Le participe passé dans un grand corpus littéraire (1700-2019) : enjeux et limites du traitement textométrique dans Hyperbase

Federica Beghini, Laurent Vanni

Le participe passé dans un grand corpus littéraire (1700-2019) : enjeux et limites du traitement textométrique dans Hyperbase[1]

Federica Beghini[1], Laurent Vanni[2]

[1] Università degli Studi di Padova, Université Côte d'Azur,
BCL – federica.beghini@phd.unipd.it
[2] Université Côte d'Azur, CNRS, BCL – laurent.vanni@univ-cotedazur.fr

RÉSUMÉ

L'étude du participe passé a fait l'objet de nombreuses réflexions dans le domaine de l'analyse des données textuelles : en raison des ambigüités de son profil grammatical, il a été depuis toujours considéré comme difficile à définir pour la recherche textométrique (Brunet, 1988 ; Engwall, 1966). Dans cet article, nous aborderons cette question en soulignant les enjeux de la méthodologie, en décrivant ses limites et ses atouts et en proposant également des solutions à ses limites.

Pour ce faire, nous mettrons à l'épreuve Hyperbase, l'un des logiciels d'ADT historiques en France, sur un corpus de littérature française couvrant une période de plus de trois siècles. De la distribution statistique à l'Analyse Factorielle des Correspondances (AFC), nous montrerons que le logiciel permet d'interroger la linguistique concernant les variations d'usage du participe passé en fonction du temps, des œuvres ou des auteurs. Enfin, une étude de cas viendra illustrer ce parcours interprétatif sur un auteur dont l'œuvre se caractérise par plusieurs variables (diachronique, linguistique et générique).

Ainsi, outre l'exploration des enjeux de la méthode, nous proposerons également un large éventail de pistes de recherche pour l'étude textométrique de cette catégorie grammaticale dans le domaine littéraire.

1. INTRODUCTION

L'étude du participe passé en analyse de données textuelles (ADT) combine deux approches : d'une part, l'informatique et la statistique qui autorisent des explorations quantitatives des textes et d'autre part, la linguistique qui non seulement définit les lignes directrices de la compilation des corpus numériques et les démarches quantitatives à suivre, mais s'occupe aussi de l'interprétation qualitative (Lebart *et al*., 2019). Hyperbase est l'un des logiciels d'ADT historiques en France (Pincemin, 2018), pionnier dans l'utilisation de lemmatiseurs et d'analyseurs morphosyntaxiques (Brunet, 1999). Ce logiciel s'est vu doté au fil des années d'outils permettant, par exemple, le repérage et l'étude du participé passé dans de larges corpus : il met en œuvre de nombreuses solutions de visualisation qui permettent divers parcours interprétatifs, ainsi que l'utilisation de *métadonnées*, qui permettent de définir des sous-corpus concernant par exemple un auteur, un genre ou une période déterminée (année, décennie, etc.). Aujourd'hui, la version accessible en ligne[2] facilite l'interrogation de nombreux corpus littéraires, médiatiques ou politiques et la création de bases de données personnalisées à partir de fichiers textes.

Dans cette contribution, après avoir brièvement présenté les composantes qualitatives à considérer en linguistique et en informatique, nous explorerons les pistes de recherche que les fonctionnalités

[1] Les parties 1, 2.2 et 3.1 ont été écrites à quatre mains ; les sections 2.1 et 4 et 5 ont été rédigées par F. Beghini.

[2] http://hyperbase.unice.fr.

du logiciel peuvent offrir aux linguistes pour l'étude du participe passé. Pour les illustrer, nous utiliserons une base comprenant des textes littéraires du XVIII^{e} siècle à nos jours.

Nous présenterons d'abord des parcours de recherche concernant la base dans son ensemble, pour ensuite nous focaliser sur un sous-corpus de littérature contemporaine. De cette façon, le logiciel sera d'abord mis à l'épreuve par l'étude des tendances linguistiques couvrant une période de trois siècles, ensuite par l'analyse d'un sous-corpus spécifique – l'œuvre d'un auteur contemporain. Notre choix s'est porté sur Milan Kundera compte tenu des spécificités linguistiques de son œuvre qui nous permettent d'explorer un plus grand nombre de fonctionnalités du logiciel.

2. MÉTHODE

2.1 Le participe passé en linguistique

Pour mener des analyses quantitatives du participe passé, il faut d'abord élaborer des requêtes de recherche adaptées au logiciel. Pour ce faire, ses fonctions et ses contextes d'usage doivent être analysées dans le détail. Plus précisément, l'étude du participe passé nécessite de considérer à la fois son **emploi nu** et son utilisation dans une **construction périphrastique**. Dans les deux cas, il peut remplir aussi bien une fonction adjectivale qu'une fonction verbale.

Le participe passé intervient dans des constructions périphrastiques : il sert à constituer les temps composés des verbes avec les auxiliaires *avoir* ou *être*. Plus précisément, les temps composés à la forme active expriment l'aspect accompli ou marquent une antériorité. Quand le participe passé forme, à l'aide d'un auxiliaire, le passif des verbes transitifs, il peut signifier un procès en cours ou l'état résultant de l'achèvement du procès. Toutefois, sa valeur verbale est annulée quand il remplit la fonction d'attribut du sujet. Dans ce cas, il peut être remplacé par un adjectif ou être modifié par un adverbe marquant le degré : *La pelouse était couverte de faibles vapeurs condensées* (Nerval). En revanche, il est impossible de lui assigner une valeur adjectivale quand il exprime un procès au passif avec un complément d'agent (*Mon repas a été cuisiné par sa mère*) ou quand il entre dans la forme composée d'un verbe actif (*Il est parti à l'heure*) (Riegel *et al*., 1994, p. 593-6).

L'emploi nu du participe passé peut jouer le rôle d'un adjectif qualificatif quand il peut être remplacé par une subordonnée relative contenant une forme verbale comportant l'auxiliaire *être* : *Les manifestant [qui sont] partis de Milan sont arrivés à midi* ou *Les candidats [qui sont] admis à l'épreuve sont au nombre de quarante*. Toutefois, dans ce dernier cas, en tant que verbe transitif à la voix passive, le participe passé peut conserver un certain dynamisme verbal, parce que le complément d'agent est virtuellement présent, même s'il n'est pas actualisé. Le participe passé à fonction adjectivale peut constituer un groupe épithète du nom (*Les moissonneurs* ***couchés*** *faisaient des groupes sombres*, Hugo), apposé (*Gavroche,* ***fusillé****, taquinait la fusillade*, Hugo) ou attribut d'un complément d'objet direct (*Je le trouve très* ***énervé***). Quand il est complètement détaché du verbe d'origine, il bénéficie d'une entrée distincte de celle du verbe dans le dictionnaire : c'est le cas par exemple d'*énervé*. Quand il est employé sans auxiliaire, il ne peut que remplir une fonction verbale quand il constitue la tête du groupe verbal d'une proposition subordonnée participiale, à la voix active (*Le moment venu, il envoya le signal convenu*) comme à la voix passive (*Le projet achevé, ils sont rentrés chez eux*) (Riegel *et al*., 1994, p. 593-5).

2.2 Le participe passé avec Hyperbase

L'étude du participe passé au moyen d'outils informatiques exige le repérage et l'annotation systématique de cette catégorie grammaticale. Si le traitement peut être manuel, il se révèle long et fastidieux pour de "grands corpus". On note toutefois quelques réalisations d'envergure comme le corpus latin du LASLA qui s'est doté au fil des ans d'un corpus annoté manuellement de plus de 2 millions de mots, qui autorise le repérage et le comptage de tous les usages du participe en latin classique.

Mais si le logiciel Hyperbase accepte l'annotation manuelle, il propose aussi une méthode automatique qui repose sur des algorithmes d'apprentissage profonds et permet l'annotation morphosyntaxique de vastes corpus. De nombreuses langues sont prises en charge par ce système et le gain de temps par rapport à une annotation manuelle est considérable (quelques secondes pour plusieurs millions de mots). Néanmoins, les performances de ces algorithmes sont à relativiser au regard des résultats obtenus. En effet, ces méthodes sont bien moins fiables qu'une exécution humaine, avec un taux d'erreur qui se situe en moyenne à 10 % suivant les corpus, les langues et les modèles entraînés. Toutefois, lorsque le volume est suffisant et que les données sont homogènes, la statistique permet de négliger ces erreurs et donne malgré tout des tendances stables et robustes.

Le repérage du participe passé concerne uniquement la catégorie « participe » et le temps « passé » (qui correspond dans le logiciel à la combinaison d'étiquettes *Past:Part*), mais ne différencie pas l'emploi nu et les constructions périphrastiques. De même, la machine ne discrimine pas l'emploi du participe passé en tant qu'adjectif ou verbe. Tout au plus l'étiquette *Pass* peut-elle être utilisée pour isoler les constructions à la voix passive, mais, on l'a vu, cette information ne peut pas à elle seule désambigüiser les deux valeurs adjectivale et verbale. Cette question n'est pas nouvelle dans le domaine de l'ADT (Brunet, 1988, p. 183) :

> Dans le cas du participe passé, plusieurs solutions ont été proposées en vue de faire le départ entre les valeurs verbale et adjectivale : relevé de la présence ou non-présence de l'auxiliaire ou établissement d'une liste de participes à considérer comme adjectifs. Comme nous, Lyne (1973, p. 88) insiste sur l'impossibilité de distinguer les deux valeurs. Et il ramène aussi au verbe toute occurrence du participe passé. (Engwall, 1962-1968, p. XXX)

Ainsi, pour contourner ce problème, nous proposons une solution permettant de réduire le pourcentage des résultats fautifs tributaires de l'ambigüité entre la valeur verbale et adjectivale. Pour isoler les participes passés dans les constructions périphrastiques, nous avons utilisé la requête "AUX *** Past:Part", qui permet de relever la suite auxiliaire+participe passé, où les deux unités sont contiguës ou non. On l'a vu, le participe passé remplit une fonction verbale dans presque tous les cas où il se trouve dans une telle construction, sauf lorsqu'il remplit la fonction d'attribut du sujet. C'est sur ce contexte d'usage que nous nous sommes concentrés dans notre étude à l'aide d'Hyperbase, limitant ainsi le champ d'étude aux seules formes périphrastiques.

Les outils proposés par le logiciel sont nombreux et variés. Ils reposent pour la plupart sur l'analyse de la fréquence et de la spécificité des occurrences dans le corpus. Cette approche est issue de l'hypothèse mathématique selon laquelle les textes suivent des lois statistiques telles que la loi normale dans un contexte d'équidistribution des mots, qui permet d'apprécier la sur/sous-utilisation des formes, lemmes ou catégories grammaticales dans les textes. Plus précisément, le calcul des spécificités (Lafon 1980) et l'Analyse Factorielle des Correspondances (AFC) (Benzécri, 1973), appliqués à la requête "AUX *** Past:Part", sont des moyens efficaces d'aborder l'étude du participe passé employé dans des constructions périphrastiques.

3. LE PARTICIPE PASSÉ AU FIL DES SIÈCLES

3.1 La littérature française de 1700 à aujourd'hui

Notre corpus littéraire comprend environ 90 millions de mots et 1251 œuvres représentant neuf genres littéraires différents [3] et une centaine d'auteurs. Sans être exhaustif, ce corpus se veut représentatif des productions littéraires des trois derniers siècles.

Nous avons choisi de nous concentrer sur la variable diachronique pour observer l'évolution des formes composées au fil des trois derniers siècles. Le calcul des spécificités illustre la distribution par décennie des formes composées (fig. 1). Ainsi, des variations significatives se dessinent : une

[3] Conte, correspondance, essai, mémoire, nouvelle, poésie, roman, théâtre, vers.

surutilisation de la construction périphrastique est observée pendant la seconde moitié du XVIII[e] siècle, ainsi que de 1950 jusqu'à nos jours. Enfin, le XIX[e] siècle se distingue par deux périodes où l'on trouve des pics de surutilisation.

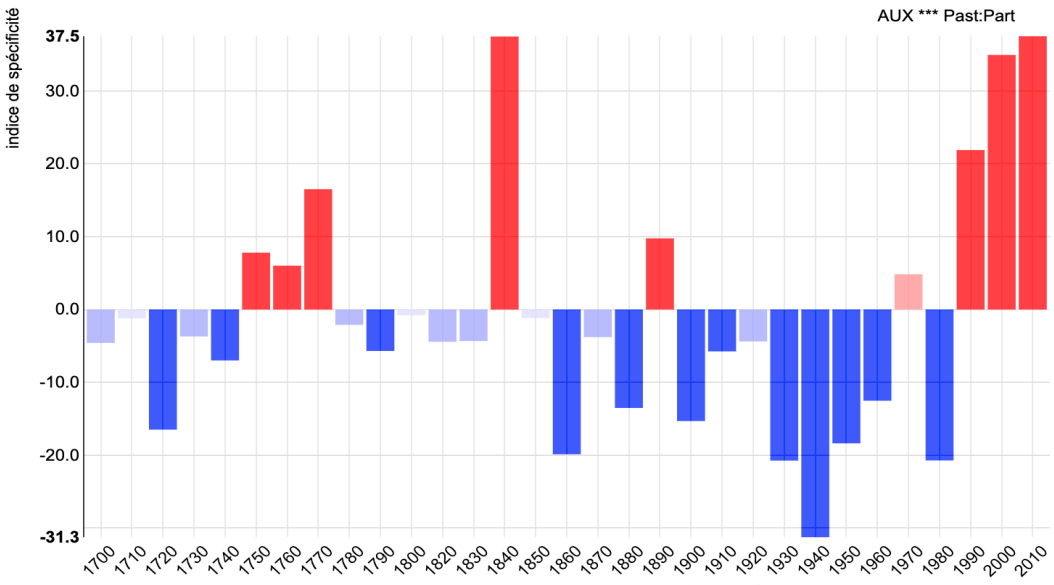

Figure 1 : Spécificités de la construction périphrastique Aux+Part Past de 1700 à 2019

Pour interpréter ces données, il est nécessaire de considérer les différentes variables du corpus (les genres, la diachronie, …) et de mener une étude plus approfondie afin d'observer les variations au fil des décennies.

Par exemple, pour les pics de surutilisation du XIX[e] siècle, une influence potentielle du romantisme sur l'emploi des formes composées pourrait constituer une hypothèse à vérifier dans d'ultérieures études intégrées : un penchant pour l'expression de l'intériorité, typique d'une partie de la production littéraire de la période, pourrait expliquer cette préférence pour le passé composé au détriment du passé simple. En effet, les études textométriques remarquent une corrélation positive entre le passé composé et la première personne du singulier ; cette forme composée permet en effet de mettre en évidence une proximité psychologique de l'énonciateur vis-à-vis des faits narrés (Imbs, 1960).[4] Une première étape pour vérifier cette hypothèse à l'aide d'Hyperbase comporterait l'étude de la distribution générique du corpus – avec un intérêt particulier pour les sous-corpus du genre de la poésie et de la correspondance – pendant les périodes où l'on observe cette surutilisation. Ensuite, chaque genre pourrait être considéré séparément au fil des décennies et/ou des années, de façon à examiner plus attentivement les différents textes qui composent le corpus et le profil stylistique de leurs auteurs.

Quant à la surutilisation vers la fin du XXe, il semble qu'il dissimule en fait le sous-emploi du passé simple comme en atteste la figure 2. Ce temps verbal se serait donc effacé à partir des années 80 au profit du passé composé, devenu plus populaire à l'oral en raison de sa simplicité d'emploi (moins de variations morphologiques), mais aussi visiblement dans la littérature. Pour aller plus loin dans cette analyse, il faudrait considérer la variable générique, la progression par année et examiner attentivement les textes composant les corpus.

Dans les deux cas, l'étude peut isoler les différentes variables en fonction de l'objectif de la recherche, si l'on souhaite se concentrer sur l'étude d'une période plutôt que d'une autre, sur un

[4] L'énonciateur veut faire retentir jusqu'à nous l'événement rapporté et le rattacher à notre présent. « C'est le temps de celui qui relate en témoin », en participant à l'action qu'il est en train de raconter. (Benveniste 1966, p. 244).

genre ou sur un auteur en particulier. Pour illustrer cette approche, nous avons délimité le champ de recherche en sélectionnant le sous-corpus textuel d'un des auteurs du corpus, Milan Kundera. Nous avons choisi l'œuvre de cet auteur parce que la présence de plusieurs variables en fait un enjeu plus complexe pour l'étude textométrique. En effet, outre la variable générique et diachronique, une autre variable linguistique doit être prise en compte, étant donné que la version définitive de son œuvre (*Œuvre I et II*, Gallimard, La Pléiade) est composée à la fois de traductions du tchèque et de textes rédigés directement en français.[5] De plus, la production de cet auteur disposait d'un modèle de référence linguistique (61 auteurs, 11 millions de mots) qui a été créé *ad hoc* dans le cadre d'un projet de recherche portant spécifiquement sur son œuvre[6]. Ce modèle de référence a été récemment intégré dans la base littéraire que nous avons présentée dans la section 3.1.

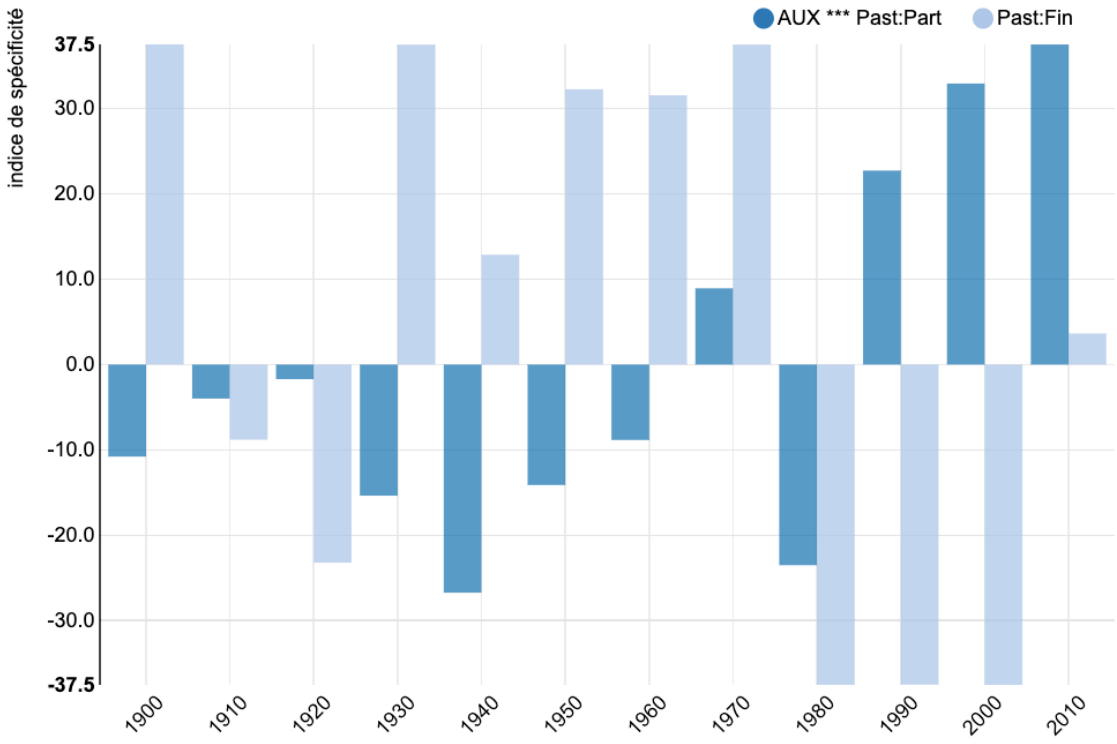

Figure 2 : Spécificités du passé simple et des formes composées (1900-2019)

[5] Les traductions des sept premiers textes ont été révisées par l'auteur et sont par-là considérées comme des versions françaises originales et même plus originales que les versions tchèques, également parce que ces dernières ont été ultérieurement révisées à partir des versions françaises (Woods M. 2006, *Translating Milan Kundera*, Clevedon, Multilingual Matters, Ltd, p. 2). Cependant, pour les besoins de cette analyse, il faut prendre en compte la variable linguistique qui distingue les textes traduits de ceux étant écrits directement en français, étant donné que le processus de rédaction des premiers textes du cycle tchèque est différent de celui des derniers textes du cycle français.

[6] Beghini, F., *Étude textométrique de l'œuvre de Milan Kundera. À la recherche de la pépite d'or*. La numérisation des textes des 61 auteurs a été réalisée à des fins académiques uniquement, l'accès est interdit au public et il ne sera pas utilisé à des fins commerciales.

4. L'USAGE DES FORMES COMPOSÉES CHEZ KUNDERA

« Pour comprendre, il faut comparer », écrit Kundera (*Œuvre II*, p. 1068) pour introduire sa réflexion sur la figure du romancier. Cette phrase n'a bien évidemment rien à voir avec la statistique textuelle et en faire une devise de la textométrie pourrait être considéré de la même manière que les nombreux cas de malentendu et de mésentente qui parsèment son œuvre et dont l'*humour* jette sa lumière discrète sur les situations existentielles les plus disparates. Cependant, même si le but de sa réflexion ne concerne pas la statistique textuelle, il est tout de même vrai que, pour que celle-ci puisse accompagner et soutenir l'étude linguistique d'un corpus textuel, il est nécessaire de disposer d'un modèle linguistique auquel pouvoir le comparer, à savoir un corpus de référence. En d'autres termes, pour qu'une analyse à l'aide d'Hyperbase puisse relever les spécificités de l'œuvre de Kundera quant à l'emploi des formes composées, il faut se référer à une norme contrastive.

Ainsi, nous allons comparer le sous-corpus contenant la quasi-totalité de l'œuvre de Kundera[7], comprenant quinze textes (dix romans, quatre essais et un recueil de nouvelles) à la littérature contemporaine à cet auteur, à savoir celle de la période 1960-2019.

Pour ce modèle de référence, nous avons sélectionné des textes des auteurs contemporains des années 60 jusqu'à 2019, dont la qualité littéraire a été reconnue[8], écrits en français et qui n'appartiennent pas à des *discours*, *genres* et *champs génériques* absents du corpus de Kundera (Rastier, 2009). Enfin, pour assurer l'homogénéité du modèle, chaque décennie doit comprendre un nombre semblable de mots (un million d'occurrences). Au total, ce corpus de référence contient dix millions de mots ; le sous-corpus de Kundera en compte environ un million.

Nous avons créé deux partitions déterminées par les métadonnées relatives à l'*auteur*, en choisissant, pour une partition, tous les textes de Kundera et, pour l'autre, ceux des 60 autres auteurs contemporains qui avaient été sélectionnés selon les lignes directrices susmentionnées.

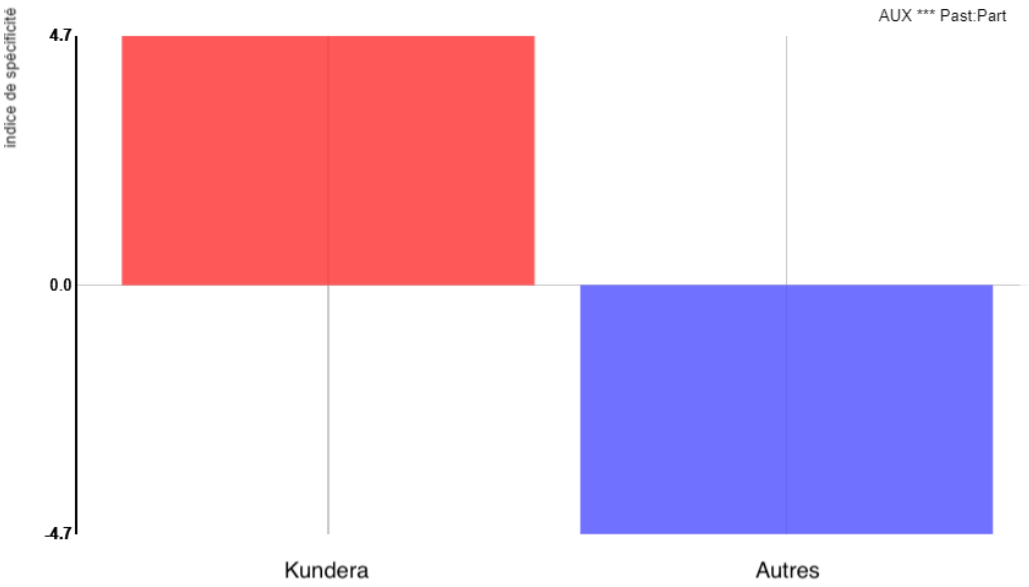

Fig. 3 – Suremploi des formes composées : le sous-corpus de Kundera *vs* le modèle de référence (1960-2019)

[7] Sauf la pièce théâtrale *Jacques le fataliste et son maître,* parce qu'il s'agit du seul texte de ce genre et, étant donné que l'on étudie les genres séparément, le sous-corpus qu'il aurait constitué aurait été trop petit.
[8] Pour ce faire, je me suis basée sur quatre études d'histoire des lettres, sur deux encyclopédies en ligne, sur les sites Internet de plusieurs maisons d'édition françaises et sur le palmarès des prix littéraires.

Tout d'abord, le calcul des spécificités relève **une surutilisation des formes composées dans son œuvre par rapport à la littérature française qui lui est contemporaine** : chez Kundera, la construction périphrastique est encore plus significative que dans la littérature contemporaine, avec un écart réduit de 4.7 (fig. 3). De plus, si l'on partitionne ces deux corpus en décennies, l'évolution des formes composées chez Kundera reflète celle de la langue littéraire contemporaine : plus précisément, son utilisation augment progressivement, surtout pendant les trois dernières décennies.

Toutefois, pour pouvoir interpréter ces données, il faut considérer **les différentes variables du sous-corpus kundérien**, pour voir si l'une d'elles influence davantage son emploi. Ainsi, nous avons examiné les différents sous-sous-corpus du corpus de Kundera, constitués sur la base du genre, de la période et de la langue, et nous les avons comparés à la norme contrastive représentée par le corpus de référence. En particulier, comme les sept traductions (un recueil de nouvelles et six romans) appartiennent à la période 1968-1990 et les huit derniers textes (quatre romans et quatre essais) écrits en français, à la période suivante 1990-2013, il n'est pas possible d'isoler la variable linguistique de la variable temporelle. En revanche, il est possible de l'isoler de la variable générique dans le cas des romans, dont six ont été traduits du tchèque et quatre rédigés directement en français. Pour les autres genres, cette distinction n'est pas possible, étant donné que les essais ont été écrits directement en français et que les nouvelles sont toutes des traductions.

Des analyses comparatives entre ces sous-corpus, il ressort que l'utilisation de la construction périphrastique du participe passé est **excédentaire dans les textes appartenant à la seconde période** (1990-2013), à savoir les romans écrits initialement en français et les autres **textes écrits à l'origine en français** (les essais). En revanche, **les textes écrits à l'origine en tchèque** – les nouvelles et les romans de la première période (1968-1990) – se caractérisent par **un usage déficitaire**. Ces analyses exogènes ont ainsi révélé des différences d'usage qui dépendent de **variables temporelle et linguistique**. Afin d'approfondir l'étude de ces constructions périphrastiques dans l'œuvre de Kundera, nous nous sommes donc concentrés sur une **étude endogène** du corpus de l'auteur.

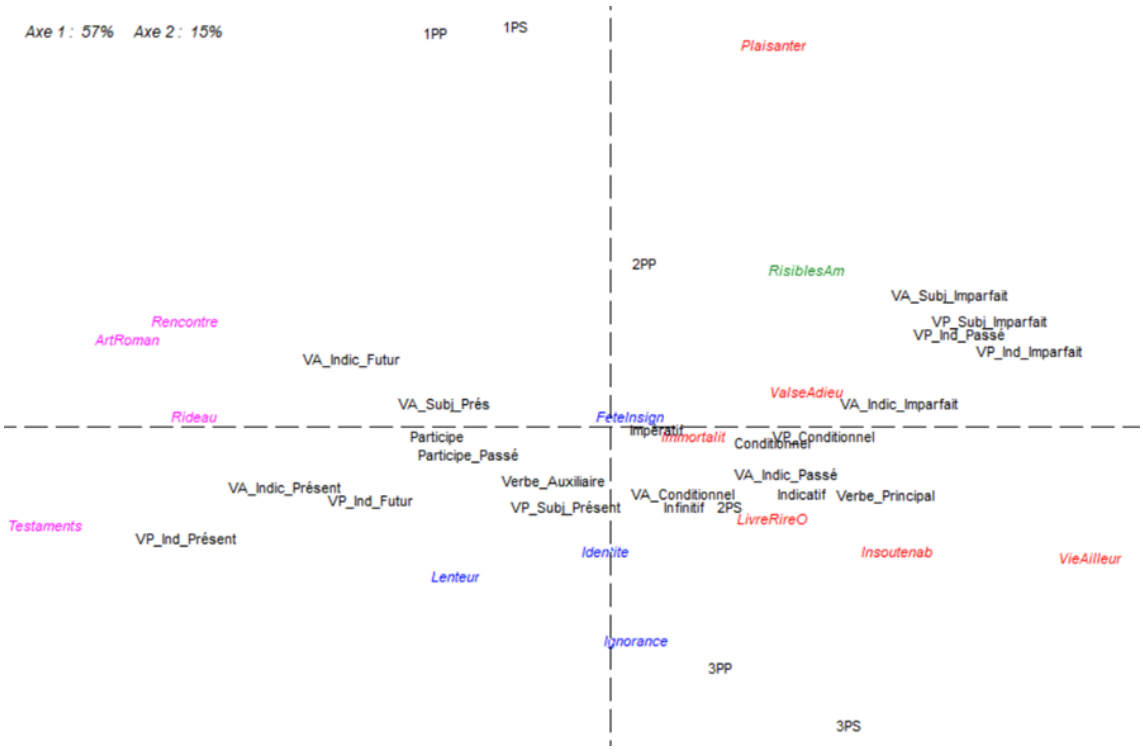

Fig. 4 – Les modes et les temps verbaux dans l'œuvre de M. Kundera

Si l'on croise les textes de Kundera avec la distribution des temps et des modes verbaux à l'aide d'une AFC, on constate que l'auxiliaire et le participe passé gravitent à proximité des textes du cycle français (**fig. 4)**. L'AFC s'interprète communément en opérant des regroupements manuels entre les points dont la distance est réduite (distance au sens euclidien du terme) ; en effet, les proximités représentent un profil distributionnel commun. Dans la fig. 4, ces points correspondent à la fois aux textes (en couleurs) et aux temps/modes (en noir). En d'autres termes, d'une part, la proximité de deux textes dans le graphique indique qu'ils partagent la même distribution des temps/modes ; d'autre part, si deux temps/modes sont proches, alors ils partagent la même distribution dans le corpus. Dans cette étude, nous avons aussi séparé visuellement les romans français des essais en français (respectivement en bleu et rose) et les romans et nouvelles en tchèques (respectivement rouge et vert). Il se confirme alors l'hypothèse de la variable linguistique : les romans tchèques sont proches des temps du récit – l'imparfait et le passé simple – tandis que les formes composées (le temps du discours) se rapprochent des romans français (Benveniste, 1966).

Par conséquent, il semble que, pour interpréter la distribution des formes composées, les variables linguistiques et temporelles l'emportent sur la variable générique et, plus précisément, que celles-ci sont plus employées dans le cycle français que dans le cycle tchèque. Quelle pourrait en être la cause ? Il se peut que l'évolution de la distribution de cette construction périphrastique chez Kundera ne fasse que suivre l'évolution de la langue française de la littérature contemporaine. Cependant, dans l'œuvre de Kundera, on relève même une présence excédentaire par rapport aux textes littéraires contemporains. Ainsi, deux autres interprétations possibles sont à considérer : en premier lieu, l'influence de la langue native, le tchèque, lorsque Kundera écrit en français ; ensuite, la possible interférence de la figure du traducteur, malgré les révisions postérieures de l'auteur.

En ce qui concerne le premier aspect, une étude de linguistique contrastive entre la langue tchèque et la langue française pourrait être envisagée pour considérer une éventuelle interférence entre les systèmes verbaux, notamment d'une possible influence de la langue source et de son substrat temporel. En effet, il n'existe pas en tchèque de distinction comparable à celle que présente le français entre le passé simple et le passé composé. Toutefois, seule une étude de linguistique contrastive pourrait vérifier le bien fondé de cette hypothèse.

En ce qui concerne le deuxième aspect, c'est-à-dire la figure du traducteur, une étude de traductologie serait nécessaire. Par exemple, une comparaison entre les traductions avant et après la révision de Kundera pourrait être envisagée et, dans une perspective exogène, il serait utile d'effectuer des analyses comparatives entre les traductions françaises des textes tchèques de Kundera et les traductions françaises de textes tchèques d'autres auteurs ayant rédigé leurs œuvres pendant la même période où Kundera écrivait en tchèque. Enfin, une valeur ajoutée serait apportée par une autre analyse comparative entre les textes français de Kundera et des œuvres écrites directement en français par des auteurs tchèques, tels que Věra Linhartová, Václav Jamek et Patrik Ouředník. Il s'agit donc d'une question à approfondir, qui pourrait constituer l'une des suites potentielles de cette étude.

Ces perspectives de recherche concernent l'approfondissement de nos hypothèses de travail. Toutefois, si l'on veut approfondir l'étude de l'usage de la construction périphrastique chez Kundera par rapport aux autres auteurs et ne pas se limiter à l'étude de la fréquence et de la spécificité, les outils textométriques offrent d'autres pistes de recherche.

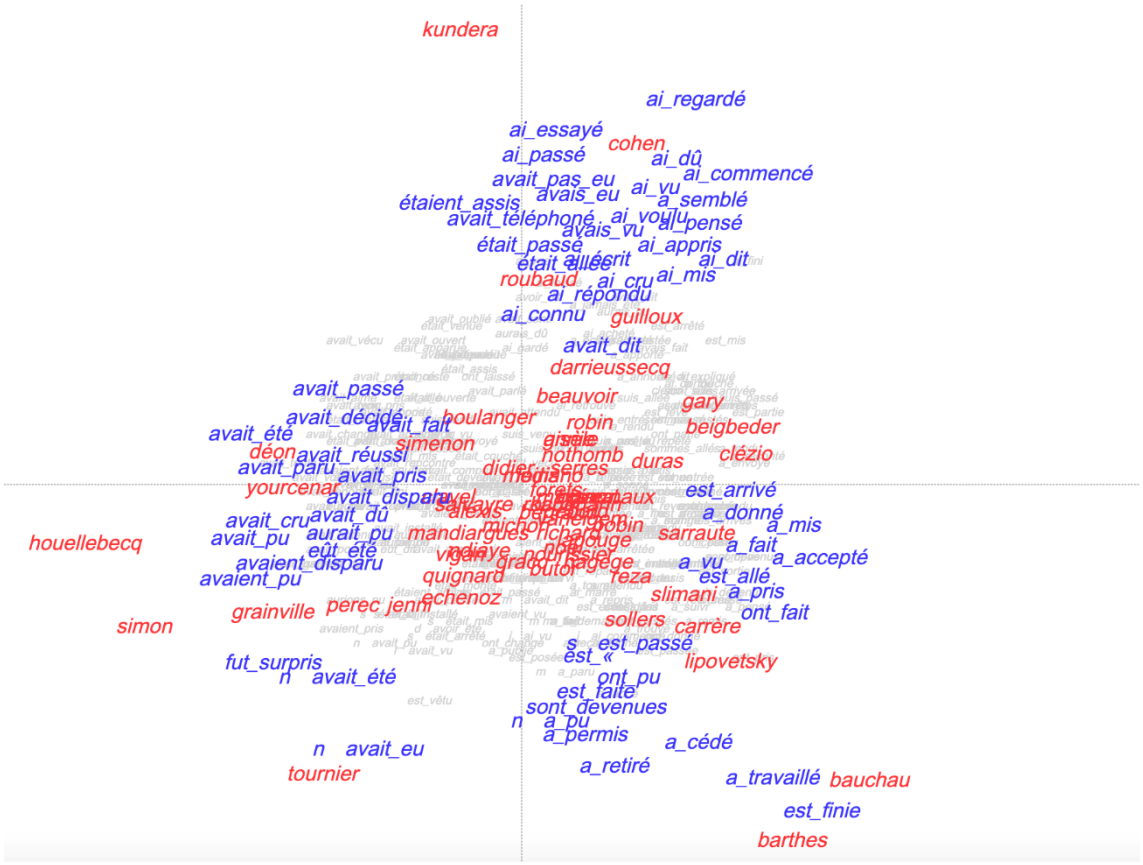

Fig. 5 – AFC des 300 formes composées les plus significatives

Afin d'observer le positionnement de l'œuvre de Kundera par rapport à celle des autres auteurs, outre le calcul de spécificités présenté au début de cette section, qui donne des indications sur la fréquence et la significativité, une autre étude a été réalisée sur la typologie des formes composées. Par le biais d'une AFC, les textes des auteurs du corpus de référence et ceux de Kundera ont été croisés avec leurs 300 constructions périphrastiques les plus significatives (fig. 4). La position de l'œuvre d'un auteur dans le graphique nous donne des indications sur sa particularité quant à l'usage de ces formes composées par rapport à la norme représentée par les autres textes : si une œuvre est en position centrale, cela signifie que cet usage ne s'écarte pas du modèle de référence ; les positions éloignées du centre, en revanche, montrent un usage qui se distingue de la norme. L'œuvre de Kundera relève de ce dernier cas puisqu'elle se situe dans une position fortement excentrée.

Grâce à cette autre analyse quantitative, l'étude qualitative plus approfondie consisterait à dresser une liste des auxiliaires et des participes les plus significatifs chez Kundera et à l'examiner en tenant compte des modes et des temps verbaux ainsi que du contenu sémantique. Quant au dernier point, il serait envisageable de décomposer les sèmes de chacun d'entre eux selon la sémantique interprétative de Rastier et de les classer dans différentes classes sémantiques. À partir de ces données, il serait possible de mener une étude qualitative sur les traits caractéristiques des formes composées les plus utilisées chez Kundera, du point de vue tant grammatical que lexical et sémantique.

5. CONCLUSION

Cet article se situe à la lisière des champs quantitatif et qualitatif, afin d'illustrer et d'interpréter les fonctionnalités d'un logiciel d'analyse statistique de données textuelles (Hyperbase) du point de vue des besoins de la linguistique. Cette opération a intéressé le cadre d'une étude de cas spécifique, celui d'une catégorie grammaticale depuis toujours considérée comme difficile à

définir pour la recherche textométrique, en raison de difficultés de désambiguïsation. Pour ce faire, nous avons proposé des solutions de recherche visant à limiter ces ambigüités (2.2) et des exemples d'application de son analyse (3 ; 4).

En nous appuyant sur une base littéraire de 90 millions d'occurrences, nous avons également présenté les différentes pistes de recherche quantitatives et qualitatives concernant l'étude du participe passé, allant de l'analyse d'un grand empan chronologique à la sélection de périodes plus courtes grâce à la souplesse d'utilisation offerte par le système des métadonnées.

En particulier, en ce qui concerne l'étude de l'œuvre littéraire d'un seul auteur contemporain (4), Milan Kundera, on a pu suivre différents chemins de recherche. On a présenté des analyses comparatives entre les sous-corpus de Kundera et le modèle de référence et entre leurs sous-sous-corpus respectifs, regroupés selon les variables génériques, diachroniques et linguistiques. Ensuite, nous avons présenté d'autres démarches pour l'observation des éléments grammaticaux caractéristiques des formes composées chez Kundera par rapport à la norme linguistique du corpus de référence (AFC, étude croisée des auteurs avec les formes composées les plus significatives). Chacune de ces analyses pourrait aussi être appliquée dans une perspective endogène : les mêmes études comparatives pourraient être effectuées au sein du seul sous-corpus kundérien, plus précisément entre les sous-corpus de l'œuvre de Kundera ou entre un texte de Kundera et le reste de sa production littéraire. Par exemple, on pourrait comparer les spécificités d'un roman de Kundera avec le reste de son œuvre ou seulement avec ses autres textes appartenant au genre romanesque, ou comparer le sous-corpus des romans avec celui des essais, ou le sous-corpus des traductions avec celui des textes écrits directement en français, et ainsi de suite.

BIBLIOGRAPHIE

BENVENISTE É. (1966-1974), *Problèmes de linguistique générale*, I – II, Gallimard.

BENZÉCRI J.-P. (1973), *L'Analyse des données.* T. 2 : *L'analyse des correspondances.* Dunod.

BRUNET É. (1988), *Le vocabulaire de Victor Hugo*, Slatkine-Champion, Paris-Genève.

BRUNET É. (1999), « Qui lemmatise dilemme attise », *11e Rencontres linguistiques en pays rhénan*, L. Kosé, A. Theissen, Strasbourg, France. p.7-32. ⟨hal-01575442⟩

ENGWALL G. (1962-1968), *Vocabulaire du roman français*, (1962–1968). Dictionnaire des fréquences, (Data linguistica n°17, Almqvist & Wiksell International, Stockholm, 1984, LXVIII, 427 p. + 43 microfiches, chacune de 207 p.)

IMBS P. (1960), *L'Emploi des temps verbaux en français moderne, essai de grammaire descriptive*, Klincksieck, Paris.

KUNDERA M. (2011), *Œuvre I et II*, Paris, Gallimard, Coll. Bibliothèque de la Pléiade, 2017.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, octobre 1980, p. 127-165.

LEBART L., PINCEMIN B., POUDAT C. (2019), *Analyse des données textuelles*, Québec, Presses de l'Université du Québec. ⟨hal-02416659⟩

MAGRI V. (2010), « Stylistique et statistiques. Le corpus textuel et Hyperbase », *Stylistiques ?*, PUR, p. 377-393, hal-01226831.

PINCEMIN B. (2018), « Sept logiciels de textométrie », Archives ouverts HAL, CNRS/CCSD, Villeurbanne, https://halshs.archives-ouvertes.fr/halshs01843695, consulté le 26 avril 2020.

RASTIER FR. (2011), *La mesure et le grain. Sémantique de corpus*, Paris, Honoré Champion, Collection Lettres numériques.

RIEGEL, M., PELLAT J.-C., RIOUL R. (1994), *Grammaire méthodique du français*, Paris, Presses Universitaires de France, 2011.