Vincenzo Lambertini
Université de Ferrare
Italie
https://orcid.org/0000–0002–4562–9967

Approche *corpus-driven* à l'étude du proverbe italien et français

Résumé. En règle générale, la parémiologie définit et catégorise les proverbes. Mais d'autres aspects également importants comme leur fréquence d'utilisation, leur contexte d'emploi, leurs significations les plus communes, leurs variantes les plus connues ou les modifications les plus fréquentes qu'on leur fait subir doivent être prises en compte. Cette étude présente les données obtenues grâce à l'application des principes de l'approche *corpus-driven* à deux corpus comparables, l'un italien et l'autre français, créés à partir du *web* et propose une méthodologie d'analyse des proverbes fondée sur les données réelles afin de répondre aux exigences des professionnels des langues, comme les interprètes ou les traducteurs, mais aussi les étudiants, qui sont confrontés à la compréhension, à la traduction, voire à l'énonciation des proverbes.

Mots clés: *proverbes italiens, proverbes français, corpus linguistics, approche corpus-driven, corpus comparables*

Introduction

Traditionnellement, la plupart des études portant sur la parémiologie ont été marquées par des *démarches top-down*, ou *démarches descendantes*[1]. Autrement dit, la parémiologie s'est spécialement concentrée sur des proverbes tirés de listes ou de dictionnaires existants pour confirmer des théories élaborées au préalable. Les études ont très souvent porté sur des proverbes sélectionnés *a priori* ou sur la base d'enquêtes menées auprès d'échantillons

[1] Cet article faisant référence à la linguistique des corpus, qui a trouvé son épanouissement dans le monde anglo-saxon, il nous a semblé naturel de maintenir la terminologie anglaise qui lui est associée.

de population et souvent en présence du chercheur, ce qui pourrait entraîner en quelque sorte une carence d'objectivité à l'égard des éléments analysés et des conclusions atteintes.

Une approche linguistique descendante ou *top-down* analyse d'abord la langue envisagée comme système pour parvenir à la langue en tant que réalisation par le biais de données réelles qui visent à confirmer les théories de départ. Cette démarche est suivie fréquemment en parémiologie : on produit avant tout des théories qui sont par la suite démontrées par des données authentiques sélectionnées en aval de ce processus. Par exemple, les parémiologues visant à démontrer que la brévité, la concision et le rythme sont des conditions nécessaires des proverbes, le feront sur la base d'un ensemble de proverbes affichant principalement ces caractéristiques. Toutefois, il suffit de présenter des contre-exemples pour démentir ces théories[2]. Voilà pourquoi, nous sommes persuadés que les approches de type *top-down* présentent des limites non négligeables.

En effet, le point de départ de l'approche *top-down* n'est pas la langue en usage, mais des théories élaborées au préalable, ce qui empêche au chercheur de se focaliser seulement sur le matériel parémiologique exploité par les locuteurs.

Or, à notre sens, dans le domaine de la parémiologie, il est incontournable de se pencher sur ce que les locuteurs natifs réalisent en termes linguistiques : autrement dit, il faut se demander, par exemple, quels sont les proverbes qu'ils utilisent le plus fréquemment, mais aussi comment ils s'en servent, dans quels contextes et à quelles fins. En effet, il n'est pas dit que tous les proverbes qui se trouvent dans un dictionnaire de proverbes ou dans une liste de proverbes soient effectivement employés par les locuteurs ; en même temps, il est tout aussi probable (comme il sera par ailleurs démontré en § 4.2) que des proverbes utilisés, voire très fréquemment utilisés, par les locuteurs natifs n'apparaissent pas dans les œuvres consacrées aux proverbes.

Limites des approches *top-down*

L'un des objectifs de notre recherche a été d'explorer les limites des approches de type *top-down* par le biais d'expérimentations conçues à ces fins. En particulier, nous visions à vérifier dans quelle mesure les proverbes

[2] A titre d'exemple, cette démarche est suivie par Anscombre (2000 : 12–13) lorsqu'il dément la théorie selon laquelle les proverbes ne sont que des phrases, sur la base de contre-exemples sélectionnés *ad hoc*, en démontrant ainsi qu'ils peuvent être des « discours ».

recensés par les principaux dictionnaires ou récoltes de proverbes étaient présents dans des corpus linguistiques contemporains.

Pour ce faire, nous sommes partis d'un échantillon général et représentatif de proverbes. Une liste de proverbes a été créée à partir d'un des dictionnaires de proverbes italiens les plus complets : le *Dizionario dei proverbi italiani* élaboré par Carlo Lapucci (2006). Pour faire en sorte que la subjectivité et le bagage culturel du chercheur n'influencent pas le choix des proverbes pris en compte, ces proverbes ont été sélectionnés au hasard. Pour ce faire, une liste de 500 nombres aléatoires a été générée : les proverbes du dictionnaire de Lapucci (2006) étant numérotés, on a attribué à chaque nombre sélectionné un proverbe issu du dictionnaire lui-même.

Le résultat de cette expérience a été éloquent : seuls 6% des proverbes sélectionnés (soit 33 proverbes sur les 500 repérés) étaient présents dans le corpus italien itWaC (un corpus de référence et synchronique qui sera décrit en § 3.2). La raison d'un pourcentage aussi faible est probablement liée à l'erreur méthodologique suivante : on a tâché de repérer des objets théoriques (les proverbes issus d'un dictionnaire) dans un répertoire de textes authentiques (le corpus itWaC). Autrement dit, la présupposition de base, selon laquelle les proverbes sélectionnés étaient effectivement utilisés dans le corpus analysé, était erronée. Néanmoins, rien n'indiquait au préalable que les proverbes issus du dictionnaire de Lapucci (2006) sont encore utilisés de nos jours : en effet, dans les dictionnaires que nous avons analysés[3] il est possible de remarquer une pénurie d'annotations concernant l'usage des proverbes. Il semble que les dictionnaires de proverbes aient pour but de recueillir des proverbes (parfois même de grandes quantités de proverbes[4]) et d'en fournir des explications, à l'aide d'exemples tirés de sources littéraires, sans pour autant les considérer sous une optique plus communicative, pragmatique ou synchronique. Etant donné que les dictionnaires n'établissent pas toujours une distinction nette entre les proverbes qui sont encore employés et les proverbes qui sont par contre désuets[5], il est impossible de comprendre s'ils sont utilisés de nos jours et, le cas échéant, dans quelles situations et en raison de quelles significations.

[3] La liste des dictionnaires examinés se trouve en bibliographie (A).

[4] Par exemple, le dictionnaire de Lapucci (2006) contient environ 25.000 proverbes.

[5] Il serait par exemple intéressant, voire extrêmement utile, de diviser les proverbes en « proverbes du passé » et « proverbes actuels » pour répondre aux exigences non seulement de ceux qui désirent approfondir leurs connaissances en matière de traditions ou de folklore, mais aussi de ceux qui nécessitent de se pencher sur les proverbes qui sont utilisés le plus fréquemment à l'heure actuelle.

Cette démarche de type *top-down* ne permet pas toujours de repérer les proverbes souhaités dans un corpus et, qui plus est, elle ne fournit pas assez d'éléments pour en comprendre la raison. En effet, rien ne permet de savoir si les proverbes cherchés ne sont effectivement pas présents dans le corpus ou bien s'ils le sont mais sous d'autres variantes ou encore s'ils y apparaissent mais de manière légèrement modifiée (par exemple, avec des mots, des signes de ponctuation ou des incises ajoutés à leur intérieur)[6]. Voici quelques exemples qui clarifient ce dernier point :

(1) *I primi amori sono i migliori* (trad. lit. : Les premières amours sont les meilleures)

(2) *Il primo amore non si scorda mai* (trad. lit. : On n'oublie jamais le premier amour)

Comme ces deux proverbes ont à peu près la même signification parémiologique, on peut les considérer comme des synonymes. Toutefois, il y a une différence importante entre ces deux proverbes : le proverbe (1) n'est pas présent dans le corpus itWaC, alors que le proverbe (2) est attesté et produit 70 résultats (il y a donc 70 occurrences de la séquence exacte *"il" "primo" "amore" "non" "si" "scorda" "mai"*). Ces deux proverbes, qui se ressemblent du point de vue de leur signification, sont placés l'un après l'autre dans le dictionnaire de Lapucci (2006) mais comme il s'agit de proverbes différents, ils ont été identifiés par deux numéros différents : le proverbe (1) est le 751 de la lettre A, le proverbe (2) est le 750 de la lettre A. Le nombre aléatoire obtenu dans la phase préliminaire était le 751 de la lettre A et non pas le 750. Un italophone quelconque reconnaîtrait immédiatement le proverbe (2) comme un véritable proverbe, alors que le même individu considérerait (1) comme une phrase libre[7]. Cette distinction peut être déduite de la signification[8] du proverbe (2) fournie par le dictionnaire de Lapucci (2006).

---

[6] La recherche des proverbes étant effectuée sur la base de leurs mots, et donc d'une combinaison de signes, il suffit en effet qu'il y ait un signe de ponctuation de plus, ou un blanc de plus ou encore un mot légèrement différent pour que le logiciel de recherche n'arrive pas à détecter les proverbes souhaités.

[7] Il est question d'un constat majeur qui a des retombées non négligeables, entre autres, sur la traduction : en effet, pour garder en langue cible le même effet et la même force du proverbe en langue source, le traducteur devra opter pour des proverbes ayant le même sens mais étant effectivement connus par les destinataires de sa traduction.

[8] En effet, au début de l'explication fournie par l'auteur, on lit : « *Molto vivo e diffusissimo* » (Lapucci, 2006 : 42 ; trad. : « Très courant et très répandu »). Il s'agit d'une remarque concernant

Voici maintenant un autre exemple de proverbes qui n'ont pas été repérés dans le corpus itWaC.

(3) *Ti piace il cacio?* (trad. lit. : Tu aimes le fromage *cacio* ?)

(4) *Hai voluto la bicicletta, pedala* (trad. lit. : Tu as voulu un vélo, pédale)

Si l'on cherche ces proverbes dans le logiciel de traitement du corpus itWaC[9], en utilisant exactement les mots de (3) et de (4), on ne parvient à aucun résultat. Certes, le proverbe (3) serait inconnu des locuteurs italophones, qui auraient par conséquent du mal à en saisir le sens, mais (4) leur serait sans aucun doute familier. Cela étant, pourquoi est-il impossible de le repérer ? La raison de cet échec réside probablement dans la considération suivante : effectivement, rien ne nous dit que le proverbe cherché ne soit pas présent dans le corpus. La seule conclusion que l'on puisse en tirer est que l'exacte séquence de mots des proverbes cherchés est absente du corpus lui-même.

En effet, en variant les critères de recherche, les résultats obtenus sont différents : il suffit de chercher toute séquence composée des mots "*hai*" "*voluto*" "*la*" "*bicicletta*" pour parvenir à 21 résultats, dont les plus représentatifs sont énumérés ci-après :

(5) a. *Hai voluto la bicicletta? Adesso pedali!* (trad. lit. : Tu as voulu ton vélo ? Maintenant tu pédales !)

b. *Hai voluto la bicicletta? Pedala!* (trad. lit. : Tu as voulu ton vélo ? Pédale !)

c. *Hai voluto la bicicletta adesso pedala!* (trad. lit. : Tu as voulu ton vélo maintenant pédale !)

d. *Hai voluto la bicicletta? E adesso pedala.* (trad. lit. : Tu as voulu ton vélo ? Et maintenant pédale.)

Les auteurs des exemples (5) utilisent des variantes du proverbe analysé qui prévoient l'insertion de quelques éléments entre le mot *bicicletta* et le verbe *pedala* (sans oublier que dans la plupart des cas, c'est l'impératif du verbe

l'usage du proverbe qui n'est pour autant présente que pour le proverbe (2) ; en revanche, l'auteur ne se penche pas sur l'usage du proverbe (1), ce qui confirme une nouvelle fois que l'usage n'est pas toujours au premier rang des préoccupations des dictionnaires de proverbes.

[9] A savoir NoSketch Engine, URL : http://nl.ijs.si/noske/index-en.html. Consulté le 15 septembre 2016.

pedalare qui est utilisé, sauf dans (5a) où c'est l'indicatif *pedali* qui est employé). On peut par exemple remarquer que certains locuteurs ont ajouté un simple point d'interrogation, ou l'adverbe *adesso* (trad. : maintenant) ou encore la conjonction *e* (trad. : et) suivie de l'adverbe *adesso*. Il s'ensuit qu'il est presque impossible de savoir *a priori* sous quelle forme ou variante un proverbe sera retrouvé dans un corpus ; de la même manière, un insuccès concernant le repérage d'un proverbe sur la base des mots dont il est composé pourrait être dû à une série de raisons qui vont de l'absence du proverbe à la présence du proverbe mais avec des variations (pouvant être minimales, comme l'ajout d'un signe de ponctuation ou d'un blanc).

Pour toutes ces raisons, une variation d'approche s'impose, en privilégiant une démarche de type *bottom-up* ou *ascendante*. Du point de vue linguistique et parémiologique, une approche *bottom-up* présente plusieurs atouts : elle permet de formuler des hypothèses en s'appuyant sur des données réelles, authentiques et non influencées par le chercheur ; elle octroie aux professionnels des langues (interprètes ou traducteurs qu'ils soient) les clefs pour mieux comprendre la réalité linguistique telle qu'elle est, en mettant en évidence les points cruciaux dont il faut tenir compte en traduction et en interprétation.

Vers une parémiologie *bottom-up*

Appliquer la méthode *bottom-up* à la parémiologie ne signifie pas seulement repérer des proverbes dans des textes (ou mieux des corpus), mais surtout retrouver de manière automatique (ou semi-automatique) des proverbes dans des corpus linguistiques.

Un corpus n'est qu'une récolte de textes qui sont censés être représentatifs d'une langue donnée (Tognini-Bonelli 2001 : 52–59). La représentativité est alors l'une des caractéristiques qui ne peuvent jamais faire défaut : on interroge un corpus en raison de sa représentativité. Néanmoins, un corpus n'est qu'un échantillon d'une langue, vu qu'il est impossible de recueillir toutes les occurrences linguistiques d'une langue (Chiari 2007 : 41–43). Voilà pourquoi, les proverbes repérés dans un corpus linguistique[10] synchronique seront les plus représentatifs des proverbes utilisés actuellement dans cette langue.

[10] Ces critères concernent notamment l'authenticité, la représentativité et la finalité. Voir à ce titre : Tognini-Bonelli (2001 : 55) ; Chiari (2007 : 42–43 et 51) ; Leech (1991) ; Biber (1994).

Le premier défi à relever est sans aucun doute le repérage automatique des proverbes. Pour ce faire, il sera nécessaire de se tourner vers la linguistique de corpus, en considérant spécialement les avancées qui ont été accomplies dans le repérage automatique des expressions figées et des métaphores.

Approche *corpus-based* ou *corpus-driven* ?

En linguistique de corpus, les approches *top-down* et *bottom-up*, qui ont été abordées ci-dessus, correspondent à deux démarches[11] connues respectivement sous le nom d'*approche corpus-based*[12] et *corpus-driven*[13].

On peut considérer la première approche comme la plus traditionnelle : c'est la démarche typique des grammairiens et des linguistes du passé qui formulaient des théories en s'appuyant sur leur expérience acquise dans le domaine linguistique et qui cherchaient par la suite des preuves qui puissent les confirmer.

La seconde approche, dite *corpus-driven*, est opposée : le chercheur interroge un corpus pour repérer des *pattern* récurrents, qui lui serviront pour formuler ses théories. Cette approche est donc de type *bottom-up* : on part des données fournies par le corpus pour formuler des hypothèses qui mènent à des généralisations en termes de règles, qui permettent à leur tour de formuler une théorie (Tognini-Bonelli 2001 : 17).

Or, dans la pratique, l'approche *corpus-driven* impose au chercheur de détecter le matériel à analyser sans qu'il intervienne dans son repérage, pour ne pas fausser les conclusions auxquelles il parvient. Cela se traduit par une recherche automatique ou semi-automatique qui se sert de critères objectifs pour la détection du matériel à examiner. Très souvent, grâce à l'annotation morpho-syntactique des corpus, il est possible de repérer des catégories de mots à l'aide des étiquettes qu'on leur a attribuées. Cette approche est très fréquemment utilisée pour la détection de co-occurences d'éléments linguistiques tels que les collocations ou les colligations[14]. En effet, elle est à même de mettre en évidence tous les mots qui entourent un mot sélectionné, ce qui permet d'avoir une vision précise des relations sémantiques et syntaxiques du mot cherché. Mais que se passe-t-il lorsqu'on doit chercher non pas un mot mais une série de mots, voire une phrase ?

---

[11] Pour cette comparaison, voir McEnery & Hardie (2012 : 150–152).

[12] Voir Tognini-Bonelli (2001 : 65–81).

[13] Voir Tognini-Bonelli (2001 : 84–99).

[14] Voir à ce titre : Sinclair (1996) et Chiari (2007 : 77–78).

Il convient de souligner que les proverbes, contrairement à d'autres unités phraséologiques, comme les collocations ou les expressions figées, sont des phrases. Selon nos recherches, en effet, d'un point de vue linguistique ce qui distingue véritablement les proverbes des expressions figées, par exemple, est le fait que les proverbes sont des phrases, alors que les expressions figées sont des constituants, à l'instar des substantifs, des verbes ou des locutions, entre autres. C'est pour cette raison qu'il est impossible de détecter des proverbes sur la base de critères morpho-syntactiques ou lexicaux, étant donné que les proverbes sont des phrases (qui sont autonomes) et qu'ils n'ont pas de schémas morpho-syntactiques fixes (par exemple, verbe + déterminant + substantif) comme peuvent en présenter les expressions figées[15]. En outre, plusieurs études[16] montrent que les proverbes peuvent faire l'objet de modifications importantes et qu'ils sont moins figés que d'autres phrases ou expressions, étant donné que tout proverbe peut avoir un nombre indéfini de variantes. Ces caractéristiques peuvent entraver remarquablement la recherche des proverbes dans un corpus.

Pour toutes les raisons qui viennent d'être énumérées, il est manifeste que le repérage automatique des proverbes pose de nombreux problèmes aux chercheurs qui visent à appliquer l'approche *corpus-driven* à l'étude des proverbes. Toutefois, avant de montrer l'une des solutions possibles pour surmonter ces obstacles, il convient de se pencher sur les deux corpus linguistiques qui ont été choisis pour mener cette recherche.

itWaC et *frWaC* : deux corpus comparables pour une analyse parémiologique

Le choix des corpus à utiliser revêt une importance cruciale au sein de la linguistique de corpus. Dans cette étude, les corpus à analyser ont été sélectionnés en fonction des objectifs établis dès le début : mener une analyse comparative sur les proverbes italiens et français sous une optique synchronique. Pour ce faire, nos corpus devaient remplir certaines conditions de base pour répondre aux nécessités qu'imposait notre recherche.

Premièrement, il était nécessaire d'avoir un corpus de grandes dimensions permettant de compenser la faible fréquence d'utilisation des proverbes, bien qu'en littérature parémiologique il soit difficile de trouver

[15] Voir à ce titre Gross (1996) qui analyse des locutions en les regroupant selon leurs structures morpho-syntactiques.

[16] Voir Schapira (2000 : 81–97) et Michaux (1999) entre autres.

des données certaines concernant la fréquence d'utilisation moyenne des proverbes[17]. Cependant, pour de simples raisons statistiques, nous avons supposé que puisque les proverbes sont des phrases et non pas des constituants, ils doivent être fatalement moins fréquents que les expressions figées.

Deuxièmement, notre but était de mener une étude comparée de proverbes italiens et français sur des données authentiques et non pas influencées (et potentiellement faussées) par la traduction. Voilà pourquoi il était nécessaire d'exclure a priori des corpus parallèles et de privilégier à leur place des corpus comparables : les premiers sont des corpus comprenant des textes en langue source et les traductions correspondantes en langue cible alors que les seconds sont des corpus composés de textes, en deux ou plusieurs langues, avec des caractéristiques comparables[18].

Enfin, nous visions à repérer des corpus assez généraux pour mener une analyse parémiologique tous azimuts. En particulier, nous souhaitions travailler sur des corpus généraux comprenant différentes typologies de texte, différents contextes, voire différents registres linguistiques. En outre, comme il serait impossible de prendre en considération des discours oraux et spontanés, et que les corpus oraux sont encore trop petits pour mener une analyse satisfaisante sur les proverbes, nous aurions souhaité analyser des corpus contenant des textes à la frontière entre la langue écrite et la langue orale.

Pour toutes ces raisons, nous avons décidé d'utiliser les corpus frWaC et itWaC[19]. Il est question de deux corpus comparables, et donc construits sur la base des mêmes critères, de très grandes dimensions[20], composés de textes repérés automatiquement sur la Toile, étiquetés sur le plan morphosyntaxique (ce qu'on appelle en linguistique de corpus *POS tagging*) et lemmatisés. Il s'agit pour la plupart d'articles, de blogs et de forums publiés sur Internet au cours des premières années 2000[21]. Les blogs et les forums permettent notamment d'avoir accès à des données linguistiques écrites qui tentent parfois de calquer la spontanéité ainsi que la rapidité de la langue

[17] Norrick (1985 : 6), par exemple, affirme que dans un corpus de conversations en anglais (qui est le suivant : A corpus of English Conversation) constitué de 43.165 lignes, qui correspondent à 891 pages, on ne peut repérer qu'un seul proverbe.

[18] Pour plus d'informations, voir Tognini-Bonelli (2001 : 6–7) et Chiari (2007 : 53–54).

[19] Voir Baroni et al. (2008).

[20] Plus de détails peuvent être trouvés dans Baroni et al. (2008).

[21] Sur la base des concordances du mot proverbe et du mot proverbio on a remarqué que la plupart des textes où se trouve le mot proverbe ont été publiés en 2007, alors que la majorité des textes contenant le mot proverbio sont apparus en 2005 (ce qui confirme une correspondance quant à la datation moyenne des textes). A ce titre, on renvoie à Lambertini (2016 : 115–119).

orale, ce qui permet d'élargir les résultats obtenus à la langue parlée, même s'il ne faut pas oublier qu'il ne s'agit que d'une approximation qui ne saurait remplacer des transcriptions de conversations authentiques.

Détection automatique des proverbes

Le repérage automatique des proverbes dans des corpus comme itWaC et frWaC ayant environ deux milliards de mots pose évidemment des problèmes : d'une part, on ne peut pas lire tous les textes des corpus pour détecter les proverbes qu'ils contiennent ; d'autre part, il est impossible de repérer des proverbes sur la base de l'annotation morphosyntaxique et de la lemmatisation, puisque rien du point de vue morphosyntaxique ou lexical ne permet de discriminer une phrase libre d'un proverbe.

Pour sortir de cette impasse, nous nous sommes inspirés des expériences menées en linguistique de corpus, afin d'étudier certains phénomènes, comme le figement et la métaphore. En particulier, nous avons analysé les expériences de repérage de métaphores, qui partagent avec les proverbes l'impossibilité d'être détectées automatiquement sur la base de critères morphosyntaxiques.

En abordant le sujet de la détection de métaphores dans des corpus linguistiques, Stefanowitsch (2006) remarque que rien n'empêche que les métaphores soient accompagnées d'expressions telles que *kind of* (une sorte de) et *so to speak* (pour ainsi dire). Ces expressions dites *marqueurs de métaphore* introduisent ou indiquent la présence d'un ou plusieurs mots utilisés de manière métaphorique. Parallèlement, en parémiologie[22] on sait que les proverbes peuvent être introduits ou accompagnés d'expressions indiquant leur statut proverbial. Les auteurs que nous avons considérés indiquent plusieurs expressions de ce type, dont les suivantes sont les plus représentatives : « comme on dit », « comme dit le proverbe », « on le sait », « la bonne sagesse populaire » (Schapira 2000) ; « on a bien raison de dire que », « si j'en crois la sagesse populaire », « comme (le) dit le proverbe », « comme dit un proverbe » (Kleiber 1999). Par analogie avec la métaphore, on pourrait nommer ces expressions *marqueurs de proverbe*.

Ce qui n'est pas suffisamment étudié en parémiologie c'est l'usage des marqueurs de proverbe et notamment une analyse du marqueur le plus fréquent. Certes, Schapira (2000 : 89–90) remarque que la pratique de signa-

[22] Voir à ce titre Shapira (2000), Kleiber (1999), Cram (1983), entre autres.

ler la présence d'un proverbe par une expression spécifique était certainement suivie au XVII[e] siècle, quoique alors « comme de nos jours, le proverbe s'insère directement dans le discours, sans aucune formule introductrice. Sa notoriété seule garantit dans ce cas son statut de citation » (Schapira 2000 : 90). En dépit de ce constat, il est possible de trouver des marqueurs de proverbe, ce qui constitue la seule étiquette formelle permettant de repérer automatiquement des proverbes.

Déterminer le marqueur de proverbe le plus utilisé et donc le plus prolifique est une lourde tâche. Pour ce faire, nous avons mené de brèves recherches exploratoires et nous sommes parvenus à la conclusion que le marqueur le plus fréquent et le plus « transversal » (à savoir, le plus utilisé par n'importe quel locuteur et dans n'importe quel domaine ou situation) est le mot *proverbio*, en italien, et *proverbe*, en français.

En cherchant les mots *proverbio* et *proverbe* respectivement dans le corpus italien itWaC et dans le corpus français frWaC, nous avons obtenu une série de proverbes co-occurrents. Il s'agit d'un petit pas vers la détection automatique des proverbes dans des corpus de très grandes dimensions à l'aide de critères d'interrogation des corpus, ce qui permet d'appliquer les principes de la linguistique de corpus et l'approche *corpus-driven* à l'étude du proverbe.

Premiers résultats

Cette méthode d'analyse linguistique des proverbes effectuée sur la base de données authentiques mène à une série de considérations cruciales. D'abord, les premiers résultats obtenus grâce à la méthode de repérage automatique de proverbes appliquée aux deux corpus synchroniques mentionnés ci-dessus démontrent que les proverbes sont encore utilisés dans la communication courante.

Ensuite, les concordances des mots *proverbio* et *proverbe* mettent en évidence que l'analyse des proverbes reposant sur la démarche *corpus-driven* commentée ci-dessus et appliquée aux deux corpus italien et français est comparable. En effet, sous une optique quantitative, il convient de souligner qu'un total de 617 proverbes italiens sont obtenus à partir du corpus itWaC, face à 630 proverbes français détectés dans le corpus frWaC. Ces chiffres font référence aux proverbes énoncés en combinaison avec le marqueur, hormis les répétitions : en effet, si on comptait également ces dernières, l'on obtiendrait 1 110 proverbes pour l'italien et 974 proverbes pour le français. Certes, les données quantitatives issues des corpus ne peuvent être prises que pour des indications générales et approximatives, ayant une simple valeur statistique.

Toutefois, on ne peut pas ignorer que le pourcentage des proverbes italiens et français qui s'accompagnent respectivement du mot *proverbio* et *proverbe* est équivalent, ce qui assure une analyse pleinement comparative.

Nature des proverbes repérés

Il convient toutefois de se pencher sur la nature des proverbes repérés. Il est en effet nécessaire de se demander si le marqueur utilisé pour extraire automatiquement les proverbes des deux corpus n'accompagne que des proverbes ou bien s'il peut également introduire des phrases ou d'autres types d'expressions. Cette question est légitime, puisque comme on l'a souligné précédemment, les données des deux corpus sont issues de textes rédigés par des locuteurs qui n'ont pas forcément une compétence parémiologique spécifique.

Il est donc nécessaire de remarquer que parmi les résultats obtenus figurent des phrases qui ne peuvent pas être considérées comme des proverbes ou bien des expressions qui ne respectent pas le seuil minimal des proverbes, à savoir le fait d'être au moins une phrase (Lambertini 2016). Voici quelques exemples extraits du corpus frWaC.

(6) a. Il se plaisait à demander qu'on laissât du temps au temps. François Mitterrand appliqua ce proverbe à l'Europe et il surmonta ses contradictions avec brio [...].

b. José Bové voit plutôt d'un bon œil cette proposition qui, à rebours du proverbe, lui ferait lâcher l'ombre de l'Elysée pour la proie du Palais Bourbon...

c. C'était le calme avant la tempête, comme le dit bien le proverbe.

Dans l'exemple (6a), le mot *proverbe* fait référence à « qu'on laissât du temps au temps » : *laisser du temps au temps* n'est pas un proverbe, puisque ce n'est pas une phrase mais un constituant. Cela vaut également pour l'exemple (6b), où cette expression figée modifiée s'avère très intéressante sous plusieurs aspects. D'abord, l'expression standard *lâcher la proie pour l'ombre* a été renversée, ce qui est souligné par le marqueur *à rebours du proverbe.* En outre, les deux métaphores de l'ombre et de la proie sont actualisées, spécifiant de quelle ombre et de quelle proie il s'agit (*l'ombre de l'Elysée* et *la proie du Palais Bourbon*). Ensuite, le marqueur du proverbe est ici utilisé pour indiquer une expression et non pas une phrase. Il s'ensuit que contrairement à ce

qu'affirment nombre de chercheurs[23], les expressions figées peuvent être modifiées et que parfois « l'homme de la rue » peut les considérer comme des proverbes. Ces réflexions peuvent être également appliquées à l'exemple (6c), où nous avons une expression (*le calme après la tempête*) modifiée (*c'était le calme avant la tempête*). Pour connaître la variante que les dictionnaires considèrent comme la forme base de cette expression, nous avons consulté le dictionnaire *Le petit Robert de la langue française* (2006), qui place cette expression sous l'entrée *calme*. Ce dictionnaire indique que la seule variante de l'expression prise en examen est *le calme après la tempête*. En consultant le corpus frWaC, il se trouve que la variante la plus commune est *le calme avant la tempête* avec 63 occurrences (cette variante n'est pas censée exister selon le dictionnaire qui ne contemple pour cette expression qu'une forme) contre les 24 occurrences de l'expression *le calme après la tempête*.

Grâce à l'utilisation de l'approche *corpus-driven*, des faits récurrents sont mis en évidence. Premièrement, on constate que les expressions figées peuvent être modifiées : si les modifications sont souvent claires, irréfutables, voire explicitées par le marqueur qui les accompagne, comme dans l'exemple (6b), elles peuvent parfois être moins nettes et plus indéfinies, surtout si dans la pratique elles sont plus fréquentes que les variantes présentées par les dictionnaires (voir à ce titre l'exemple (6c) et les différentes fréquences des deux versions repérées). Deuxièmement, on assiste parfois à un glissement de la notion de proverbe qui parvient dans certains cas à être appliquée à des expressions plus ou moins figées.

Après avoir étudié de plus près les données françaises et italiennes qui ont été collectées, la présente recherche a mis en évidence un fait essentiel, à savoir ce que nous avons nommé *compétence parémiologique*. Autrement dit, il a été possible, grâce aux corpus linguistiques sélectionnés, d'étudier la capacité des locuteurs à reconnaître et à utiliser les proverbes. Notre hypothèse de base était la suivante : si une recherche *corpus-driven* ne retient que les proverbes qui sont considérés comme tels par un nombre minimal de locuteurs, il est probable qu'on parvient à détecter un ensemble de proverbes qui réunissent effectivement les conditions nécessaires et suffisantes des proverbes[24]. Il était donc fondamental de déterminer le seuil en-dessus duquel il y avait plus de 50% de probabilités de repérer de véritables proverbes.

---

[23] Voir par exemple Gross (1996 : 9–23), Anscombre (2005 : 24).

[24] Ces critères ne sont pour autant pas évidents. En général, les proverbes remplissent les conditions suivantes : ils doivent être des phrases « ON-sentencieuses », connues et dont le sens est restreint à l'homme. Pour une explication plus approfondie, nous renvoyons à Lambertini (2016 : 7–46).

En linguistique de corpus, pour formuler une description lexicographique d'un mot, pas moins de 20 occurrences de ce mot (Sinclair 2005) sont nécessaires. Bien que conscients de la différence qu'il y a entre un mot (constituant) et un proverbe (phrase) et de la difficulté à en satisfaire ce critère des 20 occurrences, nous avons opté pour l'appliquer telle quelle à l'analyse des proverbes. Nous avons déjà souligné que les concordances du mot *proverbio* et *proverbe* amènent au repérage des proverbes et de leurs occurrences. A ce titre, le proverbe le plus fréquent en français, accompagné du marqueur *proverbe*, était *l'union fait la force*, avec 11 occurrences et le proverbe italien le plus récurrent était *l'unione fa la forza* 13 fois employé[25]. On était toutefois bien loin du seuil des 20 occurrences. Il était donc nécessaire de chercher les proverbes que nous avions obtenus dans les deux corpus sans utiliser aucun marqueur de proverbe. Pour ce faire, nous avons cherché ces proverbes sur la base des variantes détectées au début. Nous avons évité les proverbes ayant une seule occurrence initiale, afin de ne pas répéter les erreurs méthodologiques commises lorsque nous avons essayé de chercher les proverbes sur la base des variantes uniques présentées par les dictionnaires (voir § 2).

Ainsi, a-t-il été possible de calculer le seuil minimal de fréquence initiale des proverbes qui était de 5 occurrences. Autrement dit, dans nos corpus les proverbes ayant une fréquence à partir de 5 ont une probabilité supérieure à 60% d'avoir plus de 20 occurrences réelles dans les corpus ; ce pourcentage s'élève à 100% à partir du seuil de 7 occurrences. La fréquence initiale associée aux marqueurs de proverbe est très utile pour construire automatiquement un corpus de proverbes, bien qu'elle n'indique pas la fréquence réelle des proverbes[26].

En outre, une fréquence initiale de 5 ou plus assure de repérer de véritables proverbes, en évitant de mêler les proverbes aux expressions figées, par exemple. Cela indique que si la compétence parémiologique individuelle peut être parfois assez faible, la compétence parémiologique collective, quant à elle, est bien plus précise. L'explication réside probablement dans la nature

[25] Remarquons qu'il s'agit du même proverbe, ce qui souligne encore une fois le degré de ressemblance entre les parémiologies française et italienne.

[26] Par exemple, le proverbe italien *Meglio tardi che mai* présente 5 occurrences avec le mot *proverbio* et 460 occurrences totales sans utiliser ce marqueur de proverbe ; en revanche, le proverbe *Il buon giorno di vede dal mattino* a une fréquence de 7 avec le mot *proverbio* et 366 occurrences sans ce marqueur (a savoir presque 100 occurrences moins par rapport au premier proverbe qui était plus fréquent au début) ; et encore, l'un des proverbes les plus fréquents en co-occurrence avec le mot *proverbio*, à savoir *Tra il dire e il fare c'è di mezzo il mare*, a 11 occurrences, mais il ne présente que 213 occurrences sans le mot *proverbio*.

des proverbes : ce ne sont pas seulement des phrases mais plus précisément des phrases génériques typifiantes à priori (Anscombre 2000 : 12) ayant un ON-énonciateur (Anscombre 2000 : 11–12). Cela veut dire que l'auteur du proverbe est inconnu, étant donné que c'est le « savoir partagé, la science populaire, l'observation quotidienne » (Anscombre 2000 : 11) qui est l'auteur des proverbes. Autrement dit, il y a « un énonciateur premier, même s'il est indéfini, diffus, non spécifique, et qui met à la disposition de la communauté linguistique un principe général dont il autorise ainsi l'application à des cas particuliers » (*ibid.*). C'est donc la communauté tout entière qui a le droit d'utiliser un proverbe et en même temps de certifier la nature proverbiale d'une phrase.

Ecart entre réalité linguistique et parémiologie *top-down*

En suivant cette démarche, il a été possible de repérer des proverbes qui remplissent effectivement toutes les conditions nécessaires pour être considérés tels mais qui sont ignorés par les principaux dictionnaires de proverbes que nous avons examinés. Nous présentons ici quelques exemples issus du corpus frWaC.

(7) a. Tel père, tel fils (occurrences : 22 totales, 6 initiales)

b. Quand le sage montre la lune, l'idiot regarde le doigt (occurrences : 66 totales, 6 initiales)

c. Quand on aime on ne compte pas (occurrences : 89 totales, 7 initiales)

d. Qui ne tente rien n'a rien (occurrences : 128 totales, 10 initiales)

e. Le temps c'est de l'argent (occurrences : 158 totales, 5 initiales)

Il n'est pas facile de formuler des hypothèses convaincantes pour expliquer l'absence des proverbes (7) dans les dictionnaires papiers utilisés pour cette recherche. On pourrait penser que ces ressources estiment que certains proverbes ne sont pas français et qu'il ne méritent pas d'être pris en compte (*tel père, tel fils* est en réalité d'origine latine, comme nombre d'autres proverbes d'ailleurs, et *Quand le sage montre la lune, l'idiot regarde le doigt* relève plutôt de la sagesse confucéenne). Toutefois, cet argument est discutable, vu que ces dictionnaires considèrent également des proverbes étrangers. On pourrait alors supposer que, d'après ces dictionnaires, les proverbes

énumérés en (7) ne sont pas de véritables proverbes ou bien qu'ils commencent à être fréquemment utilisés de nos jours, mais qu'ils ne l'étaient pas par le passé, ce qui pourrait indiquer une mise à jour inachevée des dictionnaires. Quelle que soit la réponse, il faudrait se pencher sur les raisons de ce manque et chercher à faire en sorte que les dictionnaires soient plus sensibles à la réalité linguistique et fassent une attention accrue aux changements en cours.

Il a déjà été rappelé qu'il n'est pas dit que les proverbes obtenus grâce à leur marqueur, ayant une occurrence initiale inférieure à 5, soient véritablement des proverbes. Toutefois, parmi ces résultats il est aussi possible de rencontrer des proverbes « en puissance », à savoir des phrases qui ne sont pas encore des proverbes (notamment car leur fréquence est encore trop faible pour qu'elles relèvent de la compétence parémiologique d'une communauté) mais qui pourraient le devenir dans l'avenir. C'est le cas par exemple de ce « proverbe boursier » qui a été repéré dans le corpus frWaC : *Quand Wall Street éternue, Paris s'enrhume*. Certes, on ne peut pas encore le considérer comme un proverbe puisque sa fréquence d'utilisation est très réduite. Toutefois il s'agit d'une nouvelle forme de sagesse qui n'a plus affaire au monde rural d'autrefois mais qui cherche à transférer la force des proverbes traditionnels sur des sujets plus modernes et plus proches de notre réalité.

Conclusions

Une analyse des proverbes menée sur des bases strictement linguistiques est possible. Une méthodologie de recherche semi-automatisée capable de trouver et, ensuite, analyser des proverbes sans les connaître au préalable est qui plus est concevable.

Pour mener une étude de ce type, il faut pouvoir compter sur des corpus très grands, comme les deux corpus dont nous nous sommes servis, itWaC et frWaC, qui contiennent environ 2 milliards de mots chacun. Du point de vue méthodologique, il est incontournable d'adopter une approche *corpus-driven*, à savoir une démarche *bottom-up* qui part des données et de leur observation pour formuler des théories.

Cette méthode n'est certes pas sans insuccès mais les obstacles sont surmontables. Afin de repérer des proverbes de manière (semi-) automatique, les simples marqueurs de proverbe que sont, pour l'italien, le mot *proverbio* et, pour le français, le mot *proverbe*, s'avèrent à ce propos très performants.

Cette méthodologie permet en outre d'attirer l'attention sur une caractéristique peu mise en relief mais qui revêt une importance majeure dans le domaine des proverbes, à savoir la *compétence parémiologique* des locuteurs. En effet, grâce à cette méthodologie de recherche, l'on constate que souvent les locuteurs confondent les proverbes avec d'autres phénomènes de figement linguistique (par exemple les expression figées), ce qui n'est pas acceptable du point de vue linguistique, étant donné que les proverbes sont des phrases alors que les expressions figées sont des constituants. Pour contourner cet écueil, il suffit de repérer des proverbes fréquemment utilisés qui montrent une forte probabilité d'être de véritables proverbes.

Les socles de cette étude ayant été jetés, une série de possibilités s'ouvrent : par exemple, l'analyse du fonctionnement de phénomènes tels que le détournement des proverbes (Schapira 2000), mais aussi l'obtention d'une liste de fréquence des proverbes, ou encore l'analyse des différents contextes d'utilisation des proverbes pour en saisir la signification parémiologique la plus authentique.

A notre avis, toutefois, l'un des acquis principaux de cette méthodologie est d'avoir démontré qu'il est possible (voire qu'il convient) d'analyser les proverbes tels qu'ils sont utilisés, sans s'appuyer sur aucune théorie préétablie mais seulement sur des données linguistiques. En effet, par ce biais, il est possible de certifier que les proverbes ne sont pas en voie de disparition : ils ne traversent qu'une phase de changements liés sans doute aux variations de notre société. Comme il a été souligné dans cet article, de vieux proverbes disparaissent, mais en même temps d'autres commencent à se frayer un chemin. Nos données semblent donc suggérer qu'il faut appliquer à la parémiologie cette maxime empruntée à la physique : *Rien ne se perd, rien ne se crée, tout se transforme*.

Abréviations

Trad. lit. – traduction littéraire

Trad. – traduction

Bibliographie

Dictionnaires

Boggione, Valter ; Massobrio, Lorenzo. 2004. *Dizionario dei proverbi*. Torino : Utet.

Dournon, Jean-Yves. 1986. *Dictionnaire des proverbes et dictons de France*. Paris : Hachette.

Guazzotti, Paola; Oddera, M. Federica. 2010. *Il grande dizionario dei proverbi italiani*. Bologna : Zanichelli.

Lapucci, Carlo. 2006. *Dizionario dei proverbi italiani*. Firenze : Le Monnier.

Maloux, Maurice. 1980 [2009]. *Dictionnaire des proverbes, sentences et maximes*. Paris : Larousse.

Montreynaud, Florence Pierron, Agnès ; Suzzoni, François. 1989. *Dictionnaire de proverbes et dictons*. Paris : Dictionnaires Le Robert.

Littérature

Anscombre, Jean-Claude. 2000. Parole proverbiale et structures métriques. *Langages*, vol. 34, 139 : 6–26.

Anscombre, Jean-Claude. 2005. Les proverbes : un figement du deuxième type ? *Linx*, vol. 53, 17–33.

Biber, David. 1994. Representativeness in Corpus Design. In : Zampolli, A., Calzolari, N. e M. Palmer (eds.) *Current Issues in Computational Linguistics: in Honour of Don Walker*. Linguistica Computazionale IX.X. Giardini Editori e Stampatori in Pisa e Kluwer Academic Publishers.

Casadei, Federica. 1996. *Metafore ed espressioni idiomatiche: uno studio semantico sull'italiano*. Roma : Bulzoni.

Chiari, Isabella. 2007. *Introduzione alla linguistica computazionale*. Roma : GLF Editori Laterza.

Cram, David. 1983. The linguistic status of the proverb. *Cahiers de lexicologie. Revue internationale de lexicologie et de lexicographie*. Vol. XVLIII (II) : 53–71.

Deignan, Alice. 2009. Searching for Metaphorical Patterns in Corpora. In : Baker, P. (ed.) *Contemporary Corpus Linguistics*. London; New York: Continuum : 9–31.

Gross, Gaston. 1996. *Les expressions figées en français*. Gap, Paris: Ophrys.

Kleiber, George. 1999.. Les proverbes : des dénominations d'un type « très très spécial ». *Langue française*, 123 : 52–69.

Kleiber, Gorge. 2000. Sur le sens des proverbes. *Langages*, vol. 34, 139 : 39–58.

Lakoff, George ; Johnson, Mark. 1980 [2003]. *Metaphors we live by*. Chicago ; London : The University of Chicago Press.

Lambertini, Vincenzo. 2016. *Approccio linguistico e* corpus-driven *al proverbio italiano e francese: alla ricerca della forma perduta*. Thèse de doctorat, Université de Bologne.

Leech, Geoffrey. 1991. The State of the Art in Corpus Linguistics. In : Aijmer, K. e B. Aitenberg (eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London, New York : Longman : 8–29.

McEnery, Tony; Hardie, Andrew. 2012. *Corpus linguistics: method, theory and practice*. Cambridge ; New York : Cambridge University Press.

Michaux, Christine. 1999. Proverbes et structures stéréotypées. *Langue française*, 123, : 85–104.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English. A Corpus-based Approach.* Oxford: Clarendon Press.

Norrick, Neal R. 1985. *How proverbs mean: semantic studies in English proverbs*. Berlin : Mouton.

Perrin, L. (2000). Remarques sur la dimension générique et sur la dimension dénominative des proverbes. *Langages*, vol. 34, 139 : 69–80.

Prandi, Michele 2006. *Le regole e le scelte. Introduzione alla grammatica italiana.* Torino: UTET.

Privat, M. 1997. Proverbes, metaphors et traduction. *Paremia*, 6 : 511–154.

Shapira, Claudia. 2000. Proverbe, proverbialisation et déproverbialisation. *Langages*, vol. 34, 139 : 81–97.

Sinclair, John McH. 1996. The search for units of meaning. *Textus*, vol. 9, 1 : 71–106.

Sinclair, John McH. 2005. Corpus and text – Basic principles. In : Wynne, M. (eds.) *Developing linguistic corpora A guide to good practice*. Oxford: Oxbow Books : 1–16.

Stefanowitsch, Anatol; Gries, Stephan T. (a cura di). 2006. *Corpus-based approaches to metaphor and metonymy*. Berlin ; New York : Mouton de Gruyter.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam : John Benjamins.

Vietri, Simonetta (1985). *Lessico e sintassi delle espressioni idiomatiche. Una tipologia tassonomica dell'italiano.* Napoli : Liguori Editore.

Sources en ligne

Barbadillo de la Fuente, M. Teresa et al. *Refranero Multilingüe*. http://cvc.cervantes.es/lengua/refranero/Default.aspx (consulté le 10 janvier 2013).

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano; Zanchetta, Eros. 2008. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. In : *Language Resources and Evaluation*, vol. 43, 3 : 209–226. http://wacky.sslmit.unibo.it/doku.php?id=publications (consulté le 9 octobre 2014).

frWaC (French Web). URL : http://nl.ijs.si/noske/wacs.cgi/first?corpname=frwac&reload=

itWaC (Italian Web). URL : http://nl.ijs.si/noske/wacs.cgi/first?corpname=itwac&reload=

The *corpus-driven* approach in studies on Italian and French proverbs

Summary

As a general rule, paremiology aims to collect, define and categorize proverbs, in particular by using literary sources or surveys conducted on the basis of samples of language users. However, in order to carry out researches in paremiology producing results that are beneficial to the fields of translation, interpretation and language learning, it is necessary to take into consideration other equally important aspects such as frequency of use, meanings or variations that characterize contemporary proverbs. To do this, researchers should use authentic linguistic data, favoring bottom-up analysis approaches. Within the framework of this research, two corpora were used: the Italian reference corpus called itWaC and the French reference corpus called frWaC. These are two comparable, synchronic and very large corpora, which ensures an objective description of contemporary Italian and French paremiology.