours attestées dans les recueils :

> « In cases where there where distinct variants, both were searched and the results were combined (e.g. "All that glitters is not gold" and "All that glisters is not gold") » (*ibid*.).

Lau souligne que le moteur de recherche de la base *LEXIS*/*NEXIS* l'a empêché de saisir certains mots à haute fréquence ('*and*', '*be*', '*is*'). Ce qui n'est allé pas sans conséquence pour l'estimation de la fréquence de certains proverbes (§ 3.2.4.).

4.2.1.4. Moon (1998)

Moon attribue la responsabilité du non-repérage d'expressions figées en discours tout-venant aux contraintes des outils de traitement automatique de corpus (1998 : 48). Elle pointe aussi le doigt contre les restrictions que soulève l'interrogation d'un corpus :

> « Searches are deterministic, and only report what has been sought, not what should or could have been looked for » (1998 : 49).

Autrement dit, il est nécessaire de savoir lire d'avance entre les lignes d'un corpus, que cette lecture soit guidée par des prémisses linguistiques théoriques et/ou supportée par l'observation empirique. L'intuition comme précipité d'un acquis (théorique et empirique) de

la langue et du fait linguistique qu'on essaie d'extraire du corpus, aide ainsi à dessiner les requêtes informatiques les plus satisfaisantes en termes de précision et de rappel :

> « however much corpora provide data and strong evidence which can prove or disprove intuition, intuition is also necessary or variations will not be found » (1998 : 49).

Il est certes compliqué de se pencher sur les processus de variation et sur tous les détournements possibles des proverbes (comme de toutes les autres expressions figées et idiomatiques et des parémies) par la composition d'expressions régulières détaillées qui prévoient, dans une fenêtre donnée de mots graphiques :

- des formes lemmatisées ;
- des contraintes syntagmatiques (catégories grammaticales) sur les cooccurrents ;
- des contraintes paradigmatiques, c'est-à-dire des ensembles de mots apparentés par des liens sémantiques et/ou ontologiques[179] (1998 : 49-51).

Moon admet aussi à diverses reprises qu'un certain degré de généricité lexicale et syntaxique, ainsi qu'un certain degré de liberté de la contrainte linéaire permettent d'atteindre un résultat aussi correct que des requêtes formellement précises (1998 : 49).

Du potentiel des requêtes 'assez génériques' il en suit, en outre, le repérage de variations syntagmatiques qui témoigne d'une réutilisation créative de l'expression recherchée par rapport à son emploi conventionnel :

> « More loosely defined queries generally proved better for finding syntagmatic variations […] such searches yielded strong evidence of other structures or uses, resulting in redefinition […] and loss of lexicalized or coded status » (1998 : 50).

Malheureusement, Moon ne rentre ni dans les détails des intuitions pour repérer les variations, les permutations et les détournements des proverbes (et des autres expressions

[179] Moon cite, à titre d'exemple, des ensembles lexicaux (*lexical sets*) qui contiennent des différents types de boissons, ou d'aliments ou de vêtements (1998 : 50), sans mieux éclairer le contenu de ces ensembles et comment ils ont été conçus. À son avis, ces ensembles lexicaux facilitent le repérage de variations thématiques (*thematically bound variations*) (*ibid.*).

figées et idiomatiques) ni dans la mesure de la généricité et de la spécificité des requêtes informatiques à modéliser.

D'ailleurs, pour rendre compte de la difficulté que la recherche des proverbes entraîne, elle donne l'exemple du proverbe :

> *a bird in the hand is worth two in the bush*
> [trad. litt. 'un oiseau dans la main en vaut deux dans un buisson] (1998 : 51).

Lorsqu'elle interroge son corpus – sans détailler ses techniques, elle tombe sur l'occurrence du nom d'un pub anglais appelé :

> *The Bird in Hand*
> [trad. litt. L'Oiseau en Main].

Tout en délaissant l'humour, on pourrait en conclure qu'à ce stade, sa requête est pour l'essentiel de matrice lexicale, à savoir elle vise une cooccurrence de deux mots-clés lexicaux dans une fenêtre donnée de mots graphiques.

D'après ce qui est encore donné par Moon comme exemple, on pourrait faire l'hypothèse que le détournement suivant :

> « **A sheep in the pen is worth two in the field** could be our adaptation of the popular saw. (OHPC : journalism) » (*ibid.*)[180]

en s'appuyant encore sur la composante lexicale du proverbe. Nous supposons qu'elle peut avoir recherché des variantes paradigmatiques grâce à l'ensemble lexical 'animal' (*bird*/*sheep*), grâce à l'ensemble lexical 'morceaux de terre' (*bush*/*field*) ou peut-être grâce à l'ensemble lexical 'objets qu'on tient dans ses mains' (*hand*/*pen*)[181]. L'insistance d'une requête informatique guidée par le lexique a certainement favorisé le repérage d'un détournement qui aurait diversement été « easy to miss » (*ibid.*) et témoigne de la nécessité de bien concevoir ces ensembles lexicaux qui nous rappellent, d'ailleurs, les *classes d'objets* (G. Gross & Clas 1997 ; Le Pesant & Mathieu-Colas 1998). Pourtant, ce déterminisme lexical

[180] « Une brebis dans le stylo en vaut deux dans le champ pourrait être notre adaptation de la maxime populaire » (c'est nous qui traduisons).
[181] Probable, mais sans aucun résultat, pourrait être la recherche à l'aide d'un ensemble lexical 'chiffre' (*two*).

passe sous silence une évidence que ce même exemple éclaire sans aucun doute : l'invariance (ou, du moins, la stabilité) syntaxique. L'élaboration des requêtes informatiques donne l'impression d'ignorer que la syntaxe déclenche ou, en tout cas, participe au processus de cosélection du lexique. À cet égard, la description lexico-grammaticale d'au moins une partie du répertoire parémiographique anglais aurait mis en relief que *to be worth* contribue à 'créer d'une certaine façon' des proverbes, et ce, avant même l'interrogation du corpus[182]. Si on ne pouvait le prévoir – malgré l'étude de Sinclair & Renouf (1991) sur les *cadres collocationnels* (§ 1.1.4.), la consultation du corpus aurait offert une piste pour relancer une nouvelle recherche et pour élaborer des nouvelles requêtes lexico-grammaticales.

Mise de côté la syntaxe, la concordance du détournement proverbial nous signale un autre indice – cette fois-ci lexical – qui aurait facilité son repérage. Nous faisons référence au groupe nominal *popular saw* (proverbe, maxime populaire), c'est-à-dire une formule d'introduction. La recherche de ce groupe nominal et d'autres groupes similaires aurait permis de dégager des détournements et d'autres variantes proverbiales par un menu effort[183].

4.2.1.5. Corpas Pastor (1998)

L'optimisme quantitatif de Corpas Pastor à l'égard des fréquences repérées pour les parémies espagnoles (§ 3.2.6.) pourrait se fonder sur l'hypothèse que certaines formes non canoniques des parémies sont passées sous silence. Comme pour Moon (§ 4.2.1.5.), ce manque de détection est reconduit aux outils informatiques, même si Corpas Pastor fait confiance à son logiciel *WordCruncher*. À propos des techniques d'interrogation, elle donne quelques indices, notamment dans la conclusion de son étude, où elle précise que :

> « ha sido necesario extraer concordancias para los componentes aislados de cada unidad » (2003 [1998] : 98).

Ce qui nous fait croire que la recherche des parémies espagnoles a porté sur la facette lexicale.

Corpas Pastor précise encore que :

[182] Nous tenons à préciser que Moon présente l'analyse lexico-grammaticale de toutes les expressions figées et idiomatiques (1998 : 75-119) et de leurs modifications et variantes (1998 : 120-177), mais elles sont données en aval de son interrogation du corpus. Norrick (1985), cité d'ailleurs par Moon même, aurait offert quelques suggestions syntaxiques éventuellement à confirmer et/ou à réfuter.

[183] On précise que cette technique de recherche n'est pas mentionnée. Il n'est donc pas exclu *a priori* que Moon l'ait utilisée.

« la localización de las paremias en el corpus se ve dificultada por problemas de lematización » (2003 [1998] : 92-94).

La lemmatisation joue un rôle fondamental pour la détection automatique sur corpus (§ 6.3.) et Corpas Pastor pointe le doigt sur son impossibilité à juste titre.

Elle recense aussi les différents processus de modification des parémies (2003 [1998] : 92-94) et remarque que le processus le plus rencontré dans le CVB est la réduction, soit par effacement de la partie finale[184] en guise d'allusion, soit par le maintien d'un groupe syntaxique qui facilite son insertion et réemploi dans un discours suivi. Quant à ce processus de réduction, elle précise que :

« En algunos casos, tales usos llegan a institucionalizarse en la lengua, dando lugar al nacimiento de una nueva unidad » (2003 [1998] : 92).

Corpas Pastor donne l'exemple de :

A enemigo que huye, puente de plata
[trad. litt. : À ennemi qui s'enfuit, pont d'argent][185].

Cette parémie compte 3 occurrences dans son corpus, dont 1 en tant que groupe nominal (GN) isolé (*puente de plata*) et 2 autres où ce GN participe avec un verbe support (V_{sup}) à la création de :

poner puente de plata a alguien o a algo
[trad. litt. : mettre pont d'argent à quelqu'un ou à quelque chose] (2003 [1998] : 92).

Ainsi, des locutions nominales ou verbales de matrice parémique se phraséologisent pour s'accoutumer à la combinatoire brève propre à d'autres expressions idiomatiques. C'est à ce propos et en perspective qu'elle ajoute que :

[184] Comme Corpas Pastor, nous préférons ne pas parler de 'deuxième partie' d'une parémie pour éviter le rapprochement systématique (et parfois non pertinent) au schéma binaire propositionnel des proverbes.
[185] Voir le traitement réservé à cette même parémie par Navarro Brotons (§ 3.2.15.).

« la búsqueda automática en el corpus deberá atender tanto a la forma canónica como al núcleo fraseológico residual » (*ibid*.).

Au-delà de la forme canonique proposée par les recueils parémiographiques, la détection des parémies sur corpus ne peut se passer des variations, notamment de ce « *núcleo fraseológico residual* » (*noyau phraséologique résiduel*) dont Corpas Pastor ne donne pas de définition ou d'explication autres au fait que ce noyau « permanece constante en todos los casos » (2003 [1998] : 94). Corpas Pastor donne encore l'exemple des 6 occurrences repérées de :

A río revuelto, ganancia de pescadores
[trad. litt. : À fleuve troublé, gain de pêcheurs]

dont une seule est à la forme canonique alors que les 5 restantes présentent des modifications diverses (2003 [1998] : 94-95). Ce « noyau phraséologique résiduel » favorise la « manipulación creativa de estas unidades en el discurso » ainsi que la création de néologismes phraséologiques (2003 [1998] : 98). Cette observation s'appuie sur la lecture des concordances qui montrent qu'autour de la collocation *río revuelto* s'établissent des exploitations de ce proverbe. Le noyau *río revuelto* cosélectionne des unités lexicales qui représentent des dérivations morphosyntaxiques et des variantes flexionnelles de l'unité lexicale *pescadores* (*pescaban*, *pescando*, *pescar*) dans des séquences avec un verbe (V) et éloignées donc de la forme canonique proposée. Corpas Pastor délaisse, toutefois, les motivations qui permettent de 'créer en discours' autour de la constante lexicale *río revuelto*. Elle survole aussi pourquoi ce proverbe garde à la fois son statut phraséologique et son statut idiomatique, vu le démantèlement de sa forme de départ. Nous pouvons faire l'hypothèse que ce noyau peut concerner la partie initiale de la parémie qui garde et transmet ses informations lexico-grammaticales et co(n)textuelles canoniques (voir prototypiques)[186]. Il ne s'agit que d'une hypothèse, certes, très partielle et qui ne peut rendre compte de tous les « noyaux phraséologiques résiduels ».

D'après sa recherche sur corpus, Corpas Pastor a ainsi perçu des fondements linguistiques minimaux qui font que la parémie, soit-elle morphosyntaxiquement adaptée ou détournée, garde son statut d'unité phraséologique, d'une part, et son statut idiomatique d'origine, d'autre part. À ce propos, on est déjà en mesure de percevoir un rapprochement

[186] En ce sens, ce maintien lexical peut éventuellement être interprété en faisant recours à la théorie du *priming* lexical défendue par Hoey (2005).

entre parémies anglaises, espagnoles et françaises. L'observation de Corpas Pastor nous rappelle en effet les constats similaires sur le *cliché proverbial* par Arnaud & Moon (1993) (§ 4.2.1.1.), mais aussi le *noyau proverbial* observé par Schulze-Busacker (§ 4.1.3.). Il y a une constante interlinguistique et diachronique : la permanence de certaines unités lexicales cooccurrentes, continues ou discontinues.

Pour finir, on souligne que nulle part dans son étude, Corpas Pastor fait référence à la syntaxe pour ses techniques de recherche. Malgré les cas où elle constate que certaines modifications par réduction entraînent des modifications grammaticales, la syntaxe est ignorée comme heuristique de recherche et reléguée en arrière-plan par rapport au lexique.

4.2.1.6. Conenna (1998b, 2000c)

L'approche par expressions régulières marque également les premières expériences de reconnaissance automatique des proverbes français dans un corpus par Conenna. Le binôme solide entre la méthode du Lexique-Grammaire (M. Gross 1975) et le logiciel INTEX par Max Silberztein[187] sert de support pour repérer des occurrences des proverbes.

D'abord, elle recherche les occurrences de :

proverbe

dans le corpus *Le Monde*, notamment pour les années 1992-1994 et qu'elle saisit comme expression régulière. Conenna repère ainsi : 59 occurrences pour l'année 1992 ; 79 pour 1993 ; 49 pour 1994 (1998b : 100).

Par la suite, elle s'intéresse aux occurrences des proverbes qui appartiennent à la classe d'équivalence syntaxique *Qui*. C'est par une expression régulière, notamment le patron :

<^> *Qui*

que Conenna commence sa détection dans le corpus du quotidien *Le Monde* (1998b : 103). Dans la syntaxe des expressions régulières, le symbole ^ sert à contraindre le début d'une suite de caractères recherchée : en l'occurrence, le pronom *Qui* après un point. Il est évident

[187] Silberztein a récemment mis au point le logiciel *Nooj* (Silberztein 2003). La plupart des fonctions de *Nooj* reviennent aux expériences et aux programmes d'INTEX.

que ce patron très générique produit du bruit dans les résultats. Plus précisément : pour l'année 1992, sur un total de 1.342 occurrences repérées, la désambiguïsation manuelle pour l'effacement de toutes les formes interrogatives ramène le chiffre à 406 (1998b : 104) ; pour 1993, de 1.176 occurrences, on en obtient 869 (1998b : 103) ; pour 1994, on passe de 940 à 256 occurrences (1998b : 105). Outre l'ambigüité soulevée par la généricité du patron, nous soulignons que la contrainte positionnelle du pronom (après un point) empêche au logiciel de repérer les occurrences du pronom en d'autres positions, comme le montre l'Annexe 5 de la contribution. Ce qui va au détriment d'autres occurrences des proverbes de la classe *Qui* dans le corps du texte. La contrainte positionnelle disparaît dans l'étude de 2000 où la recherche se limite au patron :

il faut

qui renvoie, pourtant, 8.773 occurrences à désambiguïser, et ce, rien que pour l'année 1994 du même corpus (2000c : 289).

Pour éviter tout autre passage de désambiguïsation manuelle, Conenna exploite, déjà en 1998, une autre fonction d'INTEX : l'application d'un dictionnaire électronique (1998b : 103) qui suit le formalisme DELA (Courtois 1994-1995). Les tables du lexique-grammaire des proverbes français de la classe *Qui* (Conenna 1988) représentent son point de départ pour établir le dictionnaire électronique *ProverbesQui*[188] :

Qui a de l'argent a des pirouettes,. PROVERBE
Qui sème le vent récolte la tempête,. PROVERBE

où chacune de ses 200 entrées correspond à une seule ligne pour forme canonique, suivie de la signalisation de son statut linguistique (*PROVERBE*) (1998b : 102). Il en va de même pour l'étude de 2000, sauf que le point de départ pour le dictionnaire *ProverbesIlfaut* est une liste à plat de 240 proverbes (2000c : 289). Par une telle approche 'bloquée' sur une seule surface, des proverbes passent sous silence et ne sont pas repérés. Dans ce cas, c'est la lecture des concordances qui favorise la reconnaissance 'à vue' d'autres proverbes (1998b : 104), de variantes morphosyntaxiques et lexicales et des proverbes dont la linéarité est cassée pour d'autres insertions (1998b : 105 ; 2000c : 289-291) ou pour des allusions et des troncations

[188] Ce dictionnaire sera complété par des automates conséquents en (2000b) (§ 4.2.2.1.).

des proverbes mêmes (1998b : 107 ; 2000c : 289). Il nous faut donc mettre en évidence que les dictionnaires électroniques (surtout ces dictionnaires conçus pour la reconnaissance et pour l'extraction d'information) devraient prévoir non seulement la codification d'une norme parémiographique choisie de manière (plus ou moins) arbitraire, mais aussi la codification (au moins) de quelques variantes ou exploitations de cette norme. À ce propos, c'est l'usage observé (beaucoup plus que l'introspection) qui peut contribuer à l'accomplissement de cette tâche[189].

L'étude de 1998b offre encore d'autres suggestions syntaxiques pour des opérationnalisations informatiques ainsi que pour des réflexions parémiologiques. La lecture des concordances emmène Conenna à constater que :

> « […] un nombre considérable de phrases [a] la même structure des proverbes : QUI V_1 N_1 V_2 N_2 » (1998b : 106).

Même si l'on ne dispose pas du chiffre exact de cette observation des données linguistiques, la syntaxe de cette classe n'est pas vraiment loin de celle d'autres séquences. C'est un aspect qui sert à faire sortir le proverbe du lot de l'exclusivité, du moins en ce qui concerne sa surface[190]. En même temps, c'est une observation qui permet d'envisager d'autres traitements linguistiques qui pourraient rapprocher les proverbes d'autres séquences formulaires (Marcon 2011, 2013).

Conenna continue et affirme que :

> « la structure QUI V_1 V_2 est un indicateur de citation proverbiale [et qu'elle] est donc productive, surtout pour ce qui concerne la fabrication des proverbes de nos jours, autrement dit, les slogans » (*ibid*.).

Une combinatoire lexico-syntaxique (qu'elle nommera par la suite *moule syntaxique* (§ 1.1.4.)) où seulement l'unité initiale est lexicalement remplie, alors que le restant est contraint par une cooccurrence de groupes syntaxiques (en l'occurrence, deux groupes

[189] En ce sens, nous rejoignons l'approche théorique et méthodologique lexicale récente proposée par Hanks dans sa Théorie des Normes et des Exploitations qui suggère un nouvel enregistrement lexicographique de normes (ou tendances prototypiques) et d'exploitations (ou tendances préférentielles) pour toute unité lexicale, et ce, sur la seule base de l'usage observé dans un corpus.

[190] Pour cette raison, Conenna suggèrera l'étiquette de *parémiologie linguistique* pour expliciter les apports de toute analyse linguistique du proverbe (§ 2.1.).

verbaux), permet de reconnaître des *expressions*[191] proverbiales et non proverbiales. Nous observons tout de suite que nous nous trouvons vis-à-vis d'une facette (pour ainsi dire) complémentaire du cliché proverbial observé par Arnaud & Moon (§ 4.2.1.1.). Le *cliché proverbial* de ces deux chercheurs est aussi une combinatoire lexico-syntaxique. Pourtant, dans ce cas de figure, seulement une unité (peu importe sa position) est remplacée par d'autres paradigmes lexicaux alors que le restant est rempli autant du point de vue lexical que du point de vue syntaxique. Dans les deux cas, on assiste à la réutilisation de ces cooccurrences pour engendrer de nouvelles combinatoires lexico-syntaxiques dont la matrice est d'origine proverbiale. Qu'elles détournent un proverbe ou qu'elles forgent de nouvelles séquences, ces deux facettes de la cooccurrence lexico-syntaxique de matrice proverbiale confirment notre propos d'envisager d'autres approches pour aborder les proverbes, notamment celles qui sont employées pour d'autres séquences formulaires. Ainsi, nous croyons que ces cooccurrences lexico-syntaxiques de matrice proverbiale peuvent (nous dirions même qu'elles doivent) être modélisées sous forme de requêtes informatiques. D'une part, cela peut supporter la reconnaissance du nombre le plus élevé d'occurrences proprement proverbiales, et ce, en vue d'une étude de fréquence. D'autre part, leur modélisation permettra une détection plus aisée de toute autre cooccurrence lexico-syntaxique qui pourrait être formellement (et aussi sémantiquement) apparentée avec les proverbes qui sont à l'origine des cooccurrences lexico-syntaxiques recherchées[192].

Il demeure essentiel, donc, de fournir une description lexico-syntaxique préalable, comme le fait Conenna dans le cadre du Lexique-Grammaire. Nous soulignons, pourtant, que la généricité des moules syntaxiques tels qu'ils sont proposés par elle, et que la modélisation de la facette syntaxique à elle seule produisent trop de bruit dans les résultats (Marcon 2011, 2013). D'où la nécessité d'une description qui prend en compte l'interface lexique-syntaxe de manière systématique, mais aussi simultanée (§ 5).

[191] C'est le terme délibérément générique employé par Conenna.

[192] Outre à évaluer la synchronie et la microdiachronie, comme sera le cas de notre recherche, ces requêtes informatiques pourraient mettre en évidence les évolutions lexico-syntaxiques des proverbes et, en général, d'autres séquences formulaires dans des corpus diachroniques. Elles fourniront une preuve ultérieure de ce qu'Anscombre a justement constaté : « Les proverbes actuels ne nous ont pas été transmis tels quels, et ont été constamment réactualisés » (Anscombre 1994a : 96). Ces requêtes pourront aussi servir de support à l'évaluation de l'histoire des proverbes, notamment pour des améliorations de recueils parémiographiques à visée philologique, comme *DicAuPro* (Conenna *et al.* 2006).

4.2.1.7. Järv (1999)

Järv contourne les proverbes dans les corpus exclusivement par le recours aux marqueurs discursifs qui signalent la volonté du locuteur d'introduire un proverbe dans le discours. En ce sens, Järv recherche quatre mots-clés qui caractérisent la plupart de ces marqueurs estoniens :

> « *vanasõna* ('proverb'), *rahvatarkus* ('folk wisdom'), *kõnekäänd* ('proverbial phrase'), *vanarahvas* ('old folk') » (1999 : 81).

Il est évident que cette technique simple – et simpliste – de détection nuit à une étude de fréquence des proverbes estoniens. Il n'est pas nécessaire qu'un locuteur mentionne un de ces marqueurs pour introduire un proverbe dans un discours. De plus, certains locuteurs peuvent volontairement signaler des proverbes qui ne sont guère des proverbes (Marcon 2012). Ce qui comporte un jugement de proverbialité de la part du parémiologue (Järv 1999 : 82-83). Pourtant, Järv lui-même avoue que limiter la détection aux marqueurs ne constitue pas une méthode achevée et que :

> « comparison of the newspaper material with each of the 15.000 types of proverbs identified in the volume *Eesti vanasõnad* would be more resultative » (1999 : 82).

Järv sait donc qu'une liste de départ aurait garanti des résultats plus performants. Il évite, pourtant, l'entreprise parce qu'elle représente « too massive amount of work » (*ibid*.), surtout pour la visée qualitative de sa recherche.

4.2.1.8. Cignoni & Coffey (2000)

Dans le cadre d'une étude sur corpus des proverbes italiens, Cignoni & Coffey (2000) adoptent les *keywords* (mots-clés) pour leur détection. Eux aussi, comme d'autres, n'en donnent pas de définition. De leur contribution, nous comprenons qu'ils recherchent des cooccurrences de 2 à 3 mots-clés, sans contraindre l'ordre linéaire dans l'expression régulière. À ce propos, ils commentent :

« Search flexibility of this type allowed us to retrieve many variations [...] » (2000 :550).

4.2.1.9. Maniez (2000)[193]

L'étude de Maniez embrasse un défi stimulant : le repérage automatique des défigements (ou des *palimpsestes verbaux* « peu remarqués… », d'après Galisson (1994)) dans la presse étatsunienne. En s'appuyant sur une liste de 10.532 titres collectés et analysés manuellement (2000 : 20) dont 171 font référence à des proverbes anglais (2000 : 21), l'auteur s'interroge sur la manière d'automatiser leur reconnaissance dans des corpus.

C'est par l'analyse préalable de ces données qu'il peut dégager des régularités linguistiques sur lesquelles axer ses expressions régulières. Outre l'approche lexico-centrique fondée sur «les cooccurrences de deux ou trois mots » (2000 : 27) parce qu'elle produit moins de bruit que la recherche d'un seul mot (2000 : 28), il réfléchit aussi sur les « structures grammaticales » (2000 : 27) couramment employées dans les titres de presse, mais non pas dans les proverbes.

Néanmoins, l'originalité de Maniez réside dans deux aspects novateurs : (i) la structuration de 800 proverbes anglais sous la forme d'une base de données (2000 : 29) et (ii) la rédaction de programmes informatiques pour le balayage simultané de la liste des titres et de la base de données (2000 : 29-30). La détection des allusions proverbiales passe ainsi par une modélisation algorithmique réfléchie (sur base empirique et statistique) portant à la fois sur la recherche croisée de mots simples et de séquences de mots. Par cette hybridation, son programme apparie correctement 128 sur 174 titres défigés (2000 : 30). Malgré cette performance satisfaisante, le chercheur conclut que le repérage des allusions nécessite un travail fondamental de formalisation des connaissances, surtout sémantiques, tout comme des « caractéristiques lexicales et grammaticales des titres de la presse » (2000 : 31). Nous nous permettons d'ajouter qu'un travail pareil doit également intéresser les sources à l'origine de ces allusions, comme justement les proverbes.

[193] Nous remercions François Maniez pour sa disponibilité et pour avoir partagé avec nous les avancements récents entrepris en la matière avec Pierre Arnaud. Nous remercions aussi la direction de la *Revue française de linguistique appliquée* qui nous a généreusement offert le numéro intégral du volume où est reprise l'étude de Maniez que nous présentons dans ce paragraphe.

4.2.1.10. Anderson (2006)

Pour la détection des locutions et des proverbes, Anderson fait recours à *WordSmith Tools 3*, notamment au programme *Concord* (le véritable concordancier de la suite). Comme elle est consciente des variantes formelles, outre la simple recherche des formes canoniques, Anderson envisage une technique de recherche par l'utilisation de :

> « word forms which were most essential to the phrase » (*ibid*.)

pour après :

> « manually […] discard concordance lines which were not in fact instances of the phrase sought » (2006 : 129-130).

Elle privilégie cette technique à la création de requêtes informatiques plus longues et détaillées pour des soucis de précision, mais surtout de silence : elle ne veut pas négliger les occurrences de certaines variantes ou actualisations (2006 : 130). Cette technique de détection semi-automatique ne clarifie pas, pourtant, de quelle façon identifier les mots 'les plus essentiels' (*most essential*) pour élaborer ces requêtes informatiques. Anderson ne donne pas non plus des exemples à cet égard pour en déduire une logique de création.

Anderson fait également recours à la troncation et à la recherche de plusieurs formes fléchies pour les seuls cas des verbes (2006 : 129).

Autre attention typographique et la seule (explicitée) dans notre revue : elle saisit des expressions régulières, en l'occurrence, sans ou avec accent éventuel pour éviter les traitements variés qui concernent surtout les mots en lettres capitales (*ibid*.).

Malgré toutes ces attentions, Anderson veut souligner qu'il se peut que certaines locutions soient restées 'méconnues' dans les corpus parce que ses requêtes n'ont pas été en mesure de reconnaître plus de 5 insertions d'autres unités lexicales entre les « *constituent parts* » (2006 : 130) des expressions recherchées. Ce qui nous étonne, toutefois, est la deuxième raison de non-identification des locutions. Anderson dit que ce silence présupposé peut s'expliquer :

> « because [these expressions] involve some creative manipulation of the dictionary form of the phrase with the result that only the grammatical structure, below the surface, remains constant » (*ibid*.).

Il y a une prise de conscience explicite sur le fait que les locutions et les proverbes subissent des mutations qui laissent leur armature syntaxique inchangée, et ce, malgré les variations qui intéressent leur surface. Tout en étant conscient de la lourdeur que la description syntaxique de 11.647 locutions comporterait, il est étonnant qu'Anderson n'ait pas fait une telle expérience pour un échantillon de sa liste et de son corpus, éventuellement annoté morphosyntaxiquement.

4.2.1.11. Čermák (1998, 2006)

Vers la fin des années 1990, Čermák constate le manque de requêtes et d'outils informatiques pour la recherche de proverbes tchèques :

> « there is as yet no discovery procedure or even software programme available enabling one to find proverbs in a corpus » (2007 [1998] : 569).

Après la vulgarisation des concordanciers et l'interrogation des corpus par des expressions régulières, Čermák se questionne sur comment élaborer une requête informatique pour la reconnaissance de proverbes. Il observe que la recherche restreinte à la première partie du proverbe ne donne pas des résultats performants (2007 [2006] : 538). Il ajoute que :

> « The search method […] is based on a **prominent part** of the proverb (chosen ad hoc), though it may be difficult to define it in general and say what it exactly consists in » (*ibid*.).

La *partie proéminente* ressemble aux mots-clés des études de Cignoni & Coffey (§ 4.2.1.8.) ou au *noyau phraséologique résiduel* de Corpas Pastor (§ 4.2.1.5.), quoiqu'on ne dispose pas vraiment de sa définition par Čermák (ainsi que par les autres, d'ailleurs).

4.2.1.12. Ďurčo (2006)

On ne connaît pas exactement toutes les techniques d'interrogation du corpus que Ďurčo a employées pour le repérage des proverbes allemands sur corpus. Malgré cela, on apprend de Zouogbo (2011 : 104) que Ďurčo a commencé à mettre au point des patrons formulaires complexes pour la détection automatique de proverbes sur corpus (*Complex Pattern Matching*) à l'aide d'expressions régulières. Zouogbo commente ce choix comme suit :

> « […] puisque les corpus électroniques ne possèdent pas d'annotations spécifiques pour les proverbes, établir une série de modèles à partir de la typologie des moules morphologiques de tous les proverbes peut les y aider » (*ibid*.).

4.2.1.13. Gómez-Jordana Ferary (2006)

À aucun endroit de sa thèse, Gómez-Jordana Ferary traite les aspects qui concernent les techniques d'interrogation de ses corpus. Nous estimons, d'ailleurs, que le nombre trop bas (environ 800 occurrences) d'occurrences pourrait dépendre d'une réflexion peu développée sur ce volet.

En tout cas, d'après la lecture de quelques contextes d'occurrence qu'elle a repérés, il nous est possible de faire quelques hypothèses sur ses techniques de recherche. Ces contextes présentent des exemples de proverbes détournés, et ce, dès la première page des Annexes (2006 : 130) dans la section consacrée au corpus français. Par exemple :

> « Chez le pape lazuli, **l'habit ne fait pas le séducteur** » - titre d'un article sur un type d'oiseau et son plumage et comment ils séduisent les femelles ; *Le Monde* 29 oct. 2000, p. 23

Parfois des contextes montrent des proverbes qui ne sont même pas mentionnés dans la liste qu'elle a établie, comme dans les deux cas ci-dessous, toujours dans les Annexes de sa thèse (2006 : 130-131) :

> « « -Journaliste : Avez-vous une devise ? - Bernard Lavilliers (chanteur) : **Mieux vaut jeuner avec les aigles que picorer avec les poules** ! Je dis que c'est un proverbe indien, mais c'est

archi-faux. J'ai une énorme collection de faux proverbes chinois, indiens, juifs… » ; entretien avec le chanteur Lavilliers dans *Paris Match*, 21 décembre 2000

« Au Laos où il faut savoir prendre son temps, la nonchalance est une vertu. Et nul ne voudrait contredire le proverbe : « **Le Vietnamien plante le riz, le Cambodgien le coupe, le Laotien l'écoute pousser** ! » ; (Reportage de deux pages sur le Laos qui finit avec ce proverbe). *Femme Actuell*e 847; du 18 au 24 décembre 2000, p. 97

Outre à constater que certaines sources sont absentes parmi celles qui sont mentionnées comme corpus, notamment les magazines *Paris Match* et *Femme Actuelle* (§ 3.2.11.), ces contextes nous suggèrent que la recherche des proverbes n'a pas intéressé la forme canonique ou une variante. Le premier et le deuxième exemple nous font supposer que Gómez-Jordana Ferary aurait saisi certains mots-clés et/ou ciblé la partie initiale des proverbes, comme le schéma prédicatif incomplet :

l'habit ne fait pas

ou la forme impersonnelle :

Mieux vaut.

Cette dernière appartient aussi à la liste de moules proverbiaux qu'elle identifie dans sa liste de départ (2012 : 128). En guise de confirmation de cette technique d'interrogation, nous pouvons également mentionner l'exemple du contexte suivant :

« Pour oser, celui qui depuis 1981, a fait de la météo un spectacle, voire un divertissement – et fait école en la matière – a toujours osé. Au point de s'égarer, en allant, sur feue La Cinq, animer un jeu « d'une débilité profonde », selon Le Monde de l'époque, ou en rédigeant dans Libération une chronique d'un genre souvent douteux, achevé sur cet aveu : « Il faut savoir s'arrêter avant de faire la merde ». **Faute avouée…** Mais d'où vient cette audace et paradoxalement, cette « réserve qu'il revendique ? » ; *Le Monde*, 5 avril 1999, p.6, « Dans l'œil de M. Cyclone », Cornu, F.

à savoir un exemple de troncation du proverbe de sa liste :

Faute avouée est à demi pardonnée.

En revanche, le troisième exemple ne peut s'expliquer que par le recours aux introducteurs des proverbes en discours. En l'occurrence, il s'agirait tout simplement de la suite de caractères :

proverbe.

Le proverbe (apparemment) laotien cité n'est pas dans la liste des proverbes contemporains de Gómez-Jordana Ferary. Seulement une recherche des candidats-voisins du proverbe aurait permis son repérage.

4.2.1.14. Hrisztova-Gotthardt & Gotthardt (2011)

Pour leur étude de fréquence des parémies bulgares sur le Web à l'aide des moteurs de recherche *Bing* et *Google*, Hrisztova-Gotthardt & Gotthradt entreprennent la création d'un script PERL. Le script prend comme argument une liste tokenisée de 2.301 parémies et donne en sortie la même liste enrichie du nombre des résultats renvoyés par chaque moteur de recherche (2011 : 254). Comme au § 4.2.1.9., on mentionne ouvertement la programmation informatique pour repérer les parémies. Contrairement au § 4.2.1.9., l'expression régulière correspond à toute la séquence parémique. De plus, pour chaque séquence parémique, les auteurs distinguent la recherche d'un lemme parémique et de ses variantes (2011 : 256).

Malgré ces diversités majeures par rapport au panorama esquissé jusqu'à présent, il faut souligner que la recherche vise toujours l'ordre des constituants et leurs formes exactes. Ce qui sous-estime par conséquent le potentiel créatif des séquences parémiques.

4.2.1.15. Barani (2012)

Barani ne dispose pas d'un véritable concordancier, mais d'un moteur de recherche propre au CD-Rom de son corpus (§ 3.2.15.). Elle doit adapter ses techniques d'interrogation aux fonctions du moteur de recherche.

La première technique consiste en la recherche contrainte (avec guillemets) à la partie initiale de la parémie (2012 : 43). À ce propos, elle cite l'exemple de la requête :

« “*a buen entendedor*” »

pour repérer toutes les occurrences de :

« *a buen entendedor, pocas palabras* » [à bon entendeur, peu de paroles],

de ses (quelques[194]) variantes et détournements (*ibid*.).

La deuxième technique vise le contenu lexical des parémies, à savoir la création d'une requête informatique qui met en série certaines unités lexicales d'une parémie. Elle fait l'exemple de :

« “*predicar trigo*” » [trad. litt. : prêcher blé]

pour repérer toute occurrence où certains éléments formels de la parémie :

« *no es lo mismo predicar que dar trigo* » [trad. litt. : ce n'est pas la même chose de prêcher que de donner du blé]

et son sens formulaire sont actualisés (décomposés, recomposés et exploités) en discours (2012 : 42-43). On observe que l'auteure délaisse les critères de sélection de ces unités lexicales, mais nous percevons que Barani a exclu tous les mots grammaticaux (déterminants ou prépositions, par exemple) ainsi que les mots qui sont ressentis comme 'à haute fréquence' et donc peu discriminants en discours (comme le verbe *dar*).

Une troisième technique concerne la troncation de certaines unités lexicales pour le repérage de variantes morphosyntaxiques, outre que d'autres actualisations en discours. Cette technique prend en compte le schéma prédicatif (ou un des schémas prédicatifs) propre à une parémie. Pour la parémie :

« *muerto el perro, se acabó la rabia* » [trad. litt. : mort le chien, s'épuisa la rage],

Barani recherche le schéma prédicatif :

[194] C'est nous qui précisons. La variation et le détournement peuvent également concerner la partie de parémie contrainte au moment de l'interrogation.

« “*acab* la rabia*” »

afin de repérer la forme canonique ainsi que d’autres adaptations morphosyntaxiques et détournements en discours (2012 : 44).

Barani envisage aussi la possibilité de créer des requêtes informatiques plus génériques, mais elle s’y oppose parce qu’elles comportent un gaspillage de temps pour la désambiguïsation manuelle des résultats (2012 : 44, 46), à savoir pour le bruit que des requêtes trop génériques causent pour la reconnaissance des séquences formulaires (Marcon 2013)[195].

Malgré ces outils à sa disposition, Barani avoue quand même que la variation et le détournement sont des phénomènes difficiles à cerner, surtout en raison des limitations du moteur de recherche utilisé (2012 : 44). La forme exacte, d’une part, et la créativité linguistique (outre que technique) du chercheur, d’autre part, sont deux éléments qui doivent aller de pair lors de la modélisation informatique. En ce sens, Barani donne un exemple de cas extrême de détournement dont le repérage serait presque impossible. Par un rapprochement sémantique, elle conclut que la séquence formulaire :

A coño usado, televisión nueva [trad. litt. : À chatte usée, télévision nouvelle]

serait une exploitation de la parémie :

A rey muerto, rey puesto [trad. litt. : À roi mort, roi habillé] (2012 : 45).

Nous ne sommes pas en mesure d’avancer des arguments pour ou contre ce rapprochement en termes de sens formulaire, quoiqu’il nous paraît assez évident qu’un ‘remplacement dû à l’usure’, comme le suggère la première formule, ne correspond pas exactement à un ‘remplacement dû à une disparition’, comme le veut la parémie. De toute façon, cette constatation trahit plutôt une évidence (qui reste implicite dans la thèse de Barani), à savoir qu’il y a un rapprochement, voire un parallélisme, en ce qui concerne la cooccurrence des

[195] Nous partageons le sentiment de frustration éprouvée par l’auteure mais, en même temps, nous croyons qu’une généralisation des résultats obtenus (surtout quand ils impliquent une exploitation en termes pédagogiques) se doit d’être chronophage.

parties du discours qui suivent le pivot initial inchangé : la préposition *A*. C'est la description lexico-grammaticale :

A N ADJ, N ADJ

qui apparente formellement les deux exemples mentionnés. Ce qui nous confirme dans le propos qu'une requête informatique basée sur une description lexico-grammaticale préalable permettrait la détection autant de la parémie que (si l'on accepte l'interprétation de Barani) de son détournement[196] (§ 5).

4.2.1.16. Navarro Brotons (2013)

On ignore les techniques employées par Navarro Brotons pour l'interrogation du moteur de recherche *Google* : l'insertion de mots-clés et de symboles, la restriction à certains typologies de fichiers ou à certains adresses Internet ainsi que le respect (ou non-respect) d'une certaine syntaxe de recherche altèrent sensiblement le nombre d'attestations repérées, surtout de ces « combinaisons linguistiques […] de nature assurément complexe » (2012 : 453) que sont les parémies. Un seul indice nous est fourni par une capture d'écran où l'on visualise une page des résultats renvoyés par *Google* pour la parémie espagnole :

A buen hambre no hay pan duro (trad. litt. : À bonne faim il n'y a pas de pain dur) (2013 : 84).

La parémie est saisie entre guillemets pour rechercher les formes telles qu'elles ont été repérées dans les dictionnaires consultés. On en déduit que l'utilisation des guillemets, à savoir la contrainte à la forme exacte et norm(alis)ée des parémies, constitue sa seule syntaxe d'interrogation.

[196] Par souci de clarté, on précise que l'auteure n'aurait pas pu lancer une recherche lexico-grammaticale à cause des limitations du moteur de recherche utilisé pour l'interrogation du corpus qui, d'ailleurs, n'est pas annoté du point de vue morphosyntaxique.

4.2.2. Détection automatique par automates à états finis

Un *automate à états fini*s est une machine permettant la génération (en modalité écriture) et la reconnaissance (en modalité lecture) automatiques de chaînes de caractères qui relèvent d'un alphabet fini de symboles agencés par des règles de réécriture (Gross & Lentin 1970 : 121, 124). Autrement vu, un automate à états finis est représenté par un graphe orienté qui se compose d'un état initial et d'un état final et, par conséquent, d'un nombre fini d'états concaténés entre eux par des transitions (*ivi*, 122).

Plus précisément, l'application des automates à états finis que nous allons détailler concerne les automates déterministes conçus à l'usage des linguistes en vue du traitement informatique des particularités propres à une langue ou à un objet linguistique. C'est M. Gross qui appelle ses automates sous le nom de *grammaires locales* (1993, 1997). Pour contrecarrer les descriptions trop génériques autour de la phrase de la part de la grammaire dite traditionnelle, les *grammaires locales*, comme l'indique, d'ailleurs, leur nom, décrivent les contraintes locales de toute réalisation linguistique, sans aucune exception. Ce qui implique que :

> « the global nature of language results from the interaction of a multiplicity of local finite-state schemes » (Gross 1997 : 330).

L'empirisme de Gross s'accompagne au souci de l'interopérabilité. Une grammaire locale est ainsi réutilisée autant pour enrichir d'autres grammaires locales que pour l'analyse linguistique automatique de tout corpus d'une langue donnée. Comme le précise Gross, par une métaphore souvent employée pour décrire les imbrications linguistiques :

> « This is somewhat similar to the way small molecules combine to produce much larger ones in organic chemistry » (Gross 1997 : 331).

Dans les paragraphes qui suivent, nous analyserons les expériences de conception de grammaires locales par les parémiologues qui se sont situés dans la méthode du Lexique-Grammaire. On verra que la description locale mettra en relief des astuces empiriques soignées qui permettront la reconnaissance de diverses occurrences des proverbes/parémies dans des corpus.

4.2.2.1. Conenna (1995, 1998a, 2000b, 2004)

Outre les dictionnaires électroniques au format DELA et les expressions régulières (§ 4.2.1.6.), Conenna emprunte également les grammaires locales (ou graphes) au binôme Lexique-Grammaire-logiciel INTEX[197]. En 1995, dans son étude pionnière en la matière, elle élabore les premiers graphes de proverbes français et italiens, notamment pour le repérage de formes exactes ainsi que de quelques variantes lexicales synchroniques (1995 : 208-209 ; 212-214) :

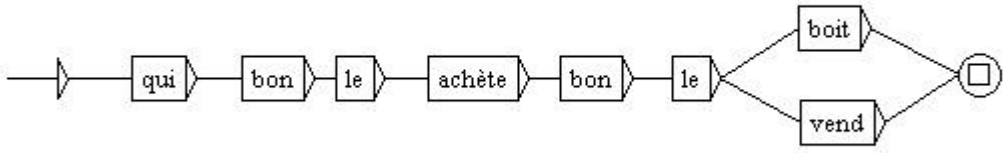


Figure 6. Graphe qui représente deux variantes lexicales synchroniques d'après Conenna (1995 : 208).

et diachroniques (1995 : 214) :

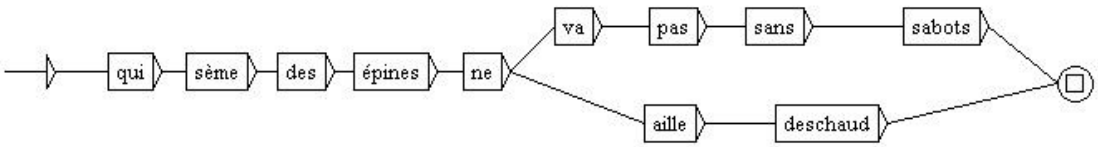


Figure 7. Graphe qui représente deux variantes lexicales diachroniques d'après Conenna (1995 : 214).

[197] Il faut préciser que INTEX tout comme maintenant *Unitex* et *Nooj* tokenisent, annotent et lemmatisent les textes à l'aide de grammaires locales et de dictionnaires électroniques des formes simples et composées au format DELA. C'est grâce à ces opérations qu'il devient possible par la suite d'effectuer des recherches qui visent exclusivement, entre autres choses, des parties du discours (§ 6.2.1.).

Suivant son maître, elle souligne :

> « On comprend alors l'efficacité d'une représentation par automates, permettant d'expliciter les régularités et les divergences formelles des proverbes, ainsi que de regrouper les exemples d'équivalence sémantique […] » (1995 : 210).

Le réseau orienté de l'automate satisfait la variation formelle et sémantique non seulement intralinguistique, mais aussi interlinguistique (1995 : 214-217). D'une part, elle remarque que les proverbes n'ont pas disparu de l'usage écrit, et ce, « malgré leur caractère 'vieillot' » (1995 : 217). D'autre part, c'est la perspective contrastive et traductologique français-italien qui pilote cette première formalisation sous forme d'automates en vue de leur repérage automatique et de leur comparaison interlinguistique, comme on lit dans les conclusions de l'étude (1995 : 218). Pour atteindre l'objectif de la comparaison interlinguistique, elle fait appel à la transduction : outre la reconnaissance d'une séquence, il est aussi possible d'insérer des données supplémentaires à la séquence reconnue. En l'occurrence, lors de la reconnaissance d'un proverbe donné en langue source, son équivalent est automatiquement identifié en langue cible parce que les deux sont associés au même automate (1995 : 215). Quelques années plus tard, Conenna commentera les avantages de l'utilisation d'automates transducteurs pour la traduction comme suit :

> « L'application d'un transducteur se révèle alors utile parce que, paradoxalement, le proverbe 'déguisé' peut être reconnu plus facilement par l'automate que par le traducteur humain ! » (2004 : 95).

Cette visée comparée se poursuit en 1998a, lors de sa discussion sur le figement proverbial. En appui de son analyse, elle utilise les automates comme outil de visualisation (didactique, dirait-on) de certains phénomènes de défigement qui intéressent le lexique et la syntaxe des proverbes autant français qu'italiens (1998a : 368-369). Dans cette étude, encore, les automates ne reprennent que les formes canoniques repérées dans des recueils parémiographiques, des variantes morphosyntaxiques et lexicales ainsi que des variantes apparentées par le sens. Une fois de plus, elle donne des exemples d'automates transducteurs.

En 2000, la présentation du dictionnaire électronique comparé français-italien des proverbes commençant par *Qui-Chi* est accompagnée d'une nouvelle série d'automates.

Outre les cas que nous avons mentionnés, Conenna présente un exemple d'automate capable de reconnaître la permutation des constituants du proverbe, comme le montre le graphe dans la Figure 7 :

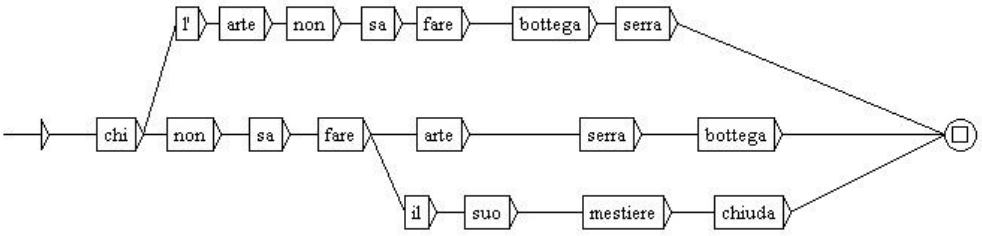


Figure 8. Graphe qui représente la permutation des constituants d'un proverbe italien d'après Conenna (2000 : 142).

Autrement dit, l'ordre linéaire canonique est remis en question, quoique la linéarité soit implicitement acquise comme une des dimensions pour créer les automates[198].

Dans l'étude de 2004, Conenna aborde la création d'automates alternatifs et plus développés par rapport à ceux qu'elle a conçus précédemment. La détection de la variation, notamment de celle morphosyntaxique, prévoit cette fois-ci sa reconnaissance par le recours à la lemmatisation des groupes verbaux.

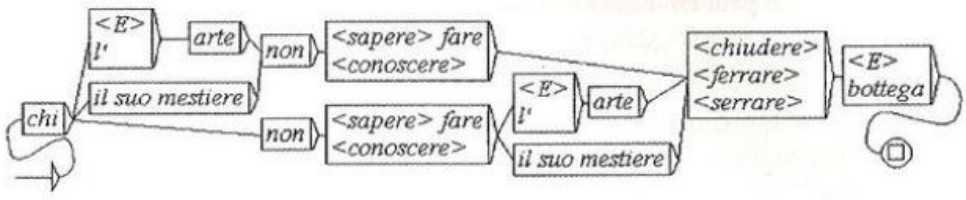


Figure 9 Automate de proverbes italiens avec lemmatisation des verbes tiré de Conenna (2004 : 94).

[198] Nous avons explicité cette dimension (§ 6.3.2.).

Comme nous avons eu l'occasion de vérifier dans nos études pilotes (§ 6.3.2.), la lemmatisation est aussi essentielle que performante parce qu'elle résume en une seule forme la possibilité de *n* variations morphosyntaxiques, parfois imprédictibles.

Un autre aspect imprédictible concerne les insertions. Pour faire face à l'imprévisibilité des modalités d'insertion d'un proverbe en discours et d'autres unités à l'intérieur du proverbe lui-même, Conenna suggère :

> « un modèle d'analyseur de proverbes, constitué par une série d'automates imbriqués » (2004 : 96)

qu'elle abrège comme MOD-INS. L'automate MOD-INS exploite le potentiel que tout automate d'INTEX (et UNITEX et *Nooj*) a d'appeler d'autres graphes, dits *sous-graphes*, créés précédemment et stockés indépendamment de l'automate MOD-INS (Figure 9).

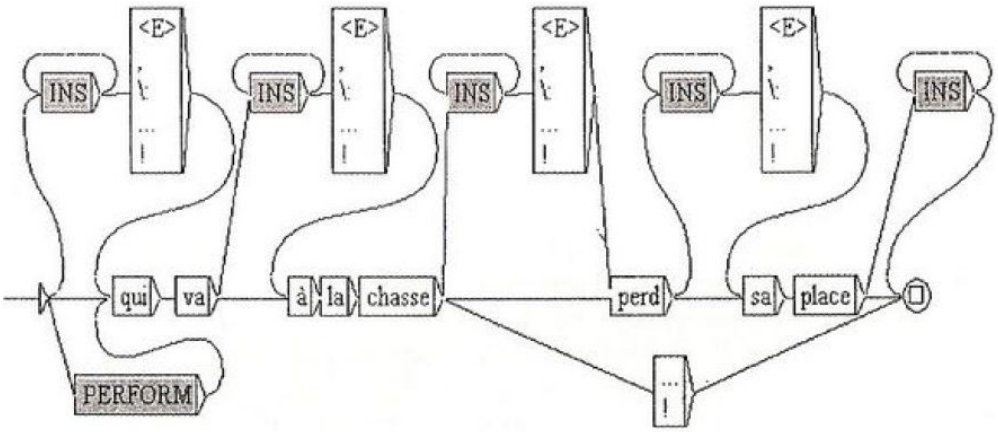


Figure 10. Automate MOD-INS tiré de Conenna (2004 : 96).

Une synergie s'établit entre plusieurs réseaux orientés de reconnaissance. Dans le cas de figure, MOD-INS vise la détection du proverbe :

Qui va à la chasse perd sa place

et appelle un sous-graphe (nommé PERFORM) contenant quelques suites de caractères qui servent normalement à introduire les proverbes en discours (2004 : 96-97) : un sous-graphe composé d'introducteurs. Un autre sous-graphe (appelé INS) plus élaboré inclut d'autres suites de caractères qu'on pourrait repérer au cœur même du proverbe (2004 : 98). À son tour (et c'est cet aspect qui est le plus puissant en termes prédictifs), le sous-graphe INS appelle d'autres sous-graphes où sont disponibles des ensembles lexicaux catégorisés d'après une référence réelle partagée par toutes les unités de ces ensembles. Il s'agit d'entités nommées, comme les noms de pays (ex. *Italie*) ou de villes (ex. *Nice*), mais aussi de noms de membres de famille (ex. *grand-père*) (2004 : 99) ou encore d'adjectifs dérivés de noms de pays (ex. *italien*) ou de langues (ex. *allemand*). Ces sous-graphes (plutôt lexicaux que syntaxiques) servent ainsi à satisfaire un éventail de modifications paradigmatiques concernant les introducteurs. Rien n'empêche qu'ils puissent remplir aussi la même fonction pour certaines unités qui relèvent de la combinatoire proverbiale à proprement parler.

À propos encore de l'automate MOD-INS, nous nous attardons quelques instants sur la position où est appelé le sous-graphe INS, suivi de l'état où est prévue une liste de signes de ponctuation. D'après son observation, Conenna souligne que :

> « la coupe est généralement syntagmatique et coïncide partiellement avec la mise en évidence des constituants » (2004 : 100).

Autrement dit, l'insertion est positionnée là où finit un groupe syntagmatique, outre le début et la fin du proverbe. Pour finir, MOD-INS permet la reconnaissance d'un usage tronqué, comme le témoigne l'état avec les points de suspension et le point d'exclamation en bas de la Figure 9. Il est dommage, en tout cas, de constater que la lemmatisation n'est pas prise en compte pour le proverbe, alors qu'elle l'est pour les autres sous-graphes.

Conenna a ainsi le mérite d'avoir ouvert la voie à une détection des proverbes par des requêtes informatiques autres que les expressions régulières. Malgré cela, il est vrai que les graphes qu'elle a conçus sous-représentent, en quelque mesure, le potentiel créateur et novateur des séquences lexico-grammaticales des proverbes. C'est ce potentiel que nous souhaitons explorer dans notre recherche.

4.2.2.2. Tsaknaki (2006)

Dans la même lignée des travaux de Conenna, nous citons la systématisation sous forme de graphes faite par Tsaknaki pour 2.500 proverbes grecs à l'aide du logiciel UNITEX. Nous n'approfondirons que trois aspects de son étude :

1. la taxinomie des variantes formelles ;
2. l'identification proverbe-phrase ;
3. et la transduction en vue d'une annotation sur corpus.

Tsaknaki établit une taxinomie des variantes formelles des proverbes grecs :

- *variantes graphiques* d'après la présence ou l'absence de signes de ponctuation ;
- *variantes orthographiques* qui peuvent être influencées par des facteurs phonologiques;
- *variantes morphologiques*, notamment en ce qui concerne les suffixes verbaux ;
- *variantes lexicales* qui affectent certaines unités lexicales ;
- *variantes morphosyntaxiques* et que nous dirions de variation paradigmatique par rapport à une même partie du discours, comme la préposition (2006 : 58).

Par cette taxinomie non exhaustive, Tsaknaki conclut, comme bien d'autres parémiologues avant elle :

> « locating proverbs yields good results to the extent that variants are exhaustively and systematically represented » (*ibid*.)

et souligne en ce sens le grand avantage d'exploiter les automates à la place des dictionnaires électroniques.

De façon originale par rapport à Conenna et à cause des particularités propres aux proverbes grecs, Tsaknaki insiste aussi sur leur statut phrastique. Elle précise que certains proverbes grecs dépassent le cadre de la phrase graphique standard et qu'ils sont composés de

deux phrases graphiques. Pour éviter des problèmes de reconnaissance, après avoir conçu un automate similaire à ceux de Conenna, Tsaknaki l'inclut comme sous-graphe appelé par l'automate de segmentation de corpus grecs en phrases (2006 : 59-60). De cette façon, elle évite que, lorsque l'automate de segmentation découpe un texte grec en phrases, ces proverbes multiphrastiques soient reconnus comme deux phrases séparées par un signe de ponctuation, comme le point final.

Pour conclure, une dernière particularité de Tsaknaki consiste à prévoir deux états (un initial et un autre final) des automates pour effectuer une transduction. Plus précisément, tout automate qu'elle conçoit permet le balisage des proverbes sur corpus comme suit :

> από κάθε άλλον – [PROV] [η καλή μέρα απ' το πρωί φαίνεται].[PROV] {S} Πήγα για να κάνω το χατίρι του
> [trad. litt. : [PROV] [la bonne journée du matin semble] [PROV]] (2006 : 61)[199].

La transduction par automate devient ainsi un outil pour une annotation et pour une extraction ciblées des proverbes sur corpus. Dans ce cas de figure, le proverbe est balisé du début à la fin et mis entre crochets.

4.2.2.3. Lacavalla (2007)

Suivant encore les recherches de Conenna, Lacavalla met au point la description en tables du Lexique-Grammaire de 293 proverbes français et 650 italiens (2007 : 94) commençant par *Quand*/*Quando*. C'est à partir de ces tables qu'elle crée des graphes pour chaque proverbe, et ce, au moyen d'un *graphe paramétré*, à savoir un graphe qui contient des variables de lecture des tables. Par cette opération de conversion, UNITEX génère un seul graphe composé de sous-graphes, chacun d'eux correspondant à chaque ligne-proverbe de la table source, comme on le voit dans la Figure 10 ci-dessous :

[199] On précise également que *{S}* est la balise de segmentation en phrases utilisée par *Unitex*.

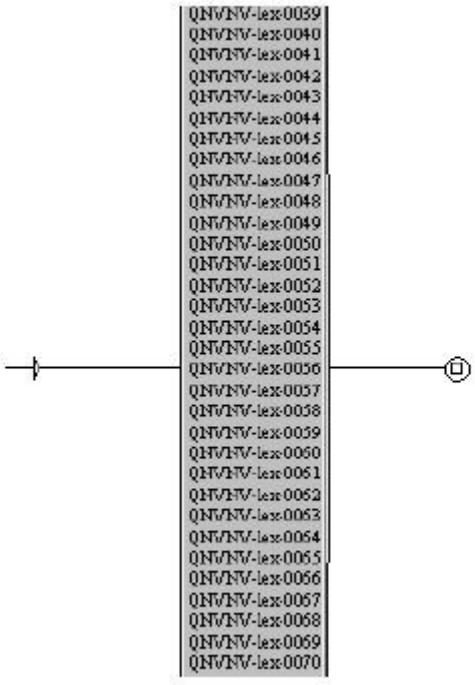


Figure 11. Graphe appelant les sous-graphes des proverbes de la classe *Quand* tiré de Lacavalla (2007 : 260).

La contribution précieuse de l'étude de Lacavalla est à repérer au niveau interlinguistique dans la création des tables des correspondances entre les proverbes français et italiens (2007 : 267-269). C'est à partir de ces tables qu'elle crée des graphes transducteurs dont les entrées sont les proverbes en langue source et les sorties leurs correspondants en langue cible.

4.3. Un point de départ empirique et informatique fédérateur

Au bout de ce recensement, nous proposons un volet complémentaire au cadre méthodologique pour les études de fréquence que nous avons présenté au § 3.3. Nous synthétisons les expériences que nous venons de présenter pour établir un cadre méthodologique informatique. En tout cas, il se veut un point de départ, non pas un éventail fermé de choix méthodologiques. La généricité et l'approximation de certaines études sur support électronique peuvent être en effet motivées par la nouveauté (tout à fait relative en

sciences du langage, bien évidemment) de l'introduction de l'ordinateur et des logiciels de traitement automatique de corpus en parémiologie. Comme on vient de le constater, les techniques de détection se sont appuyées souvent sur le bon sens ainsi que sur le 'penchant' informatique montré par les parémiologues, plutôt que sur une réflexion systématique de l'interface parémie-ordinateur.

Quant aux expériences de détection 'à vue' des parémies sur support papier, elles nous ont donné un encadrement formel générique et généralisant, à l'exception près de quelques remarques formelles pointues surtout de Schulze-Busacker (§ 4.1.3.) et que nous intégrerons dans notre état de l'art informatique.

On remarquera que se sont constituées deux grandes pratiques de repérage automatique des parémies :

A. la pratique fondée sur des expressions régulières génériques appliquées à des corpus non annotés ;
B. la pratique basée sur des automates à états finis détaillés appliqués à des corpus annotés.

4.3.1. *Expressions régulières*

Les expressions régulières sont décidément privilégiées pour le repérage des parémies dans des corpus. Ces derniers ne sont ni morphosyntaxiquement annotés ni lemmatisés. En ligne générale, les parémiologues préfèrent :

- des requêtes informatiques lexico-centriques,
- assez pauvres en ce qui concerne la syntaxe des parémies, malgré la constatation *a posteriori* de l'interaction lexique-syntaxe et
- décidément pauvres en ce qui concerne la syntaxe de recherche propre aux expressions régulières, dont la souplesse et la richesse permettraient une modélisation plus précise que celle assez générique que nous avons observée.

On distingue deux grands types de requêtes informatiques sous forme d'expressions régulières :

1. le mot graphique simple :
    a. qui relève du proverbe (ou de la parémie)
    b. ou qui cible le lexique paraproverbial (ou paraparémique), c'est-à-dire les introducteurs ;
2. la cooccurrence (référée, entre autres, comme *mots-clés*, *mots les plus essentiels*, *noyau phraséologique résiduel*, *noyau proverbial*, *partie proéminente*) entre mots graphiques :
    a. qu'elle soit continue et encore :
        i. contrainte en ce qui concerne l'ordre linéaire des mots, et ce, de façon à reproduire la cooccurrence continue de la parémie recherchée ; souvent cette contrainte intéresse la partie initiale de la parémie ;
        ii. libre, sans aucune précaution à l'égard de la syntaxe de recherche des expressions régulières ;
    b. ou qu'elle soit discontinue et, dans ce cas de figure :
        i. contrainte par la prise en compte d'un nombre *n* de chaînes de caractères non définies, à savoir par la définition d'une fenêtre de cosélection dont la distance linéaire reflète celle qui est observée dans la parémie recherchée ;
        ii. libre, sans indications d'une fenêtre de cosélection.

La cooccurrence entre mots graphiques a rarement prévu la troncature (*), c'est-à-dire la possibilité de remplacer une partie (ou l'intégralité) d'une chaîne de caractères par d'autres chaînes de caractères non définies a priori. La troncature a représenté la technique employée par les parémiologues pour assurer davantage de flexibilité aux expressions régulières, alors que moins souvent ils ont fait recours à la lemmatisation.

Plus rarement, on a pris en compte la variation paradigmatique au sein d'une cooccurrence lexicale, notamment par la recherche d'ensembles de mots apparentés du point de vue sémantique ou ontologique (§ 4.2.1.4.).

Cette insistance sur une cooccurrence lexicale et que les parémiologues ont estimée comme significative a apporté, en tout cas, la preuve d'une cooccurrence lexicale étendue à la totalité de certaines parémies, souvent répétée ou exploitée. C'est le cas du *cliché proverbial* (§ 4.2.1.1.), mais aussi (et de manière paradoxale, au premier abord) du *moule syntaxique* (§§ 4.2.1.6.). Ce qui prouve de ne pas séparer le lexique de la syntaxe au moment de l'interrogation du corpus pour davantage de précision des résultats.

La persévérance des parémiologues autour d'une cooccurrence lexicale a payé aussi en ce qui concerne l'observation sur terrain de l'évolution phraséologique d'une donnée linguistique parémiologique. À partir de la parémie, la requête informatique qui cible la cooccurrence lexicale montre le « passage en unité phraséologique » ou la *phraséologisation* d'une parémie (processus inverse à la *parémisation*).

S'écartent de ce panorama les approches ponctuelles de Maniez (§ 4.2.1.9.) et de Hrisztova-Gotthardt & Gotthardt (§ 4.2.1.14.) où les expressions régulières intègrent une réflexion informatique développée et ajustée au fur et à mesure. Dans le cas de Maniez, c'est la structuration des données parémiques dans une base de données qui facilite le repérage de mots graphiques simples et de cooccurrences. En ce qui concerne Hrisztova-Gotthardt & Gotthardt, la définition d'une liste tokenisée de parémies prétraitées et son passage comme argument à un script informatique étalent la longueur des expressions régulières jusqu'à la totalité de la séquence parémique[200].

4.3.2. *Automates à états finis*

La description locale des automates à états finis représente, en quelque mesure, un contre-exemple de spécificité et de finesse par rapport à la généricité et au sommaire des expressions régulières que nous avons examinées. De l'attention à quelques constituants lexicaux privilégiés de par leur cooccurrence, on passe ainsi à une représentation (parfois, une surreprésentation) de tout constituant parémique. Le rapprochement de la forme exacte recherchée permet de mettre au point une taxinomie des variantes parémiques, entre synchronie et diachronie, et qui correspondent, suivant les expériences de Conenna et de Tsaknaki, à :

- *variante graphique* ;
- *variante orthographique* ;
- *variante morphologique* ;
- *variante morphosyntaxique* ;
- *variante lexicale*.

[200] Navarro-Brotons recherche également les séquences parémiques dans leur intégralité, quoique sans réflexion explicite sur l'interrogation du Web.

La lemmatisation de certains constituants des parémies facilite la rencontre de la plupart de ces variantes (excepté les variantes lexicales qui devrait s'appuyer sur le bootstrapping), sans passer sous silence quelques-unes de leurs occurrences. Entre exactitude et variation formelles, le réseau orienté des grammaires locales permet ainsi de repérer le nombre plus élevé d'occurrences des parémies dans des corpus annotés et lemmatisés. Ce qui est un grand avantage lorsqu'on envisage une étude de fréquence.

Les automates prennent en compte autant la totalité que quelques constituants de la séquence formulaire parémique. Comme on l'a vu, on peut les modéliser pour qu'ils gèrent la permutation interne ainsi que la reconnaissance d'un ou d'une séquence de constituants, par leur conjonction à l'état final. L'ordre des mots et la linéarité peuvent ainsi varier.

L'imprévisibilité des insertions peut être gérée par l'imbrication d'automates simples, notamment par la synergie entre plusieurs graphes syntaxiques et/ou lexicaux qui permettent la reconnaissance d'introducteurs et/ou de mots au début et/ou après quelques-uns des constituants parémiques.

Par le recours à la transduction, outre la reconnaissance, il est aussi possible d'enrichir le corpus qu'on interroge. Ce qui facilite le balisage des formes recherchées ou l'enrichissement d'un corpus avec d'autres informations, qu'elles soient linguistiques (ex. l'équivalence sémantique interlinguistique) ou extralinguistiques.

On pourrait même prévoir que les automates conçus pour la reconnaissance soient utilisés pour segmenter un corpus en parémies avant leur consultation. Ce qui améliorerait leur calcul de la fréquence d'occurrence (§ 3.2.8., 3.3.3.).

Au cas où l'on démarrerait par une description lexico-grammaticale préalable, l'utilisation d'un graphe paramétré peut résumer une série de données linguistiques ordonnées en une seule requête informatique.

Or, les études que nous avons mentionnées n'ont pas vraiment présenté des automates qui concentrent toutes ou la plupart de ces possibilités pour concevoir des graphes hautement performants. En outre, malgré la possibilité d'exploiter le côté syntaxique, c'est le lexique qui a continué à guider la recherche sur corpus et, peut-être, à produire du silence dans les résultats. Aucun graphe n'a fait recours, par exemple, à la reconnaissance d'une partie du discours à la place d'une unité lexicale simple pour prendre en compte la variation paradigmatique d'un constituant parémique.

4.4. En guise de conclusion

Habert *et al.* (1997 : 121) nous rappellent que l'expérimentation peut aider à mieux expliquer et à mieux saisir un fait linguistique dans un corpus. C'est par l'expérimentation d'autres parémiologues et d'autres phraséologues que nous avons essayé d'esquisser un état de l'art (exhaustif, pour autant que possible) concernant la modélisation des parémies sous forme de requêtes informatiques.

La représentation informatique et les paramètres choisis pour la création des requêtes ont certes influencé les résultats quantitatifs que nous avons discutés au § 3. D'où la nécessité de consacrer une réflexion approfondie à l'acquis en matière de repérage des parémies, notamment de repérage automatique. Autant les expressions régulières que les automates à états finis se sont révélés deux manières souples pour identifier les parémies dans des corpus. Néanmoins, on a eu tendance à exploiter les premières de façon assez générique et les derniers de façon assez spécifique par rapport aux formes des parémies à détecter.

Ce sera l'équilibre entre généricité et spécificité, notamment entre lexique et syntaxe, qui sera empiriquement réglé dans nos requêtes informatiques. D'une part, nous serons guidé par le précipité de l'empirisme que nous venons de présenter. D'autre part, nous nous appuierons sur notre classification lexico-grammaticale (§ 5). Les parémies de notre liste nous aideront elles-mêmes à définir les degrés de généricité et de spécificité à adopter pour la conception de nos requêtes informatiques (§ 6).

CHAPITRE 5

LA LISTE.
ANNOTATION ET CLASSIFICATION LEXICO-GRAMMATICALES

Dans le présent chapitre, nous illustrerons notre liste ou, plus précisément, notre corpus parémique (§ 3.3.1.), qui s'appuie sur un recueil parémiographique unique en milieu francophone : *DicAuPro* (§ 5.1.). Nous motiverons ce choix et détaillerons surtout le processus de sélection des formes canoniques que nous utiliserons pour notre étude de fréquence.

Par la suite, nous expliquerons l'étape d'annotation morphosyntaxique et de lemmatisation de notre corpus parémique à l'aide de l'annotateur automatique *TreeTagger* (§ 5.2.). Loin de nous avoir mis à l'abri de notre subjectivité, l'annotation automatique a fait l'objet d'une vérification et d'une correction manuelles. D'une part, cela a permis d'évaluer la performance de *TreeTagger* à l'égard des spécificités morphosyntaxiques et lexicales de nos parémies. D'autre part, la vérification manuelle nous a confronté à des cas douteux dont la résolution a été parfois immédiate, parfois de plus longue haleine.

En conclusion, nous approfondirons la démarche de classification lexico-grammaticale (§ 5.3.) que nous avons esquissée au § 1.2.6. C'est à partir de cette classification que nous modéliserons les requêtes informatiques qui nous serviront pour interroger nos corpus (§ 6).

5.1. Liste

Nous avons décidé de partir d'une *liste parémiographique* (*lp*), à savoir d'une liste *l* dont le nombre de parémies $p \geq 1$ relève d'un recueil parémiographique (§ 3.3.1.). Nous avons choisi (et eu l'honneur d'exploiter) la base de données informatisée *Dictionnaire*

Automatique et philologique des Proverbes (désormais *DicAuPro*) (Conenna *et al*. 2006)[201]. Plus précisément, notre *lp* inclut toutes les séquences formulaires qui sont indiquées comme *formes canoniques* par l'équipe de parémiologues qui gère la base. Il s'agit donc d'une *lp intégrale* (§ 3.3.1.).

Avant d'aborder la taille de notre *lp*, nous nous penchons sur les critères de création de *DicAuPro* et qui ont motivé notre choix. Tous les encodages ainsi que toute identification du statut de forme canonique dans *DicAuPro* résultent d'une sélection minutieuse de longue durée et en cours (depuis une dizaine d'années). L'équipe s'est dotée d'un protocole commun d'encodage qui garantit l'uniformité du travail et, comme le disent les parémiologues, « constitue un des piliers de cette recherche » (Conenna *et al*. 2006 : 80). Fait partie du protocole l'inventaire des recueils parémiographiques dépouillés, à savoir des grands dictionnaires généraux de français, des dictionnaires historiques, ainsi que des textes littéraires suivis de listes de proverbes et des textes théâtraux[202]. Tous couvrent la période qui va du Moyen Âge jusqu'au XX^e^ siècle. Nous partons donc d'une *lp* qui est à présent le seul exemple français de parémiographie philologique et diachronique informatisé[203]. En ce sens, nous n'avons pas besoin de filtrer notre *lp* et de la comparer avec d'autres recueils parémiographiques.

En ce qui concerne le statut de forme canonique, le choix pose les questionnements les plus délicats. Le lemme parémiographique n'est qu'une variante privilégiée parmi d'autres, ce privilège étant néanmoins accordé sur la base de critères explicites. Le témoin parémiographique-pivot du dépouillement est représenté par le dictionnaire de *Littré* (XIX^e^ siècle) d'où ont été tirées la plupart des formes canoniques (*ivi*, 81). Dans les cas de plusieurs variantes attestées dans le même dictionnaire *Littré* sous des lemmes différents, les parémiologues ont préféré la variante attestée sous la première voix par ordre alphabétique (*ibid*.). Ce critère est certes subjectif, mais il a l'avantage de garantir un dépouillement méthodique et homogène du témoin parémiographique-pivot. Or, le XIX^e^ siècle auquel remonte le *Littré* est décidément loin de nos jours. Ce qui a comporté que certaines formes canoniques appartiennent plutôt au *Grand Larousse Encyclopédique* (GLE) et au *Grand Dictionnaire Encyclopédique Larousse* (GDLE) qui datent environ de 1960 à 1985 (*ivi*, 82). Un premier jugement introspectif sur l'usage, notamment sur la familiarité et sur la fréquence

[201] À ce propos, nous remercions Mirella Conenna, Monique Coppens d'Eeckenbrugge, Fiorella Flamini, Jean-René Klein et Jean-Marie Pierret.
[202] Par respect du travail scientifique en cours, les références exactes des témoins parémiographiques ne peuvent pas être dévoilées, sauf dans les cas où elles ont été déjà mentionnées par un des encodeurs.
[203] Pour un recueil parémiographique philologique en anglais, lire la note 103 au § 3.2.4.

supposée, s'insinue ainsi dans le choix de la forme canonique. De toute façon, à la différence d'autres recueils parémiographiques, cette subjectivité a le mérite d'être avouée et encodée. En effet, les membres de l'équipe *DicAuPro* ont systématiquement mentionné le témoin parémiographique préféré pour chaque forme canonique (*ibid*.)[204]. Klein (2006) détaille la démarche de sélection des formes canoniques que nous résumons dans le Tableau 18 qui suit :

I. Une seule forme attestée dans le dictionnaire *Littré*…	
1) absente dans GLE, GDEL et d'autres recueils parémiographiques	→ forme canonique = forme *Littré*
2) présente dans GLE, GDEL et d'autres recueils parémiographiques	→ forme canonique = forme *Littré*
3) différente de celles repérées dans GLE, GDEL et d'autres recueils parémiographiques	→ forme canonique = forme ayant *f* parémiographique la plus élevée
II. Plusieurs variantes sont attestées dans le dictionnaire *Littré*…	
1) et une variante est attestée dans (la plupart) des recueils parémiographiques (GLE, GDEL et d'autres)	→ forme canonique = variante ayant *f* parémiographique la plus élevée
2) et les variantes sont attestées dans GLE, GDEL et d'autres recueils parémiographiques	→ forme canonique = variante ayant *f* parémiographique la plus élevée et/ou la plus familière et/ou la plus attestée sur le Web
3) et aucune variante n'est attestée dans GLE, GDEL et d'autres recueils parémiographiques	→ forme canonique = forme ayant *f* parémiographique la plus élevée
III. Forme attestée dans le dictionnaire *Littré* disparu de l'usage courant	
→ forme canonique = forme ayant *f* parémiographique la plus élevée dans GLE, GDEL et d'autres recueils parémiographiques	
IV. Formes non attestées dans le dictionnaire *Littré*	
→ forme canonique = forme ayant *f* parémiographique la plus élevée dans GLE, GDEL et d'autres recueils parémiographiques	

Tableau 18. Protocole de sélection des formes canoniques de *DicAuPro* d'après Klein (2006).

[204] La documentation (titres des recueils ou ouvrages consultés, auteurs, datations) de toute forme canonique ainsi que de toute variante repérée constitue l'appui crucial pour toute spéculation successive, tout comme pour la construction de l'historique du proverbe (Conenna 2002), à savoir la reconstruction sémantique de la séquence formulaire parémique au fil de ses variantes diachroniques.

En général, on constate que la plupart des choix d'une forme canonique que l'équipe *DicAuPro* a envisagés sont confiés à *f parémiographique* dans des recueils parémiographiques plus récents par rapport à celui de *Littré*. Ils suivent la normalisation parémiographique la plus contemporaine à leur œuvre de description parémiographique. Nous soulignons le cas II.2) où, outre *f parémiographique*, les études de familiarité et le nombre d'attestations repérées à l'aide de *Google* sur le Web deviennent des heuristiques indispensables pour trancher parmi plusieurs variantes[205]. Ce filtre de *f parémiographique* en diachronie de notre *lp* nous confirme encore dans le propos d'éviter des comparaisons ultérieures avec d'autres recueils parémiographiques.

Par rapport aux classifications typologiques (§ 1.2.1.), *DicAuPro* se situe à l'extérieur dans la mesure où le terme *proverbe* subsume également ce que normalement d'autres appelleraient des dictons ou des adages. Nous traitons ainsi quelques parémies, mais, certes, nous n'avons pas d'aphorismes dans notre *lp*.

Un dernier critère extralinguistique qui a motivé notre choix de *DicAuPro* et de ses formes canoniques concerne la gratuité de la consultation tout comme la disponibilité immédiate au format électronique.

Il en suit que notre *lp* est plutôt un *corpus parémique* (§ 3.3.1.) parce qu'elle satisfait les critères (i)-(v) que nous avons définis pour qualifier la notion de *corpus* en linguistique de corpus (§ 2.2.2.). Notre liste est ainsi assimilable à un corpus.

Comme nous le disions, *DicAuPro* est une base de données en évolution continue qui essaie d'établir à sa façon un équilibre entre tradition et usage. Par conséquent, la taille de notre corpus parémique a changé au fur et à mesure de notre recherche. À la date du 1er juillet 2010, année-pivot (§ 3.3.1.) de notre *lp*, notre corpus comptait 1.689 formes canoniques (§ Annexe 1). Par souci de précision, nous avons effectué un deuxième contrôle des formes canoniques à la mi-août 2013, choisie ainsi comme année-contraste (§ 3.3.1.). Nous avons ainsi relevé 130 modifications (§ Annexe 2) qui ont affecté notre *liste bêta* (*lβ*) (§ 3.3.1.). Ces 130 modifications se répartissent en 43 effacements de formes canoniques et 87 formes canoniques différentes par rapport à notre *lβ*. De ces 87 dernières, 31 remplacent des formes canoniques mentionnées dans *lβ* et 56 sont des nouvelles formes canoniques.
Nous avons ainsi décidé d'agir comme suit :

[205] Nous avons exprimé nos perplexités quant à l'utilisation du Web comme corpus aux §§ 2-3. Néanmoins, pour l'équipe *DicAuPro*, le nombre d'attestations renvoyées par *Google* n'est pas le critère exclusif de sélection d'une variante comme forme canonique (comme dans le cas de Navarro Brotons, § 3.2.15.). Il joue plutôt le rôle d'indicateur d'usage à côté et au même titre que d'autres critères.

- pour étendre notre étude de fréquence au nombre le plus élevé de parémies, nous avons accueilli dans notre corpus parémique les 56 nouvelles formes canoniques ajoutées dans la période 2010-2013 ;
- pour garantir encore l'étendue de notre corpus parémique ainsi que pour mieux évaluer a posteriori d'après *f* que nous enregistrerons, nous avons gardé les 43 formes canoniques disparues de la mise à jour de *DicAuPro* ;
- en ce qui concerne les 31 formes canoniques remplacées, d'une part, nous avons conservé celles qui relèvent de l'année-pivot 2010 et, d'autre part, nous avons absorbé les remplaçantes de l'année-contraste 2013.

Tout compte fait, de l'année-pivot 2010 à l'année-contraste 2013, notre corpus parémique a englobé 87 autres formes canoniques.

La taille finale de notre corpus parémique s'élèverait à 1.776 formes canoniques, quoique nous avons décidé de retirer la forme canonique :

Dieu nous garde… [+ compléments désignant des désagréments]

Comme on peut le constater, il s'agit d'une forme décrite de manière partielle. Plutôt, il s'agirait d'un schéma ou d'un cadre matriciel de séquences formulaires à valeur parémique. Ce qui est d'ailleurs un des buts de notre recherche. Nous avons également mis de côté :

Bis repetita placent

parce que nous souhaitons nous concentrer seulement sur le répertoire parémiologique en français. Au final, donc, notre corpus parémique inclut 1.774 formes canoniques (14.269 occurrences[206]).

[206] La segmentation et la tokenisation de notre corpus parémique sont effectuées par *Unitex 3.1* (§ 6.2.). Nous détaillerons quelques caractéristiques de ce logiciel au § 6. Les occurrences correspondent aux mots graphiques simples.

5.2. Annotation morphosyntaxique et lemmatisation

Comme il peut arriver à tout corpus (§ 2.2.5.4.), nous avons soumis notre corpus parémique à une annotation morphosyntaxique et à une lemmatisation. Nous avons exploité l'annotateur automatique *TreeTagger* (Schmid 1994) pour (apparemment) accélérer notre description lexico-grammaticale.

Pour faciliter la tâche de reconnaissance, nous avons prétraité notre corpus parémique et ajouté un point et un retour à la ligne en fin de chaque parémie à l'aide d'un très simple script PERL[207]. Après l'annotation et la lemmatisation par *TreeTagger*, nous avons splitté le fichier de texte brut CSV grâce à d'autres lignes en PERL. Dans un nouveau fichier de texte brut, chaque parémie a donc occupé une seule ligne se terminant par le tag SENT (qui indique la fin de la phrase graphique) et par un retour à la ligne, comme le montre la figure ci-dessous.

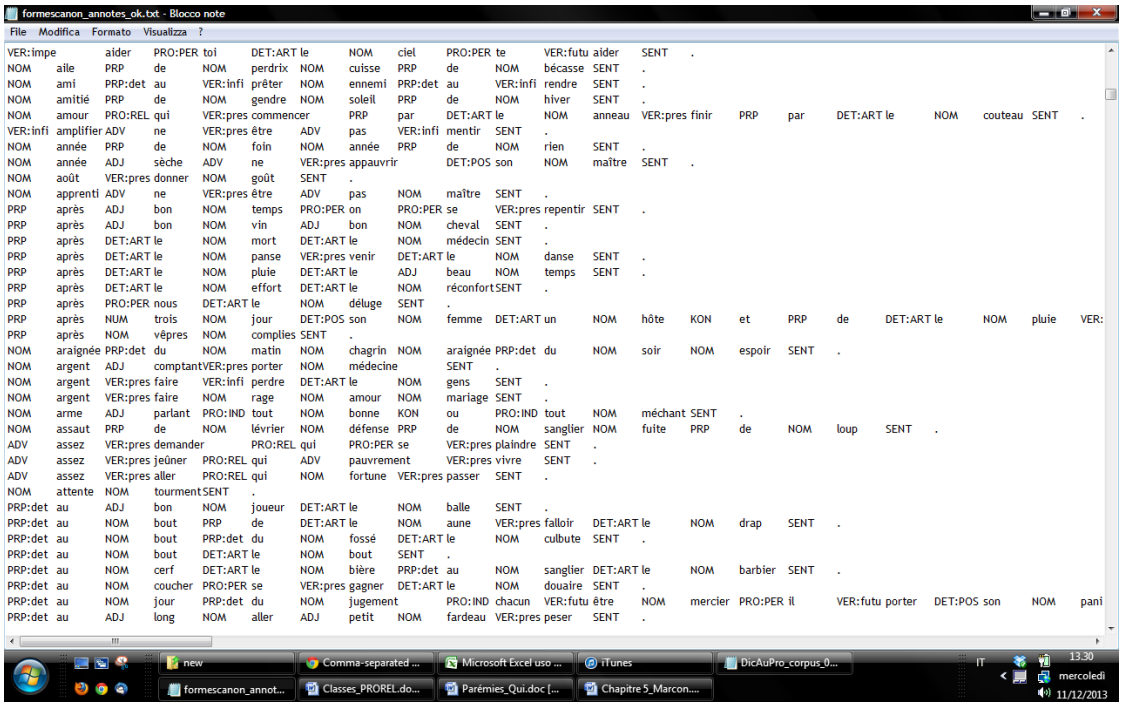
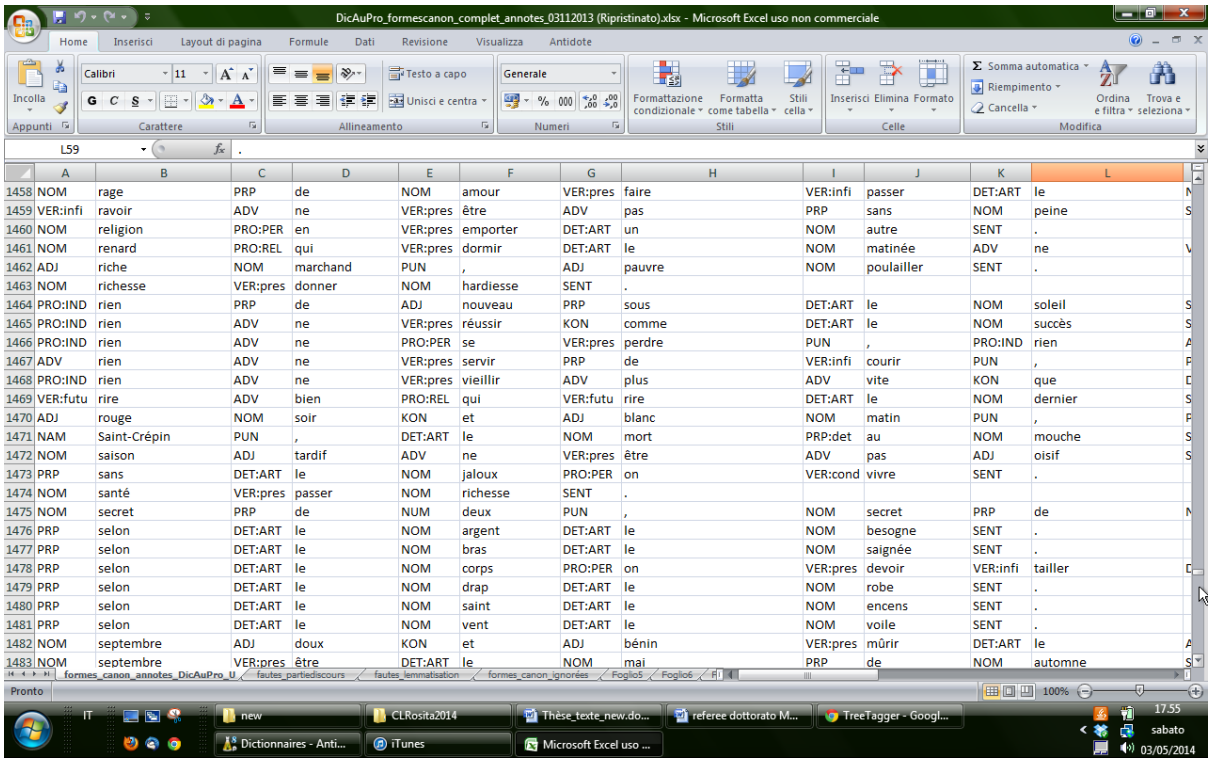


Figure 12. Fichier de texte brut contenant les formes canoniques de DicAuPro annotées et lemmatisées par TreeTagger.

[207] Par ce choix, nous avons repris la notion de *phrase*, mais seulement dans son acception *graphique* qui est, d'ailleurs, celle de ses origines (§ 1.1.1.). L'identité phrase-parémie est ici fonction d'un traitement automatique, non pas de nature linguistique.

C’est ce fichier que nous avons importé dans le tableur *Excel* :



	A	B	C	D	E	F	G	H	I	J	K	L
1458	NOM	rage	PRP	de	NOM	amour	VER:pres	faire	VER:infi	passer	DET:ART	le
1459	VER:infi	ravoir	ADV	ne	VER:pres	être	ADV	pas	PRP	sans	NOM	peine
1460	NOM	religion	PRO:PER	en	VER:pres	emporter	DET:ART	un	NOM	autre	SENT	.
1461	NOM	renard	PRO:REL	qui	VER:pres	dormir	DET:ART	le	NOM	matinée	ADV	ne
1462	ADJ	riche	NOM	marchand	PUN	,	ADJ	pauvre	NOM	poulailler	SENT	.
1463	NOM	richesse	VER:pres	donner	NOM	hardiesse	SENT	.				
1464	PRO:IND	rien	PRP	de	ADJ	nouveau	PRP	sous	DET:ART	le	NOM	soleil
1465	PRO:IND	rien	ADV	ne	VER:pres	réussir	KON	comme	DET:ART	le	NOM	succès
1466	PRO:IND	rien	ADV	ne	PRO:PER	se	VER:pres	perdre	PUN	,	PRO:IND	rien
1467	ADV	rien	ADV	ne	VER:pres	servir	PRP	de	VER:infi	courir	PUN	,
1468	PRO:IND	rien	ADV	ne	VER:pres	vieillir	ADV	plus	ADV	vite	KON	que
1469	VER:futu	rire	ADV	bien	PRO:REL	qui	VER:futu	rire	DET:ART	le	NOM	dernier
1470	ADJ	rouge	NOM	soir	KON	et	ADJ	blanc	NOM	matin	PUN	,
1471	NAM	Saint-Crépin	PUN	,	DET:ART	le	NOM	mort	PRP:det	au	NOM	mouche
1472	NOM	saison	ADJ	tardif	ADV	ne	VER:pres	être	ADV	pas	ADJ	oisif
1473	PRP	sans	DET:ART	le	NOM	jaloux	PRO:PER	on	VER:cond	vivre	SENT	.
1474	NOM	santé	VER:pres	passer	NOM	richesse	SENT	.				
1475	NOM	secret	PRP	de	NUM	deux	PUN	,	NOM	secret	PRP	de
1476	PRP	selon	DET:ART	le	NOM	argent	DET:ART	le	NOM	besogne	SENT	.
1477	PRP	selon	DET:ART	le	NOM	bras	DET:ART	le	NOM	saignée	SENT	.
1478	PRP	selon	DET:ART	le	NOM	corps	PRO:PER	on	VER:pres	devoir	VER:infi	tailler
1479	PRP	selon	DET:ART	le	NOM	drap	DET:ART	le	NOM	robe	SENT	.
1480	PRP	selon	DET:ART	le	NOM	saint	DET:ART	le	NOM	encens	SENT	.
1481	PRP	selon	DET:ART	le	NOM	vent	DET:ART	le	NOM	voile	SENT	.
1482	NOM	septembre	ADJ	doux	KON	et	ADJ	bénin	VER:pres	mûrir	DET:ART	le
1483	NOM	septembre	VER:pres	être	DET:ART	le	NOM	mai	PRP	de	NOM	automne

Figure 13. Feuille de calcul Excel contenant les formes canoniques de DicAuPro annotées et lemmatisées par TreeTagger.

pour procéder à la vérification manuelle de l’annotation et de la lemmatisation (§ 5.2.2.).

5.2.1. *TreeTagger*

TreeTagger étiquette tout corpus sur base probabiliste à l’aide d’arbres de décision *language-dependent*. Plus précisément, les contraintes distributionnelles qui concernent la cooccurrence des parties de discours en français (dont le jeu d’étiquette est repris dans l’Annexe 3) sont préalablement encodées. Ce qui convient au côté distributionnel de notre approche. De plus, son adaptabilité à 16 langues, y compris le français ancien, ainsi que son application imbriquée dans des logiciels en ligne populaires, tel *Sketch Engine* (Kilgarriff *et al*. 2004), nous ont fait pencher sur cet annotateur gratuit et en évolution continue grâce aux apports de plusieurs linguistes informaticiens.

En détail, le choix 'le plus approprié' d'étiquetage est évalué dans un trigramme[208] par des tests récursifs binaires par exclusion qui prévoient l'acceptation ou le refus d'une étiquette-partie du discours. Prenons, par exemple, la séquence :

la belle fille

TreeTagger évalue la probabilité que *fille* représente un nom (N) précédé par un déterminant (DET) et un adjectif (ADJ). Par l'acceptation de *belle* et *la* respectivement comme ADJ et DET, l'arbre de décision rejoint une feuille (à savoir un bout d'une branche de cet arbre) contenant des informations, c'est-à-dire les étiquettes et les pourcentages (préétablis) pour déterminer l'annotation morphosyntaxique 'la plus pertinente'. *TreeTagger* choisit ainsi l'étiquette la plus probable (voir l'étiquette ayant le pourcentage de probabilité le plus important) dans le trigramme-contrainte distributionnelle : *p* (N| DET, ADJ).

En même temps, à côté de l'annotation morphosyntaxique la plus probable, *TreeTagger* attribue le lemme le plus convenable par la consultation d'un lexique (restreint) formé de formes, lemmes et suffixes. Malgré la précision de cet annotateur très connu et très exploité, nous avons vérifié manuellement (et heureusement) l'annotation et la lemmatisation de chaque mot graphique (§ 5.2.2.).

[208] En informatique, un *n-gramme* consiste en une séquence consécutive de chaînes de caractères (lettres, chiffres et signes de ponctuation), chaque chaîne étant séparée par des espaces en début et fin. Un *trigramme* consiste ainsi en une séquence consécutive de 3 chaînes de caractères.

5.2.2. *Vérification manuelle : heuristiques de résolution*

Comme nous l'avons précisé au § 1.2.6., nous avons ignoré, au premier abord, la composition et toutes les formes de figement. Tout jugement sur la polylexicalité a suivi l'attribution d'une partie du discours à chaque mot graphique de nos parémies. Par conséquent, en ce qui concerne des pronoms composés, comme :

quelque chose
tout le monde

chaque mot graphique a fait l'objet d'une description atomique, pour reprendre la métaphore en chimie que M. Gross a utilisée pour décrire ses grammaires locales (§ 4.2.2.) (et nous ajouterions même à la façon de la grammaire traditionnelle qui est, d'ailleurs, un des points de départ du Lexique-Grammaire). Par conséquent, nous aurons l'annotation morphosyntaxique suivante :

quelque ADJ
chose N

tout ADJ
le DET:ART
monde N.

Ce n'est qu'au terme de notre classification lexico-grammaticale que nous avons reconstitué la polylexicalité et réfléchi sur la valeur globale (ou moléculaire) lexicale et syntaxique conventionnalisée par l'usage. Plus précisément, nous nous appuierons sur tous les DELAF et sur les DELAC disponibles dans *Unitex 3.1* (§ 6.2.) pour l'évaluation de la polylexicalité, notamment aux mots graphiques simples et composés encodés dans ces dictionnaires et reconnus dans notre corpus parémique.

Vu l'absence d'une grammaire (contemporaine) des proverbes français (§ Conclusions) ainsi que pour éviter les biais phrastico-propositionnels qui caractérisent

quelques grammaires de référence (§ 1.1.1.), nous avons établi quelques heuristiques pour faire face à des doutes, voire des fautes d'annotation de la part du *TreeTagger*.

En raison de la nature parémiographique de notre liste, nous avons consulté les ouvrages lexicographiques liés au développement même de *DicAuPro*. Plus précisément, nous avons fait recours au dictionnaire de *Littré*[209] en tant que témoin parémiographique-pivot des formes canoniques (§ 5.1), ainsi qu'au *Trésor de la Langue Française informatisé* (TLFi)[210], parce qu'il constitue un des témoins régulièrement dépouillés par l'équipe *DicAuPro* pour le repérage des variantes et des dates d'attestation. Dans ces dictionnaires, les proverbes sont souvent mentionnés comme exemples pour éclairer un comportement syntaxique ou l'usage d'un constituant dans une combinatoire polylexicale, à savoir comme séquences formulaires sous l'acception d'un lemme. Outre à garder un choix en continuité avec le protocole de *DicAuPro*, la consultation des dictionnaires nous a également servi pour observer la normalisation et l'encodage lexicographique réservé à certains constituants de nos parémies.

Pour autant que possible, nous avons également pris en compte le développement morphosyntaxique en diachronie des proverbes grâce aux historiques soignés par les auteurs de *DicAuPro*. L'observation de la reconstruction philologique au fil des siècles nous a (parfois) donné des indices pour mieux comprendre la partie du discours à préférer. Encore en diachronie, nous avons fait recours au *Dictionnaire du Moyen Français* (DMF) en ligne[211] ainsi qu'à l'archive *Gallica*[212] pour la consultation de textes anciens.

Pour satisfaire le côté contextualiste de notre approche, de temps en temps, nous avons consulté le corpus *frWaC* (Baroni *et al*. 2009) d'environ 1 milliard et 600 millions d'occurrences aspirées du Web à l'aide du logiciel en ligne *Sketch Engine* (Kilgarriff *et al*. 2004). Nous nous sommes appuyé sur ce corpus non seulement pour sa taille, mais aussi pour le fait qu'il est annoté et lemmatisé par *TreeTagger*. Ce qui nous a permis d'effectuer des contre-vérifications quant à la justesse ou aux manques d'étiquetage dans notre corpus parémique.

[209] La version XML en ligne du dictionnaire est à l'adresse : http://www.littre.org/ (date de consultation : 20/11/2013).
[210] Le TLFi est disponible à l'adresse suivante : http://atilf.atilf.fr/ (date de consultation : 20/11/2013).
[211] Le dictionnaire se trouve à la page : http://www.atilf.fr/dmf/ (date de consultation : 20/11/2013).
[212] La bibliothèque numérique de la *Bibliothèque Nationale de France* est consultable à l'adresse : http://gallica.bnf.fr/ (date de consultation : 20/11/2013).

5.2.3. *Annotation morphosyntaxique : cas exemplaires*

TreeTagger a souvent mal étiqueté notre corpus parémique. Sur un total de 1.689 formes canoniques annotées [213], 707 (environ 42%) contiennent au moins une faute d'annotation morphosyntaxique.

La performance décevante de l'annotation morphosyntaxique, semble suggérer deux observations majeures : l'une sur la nature linguistique des parémies, l'autre sur l'annotateur automatique.

En ce qui concerne la nature linguistique des parémies, il paraît qu'elles possèdent une distribution des parties du discours assez 'improbable', à savoir divergente par rapport à l'usage normalement codifié. Plutôt de qualifier ces agencements distributionnels syntaxiques comme archaïsants (comme le ferait d'ailleurs la plupart des parémiologues), nous préférons souligner qu'ils sont des agencements qui caractérisent les parémies. Ils coparticipent éventuellement à la *parémisation* d'une séquence formulaire (§ 4.3.1.), voire à la perception du statut parémique d'une séquence formulaire.

Quant à l'annotateur, il est assez évident que la performance douteuse d'annotation morphosyntaxique invite à prendre en considération la possibilité de mettre au point un annotateur *ad hoc* pour les parémies (§ Conclusions). Ce qui prouve encore la nécessité de remettre en question les évaluations traditionnelles réservées aux *f* des parémies sur corpus. Comment pourrait-on bien mesurer l'impact des parémies dans le sac de mots qu'est le corpus après la tokenisation informatique, si leurs particularités linguistiques passent sous silence par les outils informatiques eux-mêmes ?

Par la suite, nous aborderons quelques cas exemplaires parmi ceux que nous avons rencontrés au moment de notre correction manuelle. Il ne faudra pas oublier que, malgré nos redressements :

> « la représentation est arbitraire [et] doit être prise comme une représentation par simple séquence de catégories […] » (Gross 1990 : 69, 71)[214].

[213] L'annotation automatique concerne seulement les formes canoniques de notre *lβ*. En raison du nombre élevé de fautes, nous avons préféré procéder à l'annotation manuelle des 87 formes canoniques restantes, tout en veillant au respect du jeu d'étiquettes du *TreeTagger*. Par conséquent, l'estimation quantitative des fautes exclut ces dernières 87 formes canoniques.

[214] Nous remercions Michele De Gioia pour cette suggestion.

5.2.3.1. Fautes d'annotation morphosyntaxique en position initiale

En général, nous mentionnons les fautes d'annotation qui concernent les parémies commençant par la préposition (PRP) *À*, comme :

(1) *À chacun son tour*[215]

TT	VER:pres	avoir	PRO :IND	chacun	DET :POS	son	NOM	tour
CM	PRP	à						

où PRP_*À*[216] est étiquetée comme verbe (VER) au présent de l'indicatif (:pres). Au premier abord, nous avons cru que le manque de diacritique sur l'initiale dans notre corpus parémique aurait influencé le choix probabiliste. Le cas de :

(2) *A beau mentir qui vient de loin*

TT	PRP	à	ADJ	beau	VER:infi	mentir	PRO:REL	qui
CM	VER:pres	avoir	ADV					

a réfuté notre hypothèse. Comme on lit, *TreeTagger* a préféré la contrainte distributionnelle PRP-ADJ en lieu et place de la description lexico-grammaticale la plus appropriée que nous avons proposée. Or, il est évident qu'autant la position initiale de PRP_*À* que du VER:pres_avoir déstabilise l'annotateur qui, par défaut, sélectionne le choix ayant le poids (statistique) le plus important. On comprend que cette attribution immédiate n'est pas toujours applicable à l'annotation morphosyntaxique des parémies.

Toujours en position initiale, nous avons relevé des fautes d'annotation des VER à l'impératif (:imp), souvent signalés comme VER:pres. Par exemple :

(3) *Chassez le naturel, il revient au galop*

[215] Dans les tableaux que nous présenterons, la première ligne correspond à l'annotation morphosyntaxique et à la lemmatisation proposée par *TreeTagger* (TT). La deuxième ligne inclut signale à quel endroit nous sommes intervenus pour apporter notre correction manuelle (CM) et ce que nous avons corrigé.
[216] Nous utiliserons le tiret bas pour indiquer le couple partie du discours_lemme auquel nous faisons référence.

TT	VER:pres	chasser	DET:ART	le	ADJ	naturel	PUN	,
CM	VER:imp				NOM			

où l'impératif en position initiale et l'absence d'autres constituants (n-grammes) précédents pour pondérer le parcours d'annotation, ont abouti à la décision d'un étiquetage VER:pres qui n'a pas de pertinence lors d'une lecture contextuelle, à savoir dans l'ensemble de la séquence lexico-grammaticale.

Au passage, nous faisons remarquer que la forme *naturel*, nous l'avons catégorisée comme NOM, et ce, d'après le TLFi, d'une part, ainsi que d'après la consultation du *Word Sketch* de *chasser* dans le *frWaC* sous *SketchEngine*. La relation grammaticale *objet* qui est définie, entre autres, par le patron VER-NOM, confirme la collocation entre *chasser* et *naturel* (38 occurrences), ce dernier étant donc tagué comme NOM.

Après ce bref détour et pour montrer que toute personne de l'impératif est affectée par un étiquetage non pertinent, nous donnons également l'exemple (4) :

(4) *Dis-moi qui tu hantes, je te dirai qui tu es*

TT	VER:pres	dire	PRO:PER	moi	PRO:REL	qui	PRO:PER	tu
CM	VER:imp							

où le tiret entre VER et le pronom personnel (PRO:PER) n'a pas aidé à mieux calibrer l'étiquetage.

Le pronom *nul* en position initiale fait également l'objet d'une annotation non équilibrée, comme dans :

(5) *Nul n'est prophète en son pays*

TT	ADJ	nul	ADV	ne	VER:pres	être	NOM	prophète
CM	PRO:IND							

c'est-à-dire qu'il est toujours indiqué (qu'il soit en position initiale ou ailleurs dans la séquence) comme ADJ, non pas comme pronom indéfini (PRO:IND).

Un sort informatique pareil concerne plusieurs pronoms, surtout quand ils se trouvent position initiale. C'est le cas de *personne* qui est toujours reconnu comme NOM :

(6) *Personne n'en est revenu dire des nouvelles*

TT	NOM	personne	ADV	ne	PRO	en	VER:pres	être
CM	PRO:IND							

et de *rien*, à chaque fois restitué comme NOM :

(7) *Rien ne se perd, rien ne se crée*

TT	NOM	rien	ADV	ne	PRO:PER	se	VER:pres	perdre
CM	PRO:IND							

Ce qui suggère non seulement qu'il pourrait y avoir des pondérations statistiques à régler en termes de contraintes distributionnelles, mais aussi qu'il serait nécessaire d'apporter quelques retouches au lexique.

Un cas particulier d'annotation pronominale est représenté par les parémies commençant par *qui*, parfois pronom relatif (PRO:REL), parfois tout court pronom (PRO). L'annotation varie également pour les parémies où la collocation (§ 2.2.6.) initiale est la même, comme pour :

(8) *Qui va lentement, va sûrement*

TT	PRO	qui	VER:pres	aller	ADV	lentement	PUN	,
CM	PRO:REL							

et :

(9) *Qui va sans barbe et tout nu au vent de bise est morfondu*

TT	PRO:REL	qui	VER:pres	aller	PRP	sans	NOM	barbe
CM								

Comme on peut lire, nous avons harmonisé l'étiquetage et attribué la partie du discours PRO:REL.

La position initiale montre encore les cas étranges d'annotation de *tel* et *tout*. Quant au premier, qu'il occupe la position initiale de la séquence ou une autre, il n'est qu'un pronom démonstratif (PRO:DEM), alors qu'il peut être ADJ, comme dans (10) :

(10) *Tel maître, tel valet*

TT	PRO:DEM	tel	NOM	maître	PUN	,	PRO:DEM	tel
CM	ADJ						ADJ	

ou PRO:IND, comme dans (11) :

(11) *Tel est pris qui croyait prendre*

TT	PRO:DEM	tel	VER:pres	être	VER:pper	prendre	PRO:REL	qui
CM	PRO:IND							

Quant au deuxième, *TreeTagger* lui accorde de manière quasi systématique l'étiquette PRO:IND plutôt que satisfaire la contrainte distributionnelle ADJ-NOM. À ce propos, nous mentionnons :

(12) *Tout songe est mensonge*

TT	PRO:IND	tel	NOM	songe	VER:per	être	NOM	mensonge
CM	ADJ							

Assez curieusement et à l'inverse, *TreeTagger* rate l'étiquetage de *tout* comme PRO:IND pour l'annoter comme adverbe (ADV) :

(13) *Tout se fait avec le temps*

TT	ADV	tout	PRO:PER	se	VER:per	faire	PRP	avec
CM	PRO:IND							

Tout compte fait, nous avons pu apporter ces corrections (décidément chronophages), mais on a beau penser que de telles fautes pourraient intéresser des corpus de taille beaucoup plus grosse que le nôtre. Ce qui ne va pas sans conséquence sur les généralisations qu'on pourrait retirer de l'annotation morphosyntaxique . Autant vaudrait tout étiqueter et prendre le risque de l'ambiguïté, comme le fait le prétraitement linguistique sous *Unitex 3.1* (§ 6.2.1.) pour proposer des parcours adaptés de désambiguïsation, plutôt que de limiter le choix d'annotation (et le jeu d'étiquettes, aussi), sans garantir la précision et freiner la recherche. C'est ce débat que nous approfondirons dans le paragraphe suivant.

5.2.3.2. Tout autour *de* : déterminants adverbiaux

TreeTagger a alloué l'étiquette PRP à chaque occurrence des formes *d'*/*de*, sans aucune distinction. Néanmoins, cette annotation uniforme soulève des questionnements. Prenons ainsi les parémies :

(14) *À sotte demande, point de réponse*
(15) *Il est plus d'ouvriers que de maîtres*
(16) *Il y a beaucoup d'appelés, mais peu d'élus*
(17) *Les malheureux n'ont point de parents*
(18) *Trop de précaution nuit*

La *Syntaxe du nom* de Maurice Gross (1986) donne la même réponse du *TreeTagger* : toutes les occurrences des formes *d'*/*de* sont PRP. Le souci taxinomique et classificatoire qui a caractérisé son approche, amène Gross à reconnaître trois propriétés syntaxiques

distributionnelles fondamentales pour ranger tous les déterminants (*Dét*[217]). Il nous éclaire comme suit (1986 : 17) :

a. *Dét* se combine avec *N*, d'où *Dét N* ;
b. *Dét* se combine avec *GN* par la préposition *de* ; à cet égard, il précise que :
    i. dans les cas où *GN* serait pluriel et nombrable, *Dét* est à interpréter « comme un sous-ensemble propre de *GN* » (*ibid*.) qui est le domaine de définition de *Dét* ;
    ii. dans les cas où *GN* serait abstrait et/ou masculin, *Dét* ne peut être toujours interprété comme sous-ensemble de *GN*, d'où la nécessité d'établir d'autres distinctions ;

    Peu importe (i) ou (ii), la propriété syntaxique reste *Dét de GN* ;
c. *Dét* fonctionne comme *Adv*, ce qui permet la constitution de la phrase[218] N_0 *V Dét*.

Des quatre classes de déterminants, nous nous concentrons seulement sur celle des *déterminants adverbiaux* (*Dadv*) :

> « qui comprend les *Dét* ayant les propriétés *Dét de GN* et N_0 *V Dét*, mais qui n'ont pas la propriété *Dét N* » (1986 : 18).

C'est dans cette classe que Gross inclut, entre autres :

(*beaucoup* + *peu* + *plus* + *point* + *trop*) *de*

et que nous repérons dans (14)-(18).

Essayons d'approfondir la propriété distributionnelle *Dét de GN* par rapport à (14)-(18) et à notre approche, sans oublier une prise de position cruciale de Gross : la description syntaxique base l'identification de tout sens. Autrement dit : le formalisme syntaxique avant toute chose (1986 : 19).

Ce qui nous dérange de la propriété reconnue par Gross consiste en sa vision synthétique de *GN* et le parti pris sur le statut de préposition de la forme *de*. Plus précisément, le *GN* ne

[217] En l'occurrence, nous reprenons le formalisme propre au Lexique-Grammaire par respect de la méthode.
[218] Nous rappelons que la *phrase* est à concevoir comme « unité du lexique » et dans la perspective harrisienne.

convient pas à notre approche volontairement atomique et qui tend, seulement par la suite, à une approche moléculaire (§ 5.2.1.). Quant à *de*, nous sommes influencé par la grammaire traditionnelle qui nous a inculqué le *de*-déterminant partitif. Mais approfondissons encore le point de vue de Gross sur le *GN* et le *de* des *Dadv*. Il écrit :

> « […] le caractère adverbial […] (c'est-à-dire leur mobilité) et l'existence des partitifs *de Ddéf* […] suggère une analyse […] différente […] » (1986 : 54-55).

Par le recours à la construction *Dét de Ddéf N*[219] et d'exemples phrastiques autour de :

beaucoup de ces personnes,

il évalue la faisabilité de traiter *beaucoup* comme *Adv* de type général, ce qui s'aligne sur la propriété *N_0 V Dét* qui appartient aux *Dadv*, et de traiter *GN = de Ddéf N*. Par des arguments distributionnels (*ibid*.), il rejette ce traitement parce que les spécificités distributionnelles observées pour *beaucoup* vont à l'encontre de la mobilité des *Adv*. Cela équivaudrait à créer un nouveau type d'*Adv* (1986 : 56). Nous ajoutons très gentiment qu'il s'agirait, en tout cas, d'un *Adv* et que l'exigence d'une formalisation syntaxique en vue d'une classification (pour autant que possible) économe (Gross 1976 : 19) ne peut ignorer. Voilà donc une première réponse : exemples et contre-exemples, propriétés distributionnelles taxinomiques généralisables et/ou spécifiques nous guident vers l'attribution de l'étiquette ADV à :

(*beaucoup* + *peu* + *plus* + *point* + *trop*)

qui suit d'ailleurs la grammaire traditionnelle. Que faire de *de* ? La réponse est encore à Gross, non pas dans l'attribution du statut de préposition, mais dans la lecture partitive présumée par b.i. ci-dessus et qui s'applique très bien aux (14)-(18). Par conséquent, on ne qualifiera pas comme PRP, mais comme déterminant indéfini. Comme le jeu d'étiquettes de *TreeTagger* ne dispose pas de cette étiquette, on l'identifiera tout simplement comme DET. Nous nous alignons encore sur la grammaire traditionnelle, mais aussi sur la description syntaxique encodée dans les DELAF conçus dans le cadre du Lexique-Grammaire. Ce que nous suspendons est le jugement sur la composition polylexicale qui gomme (de par sa nature)

[219] *Ddéf* abrège 'déterminants définis'.

l'analyse morphosyntaxique de chaque constituant de nos parémies. Notre parémie en (14) et toutes les autres de notre corpus parémiques qui possèdent un *Dadv* sont ainsi traitées comme suit :

(14) *À sotte demande, point de réponse*

TT	PUN	,	NOM[220]	point	PRP	de	NOM	réponse
CM			ADV		DET			

5.2.3.3. Tout autour *de* : déterminant indéfini

L'abordage des *Dadv* nous a permis d'affranchir le déterminant indéfini et, par conséquent, d'entamer la revue systématique de toutes les occurrences de la forme *de* pour harmoniser notre description dans l'intégralité de notre corpus parémique. Ainsi :

(19) *Qui gagne du temps gagne tout*

TT	PRO:REL	qui	VER:pres	gagner	PRP	de	NOM	temps
CM					DET	du		

où l'annotation PRP et la lemmatisation *de* sont remplacés respectivement par DET et sa même forme fléchie à la manière proposée par le DELAF français.

Comme pour les *Dadv*, nous avons laissé de côté (pour l'instant) tout type de figement et donc pour la parémie :

(20) *Quand on prend du galon, on n'en saurait trop prendre*

TT	PRO:PER	on	VER:pres	prendre	PRP	de	NOM	galon
CM					DET	du		

[220] *TreeTagger* a toujours étiqueté *point* comme NOM.

nous avons réaménagé l'annotation DET et la lemmatisation *du* et avons ignoré l'expression figée, *prendre du galon* qui, dans le domaine militaire, indique un avancement de grade (TLFi, s.v. '*galon*').

Dans les cas, entre autres, des parémies suivantes :

(21) *La rivière ne grossit qu'il n'y entre de l'eau trouble*

(22) *On ne saurait retenir le chat quand il a goûté de la crème*

(23) *Tant qu'il y a de la vie, il y a de l'espoir*

c'est-à-dire quand le déterminant indéfini et suivi d'un article, nous annoterons DET-DET:ART et lemmatiserons respectivement par *de* et *le*. Par exemple, pour (21), nous avons :

(21) *La rivière ne grossit qu'il n'y entre de l'eau trouble*

TT	PRP	de	DET:ART	le	NOM	eau	ADJ	trouble
CM	DET							

Pour les déterminants indéfinis utilisés contextuellement à une négation, nous adoptons encore l'annotation DET et le lemme *de*, comme suit :

(24) *Il n'y a pas de feu sans fumée*

TT	PRP	de	NOM	feu	PRP	sans	NOM	fumée
CM	DET	de						

Les déterminants indéfinis pluriels sont signalés comme DET, mais lemmatisés comme *un*. À titre d'exemple :

(25) *Chantez à l'âne, il vous fera des pets*

TT	PRO:PER	vous	VER:fut	faire	PRP:det	de	NOM	pet
CM					DET	un		

5.2.3.4. ADJ ou NOM : *Impossible n'est pas français*

Notre corpus parémique inclut :

(26) *Impossible n'est pas français*

que le TLFi attribue à Napoléon et identifie par la marque d'usage [*Phrase hist*.] (s.v. '*impossible*'). *TreeTagger* a étiqueté *impossible* et *français* comme ADJ. Néanmoins, dans la même acception du TLFi, cet aphorisme passé en parémie est suivi de la citation :

> *Et voilà comme, au régiment plus qu'en nul autre lieu du monde, « impossible » n'est pas français* (COURTELINE, Gaîtés esc., Je m'en fous, 1886, I, p. 201 ; c'est nous qui soulignons).

On observe ainsi l'usage des guillemets autour de *impossible* évoqué, dans ce contexte, comme forme (*type*) ou lemme, non pas comme occurrence (*token*). L'adjectif est donc nominalisé. Cette lecture semble validée par *DicAuPro*, notamment dans la variante :

(27) *Le mot « impossible » n'est pas français*.

C'est toujours *DicAuPro* qui nous dévoile la première attestation napoléonienne :

> *Ce n'est pas possible, m'écrivez-vous ; cela n'est pas français* (Napoléon I^{er}, Lettre au général Demarois, 1813)

L'attestation originaire de Napoléon I^{er} ne prévoit pas l'adjectif *impossible*, mais son antonyme *possible*. Il en suit que (26) est le résultat d'une crase dans l'usage de deux propositions précédemment séparées où *impossible* et *français* étaient ADJ. De façon conservative par rapport à l'attestation napoléonienne et par rapport au *TreeTagger*, nous avons ainsi gardé l'annotation ADJ. Néanmoins, nous prévoyons une deuxième description où *impossible* est annoté comme NOM.

5.2.3.5. ADV ou NOM : *Peut-être garde les gens de mentir*

Un cas similaire à celui de § 5.2.3.4. est représenté par :

(28) *Peut-être garde les gens de mentir*

où *TreeTagger* attribue à *peut-être* l'étiquette de ADV. L'appel au lemme, notamment à la sémantique de cette forme, nous a conduit à proposer également l'annotation NOM. Dans ce cas, aussi, nous avons deux descriptions à tester.

5.2.3.6. VER ou NOM : *Couche-toi sans souper et tu te trouveras le matin sans dettes*

La résolution d'ambigüité n'est pas toujours supportée par nos heuristiques. C'est le cas de :

(29) *Couche-toi sans souper et tu te trouveras le matin sans dettes*

notamment du mot *souper* que *TreeTagger* étiquette comme VER à l'infinitif (:inf). Il est légitime de croire que *souper* pourrait être aussi un NOM, cette forme étant courante dans les variétés de français belge et québécoise, mais délaissée en français hexagonal.
L'analyse de *se coucher sans souper* ne suggère pas d'indices de désambiguïsation, quoique le TLFi la cite s.v. '*souper*' en tant que VER, sans donner aucune explication.
Un regard à la diachronie nous laisse également dans le doute. Malgré quelques variations morphosyntaxiques, *DicAuPro* confirme l'usage en diachronie de *se coucher sans souper* et la conclusion de la parémie par le patron PRP-DET:ART-NOM.
La vulgate – et la tentation – de la binarité symétrique invite à préférer l'hypothèse NOM à celle de VER. Nous avons choisi de retenir les deux. Nous avons ainsi accepté VER proposé par *TreeTagger* et encodée également l'annotation NOM qu'on ne peut rejeter, en principe, ni en synchronie, ni en diachronie. Ce choix implique l'acceptation de deux annotations et lemmatisations.

5.2.3.7. *Tarde qui tarde, en avril aura Pâques* : parties du discours et figement en diachronie

Comment annoter (et lemmatiser, par la suite) la séquence *tarde qui tarde* ? Le *TreeTagger* nous donne :

(30) *Tarde qui tarde, en avril aura Pâques*

TT	VER:pres	tarder	PRO:REL	qui	VER:pres	tarder	PUN	,

La synchronie ne nous aide pas à évaluer de la pertinence de cette proposition : *frWaC* ne renvoie aucune occurrence.
Le repérage de l'occurrence :

« Tarde qui tarde, venait le soir. »

dans *Gallica*, notamment dans l'œuvre littéraire *Les oisivetés du sieur du Puitspelu, Lyonnais* par Clair Tisseur au XIX^e^ siècle nous suggère que *tarde qui tarde* aurait pu être un adverbe figé et désormais désuet dans l'usage courant. Pour mieux comprendre la nature de ses constituants, nous avons fait recours à la diachronie[221] et consulté *DicAuPro*. La fiche nous a montré une variante attestée au XIV^e^ :

Tarde que tarde, en avril auras Pâques

Notre parémie remonte (au moins) au Moyen Âge. La consultation lexicographique doit ainsi intéresser des ouvrages lexicographiques qui décrivent cette période. Dans le DMF en ligne, nous découvrons s.v. '*tarde*' qu'il est NOM et signifie « retard ». Faudrait-il donc accepter le tag NOM ? Le dépouillement du DMF, notamment s.v. '*tarder*' à l'acception III.B.2., confirme, au contraire, la fixité de la séquence *tarde que tarde* qu'il qualifie comme *locution* (et nous ajoutons adverbiale), mais surtout la nature verbale de la forme *tarde*. Il reste à comprendre le temps et le mode de ce VER. Si nous lisons l'exemple du DMF :

« *...je pense qu'en ce chemin L'ait mené ; pour ç'attenderay Son retour et si m'en yray, Tarde que tarde* (Miracle de saint Panthaleon., 1364, 322) »

et pensons surtout à ce *que* déformé en *qui* par l'usage des siècles, c'est le subjonctif qui vient à l'esprit. Plus précisément, ce sont des séquences formulaires, comme *coûte que coûte* ou *vaille que vaille*, qui sont équivalentes de par leur distribution syntaxique ainsi que de par

[221] À ce propos, nous remercions Mirella Conenna, Michele De Gioia et Stefania Marzano pour leurs suggestions.

leur fonction adverbiale. L'affinité syntaxique est encore confirmée par le DMF qui inclut *vaille que vaille* s.v. '*valoir*' à l'acception I.B.1. Qui plus est, les deux séquences sont présentes à l'écrit en synchronie : *coûte que coûte* totalise 1.163 occurrences dans *frWaC*, alors que *vaille que vaille* en compte 303.

En raison de ce parcours en diachronie et en synchronie, nous envisageons la description syntaxique suivante :

TT	VER:pres	tarder	PRO:REL	qui	VER:pres	tarder	PUN	,
CM	VER:subp		KON	que	VER:subp			

où les formes *tarde* correspondent à VER au subjonctif présent (:subp) et la forme *qui* est plutôt considérée, de manière conservative, une déformation de la conjonction (KON) *que*. En d'autres termes, notre description syntaxique a restauré la variante du XIV^{e} siècle, et ce, sur la base de séquences formulaires assimilable du point de vue de la syntaxe et en usage au moment de notre étude.

5.2.4. *Lemmatisation : cas exemplaires*

Sur le total de 1.689 formes canoniques lemmatisées, 312 (environ 18%) présentent au moins une faute de lemmatisation. Comme on a vu au § 5.2.3., quelques fautes d'annotation morphosyntaxique ont comporté en même temps des fautes de lemmatisation. Dans ce paragraphe, nous souhaitons compléter l'éventail des fautes de lemmatisation relevant directement du lexique employé par *TreeTagger*.

5.2.4.1. Lexique désuet

Notre annotateur automatique a parfois lemmatisé avec le tag :

<*unknown*>

des unités lexicales désuètes. Par exemple, pour :

(31) *À brusquin brusquet*

TT	PRP	à	NOM	<unknown>	ADJ	brusquet		
CM			ADJ	brusquin				

brusquin n'a pas fait l'objet de lemmatisation. À ce propos, l'interrogation de *frWaC* ne renvoie aucune occurrence de cette forme. Ce qui nous suggère sa *f* éventuelle au moment de l'interrogation de nos corpus (§ 6.1.). D'après le TLFi, s.v. '*brusquet*', autant *brusquin* que *brusquet* dérivent de l'ADJ *brusque*, mais le TLFi n'a pas d'entrée pour *brusquin*. Malgré cette explication, nous avons attribué le lemme *brusquin*.

Dans le cas de :

(32) *Cheval faisant la peine ne mange pas l'aveine*

TT	VER:pres	manger	ADV	pas	DET:ART	le	NOM	<unknown>
CM								aveine

c'est *aveine* qui n'est pas lemmatisé. Aucune entrée lui est réservée dans le TLFi, alors que le *Littré* mentionne clairement que cette forme tombe en désuétude (6 occurrences dans *frWaC*[222]) et a cédé la place à la forme la plus récente *avoine* (3.429 occurrences). C'est en tout cas à *Littré* qui revient cette forme canonique. Par fidélité au choix de l'équipe *DicAuPro* et à leur témoin parémiographique-pivot, nous avons choisi le lemme *aveine*.

5.2.4.2. Lexique courant non lemmatisé

L'annotation <*unknown*> intéresse aussi des unités lexicales courantes en français. Prenons :

(33) *Epargne de bouche vaut rente de pré*

TT	NOM	<unknown>	PRP	de	NOM	bouche	VER:pres	valoir
CM		épargne						

[222] Sauf indication explicite, toutes les occurrences que nous reprenons de *frWaC* se réfèrent aux *f* brutes qui résultent de la recherche du lemme, non pas de la forme exacte.

où, contrairement à ce qu'on aurait cru, ni *rente* (15.706 occurrences *dans frWaC*) ni *pré* (26.916 occurrences) n'ont créé des problèmes d'annotation, mais *épargne* (31.469 occurrences comme NOM) pour lequel nous avons inséré manuellement le lemme.

Décidément imprévue la méconnaissance du pronom *soi-même* dans :

(34) *On ne trouve jamais meilleur messager que soi-même*

TT	ADJ	meilleur	NOM	messager	KON	que	ADJ	<unknown>
CM							PRO:PER	soi-même

qu'on pourrait expliquer autant par l'annotation morphosyntaxique incorrecte que par la taille restreinte du corpus français d'entraînement de *TreeTagger*. D'ailleurs, une interrogation de *frWaC* de la forme *soi-même* avec une contrainte de recherche sur la partie du discours (PRO) ne renvoie aucune occurrence. En revanche, l'interrogation sans contrainte produit 20.283 occurrences. Ce qui suggère évidemment un défaut de l'annotateur.

5.2.4.3. Entités nommées

La quasi-totalité des entités nommés (NAM) a fait l'objet de la même méconnaissance de la part de *TreeTagger*, qu'il s'agisse des saints du calendrier :

(35) *À la Sainte-Catherine tout bois prend racine*

des villes françaises moins mentionnées, comme *Aubervilliers* (non pas de *Paris*) :

(36) *Chou pour chou, Aubervilliers vaut bien Paris*

ou encore de certains gentilés, comme *Gascon* dans :

(37) *Garde-toi d'un Gascon ou Normand ; l'un hâble trop et l'autre ment*

mais non pas *Normand*, reconnu comme NAM et lemmatisé. Au passage encore pour (36), on souligne que le verbe *hâbler* (seulement 15 occurrences dans *frWaC*) est tagué correctement.

5.2.4.4. Double lemmatisation

Pour conclure, nous nous attardons sur des doubles lemmatisations proposées par *TreeTagger*. Pour la parémie :

(38) *Nous sommes tous parents en Adam*

TT	PRO:PER	nous	VER:pres	sommer | être	PRO:IND	tout	NOM	parent | parents
CM				être				parent

TreeTagger indique, outre *être*, le lemme *sommer* pour *sommes*. Ce qui ne pourrait même pas se justifier par le recours à d'autres temps et modes de la conjugaison de VER_*sommer* pour la 1ère personne plurielle. Pour la forme *parents*, *TreeTagger* donne le lemme autant au singulier qu'au pluriel de manière étrange par rapport aux conventions habituelles de lemmatisation qui préfèrent le singulier.
Contraire à ces conventions est également la double lemmatisation de l'ADJ *folles* de la parémie :

(39) *Vides chambres font dames folles*

TT	NOM	chambre	VER:pres	chambre	NOM	dame	ADJ	folle | fou
CM								fol

ou l'annotateur signale autant le féminin que le masculin. Au plus, on aurait conçu le masculin *fol*. Comme nous croyons que le *TreeTagger* n'a pas été implémenté dans une optique (qui est raisonnable, sensible et sensée) sociolinguistique attentive aux marques de genre (non strictement grammatical) en lexicographie, nous avons retenu seulement le masculin *fol*.

5.3. Classification lexico-grammaticale[223]

Après avoir corrigé manuellement l'annotation et la lemmatisation automatiques de *TreeTagger* (§ Annexe 4), nous entamerons la classification lexico-grammaticale des parémies en faisant recours à notre unité : la *séquence lexico-grammaticale* (§§ 1.2.6., 2.3.) (désormais SLG). Au fur et à mesure, nous discuterons notre démarche de classification et ses étapes per l'illustration de quelques exemples.

Précisons que notre classification se veut un guide lexico-grammatical pour la modélisation des requêtes informatiques de reconnaissance des parémies sur corpus. Plus précisément, la description des SLG fera ressortir, d'une part, ces schémas intégraux que nous avons mentionnés au § 1.1.4. et au § 4. et nous en donnera une énumération précise. D'autre part, les SLG nous aideront à prendre en compte ces amorces encadrant les parémies le plus fréquemment et, par conséquent, susceptibles d'une attention particulière lors de la modélisation de nos requêtes.

À propos de fréquence, la répétition des SLG dans notre corpus parémique nous aidera à mettre en évidence les propriétés lexico-grammaticales prototypiques des parémies-Gestalt linguistiques. Nous essayerons de dévoiler la *mémoire lexico-grammaticale* (Legallois 2009 : 9) que restituera notre corpus parémique qui est un échantillon du répertoire parémiologique francophone. Autrement dit, nous cherchons à extraire une mémoire lexico-grammaticale à partir de l'acquis formulaire francophone.

Notre tentative de fragmenter et d'analyser en constituants les parémies que l'on perçoit et utilise comme des séquences synthétiques et holistiques, veut justement révéler les mécanismes lexico-grammaticaux qui participent à cette recomposition et à cette perception unitaire. Ce n'est que par une approche empirique et descriptive que l'on pourra repérer ces « régularités autonomes des règles *a priori* de constitution (de production) » (Legallois 2009 : 10) qui caractérisent les parémies françaises.

Nous précisons encore que, pour toute étape de la classification, nous avons exploité le tableur *Excel*. Plus précisément, nous avons converti la feuille de calcul contenant les annotations et les lemmatisations sous forme d'un tableau croisé dynamique. Nous avons ainsi classé nos parémies par filtrages successifs sur la base d'une partie du discours et/ou d'une unité lexicale lemmatisée à une position donnée.

[223] Ce paragraphe réélabore les principes et les classes lexico-grammaticales dans Marcon (à paraître).

5.3.1. *Entrée syntaxique et entrée lexicale : classes lexico-grammaticales (niveau 1)*

Le premier niveau de notre classification se fonde sur la partie du discours en position initiale de chaque parémie. Nous l'appelons *entrée syntaxique*. Commençons par l'analyse des entrées syntaxiques que nous regroupons dans le Tableau 19 ci-dessous :

¶Entrée syntaxique	Occurrences
ADJ	96
ADV	122
DET	3
DET:ART	400
DET:POS	1
INT	2
KON	62
NAM	8
NOM	196
NUM	11
PRO	2
PRO:DEM	55
PRO:IND	49
PRO:PER	377
PRO:REL	138
PRP	171
PRP:det	27
VER:futu	1
VER:impe	27
VER:infi	16
VER:pper	2
VER:pres	10
VER:subp	1

Tableau 19. Entrées syntaxiques triées par ordre alphabétique avec leurs *f* respectives dans notre corpus parémique[224].

De ce Tableau 19, il résulte que certaines parties du discours sont censées démarrer la séquence parémique plus que d'autres. DET:ART et PRO:PER s'emparent de 777 parémies, à savoir d'environ 44% de notre corpus parémique. Elles sont deux entrées syntaxiques prototypiques de notre corpus. On entrevoit donc une première raison syntaxique à la base de toute spéculation sur la généricité sémantique de la séquence parémique. Cela est renforcé par l'ensemble des pronoms qui, à eux seuls, regroupent 621 parémies. En revanche, toutes les entrées syntaxiques VER sont celles qui moins caractérisent le début des parémies.

Nous faisons en même temps recours aux unités lexicales lemmatisées qui actualisent les entrées syntaxiques Chacune de ces unités lexicales nous la définissons *entrée lexicale*. Pourtant, seulement les entrées lexicales dont $f > 1$ dans notre corpus parémique permettront la reconnaissance d'une classe lexico-grammaticale de niveau 1.

Prenons le cas de l'entrée syntaxique DET:ART :

[224] Le total des entrées syntaxiques s'élève à 1.777, et ce, en raison de certains choix descriptifs que nous avons éclairés au § 5.2.

DET:ART	
le	327
un	73
	400

Tableau 20. Entrées lexicales de DET:ART triées par ordre alphabétique avec leurs *f* respectives dans notre corpus parémique.

L'entrée syntaxique DET:ART est actualisée par les lemmes *le* et *un*. L'entrée lexicale prototypique[225] est donc *le*. Néanmoins, comme les deux entrées lexicales ont f > 1 dans notre corpus parémique, nous avons deux classes de niveau 1 : la classe [*Le_Dét*] et la classe [*Un_Dét*] qui comprennent 327 et 73 parémies, respectivement[226]. Ainsi, pour l'entrée syntaxique PRO:PER :

PRO:PER	
il	273
nous	2
on	102
	377

Tableau 21. Entrées lexicales de PRO:PER triées par ordre alphabétique avec leurs *f* respectives dans notre corpus parémique.

nous avons les classes : [*Il_Pro*$_{pers}$] (273 parémies), [*On_Pro*$_{pers}$] (102 parémies) et [*Nous_Pro*$_{pers}$] (2 parémies).

Faisons maintenant le cas le plus complexe de l'entrée syntaxique NOM :

[225] Nous ferons souvent référence à la *prototypicité* d'une entrée lexicale par rapport à une autre ainsi qu'à celle de certaines classes lexico-grammaticales, voir de certaines SLG. Par *prototypicité* nous signifions la centralité d'une entrée, d'une classe ou d'une SLG, c'est-à-dire la capacité d'une d'entre elles de décrire le nombre le plus élevé de parémies par rapport à d'autres entrées, à d'autres classes ou SLG. Les entrées, les classes et les SLG prototypiques agissent ainsi comme des pôles d'attraction des parémies ainsi que d'autres séquences formulaires. Autrement dit, elles sont censées aider à reconnaître plus aisément des parémies et des séquences formulaires dans un corpus.

[226] Désormais, les crochets indiquent le statut de classe lexico-grammaticale. Entre crochets, nous insérons le lemme et une forme abrégée de la partie du discours, les deux unis par un tiret bas. Pour faciliter la lecture des parties du discours, nous simplifierons le jeu d'étiquettes du *TreeTagger* et adoptons la légende suivante : *Adj* = adjectif ; *Adv* = adverbe ; *Dét* = déterminant ; *N* = nom ; *Prép* = préposition ; *Pro* = pronom ; *V* = verbe. Toute précision sera abrégée en indice de chaque partie du discours. Les conversions entre les jeux d'étiquettes sont explicitées dans l'Annexe 3.

NOM	
abondance	1
aile	1
ami	1
amitié	1
amour	1
année	2
août	1
apprenti	1
araignée	1
argent	3
arme	1
assaut	1
attente	1
avril	2
bien	2
bœuf	1
bouche	1
brebis	3
chair	1
chance	1
changement	4
charité	1
chat	1
chêne	1
cherté	1
cheval	2
chien	3
chose	2
chou	1
cœur	2
comparaison	1
contentement	1
cordonnier	2
corsaire	1
couteau	1
crédit	1
crosse	1
croûte	1
déjeuner	1
demain	1
Dieu	7
écu	1
épargne	1
erreur	1
étrennes	1
expérience	1
face	1
fagot	1
familiarité	1
faute	3
faveur	1
femme	2
février	1
fille	1
fin	1
fou	1
fromage	1
gelée	1
goutte	2
grain	1
guerre	1
impossible	1
janvier	2
jeu	1
jeune	1
jeunesse	2
lever	1
liberté	1
mai	3
maille	1
maison	1
maître	1
mal	4
marchand	2
marchandise	1
mariage	2
mars	1
matines	1
médecin	1
mémoire	1
mi-juin	1
mi-mai	1
morceau	1
nature	2
nécessité	5
neige	1
netteté	1
noblesse	2
noce	1
nourriture	1
obéissance	1

œil	1
or	1
pain	3
paix	1
parole	1
pas	1
patience	2
pauvreté	1
pays	1
péché	1
peine	1
peu	1
peut-être	1
pierre	1
plaie	1
pluie	4
provision	1
prudence	1
rage	1
religion	1
renard	1
richesse	1
saison	1
santé	1
secret	1
septembre	2
serein	1
service	1
soldat	1
soleil	1
souhait	1
source	1
souris	1
témoin	1
temps	1
terre	1
tête	1
tricherie	1
usage	1
vache	1
vallée	1
vent	1
ventre	3
vérité	2
vie	1
vigne	1
ville	2
vin	3
voisin	1
voyage	1
	195

Tableau 22. Entrées lexicales de NOM triées par ordre alphabétique avec leurs f respectives dans notre corpus parémique.

Malgré les 197 parémies ayant NOM comme entrée syntaxique, la dispersion des unités lexicales ne crée qu'un nombre limité de classes de niveau 1. Les classes lexico-grammaticales prototypiques de premier niveau sont : [*Dieu_N*] (7 parémies) ; [*Nécessité_N*] (5 parémies) ; [*Changement_N*], [*Mal_N*] et [*Pluie_N*] (4 parémies) ; [*Argent_N*], [*Brebis_N*], [*Chien_N*], [*Faute_N*], [*Mai_N*], [*Pain_N*], [*Ventre_N*] et [*Vin_N*] (3 parémies) ; [*Année_N*], [*Avril_N*], [*Bien_N*], [*Cheval_N*], [*Chose_N*], [*Cœur_N*], [*Cordonnier_N*], [*Femme_N*], [*Goutte_N*], [*Janvier_N*], [*Jeunesse_N*], [*Marchand_N*], [*Mariage_N*], [*Nature_N*], [*Noblesse_N*], [*Patience_N*], [*Septembre_N*], [*Vérité_N*] et [*Ville_N*] (2 parémies). Les entrées lexicales dont $f \leq 1$, notamment leurs parémies, seront classées sous l'entrée syntaxique correspondante, c'est-à-dire sous NOM (§ Annexe 5 pour les autres classes lexico-grammaticales de niveau 1).

5.3.2. *Typologie des séquences lexico-grammaticales : classes lexico-grammaticales (niveaux 2 et suivants)*

Une fois reconnues les 99 classes lexico-grammaticales de premier niveau, nous observons les parties du discours en première position à droite de l'entrée lexicale et constatons, par exemple, pour la classe [*À_Prép*], que *à* est de préférence suivie par :

- *Adj* (31 occurrences) ;
- *N* (24 occurrences) ;
- $Dét_{art}$ (21 occurrences) ;
- *Adv* (5 occurrences) ;
- Pro_{ind} (5 occurrences) ;
- V_{inf} (4 occurrences) ;
- *Nam* (3 occurrences) ;
- Pro_{rel} (1 occurrence).

Notre première *séquence lexico-grammaticale générique* (SLG-g) se compose ainsi :

- de l'entrée lexicale qui définit la classe lexico-grammaticale de niveau 1
- et de la partie du discours en première position à droite dont $f > 1$.

Cette SLG-g constitue la classe lexico-grammaticale de niveau 2. Par conséquent, pour développer notre exemple de [*À_Prép*], nous avons les 7 sous-classes suivantes :

- [*À Adj*][227] ;
- [*À N*] ;
- [*À Dét*$_{art}$] ;
- [*À Adv*] ;
- [*À Pro*$_{ind}$] ;
- [*À V*$_{inf}$] ;
- [*À Nam*].

L'agencement /*À Pro*$_{rel}$/[228] dont $f \leq 1$ est ainsi classé sous [*À_Prép*].

Avant de poursuivre, nous considérons les unités lexicales lemmatisées en première position à droite. Au cas où elles auraient $f > 1$, la *séquence lexico-grammaticale* se dit *actualisée* (SLG-a) par une forme. Elle est, toutefois, *partielle* (SLG-ap) parce qu'elle ne rend compte que d'une partie de la séquence parémique. Pour les 7 sous-classes de [*À_Prép*], on obtient les SLG-ap suivantes :

[*À Adj*]	**[*À N*]**	**[*À Dét*$_{art}$]**	**[*À Adv*]**	**[*À Pro*$_{ind}$]**
{*À bon*} (7)	{*À cheval*} (3)	{*À le*}[229] (20)	{*À mal*} (3)	{*À chacun*} (4)
{*À chaque*} (5)	{*À barbe*} (2)			
{*À tout*} (3)				
{*À dur*} (2)				
{*À mauvais*} (2)				
{*À petit*} (2)				

Tableau 23. Séquences lexico-grammaticales actualisées partielles des sous-classes *[À Adj]*, *[À N]*, *[À Dét*$_{art}$*]*, *[À Adv]* et *[À Pro*$_{ind}$*]* (triées par ordre décroissant de *f*)[230].

[227] Les classes lexico-grammaticales des niveaux 2 et suivants sont toujours indiquées entre crochets. Nous spécifions d'abord l'entrée lexicale qui renvoie à la classe de niveau 1 et, par la suite, nous insérons la(les) partie(s) du discours qui constituent la SLG.

[228] Les agencements des parties de discours et des unités lexicales qui ne correspondent pas à une classe ou à une SLG sont marqués entre barres obliques.

[229] Dans ce cas, nous avertissons qu'on pourrait plutôt parler de la SLG-ap {*À la*}. *TreeTagger* étiquette toute *Prép* suivie par un *Dét*$_{art}$ contracté avec PRP:det.

On arrête quelques instants notre démarche pour souligner que ce sont la cooccurrence et *f* qui gèrent l'interface lexique-syntaxe. En raison du nombre fermé des parties du discours, la syntaxe se charge de faire de l'économie, de généraliser et de ranger toutes les parémies sous un nombre limité de classes. En revanche, le lexique actualise la syntaxe et suggère les formes lemmatisées (encore, par souci de généralisation des résultats) les plus prototypiques de telle ou telle autre distribution syntaxique.

Reprenons notre classification. Pour identifier les autres classes lexico-grammaticales, nous appliquons la procédure suivante de manière récursive. En général, toutes les SLG-g en position n_i (i ≥ 3) dont la partie du discours en position n+1 à droite a $f > 1$, acquièrent le statut de classes. Poursuivons l'exemple de [*À_Prép*]. Plus précisément, pour la classe [*À Adj*] de niveau 2, les cooccurrents continus à droite de *Adj* se divisent entre :

- *N* (29 occurrences)
- *Adj* (2 occurrences).

Nous avons donc les classes [*À Adj N*] et [*À Adj Adj*].

En général, toutes les SLG-a héritent les propriétés distributionnelles syntaxiques de leurs SLG-g respectives en position n_i (i ≥ 3), la partie du discours en position n_i (i ≥ 3) faisant l'objet d'une actualisation lexicale si et seulement si $f > 1$. Pour exemplifier, prenons :

(41) *À chaque fou sa marotte*
(42) *À chaque jour suffit sa peine*
(43) *À chaque oiseau son nid est beau*
(44) *À chaque porc vient la Saint-Martin*
(45) *À chaque saint sa chandelle*

qui relèvent de [*À Adj N*]. La SLG-ap :

{*À chaque N*}

décrit (41)-(45). Si nous continuons la description de SLG-g en position 4, nous obtenons :

[230] Tout type de SLG-a est inséré entre accolades. Entre parenthèses, nous indiquons *f*.

[*À Adj N* $Dét_{pos}$]

qui regroupe (41), (43) et (45), ainsi que :

[*À Adj N* V_{pres}]

qui rassemble (42) et (44). Ces deux SLG-g correspondent aux SLG-ap :

{*À chaque N son*}

pour (41), (43) et (45) où l'unité lexicale lemmatisée en position 4 est le déterminant possessif *son* ayant $f > 1$, alors que :

{*À chaque N* V_{pres}}

est valable pour (42) et (44) où nous n'avons pas une unité lexicale lemmatisée en position 4 dont $f > 1$. Appliquons de manière récursive notre procédure pour caractériser les SLG-g en position 5 de (41)-(45). Nous ne remportons que :

[*À Adj N* $Dét_{pos}$ *N*]

parce que la seule partie du discours en position 5 ayant $f > 1$ est *N*. Quant à SLG-a, on a :

{*À chaque N son N*}

parce qu'aucune unité lexicale lemmatisée en position 5 de (41), (43) et (45) n'a $f > 1$. D'une part, donc (42) et (44) sont assemblés sous la classe lexico-grammaticale de niveau 4 [*À Adj N* V_{pres}] et partagent la SLG-ap {*À chaque N* V_{pres}}. D'autre part, (41), (43) et (45) sont réunis sous la classe lexico-grammaticale de niveau 5 [*À Adj N* $Dét_{pos}$ *N*]. Néanmoins, la SLG-a {*À chaque N son N*} décrit entièrement (41) et (45), mais seulement une partie de (43). Alors {*À chaque* N *son* N} est une SLG-ap pour (43) et une SLG-a *complète* (SLG-ac) pour (41) et (45). Une SLG-ac sature l'interface lexique-syntaxe d'un nombre *p* de parémies d'une classe lexico-grammaticale. Elle devient une matrice descriptive intégrale et créative potentielle de

parémies ainsi que d'autres séquences formulaires. Ce cas très particulier rejoint tous les schémas et les cadres formels que nous avons rencontrés au § 1.2.5. Les autres SLG ne représentent que des matrices descriptives et créatives partielles. Toutes les SLG ainsi dégagées seront modélisées sous forme de requêtes informatiques (§ 6.3.) de façon à saisir la plupart des variations dans l'usage écrit des parémies.

Pour conclure, nous précisons que toute SLG-g décrivant l'intégralité d'au moins 2 parémies dans notre corpus est assimilée à une SLG-ac. Ce qui veut dire que, par exemple, la SLG-g :

[*À Adj N N*]

décrit entièrement les parémies :

(46) *À bon entendeur demi-mot*
(47) *À bon entendeur salut*
(48) *À tout péché miséricorde.*

Nous aurons donc la SLG-g assimilée à une SLG-ac :

{*À Adj N N*}

qui rend compte de (46)-(48) ainsi que de la SLG-ac :

{*À bon entendeur N*}

qui mieux précise (47) et (48).

Toutes les classes lexico-grammaticales et les SLG sont consultables à l'Annexe 6. Nous précisons ici que notre corpus parémique nous a permis d'isoler 160 SLG-ac sur un total de 1.778 parémies. Plus précisément, ces 160 SLG-ac correspondent à 403 parémies, à savoir une moyenne d'environ 2,5 parémies par SLG-ac.

5.3.3. *Rythme syntaxique*

L'analyse des résultats de notre classification, notamment des SLG-g et des SLG-ac, nous montre des régularités d'agencement privilégiées dans chaque classe. Continuons l'exemple de la classe lexico-grammaticale [*À_Prép*]. D'après nos critères de classification, on découvre que [*À_Prép*] est prototypiquement décrite par les 15 SLG-g complètes qui suivent :

[*À Adj*] (31)	**[*À N*] (24)**	**[*À Dét$_{art}$*] (21)**	**[*À Adv*] (5)**	**[*À Pro$_{ind}$*] (5)**
[*À Adj N Adj N*] (13)	[*À N Adj Adj N*] (2)	[*À Dét$_{art}$ N V$_{pres}$ Dét$_{art}$ N*] (2)	[*À Adv V$_{inf}$ Adv V$_{inf}$*] (2)	[*À Pro$_{ind}$ Dét$_{poss}$ N*] (3)
[*À Adj N Adv Prép N*] (4)	[*À N Adj N Adj*] (2)	[*À Dét$_{art}$ N Pro$_{per}$ V$_{pres}$ Dét$_{art}$ N*] (2)		
[*À Adj N N*] (4)	[*À N Adj N V$_{pres}$ Dét$_{art}$ N*] (2)	[*À Dét$_{art}$ Nam Adj N V$_{pres}$ N*] (2)		
[*À Adj N Dét$_{poss}$ N*] (2)	[*À N Adj Pro$_{per}$ Adv V$_{pres}$ Adv Prép N*] (2)			
	[*À N N Prép N*] (2)			
	[*À N Prép N N Prép N*] (2)			

Tableau 24. SLG-g complètes de la classe *[À_Prép]*.

Dans le Tableau 24 ci-dessus, nous remarquons que *V* intervient seulement dans 4 SLG-g. *V* ne semble donc pas vraiment participer de manière significative dans les agencements syntaxiques les plus représentatifs de [*À_Prép*].

Nous observons encore que la répartition des SLG-g complètes de [*À_Prép*] nous aide à mieux saisir la matrice lexico-grammaticale de chaque sous-classe. Par exemple, les 4 SLG-g de [*À Adj*] décrivent la quasi-totalité des parémies concernées (24 sur un total de 31) où [*À*

Adj N Adj N] représente le pivot prototypique par excellence. Dans les 4 SLG-g, *N* se trouve toujours en position finale. Après la préposition, c'est la colligation /*Adj N*/ qu'on reconnaît dans chaque SLG-g. À propos de [*À N*], on constate plutôt qu'après la préposition, c'est la colligation /*N Adj*/ qui caractérise 4 de ses 6 SLG-g. Cette colligation distingue les parémies de [*À N*] et [*À Adj*] : dans cette dernière, elle est absente. On remarque aussi que les 6 SLG-g de [*À N*] assemblent la moitié des parémies sous [*À N*] (12 sur 24), moins par rapport aux 4 SLG-g de [*À Adj*]. Ce qui nous suggère que les agencements syntaxiques saturent de manière différente chaque classe. Chacune d'elles peut soit s'épuiser avec un nombre restreint de SLG-g soit s'éparpiller en agencements syntaxiques non spécifiques. Ce comportement est aussi illustré par les classes [*À* $Dét_{art}$], [*À Adv*] et [*À* $Pro_{ind.}$]. Les deux dernières sont résumées respectivement par une seule SLG-g. Ces SLG-g fonctionnent comme des préférences quasi exclusives pour les agencements syntaxiques de [*À Adv*] et [*À* $Pro_{ind.}$]. En revanche, [*À* $Dét_{art}$] n'accorde pas de préférences particulières à tel ou tel autre agencement. Ses 3 SLG-g regroupent moins d'un tiers des parémies intéressées (6 sur 21)[231].

Il y a ainsi des enchaînements préférés de colligations qui ressortissent de notre description. Ils ne se limitent pas à satisfaire les règles combinatoires de la syntaxe traditionnelle. Ils indiquent plutôt une préférence d'agencement des parties du discours à l'intérieur d'une SLG. De plus, ces enchaînements se répètent dans notre corpus, c'est-à-dire qu'ils ont $f \geq 2$ et ne sont pas à imputer au hasard, mais plutôt caractérisent le *discours parémique canonique*, à savoir l'« acquis de propriétés lexico-syntaxiques se fondant sur une *perception parémiographique* » (Marcon, sous évaluation) des parémies qu'on repère justement dans une source parémiographique. Ce sont ces préférences d'agencement répétées qui explicitent les *rythmes syntaxiques* des parémies et qui qualifient le statut parémique d'une séquence formulaire (voire pseudo-parémique, dans les cas de détournement et d'exploitation des parémies en discours).

Comme on vient de le constater pour [*À_Prép*], certains rythmes peuvent être plus prototypiques que d'autres dans une classe lexico-grammaticale de premier niveau ainsi que dans ses sous-classes. Il s'établit ainsi une hiérarchie de préférences. C'est à ce stade que nous pouvons essayer de généraliser par groupes syntaxiques (GN, GV, etc.). Par exemple, [*À_Prép*] privilégie les agencements syntaxiques qui commencent par [*À Adj*] et qui se

[231] On rappelle que nous avons distingué les prépositions simples de celles qui sont suivies d'un déterminant contracté ($Prép_{det}$). La classe [*À* $Dét_{art}$] réunit toutes les parémies qui commencent par *à la*, *à l'* ou *à un(e)*. Pour les autres formes, voir la classe [*Au_*$Prép_{det}$].

terminent par un *N*. Dans toutes les classes de [*À_Prép*], les rythmes syntaxiques se manifestent :

- pour la classe [*À Adj*] :
  - en général : sans *V* ; quand GN_1 est rempli par /*Adj N*/ ; quand tous les modifieurs précèdent *N* en position finale dans GN_2 ;
  - de préférence, quand GN_1 correspond exactement à GN_2 et seulement si *Adj* précède *N*, comme le montre [*À Adj N Adj N*] qui est la SLG-g prototypique du rythme syntaxique de la classe [*À Adj*] ;
- pour la classe [*À N*] :
  - de préférence : quand GN_1 est occupé par /*N Adj*/ ; quand les modifieurs *Prép* ou /*N Prép*/ précèdent *N* en position finale dans GN_2 de façon spéculaire à GN_1 ; quand GN_1 rempli par /*N Adj*/ est suivi de GV qui comprend V_{pres} ;
- pour la classe [*À* $Dét_{art}$] :
  - de préférence, quand GN_1 est rempli par /$Dét_{art}$ *N*/ et suivi d'un GV ; quand $Dét_{art}$ précède *N* en position finale dans GV et il prévoit un V_{pres} ;
- pour la classe [*À Adv*] :
  - de préférence, quand GV_1 correspond à GV_2 : /*Adv* V_{inf}/ ;
- pour la classe [*À* Pro_{ind}] :
  - de préférence, sans *V* et quand un $Dét_{poss}$ précède *N* en position finale.

Les rythmes syntaxiques confirment ainsi l'existence de combinatoires de proverbialisation et, en général, de parémisation. Ces combinatoires se joignent ainsi aux schémas métriques (Anscombre 2000 ; D'Andrea 2007, 2008) des parémies et interagissent avec tous les autres points de vue qui créent chaque parémie-Gestalt linguistique.

5.3.4. *Rythme lexical*

À propos de l'interaction entre différents points de vue linguistiques, revenons à nos 15 SLG-g et indiquons toutes leurs SLG-ac correspondantes :

[*À Adj*] (31)	
[*À Adj N Adj N*] (13)	{*À bon N bon N*} (3) {*À petit N petit N*} (2) {*À Adj N court N*} (2)
[*À Adj N Adv Prép N*] (4)	{*À Adj N peu de paroles*} (2) {*À Adj N point de N*} (2)
[*À Adj N N*] (4)	{*À bon entendeur N*} (2)
[*À Adj N $Dét_{poss}$ N*] (2)	{*À chaque N son N*} (2)

[*À N*] (24)	
[*À N Adj Adj N*] (2)	//
[*À N Adj N Adj*] (2)	//
[*À N Adj N V_{pres} $Dét_{art}$ N*] (2)	{*À N Adj Dieu V_{pres} le N*} (2)
[*À N Adj Pro_{per} Adv V_{pres} Adv Prép N*] (2)	{*À N Adj il ne V_{pres} Adv de N*} (2)
[*À N N Prép N*] (2)	{*À N N et demi*} (2)
[*À N Prép N N Prép N*] (2)	{*À N de N N de N*} (2)

[*À $Dét_{art}$*] (21)	
[*À $Dét_{art}$ N V_{pres} $Dét_{art}$ N*] (2)	{*À $Dét_{art}$ N V_{pres} le N*} (2)
[*À $Dét_{art}$ N Pro_{per} V_{pres} $Dét_{art}$ N*] (2)	{*À $Dét_{art}$ N on connaître le N*} (2)
[*À $Dét_{art}$ Nam Adj N V_{pres} N*] (2)	{*À $Dét_{art}$ Nam tout N V_{pres} N*} (2)

[*À Adv*] (5)	
[*À Adv V_{inf} Adv V_{inf}*] (2)	{*À Adv V_{inf} bien V_{inf}*} (2)

[*À Pro_{ind}*] (5)	
[*À Pro_{ind} $Dét_{poss}$ N*] (3)	{*À chacun son N*} (3)

Tableau 25. SLG-g suivies de leurs SLG-ac respectives de la classe *[À_Prép]*.

Dans une vue d'ensemble, on constate que la saturation lexicale concerne de préférence les classes et les SLG-g les moins populeuses. Certaines SLG-g tendent à être satisfaites par des SLG-ac où au moins deux parties du discours sont toujours remplies par des unités lexicales précises. C'est le cas de :

{*À Adj N peu de paroles*} (2) et
{*À Adj N point de N*}

pour la classe [*À Adj*], tout comme de :

{*À N Adj Dieu* V_{pres} *le N*} et
{*À N N et demi*}

pour la classe [*À N*]. Ou encore :

{*À* $Dét_{art}$ *N on connaître le N*}

pour la classe [*À* $Dét_{art}$]. Il arrive aussi que les unités lexicales récurrentes concernent seulement une partie du discours *Prép* :

{*À N de N N de N*}

pour la classe [*À N*], ou tout juste $Dét_{art}$, comme dans le cas de :

{*À* $Dét_{art}$ *N* V_{pres} *le N*}.

La répétition de la même unité lexicale à l'intérieur des parémies intéresse seulement *Adj* (*bon*, *petit*) dans la classe [*À Adj*] et {*À N de N N de N*} pour [*À N*]. Il y a donc des *rythmes lexicaux*, c'est-à-dire des actualisations lexicales répétées dans notre corpus qui s'entrelacent avec les rythmes syntaxiques d'une classe. Les rythmes lexicaux peuvent saturer le potentiel d'un rythme syntaxique ou tout simplement indiquer des préférences d'actualisation lexicale. Comme et avec les rythmes syntaxiques, les rythmes lexicaux aussi dévoilent et participent à la reconnaissance et à la création de combinatoires pour la parémisation d'une séquence formulaire.

5.4. En guise de conclusion

Nous avons consacré ce chapitre à la présentation de notre corpus parémique composé de 1.774 formes canoniques. Les soins scientifiques déclarés par l'équipe *DicAuPro* nous assurent un éventail travaillé (et retravaillé) de parémies qui traversent les siècles et dont nous essayerons de quantifier l'usage sur corpus de français contemporain (§ 6).

L'expérience d'annotation et de lemmatisation automatiques, notamment la performance décevante de l'annotation morphosyntaxique, nous a suggéré, d'une part, que les parémies possèdent des agencements distributionnels syntaxiques particuliers. D'autre part, les fautes de *TreeTagger* incitent l'implémentation d'un annotateur pour les parémies, voir l'intégration d'un module spécialisé dans des annotateurs existants (§ Conclusions).
Outre à redresser des fautes évidentes, la correction manuelle a laissé des traces de notre intervention, quoique nous ayons veillé à mitiger notre subjectivité par des heuristiques diverses, en synchronie et en diachronie.

La classification lexico-grammaticale que nous avons développée s'ajoute aux classifications mentionnées au § 1.2. C'est une autre proposition pour aborder l'interface lexique-syntaxe des parémies par une approche à la fois distributionnelle et contextualiste, où la cooccurrence et la fréquence interviennent au cours de l'aménagement des parémies en classes.

Les types de SLG que nous avons identifiés servent à mieux souligner les propriétés lexicales et syntaxiques prototypiques des parémies (en l'occurrence, de notre corpus parémique). Autrement dit, les SLG servent à mieux saisir la Gestalt linguistique des parémies du point de vue lexico-grammatical (§ 1.2.6.). Plus en détail, les SLG-g complètes et les SLG-ac ont mis en évidence des préférences d'agencement et d'actualisation des constituants des parémies. Nous les avons dénommées *rythme syntaxique* et *rythme lexical*. Ils peuvent se présenter isolément ou interagir entre eux. Ce qui arrive assez souvent, comme on le voit dans nos classifications dans l'Annexe 6. Chaque classe dépend de l'identité de départ entrée syntaxique-entrée lexicale qui traîne le reste de la description. Autrement dit, chaque rythme inclut donc au moins une unité lexicale et une partie du discours. En particulier, les SLG-ac révèlent des *rythmes lexico-syntaxiques* (rythmes syntaxiques + rythmes lexicaux) qui se joignent donc aux schémas métriques et confirment l'existence de combinatoires privilégiées de parémisation ou, en tout cas, de 'formulaicité parémique'.

Ce travail sur les SLG nous permettra de modéliser des requêtes informatiques mieux élaborées et mieux pondérées en vue de leur reconnaissance dans des corpus (§ 6).

CHAPITRE 6

CORPUS ET MODELISATION DES REQUETES INFORMATIQUES

Nous consacrerons le présent chapitre à la présentation des corpus de français contemporain que nous interrogerons en vue de l'établissement de *f* des parémies (§ 6.1.). Nous illustrerons les programmes du logiciel *Unitex* que nous utiliserons pour modéliser nos requêtes informatiques ainsi que pour lancer nos recherches (§ 6.2.). Pour conclure, nous dévoilerons notre démarche de modélisation et les modèles de requêtes informatiques, tout en mettant en évidence le rôle de notre classification lexico-grammaticale (§ 6.3.)

6.1. Corpus

Par rapport aux critères envisagés pour les corpus à exploiter dans le cadre méthodologique fédérateur (§ 3.3.), nous avons choisi d'utiliser deux corpus : l'un existant, l'autre collecté *ad hoc*. Nous précisons, d'ailleurs, qu'en 2011, nous avons mené des études pilotes dans le CLAPI (*Corpus de Langue Parlée en Interaction*) de l'ICAR[232] aux résultats (malheureusement) peu prometteurs. Par conséquent, nous avons privilégié le médium écrit à l'oral.

6.1.1. *Leipzig Corpora Collection*

En ce qui concerne le premier, nous avons exploité la collection des corpus de Leipzig (*Leipzig Corpora Collection*) (§ 2.2.5.) conçue pour la conception de ressources

[232] Nous remercions Carole Étienne et Véronique Traverso pour leur disponibilité, leur intérêt et leurs encouragements. Le CLAPI est consultable à l'adresse suivante : http://clapi.univ-lyon2.fr/ (date de consultation : 12/11/2013).

lexicographiques et pour la création de données structurées dans une base de données[233]. La collection multilingue inclut une sous-collection de corpus en français contemporain qui couvre la période 2002-2010. Chaque corpus se compose pour l'essentiel de textes bruts aspirés du Web qui intéressent davantage la presse en ligne, mais aussi les articles encyclopédiques de *Wikipédia*. Toutes les sources, à savoir les URL, sont documentées pour chaque corpus.

La véritable particularité de toute la collection Leipzig concerne son prétraitement textuel qui rejoint un de nos intérêts théorico-méthodologiques. Les corpus sont échantillonnés d'après la notion de *phrase graphique* : chaque corpus implique un recueil de phrases graphiques, non pas de texte tout venant. Au-delà de la segmentation automatique en phrases, l'équipe du Département d'Informatique de l'Université de Leipzig a effectué une deuxième sélection. Par des scripts, ils ont effacé :

- des phrases trop longues (apparemment en raison de fautes de segmentation). Ce qui va nous éviter des problèmes de lecture des concordances et établir l'identité : ligne de concordance = phrase graphique ;
- des phrases redondantes ;
- des phrases contenant trop de mots qui ne relèvent pas de la langue visée. Ce qui nous assure une meilleure viabilité des résultats et moins d'échecs lors de l'interrogation par nos requêtes informatiques ;
- des phrases se différenciant par un seul mot graphique ainsi que des doublons. Notre fréquence est ainsi à l'abri de répétitions non souhaitées dans le corpus, voir de l'effet copier-coller qui affecte la presse en ligne.

Par la suite, un identifiant chiffré est attribué à chaque phrase graphique par un choix randomisé. D'après cet identifiant, les phrases ont été triées par ordre croissant et subdivisées en corpus ayant la même 'taille phrastique'. Chaque corpus contient donc la même quantité de phrases. La mesure quantitative de référence n'est pas, comme l'on fait couramment, l'occurrence du mot graphique simple (voir du n-gramme), mais l'occurrence de la phrase graphique. Nous avons ainsi une unité de mesure formulaire pour estimer *f* de nos parémies,

[233] Toutes les informations sur les corpus que nous donnerons par la suite sont tirées du manuel pour les utilisateurs. Le manuel est disponible à l'adresse : http://corpora.uni-leipzig.de/download.html (date de consultation : 12/11/2013).

ainsi que pour évaluer les phrases graphiques qu'elles remplissent (et si elles les remplissent de façon intégrale).

Comme nous le mentionnions ci-dessus, les corpus sont intégrés par une documentation dans un fichier *.txt* séparé. Elle comprend les URL, les dates d'aspiration des phrases ainsi que des identifiants chiffrés. Ces derniers sont jumelés à chaque identifiant chiffré des phrases. La source et la phrase sont ainsi liées. Ce qui permet de remonter aux textes dont relèvent les phrases échantillonnées. Le risque de la perte d'information (accusation contre toute pratique d'échantillonnage) est ainsi balancé par ce soin documentaire. Au cas où il serait nécessaire, on pourrait envisager l'élargissement du contexte d'une phrase par la référence directe au texte-source[234].

C'est par la lecture de la documentation que nous avons sélectionné deux corpus de la sous-collection française : les corpus de presse en ligne de 2009 (*fra_news_2009_1M-text*) et de 2010 (*fra_news_2010_1M-text*). Essayons de motiver notre décision suivant les critères énumérés au § 3.3.2.

- *Taille* : la 'taille phrastique' de chaque corpus s'élève à 1 million[235]. Pour revenir à la mesure traditionnelle, le corpus de 2009 compte 54.467.498 occurrences et celui de 2010 en totalise 54.281.586. Malgré le fait que l'avenir appartient à l'exploitation (et à la structuration) des *Big Data*, nous avons préféré et dû calibrer notre étude sur des corpus de taille assez restreinte. D'une part, ce choix testera les tendances et les plages de valeurs que nous avons synthétisées au § 3.3.3. et confirmer (ou réfuter) la nécessité d'adopter une *optique parémiométrique* qui problématise les acquis de la lexicométrie traditionnelle. D'autre part, les capacités limitées de notre ordinateur (notamment de la RAM) nous ont empêché de dépouiller des corpus de taille plus importante[236] (§ Conclusions).
- *Temps* : nous avons ciblé la microdiachronie et obtenu un *corpus strictement microdiachronique* (§ 3.3.2.). Nous estimerons la *f* des parémies dans une période

[234] Nous précisons qu'à côté de la documentation, les corpus disposent de fichiers contenant les index des mots simples ainsi que les cooccurrences statistiquement les plus significatives (test *log-likelihood*).
[235] D'après la segmentation d'*Unitex* 3.1 (§ 6.2), les tags qui reconnaissent une phrase sont un peu plus élevés (environ une quarantaine de milliers pour chaque corpus. Cela dépend des règles de segmentation divergentes. Plus précisément, *Unitex* insère le tag aussi après un point virgule (*;*), alors que la phrase graphique des corpus Leipzig correspond seulement à la suite se terminant par un point final (.).
[236] Nous avons essayé d'utiliser le corpus *frWaC*. Malheureusement, même le splittage du corpus a créé des soucis de prétraitement par *Unitex*. En tout cas, nous remercions Eros Zanchetta au nom de toute l'équipe *WaCky* pour le téléchargement gratuit du corpus. Comme nous l'avons montré au § 5, nous l'avons quand-même consulté pour résoudre des cas douteux d'annotation via l'interface *Sketch Engine*.

relativement récente et dans un arc de temps quasi synchronique. La liste de *f* pourra faire immédiatement l'objet d'enquêtes de familiarité ou être confrontée à des études disponibles et quasi contemporaines (Sevilla Muñoz & García Yelo 2008, García Yelo 2009). Dans ce cas, il faudra tenir compte du fait que l'étendue des sources cible la francophonie dans son intégralité, voir la dépasse ;

- *Médium* : le recours au Web pour les corpus nous mettra face à face de l'écrit numérique. Ce qui écarte notre étude des expériences fondées sur la version numérisée de sources textuelles au format papier (de la presse, aussi). Il en va de même pour celles où le Web est employé comme un corpus ;
- *Langue* : de manière conventionnelle, le corpus est monolingue et traite seulement le français ;
- *Discours et Genre* : le choix des genres de la presse (quoiqu'en ligne) nous situe en continuité par rapport à la plupart des expériences que nous avons relatées au § 3. Cela comble ainsi un vide dans la littérature parémiologique francophone. De plus, la variété des discours qu'assure la presse nous donnera une première liste de *f* des parémies qui ne sera pas affectée par telle ou telle autre compétence encyclopédique, à savoir par un public ciblé et restreint.

6.1.2. *LiPaF : un corpus ad hoc*

À côté des deux corpus de la collection Leipzig, nous avons décidé de juxtaposer la liste de *f* des parémies d'après leur nombre d'occurrences dans la littérature parémiologique francophone. Nous avons ainsi crée le corpus *LiPaF* (*Littérature Parémiologique sur le répertoire parémiologique Français*) pour comprendre dans quelle mesure les parémiologues contribuent à la promotion ou à la disparition des membres du répertoire parémiologique en français.

Nous avons donc sélectionné des textes tirés de la littérature parémiologique comme suit :

- *Taille* : le corpus inclut 62 textes scientifiques de parémiologie pour un total de 1.045.203 occurrences. La taille restreinte révèle déjà le caractère spécialisé de *LiPaF*.

- *Temps* : *LiPaF* part de la fin des années 1960 et arrive jusqu'à l'an 2012. Nous avons ici un *corpus approximativement microdiachronique* (§ 3.3.2.), à savoir un corpus couvrant quelques décennies du même siècle.
- *Médium* : nous avons inséré de l'écrit numérisé en vue d'une publication au format papier ainsi que de l'écrit numérisé à partir de sources textuelles au format papier. Aucun processus de numérisation n'a été mené de notre part.
- *Langue* : toutes les contributions scientifiques choisies traitent des parémies françaises. Cela n'empêche pas d'autres langues de rédaction. En ce sens, nous avons créé un corpus multilingue qui garde le français pour les parémies.
- *Discours* : *LiPaF* est un corpus spécialisé consacré au discours parémiologique. Il regroupe des textes conçus par et s'adressant à la communauté des parémiologues. En général, ils sont destinés à l'académie. Ces textes peuvent intéresser aussi un public d'apprentis et d'étudiants, ainsi qu'éventuellement d'amateurs passionnés en la matière ;
- *Genre* : excepté deux chapitres tirés et réélaborés d'une monographie, les 60 textes restants sont des articles parus dans des revues scientifiques (§ Annexe 7 pour les détails).

Le choix des études n'a pas pointé des parémiologues en particulier[237].

Le seul prétraitement réservé à *LiPaF*, outre ceux d'*Unitex*, concerne le balisage du début de texte par un identifiant chiffré (ex. <TEXT FILE 1>) pour faciliter l'identification des études a posteriori. Nous n'avons ni ajouté d'autres informations ni enlevé des éléments textuels. Nous avons gardé tableaux, titres et références. Contrairement à d'autres parémiologues (§ 3.2.15.), nous calculerons toutes les occurrences repérées, sans aucune distinction.

En gros, à l'aide de *LiPaF*, nous essayerons de mieux saisir les parémies privilégiées par les spécialistes au moment de la rédaction de leurs études. Qu'elles soient employées comme points de départ ou comme exemples, nous chercherons à mettre en relief le penchant parémique d'usage au sein de la littérature parémiologique. Il sera intéressant de comparer les *f* issues du corpus *LiPaF* avec celles des corpus Leipzig (malgré les périodes non directement superposables). La comparaison permettra de mieux évaluer si le suremploi et le sous-emploi d'une parémie sont fonction d'une compétence parémiologique plus ou moins développée.

[237] Pour une revue de la littérature parémiologique en français jusqu'à 2007, voir Gutiérrez Sánchez (2008).

De même, elle nous pourra révéler si la préférence pour les unes et le délaissement des autres sont transversaux au grand public et aux parémiologues. Autrement dit, la comparaison entre liste de *f* nous éclairera si la connaissance encyclopédique (passive) et la recherche sur le répertoire parémiologique influencent vraiment l'usage des parémies. Ou si plutôt l'usage des parémies ne dépend pas seulement de l'étendue du répertoire parémiologique connu, mais de la perception d'une *f* subjective (surtout de la part des parémiologues) qui amène à préférer la parémie qui est ressentie comme la plus 'populaire' par introspection ainsi que par vécu linguistique (production et compréhension) personnel.

6.2. *Unitex*

Unitex[238] est un gratuiciel (licence LGPL-LR) conçu pour le traitement automatique de textes et maintenu par le Laboratoire d'Informatique Gaspard-Monge de l'Université Paris-Est Marne-la-Vallée. Depuis sa première version par Paumier en 2002 qui suit l'expérience d'INTEX par Silberztein dans les années 1990, des améliorations et des enrichissements sont intervenus par la suite. Nous utiliserons la dernière version disponible (3.1 bêta) qui s'insère dans le tout récent environnement de développement intégré *GramLab* pour l'implémentation et pour l'exploitation de programmes informatiques et de ressources lexicales libres de droits.

L'interface du logiciel (ou simplement l'invite de commande) permet d'appeler plusieurs programmes de prétraitement ou de traitement informatiques ainsi que des ressources linguistiques spécifiques. À ce propos, on peut définir *Unitex* comme un logiciel *language-dependant*, c'est-à-dire qu'il applique à chaque langue des ressources qui encodent le lexique, la syntaxe et leurs particularités. Plus précisément, le traitement automatique sous *Unitex* tire parti d'une série de descriptions linguistiques informatisées mises au point dans le cadre du Lexique-Grammaire. Font partie de ces ressources : les tables[239], les dictionnaires électroniques au format DELA (Courtois 1994-1995)[240] ainsi que les grammaires locales (§§

[238] http://www-igm.univ-mlv.fr/~unitex/ (date de consultation : 12/11/2013).
[239] Elles sont téléchargeables à partir du site du LIGM – équipe *Modèles et Algorithmes* / *Linguistique pour le traitement des langues* : http://infolingu.univ-mlv.fr/ (date de consultation : 12/11/2013).
[240] Les dictionnaires sont normalement intégrés dans *Unitex* pour chaque langue. Des dictionnaires spécialisés sont également conçus par des chercheurs, comme dans le cas, entre autres, de Cetro (2013).

4.2.2., 6.3.). La création de ressources adaptées se doit au réseau informel RELEX qui rassemble des laboratoires de linguistique computationnelle partout dans le monde[241].

Rien que cette présentation sommaire d'*Unitex* nous permet de distinguer notre étude de celles que nous avons mentionnées au § 4.2.1. À la place de logiciels payants ou de moteurs de recherche non adaptés à l'usager, nous ferons recours à un gratuiciel souple. Plutôt qu'un logiciel *language-independant*, le plus souvent à base statistique, *Unitex* favorise une analyse linguistique ciblée et détaillée, notamment au moment de la tokenisation. L'utilisateur peut en outre enrichir les outils d'analyse par l'implémentation de ressources personnelles. Ce qui lui permet de prendre le dessus sur l'automatisation aveugle (comme l'a été pour l'annotation à l'aide du *TreeTagger*) et de mieux contrôler toutes les étapes du traitement automatique au fur et à mesure.

6.2.1. *Prétraitement des corpus*

Unitex analyse des textes bruts, à savoir sans aucune mise en forme ou en page des données soumises (fichiers *.txt*). Par défaut, *Unitex* se sert d'une variante du codage *Unicode*, c'est-à-dire le codage *Little-Endian Unicode*. Normalement, donc, la première étape de prétraitement concerne la transcodification de fichiers *.txt*. Ce que nous avons fait pour notre corpus parémique ainsi que pour les corpus Leipzig et *LiPaF*.

Après la transcodification, *Unitex* procède à la segmentation en phrases graphiques par une grammaire locale propre à chaque langue. Par la complexité de la grammaire de segmentation française (Paumier 2013 : 39), on comprend que la notion de *phrase graphique* peut varier d'un logiciel/annotateur à un autre. Par exemple, elle est plus mécanique dans le cas du tokeniseur utilisé pour constituer les corpus Leipzig. En revanche, elle est décidément plus facettée sous *Unitex* qui prend en compte, entre autres, les points de suspension entre parenthèses et désambiguïse les cas d'abréviations ou de sigles. D'où la différence quantitative de tags pour marquer les phrases. Vu que le processus de segmentation est *language-dependant* et malgré la nature multilingue du *LiPaF*, nous avons évidemment privilégié la grammaire de segmentation française. Notre intérêt de recherche vise les parémies en français, non pas d'autres aspects du corpus.

Au-delà de la segmentation en phrases, c'est la tokenisation qui occupe la place la plus fondamentale. Par l'appel du programme *Tokenize* et de la grammaire locale

[241] Une liste des ressources lexicales implémentées par les partenaires du RELEX est consultable à l'adresse : http://www-igm.univ-mlv.fr/~unitex/index.php?page=11 (date de consultation : 12/11/2013).

correspondante à la langue du texte à prétraiter, *Unitex* découpe le texte soumis en *tokens*. En français, ils peuvent correspondre à une chaîne de caractères relevant de l'alphabet d'une langue, un espace ou tout autre caractère (Paumier 2013 : 42). En sortie, *Unitex* crée des fichiers *.txt* contenant l'index et *f* des *tokens* ainsi que la liste des *tokens* triées par ordre alphabétique et par ordre décroissant de *f*. Outre les fichiers, ces listes sont affichées dans une fenêtre à part dans l'interface. Pour nos corpus, encore une fois, la grammaire de tokenisation appliquée est celle française, malgré la nature multilingue du *LiPaF*.

La tokenisation suit de pair l'application des DELA pour l'étiquetage des *tokens*. Le résultat de l'application des DELA par défaut est affiché dans une fenêtre de l'interface qui reprend en trois sous-parties : les mots (graphiques) simples et les mots (graphiques) composés reconnus dans les DELA, ainsi que les mots non reconnus parce qu'ils ne sont pas inclus dans les DELA. Par la suite, on peut appliquer d'autres dictionnaires disponibles sous *Unitex* ou créés par l'utilisateur (toujours au format DELA) pour améliorer la lemmatisation et l'annotation morphosyntaxique (Paumier 2013 : 43-45). Les mots étiquetés suivent le formalisme propre au DELA qui prévoit (en ordre) :

- une entrée, soit-elle une forme fléchie ou un lemme ;
- le lemme ;
- le code grammatical (§ Annexe 3), voir tous les codes grammaticaux qu'on peut attribuer à l'entrée, sans désambiguïsation préalable ;
- les codes flexionnels, notamment : la précision de temps, mode (§ Annexe 3), personne (1, 2, 3) et nombre (*:s* pour le singulier, *:p* pour le pluriel) pour les verbes ; l'indication du genre (*:f* pour le féminin et *:m* pour le masculin) et du nombre pour les adjectifs, les déterminants et les noms ;
- un code indiquant une marque d'usage approximatif (*z1*, *z2*, *z3*), à savoir une information sur l'usage ordinaire ou spécialisé de l'entrée ;
- un code sémantique générique sur la nature abstraite (*Abs*), concrète (*Conc*), humaine (*Hum*), animale (*Anl*) ou collective (*Coll*) de l'entrée (Paumier 2013 : 47-52).

Suivant ce formalisme de codification de l'information, « chaque utilisateur peut introduire ses propres codes, et créer ses propres dictionnaires » (Paumier 2013 : 52).

En ce qui concerne notre corpus parémique, nous avons préféré que l'annotation et la lemmatisation touchent toute unité lexicale, et ce, vu les corrections que nous avons

documentées au § 5.2. D'abord, nous avons appliqué tous les DELA sous *Unitex*, y inclus deux DELA consacrés aux pays et aux villes (*Prolex-PaysCapitales*) et aux toponymes (*Prolex-Toponymes*). La figure ci-dessous montre la fenêtre tripartite que nous avons obtenue :

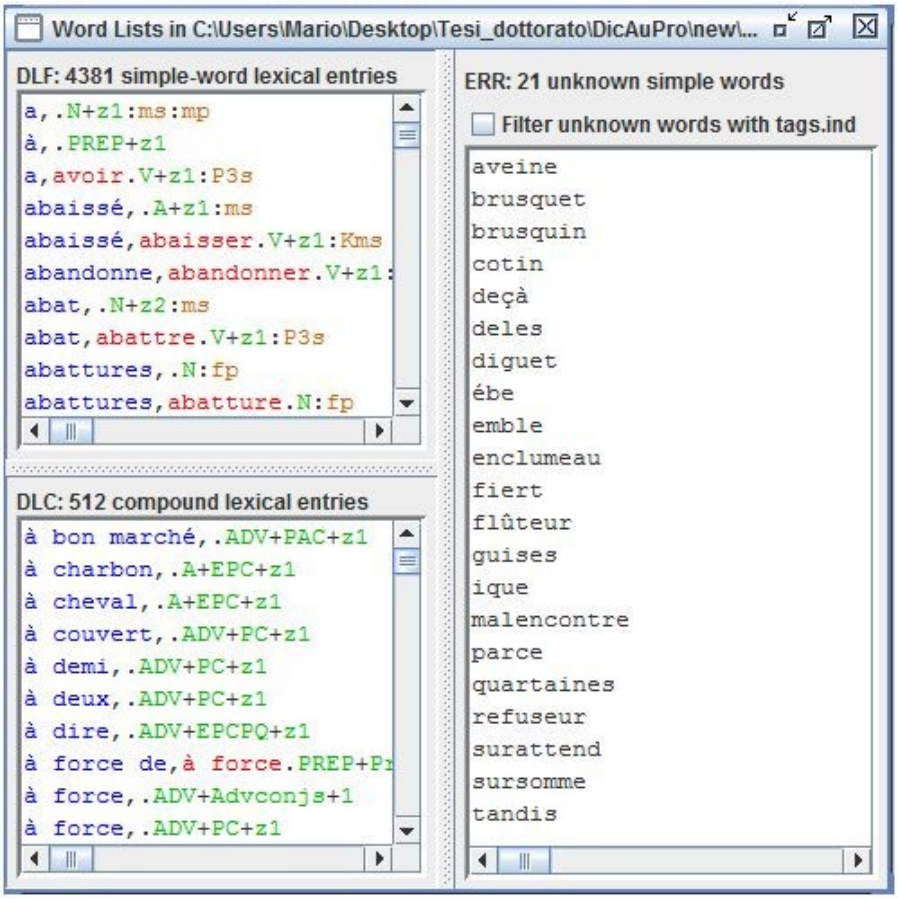


Figure 14. Listes des mots simples (encadré en haut à gauche) et composés (encadré en bas à gauche) reconnus et des mots non reconnus (encadré à droite) dans notre corpus parémique.

On remarque qu'à côté des 4.380 mots simples et des 512 mots composés reconnus dans les DELA, il reste encore 21 mots non reconnus[242]. Parmi ceux-ci, on reconnaît des unités lexicales désuètes que nous avons discutées au § 5.2.4. En revanche (et avec surprise), on constate que les formes *deçà*, *parce* et *tandis* ne sont ni enregistrées comme mots simples ni comme mots composés (l'adverbe *en deça* et les conjonctions *parce que* et *tandis que*) dans

[242] Les DELA sont, certes, exhaustifs. Cela ne nous empêche pas de souligner que le caractère archaïsant que l'on colle souvent au lexique parémique est en bonne partie remis en question. On ne saurait en effet expliquer la performance d'étiquetage d'*Unitex* que par le caractère assez ordinaire (ou du moins saisissable) du lexique parémique.

les DELA. Nous avons décidé d'encoder ces mots dans deux DELA : *dico_LexParemia_simp* et *dico_LexParemia_comp*.

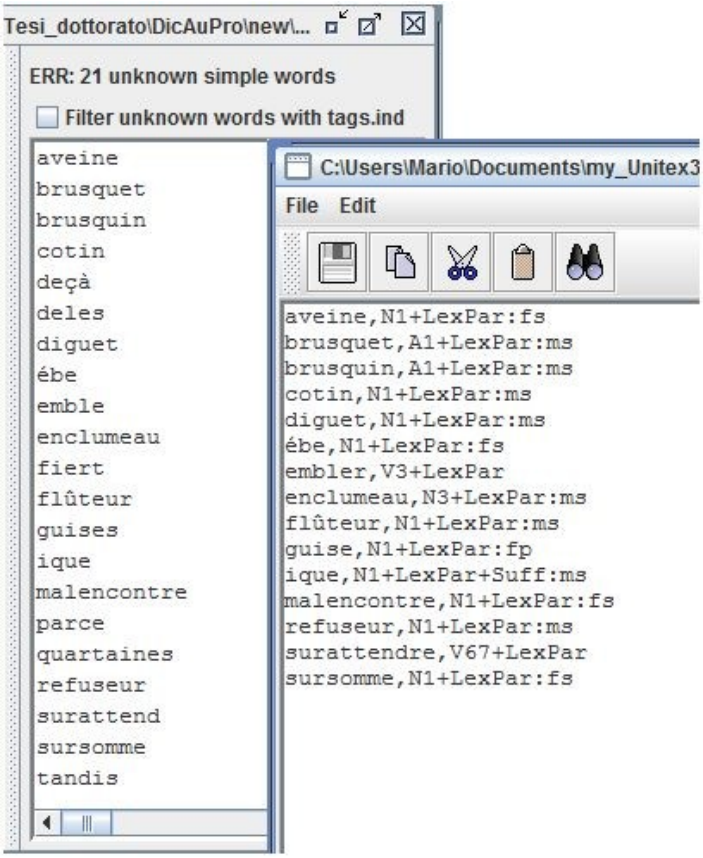


Figure 15. Création du DELA *dico_LexParemia_simp*.

Pour *dico_LexParemia_simp* (Figure 14), nous avons repris seulement les mots simples, peu usités. Dans *dico_LexParemia_comp*, nous avons inséré l'adverbe et les conjonctions mentionnées ainsi que *fièvres quartaines*, l'emploi de *quartaine* n'étant attesté que dans ce nom composé. À chaque entrée nous avons attribué le code *+LexPar* pour indiquer la source parémique à l'origine de la création du DELA.

Après avoir fléchi[243] (Paumier 2013 : 57-61) et compressé nos DELA (Paumier 2013 : 66-67), nous les avons appliqués à l'étiquetage de notre corpus parémique par la fonction *Apply*

[243] La flexion de *férir* n'a pas eu lieu. Le *Dictionnaire du Littré* (d'où l'équipe *DicAuPro* a tiré la parémie contenant ce verbe, § 5.1.) signale déjà son caractère désuet (s.v. '*férir*'). Un cas à part entière intéresse la forme *deles*. Malgré la forme correcte dans notre corpus parémique, *Unitex* transcode et tokenise comme *deles* la préposition avec déterminant pluriel contracté *des*. Cela se vérifie seulement pour 2 occurrences. Nous ignorerons cette forme que nous ne savons pas motiver. D'ailleurs, elle est bien présente dans les DELA. Nous avons enregistrée la seule entrée à la 3e personne du singulier du verbe *férir* et la forme immotivée *deles* dans un

Lexical Resources... dans le menu *Text* (Paumier 2013 : 67). Par exemple, suite à l'application du DELA *dico_LexParemia_simp*, les mots non reconnus se réduisent à 6, comme le montre la Figure 15 :

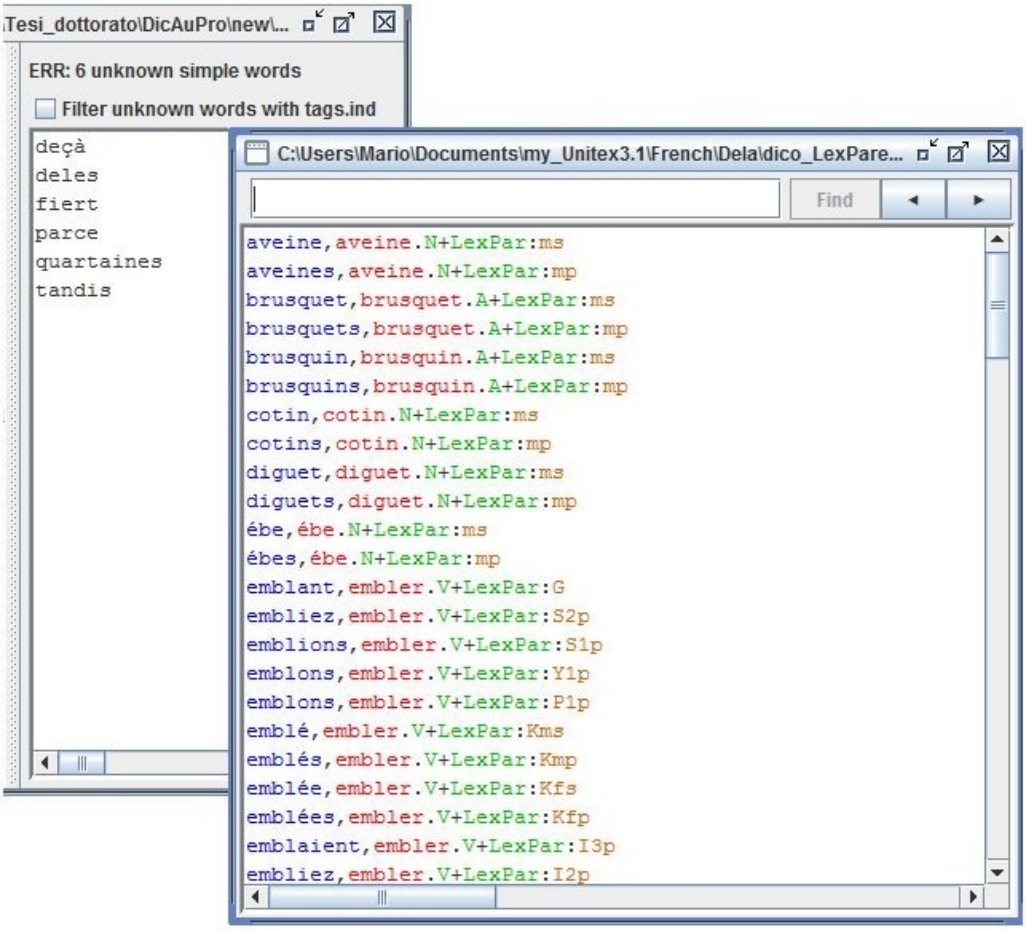


Figure 16. Résultat de l'application du DELA *dico_LexParemia_simp* à notre corpus parémique.

Pour l'étiquetage des corpus Leipzig et *LiPaF*, nous avons seulement accepté les résultats renvoyés par *Unitex* après l'application de tous les DELA, y compris *dico_LexParemia_simp* et *dico_LexParemia_comp*. À ce propos, l'annotation et la lemmatisation d'*Unitex* ont abouti à ce scénario :

DELA *dico_LexParemia_exc*. Dans ce même dictionnaire, nous avons encodé *chaque* comme adjectif. Les DELA l'enregistre en effet comme déterminant.

	Mots simples reconnus	Mots composés reconnus	Mots non reconnus
fra_news_2009_1M-text	255.326	47.552	51.623
fra_news_2010_1M-text	248.205	46.723	48.392
LiPaF	31.670	3.973	15.503

Tableau 26. Résultats de l'annotation et de la lemmatisation des corpus Leipzig et du *LiPaF* par les DELA français.

Dans les corpus Leipzig, la quasi-totalité des mots non reconnus sont des fautes d'orthographe (ex. *annnonce*, *incription*, *indémnité*) ainsi que des fautes de tokenisation présentes (ex. *publicitéferait*, *sappuie*). Parfois nous rencontrons des néologismes et des termes (ex. *biométhanisation*, *cochercheur*, *colorature*)[244] et, plus rarement, des emprunts (ex. *amaretto*). Quant au *LiPaF*, la totalité des mots non reconnus est à reconduire exclusivement au multilinguisme du corpus (à savoir aux mots qui ne sont pas français) et à quelques fautes d'orthographe en français.

Malgré l'ensemble des mots non reconnus ne soit pas à sous-évaluer, il est vrai que la plupart des mots font l'objet d'une description dans les DELA. Le manque d'annotation et de lemmatisation, en tout cas, ne compromet pas vraiment notre recherche parce que tout le lexique et toute la syntaxe parémiques sont gérés par *Unitex*.

Nous travaillerons ainsi sur des corpus étiquetés. Ce qui différencie notre étude de la plupart des recherches présentées au § 4.

Nous tenons encore à préciser qu'a priori, la taille des corpus soumis ne pose pas de problèmes. Des expériences préalables nous ont obligé à renoncer à des corpus de taille 'milliardaire' parce que le prétraitement d'*Unitex* n'arrivait pas jusqu'au bout (§ 6.1.1.). Il se peut que les difficultés de prétraitement soient à reconduire à la puissance et aux capacités de stockage de notre ordinateur plutôt qu'à la gestion des données par *Unitex*.

[244] La mise à jour des DELA est un sujet de recherche et un souci descriptif qu'on rencontre souvent parmi les membres du RELEX. Nous mentionnons, entre autres, Fairon & Cougnon (2009).

6.2.2. *Repérage sur corpus*

Outre la fonction d'édition des DELA (§ 6.2.1.) et des grammaires locales (§ 6.3.), nous nous limiterons à appeler exclusivement le programme de recherche d'*Unitex* : *Locate Pattern*. Ce programme accepte autant des expressions régulières que des graphes comme requêtes informatiques. La fenêtre de l'interface dédiée à ce programme est reproduite dans la Figure 16 ci-dessous :

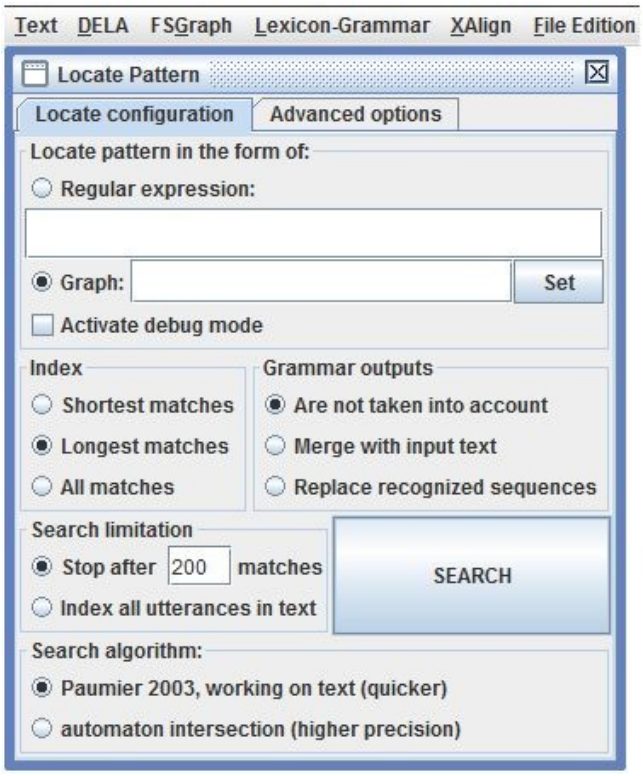


Figure 17. Fenêtre du programme *Locate Pattern*.

L'onglet *Locate Configuration* permet de façonner les modalités d'interrogation du corpus traité. Au-delà de la sélection du type de requête (encadré *Locate pattern in the form of*), l'encadré *Index* offre trois modalités de renvoi et de comptage des séquences reconnues :

- ou la présentation des séquences les plus courtes (*Shortest matches*) ;
- ou la présentation des séquences les plus longues (*Longest matches*) ;
- ou la présentation des deux (*All matches*) (Paumier 2013 : 84).

En raison de la variation qui affecte les parémies ainsi que de notre intérêt quantitatif, nous choisirons toujours la présentation des séquences les plus longues. Ce qui nous évitera de compter des doublons.

L'encadré *Grammar outputs* concerne directement les graphes. Plus précisément, *Unitex* insère (*Merge with input text*) ou remplace (*Replace recognized sequences*) les séquences reconnues par les sorties prévues par le graphe (Paumier 2013 : 147). Comme nous ne sommes pas intéressé à insérer des données ou à remplacer nos séquences parémiques, nous cocherons l'option *Are not taken into account*.

L'encadré *Search limitation* regroupe deux alternatives :

- soit arrêter le nombre d'occurrences à un seuil préétabli (*Stop after ___ matches*) ;
- soit demander à *Unitex* de présenter toutes les occurrences repérées (*Index all utterances in text*) (*ibid*.).

Cela va de soi que nous nous orienterons pour l'énumération de toutes les occurrences, et ce, afin d'apprêter une liste de *f* complète.

Pour conclure, l'encadré *Search algorithm* offre la possibilité d'effectuer l'interrogation du corpus en tant que texte (*Paumier 2003, working on text*) ou en tant qu'automate (*automaton intersection*) (Paumier 2013 : 148). Malgré notre préférence accordée aux graphes, nous pencherons pour l'interrogation du texte.

L'onglet *Advanced options* sert, d'une part, à définir le degré d'ambiguïté de l'étiquetage par *Unitex* et, d'autre part, à gérer les variables éventuellement encodées dans les graphes (Paumier 2013 : 148-151). Nous ne pratiquerons aucune politique de désambiguïsation parce que l'ambiguïté pourrait (paradoxalement) nous aider à rattraper quelques occurrences. Comme nous venons de le déclarer et comme on le verra, la modélisation des graphes pour la reconnaissance des parémies écartera toute variable (et, par conséquent, toute sortie) (§ Conclusions).

6.2.3. *Concordancier*

À l'issue de l'interrogation, les séquences reconnues peuvent être affichées dans une fenêtre comme celle qui suit :

Figure 18. Exemple de concordance sous *Unitex* (SLG-ac recherchée : *{À chaque N son N}* ; *LiPaF*).

par le recours à la fonction *Display Located Sequences…* sous le menu *Text* (Paumier 2013 : 85). La fenêtre est créée au format HTML à partir d'un index KWIC stocké dans le répertoire du corpus interrogé. L'utilisateur peut également effectuer des tris pour faciliter la lecture contextuelle des séquences reconnues (*Show Matching Sequences in Context*) (*ibid.*). Le cas échéant, il sera possible d'insérer des séquences en sortie, lorsqu'on emploie des graphes (*Modify text*), ainsi que de sauvegarder les séquences reconnues ou non reconnues dans un fichier *.txt* (*Extract units*) (*ibid.*).

Nous choisirons de visualiser les résultats dans l'interface d'*Unitex* au moment de l'interrogation des corpus Leipzig et du *LiPaF* pour évaluer la précision des résultats quantitatifs renvoyés (§ 6.2.2.). Pour ce qui concerne les corpus Leipzig, l'affichage des concordances nous aidera à mieux déchiffrer la relation entre parémie et phrase graphique.

6.3. Modélisation des requêtes informatiques

Quoique nos SLG se prêtent bien à une codification sous forme d'expressions régulières d'après la syntaxe de recherche d'*Unitex* (Paumier 2013 : 75-83) tout comme d'après toute autre syntaxe, nous préférerons nous concentrer sur la modélisation des graphes. Leur potentiel et leur souplesse ont été évidemment sous-estimés par bon nombre de parémiologues (§ 4.2.1.).

Comme nous le rappelions au § 4.4., Habert *et al.* (1997 : 121) nous invitent à expérimenter à maintes reprises pour se rapprocher d'un fait linguistique dans un corpus. Nous présenterons ainsi nos premiers exemplaires (§ 6.3.2.) qui ont délibérément poussé à l'extrême les contenus lexical et syntaxique ainsi que la contrainte linéaire. Ce pour mieux saisir l'interaction entre ces dimensions. À partir de notre expérience ainsi que de celles que nous avons illustrées au § 4.2.2.2., tout comme grâce à notre classification lexico-grammaticale (§§ Annexes 5-6), nous essayerons de modéliser les graphes syntaxiques les plus performants pour repérer les parémies et leurs variantes dans nos corpus (§ 6.3.3.). D'abord, nous détaillerons comment nous avons conçu nos graphes syntaxiques par l'éditeur d'*Unitex* (§ 6.3.1.).

6.3.1. *Graphes syntaxiques : l'éditeur d'Unitex et ses avantages pour les parémiologues*

Parmi les types de graphes qu'on peut créer (Paumier 2013 : 117-121), nous mettrons au point seulement des *graphes syntaxiques* ou *grammaires locales* (§ 4.2.2.).

> « Les graphes syntaxiques, également appelés grammaires locales, permettent de décrire des motifs syntaxiques qui pourront ensuite être recherchés dans des textes. De tous les types de graphe, ceux-ci possèdent la plus grande puissance d'expressions, car ils permettent de faire référence aux dictionnaires » (Paumier 2013 : 120).

Bien au-delà de la séquence syntaxique, les graphes syntaxiques permettent de faire appel aux codes des DELA. En outre, ils peuvent recourir à la syntaxe de recherche propre aux expressions régulières, y compris des symboles spéciaux (Paumier 2013 : 76-77) ainsi que des filtres morphologiques (Paumier 2013 : 82-83). À l'immédiateté des expressions

régulières s'ajoute la multiplicité des chemin de reconnaissance. Ce qui est un avantage précieux quand le linguiste doit gérer des séquences formulaires et 'prédire' leurs variantes.

Un éditeur de graphes est intégré à l'interface d'*Unitex* (sous le menu *FSGraph*) (Paumier 2013 :94-110) :

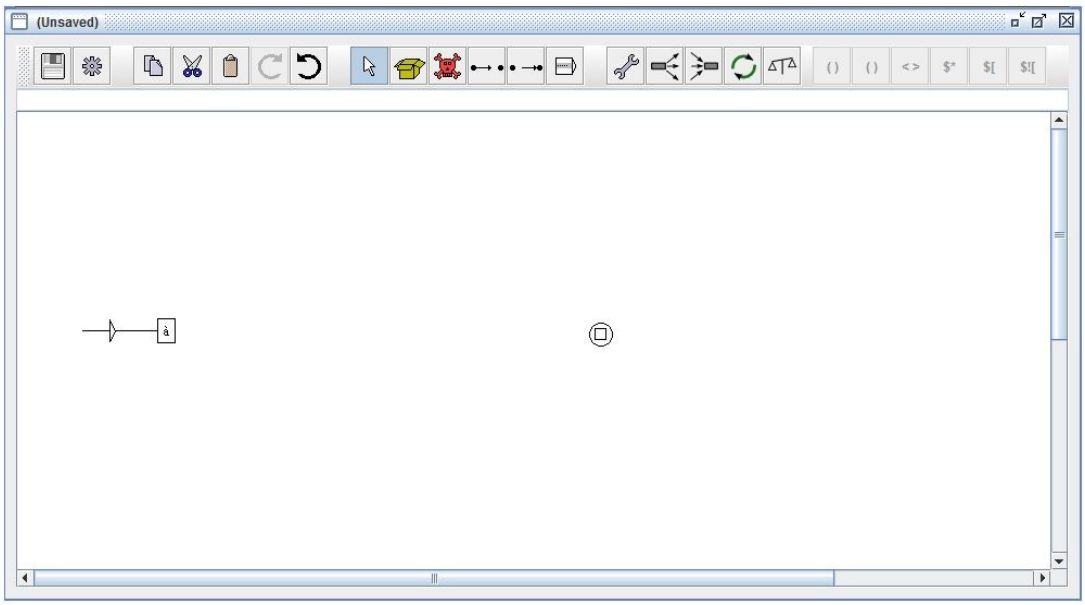


Figure 19. Éditeur de graphes sous *Unitex*.

La flèche à gauche représente l'état initial (le début de la séquence à reconnaître ou à produire) alors que le carré encerclé est l'état final (la fin de la séquence à reconnaître ou à produire) du graphe. La lecture de la séquence se fait de la gauche vers la droite. Dans cet espace idéal, il faut construire le(s) chemin de reconnaissance ou de production souhaité(s). Dans la Figure 18, nous avons créé un état contenant la préposition *à* et relié à l'état initial. Nous avons commencé à construire le graphe pour reconnaître les parémies de la classe [*À_Prép*].

Chaque état peut contenir plusieurs séquences (ou symboles). Il suffit de les séparer par le caractère +. En guise d'exemple, prenons la classe [*À Adj*] :

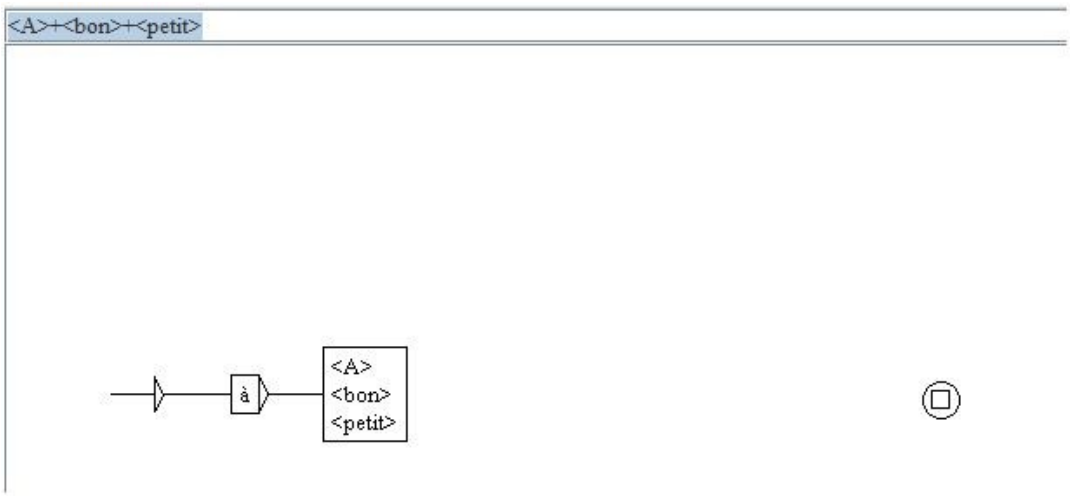


Figure 20. Exemple de graphe syntaxique avec état contenant le séparateur +.

Si l'on imagine que le dernier état est relié à l'état final, le graphe aurait reconnu toute occurrence de la préposition *à* (sans aucune contrainte de casse) suivie de n'importe quelle forme des adjectifs *bon* (<*bon*>) et *petit* (<*petit*>), voir de tout adjectif (<A>) présent dans les DELA. Les chevrons (<…>) permettent ainsi d'appeler le lemme de la forme et le code du DELA recherchés[245].

On pourrait éventuellement relier un état à lui-même comme le montre la Figure 20 :

Figure 21. Exemple de reliure d'un état à lui-même.

Ce qui favorisera, en l'occurrence, la reconnaissance de deux (ou plusieurs) adjectifs après la préposition à. En général, la récursivité d'un état pourra nous aider à déceler des adjonctions

[245] Les chevrons appellent d'autres commandes encodées comme symboles spéciaux. Dans ce cas de figure, il suffirait <A> pour reconnaître toutes les formes de *bon* et *petit*. On aurait éventuellement pu attribuer un *poids* à des états (Paumier 2013 : 104-105), c'est-à-dire donner la priorité à un chemin de reconnaissance. Ce permet de distinguer la valeur des chemins lorsque deux ou plusieurs d'entre eux reconnaissent la même séquence. L'utilisation des poids prend son sens quand on prévoit des sorties.

d'une même unité lexicale ou d'une même partie du discours à une position p donnée de la séquence-parémie.

Comme nous l'avons vu au § 4.2.2.1., un graphe peut en appeler un autre. Ce dernier est nommé *sous-graphe*. L'appel se fait par la création d'un état où l'on inclut le caractère *:* suivi de l'adresse du répertoire du PC qui contient le graphe concerné. Suivons notre exemple et faisons l'hypothèse d'insérer le sous-graphe *N_3_exemple* contenant la liste de tous les noms qui suivent les adjectifs dans la classe de parémies [*À Adj*]. Créons donc le sous-graphe faisant recours à la fonction de copie de liste (Paumier 2013 : 106-107) et insérons un chevron à gauche et à droite de chaque élément de notre liste :

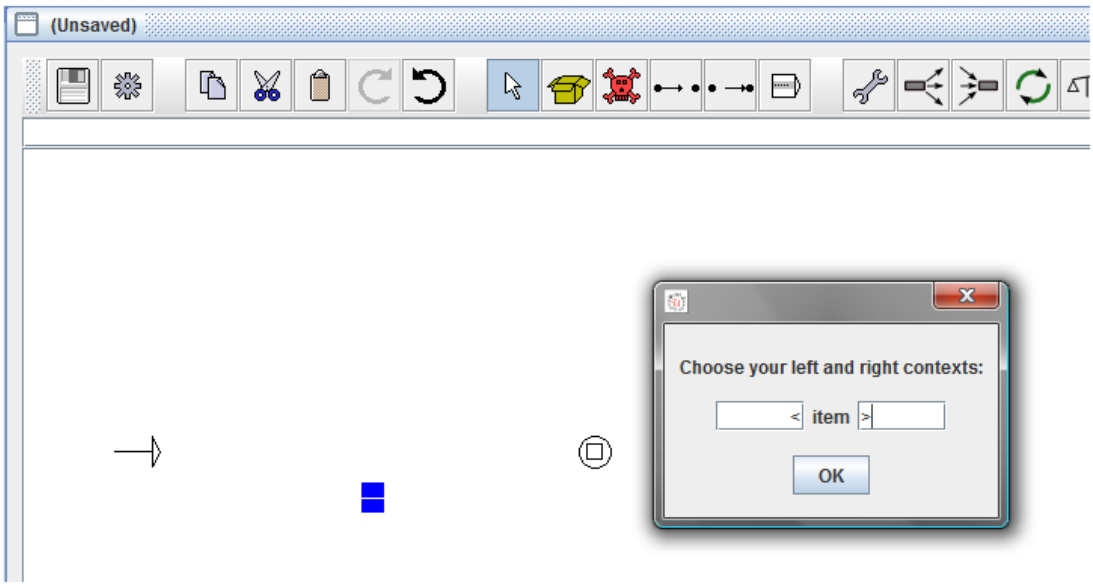


Figure 22. Copie de liste sous *Unitex*.

Nous obtenons ce qui suit :

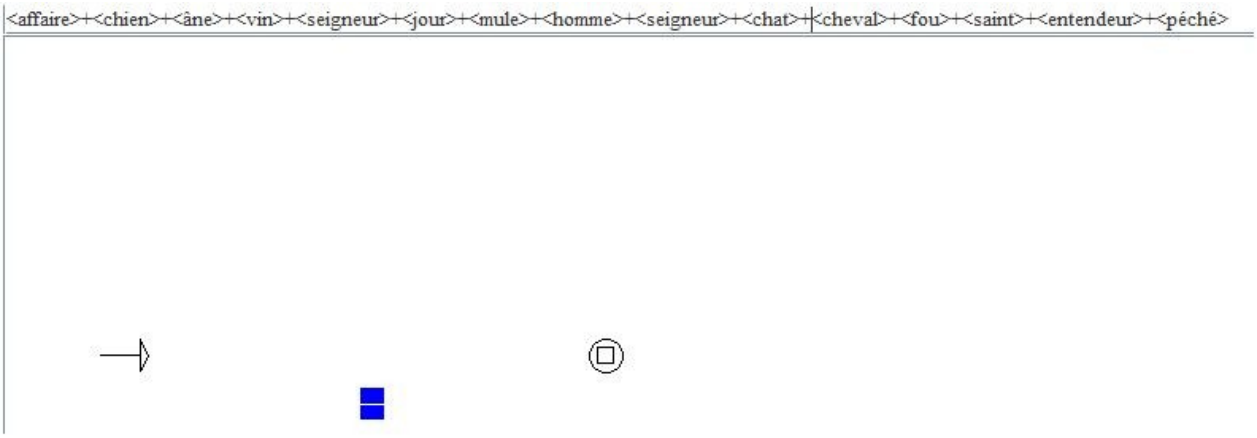


Figure 23. Résultat de la copie de liste sous *Unitex*.

Le contenu de la barre sera par la suite inséré dans l'état en bleu. Nous sauvegardons ainsi le sous-graphe *N_3_exemple* dans un répertoire de notre PC.

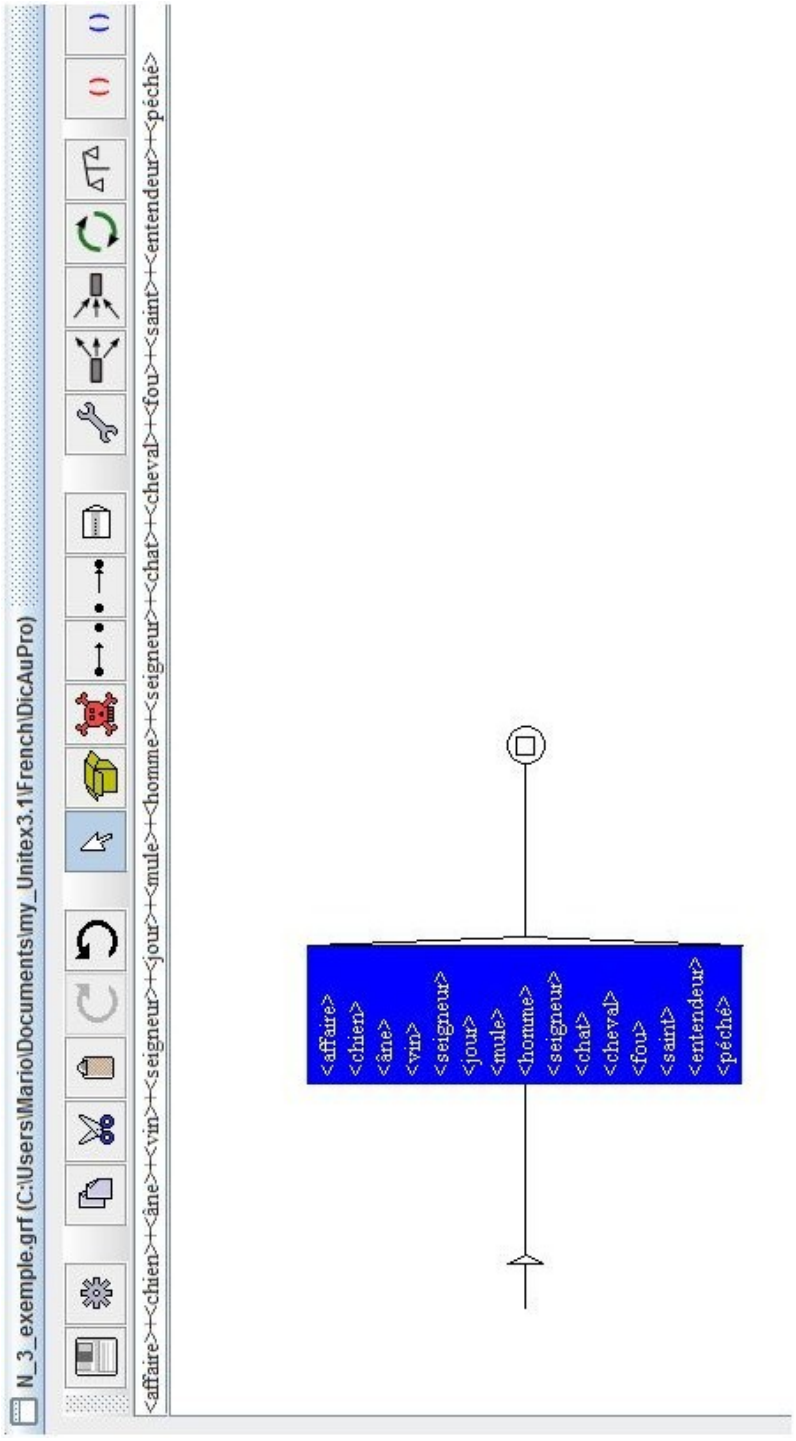


Figure 24. Sous-graphe *N_3_exemple*.

À ce stade, nous pouvons appeler *N_3_exemple* dans notre graphe qui s'enrichit, comme le montre la Figure 24 ci-dessous :

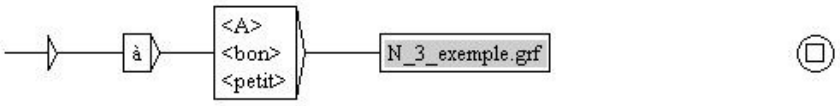


Figure 25. Exemple d'appel d'un sous-graphe.

Comme nous l'a suggéré l'expérience de Conenna (§ 4.2.2.1.), l'appel de sous-graphes nous servira à repérer des occurrences de parémies avec des insertions ou précédées par des introducteurs.

Ouvrons une parenthèse pour des considérations qui mettront en évidence quelques avantages de l'utilisation des graphes pour le repérage des parémies. Si l'on relie l'état appelant le sous-graphe à l'état final, le graphe en Figure 25 reconnaît 41 séquences dans le *LiPaF*. Nous en affichons une partie :

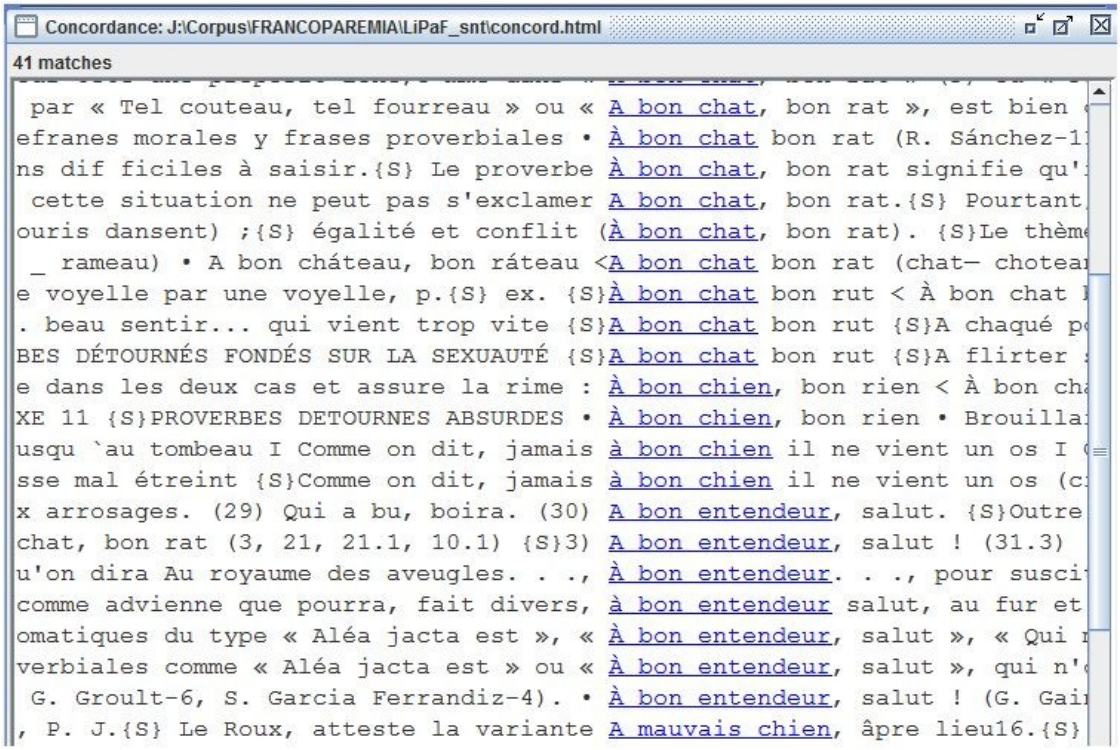


Figure 26. Concordances du *LiPaF* obtenues après l'application du graphe *exemple.grf* contenant un appel au sous-graphe *N_3_exemple.grf*.

Comme on peut constater de ces concordances, rien que notre simple graphe syntaxique donne des suggestions quant à *f* de certaines parémies (en l'occurrence, *À bon chat, bon rat* et *À bon entendeur, salut*).

Elles expliquent le fonctionnement de quelques variantes, notamment les *variantes graphiques* (§ 4.3.2.) (présence et/ou absence de signes de ponctuation).

En outre, ces concordances soulignent l'importance du processus de substitution d'unités (§ 3.3.3.) qu'Arnaud & Moon avaient déjà observé en français (§ 3.2.3.). Nous faisons allusion au détournement : *À bon chat, bon rut*. Il se forme ici par référence à la SLG-ac : {*À bon N bon N*} que nous avons identifiée dans notre corpus parémique.

Pour finir, on voit que « certains syntagmes », comme le disaient Cadiot & Visetti (2001) (§ 1.1.4.) (*incipit* ou non), sont des *parties proéminentes* ou des *noyaux phraséologiques résiduels* ou encore des *mots-clés* (§ 4.2.10.) non seulement pour une parémie, mais aussi pour tout le répertoire parémiologique. C'est le cas de /*à bon N*/ qui favorise la reconnaissance de *À bon chat*(,) *bon rat* ainsi que de *À bon chien bon rien* (détournement qui exploite encore {*À bon N bon N*}) et *Jamais à bon chien il ne vient un os*

(parémie, d'ailleurs, absente de notre corpus parémique). Nous avons une preuve empirique de l'existence de *collocations* et de *colligations parémiques* (§ 2.2.6.) qui finissent par se souder à tel point qu'elles sont interprétables comme des *préférences sémantiques parémiques* (§ 2.2.6.). De plus, nous avons une preuve ultérieure de la nécessité d'une description lexico-grammaticale préalable.

6.3.2. *Premiers exemplaires*[246]

Comme on l'a vu, les graphes et les manipulations que nous avons présentés ne requièrent ni une préparation informatique très avancée ni une formation approfondie sur les programmes d'*Unitex*. Notre souci est de concevoir une filière méthodologique qui permet de mettre au point des requêtes informatiques réexploitables et transposables à l'analyse d'autres répertoires parémiologiques. La conception de graphes syntaxiques complexes pourra représenter une étape successive de peaufinage qui passe par des expérimentations plus ciblées.

Or, à propos d'expérimentation, nous présenterons nos premiers exemplaires de graphes. Notre expérience a été quasi totalement empirique dans la mesure où nos connaissances des parémies et des techniques de repérage étaient très limitées.

Nous avons modélisé une liste de 32 parémies tirées de deux études de familiarité en Belgique (García Yelo 2009) et en France (Sevilla Muñoz & García Yelo 2008). Nous avons enrichi cette liste de 3 variantes que nous avons repérées dans *DicAuPro*. Par la suite, comme on l'a montré au § 5, nous avons annoté et lemmatisé ces parémies grâce à *TreeTagger* et corrigé quelques fautes.

Au départ, notre principe-clé a été le rétablissement de l'image formelle originaire (= forme canonique) des parémies. Nous avons distingué trois dimensions (ou facettes) linguistiques :

1. la dimension (ou facette) morphosyntaxique (dSyn) ;
2. la dimension (ou facette) lexicale (dLex) ;
3. la dimension (ou facette) linéaire (dLin) ou ordre des mots.

[246] Ce paragraphe réélabore quelques sections de Marcon (2013) décrivant une expérience qui remonte au 2011. Nous avons conduit une étude similaire déjà en juillet 2010 et dont les résultats sont présentés dans Marcon (2011).

Par la suite, nous avons pris en considération différents degrés de fidélité de ces dimensions par rapport à l'image parémique initiale. Une dimension a ainsi été :

- *annulée* (− −), à savoir ignorée ;
- *minimale* (−), c'est-à-dire présente, mais peu contraignante ;
- *élevée* (+), quand elle caractérise la description de la plupart des unités discrètes de chaque parémie ;
- *maximale* (+ +), si elle intéresse entièrement les chemins dessinés par les graphes, de leur état initial à leur état final.

Le réglage des degrés de fidélité pour chaque dimension nous a permis de mieux évaluer la précision des résultats renvoyés. En quelque mesure, la variation des degrés a assoupli la rigidité des chemins unidirectionnels des graphes, ce qui a amplifié (délibérément, vu les finalités exploratoires de notre étude) le bruit dans les données repérées. Le dépouillement manuel des résultats obtenus a focalisé notre attention sur certains processus qui seraient passés sous silence autrement. En particulier, le réglage des degrés a facilité non seulement la reconnaissance sur corpus des formes canoniques, mais aussi (et surtout) des variantes, des allusions, des détournements et de séquences formulaires ayant matrice parémique.

À l'aide de l'éditeur sous *Unitex*, nous avons modélisé 4 familles d'exemplaires de graphes que nous avons appelés :

1. prototypes lemmatisés (pLem) ;
2. prototypes partiels (pPart) ;
3. prototypes paradigmatiques (pPar) ;
4. prototypes syntaxiques (pSyn)
    a. à description générique (pSynG) ;
    b. à description fine (pSynF).

Chaque famille a combiné différents degrés de fidélité des trois dimensions (dLex, dSyn et dLin) :

	dLex	*dSyn*	*dLin*
prototypes lemmatisés (pLem)	+	−	++
prototypes partiels (pPart)	++	++	−
prototypes paradigmatiques (pPar)	++	++	+
prototypes syntaxiques (pSyn)			
à description générique **(pSynG)**	− −	+	++
à description fine **(pSynF)**	− −	++	++

Tableau 27. Prototypes et leur degrés dLex, dSyn et dLin respectifs.

Les prototypes lemmatisés ont envisagé la reconnaissance des formes canoniques, ainsi que le repérage des variantes et des actualisations morphosyntaxiques. La Figure 26 montre la forme lemmatisée de la parémie *Les bons comptes font les bons amis*. Aux prototypes lemmatisés nous avons confié la détection des variantes morphosyntaxiques.

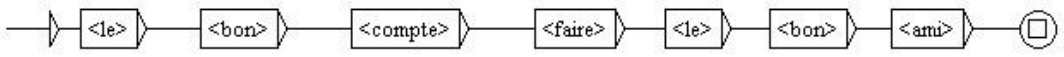


Figure 27. Exemple de prototype lemmatisé (pLem).

Dans nos intentions, les prototypes partiels comme celui de la Figure 27 auraient dû cibler les formes canoniques, les allusions et les détournements.

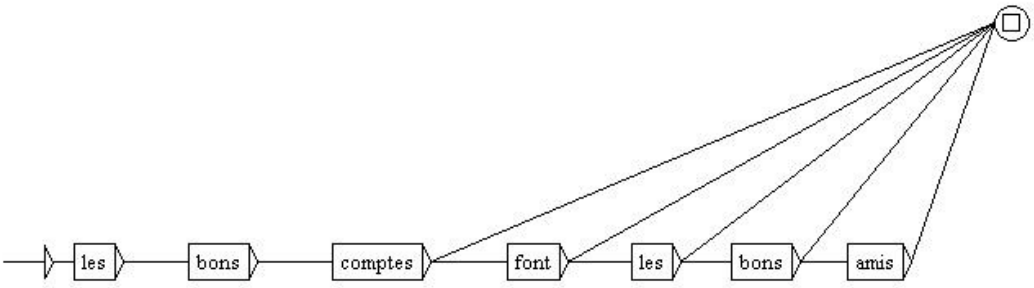


Figure 28. Exemple de prototype partiel (pPart).

Dans le cas de notre exemple, le prototype peut reconnaître les séquences :

Les bons comptes
Les bons comptes font
Les bons comptes font les
Les bons comptes font les bons
Les bons comptes font les bons amis

Telle famille de graphes aurait facilité l'observation des positions d'ancrage privilégiées des adjonctions et des substitutions (en gros, des césures de la séquence parémique), tout comme la reconnaissance de séquences formulaires à matrice parémique.

En revanche, les prototypes paradigmatiques ont exploité à l'extrême les contraintes de l'ordre linéaire et positionnelles. Outre les formes canoniques et les allusions, ce prototype était censé identifier des séquences formulaires (surtout non parémiques) générées à partir du matériel lexico-syntaxique de notre liste. Pour ces prototypes, nous avons aussi considéré les variantes graphiques, à savoir la présence et l'absence d'une virgule ou d'un apostrophe ainsi que la présence ou l'absence de la ligature *œ*/*oe*[247].

[247] Ce prototype de graphe montre au plan visuel la longueur moyenne (en nombre de mots graphiques) des proverbes de notre liste, à savoir entre 5 et 7 mots.

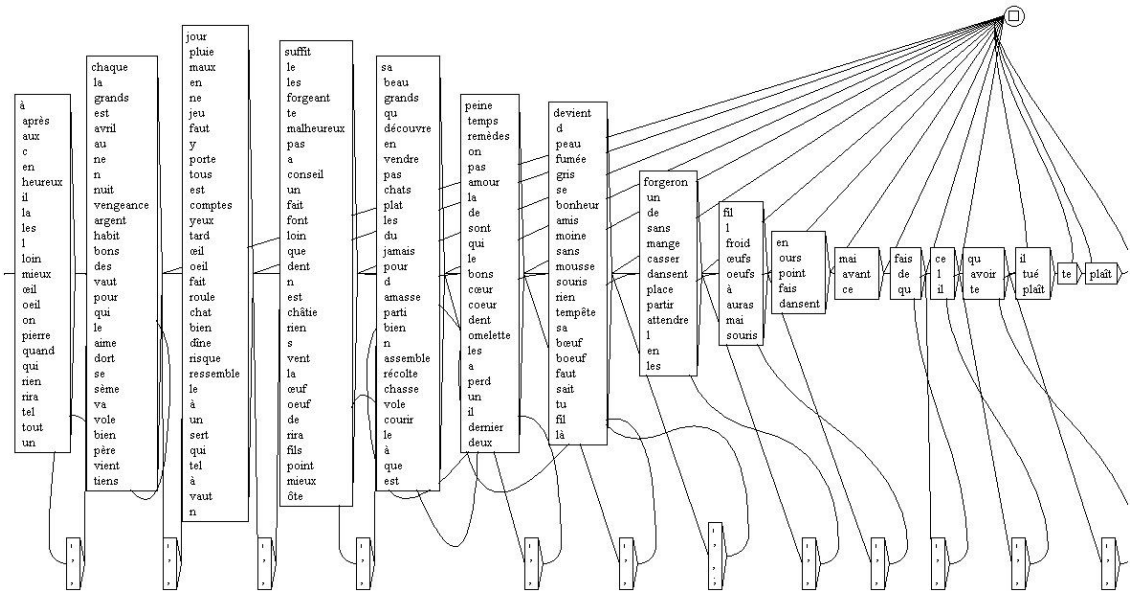


Figure 29. Exemple de prototype paradigmatique (pPar).

Pour finir, les deux familles de prototypes syntaxiques ont complètement annulé la dimension lexicale des parémies. Les deux se sont différenciés par la précision (pSynF) ou le manque de précision (pSynG) : du genre et du nombre pour les déterminants, les noms et les adjectifs d'une part ; de la personne et du nombre pour les verbes, d'autre part. Le potentiel présumé d'une requête tellement générique n'allait pas sans la conscience d'une désambiguïsation manuelle de tous les résultats.

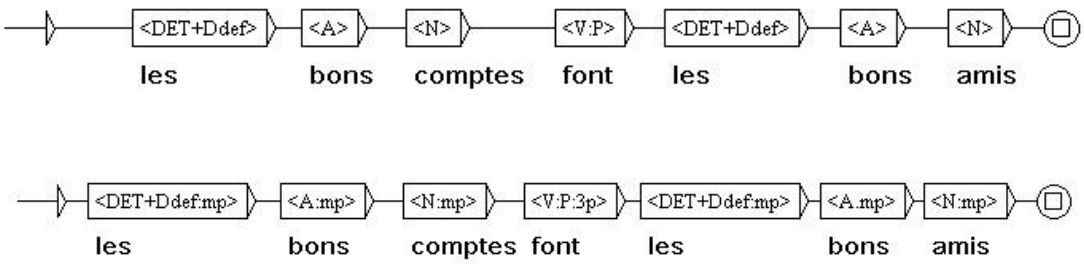


Figure 30. Exemples de prototypes syntaxiques (pSynG et pSynF).

Pendant une vingtaine de jours en 2011, nous avons lancé une recherche sur 36 sites de presse francophone grâce à *GlossaNet* (Fairon *et al*. 2008). Ce dernier utilise les DELA et

permet l'interrogation du Web (entre autres) par des graphes. Le Tableau 28 résume les occurrences totales renvoyées par chaque famille de prototypes.

	Résultats GlossaNet
prototypes lemmatisés (pLem)	124
prototypes partiels (pPart)	2.314
prototypes paradigmatiques (pPar)	3.556
prototypes syntaxiques (pSyn)	
à description générique **(pSynG)**	8.189
à description fine **(pSynF)**	6.625

Tableau 28. Nombre d'occurrences renvoyées par nos prototypes après la consultation de sites de presse en ligne à l'aide de GlossaNet.

La quantité des résultats obtenus par chaque famille de prototypes est inversement proportionnelle à leur précision. Seulement les prototypes lemmatisés ne montrent presque jamais de défaillances[248]. Son utilisation sera ainsi prioritaire dans nos requêtes.

Malgré le bruit produit, les prototypes partiels ont reconnu des allusions, des détournements et des candidates-parémies que les autres prototypes ont, pour la plupart, ignorés. Ces prototypes ont également détecté 2 parémies non mentionnées dans notre liste, et ce, grâce à la séquence déontique initiale *Il ne faut pas* qui caractérise la classe [*Il_Pro*$_{pers}$] (§ Annexe 6 – entrée syntaxique PRO:) et, en général, le répertoire parémiologique français. D'où l'importance de « certains syntagmes », comme nous l'avons rappelé à la fin du § 6.3.1. que nous avons interprété comme un indice pour mieux décrire les schémas lexico-grammaticaux propres aux parémies avant la modélisation des graphes.

Les prototypes paradigmatiques (visionnaires) n'ont créé que du bruit, tout comme les prototypes syntaxiques. En raison de l'absence du lexique ainsi que de l'étiquetage ambigu par les DELA, les prototypes syntaxiques ont offert les performances les moins encourageantes (Tableau 29). En tout cas, la reconnaissance de quelques parémies par pSynF nous a suggéré de mieux interpréter le rôle de la syntaxe. Une généralisation massive peut

[248] La différence entre nombre des résultats totaux (124) et occurrences totales des parémies reconnues (60) qu'on lit dans le Tableau 29 relève des doublons renvoyés par *GlossaNet*, non pas du repérage de séquences non pertinentes.

s'expliquer en termes de classification, comme le fait le Lexique-Grammaire, mais moins quand on recherche des données précises dans des textes.

	pLem	**pPart**	**pPar**	**pSynG**	**PSynF**	***f***
A chaque jour suffit sa peine	X	X				3
Après la pluie, le beau temps		X				1
Il n'y a pas de fumée sans feu	X					5
Il ne faut pas vendre la peau de l'ours avant de l'avoir tué	X	X				2
L'argent ne fait pas le bonheur	X	X				2
L'habit ne fait pas le moine	X					2
La vengeance est un plat qui se mange froid	X				X	3
Loin des yeux, loin du cœur	X	X				2
Mieux vaut tard que jamais	X				X	11
Œil pour œil dent pour dent	X	X				6
On ne fait pas d'omelette sans casser des œufs	X	X				1
Pierre qui roule n'amasse pas mousse	X					1
Qui se ressemble s'assemble	X	X				10
Qui sème le vent récolte la tempête	X	X	X	X	X	1
Rira bien qui rira le dernier	X	X				2
Tel père, tel fils	X					5
Tout vient à point à qui sait attendre	X	X				2
Un tiens vaut mieux que deux tu l'auras	X	X				1
						60

Tableau 29. Liste des parémies reconnues[249] par ordre alphabétique et *f* respectives avec indication (X) de la reconnaissance d'une parémie par une famille de prototypes.

Quoiqu'elle soit la seule, c'est l'occurrence de :

Qui sème le vent récolte la tempête

[249] Nous assemblons les formes canoniques, les allusions, les variantes graphiques et celles morphosyntaxiques.

qui nous a fourni, d'une part, une preuve de la viabilité de nos paramètres et de la bonne compilation des graphes. D'autre part, c'est cette occurrence (ainsi que les autres bonnes performances, bien évidemment) qui nous a motivé à creuser l'interface lexique-syntaxe et sa relation avec la linéarité de la séquence. Ce dont nous avons tenu compte au moment de la description et de la classification lexico-grammaticales de notre corpus parémique (§ 5).

6.3.3. *Modélisation des graphes*

Comme l'a montré notre expérience (§ 6.2.2.) ainsi que la littérature en la matière (§ 4.2.2.), le caractère local des graphes requiert un équilibre entre spécificité et finesse, d'une part, et généricité et souplesse, d'autre part. Le rapprochement de la forme exacte recherchée ne peut se faire ni au détriment de l'une ou de l'autre facette (en l'occurrence, lexicale et syntaxique) ni sans prendre en compte les agencements les plus récurrents qui caractérisent les parémies. Nous avons décidé d'ancrer la modélisation de nos graphes dans la description et dans la classification systématique des parémies à partir de leurs propriétés lexico-syntaxiques prototypiques (§ 5). Elle représentera le pivot autour duquel nous établirons les états de tous les chemins de reconnaissance.

1) D'une part, elle adopte la *f* d'occurrence comme critère servant à déceler des SLG récurrentes, c'est-à-dire des séquences d'unités lexicales et de parties du discours qui se co-sélectionnent et occupent des positions stables. Toutes nos SLG (mais surtout les SLG-ac et les SLG-g assimilées à celles-ci) sont des exemples de *lexico-grammaticalisation* : c'est le répertoire parémiologique qui révèle leur stabilité syntaxique et sémantique. Nous codifierons dans nos graphes cette stabilité lexico-grammaticale de cooccurrence (qu'on pourrait éventuellement caractériser comme semi-figement).
2) D'autre part, notre classification suit la linéarité des séquences parémiques, vu que chaque niveau relève de la position *n* qu'occupent une partie du discours et unité lexicale dans les séquences analysées. C'est pourquoi nous utiliserons la classification également pour de fins documentaires, à savoir pour organiser notre *bibliothèque de graphes* dans notre PC.

6.3.3.1. Quelques repères fondamentaux

Néanmoins, fixons dès maintenant des lignes directrices qui seront valables tout au long de la modélisation.

A. Chaque graphe regroupera les parémies d'après leurs classes lexico-grammaticales. Nous trierons manuellement les résultats.
B. En raison des performances excellentes observées, la lemmatisation de tout état sera incontournable ; elle favorisera surtout la reconnaissance de toutes les variantes syntagmatiques (graphique, orthographique et morphosyntaxique).
C. La partie du discours remplacera les unités lexicales les plus récurrentes à un état donné si et seulement si des SLG-ap et des SLG-ac plus génériques décrivent un nombre plus élevé de SLG-ap et de SLG-ac plus spécifiques. Par exemple, dans le cas de [*À Adj*] et notamment des SLG-ac :

   {*À Adj N N*} (3)
   {*À bon entendeur N*} (2)

   nous privilégierons la modélisation de {*À Adj N N*} parce que cette SLG-ac subsume {*À bon entendeur N*}.
D. Par le même souci de généricité et de précision des résultats, la finesse de la description syntaxique s'arrêtera aux parties du discours employées pour la classification. Nous prendrons éventuellement en compte seulement le temps et le mode des verbes. Toutes les autres informations flexionnelles et morphologiques ne seront pas encodées.
E. Les graphes n'auront aucune sortie.
F. Tout graphe fera appel à deux sous-graphes :
   a. un sous-graphe appelé *intro.grf* contenant des introducteurs ;
   b. un sous-graphe dénommé *punct.grf* qui inclut des signes de ponctuation ;

   Ils seront insérés après tout état à cause de l'imprévisibilité des césures de la séquence parémique. Pour le sous-graphe *intro.grf*, nous adapterons ceux que Conenna a conçus dans son étude de 2004 (§ 4.2.2.1.).

G. Seront éventuellement créés d'autres sous-graphes comprenant les unités lexicales (en position 2 et suivantes) des parémies d'une classe lexico-grammaticale. Ce qui améliorera la lisibilité du graphe.

6.3.3.2. Sous-graphe *intro.grf*

Partons par la modélisation du sous-graphe *intro.grf* qui s'inspire des sous-graphes PERFORM et INS de Conenna (2004 : 97-98, § 4.2.2.1.). Il suivra tout état, et ce, à partir du deuxième niveau de classification. L'appel du sous-graphe *intro.grf* s'arrêtera à l'avant-dernier état de chaque chemin que nous dessinerons (§ 6.3.3.5.).

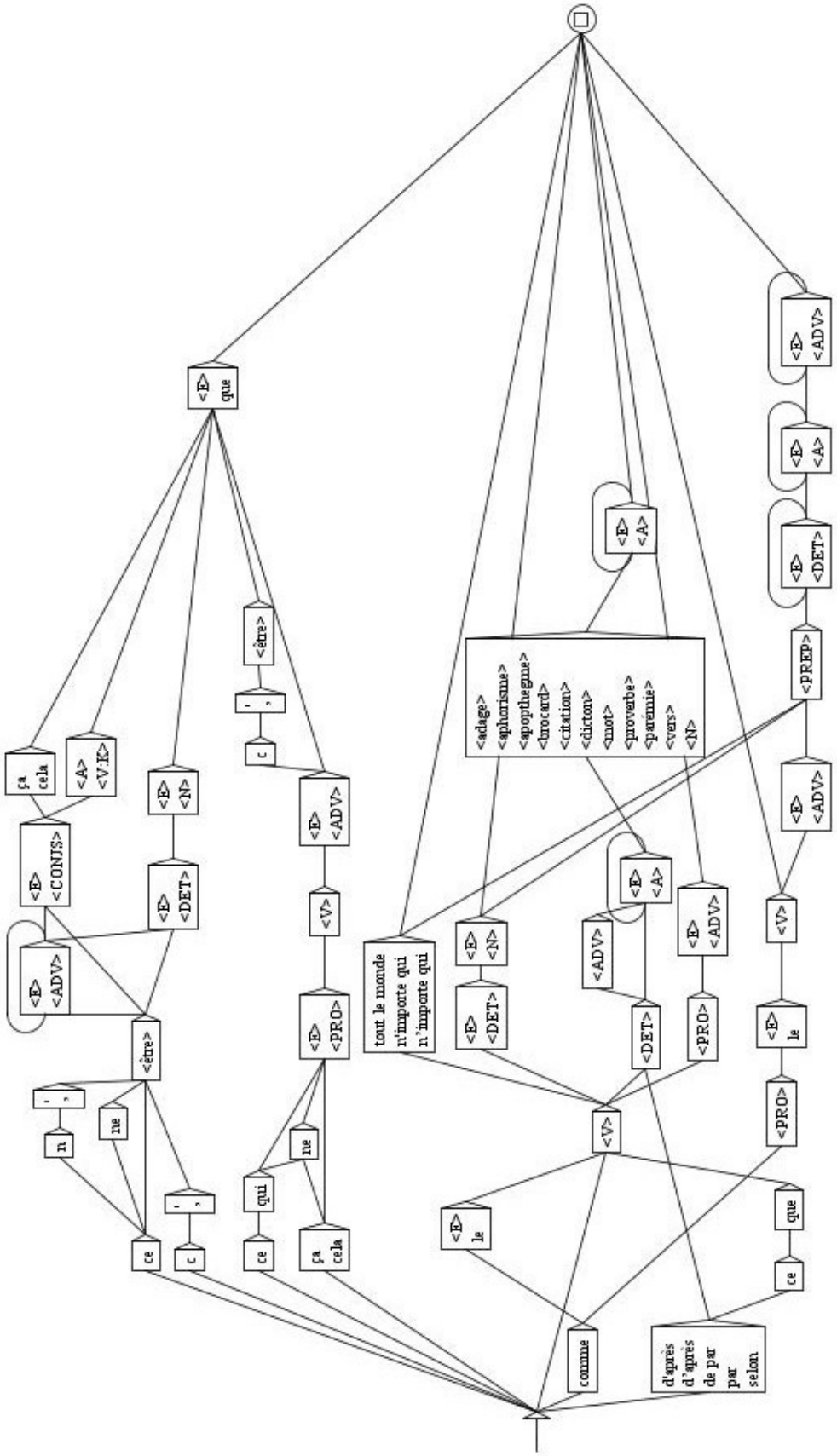


Figure 31. Sous-graphe *intro.grf*.

Nous avons condensé les deux graphes de Conenna en un seul. Par rapport aux graphes originaires, entre autres :

- nous avons effacé l'état initial contenant quelques signes de ponctuation ; il sera remplacé par le sous-graphe *punct.grf* (§ 6.3.3.3.) ;
- nous avons également effacé les chemins :

  qu'est-ce que (*tu*+*vous*) (<*E*>+*t'*) <*imaginer:2*>
  qu'est-ce que (*tu*+*vous*) (<*E*>+*vous*)
  (<*croire:2*>+<*imaginer:2*>+<*penser:2*>+<*vouloir:2*>)

  D'une part, nous les avons considérés trop spécifiques. D'autre part, d'après notre expérience de dépouillement manuel des concordances, nous n'avons jamais rencontré cette insertion.
- en revanche, nous avons ajouté d'autres chemins, c'est-à-dire :
  - celui qui commence par (*ce*+*c'*) <*être*> qui prend en compte la variation morphologique du pronom et la flexion du verbe ;
  - le chemin :

    ce qui (<E>+*ne*) (<E>+<PRO>) <V> (<E>+<ADV>) (<E>+*c'est*)

    dont la généricité pronominale bien satisfait la variété des insertions ;
  - tous les chemins qui commencent par :

    d'après ce que <*V*>

  - ainsi que tous les chemins commençants par <V> pour toutes les incises avec inversion de l'ordre non marqué des constituants ;
- nous avons réduit le nombre de formes lexicales prévues (notamment des déterminants, des pronoms et des verbes) et les avons remplacées par leurs parties du discours correspondantes. Cette montée en généricité syntaxique reste, pourtant, encadrée par les unités lexicales dans les états en début (et parfois en fin) de chemins. Nous avons gardé et enrichi l'état contenant différentes étiquettes parémiologiques

pour des raisons de lisibilité des chemins. D'ailleurs, pour ne pas restreindre la variation paradigmatique, nous avons ajouté le plus générique <N>.

- dans la même logique, nous avons effacé les appels aux sous-graphes contenant classes d'adjectifs et de noms (§ 4.2.2.1.). Certes, ils sont pertinents, mais redondants et remplaçables par <A> et <N>.
- nous avons en outre enrichi quelques chemins prévus par Conenna avec :
    - la possibilité de reconnaissance de la conjonction *que* en fin de certains chemins ;
    - ajouts de reconnaissance de la négation pour certains chemins ;
    - la reconnaissance d'espaces (<E>) supplémentaires ;
    - des états pour la reconnaissance d'adverbes et de conjonctions ;
    - de quelques reliures pour permettre la reconnaissance de deux ou plusieurs espaces ou parties du discours.

Loin d'être un graphe exhaustif ou meilleur de ce qu'a proposé Conenna, le sous-graphe *intro.grf* essaie plutôt de satisfaire notre observation empirique quant aux césures des séquences parémiques.

De plus, il cherche à établir un équilibre pondéré entre actualisation lexicale et généricité syntaxique, où la précision de la première encadre (en début et fin de chemins) la platitude de la deuxième. S'il est vrai que la généricité créera du bruit dans les résultats, elle nous assurera de ne pas passer à côté d'occurrences créatives de la séquence parémique.

6.3.3.3. Sous-graphe *punct.grf*

La Figure 31 ci-dessous illustre le simple sous-graphe *punct.grf* que nous insérerons :

- comme pour le sous-graphe *intro.grf*, après l'état qui représente la modélisation du deuxième niveau de notre classification lexico-grammaticale ;
- devant et après chaque état appelant le sous-graphe *intro.grf*.

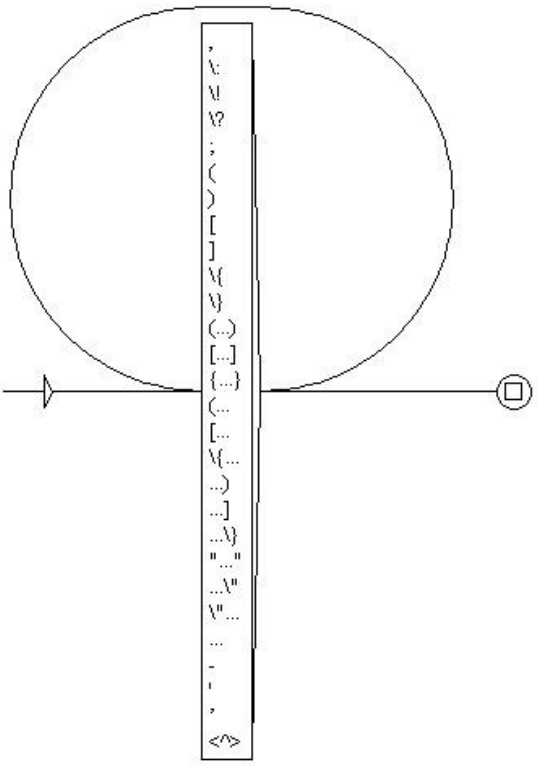


Figure 32. Sous-graphe *punct.grf*.

Il se compose d'un seul état avec reliure. L'état inclut la quasi-totalité des signes de ponctuation que l'on pourrait rencontrer en début ou en pleine séquence parémique.

Le symbole spécial <^> reconnaît, par contre, des retours à la ligne.

Les parenthèses, les crochets et les accolades fermées, ouvertes ou cernant des points de suspension pourront nous aider à repérer des incises.

Les points de suspension ont été prévus pour des allusions, voir des réductions de la séquence parémique.

Pour finir, les guillemets peuvent encore détecter des incises sous forme de citations, mais aussi nous indiquer que notre séquence parémique est employée comme citation dans un texte.

6.3.3.4. Graphes des parémies : l'exemple de la classe [*À Adj*]

Passons maintenant à illustrer la création des graphes des parémies. Nous poursuivons notre exemple de la classe [*À_Prép*] (§ 5.3.) pour montrer le rôle central de la classification lexico-grammaticale pour la modélisation.

Rappelons que la classe [*À_Prép*] prévoit les sous-classes de deuxième niveau :

- [*À Adj*] (31 parémies) ;
- [*À N*] (24 parémies) ;
- [*À Dét*$_{art}$] (21 parémies) ;
- [*À Adv*] (5 parémies) ;
- [*À Pro*$_{ind}$] (5 parémies) ;
- [*À Nam*] (3 parémies) ;

Nous créerons ainsi un répertoire *À_Prép* dans notre PC où nous sauvegarderons les 6 graphes *À Adj*, *À N*, *À Dét*$_{art}$, *À Adv*, *À Pro*$_{ind}$ et *À Nam*.

Commençons par le graphe de la classe [*À Adj*] et ouvrons la grille de classification (§ Annexe 6 – entrée syntaxique PREP) ainsi que la liste correspondante de parémies extraite à l'aide du tableau croisé dynamique sous *Excel*. Le premier état est occupé par la préposition *à*. Nous ajoutons un deuxième état qui contient <A> pour la reconnaissance de tous les adjectifs. Nous suivons donc le principe D (§ 6.3.3.1.) de prédilection de la généricité syntaxique à l'actualisation lexicale. Ce malgré les SLG-ap :

{*À bon*} (7)
{*À chaque*} (5)
{*À tout*} (3)
{*À dur*} (2)
{*À mauvais*} (2)
{*À petit*} (2)

saturent lexicalement 21 parémies sur 31.

Figure 33. Graphe de la classe [*À Adj*] *:* modélisation des positions 1 et 2 des SLG correspondantes.

Le troisième niveau de la classification est rempli par [*À Adj N*] qui assemble 29 parémies et par [*À Adj Adj*] qui en regroupe 2.

- Pour ce qui concerne les parémies de [*À Adj N*], nous suivons le même raisonnement qui nous a conduit à modéliser les deux premières positions. Les SLG-ap que nous avons observées sont :

  {*À bon N*} (7) [{*À bon entendeur*} (3), {*À bon vin*} (2)]
  {*À chaque N*} (5)
  {*À Adj chien*} (2)
  {*À Adj seigneur*} (2)
  {*À dur N*} (2)
  {*À petit N*} (2)
  {*À tout N*} (2)

  qui actualisent quelques membres de la classe. Par conséquent, nous saisissons <N> dans le troisième état.
- Les parémies :

  À tout bon compte revenir
  À brusquin brusquet

  partagent tout simplement la SLG-g initiale qui renvoie à leur classe. Nous encodons <A>.

Nous distinguons deux chemins et obtenons ce que montre la Figure 33 ci-dessous :

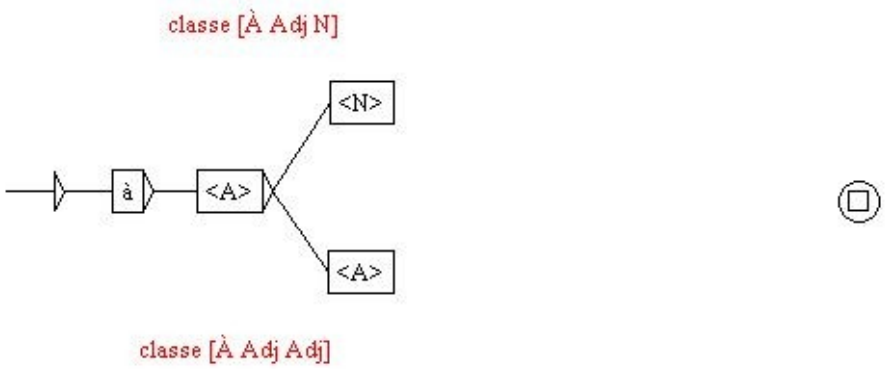


Figure 34. Graphe de la classe [*À Adj*] *:* modélisation des sous-classes [*À Adj N*] et [*À Adj Adj*].

Regardons ce qui se passe maintenant au quatrième niveau de la classe [*À Adj*]. La sous-classe [*À Adj N*] se fragmente en d'autres sous-classes :

- [*À Adj N Adj*] (13 parémies) ;
- [*À Adj N N*] (6 parémies) ;
- [*À Adj N Adv*] (4 parémies) ;
- [*À Adj N* $Dét_{poss}$] (3 parémies) ;
- [*À Adj N* V_{pres}] (2 parémies) ;

qui réunissent 28 des 29 parémies appartenant à [*À Adj N*]. La parémie :

À mauvais chien on ne peut montrer le loup

ne rentre dans aucune autre sous-classe. Quant aux deux parémies de [*À Adj Adj*] :

À brusquin brusquet
À tout bon compte revenir

elles ne constituent pas une sous-classe de quatrième niveau. Que faire de ces parémies ?

- Pour les sous-classes de [*À Adj N*], nous appliquons le même principe d'évaluation sur la saturation lexicale de la partie du discours la plus terminale de la séquence en nous appuyant sur les SLG-ac et SLG-ap repérées. Voyons-les en détail.
    - [*À Adj N Adj*] : les SLG-ap décrivent 7 sur 13 parémies. Nous ajoutons l'état <A>.
    - [*À Adj N N*] : la SLG-g assimilée à une SLG-ac {*À Adj N N*} rend compte de 3 parémies, dont 2 sont représentées par la SLG-ac {*À bon entendeur N*}. Dans ce cas, nous ajoutons l'état <N> lié à l'état final qui nous aidera à détecter ces 3 parémies. En revanche, pour les 3 parémies :

      À dure enclume, marteau de plume
      À quelque chose malheur est bon
      À vieille mule frein doré

      nous procédons à la création d'états lemmatisés qui encodent la séquence jusqu'à épuisement. Ce qui est en continuité avec le principe B (§ 6.3.3.1.).
    - [*À Adj N Adv*] : les SLG-ap {*À Adj N peu*} et {*À Adj N point*} comblent le potentiel syntaxique de la SLG-g. Le quatrième état s'actualise lexicalement avec les formes *peu* et *point*.
    - [*À Adj N Dét$_{poss}$*] : comme il est arrivé à [*À Adj N Adv*], la SLG-ap {*À chaque N son*} retrace toutes les parémies qui appartiennent à cette sous-classe. Le quatrième état correspondant contient ainsi la forme lemmatisée <*son*>. L'adjectif *chaque* est déjà pris en compte par le deuxième état <A> de notre graphe.
    - [*À Adj N V$_{pres}$*] : de même que pour [*À Adj N Dét$_{poss}$*], la SLG-ap {*À chaque N V$_{pres}$*} signale les deux parémies qui relèvent de cette sous-classe. Le quatrième état implique seulement <V:P>, c'est-à-dire la reconnaissance d'un verbe au présent de l'indicatif. En vue de la détection de variantes morphosyntaxiques, nous y ajoutons (de manière, certes, redondante) <V> pour ne pas contraindre à tout prix le mode et le temps. Nous tendons vers la description la moins fine,

comme déclaré par notre principe E (§ 6.3.3.1.). L'adjectif *chaque* est représenté par <A> dans le deuxième état de notre graphe.

Quant à la parémie sans sous-classe :

À mauvais chien on ne peut montrer le loup

nous lemmatisons chaque unité jusqu'à sa fin en ligne avec notre principe B (§ 6.3.3.1.).

- Les parémies :

À brusquin brusquet
À tout bon compte revenir

n'engendrent aucune sous-classe de quatrième niveau. Comme pour les autres cas de parémies sans sous-classes ultérieures, nous lemmatisons les éléments qui restent jusqu'au bout *À tout bon compte revenir* et relions le chemin :

à <A> <A>

à l'état final (§ 6.3.3.1.).

En conclusion de la modélisation du quatrième niveau de classification de [*À Adj*], le graphe prend la forme qui est reproduite ci-dessous :

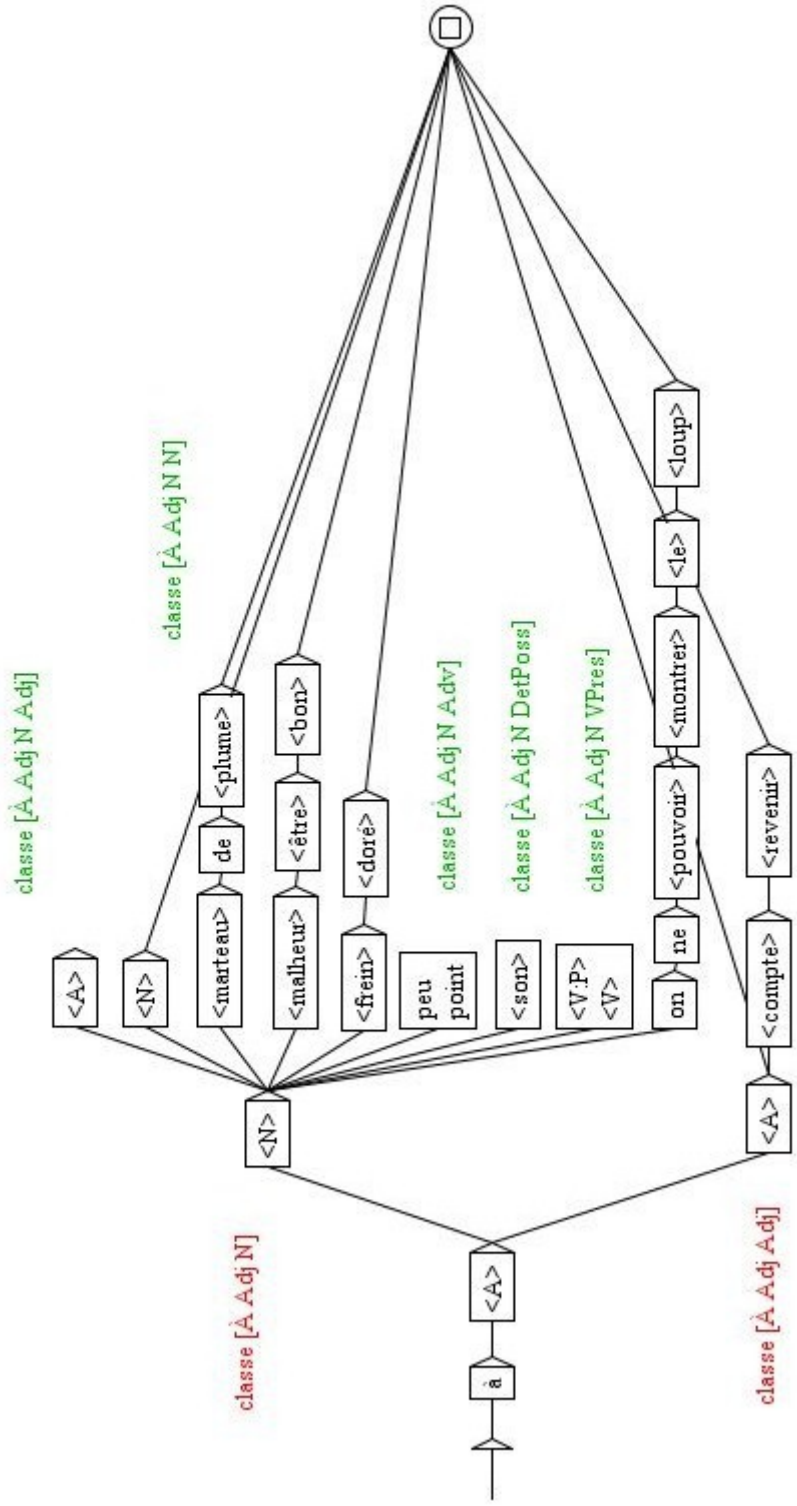


Figure 35. Graphe de la classe [*À Adj*] *:* modélisation des sous-classes appartenant au 4^e^ niveau de la classification lexico-grammaticale.

Concluons notre graphe de la classe [*À Adj*] faisant référence aux niveaux qui restent à préciser et aux chemins à clôturer.

- Prenons [*À Adj N Adj*] et notons que ses 13 parémies sont décrites par la SLG-g assimilée à une SLG-ac : {*À Adj N Adj N*}. Nous dessinons ainsi un cinquième état avec <N> et le relions à l'état final.
- Nous avons discuté le cas de [*À Adj N* $Dét_{poss}$ *N*] au § 5.3 : {*À chaque N son N*} est une SLG-ac pour 2 parémies, alors qu'elle est une SLG-ap pour la troisième. Nous insérons un état qui contient <N> et l'enchaînons à l'état final. Pour la parémie restante (*À chaque oiseau son nid est beau*), nous saisissons le lemme de toute unité lexicale dans des états séparés et la connectons à l'état final.
- Quant à [*À Adj N Adv*], on remarque que les SLG-ac {*À Adj N peu de paroles*} et {*À Adj N point de N*} remplissent entièrement la dernière SLG-g [*À Adj N Adv Prép N*]. Par conséquent, nous rentrons les états qui manquent et renvoient aux SLG-ap.
- Au-delà de [*À Adj N* V_{pres}], il n'y a pas d'autres sous-classes. Par conséquent, les unités qui restent des parémies :

  À chaque jour suffit sa peine
  À chaque porc vient la Saint-Martin

  seront lemmatisées.

D'après notre classification de la classe [*À Adj*], le graphe se compose comme suit :

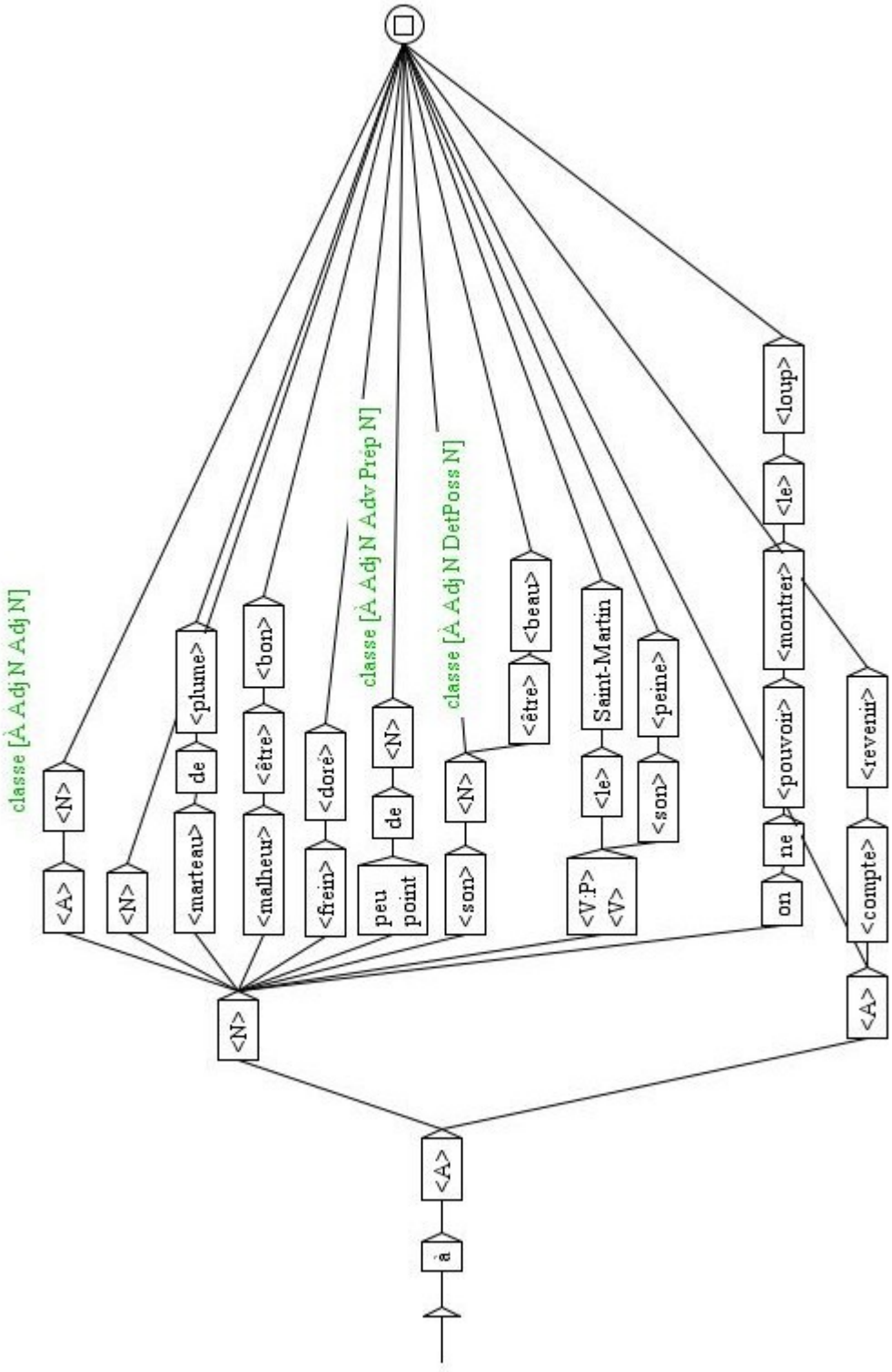


Figure 36. Graphe de la classe [*À Adj*] d'après notre classification lexico-grammaticale.

Il ne reste qu'à intégrer les sous-graphes *intro.grf* et *punct.grf* aux endroits que nous avons mentionnés aux §§ 6.3.3.1.-6.3.3.2. Nous introduisons les premiers états :

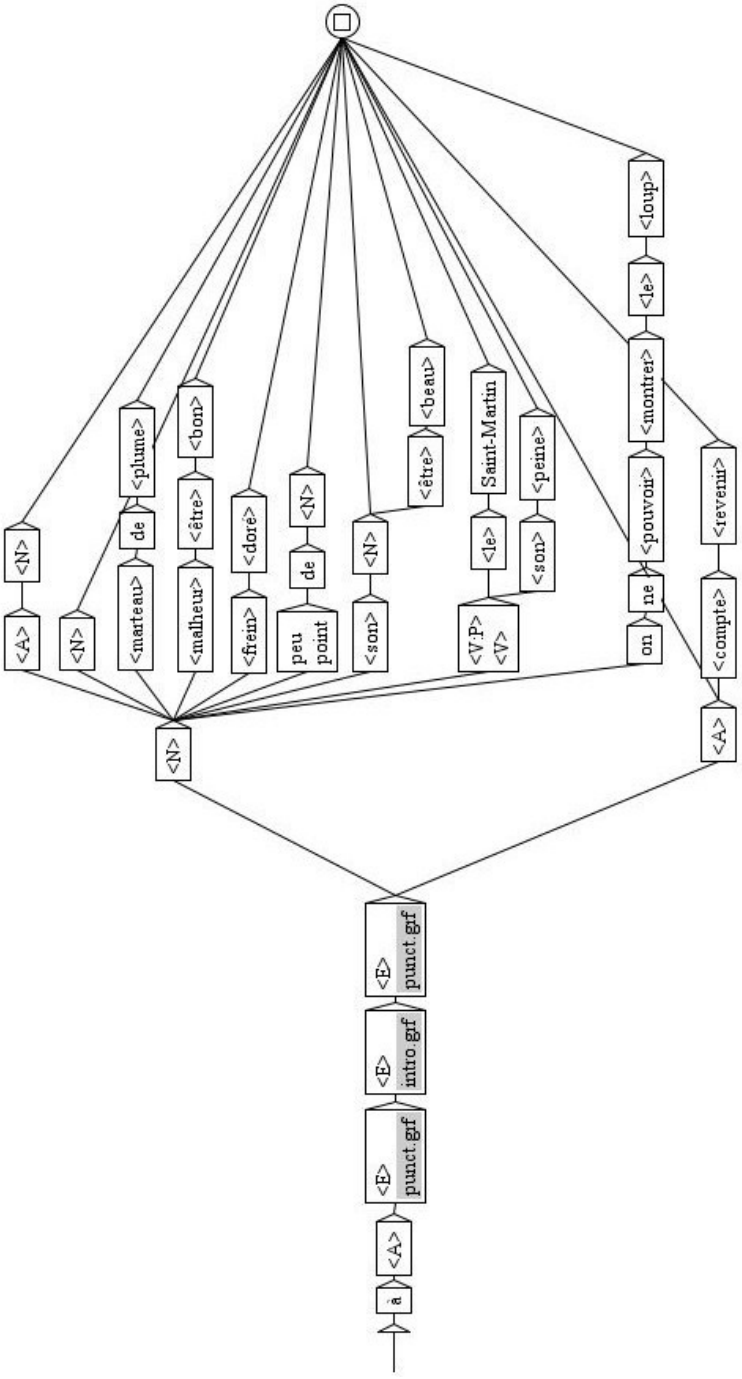


Figure 37. Graphe de la classe [*À Adj*] : insertion des appels aux sous-graphes *intro.grf* et *punct.grf*.

On voit que, au fur et à mesure, la triplette d'états appelant les sous-graphes *intro.grf* et *punct.grf* endommagerait la lisibilité des chemins et compliquerait la gestion du graphe en soi. Pour pallier ce problème, nous créons le graphe *introponct.grf* qui se chargera d'appeler la triplette en question.

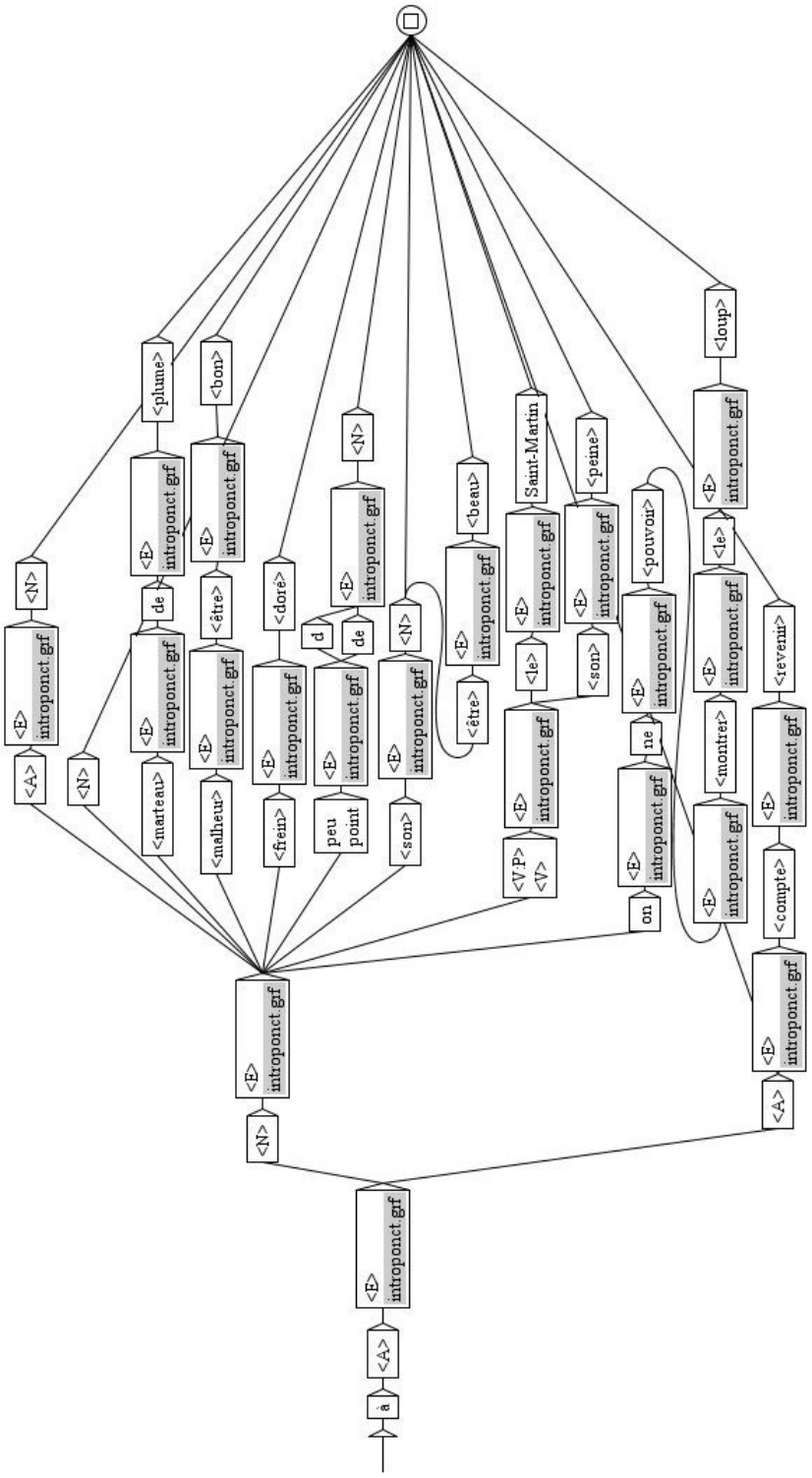


Figure 38. Graphe de la classe [*À Adj*] : insertion du sous-graphe *introponct.grf*.

Figure 39. Sous-graphe *introponct.grf*.

Nous avons ainsi une imbrication de graphe : *À Adj.grf* appelle *introponct.grf* comme sous-graphe qui, à son tour, appelle *punct.grf* et *intro.grf*.

En Figure 38, nous avons un modèle de graphe qui veut repérer le nombre le plus élevé d'occurrences de parémies (en l'occurrence, commençant par /*À Adj*/)[250]. Loin d'atteindre une précision optimale, ce modèle (et notre démarche, en général) exploite l'ambiguïté des DELA d'*Unitex* et provoque délibérément du bruit dans les résultats. En même temps, nous avons veillé à l'atténuer par des contraintes de cooccurrence que nous avons identifiées en amont dans notre classification[251].

D'après le programme d'exploration des chemins d'un graphe (Paumier 2013 : 136-138), *À Adj.grf* compte 1.408 chemins, excepté les appels aux sous-graphes. Pour les 31 parémies qui appartiennent à la classe [*À Adj*], nous avons une étendue combinatoire décidément plus importante que celle d'une recherche menée grâce à des expressions régulières. L'imprévisibilité de l'usage parémique que tous les parémiologues ont souligné (§§ 3, 4). Pour y faire face, nous préférons gérer du bruit (filtré au préalable par une description systématique des parémies) dans les concordances (qu'on peut trier) plutôt qu'être confronté au silence (non réel) du corpus, comme beaucoup d'autres parémiologues.

Rien n'empêche qu'à partir de notre modèle on puisse :

- réduire le nombre de chemins et calibrer les requêtes sur une ou deux SLG ;

[250] Nous aurions pu également créer un *automate de séquence* (Paumier 2013 : 194-196) à partir d'une sélection raisonnée de SLG dans un *corpus de séquences* ou *corpus qualifié* (Paumier 2013 : 194). C'est d'ailleurs ce que nous avons fait pour d'autres classes.

[251] Rien n'empêche de créer des graphes suivant le modèle de Conenna pour chaque parémie ou encore d'ajuster notre modèle en reprenant, par exemple, la sortie anticipée à une boîte précise ou en exploitant les paradigmes des mots à telle ou telle autre position dans une classe lexico-grammaticale, comme nous l'avons dans nos prototypes (§ 6.3.2.).

- réduire l'ambiguïté des DELA. Par exemple, l'interrogation du *LiPaF* avec le graphe *À Adj.grf* renvoie 365 résultats. La plupart d'entre eux est à attribuer à la généricité de la partie initiale /*À Adj*/ et à l'étiquetage : *un,.A+z2* dans le DELA *motsGramf-.bin*[252]. Par le recours au symbole spécial **~** servant à exclure les codes sémantiques (ainsi que grammaticaux et flexionnels) des DELA (Paumier 2013 : 79), nous excluons la reconnaissance de *un,.A+z2* par une négation de la reconnaissance des adjectifs encodés avec le code sémantique *z2*. Ce qui baisse les résultats renvoyés par *Unitex* à un total de 132 et simplifie la lecture des concordances[253].

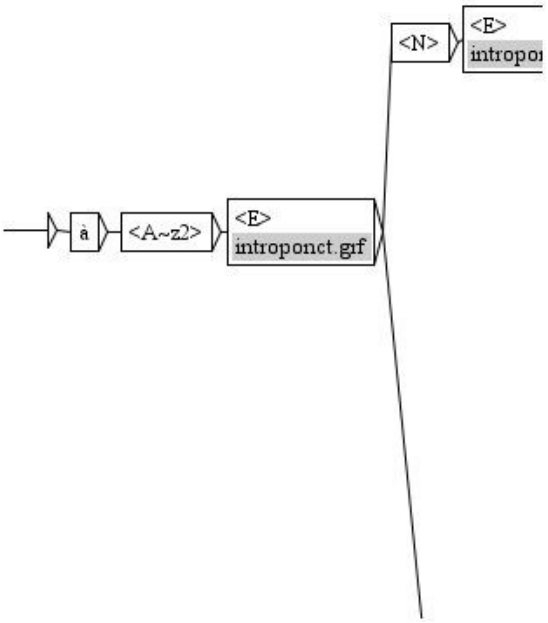


Figure 40. Exemple de négation de patron.

- multiplier les chemins par des reliures d'un état à lui-même ou par la création de sous-graphes pour reconnaître plus aisément, par exemple, des adjonctions.

Bien évidemment, toutes ces modifications dépendent des finalités qui motivent l'interrogation du corpus.

[252] On peut consulter les DELA sous *Unitex* à l'aide du programme *Lookup* (Paumier 2013 : 53).
[253] Ce type d'opération demande une consultation soignée des DELA, le risque étant d'ignorer des occurrences.

6.3.3.5. Modélisation des parémies non classées : quelques remarques

Avant de conclure, nous consacrons quelques mots au pourquoi de la lemmatisation pour les parémies qui 's'égarent' au cours de notre classification ou qui n'y sont pas rentrées (199, pour la plupart commençant par un nom et par un verbe).

¶Entrée syntaxique	Total parémies	Parémies sans classe
ADJ	96	17
ADV	122	15
DET	3	1
DET:ART	400	//
DET:POS	1	1
INT	2	2
KON	62	2
NAM	8	3
NOM	196	110
NUM	11	4
PRO	2	2
PRO:DEM	55	//
PRO:IND	49	//
PRO:PER	377	//
PRO:REL	138	1
PRP	171	5
PRP:det	27	//
VER:futu	1	1
VER:impe	27	17
VER:infi	16	10
VER:pper	2	2
VER:pres	10	5
VER:subp	1	1

Tableau 30. Comparaison entre le total des parémies (première colonne) et les parémies sans classe lexico-grammaticale (deuxième colonne) d'après leur entrée syntaxique.

Tout d'abord, celles que nous avons montrées durant notre modélisation ne 's'égarent' pas vraiment dans la mesure où elles appartiennent toujours du moins à une classe. Pour ce qui a trait à toutes les parémies sans classe (Tableau 30) (et que nous modélisons avec lemmatisation), elles dépendent quand même d'une entrée syntaxique. Tout notre corpus parémique est ainsi organisé, quoiqu'il ne soit pas entièrement classé. bœuf

Répétons que notre classification se veut une manière autre d'aborder les parémies, et ce, pour découvrir les cadres/schémas lexico-grammaticaux en œuvre ainsi que pour leur modélisation informatique. L'absence de ces 199 parémies est conséquente à *f* observée dans notre corpus parémique qui nous offre un indice de productivité de telle ou telle autre SLG. Or, les taux d'exclusion très élevés des noms et des verbes nous signalent que la contrainte de *f* posée sur l'entrée lexicale pourrait nuire à la description lexico-grammaticale de ces parémies (§ Conclusions). On remarque, par exemple, que la séquence {*N* V_{pres} *N*} représente l'intégralité de 13 parémies dont 7 ont V_{pres} = *passer*. Ce qui est indépendant de *f* d'une

entrée lexicale. En tout cas, il est évident que la modélisation informatique d'une séquence pareille aurait produit trop de bruit dans les résultats. Comme nous voulons assurer le repérage le plus correct et vaste de ces parémies, nous choisissons ainsi de faire recours à la lemmatisation qui – on le rappelle – a obtenu les meilleures performances dans notre expérience et dans la littérature en la matière (§ 4). À la lemmatisation nous ajoutons aussi nos sous-graphes *intro.grf*, *punct.grf* et *introponct.grf*.

6.4. En guise de conclusion

Nous avons illustré les deux corpus de français contemporain pour calculer *f* de notre corpus parémique. En continuité avec notre cadre méthodologique fédérateur (§ 3.3.), nous avons choisi et décrit les deux sous-corpus de presse en ligne de 2009 (*fra_news_2009_1M-text*) et de 2010 (*fra_news_2010_1M-text*) de la collection française Leipzig. Les soins de documentation, de prétraitement (malgré les fautes de tokenisation que nous avons repérées, § 6.2.1.) et l'échantillonnage suivant la notion de *phrase graphique* nous ont paru fiables du côté informatique et linguistique. L'attention à la presse en ligne nous aidera à combler le vide de la littérature parémiologique francophone en matière de liste de f des parémies basée sur corpus. La mesure de la *f* par la phrase graphique plutôt que par l'occurrence du mot graphique, mais aussi la taille globale d'environ 100 millions d'occurrences que nous avons souvent rencontrées (§ 3) nous permettent de tester la validité des plages de *f* reconnues (§ 3.3.3.) et le besoin d'une optique parémiométrique (§ 7).

À côté, nous avons constitué un corpus spécialisé de littérature parémiologique francophone (*LiPaF*). Il nous supportera pour évaluer les parémies que les parémiologues préfèrent ou choisissent de mentionner dans leurs articles scientifiques. Malgré la taille réduite, nous croyons qu'il nous donnera des indications de tendance d'usage au sein de la communauté parémiologique. Il sera intéressant de comparer les liste de *f* issues des corpus pour mieux comprendre le rôle que joue la compétence parémiologique sur la variété d'usage des parémies. Le croisement des tendances entre grand public et parémiologues sera en ce sens révélateur.

Nous avons présenté le logiciel *Unitex*. Nous avons illustré les programmes que nous utiliserons pour interroger nos corpus et bien défini nos paramètres de recherche, de compilation de l'index des concordances et de prétraitement des corpus. Ceux-ci sont étiqueté

avec les DELA que nous avons aussi créés pour assurer la reconnaissance de la totalité des occurrences de notre corpus parémique.

Par la suite, nous avons approfondi les manipulations que l'on peut envisager dans l'éditeur de grammaires locales ou graphes syntaxiques. Nous avons insisté sur la nécessité de conjuguer le caractère local (quoique souple) des graphes et l'incertitude de la fixité de la séquence parémique dans son intégralité en discours. Ce qui est vrai pour les deux axes syntagmatique et paradigmatique.

Ancrer la modélisation dans la classification des parémies conçue à partir de leurs propriétés lexico-syntaxiques d'agencement récurrentes représente cette recherche d'équilibre entre la précision et la variation. Les SLG issues du répertoire parémiologique sont en effet des exemples de *lexico-grammaticalisation* qui révèlent à la fois stabilité et productivité syntaxiques et sémantiques. Leur modélisation assure, d'une part, de saisir chaque parémie à la forme souhaitée et, d'autre part, de laisser la porte ouverte à des variantes.

Nous avons illustré notre démarche de création des graphes guidée par notre classification. Ce qui a abouti à un modèle qui joue avec l'interface lexique-syntaxe des SLG parémiques ainsi qu'avec l'ambiguïté des DELA d'*Unitex*. Il engendre délibérément du bruit dans les résultats en raison, aussi, des combinatoires qu'il peut reconnaître pour prévoir (pour autant que possible) l'imprévisible parémique. Quoique nous puissions réduire le nombre de chemins et contrôler l'étiquetage du DELA, nous gérerons ce bruit lors du tri des concordances calculées à partir de nos corpus (§ 7). De cette façon, nous estimons extraire le nombre le plus élevé d'occurrences, et ce, malgré d'autres chemins de reconnaissance (d'adjonctions, par exemple) soient envisageables.

CHAPITRE 7

FREQUENCE ET USAGE DES PAREMIES. UNE APPLICATION

Le présent chapitre sera consacré à l'application de notre cadre fédérateur, de notre classification lexico-grammaticale et de nos modèles de grammaires locales à une partie de notre liste parémiographique (§ 7.1.) pour estimer la fréquence de chaque parémie dans les deux corpus Leipzig et dans le corpus *LiPaF*. Nous interpréterons autant les fréquences obtenues que quelques occurrences parémiques repérées (§ 7.2.)

7.1. Fréquence sur corpus : un test pour notre cadre fédérateur

Suivant nos travaux précédents, nous nous concentrerons sur l'établissement des *f* des 32 parémies et 3 variantes mentionnées au § 6.3.2. Nous souhaitons porter attention à ces parémies parce qu'elles représentent le banc d'essai idéal non seulement pour comparer et compléter les résultats des enquêtes de familiarité avec les études de fréquence sur corpus, mais aussi pour tester rapidement la validité de notre cadre méthodologique fédérateur (§ 3.3) et de nos requêtes informatiques (§ 6).

Voici donc la liste des parémies :

1. *À chaque jour suffit sa peine*	17. *Œil pour œil, dent pour dent*
2. *Après la pluie, le beau temps*	18. *On ne fait pas d'omelette sans casser des œufs*
3. *Aux grands maux les grands remèdes*	19. *Pierre qui roule n'amasse pas mousse*
4. *C'est en forgeant qu'on devient forgeron*	20. *a) Quand le chat est parti, les souris dansent* *b) Quand le chat n'est pas là, les souris dansent*
5. *a) En avril ne te découvre pas d'un fil, en mai fais ce qu'il te plaît* *b) En avril, n'ôte pas un fil ; en mai, fais ce qu'il te plaît*	21. *Qui aime bien châtie bien*
6. *Heureux au jeu, malheureux en amour*	22. *Qui dort dîne*
7. *Il ne faut pas vendre la peau de l'ours avant de l'avoir tué*	23. *Qui ne risque rien n'a rien*
8. *Il n'y a pas de fumée sans feu*	24. *Qui se ressemble s'assemble*
9. *La nuit porte conseil*	25. *Qui sème le vent récolte la tempête*
10. *La nuit, tous les chats sont gris*	26. *Qui va à la chasse perd sa place*
11. *La vengeance est un plat qui se mange froid*	27. *Qui vole un œuf vole un bœuf*
12. *L'argent ne fait pas le bonheur*	28. *Rien ne sert de courir il faut partir à point*
13. *Les bons comptes font les bons amis*	29. *Rira bien qui rira le dernier*
14. *L'habit ne fait pas le moine*	30. *Tel père, tel fils*
15. *Loin des yeux, loin du cœur*	31. *Tout vient à point à qui sait attendre*
16. *Mieux vaut tard que jamais*	32. *a) Un tiens vaut mieux que deux tu l'auras* *b) Un « tiens » vaut mieux que deux « tu l'auras »*

Tableau 31. Liste des 32 parémies (et 3 variantes) sélectionnées pour notre étude de *f* et tirée de Marcon (2013).

Notre étude pilote datant 2011 (pour la période 9-30 juin) nous a suggéré une corrélation potentielle entre fréquence et familiarité (Marcon 2013 : 310-312). Si l'on regarde aux données de *f* brute et l'on trie notre liste par ordre décroissant de *f*, on obtient le tableau suivant :

Parémies	***f***
Mieux vaut tard que jamais	11
Qui se ressemble s'assemble	10
Œil pour œil dent pour dent	6
Il n'y a pas de fumée sans feu	5
Tel père, tel fils	5

A chaque jour suffit sa peine	3
La vengeance est un plat qui se mange froid	3
Il ne faut pas vendre la peau de l'ours avant de l'avoir tué	2
L'argent ne fait pas le bonheur	2
L'habit ne fait pas le moine	2
Loin des yeux, loin du cœur	2
Rira bien qui rira le dernier	2
Tout vient à point à qui sait attendre	2
Après la pluie, le beau temps	1
On ne fait pas d'omelette sans casser des œufs	1
Pierre qui roule n'amasse pas mousse	1
Qui sème le vent récolte la tempête	1
Un tiens vaut mieux que deux tu l'auras	1

Tableau 32. Liste de *f* des parémies au Tableau 31 dans Marcon (2013).

Dans notre corpus dynamique, on observe les tendances que nous avons décrites au § 3.3.3 :

- 17 parémies ont $f = 0$, soit environ 50% de notre liste, ce qui confirme le *silence parémique* dans l'usage ;
- la distribution des *f* des parémies est assez homogène, c'est-à-dire que la plupart des parémies repérées se regroupent autour des mêmes valeurs chiffrées basses ;
- la plage de valeurs $1 \leq f \geq 5$ rend compte de 15 sur 18 parémies reconnues ;
- les 2 parémies-*outliers* en tête de liste suggèrent des situations extralinguistiques similaires et pointent des dynamiques situationnelles ou des évaluations similaires sur le fond et récurrentes dans notre corpus.

Même si nous avons travaillé avec un corpus dynamique dont nous n'avons pu apprécier la taille (la collecte en ligne ne comptabilisant pas les occurrences), il s'avère que les tendances majeures que nous avons fait ressortir de la littérature témoignent d'une normalité d'usage parémique : il fallait donc s'attendre à des résultats pareils dans une optique parémiométrique.

Comme on croit que la modélisation des requêtes peut jouer elle aussi sur le décompte de nos parémies, comparons maintenant les données précédentes avec *f* dans les corpus *Leipzig Corpora Collection* (*fra_news_2009_1M-text* et *fra_news_2010_1M-text*) et de notre corpus *LiPaF*. Dans ce cas de figure, notre liste sera triée par ordre alphabétique. Par rapport

à notre étude pilote, nous distinguerons en outre la *fréquence d'usage standard*, c'est-à-dire toutes les occurrences de la forme exacte recherchée, de la *fréquence d'usage créatif*, à savoir toutes les occurrences parémiques qui utilisent un (ou plusieurs) processus de variation (adjonction, réduction, permutation et substitution paradigmatique/flexionnelle). Nous tenons à préciser que, lors du repérage des parémies, nous avons fait recours à la fois à nos graphes ainsi qu'aux expressions régulières. La recherche par expressions régulières nous a servi seulement de vérification pour l'ajout éventuel d'occurrences passées sous silence par les graphes. Dans les cas où il a été envisageable, nous avons accordé une préférence particulière à la saisie des SLG-g, SLG-ap et SLG-ac dégagées de notre corpus parémique (voir tables dans l'Annexe 6).

Parémies	***f* Leipzig 2009**		***f* Leipzig 2010**		***f* LiPaF**		***Total f*** **(par parémie)**
	f **us**	***f*** **uc**	***f*** **us**	***f*** **uc**	***f*** **us**	***f*** **uc**	
À chaque jour suffit sa peine	**2**	**0**	**0**	**0**	**8**	**4 (ana)**	**14**
Après la pluie, le beau temps	**1**	**1 (adj)**	**1**	**0**	**8**	**3 (adj)**	**14**
Aux grands maux les grands remèdes	**1**	**1 (subpar)**	**1**	**1 (subpar)**	**8**	**4 (ana)**	**16**
C'est en forgeant qu'on devient forgeron	**0**	**0**	**1**	**0**	**14**	**12 (ana)**	**27**
a) En avril ne te découvre pas d'un fil, en mai fais ce qu'il te plaît *b) En avril, n'ôte pas un fil ; en mai, fais ce qu'il te plaît*	**0**	**0**	**0**	**0**	**a) 7** **b) 0**	**a) 5 (réd)** **4 (réd + ana)** **b) 0**	**16**
Heureux au jeu, malheureux en amour	**0**	**0**	**0**	**0**	**3**	**1 (subpar)**	**4**
Il ne faut pas vendre	**1**	**3 (ana)**	**0**	**2 (ana)**	**13**	**9 (ana)**	**34**

la peau de l'ours avant de l'avoir tué		**1 (réd) 1 (subpar)**				**4 (subpar)**	
Il n'y a pas de fumée sans feu	**0**	**2 (ana)**	**4**	**0**	**17**	**3 (ana) 1 (perm)**	**27**
La nuit porte conseil	**2**	**0**	**2**	**0**	**6**	**0**	**10**
La nuit, tous les chats sont gris	**0**	**1 (réd+ ana)**	**0**	**1 (ana)**	**11**	**2 (ana)**	**15**
La vengeance est un plat qui se mange froid	**2**	**1 (subpar)**	**0**	**1 (subpar)**	**7**	**3 (réd) 1 (subpar)**	**15**
L'argent ne fait pas le bonheur	**1**	**1 (ana) 1 (subpar)**	**5**	**0**	**28**	**3 (adj)**	**39**
Les bons comptes font les bons amis	**0**	**0**	**1**	**0**	**6**	**0**	**7**
L'habit ne fait pas le moine	**1**	**3 (ana) 2 (adj + subpar)**	**2**	**0**	**44**	**2 (adj)**	**54**
Loin des yeux, loin du cœur	**0**	**1 (ana)**	**0**	**2 (ana) 1 (réd) 1 (réd+adj)**	**12**	**2 (adj) 2 (ana)**	**21**
Mieux vaut tard que jamais	**6**	**1 (ana)**	**7**	**0**	**8**	**1 (réd)**	**23**
Œil pour œil, dent pour dent	**2**	**0**	**2**	**0**	**4**	**0**	**8**
On ne fait pas d'omelette sans casser des œufs	**0**	**1 (adj)**	**1**	**0**	**7**	**3 (ana)**	**12**
Pierre qui roule n'amasse pas mousse	**0**	**0**	**0**	**0**	**37**	**5 (ana) 3 (subpar) 2 (adj)**	**47**
a) Quand le chat est	**0**	**0**	**a) 0**	**0**	**a) 4**	**a) 2**	**21**

parti, les souris dansent b) Quand le chat n'est pas là, les souris dansent			**b) 2**		**b) 11**	**(subpar), 1 (ana) b) 1 (ana)**	
Qui aime bien châtie bien	**1**	**0**	**1**	**0**	**23**	**1 (adj) 1 (subpar)**	**27**
Qui dort dîne	**0**	**0**	**0**	**0**	**40**	**2 (subpar)**	**42**
Qui ne risque rien n'a rien	**0**	**0**	**2**	**0**	**0**	**0**	**2**
Qui se ressemble s'assemble	**0**	**0**	**1**	**0**	**9**	**0**	**10**
Qui sème le vent récolte la tempête	**0**	**0**	**1**	**0**	**10**	**1 (ana)**	**12**
Qui va à la chasse perd sa place	**0**	**0**	**0**	**0**	**34**	**5 (ana)**	**39**
Qui vole un œuf vole un bœuf	**0**	**1 (ana)**	**0**	**0**	**28**	**11 (ana)**	**40**
Rien ne sert de courir il faut partir à point	**1**	**1 (réd)**	**1**	**1 (réd)**	**12**	**2 (subpar) 1 (réd) 1(ana)**	**20**
Rira bien qui rira le dernier	**2**	**0**	**2**	**0**	**13**	**0**	**17**
Tel père, tel fils	**3**	**0**	**1**	**1 (ana)**	**25**	**11 (ana) 11 (subpar)**	**52**
Tout vient à point à qui sait attendre	**2**	**1 (subpar)**	**0**	**0**	**4**	**2 (réd) 1 (adj)**	**10**
a) Un tiens vaut mieux que deux tu l'auras b) Un « tiens » vaut	**0**	**0**	**a) 1 b) 0**	**0**	**a) 10 b) 8**	**a) 2 (ana) b) 0**	**21**

mieux que deux « tu l'auras »							
***Total f* (par corpus)**	**28**	**24**	**39**	**11**	**479**	**135**	**716**

Tableau 33. *f* comparées de la liste des parémies au Tableau 31 (tri par ordre alphabétique) [Légende : adj = adjonction ; ana = analogie formelle/sémantique ; réd = réduction ; perm = permutation ; subflex = substitution flexionnelle ; subpar = substitution paradigmatique].

7.2. Application de notre cadre fédérateur : quelques remarques

Le Tableau 33 ci-dessus présente une vue d'ensemble sur les *f* brutes de notre liste de parémies dans nos 3 corpus. Nous avons repéré 716 occurrences parémiques qui se répartissent comme suit :

- 52 occurrences pour le corpus Leipzig 2009 ;
- 50 occurrences pour le corpus Leipzig 2010 ;
- 614 occurrences pour le corpus *LiPaF*.

Comme nous l'avons anticipé, nous avons distingué la *fréquence d'usage standard* de la *fréquence d'usage créatif*. Les résultats montrent que l'*autorité de la norme* a une emprise décidément plus significative que l'*autorité de l'usage* : 546 occurrences relèvent de la fréquence d'usage standard alors que seulement 170 de la fréquence d'usage créatif. Plus précisément, si l'on assiste à un tiraillement entre les deux autorités dans les corpus Leipzig, la préférence pour la forme exacte parémiographique est en revanche nette dans le *LiPaF*. Qu'elle serve ou non en guise d'appui, par exemple, pour argumenter sur le détournement parémique, il est évident que la source parémiographique, ou plutôt, qu'une *perception parémiographique* (Marcon, sous évaluation) de la séquence-parémie est centrale dans le discours parémiologique (et dans l'esprit des parémiologues) beaucoup plus que dans le discours de presse que représentent les corpus Leipzig.

Par rapport à notre cadre fédérateur, notamment aux tendances observées dans la littérature à l'égard de *f* des parémies, ainsi que par rapport à la nature de nos corpus :

1) une partie de la liste de parémies a $f = 0$ dans les deux corpus de la collection Leipzig. Quoiqu'on travaille sur deux corpus de taille comparable (environ 50

millions d'occurrences, § 6.6.1.), le pourcentage des **parémies à *f* zéro** varie dans une plage de valeurs plutôt que s'axer autour d'une valeur précise. Le silence parémique touche 12/(32 + 3) parémies pour le corpus de 2009 et 8/(32 + 3) pour le corpus de 2010, respectivement 34% et 22% de la liste de parémies. On pourrait donc en conclure qu'environ 30% des parémies d'une liste filtrée à la fois sur la base d'études de familiarité et de ressources parémiographiques, résultera absente dans des corpus généraux, notamment des corpus de presse en ligne, d'environ 50 millions d'occurrences. Bien évidemment, il s'agit d'une tendance potentielle qu'il faudrait valider par la répétition de cette expérience à l'aide de corpus pareils. Cette expérience confirme encore que les *parémies à f zéro* ne sont pas méconnues : nos parémies résultent d'enquêtes sur la familiarité et n'ont pas donc disparu. Ce qui est d'ailleurs prouvé par leur vivacité dans le *LiPaF* où 1 parémie seulement (*Qui ne risque rien n'a rien*) à $f = 0$. Par une démonstration à l'extrême, les occurrences parémiques du *LiPaF* dévoilent que le silence parémique est à corréler à l'*ancrage pragmatique* (Schmale 2013 : 36-37) que les parémies peuvent repérer dans les textes du corpus. De plus, les résultats issus d'un corpus d'environ 1 million d'occurrences (§ 6.1.2.) comme le *LiPaF*, montrent que les tendances quantitatives d'usage à l'écrit doivent nécessairement être liées à la taille du corpus et à la nature des textes qui le composent ;

2) comme nous l'avons constaté dans la littérature, la distribution des *f* des parémies s'assemble autour de quelques valeurs : 1, 2 ou 3 occurrences pour les corpus Leipzig. On constate tout de suite que cette tendance ne semble valable que pour les corpus généraux. La distribution des *f* dans le *LiPaF* suggère que cette tendance ne vaut pas pour des corpus hautement spécialisés, comme l'est le *LiPaF* ;
3) pour les corpus Leipzig, plus de 50% des parémies ayant $f \geq 1$ privilégient la plage $1 \leq f \geq 5$ pour les corpus de taille jusqu'à 50 millions d'occurrences. Les résultats reflètent les données repérées dans la littérature. S'écarte évidemment de ce raisonnement le *LiPaF* à cause du discours spécialisé qu'il représente ;
4) dans les trois distributions des *f* on reconnaît des parémies-*outliers* qui diffèrent dans les 3 corpus. D'une part, les corpus Leipzig enregistrent un total de 14 occurrences (7 respectivement en 2009 et en 2010) pour *Mieux vaut tard*

que jamais. Encore, 6 occurrences en 2009 pour *Il ne faut pas vendre la peau de l'ours avant de l'avoir tué* et *L'habit ne fait pas le moine*, alors qu'en 2010 on compte 5 occurrences pour *L'argent ne fait par le bonheur*. D'autre part, les parémies *L'habit ne fait pas le moine* (46 occurrences), *Pierre qui roule n'amasse pas mousse* (45 occurrences) et *Qui dort dîne* (42 occurrences) dépassent le seuil des 40 occurrences dans le *LiPaF* et s'écartent des parémies restantes.

Dans nos corpus généraux, les *f* des parémies qui se situent dans les plages de valeurs mentionnées au 3) sont attendues, voir normales. Il s'agit de *f* probables et prévisibles pour ces séquences formulaires. L'invitation à adopter donc une *optique parémiométrique* se voit confirmée et la disparition des parémies (familières, d'ailleurs, dans notre cas de figure) dans l'usage écrit est bien conjurée.

Si nous argumentions encore à partir des données de la colonne *Total f* (*par parémie*) du Tableau 33, nous donnerions une interprétation sommaire et surtout fautive de l'usage des parémies, le décompte étant visiblement altéré par les résultats du *LiPaF*. Nous préférons discuter les performances des corpus Leipzig par rapport au corpus dynamique que nous avons constitué pour notre étude pilote. Ces trois corpus sont en effet comparables en termes de discours (presse française), malgré leurs différences de constitution que nous n'oublions guère. Pour cette comparaison, nous prendrons en compte la somme de la *fréquence d'usage standard* et de la *fréquence d'usage créatif* dans les corpus Leipzig.

Parémies	***f*** **Leipzig 2009**	***f*** **Leipzig 2010**	***f*** **(Marcon 2013)**
Mieux vaut tard que jamais	7	7	11
Qui se ressemble s'assemble	0	1	10
Œil pour œil dent pour dent	2	2	6
Il n'y a pas de fumée sans feu	2	4	5
Tel père, tel fils	3	2	5
A chaque jour suffit sa peine	2	0	3
La vengeance est un plat qui se mange froid	3	1	3
Il ne faut pas vendre la peau de l'ours avant de l'avoir tué	6	2	2
L'argent ne fait pas le bonheur	3	5	2
L'habit ne fait pas le moine	6	2	2
Loin des yeux, loin du cœur	1	4	2
Rira bien qui rira le dernier	2	2	2
Tout vient à point à qui sait attendre	3	0	2
Après la pluie, le beau temps	2	1	1

On ne fait pas d'omelette sans casser des œufs	1	1	1
Pierre qui roule n'amasse pas mousse	0	0	1
Qui sème le vent récolte la tempête	0	1	1
Un tiens vaut mieux que deux tu l'auras	0	1	1
Aux grands maux les grands remèdes	2	2	0
C'est en forgeant qu'on devient forgeron	0	1	0
a) En avril ne te découvre pas d'un fil, en mai fais ce qu'il te plaît *b) En avril, n'ôte pas un fil ; en mai, fais ce qu'il te plaît*	0	0	0
Heureux au jeu, malheureux en amour	0	0	0
La nuit porte conseil	2	2	0
La nuit, tous les chats sont gris	1	1	0
Les bons comptes font les bons amis	0	1	0
a) Quand le chat est parti, les souris dansent *b) Quand le chat n'est pas là, les souris dansent*	0	2	0
Qui aime bien châtie bien	1	1	0
Qui dort dîne	0	0	0
Qui ne risque rien n'a rien	0	2	0
Qui se ressemble s'assemble	0	1	0
Qui va à la chasse perd sa place	0	0	0
Qui vole un œuf vole un boeuf	1	0	0
Rien ne sert de courir, il faut partir à point	2	2	0

Tableau 34. *f* comparées des parémies dans le corpus dynamique de notre étude pilote et dans les corpus Leipzig (tri par ordre de fréquence décroissant d'après les résultats issus de Marcon (2013) dans la dernière colonne à droite).

Ce parcours en microdiachronie dans le discours de la presse en ligne que nous restitue le Tableau 34, renvoie à une distribution dans l'usage où *Mieux vaut tard que jamais* représente la parémie par excellence, voir le ***stéréotype parémique*** privilégié par ce discours. En général, avec la seule exception notable de *Qui se ressemble s'assemble*, il y a un noyau parémique assez stable dans l'usage écrit de la presse francophone qui correspond à presque toutes les parémies repérées dans notre étude pilote :

A chaque jour suffit sa peine
Après la pluie, le beau temps
Il n'y a pas de fumée sans feu
Il ne faut pas vendre la peau de l'ours avant de l'avoir tué
L'argent ne fait pas le bonheur
L'habit ne fait pas le moine
La vengeance est un plat qui se mange froid

Loin des yeux, loin du cœur
Mieux vaut tard que jamais
Œil pour œil dent pour dent
On ne fait pas d'omelette sans casser des œufs
Rira bien qui rira le dernier
Tel père, tel fils
Tout vient à point à qui sait attendre

Autour de ce ***lexique parémique noyau*** composé de 14 séquences, dont l'occurrence est plus probable que celle d'autres parémies dans le discours de presse en ligne, tournent d'autres parémies. Ce qui est prouvé d'évidence par les *parémies à f zéro*, notamment :

a) En avril ne te découvre pas d'un fil, en mai fais ce qu'il te plaît *b) En avril, n'ôte pas un fil ; en mai, fais ce qu'il te plaît*
Heureux au jeu, malheureux en amour
Qui dort dîne
Qui va à la chasse perd sa place

qui ne comptabilisent aucune occurrence dans nos 3 corpus. Il en résulte qu'il y a des ***parémies 'usage-resistant'***, à savoir des parémies qui persistent à contourner l'usage à l'écrit. Leur absence est certes fonction de leur non-pertinence dans le discours, et ce malgré leur connaissance de la part de locuteurs natifs, cette dernière n'étant pas un facteur suffisant pour en déclencher l'usage. Si l'on peut motiver cette résistance en raison de la nature textuelle des corpus, il est non moins vrai que la familiarité n'assure pas l'usage. D'où la nécessité de ne pas laisser passer l'argument identitaire : *familiarité = contemporanéité* (et par conséquent insertion dans un minimum parémiologique) et de ne pas ignorer le message que l'usage nous suggère à ce propos. Autrement dit, la contemporanéité des parémies doit témoigner d'une convergence (et, pour autant que possible, d'une coïncidence) des résultats issus d'enquêtes sur la familiarité et sur la fréquence.

Il serait à mieux enquêter à l'avenir le rôle de ces *parémies* '*usage-resistant*' ainsi que des parémies :

Aux grands maux les grands remèdes
La nuit porte conseil
Rien ne sert de courir, il faut partir à point

ayant *f* = 2 dans les corpus Leipzig, mais absentes dans le corpus de notre étude pilote. Ce pour mieux comprendre si elles appartiendraient à ce lexique parémique noyau.

De même, il faudrait se pencher ultérieurement sur le silence des parémies commençant par le pronom *Qui* :

Parémies	***f* Leipzig 2009**	***f* Leipzig 2010**	***f* (Marcon 2013)**
Qui aime bien châtie bien	1	1	0
Qui dort dîne	0	0	0
Qui ne risque rien n'a rien	0	2	0
Qui se ressemble s'assemble	0	1	10
Qui se ressemble s'assemble	0	1	0
Qui sème le vent récolte la tempête	0	1	1
Qui va à la chasse perd sa place	0	0	0
Qui vole un œuf vole un boeuf	1	0	0

Tableau 35. *f* comparées des parémies commençant par le pronom *Qui* (tri par ordre alphabétique).

La classe lexico-grammaticale la plus représentée dans notre liste et mieux connue par les locuteurs natifs interviewés par Sevilla Muñoz et García Yelo en France et en Belgique, passe sous silence dans les corpus Leipzig. Nous avons déjà enregistré le *chuchotement* (Marcon 2011 : 112) de la classe *Qui* en 2010 : débutant avec une liste parémiographique de 68 parémies, nous n'en avons repéré que 7. La classe *Qui* est donc un cas exemplaire : comme la familiarité diverge sensiblement de l'usage à l'écrit, que faudrait-il en conclure en termes de contemporanéité ? Ajoutons à cela que notre liste actuelle ne mentionne pas la parémie *Qui vivra verra*, *parémie-outlier* de notre étude de 2010, comptant 12 occurrences dans un corpus dynamique conçu avec les mêmes critères que le corpus de Marcon (2013)[254]. Le cas des parémies en *Qui* exemplifie l'insuffisance des études de familiarité et de fréquence quand, isolées les unes des autres, elles servent d'appui pour des jugements de contemporanéité parémique et pour l'établissement d'un minimum parémiologique. De manière encore plus patente et urgente, les résultats concernant la classe *Qui* invitent les parémiologues à s'interroger sur l'épistémologie adoptée pour la réalisation de ces études et pour établir des

[254] Nous précisons que la parémie *Qui vivra verra* est mentionnées dans les enquêtes sur terrain de Sevilla Muñoz et García Yelo. Néanmoins, son nombre d'informateurs est inférieur par rapport au nombre d'informateurs des parémies filtrées dans notre liste.

corrélations entre familiarité et fréquence – comme l'a essayé Grzybek (§§ 2-3) – avant toute généralisation sur la disparition et la vitalité des parémies.

Certes, en microdiachronie, les corpus de presse que nous avons interrogés, signalent une chute dans l'usage à l'écrit d'une entrée lexicale (*Qui*) récurrente dans les répertoires parémiographiques. L'autorité de la norme et son discours parémique canonique diffèrent significativement de l'autorité de l'usage écrit.

Cette différence ressort encore quand on compare les parémies-*outliers* des corpus Leipzig et celles du *LiPaF*. D'un côté, donc, dans le discours de la presse, *Mieux vaut tard que jamais* est une parémie à la mode et un stéréotype parémique de ce discours, comme nous l'avons dit. De l'autre côté, *L'habit ne fait pas le moine*, *Pierre qui roule n'amasse pas mousse* et *Qui dort dîne* sont des stéréotypes parémiques du discours parémiologique. Néanmoins, par rapport à leur *f* dans les corpus Leipzig, nous pourrions en conclure qu'elles sont plutôt des ***parémies-fétiches*** qui ne reflètent pas ce qui arrive dans l'usage. Autrement dit, elles sont des stérétoypes parémiques qui se chronicisent dans le discours parémiologique, mais elles sont en même temps délégitimées par leur silence ou chuchotement dans l'usage. Cette déviance se fait frappante dans le cas de *Pierre qui roule n'amasse pas mousse* et de *Qui dort dîne* : face à aucune occurrence dans les corpus Leipzig, le premier compte 45 occurrences et le deuxième 42 occurrences dans le *LiPaF*[255]. Ce qui rejoint encore nos observations précédentes sur la classe *Qui* et sur la distance qui se crée entre le discours parémique canonique et ce qui ressort dans l'usage.

Pour conclure, nous nous pencherons sur les processus de variation qui ont contribué à l'estimation de la *f d'usage créatif*[256].

[255] Outre ces *parémies-outliers*, d'autres *parémies-fétiches* du discours parémiologique sont également *C'est en forgeant qu'on devient forgeron* ; *En avril ne te découvre pas d'un fil, en mai fais ce qu'il te plaît* ; *La nuit, tous les chats sont gris* et les parémies en *Qui*.

[256] L'analyse des processus de variation mériterait une recherche à part entière (§ Conclusions). Nous ne présentons que quelques exemples pour discuter des fréquences obtenues. Nous faisons recours aux occurrences des corpus Leipzig, vu que le *LiPaF* peut contenir des exemples de détournement parémique forgés *ad hoc* par les parémiologues.

Processus de variation	***f* Leipzig 2009**	***f* Leipzig 2010**	***f* *LiPaF***
Adjonction	2	0	14
Analogie (formelle et/ou sémantique)	12	6	77
Permutation	0	0	1
Réduction	2	2	12
Substitution paradigmatique	5	2	27
Combinaison de 2 ou plusieurs processus	3	1	4

Tableau 36. *f* d'usage créatif des parémies reparties par processus de variation et comparées dans chaque corpus.

Le Tableau 36 laisse entendre que l'ordre des mots est la facette la plus stable des séquences parémiques. Ce qui ne revient pas à dire que la permutation soit impossible dans l'absolu, mais plutôt que les locuteurs tendent à l'éviter. Nous précisons que ce résultat ne peut être vraiment imputé à notre démarche descriptive linéaire puisque notre interrogation des corpus s'est fondée à la fois sur nos modèles de graphes syntaxiques que sur des expressions régulières insistant sur le lexique, outre que sur nos SLG.

Cette tendance à éluder la permutation est d'ailleurs confirmée par le suremploi de l'analogie formelle et sémantique. C'est la régularité sémantique que nous avons évoquée dans l'Introduction qui prime et qui impose le respect de l'ordre des constituants de la séquence parémique. Et, comme on le disait au § 1.1.4., c'est en même temps l'usage qui donne une forme à l'instabilité sémantique éventuelle par l'intérmediaire de SLG de matrice parémique. Par exemple, le cadre sémantique de la parémie *Il n'y a pas de fumée sans feu* que *Littré* paraphrase comme « *Il n'y a pas d'effet sans cause* » (s.v. *fumée*)[257] et que nous avons classée sous la SLG-ac {*Il ne y avoir pas de N sans N*}, réapparaît dans ces occurrences des corpus Leipzig :

<1.> *Il n'y a pas de recherches sans expérimentation animale.* (Leipzig 2009)

<2.> *« Il ne peut y avoir de démocratie sans une bonne gestion des affaires publiques et sans respect de l'Etat de droit », a-t-il conclu.* (Leipzig 2010)

<3.> *« Il n'y a pas de monnaie unique sans politique socio-économique unique », a ajouté Yves Leterme, appelant à l'instauration « au plus vite »*

[257] Il y a aussi la parémie *Il n'y a pas de feu sans fumée* que *Littré* commente comme suit : « *Il n'y a pas de cause sans effet* » (s.v. *fumée*). Si dans le cas de <2> et <3>, le cadre sollicité est celui de la parémie *Il n'y a pas de fumée sans feu*, on peut interpréter l'occurrence en <1> en faisant recours aux deux parémies.

d'une telle gouvernance économique européenne. (Leipzig 2010)

Du point de vue formel, on constate que les concordances <2> et <3> nous permettent de reformuler notre SLG-ac en ce qui concerne *N*. En <2>, *gestion* est précédé d'un déterminant et d'un adjectif, et suivi d'une spécification : *de N Adj*. De plus, la séquence *sans N* est réitérée comme adjonction et on ajoute également le modal *pouvoir*. En <3>, *N* est un nom composé dans les deux positions. L'usage nous encourage donc à généraliser notre SLG-ac comme suit : {{*Il ne y avoir pas de N* (*sans GN*)*}}, où *GN* représente le groupe nominal et * la réitération possible de la séquence entre parenthèses[258]. On revient donc à la notion de *moule* (§ 1.1.4.), notamment à celle proposée par Conenna[259]. Cependant, notre moule résulte d'une approche dialectique entre la description lexico-grammaticale canonique de surface, d'une part, qui est vérifiée et généralisée par ce que l'usage écrit restitue, d'autre part (Marcon, sous évaluation).

Dans le calcul de *f* d'usage créatif avec le processus analogique, nous avons aussi inclus toutes les occurrences où l'analogie sémantique fonctionne par antonymie, comme le montre le renversement sémantique de la parémie *L'habit ne fait pas le moine* dans <4> et <5> :

<4.> *Hijab and the city: quand l'« habit fait la musulmane » (mis à jour)* (Leipzig 2009)

<5.> *À la longue, l'habit finit par faire le moine*. (Leipzig 2009)

L'analogie formelle peut intéresser, par contre, seulement quelques éléments de la parémie, comme dans <6> :

<6.> *L'habit ne fait pas Lemoine : enfin du talk de qualité*. (Leipzig 2009)

où l'assonance phonétique et l'homographie du déterminant et du nom avec le nom propre d'un animateur de télévision (*le moine* / *Lemoine*) permettent de faire référence à la source parémique.

[258] Nous n'avons pas codifié le modal parce qu'il s'agit d'un ajout dont l'insertion est à prévoir (§ Conclusions), mais ne participe pas à la Gestalt sémantique de la séquence.

[259] Comme on le voit, la schématisation formelle intègre la Gestalt linguistique des parémies et confirme sa participation à la transposabilité des formes sémantiques (§ 1.1.4., Gestalt *appliquée à la sémantique des proverbes : Visetti & Cadiot*).

Plus en détail, l'analogie formelle favorise la création d'expressions de matrice proverbiale : on assiste à une *déparémisation* (qui va au-delà de la *déproverbialisation* selon Kleiber[260]) en vue de la genèse lexico-syntaxique d'autres expressions dans l'usage, sans en altérer le sens formulaire originaire[261]. C'est le cas des parémies qui commencent par le déontique *falloir* et, en l'occurrence, de la parémie *Il ne faut pas vendre la peau de l'ours avant de l'avoir tué*. L'omission (ou l'érosion) de la formule impersonnelle initiale en <7> et <8> autonomise ce qu'on appellerait une expression verbale figée, tout en gardant la même prosodie sémantique (§ 2.2.6.) de prudence et de mesure grâce à la conservation de la négation (<7>) et à l'utilisation de verbes insistant sur l'idée de précaution (*garder*, <8>) :

<7.> *Ne vendons pas la peau de l'ours avant de l'avoir tué*. (Leipzig 2010)

<8.> *Soulignant que la politique réserve bien souvent des surprises, le libéral-radical bernois se gardait bien de vendre la peau de l'ours avant de l'avoir tué*. (Leipzig 2009)

La primauté de l'ordre des constituants sur l'axe syntagmatique pour favoriser l'exploitation de la régularité sémantique est aussi corroborée par la préférence accordée aux substitutions paradigmatiques. On rappelle qu'elle intéresse une partie du discours ou unité lexicale appartenant à une classe d'objets (G. Gross & Clas 1997, Le Pesant & Mathieu-Colas 1998) ou d'une colligation ou préférence sémantique (§ 2.2.6. Ce qui veut dire que nous avons considéré le sens littéral de l'unité lexicale remplacée, non pas son sens dans le cadre de la Gestalt linguistique parémique. On le voit par exemple en <9> :

<9.> *En face, malgré la ressemblance frappante d'apparence (même sponsor, même raquette, même tenue blanche et bleue), l'habit ne fait pas encore la nonne pour Wozniacki*. (Leipzig 2009)

où, outre l'adjonction de l'adverbe *encore* par rapport à notre forme standard de départ, l'auteur remplace le *moine* masculin par son correspondant féminin *nonne*, les deux appartenant à la classe d'objets <*humain religieux*>. Il en va de même pour <10> où les

[260] « La déproverbialisation est l'opération qui fait perdre au proverbe son côté dénominatif, pour ne lui laisser que son aspect de phrase » (Kleiber 1999 : 66). La déparémisation dépasse la déproverbialisation dans la mesure où même l'aspect (pseudo) phrastique des parémies disparaît et laisse la place à d'autres agencements syntaxiques.

[261] À cet égard, il serait très intéressant d'étendre à la parémiologie les observations originales de Bolly (2010) sur le rapprochement épistémologique entre phraséologie et grammaticalisation.

adjectifs *froid* et *chaud* relèvent d'une préférence sémantique pour les adjectifs désignant la <*température*> :

<10.> *Pour le légendaire Hongrois Ferenc Puskas, la vengeance est un plat qui se mange chaud*. (Leipzig 2009)

En revanche, nous avons inclus l'occurrence en <11> et en <12> sous le processus de variation par analogie parce que les unités lexicales remplaçantes (*dealers*, *réalités*) n'ont aucun lien sémantique avec le sens littéral de l'unité lexicale remplacée (*chats*, *cœur*) :

<11.> *Il faut reconnaître qu'il a du courage de venir ainsi puisque la nuit tous les dealers sont gris*. (Leipzig 2010)

<12.> *Loin des yeux, loin des réalités au 21e siècle est aussi une affirmation tout simplement abérrante* [sic]. (Leipzig 2010)

Appartient encore au processus de substitution paradigmatique l'occurrence en <13> qui porte sur le remplacement au sein d'une même classe d'objets et, en l'occurrence, de parties du discours : <*préposition*>.

<13.> *La 4e journée des Championnats du monde d'athlétisme de Berlin a été à l'enseigne du proverbe « tout vient à point pour qui sait attendre », du moins pour 4 des 5 vainqueurs, en particulier l'Américaine Sanya Richards sur 400 m.* (Leipzig 2009)

Soulignons que la littérature aurait normalement classé l'occurrence parémique <13> comme variante morphosyntaxique, au même titre que <14> et <15> :

<14.> *La nuit portant conseil, Fabrice Santoro pourrait surprendre un adversaire souvent sujet à la « gamberge » dans les moments difficiles comme l'a montré son incapacité à conclure face à un adversaire visiblement submergé par l'émotion.* (Leipzig 2009)

<15.> *Si rien ne sert de courir et qu'il faut partir à point, un léger retard n'est pas non plus forcément rédhibitoire.* (Leipzig 2009)

L'emploi cadratif à la manière d'un adverbe (Charolles & Vigier 2005) en <14> et la réadaptation comme hypothèse en <15> conservent le sens formulaire des parémies *La nuit porte conseil* et *Rien ne sert de courir, il faut partir à point* en discours. La variation morphosyntaxique se rapproche de la déproverbialisation kleiberienne cette fois-ci (surtout en <15>) et, comme « l'aspect formel fait que le caractère dénominatif n'est pas gommé dans l'histoire » (Kleiber 1999 : 67), nous avons opté pour l'insertion de ces deux occurrences dans la *f* d'usage standard.

Les processus d'adjonction et de réduction co-participent respectivement au maintien de la régularité sémantique par l'ajout ou par la suppression d'éléments de la séquence parémique. Le Tableau 38 montre plutôt qu'ils sont l'apanage du *LiPaF* que des corpus Leipzig. Rares sont les cas, comme en <16> et en <17> :

<16.> *Audio Le Duel Libé-Le Point DSK, loin des yeux, haut dans les sondages?* (Leipzig 2010)

<17.> *Contrairement à ce qu'affirme la maxime, loin des yeux ne va pas de pair avec loin du cœur en Formule Un*. (Leipzig 2010)

et encore plus occasionnelles les occurrences qui montrent la réitération d'un même processus de variation (dans ce cas de figure, de l'adjonction) pour créer des enchaînements de séquences parémiques :

<18.> *Evilspell : mieux vaut tard que jamais tuer la peau de l'ours avant les boeufs*. (Leipzig 2009)

7.3. En guise de conclusion

Dans le présent chapitre, nous avons illustré un exemple d'application de notre cadre fédérateur, des SLG et de nos modèles de grammaires locales pour l'étude de la fréquence d'une liste de 32 (+ 3 variantes) parémies familières dans les corpus de presse francophone Leipzig (2009-2010) et notre corpus spécialisé sur le discours parémiologique *LiPaF*.

716 occurrences parémiques, qui se repartissent en 546 occurrences pour la fréquence d'usage standard et 170 pour la fréquence d'usage créatif, témoignent d'une *perception*

parémiographique (Marcon, sous évaluation) à l'égard des parémies, surtout dans le discours parémiologique.

Les attentes prévues par notre cadre ont été respectées, confirmant la nécessité d'épouser et d'approfondir une *optique parémiométrique* lorsqu'on aborde l'évaluation de la vitalité des parémies.

En ce qui concerne le discours de la presse francophone, l'observation en microdiachronie nous a permis de reconnaître un *lexique parémique noyau* dont l'occurrence est plus probable que celle d'autres parémies. Au contraire, d'autres parémies ont révélé une réticence évidente à apparaître en discours (*parémies* '*usage-resistant*'), et ce, malgré leur familiarité. Ce silence (et plus particulièrement, le silence de la classe lexico-grammaticale la plus familière des parémies commençant par *Qui*) prouve qu'il faut mitiger tous les jugements de contemporanéité axés sur la seule familiarité ou sur la seule fréquence. La contemporanéité des parémies (et un minimum parémiologique français éventuel) doit donc s'implanter dans une convergence et une corrélation des résultats issus autant d'enquêtes sur la familiarité que sur la fréquence.

Autant dans les corpus Leipzig que dans le *LiPaF*, nous avons repéré des *parémies-outliers*, c'est-à-dire des parémies à la mode qui se comportent comme des *stéréotypes parémiques* privilégiés de l'un ou de l'autre discours. Cependant, de par la comparaison des *f* de ces parémies-outliers, nous avons remarqué que, dans les discours des parémiologues, elles sont plutôt des *parémies-fétiches* dont l'usage s'écarte visiblement du silence respectif dans le discours de presses. Autrement dit, elles sont des stérétoypes parémiques qui se chronicisent et se fossilisent dans l'argumentation parémiologique, tout en étant délégitimées par le manque d'usage dans le discours de presse.

Quant aux processus de variation, le sous-emploi de la permutation a montré que l'ordre des mots constitue la véritable stabilité lexico-grammaticale des séquences parémiques. En guise de support à cette stabilité par l'ordre des mots, nous avons trouvé une confirmation dans le suremploi de l'analogie formelle et sémantique, non seulement en termes d'actualisations de cadres sémantiques, mais aussi de modifications faisant appel à la facette phonétique des constituants parémiques. Nous avons également constaté une *déparémisation* de certaines séquences en faveur de la genèse d'expressions verbales figées de matrice proverbiale, comme dans le cas de la parémie *Il ne faut pas vendre la peau de l'ours avant de l'avoir tué*. La centralité de l'ordre des constituants sur l'axe syntagmatique a été renforcée par la préférence accordée aux substitutions paradigmatiques. Pour conclure, nous avons donné une interprétation aux variantes morphosyntaxiques que nous avons

comptées soit comme des occurrences créatives soit comme des occurrences standard. Ce qui revient à remettre en question la pertinence de la notion de *variante morphosyntaxique*.

CONCLUSIONS ET PERSPECTIVES

Essayons maintenant de faire brièvement le point sur les résultats de notre étude et de comprendre quelles voies ils ouvrent pour de nouvelles pistes de recherche.

Conclusions

Dans notre recherche interdisciplinaire et translinguistique, nous avons remarqué que dans la littérature parémiologique les superpositions métaterminologiques et la variété des unités descriptives exploitées (notamment phrase, proposition, énoncé et structure) pour appréhender les parémies, n'ont éclairé que quelques aspects de leur nature linguistique. Le manque de consensus sur ces unités descriptives ainsi que leurs différentes opérationnalisations ont contribué en partie à mettre en relief la combinatoire formelle des parémies. À l'exception de perplexités que certains parémiologues ont manifestées, la quasi-totalité a essayé de saisir les parémies par des unités descriptives minimales instables du point de vue notionnel. Par conséquent, la confusion autour des parémies n'a pu que s'amplifier, comme l'a aussi montré notre métaclassification qui a cherché à résumer et à organiser les diverses tentatives de ranger les répertoires parémiologiques.

Nous sommes reparti de la forme parémique. Plus précisément, nous avons abordé les parémies comme des ***Gestalts linguistiques*** au sens que Lakoff avait envisagé dans les années 1970. Parmi les plusieurs points de vue qu'on aurait pu envisager, nous avons essayé de faire ressortir la ***Gestalt lexico-grammaticale*** de notre corpus parémique. Par une approche à la fois distributionnelle et contextualiste, nous sommes reparti d'une analyse traditionnelle en constituants, notamment en *partie du discours* et *unités lexicales*. Ce retour aux origines de la description syntaxique et lexicale a été motivé par les développements récents en linguistique basée sur l'usage, en particulier sur corpus (Grammaires des Patrons,

Grammaire des Constructions, Analyse des Collostructions, Analyse des Patrons sur Corpus, Théorie des Normes et des Exploitations, etc.).

Notre but n'étant celui de « réinventer la roue » (§ Introduction) ni en parémiologie ni en syntaxe ni en lexicologie, nous avons exploré la combinatoire lexico-grammaticale parémique de manière atomique pour concevoir, par la suite, leur assemblage à l'aide de l'*idiom principle* sinclairien. Nous avons donc essayé d'aborder les parémies par le principe phraséologique par excellence. Ce qui a rencontré jusqu'à présent quelques résistances, vu la distinction (parfois trop forcée) entre phraséologie et parémiologie. Les parémies, comme toute unité phraséologique, sont des parties du discours et des unités lexicales que l'on perçoit et emploie comme un tout unitaire. Elles reflètent l'organisation ensembliste du monde réel ainsi que nos processus de catégorisation. En ce sens, nous avons gardé la vulgate *catégorielle* parémiologique qui est promue par les travaux de Kleiber, quoique par une approche tout à fait différente de la sienne.

Nous avons introduit la notion de ***séquence lexico-grammaticale*** **(SLG)** comme *unité de cosélection étendue des parties du discours et des unités lexicales*. Dépassant le *patron*, elle hérite de la rencontre entre approche distributionnelle et approche contextualiste, ainsi que de certains principes généraux de la linguistique basée sur l'usage. Plus précisément, elle est ouvertement débitrice des travaux de Maurice Gross et du courant sinclairien tout comme du *chunking* et de la *linéarité* cognitivistes. La SLG nous a servi pour décrire notre corpus parémique, sans aucune prétention de définition. Les différentes typologies de SLG nous ont permis de mettre en évidence les agencements lexico-grammaticaux conventionnalisés que nous a montrés un échantillon d'environ 1.800 parémies du répertoire parémiologique. Les SLG nous ont aidé à gérer l'interface lexique-syntaxe et la rencontre distribution-contexte et donc pour équilibrer les tendances parfois syntactico-centriques, parfois lexico-centriques, qu'on a observées respectivement dans la littérature sur le figement et sur la phraséologie.

Avant de saisir les parémies par leurs formes lexico-syntaxiques, il a fallu que nous nous situions autant par rapport à la parémiologie que par rapport à la phraséologie. Comme on l'a vu, c'est la deuxième qui a engendré une mutation de la première. C'est dans la linguistique de corpus (qui a contribué à relancer les études phraséologiques) que nous avons repéré l'outillage méthodologique (sinclairien) pour féconder la parémiologie. Le support de *corpus* conçus (ou recherchés) d'après des critères explicites ainsi que l'observation des observables parémiques en contexte ont satisfait en plein notre souci de description lexico-grammaticale sur la base de la cooccurrence et de la récurrence. De même, l'appui du corpus et le réglage de ses critères de constitution (ou de sélection) se sont révélés comme essentiels

pour le repérage des parémies dans le discours écrit et pour l'estimation de leur fréquence. En ce qui concerne la parémiologie, c'est grâce à la référence aux travaux de Grzybek & Chlosta que nous avons inséré notre étude au sein de la *parémiologie empirique*. Plus précisément, leur systématisation typologique nous a permis de faire de l'ordre dans la littérature parémiologique, notamment en ce qui concerne les notions de *familiarité* et de *fréquence*. Nous avons gardé cette distinction pour mener notre étude de fréquence d'occurrence des parémies sur corpus. Compte tenu de notre approche empirique, nous avons accordé une attention particulière au « strictement linguistique » des parémies françaises. Pour cette raison, notre étude se veut un exemple de ***parémiologie linguistique basée sur l'usage*** dans le sillage autant de la *parémiologie empirique* que de la *parémiologie linguistique* de Conenna.

Notre approche empirique et cumulative (§ Introduction) a également essayé de thésauriser les expériences d'autres parémiologues qui se sont confrontés au repérage des parémies dans des sources textuelles, soient-elles sur support papier ou électronique. Nous avons ainsi dressé un bilan critique de ces expériences qui a débouché sur la proposition d'un **cadre méthodologique fédérateur** axé sur trois lignes directrices : l'élaboration de la liste des parémies à enquêter ; la sélection des sources textuelles à interroger ; les mesures à adopter pour le calcul et pour l'interprétation de la fréquence. Ce cadre et ses suggestions métaterminologiques conséquentes sont à concevoir en continuité idéale avec la typologie d'études empiriques élaborée par Grzybek & Chlosta en parémiologie empirique. D'une part, c'est une tentative de formaliser une méthodologie commune (quoique non exhaustive) pour assurer la compréhension entre les diverses approches parémiologiques et pour améliorer l'interopérabilité de leurs résultats de recherche. D'autre part, il a représenté l'arrière-plan de toute notre étude. En tout cas, ce cadre représente un des parcours épistémologiques et herméneutiques possibles et qui suggère aux chercheurs de laisser perdre l'argument du décès à tout prix des parémies ainsi que l'argument de la contemporanéité des parémies, surtout si elle se fonde sur l'usage, comme il est bien licite de s'y attendre. Les parémies demandent de se situer dans une **optique parémiométrique** qui est à elles seulement (et aux séquences formulaires, en général). C'est cette optique qu'il faut adopter et surtout creuser.

À ce cadre, nous avons ajouté une réflexion à part entière sur les pratiques de recherche des parémies. Comme on l'a vu, les questionnements des parémiologues face au défi de la reconnaissance et du découpage de la parémie dans le discours écrit sont restés quasiment inchangés. Ce qui nous a étonné, c'est l'absence d'une réflexion systématique sur ces aspects dans la littérature, vu l'influence qu'ils exercent sur les résultats quantitatifs, mais aussi qualitatifs des recherches parémiologiques. D'où l'élaboration d'un **volet**

complémentaire informatique à notre cadre méthodologique. Il résume et réorganise l'éventail des techniques de recherche que les parémiologues ont adoptées. Ces derniers ont décidément penché pour l'interrogation des textes électroniques à l'aide des expressions régulières. En revanche, les automates à états finis, et en particulier les grammaires locales, ont suscité l'intérêt de ceux parémiologues qui ont privilégié la forme parémique, et ce, dans le cadre du Lexique-Grammaire. Quoiqu'autant les expressions régulières que les automates à états finis soient deux manières souples pour identifier les parémies dans des corpus, nous avons privilégié les derniers. Ce pour montrer la nécessité d'une réflexion non seulement informatique, mais surtout linguistique sur les parémies à repérer.

Le choix des grammaires locales nous a en effet conduit à tester la validité de la SLG et de notre **modèle de classification lexico-grammaticale**. Pour chaque parémie, suivant le principe de la linéarité, nous avons suivi l'agencement des parties du discours et des unités lexicales par la prise en compte de leur **cooccurrence** et de leur **fréquence** dans le corpus parémique (§ Annexe 1). Certes, ce dernier constitue un biais. La **classification en tables lexico-grammaticales** (§ Annexes 5-6) dépend en effet de son étendue. Pourtant, elle représenté une alternative à la littérature partielle en la matière. La double articulation lexique-syntaxe nous a en effet rapproché de la surface des parémies. Les niveaux de classification et leur numérotation sont fonction de l'analyse linéaire des constituants et de la mise en forme de la SLG. Chaque SLG a acquis le statut de *classe lexico-grammaticale* du niveau *n* si et seulement si la cosélection de ces parties de discours et d'unités lexicales ont pu décrire au moins 2 observables de notre corpus parémique. La contrainte quantitative nous a permis de repérer et de mesurer les *séquences lexico-grammaticales actualisées* (SLG-a) de matrice parémique, à savoir ces cadres ou schémas lexico-grammaticaux que quelques parémiologues ont esquissés.

Pour faciliter notre expérience de classification, nous avons créé un **corpus parémique annoté et lemmatisé** (§ Annexe 4) qui est maintenant à la disposition des parémiologues et des linguistes. Cette facilitation apparente a dû faire face à la performance douteuse, du moins en ce qui concerne l'annotation morphosyntaxique, de la part de *TreeTagger*. Outre à redresser ses fautes, nos corrections manuelles ont laissé des traces de notre subjectivité dans la description lexico-grammaticale, quoique nous les ayons contrôlées par quelques heuristiques adaptées.

Les SLG que nous avons reconnues (surtout les SLG génériques et les SLG actualisées complètes), ont mis en évidence les préférences d'agencement syntaxique et d'actualisation lexicale des parémies, permettant la constitution d'une **grammaire**

parémique. Les SLG font également ressortir des ***rythmes syntaxiques*** et des ***rythmes lexicaux*** qui sont propres aux parémies. Ces rythmes peuvent agir isolément ou, le plus souvent, ensemble. Il serait plus correct, en tout cas, de parler de *rythmes lexico-syntaxiques*, parce que chaque rythme inclut au moins une unité lexicale et une partie du discours. Sont ces rythmes qui participent, d'un côté, à la ***parémisation*** d'une séquence formulaire. D'un autre côté, ils indiquent le processus de ***lexico-grammaticalisation*** qui intéresse certaines combinatoires.

Par la suite, nous avons illustré notre démarche de modélisation des graphes. Faire pivoter celle-ci sur la classification lexico-grammaticale a assuré, d'une part, la reconnaissance des parémies de notre corpus parémique et, d'autre part, la reconnaissance de variantes éventuelles basées sur les SLG. Le guide de la classification a abouti à un **modèle de graphe** qui joue avec l'interface lexique-syntaxe des SLG parémiques ainsi qu'avec l'ambiguïté de l'annotation et de la lemmatisation par les DELA d'*Unitex*. Il a créé délibérément du bruit pour mieux gérer l'imprévisibilité des variations parémiques. Ce modèle de graphe est en effet le précipité d'un compromis entre finesse descriptive pour la recherche d'une forme exacte et généricité lexico-syntaxique en vue de la détection de détournements et d'exploitations parémiques. Quoique nous puissions l'atténuer, nous avons préféré garder ce bruit et le gérer par un tri raisonné des concordances. Celles-ci sont dérivées de l'interrogation des deux sous-corpus de presse en ligne (*fra_news_2009_1M-text* et *fra_news_2010_1M-text*) qui appartiennent à la série française du *Leipzig Corpora Collection*, ainsi que d'un corpus spécialisé de littérature parémiologique francophone, le *LiPaF*. Les premiers nous ont nous ont fourni les matériaux optimaux (notamment par l'échantillonnage du corpus en phrases graphiques ainsi que par la taille (environ 100 millions d'occurrences de mots graphiques)) pour tester les tendances quantitatives que nous avons synthétisées dans notre cadre méthodologique fédérateur. En revanche, le *LiPaF* nous a donné des indications sur la préférence d'usage des parémies au sein de la communauté parémiologique.

L'attention à la **fréquence des parémies dans des corpus** nous a donné l'occasion d'enrichir la littérature parémiologique francophone d'une première liste de fréquence des parémies basée sur corpus. Ces données sur la fréquence fourniront des données quantitatives qui pourront concourir à une meilleure compréhension de la relation entre familiarité et fréquence, notamment dans le cadre de l'élaboration d'un minimum parémiologique français. L'observation en microdiachronie nous a permis de commencer à envisager un ***lexique parémique noyau*** dont l'occurrence serait plus probable que celle d'autres parémies. Comme

étalées sur un continuum, quelques parémies ont montré une réticence évidente à apparaître en discours (***parémies 'usage-resistant'***), alors que d'autres agissent comme des ***stéréotypes parémiques*** privilégiés de l'un ou de l'autre discours, et parfois comme des ***parémies-fétiches*** d'une communauté de locuteurs/auteurs précise (en l'occurrence, de la communauté des parémiologues). En ce qui concerne les processus de variation, le sous-emploi de la permutation a montré que l'ordre des constituants parémique représente le seul élément de stabilité (le fameux figement !) lexico-grammaticale. Ce qui a trouvé une confirmation dans le suremploi de l'analogie formelle et sémantique. Comme les rythmes lexico-syntaxiques révèlent les parcours lexico-grammaticaux de parémisation d'une séquence formulaire, l'usage nous a dévoilé la ***déparémisation*** de certaines séquences en faveur de la genèse d'autres expressions (apparemment) figées de matrice proverbiale. On peut donc conclure qu'il y a des ***cycles de parémisation*** (Marcon, sous évaluation) que les régularités lexico-grammaticales et les corpus permettent de suivre de près et de cerner.

Perspectives

Notre recherche a certainement soulevé d'autres questions qui méritent des réflexions à part entière.

Comme on l'a souligné, ce sont la SLG et la classification, à savoir l'interface lexique-syntaxe, la cooccurrence et la récurrence au sein du répertoire parémiologique, qui ont joué un rôle central dans notre recherche. Néanmoins, il reste à relever un défi d'ordre psycholinguistique qui dépasse les finalités de notre étude. Nous alludons à la relation entre la fréquence que nous avons enregistrée, la familiarité qu'on pourrait observer dans un échantillon de la population francophone et les SLG, notamment les SLG-ac. Nous nous demandons comment les SLG, surtout celles qui se sont révélées comme les plus centrales de notre corpus parémique, interagissent avec la fréquence et la familiarité. En gros, nous nous interrogeons sur le stockage éventuel dans le lexique mental des SLG comme combinatoires privilégiées de reconnaissance (et de production) par analogie/imitation de séquences parémiques et, en général, formulaires. Cela équivaut à se demander si la récurrence des SLG que nous avons décelées de notre corpus parémique pourrait refléter en quelque mesure les modèles exemplaires (Bybee 2010 : 18-19) que les locuteurs mémorisent grâce à l'effet de conservation par fréquence d'occurrence ou *entrenchment* (Bybee 2010 : 24-25). On pourrait comprendre si les SLG organisent le répertoire parémiologique mental et dans quelle mesure

cette organisation est affectée par la fréquence d'occurrence des parémies et par la familiarité déclarée. En ce sens, nous nous interrogeons aussi sur la validité cognitive (autant en termes de mémorisation que de production) de la représentation par notre modèle de graphe/réseau orienté, compte tenu de l'absence de récursivité qui caractérise (pour l'instant) ses états.

Revenons à l'essentiel de notre recherche. À partir de la description lexico-grammaticale atomique, on pourrait appliquer les études sur la composition et sur le figement pour raffiner nos SLG. En outre, l'absence de 199 parémies dans notre classification, dont la plupart commencent par des noms et des verbes, invite à moduler la contrainte de *f* pour les parémies qui commencent par ces parties du discours, notamment pour les entrées lexicales.

Outre la fréquence des parémies à l'écrit, il serait convenable d'envisager une enquête similaire pour l'oral (voire pour le multimodal). Malgré la disponibilité de corpus oraux pour le français, nous avons préféré l'écrit parce qu'il présente moins de problèmes pour le traitement. L'interrogation de corpus oraux implique une connaissance approfondie des protocoles de transcription (qui diffèrent d'un centre de recherche à un autre, malgré les quelques tentatives d'harmonisation) et la disposition d'un jeu d'étiquettes propre à la syntaxe de l'oral ainsi que d'annotateurs syntaxiques qui ne sont qu'à leurs débuts. En d'autres termes, ils auraient excessivement compliqué la phase de prétraitement et la modélisation des requêtes informatiques. En outre, nous n'avions pas de repères pour discuter des fréquences. Comment évaluer la fréquence d'une parémie à l'oral ? Quelle taille pour les corpus ? Comment gérer et calculer les reprises, les autocorrections, etc. ? La piste de l'oral reste en tout cas un terrain d'enquête très stimulant (notamment en raison de l'oralité originaire des parémies) qui marquera sans aucun doute la parémiologie linguistique.

Notre focalisation sur des corpus de presse s'est insérée dans la 'tradition' des parémiologues qui nous ont précédé. Il reste à répéter notre expérience sur d'autres corpus. Notre classification lexico-grammaticale et les SLG sont maintenant à la disposition de la communauté pour la modélisation d'autres graphes et pour la création d'expressions régulières.

L'expérience défectueuse et chronophage de l'annotation par *TreeTagger* et de la correction manuelle des fautes conséquentes exprime le besoin d'implémenter un annotateur *ad hoc* pour les parémies, soit-il conçu ou intégré à d'autres annotateurs existants. Il est évident que la spécificité des agencements syntaxiques a représenté un grand obstacle pour l'automatisation de l'annotation et un ralentissement important dans notre recherche. La description lexico-grammaticale pourrait constituer un point de départ pour cette implémentation.

Il serait également intéressant de poursuivre les expériences de création de DELA des parémies entamées par Conenna et suivies par les autres parémiologues lexico-grammairiens. Par exemple, nous suggérons de codifier nos SLG comme codes, et ce, pour améliorer le formalisme essentiel que Conenna a proposé.

La modélisation des graphes pour la reconnaissance des parémies a écarté toute variable (et, par conséquent, toute sortie). Il est certainement attrayant de revenir sur notre modèle de graphe pour insérer des variables qui permettent, par exemple, le balisage et l'annotation d'un corpus en termes de parémies. Autant fascinante est la création d'un *graphe dictionnaire* (Paumier 2013 : 69-71) pour l'enrichissement et pour la mise à jour d'un DELA avec des parémies manquantes, des détournements parémiques ou des séquences formulaires qui correspondent aux SLG de notre classification. D'ailleurs, d'autres modèles de graphes seraient à élaborer à partir de nos tables descriptives, notamment des SLG-g et des SLG-ap. Ce qui permettrait de mieux saisir leur potentiel créatif formulaire. Sujet plus compliqué à cerner, on pourrait créer des sous-graphes pour améliorer la gestion des adjonctions, notamment des modaux (§ 7.2.) ou des parcours faisant recours aux études sur le figement nominal (G. Gross 1996). Par exemple, l'insertion des états :

<*Prép*$_{det}$> <*N*>

dans le graphe lemmatisé de :

Qui va à la chasse perd sa place

pourrait reconnaître dans un corpus des détournements comme celui que nous avons repéré dans un forum sur le Web consacré aux insectes :

> C'est clair... on s'absente quelques heures et d'autres ont fait le boulot pour vous 😃 Comme dit le proverbe : qui va à la chasse aux papillons perd sa place (http://www.insecte.org/forum/, date de rédaction : 08/06/2007; date de consultation : 11/10/2013, c'est nous qui soulignons).

Comme on le voit, le parcours pour la reconnaissance des parémies dans les corpus présente encore un bon nombre de défis et de problématiques à résoudre comme, entre autres,

l’appréhension détaillée et l’opérationnalisation informatique de tous les processus de variation. Nous espérons du moins avoir donné une contribution à ce chemin.

RÉFÉRENCES

Achard-Bayle, G., & Schneider, B. (2010). Les énoncés parémiques, hypo- et paratactiques: des constructions syntaxiques aux interprétations sémantiques. In M.-J. Béguelin, M. Avanzi & G. Corminboeuf (dir.), *La Parataxe. Structures, marquages et exploitations discursive. Tome 2* (p. 95-120). Berne: Peter Lang.

Alonso Pérez-Ávila, E. (2008). Paremia guardada, dos veces ganada: criterios de ordenación de paremias en los repertorios paremiográficos españoles e italianos. *Critica del testo XI(1-2)*, 447-467.

Anderson, W. J. (2006). *The Phraseology of Administrative French. A Corpus-Based Study.* Amsterdam: Rodopi.

Anscombre, J.-C. (1994a). Proverbes et formes proverbiales : valeur évidentielle et argumentative. *Langue française 102*, 95-107.

Anscombre, J.-C. (1994b). La sémantique française au XX^{e} siècle : de la théorie de la référence à la théorie des stéréotypes. In J. F. Corcuera, M. Djian, A. Gaspar (dir), *La lingüística francesa*. *Situación* y *perspectivas a finales del siglo XX / La linguistique française. Bilan et perspectives a la fin du XX*e *siecle*, Actes du *Congreso internacional de Lingüística francese* de Zaragoza (p.9-27). Zaragoza: Universidad de Zaragoza.

Anscombre, J.-C. (2000). Parole proverbiale et structures métriques. *Langages 139*, 6-26.

Anscombre, J.-C. (2003). Les proverbes sont-ils des expressions figées ? *Cahiers de Lexicologie 82*, 159-173.

Anscombre, J.-C. (2005). Les proverbes : un figement du deuxième type ? *Linx 53*, 17-33.

Anscombre, J.-C. (2006). Polyphonie et classification des énoncés sentencieux. Les marqueurs médiatifs génériques. *Le Français Moderne 74(1)*, 87-99.

Anscombre, J.-C. (2008a). Les formes sentencieuses : peut-on traduire la sagesse populaire? *META 53(2)*, 253-268.

Anscombre, J.-C. (2008b). Quelques propriétés linguistiques des formes sentencieuses, et leur application à la traduction franco-espagnole. In C. González Royo & P. Mogorrón Huerta (dir.), *Estudios y Análisis de Fraseología Contrastiva: Lexicografía y Traducción* (p. 11-36). Alicante: Universidad de Alicante.

Anscombre, J.-C. (2011a). L'introduction du pronom neutre dans les marqueurs médiatifs à verbe de dire de type « *Comme dit le proverbe* / *Como dice el refrán* ». *Langages 184*, 13-34.

Anscombre, J.-C. (2011b). Figement, idiomaticité et matrices lexicales. In J.-C. Anscombre, & S. Mejri (dir.), *Le figement linguistique: la parole entravée* (p. 17-40). Paris: Champion.

Anscombre, J.-C. (2013). Les formes sentencieuses : classes et sous-classes. In J.-M. Benayoun, N. Kübler & J.-P. Zouogbo (dir.), *Parémiologie. Proverbes et formes voisines. Tome I* (p. 93-112). Sainte-Gemme: Presses Universitaires de Sainte-Gemme.

Arnaud, P. J. (1991). Réflexions sur le proverbe. *Cahiers de Lexicologie 59(2)*, 5-27.

Arnaud, P. J.-L. (1992). La connaissance des proverbes français par les locuteurs natifs et leur sélection didactique. *Cahiers de Lexicologie 60(1)*, 195-238.

Arnaud, P. J.-L., & Moon, R. (1993). Fréquence et emploi des proverbes anglais et français. In C. Plantin (dir.), *Lieux communs, topoï, stéréotypes, clichés* (p. 323-341). Paris: Kime.

Arora, S. L. (1998). « El que nace para tamal... »: A Study in Proverb Patterning. *De Proverbio 4(1)*, 109-152.

Asensio Sánchez, M. Á. (2008). Aforismos y brocardos : la expresión de un principio jurídico meta-histórico. *Critica del Testo XI(1-2)*, 391-402.

Aussenac-Gilles, N., & Condamines, A. (2009). Variation syntaxique et contextuelle dans la mise au point de patrons de relations sémantiques. In J.-L. Minel (dir.), *Filtrage sémantique* (p. 115-149). Paris: Hermes/Lavoisier.

Aussenac-Gilles, N., Condamines, A., & Szulman, S. (2002). Prise en compte de l'application dans la constitution de produits terminologiques. In *Actes des 2e Assises Nationales du GDR I3, Nancy, Décembre 2002* (p. 289-302). Toulouse: Cépaduès Editions.

Baker, M. & Saldanha, G. (dir.) (2009). *Routledge Encyclopedia of Translation Studies*. Abingdon/New York: Routledge.

Baker, P., Hardie, A., & McEnery, T. (dir.) (2006). *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Barani, N. (2007). Cincuenta refranes entre los más frecuentes del español actual con su correspondencia en farsi. *Paremia 16*, 99-105.

Barani, N. (2012). *Aspectos de la utilización de las paremias en el diario El País: hacia el desarrollo de materiales didácticos para la enseñanza del español a hablantes de persa*. Salamanca: Universidad de Salamanca [Thèse de doctorat].

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004* (p. 1113-1116). Lisbon: ELDA.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43(3)*, 209-226.

Barta, P. (2005). Au pays des proverbes, les détournements sont rois. Contribution à l'étude des proverbes détournés du français (I). *Paremia 14*, 139-152.

Barta, P. (2006). Au pays des proverbes, les détournements sont rois. Contribution à l'étude des proverbes détournés du français (II). *Paremia 15*, 57-71.

Béguelin, M.-J. (2000). *De la phrase aux énoncés: grammaire scolaire et descriptions linguistiques.* Bruxelles: De Boeck Duculot.

Béguelin, M.-J. (2002). Clause, période ou autre ? La phrase graphique et la question des niveaux d'analyse. *Verbum XXIV(1-2)*, 85-107.

Berrendonner, A. (2002). Les deux syntaxes. *Verbum XXIV(1-2)*, 23-36.

Bessi, P. (2004). *Le strutture del proverbio monofrastico. Analisi di millecinquecento formule tratte dall'archivio dell'Atlante Paremiologico Italiano.* Torino: Dell'Orso.

Blanche-Benveniste, C. (2002). Phrase et construction verbale. *Verbum XXIV(1-2)*, 7-22.

Blanco, X., Moreno, L., & Wuattier, S. (1995). Lexique-grammaire et proverbe dans le cadre du FLE. In R. Gauchola, C. Mestreit, & M. A. Tost, *Enseignement/Apprentissage du FLE – Répéres et Applications. Recueil d'interventions aux XVIes, XVIIes et XVIIIes Journées Pédagogiques sur l'Enseignement du Français en Espagne* (p. 149-164). Barcelone: ICE/Université Autonome de Barcelone.

Bolly, C. (2010). Flou phraséologique, quasi-grammaticalisation et pseudo marqueurs de discours : un no man's land entre syntaxe et discours ? *Linx 62-62*, 11-38.

Bourigault, D., & Slodzian, M. (1999). Pour une terminologie textuelle. *Terminologies nouvelles*, 29-32.

Bowker, L. & Pearson, J. (2002). *Working with Specialized Language. A practical guide to using corpora*. Londres & New York: Routledge.

Bruxelles, S., Mondada, L. S., & Traverso, V. (dir.) (2009). Grands corpus de français parlé. Bilan historique et perspectives de recherche. *Cahiers de Linguistique 33(2)*, 268pp.

Buridant, C. (1976). Nature et fonction des proverbes dans les Jeux-Partis. *Revue des Sciences Humaines 163*, 377-418.

Buridant, C. (2011). Essai sur la proverbialité médiévale. In M.-S. Ortola (dir.), *Aliento. Corpus anciens et Bases de données* (p. 223-258). Nancy: Presses Universitaires de Nancy.

Bybee, J. (2010). *Language, Usage and Cognition.* Cambridge: Cambridge University Press.

Cabré, M. T. (1998). *La terminologie. Théorie, méthode et applications.* Ottawa/Paris: Presses de l'Université d'Ottawa/Armand Colin.

Cadiot, P., & Visetti, Y.-M. (2001). *Pour une théorie des formes sémantiques. Motifs, profils, thèmes.* Paris: PUF.

Calvo Espiga, A. (2008). En torno al origen y significado del aforismo jurídico. *Critica del testo XI(1-2)*, 413-445.

Čermák, F. (2007 [1998]). Usage of Proverbs in Today's Czech language: What the Czech National Corpus Seems to Suggest. In F. Čermák, *Frazeologie a idiomatika ceska a obecna. Czech and General Phraseology* (p. 569-583). Prague: Karolinum.

Čermák, F. (2007 [2003]). Paremiological Minimum of Czech: The Corpus Evidence. In F. Čermák, *Frazeologie a idiomatika ceska a obecna. Czech and General Phraseology* (p. 597-613). Prague: Karolinum.

Čermák, F. (2007 [2006]). What One Can Do with Proverbs in Text. In F. Čermák, *Frazeologie a idiomatika ceska a obecna. Czech and General Phraseology* (p. 536-548). Prague: Karolinum.

Cerquiglini, B., & Cerquiglini, J. (1976). L'écriture proverbiale. *Revue des Sciences Humaines 163*, 359-375.

Cetro, R. (2013). *Lexique-grammaire et Unitex : quels apports pour une description terminologique bilingue de qualité ? Analyse sur deux corpus comparables de médecine thermale*. Università degli Studi di Brescia & Université Paris-Est Marne-la-Vallée [Thèse de doctorat].

Chacoto, L. (2007). A sintaxe dos proverbios. As estruturas *quem* / *quien* en portugués e español. *Cadernos de Fraseoloxía Galega 9*, 31-53.

Charaudeau, P., & Mainguenau, D. (2002). *Dictionnaire d'analyse du discours.* Paris: Seuil.

Charolles, M. & Vigier, D. (2005). Les adverbiaux en position préverbale : portée cadrative et organisation des discours, *Langue française 148*, 9-30.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics 11(4)*, 411-433.

Chevalier, J.-C. (2006). *Histoire de la syntaxe: naissance de la notion de complément dans la grammaire française (1530-1750).* Paris: Honoré Champion.

Chlosta, C., & Grzybek, P. (1995). Empirical and Folkloristic Paremiology: Two to Quarrel orto Tango? *Proverbium 12*, 67-85.

Cignoni, L., & Coffey, S. (2000). A Corpus Study of Italian Proverbs: implications for lexicographical description. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (dir.)

Proceedings of the Ninth EURALEX International Congress 2000. Volume II (p. 549-555). Stuttgart: Universität Stuttgart.

Colombat, B. (1988). Présentation : Eléments de réflexion pour une histoire des parties du discours. *Langages 92*, 5-10.

Condamines, A. (2005). Sémantique et corpus, quelles rencontres possibles ? In A. Condamines (dir.), *Sémantique et Corpus* (p. 17-38). Londres: Hermès Science.

Conenna, M. (1988). Sur un lexique-grammaire comparé des proverbes. *Langages 90*, 99-116.

Conenna, M. (1995). Équivalence sémantique et variantes formelles dans les proverbes. In M. Margarito, & A. M. Raugei, *Studi di Linguistica Storia della lingua Filologia francesi. Atti del Convegno della Società Universitaria per gli Studi di Lingua e Letteratura francese, Torino, 16-17 giugno 1994* (p. 205-219). Torino: Dell'Orso.

Conenna, M. (1998a). Le proverbe, degré ultime du figement ? In S. Mejri, G. Gross, A. Clas, & T. Baccouche (dir.), *Le figement lexical. Actes de la 1ère Rencontres Linguistiques Méditerranéennes* (p. 361-373). Tunis: Editions du CERES.

Conenna, M. (1998b). Analyses automatiques des textes. Méthodes et perspectives. In *Atti del Convegno internazionale. Studi di linguistica francese in Italia (1960-1996)* (p. 97-117). Brescia: La Scuola.

Conenna, M. (2000a). Structure syntaxique des proverbes français et italiens. *Langages 139*, 27-38.

Conenna, M. (2000b). Dictionnaire électronique de proverbes français et italiens. In A. Englebert, M. Pierrard, L. Rosier, & D. Van Raemdonck (dir.), *Des mots aux dictionnaires. Travaux de la section « Lexicologie, lexicographie, onomastique, toponymie». Actes du XXIIe Congrès International de Linguistique et de Philologie Romanes, Bruxelles, 23-29 juillet 1998. Volume IV* (p. 137-145). Tübingen: Max Niemeyer Verlag.

Conenna, M. (2000c). Classement et traitement automatique des proverbes français et italiens. *BULAG (hors serie)*, 285-294.

Conenna, M. (2002). Sur l'historique du proverbe. In G. Maiello, & R. Stajano (dir.), *Collage. Studi in memoria di Franca Caldari Bevilacqua* (p. 35-55). Salerno/Milano: Oedipus.

Conenna, M. (2004). Principes d'analyse automatique des proverbes. *Linguisticae Investigationes Supplementa 24*, 91-103.

Conenna, M. (2010). *La salle de cours. Questions/réponses sur la grammaire française.* Berne: Peter Lang.

Conenna, M. (2011). Variantes proverbiales : classement et traduction français-italien. In C. González Royo, & P. Mogorrón Huerta (dir.), *Fraseología contrastiva: lexicografía, traducción y análisis de corpus* (p. 75-93). San Vicente del Raspeig: Publicaciones Universidad de Alicante.

Conenna, M., & Kleiber, G. (2002). De la métaphore dans les proverbes. *Langue française 134*, 58-77.

Conenna, M., Coppens d'Eeckenbrugge, M., Flamini, F., Klein, J.-R., & Pierret, J.-M. (2006). Le projet DicAuPro. Développement d'une base de données informatisée des proverbes du français. In A. Häcki Buhofer, & H. Burger (dir.), *Phraseology in Motion 1. Methoden und Kritik* (p. 79-90). Baltmannsweiler: Schneider Verlag.

Conenna, M., Lavermicocca, F., Marcon, M., Serrone, G., & Vergne, M. (2011). *Un caso paremio-giuridico*. Communication présentée à *I Jornadas Fraseología y Paremiología – Perspectivas y aplicaciones en Didáctica, Traducción y Lingüística de corpus*, Bari, 5-6 mai 2011.

Cornu, G. (2005). *Linguistique juridique.* Paris: Montchrestien.

Corpas Pastor, G. (2003 [1998]). El uso de paremias en un corpus del español peninsular actual. In G. Corpas Pastor, *Diez años de investigaciones en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos* (p. 83-107). Madrid/Frankfurt am Main: Iberoamericana/Vervuert.

Courtois, B. (1994-1995). Buts et méthodes de l'élaboration des dictionnaires électroniques du LADL. *Cahiers du CIEL*, 87-108.

Cram, D. (1983). The linguistic status of the proverb. *Cahiers de Lexicologie 43*, 53-71.

Crépeau, P. (1975). La définition du proverbe. *Fabula 16(3-4)*, 285-304.

D'Andrea, G. (2007). Sur le rôle de la répétition dans les proverbes. *Cahiers du Centre d'Études Métriques 5*, 119-129.

D'Andrea, G. (2008). *Le rythme dans les proverbes français.* Lecce: Adriatica Editrice Salentina.

De Gioia, M. (à paraître). « Mieux vaut tard que jamais ». Su alcuni proverbi francesi della collezione di Marco Besso. In L. Lalli (dir.), *La fortuna dei proverbi, le identità dei popoli: Marco Besso e la sua collezione*.

De Gioia, M., & Marcon, M. (à paraître). Discours de médiation(s). Le cas de *conflit*/*conflitto*. In M. De Gioia (dir.), *Pratiques communicatives de la médiation.*

Dundes, A. (1994 [1975]). On the Structure of the Proverb. In W. Mieder, & A. (. Dundes, *The Wisdom of Many: Essays on the Proverb* (p. 43-64). Madison: The University of Wisconsin Press.

Ďurčo, P. (2005). *Sprichwörter in der Gegenwartssprache.* Trnava: Univerzita sv. Cyrila a Metoda v Trnave.

Ďurčo, P. (2006). Methoden der Sprichwortanalysen oder Auf dem Weg zum Sprichwort-Optimum. In A. Häcki Buhofer, & H. Burger (dir.), *Phraseology in Motion 1. Methoden und Kritik* (p. 3-20). Baltmannsweiler: Schneider Verlag.

Eco, U. (2008 [1968]). *La struttura assente. La ricerca semiotica e il metodo strutturale.* Milano: Bompiani.

Fairon, C., & Cougnon, L.-A. (2009). La mise à jour d'un dictionnaire électronique. Une expérience pédagogique liée à la mise à jour du DELAF. *Arena Romanistica 4*, 58-71.

Fairon, C., Klein, J.-R., & Paumier, S. (2006). *Le langage SMS. Etude d'un corpus informatisé à partir de l'enquêtes « Faites don de vos sms à la science ».* Louvain-la-Neuve: Presses Universitaires de Louvain.

Fairon, C., Macé, K., & Naets, H. (2008). GlossaNet 2: a linguistic search engine for RSS-based corpora. *Proceedings of LREC 2008. Workshop WAC4.*

Fillmore, C. J. (1992). "Corpus Linguistics" vs. "Computer-aided Armchair Linguistics". In J. Svartvik (dir.), *Directions in Corpus Linguistics. Proceedings from a 1992 Nobel Symposium on Corpus Linguistics, Stockholm.* (p. 35-60). Berlin: Mouton de Gruyter.

Firth, J. R. (1957 [1961]). Modes of Meaning. In J. R. Firth (dir.), *Papers in Linguistics 1934-1951* (p. 190-215). Londres: Oxford University Press.

Firth, J. R. (1968 [1957]). A Synopsis of Linguistic Theory, 1930-55. In F. Palmer (dir.), *Selected Papers of J. R. Firth 1952-1959* (p. 168-205). Bloomington & Londres: Indiana University Press.

Francis, W. N. (1992). Language Corpora B.C. In J. Svartvik (dir.), *Directions in Corpus Linguistics* (p. 17-32). Berlin: Mouton de Gruyter.

Galisson, R. (1994). Les palimpsestes verbaux : des révélateurs culturels remarquables, mais peu remarqués… *Repères 8*, 41-62.

García Yelo, M. (2009). El refranero hoy en Bélgica. *Paremia 18*, 225-244.

Gardes-Tamine, J. (1990). *La Grammaire. 2/Syntaxe.* Paris: Armand Colin.

Gardes-Tamine, J. (2013). *L'ordre des mots.* Paris: Armand Colin.

Gómez-Jordana Ferary, S. (2003). Taxinomie des proverbes français et espagnols contemporains. *Revue de Sémantique et Pragmatique 13*, 69-97.

Gómez-Jordana Ferary, S. (2012). *Le proverbe: vers une définition linguistique. Etude sémantique des proverbes français et espagnols contemporains.* Paris: L'Harmattan.

Gouvard, J.-M. (1996). Les formes proverbiales. *Langue française 110*, 48-63.

Gouvard, J.-M. (1999). Les adages du droit français. *Langue française 123*, 70-84.

Granger, S. (2002). A Bird's Eye View of Learner Corpus Research. In Granger, S., Hung, J. & Petch-Tyson, S. (dir.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (p.3-33). Amsterdam & Philadelphia: John Benjamins.

Greimas, A. J. (1960). Idiotismes, proverbes et dictons. *Cahiers de Lexicologie 2*, 41-61.

Gries, S. Th. (2009). *Quantitative Corpus Linguistics with R. A Practical Introduction*. New York: Taylor & Francis.

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

Gross, G., & Clas, A. (1997). Synonymie, polysémie et classes d'objets. *META 42(1)*, 147-154.

Gross, M. (1975). *Méthodes en syntaxe. Le régime des constructions complétives.* Paris: Hermann.

Gross, M. (1976). Présentation. In J.-P. Boons, A. Guillet, & C. Leclère, *La structure des phrases simples en français : constructions intransitives* (p. 7-28). Genève: Droz.

Gross, M. (1982). Une classification des « phrases figées » du français. *Revue Québécoise de Linguistique 11(2)*, 151-185.

Gross, M. (1986a). Lexique-grammaire et adverbes : deux exemples. *Revue Québécoise de Linguistique 15(2)*, 299-311.

Gross, M. (1986b). *Grammaire transformationnelle du français. 2-Syntaxe du nom.* Malakoff: Editions Cantilène.

Gross, M. (1990). *Grammaire transformationnelle du français. 3-Syntaxe de l'adverbe.* Paris: ASSTRIL.

Gross, M. (1993). Local Grammars and Their Representations By Finite Automata. In M. Hoey (dir.), *Data, Description, Discourse. Papers on the English Language in honour of John McH. Sinclair on His Sixtieth Birthday* (p. 26-38). New York: HarperCollins.

Gross, M. (1997). The Construction of Local Grammars. In E. Roche, & Y. Schabes (dir.), *Finite-State Language Processing* (p. 329-354). Cambridge/London: The MIT Press.

Gross, M., & Lentin, A. (1970). *Introduction to Formal Grammars.* Londres/Berlin/New York: George Allen & Unwin/Springer/Heidelberg.

Grzybek, P. (2009). The Popularity of Proverbs. A Case Study of the Frequency-Familiarity Relation for German. In R. J. Soares, & O. Lauhakangas (dir.), *Proceedings of the Second Interdisciplinary Colloquium on Proverbs* (p. 214-229). Tavira: IAP.

Grzybek, P., & Chlosta, C. (1993). Grundlagen Der Empirischen Sprichwortforschung. *Proverbium 10*, 89-120.

Grzybek, P., & Chlosta, C. (2009). Some Essentials on the Popularity of (American) Proverbs. In K. J. McKenna (dir.), *The Proverbial 'Pied Piper'. A Festschrift Volume of Essays in Honor of Wolfgang Mieder on the Occasion of his 65th Birthday* (p. 95-110). New York: Peter Lang.

Gutiérrez Sánchez, E. (2008). Las paremias en revistas en lengua francesa. In J. Sevilla Muñoz, C. A. Crida Álvarez, & M. I. Zurdo Ruiz-Ayúcar (dir.), *Estudios paremiológicos. I: La investigación paremiológica en España. II. Los refranes y El Quijote* (pp. 343-374). Athènes: Ta kalós keímena.

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin.

Hanks, P. (2004a). Corpus Pattern Analysis. In G. Williams, & S. Vessier (dir..), *Proceedings of the Eleventh EURALEX International Conference 2004. Volume I* (p. 87-97). Lorient: Université de Bretagne Sud.

Hanks, P. (2004b). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography 17(3)*, 245-274.

Hanks, P. (2008). Lexical patterns: From Hornby to Hunston and beyond. In E. Bernal, & J. de Cesaris (dir.), *Proceedings of the XIII EURALEX International Congress* (p. 89-129). Barcelona: Universitat Pompeu Fabra - Documenta Universitaria.

Harris, Z.S. (1976). *Notes du cours de syntaxe*. Paris: Seuil.

Hoey, M. (2005). *Lexical Priming. A New Theory of Words and Language*. Abingdon/New York: Routledge,

Hrisztova-Gotthardt, H. & Gotthardt, Z. (2011). Коŭто мърси, намura: Searching for Bulgarian proverbs on the Web. *Jezikoslovlje 12(2)*, 249-262.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English.* Amsterdam/Philadelphia: John Benjamins.

Järv, R. (1999). Is Providing Proverbs a Tough Job? References to Proverbs in Newspaper Texts. *Folklore 10*, 77-107.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* Londres/New York: Longman.

Kerbrat-Orecchioni, C. (2003). Les genres de l'oral : types d'interactions et types d'activités. *Contributions à la journée « Les genres de l'oral » organisée par Catherine Kerbrat-Orecchioni et Véronique Traverso, Université Lumière Lyon, Campus Porte des Alpes, 18 avril 2003.*

Kerbrat-Orecchioni, C., & Traverso, V. (2004). Types d'interactions et genres de l'oral. *Langages 153*, 41-51.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the Special Issue on the Web As Corpus. *Computational Linguistics 29(3)*, 333-347.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (dir.), *Proceedings of the Eleventh EURALEX International Congress 2004. Volume I.* (p. 105-116). Lorient: Université de Bretagne Sud.

Kleiber, G. (1989). Sur la définition du proverbe. In G. Gréciano (dir.), *Europhras 88. Phraséologie contrastive. Actes du Colloque International de Klingenthal-Strasbourg, 12-16 Mai 1988* (p. 233-252). Strasbourg: Université des Sciences Humaines.

Kleiber, G. (1999). Les proverbes : des dénominations d'un type « très très spécial ». *Langue française 123*, 52-69.

Kleiber, G. (2000). Sur le sens du proverbe. *Langages 139*, 39-58.

Kleiber, G. (2003). Faut-il dire *adieu* à la phrase ? *L'Information grammaticale 98*, 17-22.

Kleiber, G. (2010a). Proverbes : transparence et opacité. *META 55(1)*, 136-146.

Kleiber, G. (2010b). Proverbes et classification. In T. Nakamura, E. Laporte, A. Dister, & C. (. Fairon, *Les tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère* (p. 155-168). Louvain-la-Neuve: Presses Universitaires de Louvain.

Klein, J.-R. (2006). Problemas relacionados coa determinación da forma «canónica» nunha base de datos de refráns franceses (*DicAuPro*). *Cadernos de Fraseoloxía Galega 8*, 147-163.

Koch, P., & Oesterreicher, W. (2001). Gesprochene Sprache und geschriebene Sprache. Langage parlé et langage écrit. In G. Holtus, M. Metzeltin, & C. Schmitt (dir.), *Lexikon der Romanistichen Linguistik. Band I.2* (p. 584-627). Tübingen: Niemeyer/Mouton de Gruyter.

Kuusi, M. (1998 [1953]). Variations in the popularity of Finnish proverbs. *De Proverbio 4(1)*, 24-40.

Lacavalla, C. (2007). *Lexique-grammaire des proverbes en Quand/Quando. Comparaison français-italien et représentation par grammaires locales*. Università degli Studi di Bari [Thèse de doctorat].

Lagarde, J.-P. (1988). Les parties du discours dans la linguistique moderne et contemporaine. *Langages 92*, 93-108.

Lakoff, G. (1977). Linguistic Gestalts. In W. A. Beach, S. E. Fox, & S. Philosoph (dir.), *Papers From the Thirteenth Regional Meeting of Chicago Linguistic Society* (p. 236-287). Chicago: University of Chicago.

Lamiroy, B. (coord.), Klein, J.-R., Labelle, J., Leclère, C., Meunier, A., & Rossari, C. (2010). *Les expressions verbales figées de la francophonie. Belgique, France, Québec et Suisse.* Paris: Ophrys.

Lau, K. J. (2003 [1996]). "It's about Time!": The Ten Proverbs Most Frequently Used in Newspapers and Their Relation to American Values. In W. Mieder (dir.), *Cognition, Comprehension, and Communication. A Decade of North American Proverb Studies (1990-2000)* (p. 231-254). Baltmannsweiler: Schneider Verlag.

Lavermicocca, F. (2011). Proverbes et altérité, entre deux langues. *Cahiers de recherche de l'école doctorale en linguistique française 5*, 64-74.

Laviosa, S. (2002). *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.

Le Goffic, P. (2005). La phrase « revisitée ». *Le français aujourd'hui 148(1)*, 55-64.

Le Pesant, D., & Mathieu-Colas, M. (1998). Introduction aux classes d'objets. *Langages 131*, 6-33.

Leech, G. (1992). Corpora and Theories of Linguistic Performance. In J. (. Svartvik, *Directions in Corpus Linguistics* (p. 105-122). Berlin: Mouton de Gruyter.

Legallois, D. (2009). Mémento sur quelques rapports entre mémoire et linguistique. *Questions de style 6*, 1-21.

Legallois, D., & François, J. (dir.) (2006). *Autour des grammaires de constructions et de patterns. Cahiers du CRISCO* 21, janvier 2006.

Legallois, D., & François, J. (2011). La Linguistique fondée sur l'usage : parcours critique. *Travaux de Linguistique 62*, 7-33.

Lerat, P. (1995). *Les langues spécialisées.* Paris: PUF.

L'Homme, M.-C. (2004). *La terminologie: principes et techniques.* Montréal: Presses de l'Université de Montréal.

Lindquist, H. (2009). *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Loiseau, S. (2011). Les faits statistiques comme objectivation ou comme interprétation : statistiques et modèles basés sur l'usage. *Travaux de Linguistique 62*, 59-78.

Longrée, D., & Mellet, S. (2013). Le motif: une unité phraséologique englobante? Etendre le champ de la phraséologie de la langue au discours. *Langages 189*, 65-79.

Mainguenau, D. (2012). *Les phrases sans texte.* Paris: Armand Colin.

Maniez, F. (2000). Le repérage par traitement automatique du défigement lexical des proverbes dans la presse américaine. *Revue française de linguistique appliquée V(2)*, 19-32.

Marchello-Nizia, C. (1979). La notion de « phrase » dans la grammaire. *Langue française 41*, 35-48.

Marcon, M. (2011). Productivité des structures proverbiales en *Qui* en français écrit contemporain. *Cahiers de recherche de l'école doctorale en linguistique française 5*, 107-121.

Marcon, M. (2012). Another proverb in the wall. Perception parémique et intéret parémiologique sur les fan pages de Facebook. In M. I. González Rey (dir.), *Unidades fraseológicas y TIC* (pp. 125-146). Madrid/Las Rozas: Centro Virtual Cervantes/Instituto Cervantes.

Marcon, M. (2013). Détection automatique des proverbes français sur corpus. Modélisation d'automates et fréquence d'usage. In Benayoun, J.-M.; Kübler, N.; Zouogbo, J.-P. (dir.) *Parémiologie. Proverbes et formes voisines. Tome II* (pp. 299-315). Sainte Gemme: Presses Universitaires de Sainte Gemme.

Marcon, M. (à paraître). Une classification lexico-grammaticale des proverbes français. In *Actes du Colloque International Jeunes Chercheurs 2012 – Les classifications en linguistique: problèmes, méthodologie, enjeux. Strasbourg, 6-8 juin 2012.*

Marcon, M. (sous évaluation). Les rythmes lexico-syntaxiques des parémies. La place du verbe. *Repères-DORIF 5*.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies. An Advanced Resource Book.* Abingdon: Routledge.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mejri, S. (1997). *Le figement lexical : descriptions linguistiques et structuration sémantique.* Tunis: Publications de la Faculté des lettres de la Manouba.

Mejri, S. (2001). La structuration sémantique des énoncés proverbiaux. *L'information grammaticale 88*, 10-15.

Mejri, S. (2008). Inférence et structuration des énoncés proverbiaux. In D. Leeman (dir.), *Des topoï à la théorie des stéréotypes en passant par la polyphonie et l'argumentation dans la langue. Hommages à Jean-Claude Anscombre* (p. 169-180). Chambéry: Université de Savoie.

Meschonnic, H. (1976). Les proverbes, actes de discours. *Revue des Sciences Humaines 163*, 419-430.

Michel, J.-B., Shen, Y. K., Aiden, A., Veres, A., Gray, M. K., Brockman, W., Lieberman Aiden, E. (2011). Quantitative Analysis of Culture Using Millions of Digitized Book. *Science 331(6014)*, 176-182.

Mieder, W. (1995 [1993]). "The apple doesn't fall far from the tree": a Historical and Contextual Proverb Study Based On Books, Archives and Databases. *De Proverbio 1(1)*, 222-275.

Mieder, W. (1995 [1994]). Paremiological Minimum and Cultural Literacy. *De Proverbio 1(1)*, 12-37.

Mieder, W. (2004). *Proverbs. A Handbook.* Westport/London: Greenwood Press.

Milner, G. B. (1969). De l'armature des locutions proverbiales. Essai de taxonomie sémantique. *L'Homme 9*, 49-70.

Mogorrón Huerta, P., & Navarro Brotons, L. (2012). Analyse contrastive et syntaxique des proverbes espagnols et français en *a*/*à* ; *más vale*/*mieux vaut* ; *no*/*ne* ; *quien*/*qui*. In J.-C. Anscombre, B. Darbord, & A. Oddo, *La Parole exemplaire - Introduction à une étude linguistique des proverbes* (p. 453-469). Paris: Armand Colin [Ebook].

Moon, R. (1998). *Fixed Expressions and Idioms. A Corpus-Based Approach.* Oxford: Clarendon Press.

Navarro Brotons, L. M. (2013). *Las paremias y sus variantes: análisis sintáctico, semántico y traductológico español/francés*. Universidad de Alicante [Thèse de doctorat].

Norrick, N. R. (1985). *How Proverbs Mean: Semantic Studies in English Proverbs.* Berlin: Mouton de Gruyter.

O'Keeffe, A., McCarthy, M. & R. Carter (dir.). (2007). *From Corpus to Classroom: language use and language teaching*. Cambridge: Cambridge University Press.

O'Reilly, T. (2005, 09 30). *What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software.* Consulté le 12/11/2013 de http://oreilly.com/web2/archive/what-is-web-20.html

Paisij, D. C. (1971). Notes sur la structure des proverbes français. In *Travaux de l'Université Saints Cyrille et Méthode de Veliko Tirnovo. Tome VIII. Volume 1.* Faculté des Lettres de l'Université Saints Cyrille et Méthode de Veliko Tirnovo.

Paulhan, J. (1993 [1925]). *L'Expérience du proverbe.* Paris: L'Échoppe.

Paumier, S. (2013). *Unitex 3.1 bêta. Manuel d'utilisation*. Marne-la-Vallée: Université Paris-Est Marne-la-Vallée.

Paveau, M.-A., & Sarfati, G.-E. (2003). *Les grandes théories de la linguistique. De la grammaire comparée à la pragmatique.* Paris: Armand Colin.

Pearson, J. (1998). *Terms in Context.* Amsterdam/Philadelphia: John Benjamins.

Permyakov, G. L. (1979). *From Proverb to Folk-Tale.* Moscou: USSR Academy of Sciences "Nauka" Publishing House.

Permyakov, G. L. (1979). Notes On Structural Paremiology. In G. L. Permyakov, *From Proverb To Folk-Tale* (p. 130-159). Moscou: USSR Academy of Sciences "Nauka" Publishing House.

Permyakov, G. L. (1997 [1982]). On the Question of a Russian Paremiological Minimum. *De Proverbio 3(2)*, 256-273.

Polguère, A. (2003). *Lexicologie et sémantique lexicale. Notions fondamentales.* Montréal: Presses de l'Université de Montréal.

Przepiórkowski, A. (2004). *The IPI PAN Corpus. Preliminary version.* Consulté le 12/11/2013 de IPI PAN Corpus: http://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/book_en.pdf

Quitout, M. (2002). *Proverbes et énoncés sentencieux.* Paris: L'Harmattan.

Quitout, M., & Sevilla Muñoz, J. (dir.) (2009). *Traductologie, proverbes et figement.* Paris: L'Harmattan.

Radziszewski, A., Kilgarriff, A., & Lew, R. (2011). *Polish Word Sketches.* Consulté le 12/11/2013 de Sketch Engine: http://www.sketchengine.co.uk/documentation/attachment/wiki/AK/Papers/2011-RadziszewskiKilgLew.pdf

Rastier, F. (s.d.). Enjeux épistémologiques de la linguistique de corpus. Consulté le 12/11/2013 de *Texto! Textes et cultures* : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html

Renouf, A., Kehoe, A., & Banerjee, J. (2007). WebCorp: an integrated system for web text search. In M. Hundt, N. Nesselhauf, & C. Biewer (dir.), *Corpus Linguistics and the Web* (p. 47-68). Amsterdam: Rodopi.

Riegel, M. (1986). « Qui dort dîne » ou le pivot implicatif dans les énoncés parémiques. *Travaux de Linguistique et de Littérature 24(1)*, 85-99.

Riegel, M., Pellat, J.-C., & Rioul, R. (2009). *Grammaire méthodique du français.* Paris: PUF.

Rodegem, F. (1972). Un problème de terminologie: les locutions sentencieuses. *Cahiers de l'Institut de Linguistique de Louvain 1(5)*, 677-703.

Rodegem, F. (1984). La parole proverbiale. In F. Suard & C. Buridant (dir.), *Richesse du proverbe. Vol. 2. Typologie et Fonctions* (p. 121-135). Lille: Presses Universitaires de Lille.

Rozumko, A. (2012). English influence on Polish proverbial language. In C. Furiassi, V. Pulcini, & F. R. González (dir.), *The Anglicization of European Lexis* (p. 261-277). Amsterdam/Philadelphia: John Benjamins.

Sarfati, G.-E. (2005). *Eléments d'analyse du discours.* Paris: Armand Colin.

Schapira, C. (1999). *Les stéréotypes en français: proverbes et autres formules.* Paris: Ophrys.

Schapira, C. (2000). Proverbe, proverbialisation et déproverbialisation. *Langages 139*, 81-97.

Schmale, G. (2013). Qu'est-ce qui est préfabriqué dans la langue ? – Réflexions au sujet d'une définition élargie de la préformation langagière. *Langages 189*, 27-45.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.* Manchester.

Schulze-Busacker, E. (1985). *Proverbes et expressions proverbiales dans la littérature narrative du moyen âge français.* Paris/Genève: Champion/Slatkine.

Schulze-Busacker, E. (2012). *La Didactique profane au Moyen Âge.* Paris: Classiques Garnier.

Serrone, G. (2013). Les adages de droit : classement et analyse de corpus. *Cahiers de recherche de l'école doctorale en linguistique française 7*, 47-58.

Sevilla Muñoz, J. (1993). Las paremias españolas: clasificación, definición y correspondencia francesa. *Paremia 2*, 15-20.

Sevilla Muñoz, J. (2000). Les proverbes et les phrases proverbiales français, et leurs équivalences en espagnol. *Langages 139*, 98-109.

Sevilla Muñoz, J. (2008). Formas paremiológicas y criterios de clasificación (francés-español). *Critica del testo XI(1-2)*, 235-248.

Sevilla Muñoz, J., & García Yelo, M. (2008). El refranero hoy en Francia. *Paremia 17*, 209-222.

Silberztein, M. (2003). *Nooj manual*. Téléchargeable à l'adresse : <www.nooj4nlp.net> (date de consultation : 10/02/2014).

Simon, A. C., & Degand, L. (2011). L'analyse en unités discursives de base : pourquoi et comment ? *Langue Française 170*, 45-59.

Sinclair, J. McH. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. McH. (1995, Mai). Corpus Typology. A Framework for Classification. In G. Melchers, & B. Warren (dir.), *Studies in Anglistics* (p. 17-33). Stockholm: Almqvist & Wiksell. Consulté le 12/11/2013 de EAGLES: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Sinclair, J. McH. (1999). A Way with Common Words. In H. Hasselgård, & S. Oksefjell (dir.), *Out of Corpora. Studies in Honour of Stig Johansson* (p. 157-179). Amsterdam: Rodopi.

Sinclair, J. McH. (2004 [1996]). The Search for Units of Meaning. In J. M. Sinclair, & R. Carter (dir.), *Trust the Text. Language, Corpus and Discourse* (p. 24-47). Londres/New York: Routledge.

Sinclair, J. McH. (2005). Corpus and Text - Basic Principles. In M. Wynne (dir.), *Developing Linguistic Corpora: a Guide to Good Practice* (p. 1-16). Oxford: Oxbow Books. Consulté le 12/11/2013 de http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

Sinclair, J. McH. (2008). Preface. In S. Granger, & F. Meunier (dir.), *Phraseology: An Interdisciplinary Perspective* (p. xv-xviii). Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. McH., & Renouf, A. (1991). Collocational Frameworks in English. In K. Ajimer, & B. Altenberg (dir.), *English Corpus Linguistics* (p. 128-143). Londres/New York: Longman.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics 8(2)*, 209-243.

Stubbs, M. (2007). On Texts, Corpora and Models of Language. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert, *Text, Discours and Corpora* (p. 127-161). Londres: Continuum.

Tamba, I. (2000). Formules et dire proverbial. *Langages 139*, 110-118.

Taylor, A. (1962 [1931]). *The Proverb and An Index to 'The Proverb'*. Hatboro/Copenaghen: Folklore Associates/Rosenkilde & Bagger.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins.

Tsaknaki, O. (2006). Locating proverbs with finite-state transducers. *Toward Computational Models of Literary Analysis. Workshop of the International Conference on Language Resources and Evaluation (LREC-2006)*, (p. 57-62).

Visetti, Y.-M., & Cadiot, P. (2006). *Motifs et proverbes. Essai de sémantique proverbiale.* Paris: PUF.

Wilmet, M. (2007). *Grammaire critique du français [4e édition]*. Bruxelles: De Boeck Duculot.

Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Zouogbo, J.-P. (2008). Traduire le proverbe : à la recherche de concordances parémiologiques en bété pour la constitution d'un corpus trilingue allemand/français/bété. *META 53(2)*, 310-323.

Zouogbo, J.-P. (2011). Prolégomènes à l'établissement d'un minimum parémiologique pour le français. In A. Pamies Bertrán, J. d. Luque Durán, & P. Fernández Martín (dir.), *Paremiología y herencia cultural* (p. 97-106). Granada: Granada Linguistica.

SITOGRAPHIE

Dernière date de consultation : 22 avril 2014.

Base de données des corpus oraux de français hors de France, Délégation Générale à la Langue Française et aux Langues de France,
<http://www.dglflf.culture.gouv.fr/ (> *Études et recherches*)>

BootCaT, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Università degli Studi di Bologna,
< http://bootcat.sslmit.unibo.it/>

Centre National de Ressources Textuelles et Lexicales, CNRS/ATILF,
<http://cnrtl.fr/>

Czech National Corpus / Český národní korpus (*CNC*), Ústav Českého národního korpusu, Filozofické fakultě, Univerzity Karlovy v Praze,
< http://ucnk.ff.cuni.cz>

Consortium Corpus Écrits, Institut de Linguistique Française,
<http://corpusecrits.corpus-ir.fr/>

Corpora List,
<http://www.hit.uib.no/corpora/>

Corpus de Langue Parlée en Interaction (CLAPI), Laboratoire ICAR, Université Lyon 2/École Normale Supérieur de Lyon,
< http://clapi.univ-lyon2.fr/>

Corpus of Contemporary American English, Brigham Young University,
<http://corpus.byu.edu/coca/>

David Lee's Corpus-based Linguistics Links,
<http://www.uow.edu.au/~dlee/CBLLinks.htm>

Dictionnaire de la langue française, É. Littré,
<http://www.littre.org/>

Dictionnaire du Moyen Français (DMF), ATILF,
<http://www.atilf.fr/dmf/>

El Refranero Multilingüe, Instituto Cervantes,
<http://cvc.cervantes.es/lengua/refranero/Default.aspx>

Frantext, ATILF,
<http://www.frantext.fr/>

Gallica, Bibliothèque Nationale de France,
<http://gallica.bnf.fr/>

GlossaNet,
<http://glossa.fltr.ucl.ac.be/>

Google Ngram Viewer, Google Ngram Viewer Team, Google Research,
<http://books.google.com/ngrams>

IPI PAN Corpus, Zespół Inżynierii Lingwistycznej, Instytucie Podstaw Informatyki PAN,
<http://korpus.pl/index.php?lang=en>

Leipzig Corpora Collection, Institut für Informatik, Universität Leipzig,
<http://corpora.uni-leipzig.de/>

LIGM – équipe *Modèles et Algorithmes* / *Linguistique pour le traitement des langues*,
Université Paris-Est Marne-la-Vallée,
<http://infolingu.univ-mlv.fr/>

Mannheim Cosmas II, Institut für Deutsche Sprache,
<http://www.ids-mannheim.de/cosmas2/>

Matti Kuusi International Database of Proverbs,
<http://lauhakan.home.cern.ch/lauhakan/int/cerpint.html>

Nooj,
< http://www.nooj4nlp.net/>

Polskiego Wydawnictwa Naukowego (PWN) *Corpus*,
<http://korpus.pwn.pl/index_en.php>

REDAC, Laboratoire CLEE-ERSS de l'Université de Toulouse II-Le Mirail,
<http://redac.univ-tlse2.fr/>

Scientext, Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (LIDILEM) de l'Université Stendhal-Grenoble 3
<http://scientext.msh-alpes.fr/>

Sketch Engine, Lexical Computing Ltd.,
<http://www.sketchengine.co.uk/>

SMS4SCIENCE, Centre de traitement automatique du langage (CENTAL) de l'Université Catholique de Louvain,
<http://www.sms4science.org>

TreeTagger, Institute for Computational Linguistics, Universität Stuttgart,
<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

Trésor de la Langue Française informatisé (TLFi), ATILF,
<http://atilf.atilf.fr/>

Unitex 3.1, LIGM, Université Paris-Est Marne-la-Vallée,
<http://www-igm.univ-mlv.fr/~unitex/>

WaCky – The Web-As-Corpus Kool Yinitiative,
<http://wacky.sslmit.unibo.it/doku.php>

WebCorp, Research and Development Unit of English Studies, Birmingham City University,
<http://www.webcorp.org.uk/live/>

Il n'y a pas de mot qui soit le premier ou le dernier,
et il n'y a pas de limites au contexte dialogique (…).
Les sens passés eux-mêmes,
ceux qui sont nés du dialogue avec les siècles passés, ne seront jamais stabilisés
(clos, achevés une fois pour toutes).
Ils se modifieront toujours (se renouvelant)
dans le déroulement du dialogue subséquent, futur. (…)
en un point donné, dans le déroulement du dialogue,
au gré de son évolution, des sens seront remémorés de nouveau
et ils renaîtront sous une forme renouvelée (dans un contexte nouveau).
Il n'est rien qui soit mort de façon absolue.
Tout sens fêtera un jour sa renaissance.

(M. Bakhtine, « Remarques sur l'épistémologie des sciences humaines », 1974
dans *Esthétique de la création verbale*, Paris, Gallimard, 1984, p. 393)

Non, rien de rien
Non, je ne regrette rien.

(E. Piaf, « *Non, je ne regrette rien* », 1956)