Sine nomine uulgus : étude des profils combinatoires des noms de la foule à partir d'un corpus arboré latin

Louis Autin[1], Kamel Bouzidi[2], Olivier Kraif[2], Julie Sorba[2]

[1] Univ. Grenoble Alpes, Litt&Arts – *Translatio* (UMR 7355) & Universität Osnabrück

[2] Univ. Grenoble Alpes, LIDILEM, F-38040 Grenoble – France

Abstract

The *Phraseotext* Project focuses on the comparison of phraseological phenomena through different literary subgenres. In this context, this paper presents the results of a preliminary experiment conducted on a Latin corpus. First, we built a dependency-parsed corpus and develop a methodology to study combinatorial profiles using textometric tools. This methodology was then applied to comparison of the profiles of three Latin semantically related "collective nouns" : *uulgus*, *turba* and *multitudo*. The detailed review of their most strongly associated collocates, as well as the frequent colligations related to these collocations, can reveal, in a contrastive way, both their semantic nuances and the peculiarities of their contexts of use

Résumé

Cet article présente les résultats d'une étude préliminaire s'inscrivant dans le cadre du projet *Phraseotext*, qui porte sur la comparaison des phénomènes phraséologiques à travers différents sous-genres littéraires. Une première étape a été consacrée à la constitution d'un corpus latin annoté en dépendances, et au développement d'une méthodologie d'observation des profils combinatoires s'appuyant sur des outils textométriques. Nous avons ensuite mis en œuvre cette approche de linguistique outillée pour comparer les profils combinatoires de trois noms collectifs latins sémantiquement proches : *uulgus*, *turba* et *multitudo*. L'étude fine de leurs cooccurrents privilégiés, mais aussi des relations fonctionnelles fréquemment associées à ces collocations, permet de révéler, de manière contrastive, à la fois leurs nuances sémantiques et les spécificités de leurs contextes d'usage.

Key words : Latin treebank, phraseology, contrastive analysis.

1. Introduction

Dans le cadre des humanités numériques et du *text mining*, notre contribution présente le travail d'annotation d'un corpus latin en vue de son intégration dans le *Lexicoscope*[1] – un outil permettant d'étudier les profils combinatoires des unités lexicales (Kraif & Diwersy 2014) –, ainsi qu'une étude pilote fondée sur les extractions obtenues. Son objectif est de montrer les apports du *Lexicoscope* à l'étude du lexique latin en s'intéressant ici aux noms de la foule qui relèvent de la catégorie des noms collectifs (Ncoll), même si leur classement à l'intérieur de cette catégorie est discuté ('Ncoll non décomposable' Jespersen 1924 ; '*ordinary group nouns'* Copestake 1995 ; 'Ncoll catégorisateur' Lecolle 1998 ; Ncoll ne constituant pas une sous-classe grammaticale distincte pour Flaux 1999 ; 'Ncoll de regroupement spatial' Lammert 2010). D'un point de vue cognitif, le collectif est une entité complexe formée d'unités plus simples et les noms qui véhiculent cette notion ont pour rôle sémantique spécifique « d'exprimer la pluralité par le biais de la singularité » (Lammert 2010, 13).

[1] Le *Lexicoscope* est librement consultable à l'adresse suivante : http://phraseotext.u-grenoble3.fr/lexicoscope.

Nous présenterons, dans un premier temps, les modalités de constitution du corpus arboré (section 2) en décrivant les opérations d'étiquetage morphosyntaxique et de lemmatisation, d'analyse en dépendances et d'intégration au *Lexicoscope*. Puis nous proposerons une étude pilote sur trois noms latins de la foule *multitudo*, *turba* et *uulgus* (section 3) afin d'établir le profil combinatoire de chacune de ces unités lexicales et de les comparer.

2. Constitution du corpus arboré

Le corpus choisi pour cette étude est composé de textes littéraires qui ont en commun leur rédaction en prose et leur dimension publique ou politique (les rhéteurs Cicéron, Sénèque le Père et Sénèque et les historiens Salluste, César, Tite-Live et Tacite). La première étape de ce travail a donc consisté à analyser ces corpus en dépendances, afin de les intégrer dans notre interface d'exploration, le *Lexicoscope* (section 2.3). Pour ce faire, nous avons entraîné un parseur stochastique, *MaltParser* (Nivre et al., 2006) sur plusieurs corpus arborés disponibles en ligne[2] : *Perseus Project Latin Dependency Treebank* (Bamman & Crane, 2011) et *Index Thomisticus Treebank*[3] (McGillivray et al., 2009), que nous abrégerons désormais respectivement par *LDT* et *IT*. Ces deux corpus ont des tailles respectives d'environ 53 000 et 244 000 mots.

2.1 Étiquetage morphosyntaxique et lemmatisation

Pour l'étiquetage morphosyntaxique, nous avons utilisé le fichier de paramètres de *Treetagger* développé par Brandolini à partir des corpus *Proiel*, *IT* et *LDT*[4]. Les corpus *LDT* et *IT* ont donc subi une première modification afin de mettre les étiquettes catégorielles (parties du discours et traits) en conformité avec le jeu d'étiquettes utilisé par *Treetagger* (nous noterons ces nouvelles versions *LDT'* et *IT'*). Comme *Treetagger* ne parvient pas à lemmatiser environ 60 % des formes, nous avons ajouté sur nos corpus une phase de lemmatisation en post-traitement : nous nous sommes appuyés sur les formes fléchies extraites des trois corpus arborés déjà mentionnés : *LDT*, *IT* et *Proiel* (Haug et al., 2009). Pour les formes fléchies inconnues, nous avons réduit les mots à leur racine (par troncation des lettres finales formant la désinence) pour rechercher un lemme probable. Enfin, nous avons traité les abréviations et segmenté certains mots agglutinés à la conjonction copulative *-que*.

[2] Nous n'avons pas utilisé les textes lemmatisés par le LASLA (Laboratoire d'Analyse Statistique des Langues Anciennes, Université de Liège), qui propose un étiquetage morphologique de grande qualité car entièrement vérifié par un philologue, car à la différence du *Lexicoscope*, son but n'est pas d'analyser les relations syntaxiques entre différents constituants (l'étiquetage se réduit sur ce point à l'architecture générale de la phrase entre propositions principales et subordonnées). C'est pourquoi nous avons choisi les corpus sus-cités, disponibles en *open source* et pré-analysés en dépendances.

[3] Bien que tardif (XIII^{e} s.), le latin scolastique de Thomas d'Aquin permet d'entrainer le parseur pour notre corpus de latin classique (I^{er} s. av. n.e.-I^{er} s.) car la syntaxe y est tout à fait similaire : « Le latin scolastique restait une langue vivante, ou plutôt, car […] il demeurait frappé par son isolement scolaire et sa spécialisation technique, il était une langue en survie, moyen véhiculaire d'une pensée réutilisant une langue ancienne, sous la pression de besoins qui en refont un parler courant, sinon populaire. C'était la langue vivante de l'Université. » (Chenu, 1993, 98).

[4] Cf. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Latin-parameter-file-readme.

2.2 Analyse en dépendances

À partir des nouveaux corpus *LDT'* et *IT'*, nous avons constitué un troisième corpus nommé *MDT'* (*Merged Dependency Treebank*). Bien que les principes d'annotation et les jeux d'étiquettes de ces deux corpus soient très similaires, il a fallu harmoniser les annotations en effectuant les opérations suivantes : (1) conversion des étiquettes en majuscule ; (2) remplacement de SB par SBJ (relation sujet) ; (3) simplification des étiquettes composées de LDT pour les cas d'ellipses combinées à l'apposition et la coordination (par ex. APOS_ExD0_PRED_CO ou NOM_ExD5_PRED_CO sont devenus : EXD et EXD_CO). Nous avons ainsi entraîné trois modèles avec *MaltParser* : M_{LDT}, M_{IT} et M_{MDT}. Ces modèles ont été optimisés sur les corpus d'entraînement grâce à l'outil *MaltOptimizer* (cf. http://nil.fdi.ucm.es/maltoptimizer/). Afin d'évaluer quantitativement les résultats de ces différents modèles, nous avons prélevé aléatoirement 10 % des phrases de chaque corpus (corpus TEST) et entraîné des modèles d'évaluation sur les 90 % restants. Les résultats sur le TEST sont donnés par l'outil *MaltEval* (Nilsson & Nivre, 2008), grâce à 3 métriques standards :

- LAS (*labelled attachment score*) : indique le pourcentage de *tokens* correctement rattachés à leur tête avec une étiquette de relations correcte ;
- UAS (*unlabelled attachment score*) : indique le pourcentage de *tokens* correctement rattachés à leur tête sans tenir compte de l'étiquette ;
- LA (*label accuracy*) : indique le pourcentage de *tokens* avec une étiquette de relation correcte (sans tenir compte de la tête).

Métrique	LAS	UAS	LA
M_{IT}	0.731	0.787	0.849
M_{LDT}	0.391	0.472	0.648
M_{MDT}	0.666	0.728	0.813

Tableau 1 : Résultat de l'évaluation quantitative pour les 3 modèles de MaltParser

Les mauvais résultats pour le corpus *LDT* s'expliquent par l'exiguïté du corpus. Pour le corpus fusionné, on pouvait s'attendre à une amélioration, mais le système d'annotation plus complexe et plus fin du projet *Perseus* n'est peut-être pas adapté à un système stochastique tel que *Malt*. Par ailleurs, la simplification des relations effectuée lors de l'harmonisation des annotations n'est peut-être pas suffisante pour rendre les deux corpus cohérents. À ce stade de notre recherche, nous nous contenterons donc d'utiliser le modèle M_{IT} entraîné sur l'*Index Thomisticus* seul, qui obtient un score d'attachement suffisant, proche de 80 %.

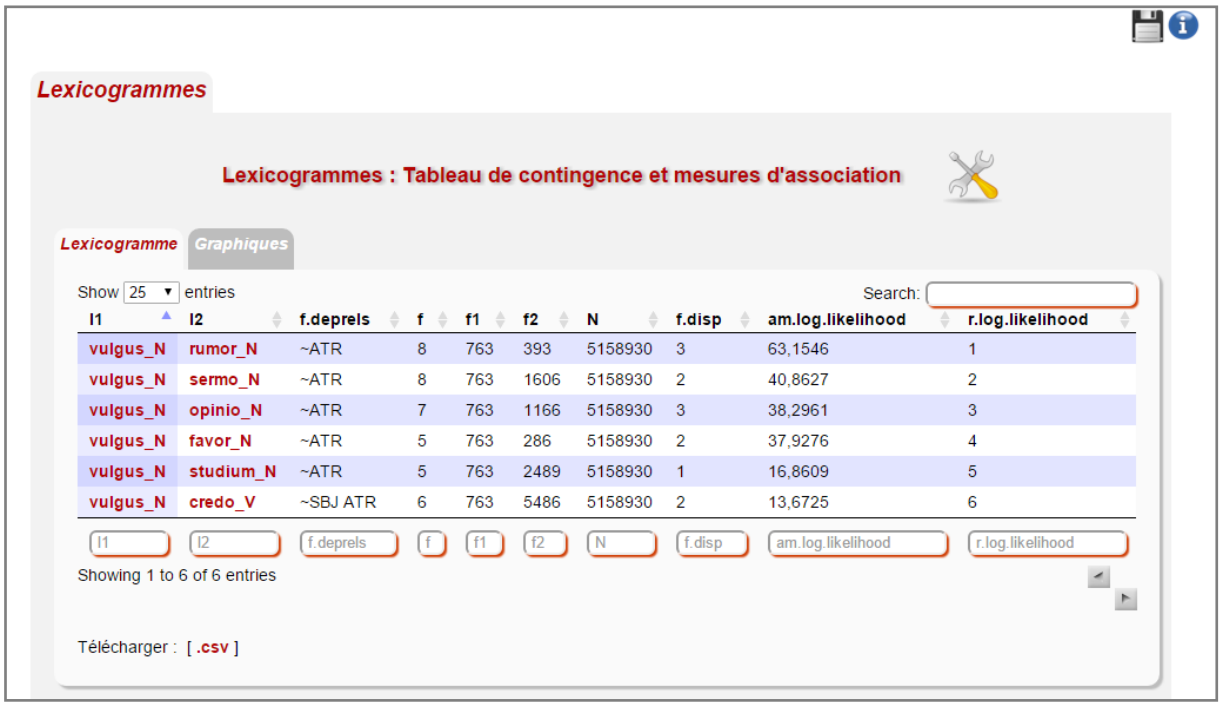
Par ailleurs, l'évaluation montre que certaines relations ne sont pas correctement acquises par le modèle, comme, par exemple, les étiquettes ATV et AtvV, indiquant une relation d'apposition à un nom (ATV) ou à un pronom sujet sous-entendu (AtvV) ayant une fonction adverbiale, ou encore l'étiquette OCOMP, utilisée pour les attributs du COD (Bamman et al., 2007, 21-25). Ces différentes relations sont en réalité assez subtiles et peu pertinentes dans le cadre de l'outil développé ; c'est pourquoi, nous avons décidé, afin d'améliorer les résultats, de procéder à la simplification suivante : ATV, AtvV et OCOMP ont été transformés en ATR, étiquette bien plus courante caractérisant une relation de qualification « standard » d'un terme sur un autre. À l'issue de la chaîne de traitement (étiquetage, lemmatisation et *parsing*), on obtient le corpus arboré suivant :

Auteur	Mots	Textes
César	98 611	2
Cicéron	1 594 938	30
Tacite	193 169	5
Tite-Live	789 059	13
Salluste	38 670	2
Sénèque	409 980	24
Sénèque le père	121 806	4
Total	3 246 233	80

Tableau 2 : Constitution du corpus arboré

2.3. Intégration au Lexicoscope

Le *Lexicoscope* permet d'extraire, pour un pivot donné, l'ensemble de ses cooccurrents syntaxiques les plus significatifs ainsi que les relations syntaxiques mises en jeu (cf. schéma 1).



l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likelihood
vulgus_N	rumor_N	~ATR	8	763	393	5158930	3	63,1546	1
vulgus_N	sermo_N	~ATR	8	763	1606	5158930	2	40,8627	2
vulgus_N	opinio_N	~ATR	7	763	1166	5158930	3	38,2961	3
vulgus_N	favor_N	~ATR	5	763	286	5158930	2	37,9276	4
vulgus_N	studium_N	~ATR	5	763	2489	5158930	1	16,8609	5
vulgus_N	credo_V	~SBJ ATR	6	763	5486	5158930	2	13,6725	6

Schéma 1 : Lexicogramme du Ncoll uulgus

Un tel tableau permet de faire figurer les cooccurrents syntaxiques du pivot de la requête triés par ordre de pertinence décroissante ; la pertinence est quantifiée par la mesure d'association nommée *log likelihood* (ou *rapport de vraisemblance*). Analogue à un *Chi2*, celle-ci exprime l'invraisemblance d'obtenir, par le seul jeu du hasard, le tableau de contingence (résumé par la fréquence de cooccurrence *f*, la fréquence du pivot *f1*, la fréquence du collocatif *f2*, et la taille de l'espace de cooccurrence *N*). Des valeurs seuils, modifiables par l'utilisateur, permettent de ne filtrer que les cooccurrences dépassant une certaine fréquence ou un certain degré d'association (d'autres mesures classiques, comme l'information mutuelle spécifique ou le t-score sont également utilisables). D'autres informations viennent compléter ce tableau,

telles que les relations qui entrent en jeu (p.ex. ATR) ou la dispersion de la cooccurrence, exprimant le nombre de sous-corpus où celle-ci apparait.

Le modèle de cooccurrence ne fait intervenir que les unités reliées par des relations de dépendance – il ne s'agit donc pas de considérer les unités voisines en surface, par exemple en recherchant les mots précédents ou suivants dans une fenêtre de largeur fixe. Comme le note Evert (2009, 1223), ce type de cooccurrence est intéressant en terme de bruit et de silence, car « à la différence des cooccurrences de surface, on ne fixe pas de limite arbitraire à l'éloignement des mots, tout en introduisant moins de bruit que dans les cooccurrences textuelles »[5].

Il est possible de rechercher les cooccurrents et les contextes d'une expression correspondant à un sous-arbre syntaxique (cf. schéma 2).

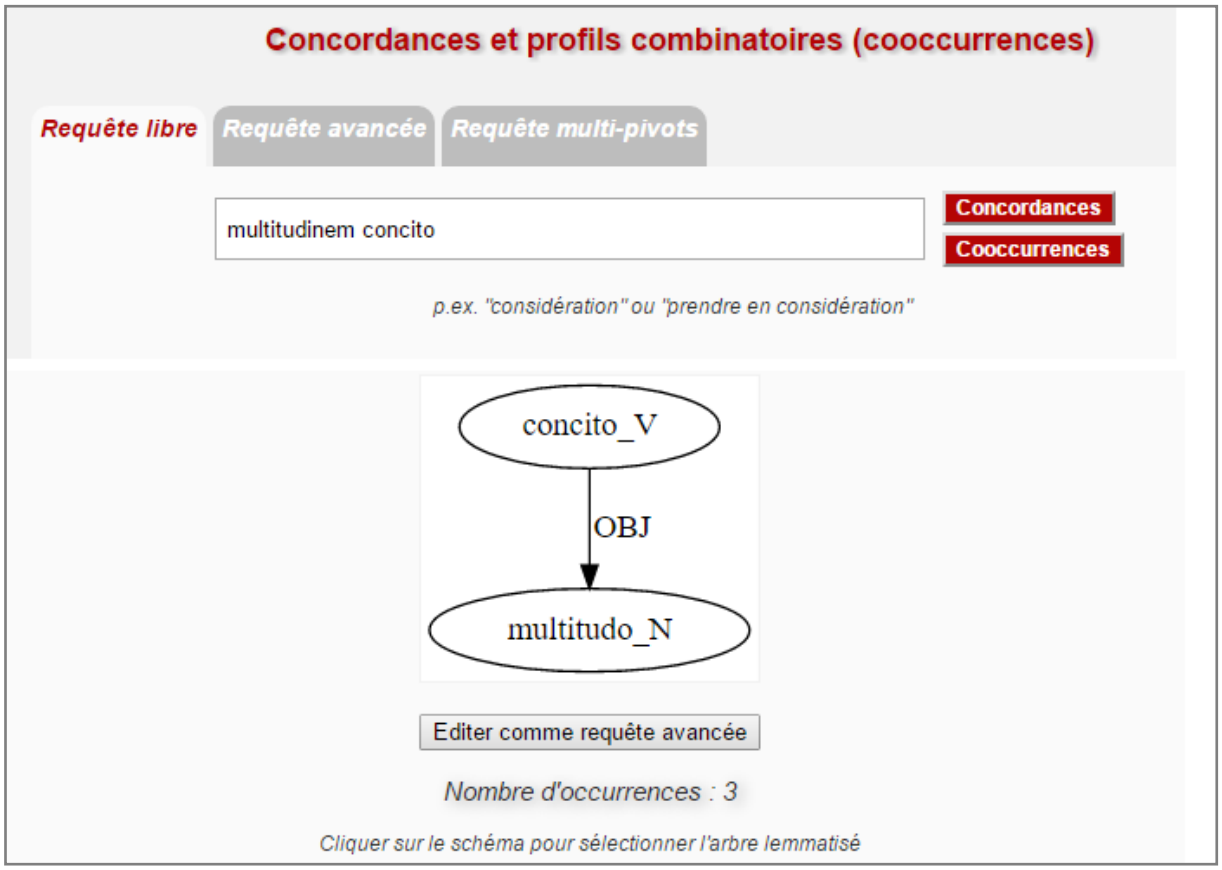


Schéma 2 : Exemple d'arbre fourni pour une recherche à contrainte syntaxique

Vu le fonctionnement du *Lexicoscope*, nous pouvons supposer que les cooccurrents extraits comme les plus fréquents correspondent à des rattachements corrects, du fait de leur forte redondance. Si des formes sont très souvent cooccurrentes, il est peu probable que ces cooccurrences soient fortuites : c'est précisément ce qu'évaluent les mesures d'association employées (rapport de vraisemblance, information mutuelle) qui constituent de fait un moyen efficace de filtrer le bruit dans les rattachements. Quant aux erreurs d'étiquetage des relations, il est possible que certaines erreurs soient suffisamment régulières pour donner des

[5] « [...] Unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time introduces less "noise" than textual cooccurrence. »

statistiques significatives : dans tous les cas, le retour au texte, via les concordances, s'impose pour interpréter les cooccurrences.

3. Étude comparée des noms de la foule *multitudo*, *turba* et *uulgus*

Notre étude pilote repose sur la mise en œuvre de l'une des fonctionnalités du *Lexicoscope* permettant d'analyser les profils combinatoires[6]. Les trois lexies choisies, *uulgus* (n), *turba* (f), *multitudo* (f), présentées comme des synonymes par les outils lexicographiques[7] (« foule, multitude » dans Gaffiot 2001 et Ernout & Meillet 2001 *s.v.* ; *« crowd »* dans Glare 1996 *s.v.*), ont, jusqu'à présent, fait l'objet de très peu d'études linguistiques. Il est vrai que d'un point de vue diachronique, l'enquête s'arrête vite. En effet, *turba* est, selon toute vraisemblance, un emprunt au grec ἡ τύρβη « désordre, tumulte » – lui-même appartenant à une famille de mots expressifs et obscurs (cf. Chantraine 2009 *s.v.*) –, tandis que *uulgus*/*uolgus* est sans correspondant connu dans le domaine indo-européen. Quant à *multitudo*, il s'agit d'un dérivé de *multus* « nombreux » (Ernout & Meillet 2001). Néanmoins, l'intérêt de ces termes anciens et usuels se révèle dans une perspective contrastive et synchronique. Ainsi, on fait l'hypothèse que le sémantisme différent de ces lexies a des répercussions sur la structuration lexicale, syntagmatique et phrastique. En effet, il semblerait que *uulgus* apparaisse plus fréquemment que *multitudo* et *turba* dans des structures à forte valeur idéologique.

L'objectif de cet étude pilote est ainsi de proposer une esquisse de typologie pour ces Ncoll de la foule en latin et de décrire leur traitement dans une perspective contrastive. Pour ce faire, nous appuyons nos analyses sur des approches lexico-statistiques (Blumenthal 2007, 2012, Novakova & Tutin 2009, Kraif & Diwersy 2014). Le fonctionnement de chacune des lexies est décrit au moyen d'un profil combinatoire qui s'appuie sur des analyses syntagmatiques et phrastiques. Après la présentation des données concernant les lexies *multitudo*, *turba* et *uulgus* dans le corpus (section 3.1), nous proposons une typologie de leurs collocatifs (section 3.2) avant d'analyser le fonctionnement des collocations ainsi formées au niveau phrastique (section 3.3). Les éléments issus de ces trois niveaux d'analyse nous permettent de dresser le profil combinatoire de chacune des lexies.

3.1 Présentation des données

Dans notre corpus, les extractions révèlent que le nom de foule le plus utilisé est *multitudo* (980 séquences), devant *turba* (472 séquences) et *uulgus* (289 séquences). La dispersion entre les auteurs du corpus est maximale pour *multitudo* et *turba* ; en revanche, *uulgus* n'apparaît pas chez Sénèque le Père. Les outils du *Lexicoscope* permettent d'étudier la répartition dans le corpus d'un lemme donné (c'est-à-dire de l'ensemble des formes qu'il peut prendre dans sa

[6] Étudier le profil combinatoire d'une lexie, c'est étudier la structure schématique de son voisinage syntaxique et sémantique qui comprend « l'ensemble des accompagnateurs stéréotypés du mot, porteurs d'associations typiques » (Blumenthal, 2007, 19). En effet, dans le cadre de la théorie du *Lexical Priming*, l'emploi d'un mot est lié à des associations sémantiques pré-activées (Hoey, 2005, 13). Nous nommerons ces « accompagnateurs stéréotypés » *collocatifs*, avec lesquels les lexies forment des « associations typiques » ou *collocations*.

[7] Comme le rappelle Moussy (2010, 25), il n'y a guère que dans les dictionnaires, c'est-à-dire hors de leur réalisation dans le discours, que les unités sont synonymes. L'étude des profils combinatoires permet précisément de révéler ces différentes réalisations dans le discours pour distinguer des lexies synonymes (cf. Sorba & Goossens, sous presse).

flexion[8]), mais aussi de sélectionner des formes fléchies spécifiques. En effet, grâce à la recherche avancée du *Lexicoscope* (« requête multi-pivots », cf. schéma 3), il est possible de distinguer plusieurs cas homonymes en latin, par exemple le génitif et le datif singulier de la première déclinaison. Ainsi, là où une simple recherche lexicale du terme *turbae* confondrait plusieurs formes homographes (génitif et datif singulier, nominatif et vocatif du pluriel), l'outil autorise, grâce à la liste déroulante « traits » (cf. schéma 3), une recherche plus fine selon le cas – et conséquemment la fonction – souhaitée :

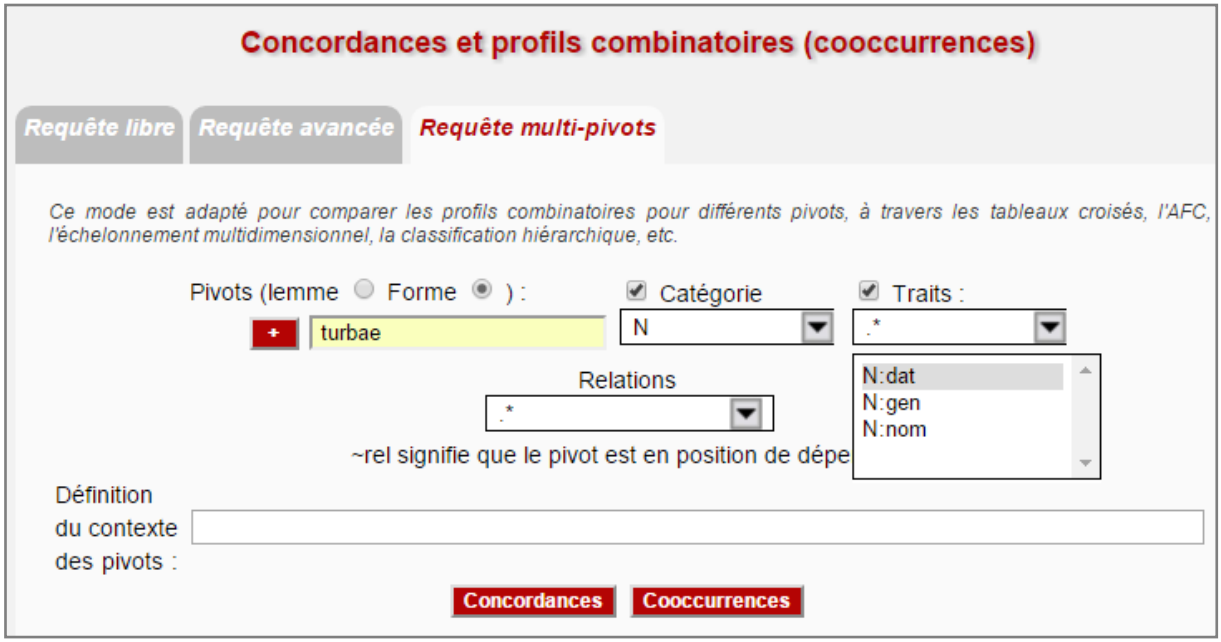


Schéma 3 : Discrimination des formes homonymes par l'outil « requête multi-pivots »

Avec l'option « catégorie » du même volet « multi-pivots », il est également loisible de distinguer deux formes semblables mais de nature différente (par exemple *malo*, verbe « je préfère » et *malo*, datif singulier de *malum*, « le mal »).

En utilisant ces différentes possibilités, nous pouvons observer, en premier lieu, la rareté du pluriel en tant que marque morphologique[9]. Si d'autres langues utilisent ce nombre pour les collectifs humains (voir par ex. le titre français de l'ouvrage de Gustave Le Bon intitulé *Psychologie des foules*), le pluriel morphologique semble presque exceptionnel en latin, qui paraît ainsi considérer la foule sous l'angle de la pluralité interne propre aux Ncoll (Lammert 2010, 64). Un effet de source doit probablement être pris en compte car tous les écrivains de notre corpus sont des aristocrates observant la foule du dessus. Au demeurant, on ne trouve

[8] Dans une langue flexionnelle telle que le latin, la désinence d'un mot est soumise à des variations morphologiques dépendant de la fonction syntaxique occupée par celui-ci dans la phrase. Les cas du latin sont le nominatif (cas du sujet), l'accusatif (cas du COD et de divers circonstants), le génitif (cas du complément du nom), le datif (cas du COI) et l'ablatif (cas de divers circonstants). Comme le soulignent Longrée et Mellet (2012, 2), les marques morphologiques ont aussi un impact sur les aptitudes cooccurrentielles de chaque forme étant donné que « la forme prise par chaque lemme dans un énoncé est la trace de son insertion en contexte ».

[9] Bien que le pluriel morphologique soit quasiment absent, l'usage d'un accord sylleptique avec un verbe au pluriel est attesté pour les trois lexies : *Locros omnis multitudo abeunt* (Tite-Live, *Ab Vrbe condita* 24.3). « Toute la foule se retira à Locres ». Les traductions présentées ici sont celles des auteurs de l'article.

qu'une seule occurrence du pluriel pour *multitudo* (chez Sénèque le Père) et aucune pour *uulgus*. En ce qui concerne *turba*, les apparitions du pluriel sont plus élevées (19 occ.), mais la proportion par rapport au total des séquences demeure faible (4% environ). Ce chiffre s'explique par la polysémie plus grande du terme qui revêt le sens de « troubles » au pluriel.

L'analyse des cas (et subséquemment des fonctions) de ces trois mots pivots aboutit à des conclusions contrastées, comme l'atteste le tableau 3 :

	Nominatif	Accusatif	Génitif	Datif	Ablatif	Total
Multitudo	306	252	125	52	245	**980**
Turba	220	101	31	19	101	**472**
Vulgus	100	63	115	3	8	**289**

Tableau 3 : Profil casuel des trois noms de la foule en nombre d'occurrences

Comme on le voit, deux profils casuels distincts émergent. Pour *multitudo* et *turba*, les trois cas les plus fréquemment utilisés sont le nominatif (respectivement 31% et 46% des occ.), l'accusatif (26% et 23%) et l'ablatif (25% et 21,5%), les deux autres cas obliques, le génitif et le datif, ne dépassant jamais 20% à eux deux. À l'inverse, pour *uulgus*, c'est bien le génitif qui est le plus employé, avec près de 40% des occurrences. Ce terme apparait en effet le plus souvent comme complément du nom de substantifs d'opinion (cf. 3.2) indiquant qu'un jugement particulier appartient à la foule. La fréquence de l'accusatif et de l'ablatif pour les deux autres termes (*multitudo* et *turba*), cas de la majorité des compléments circonstanciels, indique déjà la tendance à les employer pour spatialiser une scène, c'est-à-dire pour inscrire ou pour situer l'action relatée dans un espace donné.

3.2 Typologie des collocatifs

Le profil combinatoire lexico-syntaxique des lexies de l'étude révélé par l'extraction de leurs trois lexicogrammes comporte plusieurs faits saillants. Tout d'abord, les collocatifs apparaissent plus nombreux (15 collocatifs différents) et de nature grammaticale plus variée pour *multitudo* que pour *turba* et *uulgus* (cf. tableau 4).

	Verbe	Nom	Adj	Prép	Total
multitudo	*circumfundo* (répandre à l'entour), *concito* (exciter), *effundo* (répandre au dehors), *conuenio* (se rassembler), *traduco* (faire traverser), *supero* (dépasser)	*telum* (trait), *armatus* (armé)	*tantus* (de cette grandeur), *ingens* (énorme), *magnus* (grand), *imperitus* (ignorant), *ceterus* (tout le reste de), *reliquus* (restant)	*propter* (à cause de)	15
turba	*circumfundo* (répandre à l'entour)	*seruus* (esclave)	*inconditus* (confus), *ingens* (énorme), *alius* (autre)	ø	5
uulgus	*credo* (croire)	*rumor* (rumeur), *sermo* (discussion), *opinio* (opinion), *fauor* (faveur), *studium* (attachement)	ø	ø	6

Tableau 4 : Associations lexicales privilégiées par les 3 lexies

De plus, *multitudo* privilégie une association lexicale avec des termes de forme adjectivale (adjectif ou participe passé passif en fonction d'adjectif épithète, cf. ex. 2 et 3), tandis que *uulgus* s'associe de manière préférentielle avec des noms (cf. ex. 4). *Turba* se distingue des

deux autres par la pauvreté de sa combinatoire lexico-syntaxique (seulement 5 collocatifs statistiquement significatifs apparaissent) :

(1) *Frumenta non solum **tanta multitudine** iumentorum atque hominum consumebantur, sed etiam anni tempore atque imbribus procubuerant*. (César, *Bellum Gallicum* 6.43). « Le blé était épuisé par **un si grand nombre** de chevaux et d'hommes, mais bien plus, les pluies de cette période de l'année le détruisaient. »

(2) *Igitur Germanicus in urbe Artaxata adprobantibus nobilibus **circumfusa multitudine** insigne regium capiti eius imposuit*. (Tacite, *Annales* 2.56). « Ainsi Germanicus, dans la ville d'Artaxate, avec l'approbation des nobles et **devant la foule répandue à l'entour**, ceint la tête [de Zénon] du bandeau royal. »

(3) *Et ludicro circensium, quod adquirendis **uulgi studiis** edebatur, Britannicus in praetexta, Nero triumphali ueste trauecti sunt.* (Tacite, *Annales* 12.41). « Et lors des jeux du cirque que l'on donnait pour capter **l'affection de la foule**, Britannicus défila en toge prétexte, Néron avec un vêtement triomphal. »

Par ailleurs, d'un point de vue sémantique, *turba* et *multitudo* semblent plus proches dans le choix de leurs collocatifs. En effet, les deux lexies s'associent majoritairement avec des termes véhiculant les traits sémantiques /intensité/ (*ingens, magnus, tantus*) et /mouvement/ (*circumfundo, concito, effundo, conuenio, traduco*) :

(4) *Et [uictores] speciosum aduentum suum **ingentem turbam** captiuorum prae se agentes fecerunt.* (Tite-Live, *Ab Vrbe condita* 28.4). « Et les vainqueurs, en poussant devant eux **l'immense foule** des captifs, firent une arrivée clinquante. »

En outre, la combinatoire de *uulgus* se distingue de celle des deux autres par sa préférence pour les noms abstraits (*rumor, sermo, opinio, fauor*). Il s'agit d'une structure récurrente dans laquelle un nom d'opinion et/ou de parole (cf. ex. 6) a pour émetteur *uulgus* dont la position syntaxique privilégiée est celle de complément du nom (cf. 3.1.) :

(5) *Rursum Seianus non iam de matrimonio, sed altius metuens tacita suspicionum, **uulgi rumorem**, ingruentem inuidiam deprecatur.* (Tacite, *Annales* 4.41). « Séjan, derechef, ne parle plus de son mariage, mais ses craintes, qui s'étendaient plus loin, le poussent à détourner les silences qui soupçonnent, **les rumeurs de la foule**, la jalousie insidieuse. »

Enfin, l'étude des collocatifs montre que *uulgus* désigne systématiquement une foule comme une collection d'êtres animés (cf. ex. 4, les spectateurs des jeux du cirque et ex. 6, les gens réunis), tandis que *multitudo* et *turba* peuvent renvoyer à des collections soit d'êtres animés (cf. pour *multitudo* ex. 2, des animaux et des hommes, ex. 3, les gens réunis, et pour *turba* ex. 5, les captifs) soit d'entités inanimées (cf. pour *multitudo* ex. 7, les traits, et pour *turba* ex. 8, les idées susceptibles d'être mémorisées) :

(6) ***Multitudine telorum** ex turribus propugnantes deturbant, aggere et cratibus fossas explent, falcibus vallum ac loricam rescindunt*. (César, *Bellum Gallicum* 7.86) « Ils chassent **par une foule de traits** ceux qui combattaient du haut des tours. »

(7) *Fragile est memoria et **rerum turbae** non sufficit*. (Sénèque, *De Beneficiis* 7.28) « La mémoire est fragile et ne suffit pas à [retenir] **la foule des choses**. »

L'étude des collocatifs de ces trois Ncoll de la foule fournit des éléments pour les distinguer les uns des autres. En effet, tandis que *multitudo* privilégie les collocatifs de type verbal ou adjectival, *uulgus* s'associe préférentiellement avec des noms abstraits de parole ou d'opinion, ce qui lui donne une portée contextuelle plutôt idéologique. De plus, nous avons observé une

nette tendance pour *multitudo*, également marquée pour *turba*, à sélectionner des collocatifs dénotant l'intensité et le mouvement. Enfin, il apparaît que *uulgus* s'est spécialisé dans la désignation de collections d'êtres animés et humains, contrairement à *multitudo* et à *turba* qui sont également utilisés pour nommer des collections d'entités inanimées ou des collections constituées d'éléments hétérogènes (hommes et animaux dans l'ex.2).

3.3 Fonctionnement des collocations au niveau phrastique

Pour ce qui concerne le fonctionnement des collocations de structures N+N et N+A construites autour de ces noms de foule, l'analyse fait émerger des points de convergence, mais aussi des différences dans les fonctions grammaticales occupées par les syntagmes ainsi formés comme l'on peut le voir dans le tableau 5 :

	Sujet	COD	CdN	COI	CC	Total
multitudo	105	49	18	7	78	257
turba	14	7	2	0	6	29
uulgus	9	5	0	2	37	53

Tableau 5 : Fonctions grammaticales des collocations (en nombre d'occ.)

Premièrement, les collocations des lexies *multitudo* et *uulgus* sont régulièrement employées comme complément circonstanciel (CC), et plus spécifiquement comme CC de lieu (cf. ex. 2). On constate ainsi une convergence entre cette fonction majoritaire des collocations (CC) et celle du nom collectif *multitudo*, lui aussi fréquemment employé à l'ablatif (cf. section 3.1). Les collocations ayant *uulgus* pour pivot occupent à 70% la fonction de CC, et plus précisément, pour les deux-tiers d'entre elles, de CC de lieu (locatifs ou directifs, cf. ex. 8) :

(8) *Ignominia tu putas quemquam sapientem moueri posse, qui omnia in se reposuit, qui* ***ab opinionibus uulgi*** *secessit ?* (Sénèque, *Consolatio ad Helviam* 13). « Toi, tu penses que la honte peut émouvoir un sage, lui qui n'a de vie qu'intérieure, lui qui s'est détaché **des opinions de la foule** ? »

Pour les deux autres lexies (*multitudo* et *turba*), l'expression du lieu ne résulte pas de la fonction CC de lieu de leurs collocations, mais du sémantisme des verbes dont elles sont sujets. En effet, ceux-ci sont fréquemment des verbes de mouvement (7/14 occ. pour *turba*, 47/105 occ. pour *multitudo*), indiquant un déplacement de la foule ou sa répartition dans l'espace (cf. ex. 9 et 10) :

(9) ***Magna multitudo*** *sagittariorum ab utraque parte circumfundebatur.* (César, *Bellum Civile* 3.63). « **Une abondante foule** d'archers se répandait à l'entour des deux côtés. »

(10) *Ex propinquo cognoscit <Fuluius> (...) duo milia plaustrorum, inconditam inermemque* ***aliam turbam*** *aduenisse*. (Tite-Live, *Ab Vrbe condita* 25.13). « De près, Fulvius constate que deux mille chariots et **une autre foule** en désordre et sans armes arrivaient. »

Dans ces deux exemples, comme souvent par ailleurs, le sémantisme de mouvement du verbe dont les collocations sont sujets (*magna multitudo* en (9) et *aliam turbam* en (10)), est renforcé par un préverbe précisant la nature du mouvement (respectivement *circum-* et *ad-*).

L'accusatif est le troisième cas apparaissant le plus fréquemment dans notre corpus pour ces collocations (cf. tableau 5). Si l'on s'intéresse au sémantisme des verbes qu'il complète, on s'aperçoit que presque la moitié de ceux-ci actualisent le trait sémantique /contrôle/, parfois avec une polarité neutre, parfois avec une polarité négative, mais le plus souvent conjointement avec le trait /spatialité/, comme le résume le tableau 6 :

Traits	Verbes
/spatialité/, neutre	*capere*, contenir ; *habere*, posséder ; ***re**cipere*, retenir ; ***sus**tinere*, soutenir ; *tenere*, tenir
/spatialité/, négatif	*premere*, réprimer
/spatialité/ + /contrôle/	*agere*, mener ; ***circum**dare*, entourer ; ***co**gere*, réunir ; ***col**ligare*, rassembler ; ***con**cludere*, enclore ; *dare*, placer, *fugare*, mettre en fuite ; *fundere*, chasser ; ***pro**pellere*, pousser en avant ; *rapere*, emporter ; ***re**linquere*, laisser en arrière ; ***re**stituere*, remettre à sa place ; ***tra**ducere*, faire traverser

Tableau 6 : Répartition des traits sémantiques des verbes dont les collocations sont COD

À partir de cette liste exhaustive, trois remarques s'imposent. Premièrement, en latin, la foule est désignée comme une entité sur laquelle l'on peut exercer un contrôle allant jusqu'à la possession (voir le sémantisme des verbes ainsi que la fonction COD de ces collocations). Deuxièmement, on observe une prédominance de la spatialisation (répartition dans un lieu, mouvement, étendue), rappelée ici par le grand nombre de verbes préfixés et la variété de ces préverbes, en gras dans le tableau 6 : ils dénotent un mouvement inclusif (*circum-*), incisif (*pro-*), régressif (*re-*), ascensionnel (*sus-*), ou transgressif (*tra-*), le préfixe le plus fréquent (*cum-*, *co-*[10]) indiquant le rassemblement. Troisièmement, la fréquence des verbes renvoyant au contrôle pour les Ncoll *turba* (4 occ./ 7 collocations ayant comme fonction COD, soit 57%) et *multitudo* (20 occ./ 46, soit 43,5%) est bien plus grande que pour *uulgus* (seulement 1 occ. pour 5). Ces chiffres confortent l'idée selon laquelle l'emploi de *uulgus* se détache de ceux de *multitudo* et de *turba*, ce que mettait déjà en évidence l'analyse des collections de *turba* et de *multitudo* pouvant réunir des entités inanimées ou animées non humaines, à la différence de *uulgus* qui est utilisé pour désigner des collections humaines.

4. Conclusion & perspectives

Notre étude pilote fournit des pistes prometteuses pour explorer de manière originale la combinatoire lexico-syntaxique dans un corpus latin, mais aussi pour améliorer l'interface d'interrogation (prise en compte des variantes orthographiques, amélioration du décompte des occurrences par exemple). En effet, à l'issue de ce travail, nous pouvons constater que la fonctionnalité d'extraction, à partir d'un pivot donné, de ses cooccurrents syntaxiques les plus fréquents est opérationnelle. Nous avons ainsi pu présenter des résultats satisfaisants pour l'étude des trois Ncoll de la foule qui présentent chacun une combinatoire lexico-syntaxique propre et qui confortent notre hypothèse initiale d'un emploi de *uulgus* au sein de structures idéologique forte. Quelques ajustements nécessaires, tel qu'un nettoyage de l'apparat critique[11], qui n'apporte rien dans ce type d'exploration et provoque des erreurs d'analyse de dépendance, permettront d'exploiter de manière plus efficace le *Lexicoscope* et d'étendre ainsi l'étude de la phraséologie dans les textes latins.

[10] Nous n'incluons pas dans cette liste le verbe ***con**citare*, qui a souvent pour COD une collocation formée autour d'un Ncoll de la foule, car celui-ci est exclusivement utilisé par nos auteurs dans son sens figuré (*exciter*, *enflammer*) et non dans son sens propre (*pousser*, *lancer*).

[11] En effet, dans certains textes fournis au parseur du Lexicoscope pour l'analyse en dépendances issus de la librairie *open source* de Perseus (Tite-Live notamment), l'apparat critique était inclus au même niveau que le texte latin (plusieurs variantes d'un même terme étant énumérées successivement). Une solution simple à ce problème consiste à nettoyer ces textes de leur apparat critique, avant de les soumettre une nouvelle fois à l'analyse du logiciel pour obtenir des résultats plus précis.

Références

Bamman D. & Crane G. (2011). The Ancient Greek and Latin Dependency Treebanks. In Sporleder, C., van den Bosch A. and Zervanou K. editors, *Language Technology for Cultural Heritage, ser. Foundations of Human Language Processing and Technology*, Springer.

Bamman D., Passarotti, Crane G. & Raynaud S. (2007). *Guidelines for the Syntactic Annotation of Latin Treebanks* (v.1.3). www.nlp.perseus.tufts.edu.

Blumenthal P. (2007). Sciences de l'homme vs sciences exactes : combinatoire des mots dans la vulgarisation scientifique. *Revue française de linguistique appliquée*, vol.(2) : 15-28.

Blumenthal, P. (2012). Méthodes statistiques en lexicologie contrastive. In Begioni, L. & Bracquenier, C. editors, *Sémantique et lexicologie des langues d'Europe*. Presses Universitaires de Rennes.

Chantraine P. (2009). *Dictionnaire étymologique de la langue grecque : histoire des mots*. Klincksieck.

Chenu M.-D. (1993). *Introduction à l'étude de Saint Thomas d'Aquin*. Librairie Vrin.

Copestake A. (1995). The Representation of Group Denoting Nouns in a Lexical Knowledge Base. In Saint-Dizier, P. and Viegas, E. editors, *Computational Lexical Semantics*. Cambridge University Press.

Ernout A. & Meillet A. (2001). *Dictionnaire étymologique de la langue latine : histoire des mots*. Klincksieck.

Evert, S. (2009). Corpora and collocations. In Lüdeling, A & M. Kytö, M. editors, *Corpus Linguistics. An International Handbook.* Mouton de Gruyter.

Flaux N. (1999). À propos des noms collectifs. *Revue de linguistique romane*, vol.(63) : 471-502.

Gaffiot F. (2001). *Le grand Gaffiot : dictionnaire latin-français*. Hachette.

Glare P.G.W. (1996). *Oxford Latin Dictionary*. Oxford University Press.

Haug D.T.T., Jøhndal M. L., Eckhoff H. M., Hertzenberg M. J. & Müth A. (2009). Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *TAL* 50(2): 17-45.

Hoey M. (2005). *Lexical Priming. A New theory of Words and Language*. Routledge.

Jespersen O. (1924). *La philosophie et la grammaire*. Éditions de Minuit.

Kraif O. & Diwersy S. (2014). Explorating combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions. In Blumenthal, P., Novakova, I. & Siepmann D. editors, *Actes du colloque international Nouvelles perspectives en sémantique lexicale et en organisation du discours.* Peter Lang.

Lammert M. (2010). *Sémantique et cognition. Les noms collectifs*. Librairie Droz.

Lecolle M. (1998). Noms collectifs et méronymie. *Cahiers de grammaire*, vol.(23) : 41-65.

Longrée D. & Mellet S. (2012). Asymétrie de la cooccurrence et contextualisation. Le rôle de la flexion casuelle dans la structuration des réseaux cooccurrentiels d'un mot-pôle en latin. *Corpus* 11. http://corpus.revues.org/2230

McGillivray B., Passarotti M. & Ruffolo P. (2009). The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon. *TAL* 50(2) : 103–127.

Moussy C. (2010). *Synonymie et antonymie en latin*. L'Harmattan.

Nillson J. & Nivre J. (2008) MaltEval: An Evaluation and Visualization Tool for Dependency Parsing. *Proceedings of LREC08*, pages 161-166.

Novakova I. & Tutin A. (2009). *Le Lexique des émotions*. ELLUG.

Sorba J. & Goossens V. (sous presse). Le rôle du figement dans le traitement de la synonymie au sein du champ de la colère. *Linguisticae Investigationes*, 39(1).