Les taggers, auxiliaires heuristiques en ADT ? [1]

Dominique Longrée [1], Sylvie Mellet [2], Céline Poudat [2]

[1] LASLA – Université de Liège – Bâtiment A4 Quai Roosevelt 1B – B 4000 Liège – Belgique

[2] BCL, Université Nice – Sophia Antipolis, CNRS, MSH de Nice – Faculté des Lettres – 98, Bd Edouard Herriot – BP 3209 – 06204 Nice Cedex 3 – France

Résumé

La présente communication a pour point de départ un constat étonnant : le *Catilina* de Salluste, historien latin, s'avère aussi efficacement étiqueté par un tagger (MBT ou TNT) entraîné sur les discours de Cicéron (œuvre ne relevant pas du tout du genre historique) que sur un corpus de textes historiques comprenant une autre œuvre de Salluste lui-même (le *Jugurtha*). Dans la mesure où un étiqueteur entraîné sur un corpus donné obtient généralement de meilleures performances sur un corpus similaire, on est tenté d'en déduire qu'il y a peut-être une forme d'homogénéité, ou du moins des affinités morphosyntaxiques entre les écritures de Salluste et de Cicéron, contrairement à ce que dit la tradition philologique et rhétorique. On cherchera donc à analyser le phénomène en se fondant d'une part sur une comparaison différentielle des sorties du meilleur des étiqueteurs, TnT, et d'autre part sur un ensemble de méthodes ADT (segments répétés, repérage de *motifs*…) en insistant sur la valeur heuristique de ces outils du TAL pour une ADT, et la complémentarité des deux approches.

Abstract

The present paper starts with a surprising result : when tagging the *Catilina* by Salluste, a Latin historian, a tagger trained on Ciceron's works (not belonging to the historical genre) and a tagger trained on historical texts including another work by Salluste (the *Jugurtha*) yield results of comparable quality. This contradiction with the commonly assumed fact that a tagger gives better results when trained on similar data, might lead us to conclude that Salluste and Ciceron's writing styles are close, or at least that they have strong morphosyntactic similarities, which is contrary to what philological and rhetoric tradition say. The phenomena will be analyzed in more detail using a differential comparison of the TnT tagger output on the one hand, and a set of text data analysis (TDA) methods on the other hand (repeated segments, *motifs*…). We present heuristic values for these NLP tools for TDA and show the complementarity of the two approaches.

Keywords : morphosyntactic taggers, repeated segments, motives, latin corpora

1. Un point de départ étonnant

Un entraînement d'étiqueteurs morphosyntaxiques, ou *taggers*, sur les fichiers du Laboratoire d'analyse statistique de langues anciennes (LASLA) de l'Université de Liège a permis un constat étonnant : dans le cadre d'une étude comparative chiffrée des performances des étiqueteurs (Poudat et Longrée, 2009), on a ainsi pu observer que le *Catilina* de Salluste, historien latin, était aussi efficacement étiqueté par un tagger (MBT ou TnT) entraîné sur les discours de Cicéron (œuvre

[1] Cette communication a bénéficié du soutien de l'ANR dans le cadre du projet Textométrie ANR-06-CORP-029 et du soutien de Wallonie-Bruxelles International et du Fonds de la recherche Scientifique, du Ministère Français des Affaires étrangères et européennes, du Ministère de l'Enseignement supérieur et de la Recherche dans le cadre des Partenariats Hubert Curien.

ne relevant pas du tout du genre historique) que sur un corpus de textes historiques comprenant une autre œuvre de Salluste lui-même (le *Jugurtha*).

Il faut bien sûr intégrer la taille du corpus d'entraînement à l'analyse de ce résultat : les discours de Cicéron fournissent un corpus d'entraînement plus important que les textes historiques, ce qui peut en partie expliquer leur meilleure performance. Mais la taille n'explique pas tout : ainsi, on observe qu'elle n'améliore pas véritablement l'étiquetage d'une autre œuvre historique, le livre 3 de la *Guerre des Gaules* de César ; le phénomène n'est donc pas général et semble se localiser principalement sur l'œuvre de Salluste [2].

On est alors tenté d'en déduire qu'il y a peut-être une forme de similarité ou d'homogénéité (affinités morphosyntaxiques) entre les écritures de Salluste et de Cicéron, contrairement à ce que dit la tradition philologique et rhétorique : celle-ci oppose en effet généralement, et ce depuis l'Antiquité même, la *concinnitas* (la « symétrie ») cicéronienne à l'*inconcinnitas* (ou à la *variatio*, la « dissymétrie ») sallustéenne. On s'interrogera dès lors sur la nature des affinités pouvant exister entre les deux auteurs.

On cherchera à analyser le phénomène en s'appuyant sur la démarche suivante. Tout d'abord on s'intéressera aux mécanismes de fonctionnement du tagger ayant présenté les meilleurs résultats, à savoir TnT. On précisera les corpus d'entraînement utilisés et on procédera à une comparaison différentielle entre le texte correctement étiqueté et les résultats de l'étiquetage par TnT suivant ces divers corpus d'entraînement. On observera où sont les erreurs récurrentes et on tentera d'en préciser la nature, en cherchant à savoir s'il s'agit d'erreurs portant sur des étiquettes isolées ou sur des séquences plus complexes assimilables à des n-grammes morphosyntaxiques. On essayera par la même occasion de voir quels facteurs peuvent entrer en jeu dans la production des ces erreurs. Mais tout ceci restera insuffisant pour rendre compte de la proximité globale des textes sallustéens et cicéroniens et des résultats des taggers. On complètera donc l'analyse par l'exploitation des outils spécifiques de l'ADT. Grâce à une nouvelle fonctionnalité d'Hyperbase-Latin, on recherchera ainsi les segments répétés entendus au sens large, c'est-à-dire appliqués non pas seulement au lexique, mais aussi aux codes grammaticaux. La notion de « motifs », déjà présentée aux JADT 2008 (Longrée et al., 2008), sera par ailleurs à nouveau exploitée. En final, on évaluera la possible complémentarité entre les étiqueteurs, outils issus du TAL, et les méthodes propres à l'ADT, en vue de mieux évaluer les distances entre textes.

2. S'interroger sur les mécanismes de fonctionnement du meilleur des Taggers : TnT

Les possibilités d'entraînement ou d'apprentissage que proposent certains étiqueteurs sont de plus en plus exploitées, à l'heure où de grands corpus numérisés sont développés et rendus disponibles. Il est ainsi possible de générer un outil d'annotation à partir d'un corpus manuellement étiqueté, le système d'annotation et la langue de départ étant libres. Cette possibilité intéresse particulièrement la communauté linguistique en permettant d'annoter automatiquement la morphosyntaxe des langues anciennes ou de langues non encore prises en charge par les systèmes existants (voir par exemple Sjobergh, 2003, pour le suédois, ou Heiden et Prévost, 2002, pour l'annotation du français médiéval).

Parmi les étiqueteurs disponibles, TnT est un tagger statistique constitué d'un ensemble de méthodes de lissage (smoothing [3]) et de traitement des mots inconnus, implémenté sur un

[2] Pour le détail chiffré de cette comparaison et, plus globalement, des autres résultats obtenus avec MBT et TnT, voir (Poudat et Longrée, 2009).

[3] Ajustement de l'ensemble des données au modèle, ou à la courbe.

algorithme fondé sur les modèles de Markov (Brants, 2000). L'étiqueteur est un modèle génératif qui obtiendrait de meilleurs résultats sur des corpus d'entraînement de taille restreinte que les modèles discriminants (Clark et al., 2003) et qui a obtenu les meilleurs taux de précision dans plusieurs études (*e.g*. Zavrel et Daelemans, 2000 ; Sjobergh, 2003 ; Poudat et Longrée, 2009). Bien que TnT commette peu d'erreurs, son fonctionnement est probabiliste et donc peu transparent pour l'utilisateur ; contrairement à un tagger comme Brill, qui fonctionne par règles et génère le fichier des règles inférées des corpus, le fonctionnement de TnT ne peut qu'être déduit des erreurs d'étiquetage commises et des trigrammes générés par l'opération d'entraînement.

Si notre précédente étude avait exploité la presque totalité de la base de *prose latine classique* du LASLA (Poudat et Longrée, 2009), nous restreignons notre examen à trois textes-tests étiquetés par TnT entraîné sur trois corpus différents : le 3ème livre de la *Guerre des Gaules* (3 673 tokens), la *Première Catilinaire* de Cicéron (3 333) et la *Conjuration de Catilina* de Salluste (10 688) ont ainsi été soumis à trois étiqueteurs respectivement entraînés sur les corpus suivants : César (*Guerre des Gaules* et *Guerre civile*, 75 277 tokens), l'ensemble des *discours* de Cicéron *(*417 245) et le *Jugurtha* de Salluste (25 602), désormais TnT-César, TnT-Cicéron et TnT-Salluste. Nous disposons naturellement pour chaque texte-test d'une version manuellement annotée qui nous servira de référence ; les différences entre la version automatiquement étiquetée et la version de référence ont été extraites très simplement, au moyen d'un algorithme *diff* standard.

L'examen des différentiels entre le fichier du LASLA et 3 fichiers étiquetés avec TnT, à partir du *Jugurtha*, à partir du corpus césarien et à partir des discours de Cicéron, nous a d'abord permis de mettre en évidence des erreurs récurrentes. On s'est dès lors interrogé sur leur nature : s'agit-il d'erreurs sur des structures ou des codes fréquents chez Salluste et rares dans le corpus césarien ? Ces même structures ou codes se rencontrent-ils plus fréquemment dans les discours Cicéron ? Ou bien y a-t-il quelque chose de plus complexe en n-grammes ?

3. Erreurs liées à la fréquence

La grande finesse de la granularité des étiquettes du LASLA ne rend pas toujours aisée la comparaison des trois différentiels. Il est notamment très difficile de quantifier avec précision les différents types d'erreurs. Pour un même type de forme, les erreurs peuvent en effet porter sur l'ensemble du codage ou seulement sur une partie de celui-ci, comme par exemple l'indice du lemme [4]. Pour une évaluation d'une proximité ou d'une distance entre textes, une erreur de ce dernier type sera à l'évidence moins significative qu'une erreur sur le codage de la partie du discours. Ainsi, par exemple, pour la forme *fluxa* codée correctement C1111_2, soit adjectif (C) de la première classe (1) au nominatif (1) singulier (1) positif (1) pourvu d'un indice de lemme 2, on rencontre trois types d'erreurs : dans l'étiquetage du *Catilina* par TnT-Salluste, la forme est étiquetée comme un ablatif C1611_2, analyse morphologiquement plausible ; pour l'étiquetage par TnT-Cicéron, on trouve le codage B3_1_1423, soit un verbe (B) de la troisième conjugaison (3) au singulier (1) indicatif (1) parfait (4) passif (2) 3ème personne (3), analyse qui pourrait s'expliquer par la nature périphrastique de telles formes de parfait (type *amata est*) ; l'étiquetage par TnT-César propose A161, soit une analyse comme un substantif (A) de la première déclinaison (1) à l'ablatif (6) singulier (1). Comptabiliser avec exactitude chaque type d'erreur devient alors très délicat (il faudrait procéder à une comparaison position par position, travail compliqué par la longueur différente des étiquettes) ; cependant certaines erreurs sont suffisamment caractéristiques pour pouvoir être relevées par une simple lecture des différentiels.

[4] L'indice du lemme est ce qui permet de différencier deux entrées de dictionnaire homonymes.

Ainsi, à diverses reprises, les 1ères et 2èmes personnes présentes dans le texte de Salluste ne sont pas analysées correctement par TnT-César, mais le sont bien avec TnT-Cicéron : par exemple, *considero* est analysé comme un pronom indéfini au lieu de la première personne du verbe *considerare* ; *fecistis* ou *prouideris* sont reconnus comme des substantifs et non pas comme une 2ème personne du pluriel de l'indicatif parfait actif de *facere* ou comme une 2ème personne du singulier du subjonctif parfait actif de *prouidere* (dans ce dernier cas, TnT-Cicéron se trompe également dans l'analyse mais propose une analyse plausible comme futur antérieur, cette forme étant identique à celle du subjonctif parfait), etc. Ces erreurs récurrentes s'expliquent aisément si l'on examine la distribution des personnes verbales dans l'ensemble du corpus proposé par la base Latin du CD-Rom Hyperbase Textes-Latins. Pour ce faire, on aura recours à une simple AFC (Fig. 1).

Figure 1 : AFC des personnes verbales dans l'ensemble de la base

On remarquera que toutes les œuvres historiques (et notamment celle de César – *Guerre des Gaules* et *Guerre Civile* – se situent à gauche de l'axe vertical à proximité de la troisième personne du pluriel (B_2_3), alors que les discours de Cicéron (*1_Discours, 2_Verrines, 3_Discours, 4_Discours, 5_Discours, 6_Discours, Philipp.*) se rassemblent dans le quart inférieur droit avec la 1ère personne du singulier (B_1_1) et la 2ème personne du pluriel (B_2_2). En ce qui concerne Salluste, celui-ci occupe dans le quart inférieur gauche, une position excentrée par rapport aux autres historiens : près de l'axe vertical et fortement décentré sur l'axe 2, il est proche de plusieurs ensembles de discours de Cicéron dans une zone polarisée par la 2ème personne du singulier (B_2_2) ; vu le poids des deux premiers axes, on ne peut nier, sur le plan de l'emploi des personnes verbales, une large proximité entre Salluste et Cicéron.

De la même manière, TnT-César propose régulièrement des étiquetages erronés pour les vocatifs, formes casuelles de l'interpellation : ainsi tous les *patres conscripti* (« pères conscrits », c'est-à-dire « sénateurs ») au vocatif dans le *Catilina* sont-ils tous analysés comme des nominatifs, ce qui n'est pas le cas avec TnT-Cicéron. Un histogramme de distribution des vocatifs (Fig. 2) dans l'œuvre de César, dans les discours de Cicéron et chez Salluste (réalisé avec Hyperbase) montre que le fait peut s'expliquer par la rareté du vocatif chez César (seules 9 occurrences sur l'ensemble de l'œuvre contre 109 chez Salluste) et par son abondance chez Cicéron.

On constate aussi que TnT-César étiquette fréquemment des formes déponentes (de forme passive, mais de sens actif) comme des formes passives : par exemple, *aspernabatur*, du verbe *aspernari*, « mépriser », analysé comme le passif d'un verbe actif **aspernare* inexistant en latin classique. Cela s'explique fort probablement par l'abondance des passifs chez César (voir Fig. 3).

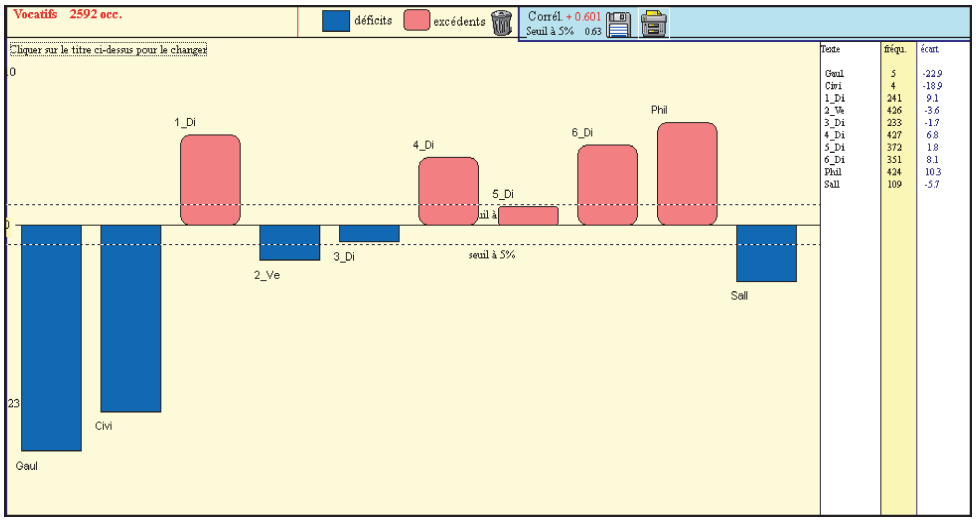

Figure 2 : Distribution des vocatifs chez César, Cicéron et Salluste [5]

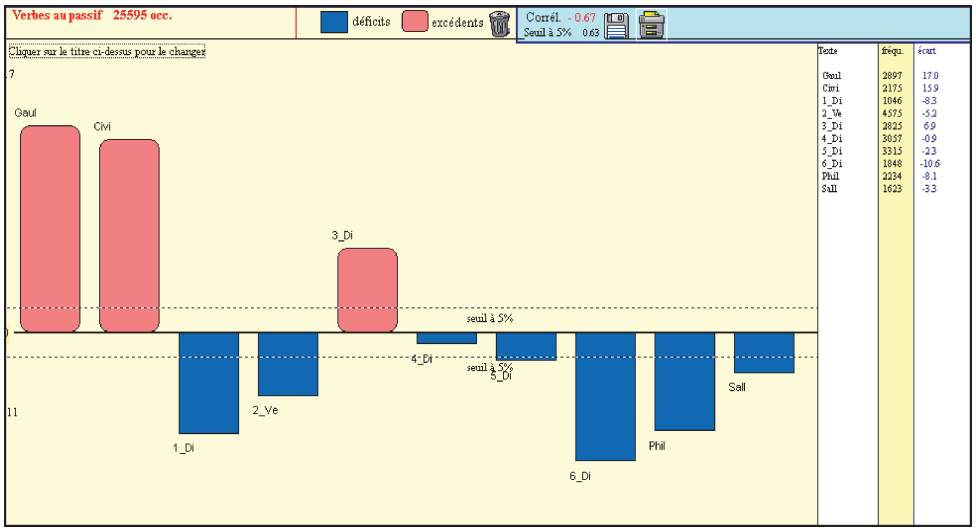

Figure 3 : Distribution des passifs chez César, Cicéron et Salluste

Si certaines erreurs, comme cette dernière, peuvent s'expliquer par des caractéristiques de l'écriture césarienne, d'autres, comme les deux premières, résultent en grande partie de la place des discours directs rapportés, plus fréquents chez Salluste que chez César. Ce goût de Salluste pour le discours direct représente la principale cause de l'utilisation de 1ère et 2ème personnes, de vocatifs et d'impératifs, évidemment présents également dans les discours de Cicéron. En revanche, César préfère le discours indirect dans lequel ces différentes formes soit disparaissent purement et simplement, comme le vocatif, soit sont remplacées par des formes transposées (3ème personnes au lieu des deux premières, subjonctif au lieu de l'impératif).

[5] Sur ce graphique et les suivants, il faut attribuer à César les deux premiers bâtons (Guerre des Gaules et Guerre Civile), à Salluste le dernier et à Cicéron tous les autres (Discours, Verrines et Philippiques)

Dans tous ces cas, c'est la fréquence d'une analyse morphologique dans le corpus d'apprentissage qui explique la présence de certaines erreurs dans Salluste étiqueté par TnT-César et leur absence dans l'étiquetage par TnT-Cicéron.

4. De la fréquence à la séquence

Mais la fréquence d'une forme isolée n'explique pas tout ; un examen détaillé des fichiers différentiels révèle des mécanismes plus complexes qui mettent en jeu la fréquence de séquences grammaticales.

Ainsi on observe de fréquentes erreurs d'analyse sur les adjectifs dans l'étiquetage du texte de Salluste par TnT-César, erreurs que l'on ne retrouve pas avec TnT-Cicéron. De prime abord, cette différence peut surprendre, car les index grammaticaux ne signalent pas d'écarts de fréquence important sur la catégorie de l'adjectif entre les trois auteurs.

Mais un examen plus attentif montre que les erreurs se produisent assez souvent dans des séquences de plusieurs adjectifs successifs, employés en parataxe (voir les deux exemples donnés dans le Tab. 1).

Exemple	*Résultat tagger*	*Analyse correcte*
Exemple 1 ligne 85,86c85,86 < aeterna D332 (numéral	< clara A332 (substantif 3e déclinaison, accusatif, pluriel) > aeterna C1111 (adjectif distributif, accusatif, pluriel)	> clara C1111 (adjectif 1re classe, nominatif singulier, au positif) 1ère classe, nominatif singulier, au positif)
Exemple 2 ligne 6130,6131c6128,6129 (adjectif 2e classe,	< ferox A311 (substantif 3e déclinaison, nominatif, singulier) < uehemens B211_411 (verbe de la 2e conjugaison, nominatif singulier présent actif)	> ferox C5111 nominatif singulier, au positif)> uehemens C5111 (adjectif 2e classe, nominatif singulier, du participe au positif)

Tableau 1 : Deux exemples d'erreurs – séquences de plusieurs adjectifs successifs

Or, les séquences /Adjectif + Adjectif/ sont distribuées de manière irrégulière dans le corpus des trois auteurs (Fig. 4, graphe produit par Hyperbase-Latin).

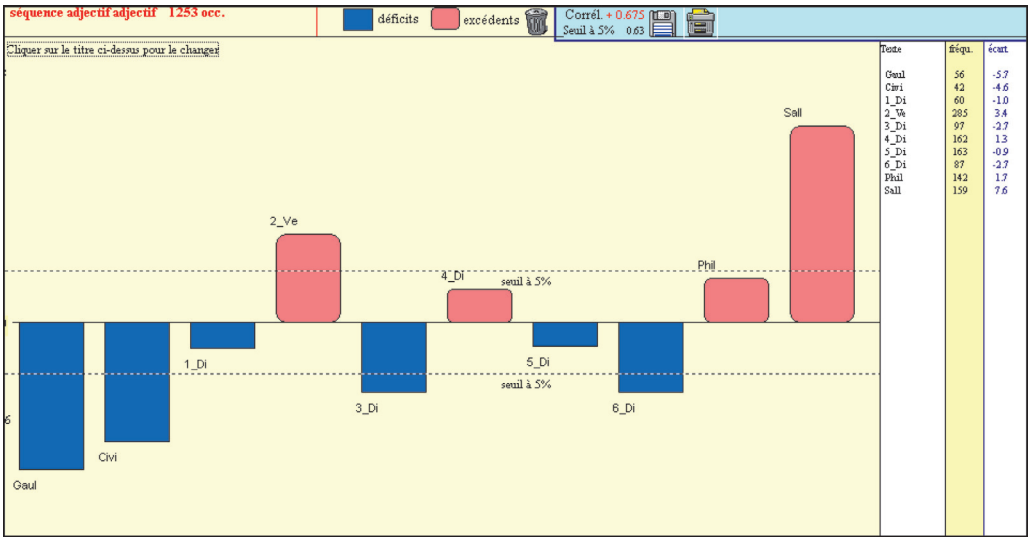

Figure 4 : Distribution des séquences adjectif-adjectif chez César, Cicéron et Salluste

De la même façon, l'analyse des séquences /Nom + Adjectif/ échoue régulièrement dans l'étiquetage de Salluste par TnT-César alors qu'elle n'offre aucune difficulté à TnT-Cicéron. Là encore, ce sont les fréquences respectives de la séquence – fréquente chez Cicéron et Salluste, plus rare chez César – qui expliquent ces résultats (Fig. 5).

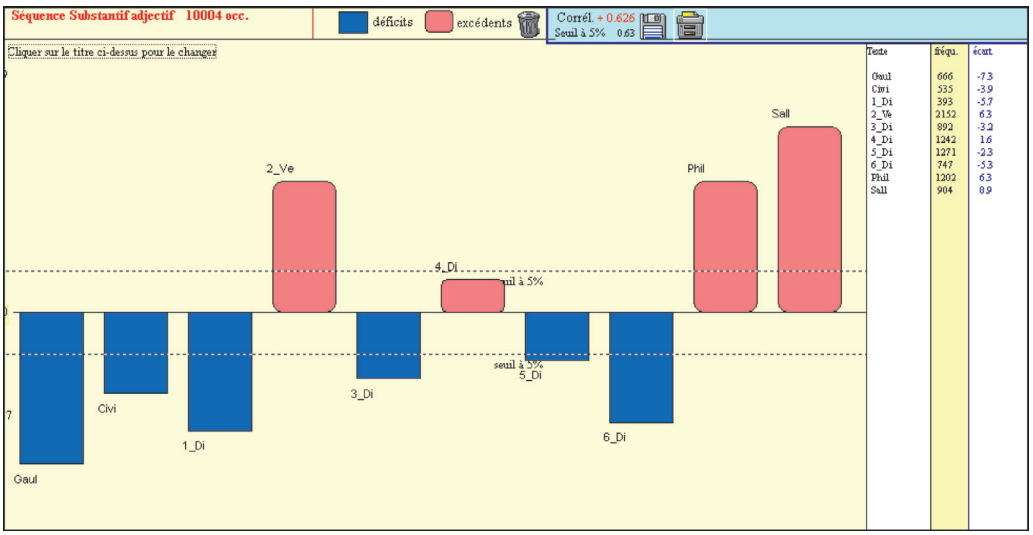

Figure 5 : Distribution des séquences substantif-adjectif chez César, Cicéron, Salluste

Face à ces séquences /Substantif + Adjectif/, TnT-César privilégie une analyse en /Substantif + Substantif/ ou en /Adjectif + Substantif/ car l'une et l'autre sont fréquentes et chez César, et chez Salluste (Tab. 2) :

Exemple	*Résultat tagger*	*Analyse correcte*
Exemple 1 ligne 211,212c211,212	< res A511 (substantif 5e déclinaison, nominatif singulier) < humanae A141 (substantif 1re déclinaison, génitif singulier)	> res A512 (substantif 5e déclinaison, nominatif pluriel) > humanae C1121 (adjectif 1re classe, nominatif pluriel au positif)
Exemple 2 ligne 23,25c23,25	< uentri C4611 (adjectif 2e classe, ablatif singulier au positif) < oboedientia A161 (substantif 1re déclinaison, ablatif singulier)	> uentri A351 (substantif 3e déclinaison, datif singulier) > oboedientia C5321 (adjectif 2e classe accusatif pluriel au positif)

Tableau 2 : Deux autres exemples d'erreurs

Ainsi, au terme d'un examen approfondi des fichiers différentiels, on retiendra que, bien naturellement, la présence et la fréquence des items linguistiques dans le corpus d'apprentissage jouent un rôle important dans les performances du tagger et que, par conséquent, les erreurs et les succès de celui-ci permettent de détecter les formes et les structures fréquemment communes à Salluste et à Cicéron et mal représentées chez César. On voit aussi qu'on ne saurait, pour bien comprendre la possible proximité d'écriture entre les deux premiers, se contenter de prendre en compte la fréquence d'items isolés : il semble tout à fait nécessaire de passer au niveau de la séquence.

Surgit alors le soupçon que les séquences de codes grammaticaux telles que nous les avons étudiées jusqu'ici présentent deux défauts, ou du moins deux limites, pour permettre d'aller encore plus loin dans la détection automatique des séquences communes aux textes et pour

rendre compte globalement de leur proximité : d'une part ces codes sont trop rigides en ce sens que chaque caractère du code fournit une information fixe et pleine et ne permet aucun regroupement de codes proches, qui partageraient quelques informations majeures ; d'autre part, ces codes sont unidimensionnels : ils sont exclusivement morphosyntaxiques et n'autorisent pas la superposition d'informations lexicales à la dimension grammaticale. Ce sont ces deux points que nous allons évoquer dans la dernière partie de notre exposé.

5. Du segment répété au motif

5.1. Nécessité d'introduire des variables libres dans les codes grammaticaux

Nous allons mettre ce point en évidence grâce à la fonction « recherche de segments répétés » de HYPERBASE-Latin, qui peut traiter aussi bien le texte « normalement » enregistré sous formes graphiques que le texte réduit à la succession de ses lemmes ou encore réduit à la succession des étiquettes morphosyntaxiques affectées à chacune de ses formes.

Cette fonction fournit donc les segments répétés spécifiques de chaque auteur étudié, qu'elle classe en fonction de l'écart réduit.

Ainsi, parmi les segments répétés de codes grammaticaux fortement caractéristiques de l'écriture sallustéenne, on relève les séquences :

a211 - a211 - a211
et
a211 - a211 - a311

soit, dans le premier segment, une séquence de trois substantifs de la deuxième déclinaison au nominatif singulier, alors que dans le deuxième segment le dernier substantif, toujours au nominatif singulier, appartient à la troisième déclinaison. Or l'appartenance d'un substantif à tel ou tel type de déclinaison est programmée en langue et la liberté de choix de l'écrivain paraît ici minimale ; le trait stylistiquement pertinent concerne uniquement l'accumulation paratactique de trois substantifs au nominatif singulier, que l'on opposera à la facilité avec laquelle Cicéron accumule, lui, les substantifs à l'accusatif singulier et celle avec laquelle César accumule les ablatifs singuliers : à chacun son cas de prédilection ! En revanche, si l'on considère les adjectifs, on retrouve une affinité entre Salluste et Cicéron, qui les oppose à César – ceci, à condition de pouvoir, là encore, introduire une sorte de caractère joker dans le code et de considérer comme équivalent les adjectifs de première classe et les adjectifs de deuxième classe.

Il s'agirait donc de développer des outils d'exploitation des données textuelles dans lesquels on puisse paramétrer la granularité des informations codées dans l'étiquette morphosyntaxique, quel que soit le format de celle-ci.

Outre la possibilité d'introduire une variable au sein d'un code grammatical, l'analyste latiniste qui observe et traite manuellement les résultats de la recherche de segments répétés est tenté par un autre type de licence : il lui apparaît en effet assez vite que c'est aussi au sein du segment que la possibilité d'introduire une variable pourrait lui permettre d'appréhender des convergences d'écriture entre deux auteurs. Ainsi les spécificités de Salluste et de Cicéron font apparaître respectivement les deux motifs suivants :

m°°°1 - c1111 - c1111
et
m°°°1 - c1111 - s - c1111

soit la succession [adverbe + deux adjectifs de la première classe au positif, nominatif singulier], sans coordonnant entre les deux adjectifs dans le premier segment, avec coordonnant dans le second – séquence qu'on ne retrouve ni sous une forme, ni sous une autre chez César. La parenté des deux structures peut bien sûr se discuter : on pourrait argumenter que la présence ou l'absence de la coordination est fondamentale, l'usage de la parataxe pouvant être considéré comme un trait stylistique majeur. Mais on peut aussi estimer que l'absence de l'un et l'autre segment chez César (et de tout segment comparable, par exemple avec des adjectifs de la deuxième classe) impose par contraste de prendre en considération la parenté de ces deux structures syntaxiques. On touche là aux limites de l'analyse automatique, non seulement parce que, bien évidemment, le tagger (ou tout autre outil d'exploration automatique) ne peut rapprocher les deux structures en leur entier (pour lui, la similitude de structure s'arrête nécessairement après le deuxième terme), mais aussi parce que l'analyste ne peut prendre de décision pertinente sur la valeur de ce rapprochement à la seule lecture des listes de spécificités : il est obligé, pour trancher, de revenir aux textes et de mobiliser sa connaissance philologique des œuvres qui l'occupent.

5.2. Le rôle des motifs, structures récurrentes hétérogènes

Le constat précédent nous conduit à terminer cet exposé en faisant appel à la notion de « motif » que nous avons présentée aux JADT 2008. En effet, le motif, tel que nous l'avons défini, a deux propriétés fondamentales : celle de prévoir en son sein la présence d'une ou plusieurs variables et celle d'associer des éléments linguistiques de nature hétérogène.

- La présence de variables se traduit par la formalisation d'une place vide au sein du motif qui peut être, ou non, instanciée par différents items. Ainsi, dans l'exemple précédent, le motif prévoira la place d'une conjonction de coordination entre les deux adjectifs, qui pourra soit être instanciée par les divers coordonnants latins, soit rester vide.
- La nature hétérogène des éléments constitutifs de tout motif permet de définir un motif particulier à partir d'une combinaison d'éléments lexicaux, morphologiques, syntaxiques, prosodiques, etc. La détection automatique de tels motifs appelle impérativement des outils capables de gérer et d'exploiter les annotations multi-niveaux, par exemple des balises XML. Malheureusement, les outils de ce type ne sont pas encore très performants ni réellement disponibles et, en attendant, on doit de nouveau prendre appui sur les compétences de l'analyste et sur sa capacité à établir des ponts entre certains résultats des relevés automatiques. Cependant, s'il est vrai que l'intervention du philologue reste à ce stade nécessaire, il n'en est pas moins vrai aussi que son regard sur les textes et son activité interprétative peuvent être considérablement réorientés par le traitement machine.

L'exemple suivant va illustrer cette double articulation méthodologique et l'apport qu'on peut en attendre.

La liste des spécificités sur codes grammaticaux chez nos trois auteurs a fait apparaître, entre autres choses, la récurrence de la succession de deux adverbes chez Salluste et chez Cicéron, opposée à un net déficit de cette structure chez César (Fig. 6) :

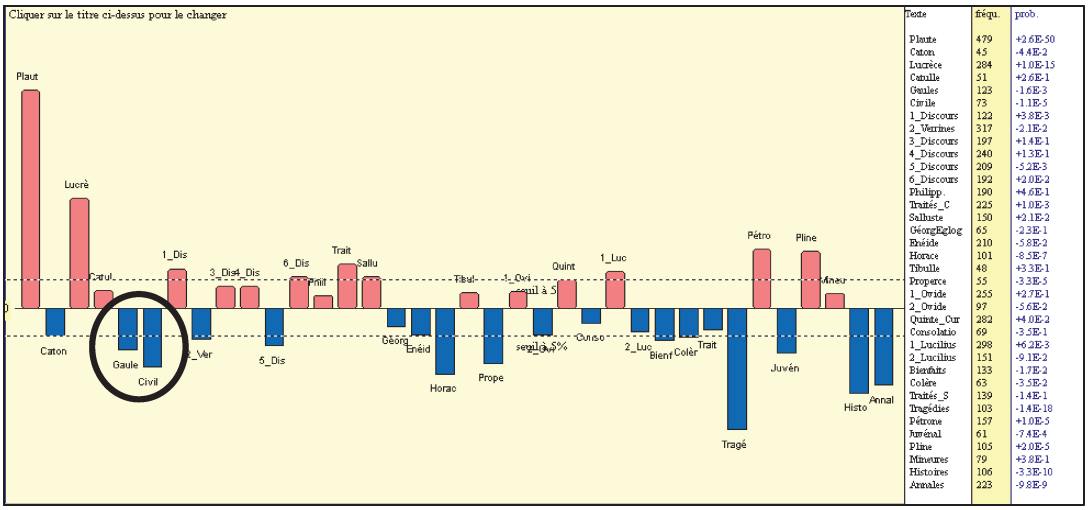

Figure 6 : Distribution de la structure [adverbe + adverbe] dans la base latine

Certes, Plaute et Lucrèce au début de la chronologie, Pétrone et Pline à la fin, ont des écarts supérieurs à ceux de Salluste et de Cicéron, mais ceux-ci restent positifs alors que ceux de César (entourés par l'ellipse) sont nettement négatifs.

Un examen plus attentif de la liste des segments répétés fait apparaître ensuite la récurrence de la structure plus complexe [pronom relatif + adverbe + adverbe] et un retour au contexte dévoile un motif remarquablement stable qui a la forme : [relatif complément d'objet + adverbe *paulo* « peu, un peu » + adverbe anaphorique intradiégétique + verbe (*com*)*memorare* « rappeler », au parfait et à la première personne].

Le Tab. 3 présente les réalisations principales de ce motif :

Salluste	*Cicéron*	*César*
quos paulo ante memoraui (« ... que j'ai rappelés/ mentionnés peu auparavant »)	*quae paulo ante commemoraui* (« ... que j'ai rappelé peu auparavant »)	*quam supra commemoraui* (« ... que j'ai rappelée plus haut »)
de quo paulo ante memoraui (« ... dont j'ai fait mention peu auparavant »)	*quos antea commemoraui* (« ... que j'ai rappelés avant cela »)	*his de causis quas commemoraui* (« ... pour ces raisons que j'ai rappelées »)
quos supra memoraui (« ... que j'ai rappelés ci-dessus »)		

Tableau 3 : Réalisations principales du motif

On tire du relevé d'occurrences exhaustif les observations suivantes :

1) la notion de motif est particulièrement bien adaptée pour rendre compte de cette structure qui se caractérise par l'équilibre entre une certaine stabilité qui assure sa reconnaissance et le jeu de ses variations, et qui sollicite divers types de paramètres, grammaticaux et lexicaux : une structure syntaxique, à savoir la proposition relative, une séquence de deux adverbes ayant des formes variables, y compris la forme Ø, mais un sens très voisin, et la forme verbale simple ou préfixée ;
2) ce motif, en raison justement de ses variantes, échappe totalement au Traitement Automatique des Langues ; en revanche, une Analyse des Données Textuelles correctement outillée et contrôlée par un retour aux textes a des chances de la détecter ;

3) la comparaison de ses différentes formes chez les trois auteurs étudiés met au jour de différents rapprochements possibles : la philologie traditionnelle avait repéré et mis en avant l'originalité de Salluste qui est le seul à employer le verbe sous sa forme non préfixée : *memoro* au lieu de *commemoro*. Cet emploi est marqué et comme il s'agit chez Salluste d'un choix constant pour bon nombre de verbes, certains commentateurs l'ont érigé en stylème. Cet usage idiosyncrasique distingue donc Salluste et l'oppose au couple Cicéron / César qui usent l'un et l'autre, très banalement, du préfixé. Tel est le trait dominant que la philologie classique a retenu.

Mais on peut voir les faits sous un autre jour et remarquer que c'est César qui s'isole en n'employant **jamais** l'adverbe *ante* ou *antea*, là où Cicéron et Salluste, une fois de plus, se rapprochent volontiers : les deux occurrences sallustéennes de *memoraui* en relative précédées de *paulo ante* sont sans doute inspirées de Cicéron.

Le nombre d'occurrences n'est bien sûr pas suffisant pour tirer des conclusions précises sur cet exemple isolé. Il faudrait étudier de manière systématique des motifs apparentés comme *ut supra diximus* (« comme nous l'avons dit plus haut »), *dixi iam antea* (« j'ai déjà dit auparavant »), etc. Mais le but n'est pas ici de produire une étude sur les manies phraséologiques des auteurs latins.

Notre objectif était purement méthodologique.

6. Conclusions

Malgré des visées différentes – applicative d'une part et analytique d'autre part –, les outils du Traitement Automatique des Langues et les méthodes d'ADT se croisent régulièrement, en mobilisant des ensembles de données comparables ou en recourant à des méthodes et des observables similaires. Il en va ainsi des *trigrammes* sur lesquels se fonde TnT pour générer ses étiqueteurs et des *segments répétés* qu'exploite l'ADT pour caractériser et contraster ses corpus d'étude : appliquées à un même corpus, les deux méthodes peuvent être exploitées de manière complémentaire pour caractériser la topologie micro-structurelle du texte.

Nous espérons avoir montré que les sorties des étiqueteurs et les erreurs d'annotation commises d'un texte et d'un corpus à l'autre pouvaient révéler une homogénéité ou une proximité textuelle, que les méthodes de l'ADT permettent ensuite de confirmer et de préciser. Cette valeur heuristique de certains outils du TAL par rapport à l'ADT permet de jeter un pont entre les deux disciplines, d'autant que les résultats obtenus au fil des analyses ADT peuvent ensuite permettre d'améliorer, voire d'optimiser l'application TAL ; dernière étape qui permet ainsi d'évaluer la pertinence applicative des analyses ADT menées.

Références

Brants Th. (2000). TnT – A Statistical Part-of-Speech Tagger. *Proceedings of the 6th Applied NLP Conference*, ANLP-2000, Seattle, WA.

Clark St., Curran J. et Osboren M. (2003). Bootstrapping POS Taggers using Unlabelled Data. In Daelemans W. et Osborne M., editors, *Proceedings of CoNLL-2003*, pp. 49-55.

Heiden S. et Prévost S. (2002). Étiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités. In Pusch, C.D. et Raible, W. editors, *Romance Corpus Linguistics – Corpora and Spoken Language.* Gunter Narr Verlag Tübingen.

Longrée D., Luong X. et Mellet S. (2008). Les motifs : un outil pour la caractérisation topologique des texts. In Heiden S. et Pincemin B., editors, *Actes des 9e JADT*, PUL, vol. 2, pp. 733-744.

Longrée D. et Mellet S. (2009). Syntactical Motifs and Textual Structures. *Belgian Journal of Linguistics*, vol.(23) : 161-173.

Poudat, C. et Longrée, D. (2009). Variations langagières et annotation morphosyntaxique du latin classique. *TAL* vol.(50-2), pp. 129-148.

Sjobergh J. (2003). Combining POS-taggers for improved accuracy on Swedish text. *Proceeding of NoDaLiDa 2003*, Reykjavik, 2003.

Zavrel J. et Daelemans W. (2000). Bootstrapping a tagged corpus through combination of existing heterogeneous taggers. *Conference on Language Resources and Evaluation (LREC 2000)*, pp. 7-20.