Beaudouin V. (2004). "Mètre en règles", *Revue française de linguistique appliquée,* Vol IX, n°1, p. 119-137.

Mètre en règles

Valérie Beaudouin
Laboratoire Usages, créativité, ergonomie
France Télécom R&D

Résumé :
Cet article rend compte d'une expérience d'analyse systématique des aspects métriques et rythmiques d'un corpus de près de 80 000 vers. Des outils d'analyse du vers ont été construits à partir de briques de Traitement Automatique des Langues déjà existantes. Nous défendons une approche expérimentale et cumulative qui consiste à enrichir la description des vers par des traits de nature différente (morphosyntaxe, accent, rime…) qui portent sur des unités de taille différente (position métrique, hémistiche, vers, couple de vers), et qui sont construits avec des outils hétérogènes. Cette intégration dans une seule base de donnée permet de valider des hypothèses classiques sur le vers et de tester de nouvelles hypothèses portant en particulier sur les corrélations entre niveaux d'analyse. Cette approche expérimentale s'appuie sur les savoirs traditionnels sur le vers, mais elle est aussi en mesure de faire évoluer les théories ou hypothèses dominantes en faisant émerger de nouvelles formes de régularités, peu visibles à l'œil nu, et des corrélations inattendues entre phénomènes relevant de plusieurs niveaux linguistiques.

Summary :
Metrics and rythmics aspects are examined on a 80 000 verses corpus, analysed with computational linguistics tools. We propose a cumulative experimental approach consisting in building a verse pattern with a series of features (morpho-syntactic, stress, rhyme…). Features may characterize units of different levels (syllables, hemi-verse, verse, etc.) and are evidenced by different tools, but all are integrated in a single database. Thus we can verify classic metric rules and hypotheses. We also document new regularities, for example on stress patterns, and we test some new hypotheses about links between features and patterns. This empirical approach on a large corpus, beyond verification of hypotheses, may lead to the construction of grounded theories.

Compter les syllabes métriques pour reconnaître les vers, identifier les formes métriques, repérer les phénomènes atypiques sont des activités qui occupent linguistes et littéraires qui se soucient quelque peu des aspects formels, pour qui la forme métrique est porteuse de signification. Dès que cette attention portée à la forme existe, le vers est observé, analysé dans toutes ses dimensions formelles (syllabes métriques, césure, accents, rime, strophes…). Ainsi, les ouvrages de métriques regorgent-ils de règles et d'exceptions. Elwert (1965) est à ce titre instructif : chacun des paragraphes est nourri d'exemples et contre-exemples, puisés dans de vastes corpus. Le lecteur imagine facilement que les textes, en nombre très élevé au vu de la diversité des citations, ont été parcourus un nombre incalculable de fois à l'affût de toutes les variations possibles pour chaque question. Derrière chaque exemple se cache une exploration aussi systématique que possible, « à la main », de grands ensembles de vers. L'exploration artisanale des textes permet d'identifier les cas rares et, dans le meilleur des cas, d'évaluer la répartition de certains critères (présence d'un accent, d'un e muet…) (Roubaud, 1988). L'exploration manuelle est coûteuse en temps, parfois incertaine (erreurs de repérage, incohérence dans les marquages). Elle peine davantage à évaluer les répartitions générales

qu'à identifier les cas exceptionnels. Elle n'est pas en mesure d'identifier des corrélations entre marquages, puisque chacun est traité isolément.
Ce sont ces limites, expérimentées concrètement pour l'étude stylistique d'un recueil de Supervielle, *Gravitations*, qui nous ont incitée à explorer les pistes pour une informatisation du découpage et de la caractérisation accentuelle des vers. Il n'existait pas au début des années 90 d'outil pour l'analyse automatique du vers, ni pour le français, ni pour d'autres langues. L'objectif principal n'était pas tant l'élaboration d'un outil que son exploitation visant à mettre en relation les aspects métriques et rythmiques avec d'autres informations relevant des différents niveaux de l'analyse linguistique (étiquettes morpho-syntaxiques des mots, groupes syntaxiques, etc.).
Grâce à la conjonction entre corpus numérisés, outils de traitement du vers, bases de données et outils statistiques, on peut systématiser les explorations manuelles et les dépasser en créant des mises en relation impensables à une telle échelle en dehors d'un environnement informatique. On a ainsi pu montrer comment le deuxième hémistiche qui paraît a priori fort semblable au premier (hormis la rime évidemment) présente en fait une forme rythmique sensiblement différente à celle du premier : moins d'accents et une répartition beaucoup plus équilibrée des accents.
Nous proposons de rendre compte de la démarche mise en place et des difficultés rencontrées pour analyser le rythme de l'alexandrin classique avec des méthodes de TAL (Traitement Automatique des Langues). Cette démarche a mobilisé des outils sophistiqués de TAL appliqués à d'importants corpus de vers.
Le projet visait à éclairer la question du rythme en considérant celui-ci comme un phénomène complexe, faisant intervenir différentes composantes linguistiques et différents niveaux d'analyse. Il était donc indispensable de concevoir une représentation informatique cohérente avec cette conception du rythme. Ce sera l'objet de la première partie.
Le métromètre, outil qui délimite les positions métriques du vers et les caractérise avec des critères linguistiques, constitue le noyau central du dispositif. Cet outil résulte d'un choix privilégiant une approche par règles (et exceptions) s'appuyant sur une représentation phonétique de vers. En cela, il est en cohérence avec son objet : le vers classique, dans sa manière d'inscrire ses douze syllabes métriques en deux segments de six syllabes, se laisse entièrement appréhender dans un ensemble fini de règles, à condition toutefois de passer par une représentation phonétique (partie 2).
Le développement des outils s'est fortement appuyé sur le savoir accumulé sur le vers et le projet a avancé grâce à la confrontation entre les traités de métrique, qui le plus souvent masquent à peine une dimension normative, et l'observation des pratiques sur grands corpus de vers. La démarche de validation et de dépassement des théories se fait au travers d'allers et retours entre règles, corpus et régularités observées (partie 3).

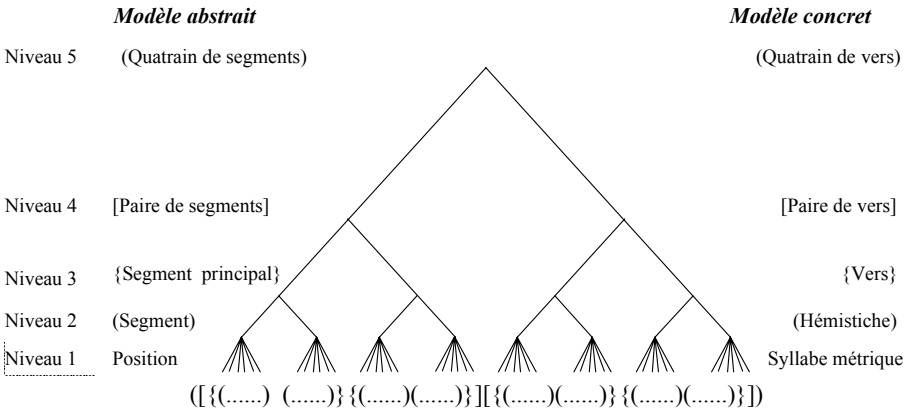
1. Au cœur du dispositif, une base de données de description du vers

Au terme de cette expérience sur l'analyse du vers, nous concevons le rythme comme un phénomène multimodal. Pour la poésie, le rythme s'appréhende autant par la vue à travers la disposition graphique que par l'oreille dans sa forme orale. Il est fait pour un « œil-oreille » comme l'indiquait Roubaud (1986). L'utilisation d'outils de TAL pour la poésie nous a conduit à ne travailler que dans la dimension graphique. A travers ce filtre, le rythme continue cependant d'apparaître comme multimodal, tout d'abord parce qu'il fait intervenir les différentes composantes de la langue ; ensuite parce qu'il résulte de phénomènes relevant de différentes niveaux d'analyse (position, hémistiche, vers, rime…) qui se renforcent mutuellement et fonctionnent en phase (ainsi quel que soit le marquage retenu et le niveau d'analyse, la fin d'hémistiche est-elle toujours marquée). Le dispositif informatique mis en

oeuvre, ici une base de données centrée sur le vers, est en harmonie avec cette représentation théorique du rythme. Il permet 1) de tenir ensemble les différentes composantes de la langue et 2) de faire varier le niveau d'analyse et de pouvoir mettre en relation ces différents niveaux. Cette représentation hiérarchisée et complexe du rythme résulte d'une confrontation entre la théorie du rythme développée par P. Lusson (1974) et J. Roubaud (1978) et l'exploration des corpus de Corneille et Racine à l'aide des outils de TAL mis en place. C'est à travers les allers-retours entre théories et corpus que s'est construite peu à peu une représentation informatique en cohérence avec la théorie tandis que s'affinait la théorie au travers des confrontations avec les corpus.

1.1. Un modèle hiérarchisé du vers

L'application au vers de la théorie du rythme de P. Lusson (1973 et 1998), telle qu'elle a été faite par J. Roubaud (1978), permet de proposer un modèle métrique de l'alexandrin classique à rimes plates alternées. Celui-ci peut être représenté comme sur la figure 1 : la structure métrico-rythmique résulte de l'articulation de différents niveaux hiérarchisés. Chaque niveau est défini par un type d'événement élémentaire, ou constituant. Par exemple, celui du vers est défini par l'hémistiche : le vers est composé par deux hémistiches (segments métriques). Au niveau inférieur, le constituant élémentaire est la position métrique qui compose les hémistiches. Cette modélisation adoptée pour le vers n'est pas sans rappeler les propositions de Benvéniste sur les niveaux d'analyse linguistique (1966) : Benvéniste indiquait la nécessité de décomposer en *niveaux* l'analyse linguistique pour mieux pouvoir recomposer le sens : « La *forme* d'une unité linguistique se définit comme sa capacité de se dissocier en constituants de niveau inférieur. Le *sens* d'une unité linguistique se définit comme sa capacité d'intégrer une unité de niveau supérieur. » (1966, 126-127)



Appliqué à une séquence de quatre vers, le parenthésage devient le suivant (pour assurer la lisibilité, les frontières des positions ou syllabes métriques ne figurent pas) :

([{ (Ariane, ma soeur,) (de quel amour blessée,) }
{ (Vous mourûtes aux bords) (où vous fûtes laissée!) }]
[{ (Que faites-vous, madame ?) (et quel mortel ennui) }
{ (Contre tout votre sang) (vous anime aujourd'hui?) }])

Phèdre, vers 253-256

Clef de lecture : le vers (niveau 3) est constitué de deux hémistiches (niveau 2), chacun étant à son tour constitué de six positions métriques (niveau 1). Deux vers unis par la rime constituent une unité d'un niveau supérieur (niveau 4) et deux paires de vers avec alternance en genre des rimes constituent un quatrain (niveau 5).

Figure 1. Le modèle de l'alexandrin classique à rimes plates alternées

Le vers classique à rimes plates constitue évidemment le cas le plus simple puisque les niveaux sont parfaitement emboîtés. A chaque niveau, le constituant élémentaire est une combinaison de constituants élémentaires du niveau inférieur. Les rimes plates nous permettent d'éviter d'avoir à rendre compte des phénomènes d'intrication (liens à distance) très complexes à traiter (Roubaud, 1979).
Le modèle théorique du vers est hiérarchisé et nous incite de ce fait à mettre systématiquement en relation des phénomènes observés à différents niveaux, ce que les traités de métrique ne font que très rarement. En effet, ces derniers consacrent un chapitre au décompte, un autre à la césure et/ou aux accents dans le vers, l'autre à la rime, sans que la question de la relation entre ces niveaux ne soit traitée. Cette structure hiérarchisée nous guide dans la structuration de la base de données, conçue pour faire dialoguer des niveaux hétérogènes et pour articuler des traits de description provenant de différents niveaux d'analyse.

1.2. Des modules hétérogènes qui contribuent à alimenter la description du vers

Nous avons construit une matrice de description du vers, qui contient en ligne les individus que sont les vers et en colonne un ensemble de variables de description. Cette matrice est alimentée par des modules hétérogènes. Les différents traits de description du vers ou marquages tels qu'ils apparaissent dans la base de données proviennent de différents modules, ont des grains de description variables, sont de nature hétérogènes. Les principaux modules producteurs de traits de description des vers sont :

- Les ***corpus*** qui fournissent les informations paratextuelles de description des vers (quel auteur, quelle pièce, dans quel acte, scène, prononcé par qui…). Nous reviendrons dans la section 2 sur les éditions utilisées. Si nous reprenons les vers cités dans la figure 1, les indications porteront sur l'auteur (Racine), la pièce (*Phèdre*), son genre (tragédie), sa date de publication (1677), les personnages qui prononcent les vers (Phèdre pour les deux premiers, Oenone pour les suivants) ;
- Le ***métromètre*** qui construit la description des positions métriques dans le vers. Ce module est le plus important. Il produit lui-même des représentations hétérogènes du vers : phonétiques, morphosyntaxiques, accentuelles… Il identifie aussi le nombre de syllabes du vers, ce qui permet de présumer de la forme du vers (octosyllabe, décasyllabe, alexandrin…). Nous y revenons dans la deuxième partie ;
- Le ***rimarium*** qui regroupe les mots pouvant rimer ensemble, en appliquant une règle de transitivité généralisée (si A rime avec B et B avec C, alors A peut rimer avec C). Le *rimarium* a été construit sur les paires de vers rimant ensemble. A chaque classe de mots-rimes correspond un *rimème*, séquence phonético-graphique qui décrit la classe. En conséquence, chaque mot-rime peut être décrit par son *rimème*. Ainsi, si « amour » rime avec « détour » et « détour » avec « jour », et que ces mots ne riment avec aucun autre, « amour », « détour » et « jour » constituent une classe, dont le *rimème* est « –our ». Chaque vers finissant par un des mots de la classe peut être qualifié par le *rimème* « –our ». Ainsi pour les vers précédents, les indications qui seront introduites dans la base de données seront le rimème, le genre et la terminaison de la rime, soit pour les deux premiers vers « sées », rime féminine, terminaison plurielle et pour les suivants : « ui », rime masculine, terminaison singulier ;
- Les ***outils de statistique textuelle*** qui permettent de construire des typologies de vers lexicalement proches et d'affecter une classe lexico-sémantique à chaque vers. En général, les outils de statistique textuelle qui exploitent les cooccurrences ont besoin d'analyser des séquences de texte d'une certaine longueur. Le vers est parfois insuffisant. Des typologies de vers ont pu être construites sur des séquences de deux vers à dix vers. Ensuite chaque séquence de vers est affectée à une classe, cette

affectation pouvant se reporter à chacun des vers de la séquence. Ainsi, suite à des traitements de statistique textuelle portant sur l'ensemble du corpus, les quatre vers de *Phèdre* sont classés dans la classe interprétée en « Mort et culpabilité ». Sur la base d'une autre typologie ne portant que sur les tragédies de Racine, ces mêmes vers sont classés dans la classe « amour-dialogue ».

Ainsi, chaque vers est décrit par un ensemble de traits de nature différente, qui s'appliquent à des composants du vers, au vers ou à des structures supérieures. Ces traits de description sont produits par des modules autonomes et de nature différente. Le corpus, constitué principalement des 80 000 vers de théâtre de Corneille et Racine, provenait de la Bibliothèque Nationale de France qui reprenait la base Frantext : les textes de Corneille ont été saisis d'après l'édition de Ch. Marty-Laveaux, ceux de Racine d'après l'édition de P. Mesnard. Il a été balisé par nos soins ce qui permet ensuite d'extraire les traits paratextuels. Le métromètre est le résultat d'une adaptation d'outils de TAL (phonétiseur et analyse syntaxique) : nous y reviendrons dans la section suivante. Le *rimarium* est le résultat d'un programme que nous avons développé dans un langage propriétaire, SAS[1] ; le programme est loin de se présenter sous une forme industrialisable et reproductible. Enfin, les outils de statistique textuelle utilisés, principalement Alceste[2], sont des outils que l'on trouve sur le marché. Les traits sont acquis par des modules hétérogènes, mais leur intégration dans une base de données permet d'une part une homogénéisation des traitements et de l'autre des transferts de la base et donc d'autres formes d'exploitation.
Nous avons une table avec une ligne par vers et une ensemble de variables de description : un premier groupe de variables décrit la pièce dont est issu le vers, la position de ce dernier dans la pièce, le personnage qui le prononce ; le suivant fournit des éléments de description de chaque position métrique (nous y revenons dans la section suivante), un troisième ensemble donne des éléments de description de l'hémistiche et un quatrième fournit les descriptifs liés à la rime (genre, terminaison, rimème) et un dernier ensemble fournit les classes lexico-sémantiques auxquelles ont été affectés les vers, sachant qu'il y a autant de variables que de classification des vers. Tout cela aurait pu être organisé dans une base de données relationnelle avec une table spécifique pour chacun des niveaux d'analyse et des liens entre les tables. Le logiciel SAS, que nous avons utilisé pour gérer la base de données et pour effectuer les traitements, permettait de traiter toute l'information dans une seule table, ce qui simplifiait la formulation des requêtes. Nous avons une seule matrice avec un vers par ligne. A partir de cette matrice, nous pouvons construire aisément d'autres matrices où l'individu est la position, l'hémistiche, le couple de vers, ce qui nous permet de faire varier le niveau d'analyse.
Nous avons au final des éléments de caractérisation de la syllabe métrique, de l'hémistiche (structure accentuelle ou morpho-syntaxique du segment), de la paire de vers (rimème), de séquences de vers (traitement de statistique textuelle) et des éléments de description de la pièce.

1.3. Faire varier les niveaux d'analyse

Nous venons de voir que pour décrire les vers, nous avions recours à des ressources hétérogènes éditoriales, métriques, morphosyntaxiques, rimiques, lexico-sémantiques, etc. Or ces ressources, en raison du modèle hiérarchisé et parfaitement imbriqué du vers classique, bien qu'elles aient des grains de description très différents peuvent être reportées sur le niveau du vers. Si la matrice principale est centrée sur le vers, on peut procéder à des calculs qui

[1] SAS : progiciel pour la constitution et le traitement statistique de bases de données (www.sas.com).
[2] ALCESTE : outil de statistique textuelle, conçu et développé par Max Reinert (1993), qui permet de constituer des classes lexico-sémantiques dans un corpus.

porteront sur des unités de taille variable comme la position métrique, l'hémistiche, la paire de vers rimant ensemble, la pièce, le genre, l'auteur. Les propriétés sont transmises d'un niveau à l'autre par translation… Ainsi, peut-on peut aisément à partir de la base centrée sur le vers, constituer une base dont « l'individu » est la position (réduction de la taille du grain) ou au contraire dont l'individu est la paire de vers rimant ensemble. Ainsi peut-on calculer la répartition de la voyelle *a*, toutes positions métriques confondues ; sa répartition en première position d'hémistiche, en première position de vers ou encore en première position de couple de vers. Parce que les effets de structure se font sentir à chacun des niveaux de l'analyse métrique, les résultats seront sensiblement différents selon le grain retenu.

Cette possibilité de variations permet de mettre en relation des phénomènes observés à différents niveaux d'analyse. Ainsi, avons-nous pu montrer avec une simple mesure du chi2 qu'il y avait des corrélations entre des champs lexico-sémantiques, calculés sur des séquences de plusieurs vers, avec la structure accentuelle des hémistiches. A titre d'exemple, la thématique de la mort est associée avec un vers plutôt régulier dans lequel dominent les hémistiches portant un « accent » en troisième position, soit un marquage de l'hémistiche de type 001001 tandis que pour la thématique de l'amour, les formes non régulières (ni de type 001001, ni de type 010101) sont significativement plus fréquentes. Nous avons aussi pu montrer que pour les tragédies de Racine, les rimes les plus spécifiques de chaque pièce correspondaient aux noms des personnages. Ainsi, les rimes en « *-ate* » et « *-ime* » sont dominantes dans *Mithridate*, dont les principaux personnages sont *Mithridate*, *Monime*. On notera que Racine s'efforçait de donner aux confidents des noms qui rimaient avec le personnage principal (*Arbate* pour *Mithridate*, *Phaedime* pour *Monime*). Certes ces noms sont souvent utilisés à la rime, mais même lorsque l'on exclut des calculs les noms de personnages, la sur-représentation de ces rimes se maintient. Tout se passe comme si Racine avait souhaité donner une coloration sonore spécifique à chacune de ses tragédies, portée par les sonorités des noms. D'autres croisements ont été faits entre différents marquages, et bien d'autres pistes restent encore à explorer comme le lien entre la forme du vers et la structure de la rime.
Les possibilités d'exploration sont très grandes grâce à cette organisation en base de données qui permet toutes formes de croisement entre les traits de description, même si ces derniers relèvent de niveaux d'analyse différents.

2. Le métromètre : une chaîne de traitement hybride

Dans la base de données, le noyau central de description métrico-rythmique du vers est le métromètre. Cette section décrit les choix qui ont été faits pour avoir un module de description métrico-phonétiques, et quelques difficultés liées à ces choix.
La question de départ était simple : où trouver un outil informatique qui soit capable de découper le vers en syllabes métriques, voire de disposer les accents sur les positions métriques ? Un tel outil n'existait pas pour le vers français, et aujourd'hui encore, en dehors du métromètre, nous n'en connaissons pas d'autres. Pourtant, le découpage des vers en syllabes métriques obéit à des règles très strictes, ce qui aurait dû rendre aisé l'automatisation du processus. Comme l'outil n'existait pas, et que nous n'étions pas en mesure d'en développer un, il a fallu faire alliance, avec des équipes informatiques, prêtes à consacrer du temps à la question du vers, qui n'est pas *a priori* un enjeu majeur de la recherche en informatique (Beaudouin & Yvon, 1996).
Le cas du vers français n'est pas isolé. Peu d'outils d'analyse du vers ont été développés pour d'autres langues. Le projet d'informatisation le plus ancien remonte au début des années 90 : D. Robey (1993) souhaitait mettre en place un système d'analyse métrique de l'hendécassyllabe de la Divine Comédie. Confronté à de grandes difficultés techniques et

poétiques, il a finalement opté pour un système interactif de codage et de marquage. En 1994, Y. Ousaka, M. Yamazaki et M. Miyao ont proposé un système d'analyse automatique d'un des principaux canons du janaïsme en moyen indo-aryen, qui permet de reconnaître les mètres à partir d'une analyse des voyelles. Cette analyse nécessite au préalable une transduction en caratères latins des textes originaux. Enfin, en 2000, P. Gervas a proposé un outil d'analyse du vers espagnol qui s'appuie directement sur la forme graphique. Il est fort probable que nous n'ayons pas une connaissance complète de tous les systèmes qui ont pu être développés. Ces derniers se distinguent cependant par leur faible nombre.

2.1. Apprentissage ou règles et exceptions

Dans le projet Dynastie (1986, 1988), Roubaud avait esquissé les premières règles qui devaient permettre, à partir de la chaîne écrite, d'identifier les positions métriques. Une première équipe d'Intelligence Artificielle [3], spécialisée en apprentissage, a accepté de collaborer sur la question : mise en place de règles et acquisition de nouvelles règles à partir de l'exploitation du corpus. Très rapidement, les difficultés posées par la graphie très complexe du français ont conduit à abandonner cette voie : il nous fallait passer par une représentation phonétique pour pouvoir traiter le problème du découpage. Et c'est bien un problème lié au français. En effet, pour l'espagnol, P. Gervás (2000) en utilisant les grammaires logiques en Prolog (Definite Clause Grammars), directement appliquées à la graphie a développé un système de règles hiérarchisées qui découpe et accentue les positions métriques correctement. Chaque mot du vers est découpé en syllabes puis sont ensuite appliquées des règles spécifiques de la métrique pour ajuster : synalèphe, accent, rime…Pour le français, on est obligé de passer par une représentation phonétique, qui elle-même s'appuie sur une analyse morpho-syntaxique. En effet, la seule analyse graphique est insuffisante car des phénomènes comme la diérèse et la synérèse, le *h* en début de mot… ne sont traitables qu'avec une qualification morpho-syntaxique du mot.
Nous avons fait alliance avec une équipe de l'Ecole Nationale Supérieure des Télécommunications (ENST), spécialisée en TAL, et disposant d'outils d'analyse syntaxique et de phonétisation. Une démarche possible était de partir d'ensemble de règles de transcription phonétique prévues pour le français ordinaire et de procéder à des ajustements vers à vers en fonction du résultat obtenu. Pour un alexandrin, qui en diction standard donnerait onze syllabes, on cherche les *e* muets qui pourraient être comptés, les diérèses qui pourraient être faites, pour que le compte soit juste. Cela sous-entend un ajustement vers après vers pour atteindre le bon compte.
Cette démarche n'a pas été retenue pour deux raisons. D'une part, le corpus contenait des vers de longueur variable. Les systèmes d'ajustements ne sont possibles que si l'on connaît a priori la longueur visée. Cela devient plus périlleux quand la longueur des vers varie de manière non prévisible. D'autre part, les règles de prononciation ou de décompte du vers classique sont fixes : un mot n'est pas prononcé avec ou sans diérèse selon le contexte ; la prononciation des e muets répond elle aussi à des règles précises. Comme les règles du vers classique peuvent être formalisées, il était dommage d'adopter un système d'ajustement au cas par cas, susceptible d'introduire des erreurs, au lieu de construire les règles.

Nous avons donc adopté un système à base de règles, avec des listes d'exceptions (principalement pour la diérèse et synérèse), adapté au caractère extrêmement réglé du vers français.

2.2. Le métromètre : s'appuyer sur l'existant, accepter quelques compromis

[3] Le Laforia à Paris VI, J. G. Ganascia.

Quelle aurait été la situation idéale ? Disposer des textes dans leur édition d'origine et avoir des outils de TAL adaptés à la langue et aux graphies du XVII^e siècle. La tradition de numérisation des textes, ancienne en France puisqu'elle remonte aux années 60, n'a pas, dans les premiers temps, été soucieuse de la qualité des éditions utilisées : les éditions de Corneille et Racine datent du XIX^e siècle. Ensuite le développement des outils de TAL est récent et les principaux domaines d'application concernent un état de langue contemporain. Il y a fort à parier que des outils de TAL adaptés à la langue du XVII^e siècle n'apparaîtront pas de sitôt. Il a donc fallu faire des compromis par rapport à cet idéal-type : les vers analysés utilisent une graphie moderne, les outils de TAL sont conçus pour la langue contemporaine.

Pour le découpage en positions métrique, les premières expériences sur la forme graphique ont convaincu qu'il était indispensable de passer par une représentation phonétique du vers pour pouvoir le découper correctement. Comme il n'était pas question de développer un nouveau phonétiseur adapté au vers, le métromètre s'appuie sur des outils existants : un analyseur syntaxique développé par Patrick Constant (1991), un phonétiseur développé par François Yvon (1995) qui s'appuie sur l'analyseur syntaxique. Ces systèmes ont été adaptés au cas particulier de la métrique du vers. Trois phénomènes métriques ont donné lieu à l'identification et à l'implémentation de règles spécifiques : la diérèse/synérèse, le e muet et la liaison.

L'outil est remarquablement performant parce qu'il s'appuie sur des composants de qualité industrielle. Une difficulté provient cependant de la non disponibilité de l'outil : en effet, l'analyseur syntaxique, Sylex, est un produit développé à l'origine à l'ENST, mais commercialisé désormais par une société privée (Ingenia). Le phonétiseur, qui s'appuie sur l'analyseur syntaxique, est un produit de laboratoire, qui en tant que tel pourrait être diffusé s'il n'était pas dépendant de cet analyseur syntaxique. Pour se libérer de cette dépendance par rapport à un analyseur syntaxique particulier, F. Yvon développe un phonétiseur qui pourrait être « branché » sur n'importe quel analyseur.

2.3. Une représentation intégrée des différents niveaux distingués par la linguistique

Dans le cadre des applications de la théorie du rythme [Lusson, 1974 ; Roubaud, 1986 1988) de nombreux marquages ou traits de description des positions métriques (*e* muet, fin de mot, accent, ponctuation…), tous représentés par le couple 1/0 (présence/absence), ont été introduits et utilisés. Le métromètre, reprenant cette notion de marquage, produit une représentation du vers organisée selon les niveaux que distingue traditionnellement la linguistique : phonétique, morpho-syntaxique, accentuel et lexical comme on peut le voir dans l'analyse que fait le métromètre du premier vers de Racine cité dans la figure 1 : La première ligne fournit la transcription phonétique du vers segmentée selon les positions métriques, conformément aux règles de la métrique classique. Les règles de la diérèse et de la synérèse ainsi que les règles concernant le traitement du *e* muet et de la liaison ont été introduites pour produire une représentation phonétique métriquement correcte. Comme les corpus correspondent à des éditions modernisées et que l'outil produit une prononciation contemporaine, la transcription phonétique est sans doute assez éloignée de ce qu'aurait pu être une transcription d'époque. La deuxième ligne extrait les voyelles métriques de chacune des syllabes, la troisième ligne indique si la syllabe correspond ou non à la syllabe finale du mot, la quatrième indique la catégorie morpho-syntaxique des mots auxquels appartiennent les syllabes ; la suivante indique les positions métriques qui portent une marque accentuelle. La dernière ligne donne le nombre de syllabe ce qui permet d'identifier aisément les erreurs

de corpus ou d'analyse[4]. Le tableau 2 présente l'ensemble des variables dans la base de données : celles que nous avons vues dans la section 1 ainsi que les variables construites à partir de l'analyse du métromètre.

Syllabes métriques	a	r i	**a**	n □f	m a	**s œ r**	d □	k ɛ l	a	**m u r**	b l e	**s e**
Voyelles métriques	a	i	**a**	□f	a	**œ**	□	ɛ	a	**u**	e	**e**
finale de mots	-	-	**-**	oui	oui	**oui**	oui	oui	-	**oui**	-	**oui**
Catégories syntaxiques	Nom propre	Nom propre	**Nom propre**	Nom propre	dét	**nom**	prép	pron	nom	**nom**	adj	**adj**
Marquage accentuel	-	-	**oui**	-	-	**oui**	-	-	-	**oui**	-	**oui**
Nb syllabes	12											

Clef de lecture : La première syllabe métrique du vers est /a/, elle ne correspond pas à une finale de mot car c'est la première syllabe de *Ariane*, elle appartient à un mot catégorisé comme « nom propre » et elle ne porte pas d'accent, l'accent porte en effet sur la troisième syllabe métrique qui est aussi /a/. Le petit */f/* accolé au signe du/ e/ muet indique qu'il s'agit d'un */e/* muet faible, qui selon les phonèmes qui le suivent constituera ou non une voyelle métrique.

Figure 2. Exemple de sortie d'analyse de vers

Cette décomposition par niveau pourrait laisser croire que les composantes linguistiques sont traitées de manière indépendante les unes des autres. En réalité, au cours du traitement, les différents niveaux d'analyse sont mis en relation. Premièrement, la représentation phonétique est construite sur la base d'une analyse syntaxique. La transcription phonétique n'aurait pu en effet se passer d'une analyse syntaxique préalable qui détermine la catégorie morpho-syntaxique de la forme graphique. En effet, le mot sous sa forme graphique isolée se révèle être une matière des plus ambiguës. Patrick Constant, dans la conception de son analyseur syntaxique Sylex (1991), prend comme point de départ cette situation maximale d'ambiguïté et met en place des couches de règles qui une à une visent à réduire l'ambiguïté. Souvent, même au terme de ces explorations, il reste des interprétations différentes proposées à l'utilisateur. L'ambiguïté des formes graphiques a des conséquences cruciales sur la transcription phonétique, car une même forme peut être « comptée » différemment selon sa catégorie : ainsi, *fier* est-il analysé comme un mot d'une syllabe quand il s'agit de l'adjectif et de deux quand il s'agit du verbe. Sans aller jusqu'au décompte, la transcription phonétique est de toute évidence différente selon l'analyse.
Deuxièmement, le marquage accentuel s'appuie d'une part sur l'analyse morphosyntaxique qui détermine la catégorie des mots et, d'autre part, sur l'analyse phonétique qui signale entre autre la position des *e* muet : en effet, si le mot finit par un *e* muet maintenu, cette dernière syllabe ne peut porter l'accent. L'existence de l'accent dépend de la catégorie du mot et sa place dans le mot de sa structure phonétique.
Ainsi, la séparation des niveaux linguistiques n'est-elle qu'une forme de représentation, l'imbrication entre les niveaux étant constitutive de l'outil.

Les analyses produites par le métromètre semblent proposer des représentations par composantes linguistiques séparées. Mais la mise sous forme de base de données unique permet de mettre en relation des description provenant de niveau différent et de faire émerger

[4] Nous n'insistons pas ici sur les limites de l'outil, longuement analysés dans [Beaudouin, 2002, 249-274].

des corrélations, certaines étant connues, d'autres plus inattendues : ainsi, sur les sixième et douzième positions, la part des mots grammaticaux est extrêmement faible, l'accent quasi-systématique et les *e* muets toujours absents. On observe des corrélations fortes et inattendues entre des marquages comme la répartition de l'accent et l'« épaisseur » des syllabes (en nombre de phonèmes).

A travers la décomposition en niveaux d'analyse et sa recomposition, nous voyons à quel point le système linguistique forme un tout et les interactions se font de fait à tous les niveaux. Quel que soit le marquage retenu, il contribue à redéfinir les contours de l'hémistiche, en marquant ses frontières par un contraste important avec les positions adjacentes. Tout se passe comme si chaque composante de la langue, tel un instrument dans un orchestre, contribuait à renforcer le mouvement d'ensemble ponctué par le mètre : ce sont bien ces séquences de six syllabes, les segments métriques constitutifs de l'alexandrin qui contribuent à organiser la matière linguistique.

	Signification
Signalétique	Nom de l'auteur
	Nom de la pièce
	Numéro du vers courant
	Numéro de l'acte
	Numéro de la scène
	Personnage
	Genre de la pièce
	Genre et auteur de la pièce
	Numéro de l'interprétation
	Vers
Description multidimensionnelle des positions	Syllabe 1
	Voyelle 1
	Catégorie syntaxique 1
	Accent 1
	Fin de mot 1
	Syllabe 2
	Voyelle 2
	Catégorie syntaxique 2
	Accent 2
	Fin de mot 2
	Syllabe 3
	Voyelle 3
	Catégorie syntaxique 3
	Accent 3
	Fin de mot 3
	Syllabe 4
	Voyelle 4
	Catégorie syntaxique 4
	Accent 4
	Fin de mot 4
	Syllabe 5
	Voyelle 5
	Catégorie syntaxique 5
	Accent 5
	Fin de mot 5
	Syllabe 6
	Voyelle 6
	Catégorie syntaxique 6
	Accent 6
	Fin de mot 6
	Syllabe 7
	Voyelle 7
	Catégorie syntaxique 7
	Accent 7
	Fin de mot 7
	Syllabe 8
	Voyelle 8
	Catégorie syntaxique 8
	Accent 8
	Fin de mot 8
	Syllabe 9
	Voyelle 9
	Catégorie syntaxique 9
	Accent 9
	Fin de mot 9
	Syllabe 10
	Voyelle 10
	Catégorie syntaxique 10
	Accent 10
	Fin de mot 10
	Syllabe 11
	Voyelle 11
	Catégorie syntaxique 11
	Accent 11
	Fin de mot 11
	Syllabe 12
	Voyelle 12
	Catégorie syntaxique 12
	Nombre de positions
	Accent 12
	Fin de mot 12
	Semi-V-Voyelle 1
	Semi-V-Voyelle 2
	Semi-V-Voyelle 3
	Semi-V-Voyelle 4
	Semi-V-Voyelle 5
	Semi-V-Voyelle 6
	Semi-V-Voyelle 7
	Semi-V-Voyelle 8
	Semi-V-Voyelle 9
	Semi-V-Voyelle 10
	Semi-V-Voyelle 11
	Semi-V-Voyelle 12
Figure	Figure accentuelle 1er hémistiche
	Figure accentuelle 2eme hémistiche
	Figure accentuelle réduite 1er hémistiche
	Figure accentuelle réduite 2eme hémistiche
	Figure des fins de mots 1er hémistiche
	Figure des fins de mots 2eme hémistiche
Rime	Nombre de mots
	Dernier mot du vers
	Genre de la rime
	Fréquence du mot-rime sur CORRAC
	Groupe de mots-rimes
	Char
Stat. textuelle	Typologie sur Corrac
	Typologie tragédies de Racine
	Autres typologies

Tableau 1. Descripteurs multidimensionnels du rythme définis dans la base de données

3. Règles, corpus, régularités

Cette expérience menée sur le vers nous a permis d'identifier deux types de configurations dans la manière d'articuler règles, corpus et régularités.

Nous nous sommes parfois trouvée devant des cas simples, où les règles identifiées de longue date, souvent sous la forme de prescriptions, pouvaient facilement être implémentées dans le système, avec des adaptations marginales. La théorie, suffisamment précise et adaptée à la réalité, rendait l'automatisation simple. Un autre cas se rapproche de cette première configuration, bien qu'il nécessite des ajustements coûteux en temps. Il arrive en effet que la théorie ou les manuels de métrique proposent des règles justes mais non exhaustives, car tous les cas ne sont pas couverts. Les règles doivent être complétées et enrichies par l'examen des corpus, et sont donc dépendante des corpus, puisque l'ajout de nouveaux textes peut modifier les règles et leurs exceptions. L'examen des corpus conduit à enrichir le système de règles et à l'ajuster marginalement.

Dans l'autre configuration, l'analyse du corpus peut conduire, comme nous allons le voir, à invalider certaines théories ou certaines hypothèses. L'analyse fait émerger des régularités,

jusque-là ignorées, qui deviennent constitutives du rythme du vers, bien qu'il ne s'agisse pas de règles métriques.
Ces deux configurations étaient aussi un moyen pour nous d'éclairer les frontières entre mètre et rythme. Quel que soit le point de vue adopté, nous avons oscillé entre l'utilisation des prescriptions et l'observation des pratiques (Habert & Zweigenbaum, 2002)

3.1. Quand les théories s'appliquent…

En construisant le métromètre, il ne s'agissait pas de faire table rase du passé, bien au contraire. Les traités et manuels de métrique ont été la ressource principale utilisée pour construire les règles permettant l'analyse syllabique du vers. Le découpage des vers en syllabes ou positions métriques est un exercice que tout étudiant initié à la question métrique réalise sans difficultés, car il n'y a pas d'aléatoire dans la manière dont le vers classique est élaboré. Le découpage répond à des règles systématiques qui ne souffrent pas d'exceptions. Reste à choisir les théories qui énoncent les règles de la manière la plus rigoureuse pour que la mise en œuvre informatique soit aisée.
Le traitement du *e* muet dans le métromètre est une application quasiment complète des propositions de Milner (1974), reprises dans *Dire le vers* de Milner et Regnault (1987). Il en est de même pour le traitement de la liaison. L'originalité de la démarche de Milner est la suivante : il définit les règles de traitement du *e* muet et de la liaison en langue en considérant comme unité de traitement le mot phonologique ; il pose que le vers constitue à lui seul un mot phonologique ; alors, les règles identifiées pour les mots phonologiques s'appliquent au vers. Dans ce cas, le modèle théorique était suffisamment générique pour pouvoir être décliné avec une grande économie de règles[5].
Pour le traitement de la diérèse et de la synérèse, la mise en œuvre des règles a été plus complexe. Doit-on ou non scinder les groupements vocaliques commençant par une voyelle haute (*i*, *ou* ou *u*) ? Dans *Ari-ane*, il y a diérèse car le groupement *ia* constitue deux voyelles métriques alors que dans *diable*, on a une synérèse, le groupe *ia* ne constituant qu'une voyelle (le *i* acquiert une valeur consonantique). Certes, certains manuels traitent de manière approfondie la question de la diérèse comme Gramont (1876) et plus récemment Elwert (1965). Mais ces ouvrages ne donnent pas une représentation exhaustive des phénomènes. En effet, la question du décompte des groupements vocaliques commençant par une voyelle haute recouvre un nombre très élevé de cas, qui souffrent en plus d'un certain nombre d'exceptions, qu'il serait très rébarbatif de trouver dans un traité. Pour mettre en œuvre les règles de la diérèse et synérèse, le recours aux traités de métrique était donc insuffisant. Un examen manuel de tous les vers comprenant un groupement vocalique commençant par un voyelle haute a été entrepris, qui a conduit à l'élaboration de règles, environ 70, et de listes d'exceptions. Bien que les règles cherchent à monter en généralité par rapport aux cas observés, elles continuent de dépendre fortement du corpus. L'extension du corpus, initialement restreint à Corneille et Racine, a en effet conduit à une extension du nombre de règles et d'exceptions.
La construction d'une représentation syllabique du vers conforme aux règles de la métrique, résulte d'une implémentation de règles générales et de règles empiriques induites de l'analyse des corpus. En ce sens, le métromètre résulte d'une forme d'hybridation entre règles théoriques et règles empiriques. En tout état de cause, pour le découpage en syllabes métriques du vers classique, nous sommes dans un modèle logique et systématique, étranger à la notion de régularité : ce sont des règles qui dictent la structure syllabique des vers. Pour des

[5] On notera que Milner propose un mode de traitement autonome du niveau phonologique, puisque ses règles ne nécessitent pas de recours à des informations provenant d'autres niveaux linguistiques, conformément à la position qu'il défend sur l'autonomisation des niveaux dans l'analyse (Milner 1989). Si cela est juste pour les phénomènes qu'il traite (*e* muet et accent), pour la diérèse, l'accès à des informations du niveau lexical est indispensable.

formes de vers plus contemporaines, la notion de régularité reprend tout son sens, comme, par exemple, lorsque G. Purnelle (2002) étudie le « quasi-alexandrin » d'Yves Bonnefoy.
Quand l'adéquation est bonne entre les règles énoncées dans les manuels et leur application, on est en mesure d'observer les rares variations par rapport à ces règles, qui peuvent être dues à des flottements dans le traitement ou à des licences. On trouve encore dans les premières pièces de Molière, des cas où les séquences consonne+liquide+voyelle haute+voyelle constituent une syllabe (*san-glier*), alors qu'à cette période se mettait en place la règle qui interdit les suites de trois consonnes et force à vocaliser les voyelles hautes après deux consonnes. Quant aux licences, elles sont rarissimes chez Corneille et Racine, un peu plus fréquentes chez Molière, qui était sans doute moins soucieux d'une application systématique des règles.
Le découpage en syllabes métriques n'a de fait pas été le lieu de débats théoriques importants. Les règles étaient plus ou moins clairement formalisées, donc plus ou moins faciles à mettre sous forme informatique, mais elles avaient le mérite d'être en cohérence avec l'examen des corpus.

3.2. Quand les régularités observées invalident les théories

Cette situation de conformité entre la théorie et les phénomènes observés n'était évidemment pas systématique. Nous évoquerons deux théories qui se sont trouvées invalidées par l'examen et l'analyse de grands corpus de vers : celle des accents dans le vers et celle de la richesse de la rime.
Commençons par la question de l'accent. Le mètre alexandrin, si l'on ne regarde que la sous-partie du modèle constitué par le vers, est composé de douze syllabes, avec un marquage ou accent en sixième et douzième positions. C'est ainsi que le vers français a été décrit par les premiers métriciens, au moins jusqu'au XIX^{e} siècle, et qu'il a été à nouveau décrit au XX^{e} (Gouvard, 1996). Entre temps, des théories accentuelles du vers ont été proposées et ont eu une forte diffusion. Quicherat (1838) considère ainsi que le vers comprend deux accents fixes (en sixième et douzième positions) et deux accents mobiles, un par hémistiche, qui se posent sur les autres positions. Cette théorie se retrouve dans des ouvrages récents (Milner et Regnault, 1987).
Nous avons cherché à voir comment se répartissaient les accents dans le vers. La notion d'accent ne va pas de soi en français, certaines approches nient même son existence. Beaucoup de linguistes s'accordent cependant à reconnaître qu'il existe un accent de groupe (mot phonologique, groupe de souffle…) qui tombe sur la dernière voyelle pleine (autre que *e* muet) du groupe (Dell, 1984 ; Vaissière, 1991). La mise en œuvre de cette notion d'accent n'est pas simple puisqu'elle impliquerait une analyse syntaxique fine pour identifier les groupes. Nous avons dû y renoncer et avons opté pour une approximation de cette notion d'accent, en plaçant un accent sur la dernière syllabe pleine des mots pleins (nom, verbe, adjectif, adverbe), reprenant ainsi la « marque fondamentale » proposée par Roubaud (1988).
En adoptant ce marquage, il apparaît qu'il n'y a pas de différence de nature entre les accents dans le vers et qu'ils sont tous d'ordre linguistique. Ce qui distingue les accents de fin d'hémistiche des accents internes relève de la différence entre règle et régularité : en fin d'hémistiche et de vers, l'accent linguistique est *toujours* présent : la fin des hémistiches coïncide sans exception avec la fin d'un mot phonologique ou groupe accentuel. Les autres accents se répartissent de manière moins systématique sur les autres positions : plutôt sur les positions 2, 3 et 4 et très rarement sur les positions 1 et 5. La présence de deux accents par hémistiche est loin de constituer une règle : la « règle » des quatre accents n'est en effet vérifiée qu'à 60%, ce qui fragilise grandement la théorie.

Le cas de la rime est lui aussi exemplaire par les écarts qu'il met à jour entre les théories et la pratique. En même temps que les traités consacraient tous des sections conséquentes à la rime, de nombreux dictionnaires de rimes ont été publiés comme soutien à l'élaboration de vers. Or, l'analyse systématique du traitement effectif de la rime concorde assez peu avec la vision que donnent les dictionnaires ou les manuels de métrique[6]. Ces derniers considèrent la rime en fonction de sa richesse et la richesse est le plus souvent, comme chez Martinon (1905) ou Grammont (1908), déterminée par le nombre de phonèmes communs. Les traités semblent considérer que la richesse de la rime peut être définie de manière globale, sans tenir compte du contexte. Or la construction des réseaux de mots-rimes sur l'ensemble des pièces de Corneille et Racine, tend au contraire à montrer que la qualité d'une rime ne peut être évaluée qu'en *contexte*, en fonction de la terminaison donnée. Contrairement à ce qu'indiquait Grammont (1908), la qualité de la rime ne peut être définie de manière générale en fonction du nombre de phonèmes communs entre les terminaisons. La qualité de la rime (pauvre, suffisante, riche) ne peut être évaluée que pour un type de terminaison donné ; c'et en ce sens qu'elle est contextuelle. Ainsi, une terminaison « –er » ne peut constituer une rime, « -ir » est à peine suffisant, tandis que « –or » constitue une rime 'normale'. En effet, on ne trouve jamais de rime en « –er » sans similitude de la consonne d'appui, alors que les cas sont rares de rimes en « –ir », et que les rimes en « –or » sans consonne d'appui sont la situation la plus commune.

Dans ce cas encore, la confrontation entre les théories et l'émergence de régularités nées de l'observation des corpus conduit à invalider certaines théories mais aussi à redéfinir des cadres théoriques davantage en adéquation avec les pratiques observées.

3.3. Quand les régularités observées conduisent à de nouvelles théories

En effet, le dialogue entre les théories sédimentées sur le vers et l'observation des pratiques effectives, en décalage avec ces théories, nous a conduit à rendre compte de régularités nouvelles et à proposer des règles ou des théories en accord avec les observations. Nous reprenons à nouveau les exemples de l'accent dans le vers et de la rime.

L'exploration du corpus a invalidé certaines des théories sur l'accent dans le vers. Elle nous a amenée à considérer qu'il était inutile de poser l'existence d'accents métriques qui auraient une forme d'autonomie par rapport aux accents linguistiques : le mètre est informé par la matière de la langue et par elle seule. Les marquages morpho-syntaxiques et accentuels mis en place dans le métromètre et projetés sur le vers sont définis par des règles *exclusivement linguistiques* qui ne tiennent pas compte du fait qu'il s'agit de vers. Or, quel que soit le marquage retenu, sa distribution sur chaque position délimite les deux segments métriques qui constituent le vers : une forme similaire se reproduit entre le premier et le second hémistiche. La frontière entre les deux hémistiches, entre la fin d'un vers et le début du suivant, se construit par le contraste maximal, quel que soit le marquage, entre la fin d'un segment et le segment suivant : en début de segment, les syllabes ne sont que très rarement accentuées, les mot-outils (articles, prépositions, pronoms, conjonctions…) sont très fréquents et les syllabes très fluettes (en nombre moyen de phonèmes) ; en fin de segment, l'accent est quasi-systématique, les mots-pleins (noms, verbes, adjectifs et adverbes) ont effacé tous les mots-outils et les syllabes sont devenues « épaisses ». Contraste important entre la fin d'un segment et le suivant, mais aussi entre la fin d'un segment et l'avant-dernière position, dont le profil est assez similaire au début de segment. Tout se passe comme si l'éminence de la fin du segment métrique était renforcée, garantie par le déficit de marquage des positions adjacentes.

[6] D. Billy (1984) a fortement contribué à renouveler la réflexion sur la nomenclature des rimes.

Le mètre définit le cadre dans lequel vient s'inscrire le rythme. Toutes les composantes de la langue projetées dans ce cadre voient leurs effets se renforcer réciproquement. C'est la distribution de ces marquages linguistiques, ajustés sur le segment du vers, qui rend visible la structure rythmique du vers, autrement dit la régularité de la répartition des marquages quel que soit le niveau linguistique retenu.
Tous les accents dans le vers sont de nature linguistique, ils se distinguent cependant par leur caractère plus ou moins systématique. Certains, comme Meschonnic (1982), considèrent que seuls les accents systématiques sont métriques, tandis que les autres, plus variables, liés à la langue, relèvent du rythme.
L'exploration de la distribution des accents a permis d'identifier d'autres régularités qui nous paraissent propres à décrire le vers même si elles ne s'expriment pas sous forme de règle. Nous avons vu que le vers est constitué de deux segments métriques : par quoi est donc assurée l'unité du vers, en dehors de la rime ? Y a-t-il des variations sensibles entre le premier et le second hémistiche du vers ? L'examen de la répartition des accents sur les deux hémistiches montre que le second comporte moins d'accents, et que ces derniers se situent de manière privilégiée sur les positions paires ou ternaires. Le vers avance en gagnant en régularité, et cette tendance se trouve confirmée sur tous les corpus étudiés. Il ne s'agit pas d'une règle, mais d'une régularité qui a émergée de l'analyse des corpus. Ainsi, par delà la rime, nous voyons comment certains marquages contribuent à garantir l'unité du vers, comme séquence de segments non indépendants, le second étant plus régulier que le premier.

Revenons à la rime. L'analyse des réseaux de mots-rimes a permis d'examiner à nouveau la nomenclature des rimes. Pour la rime, nous avons été amenée à proposer une définition contextuelle de la richesse de la rime, en décalage très sensible par rapport aux théories dominantes, et cela grâce à l'examen des régularités sur les corpus. La longueur de la rime, et par conséquent la qualité de la rime, varient selon la voyelle, la qualité de la terminaison, la fréquence de la finale. Ainsi plus le vocabulaire présentant une terminaison donnée est rare (le vocabulaire étant ici évalué par rapport au corpus des mots à la rime, et non en tant que potentialité de langue), moins l'extension de la rime est grande. L'analyse systématique des corpus conduit à un réexamen en profondeur des théories, et cela est d'autant plus vrai pour des théories qui étaient difficiles à vérifier par un examen « à la main ».

Les outils confortent une partie des savoirs partagés tout en permettant de faire émerger des savoirs nouveaux. La validation, avec les outils, d'une grande partie des règles habituellement posées sur le vers donne une certaine confiance dans les régularités observées par ailleurs, qui n'ont pas forcément été vues antérieurement, et qui peuvent modifier les théories en cours.

4. Conclusion

Cette expérience d'analyse des aspects formels du vers avec des outils de TAL appliqués à de grands corpus s'est faite sous le signe de la diversité. Tout d'abord, par la diversité des outils de TAL mobilisés : plusieurs composantes linguistiques sont traitées (phonétique, syntaxique, prosodique, lexicale, sémantique) ; les outils ont des origines très diverses (maquettes, outils ad hoc, produit commercial, produit de laboratoire adapté sur mesure). Ensuite par la diversité des questions examinées : identification des syllabes métriques, figures des vers, traitement de la rime, analyse lexico-sémantique. Cette diversité trouve son sens dans le modèle hiérarchisé du rythme qui permet de faire dialoguer toutes ces composantes. Le pendant informatique de ce modèle est la base de données des vers enrichie par de nombreux traits affectant la syllabe, l'hémistiche, le vers, la paire de vers, la pièce entière. Cette représentation informatique prend en compte différentes dimensions constitutives du rythme. Cette structure permet aussi un

enrichissement par l'introduction de nouveaux traits ou marquages, qui peuvent par exemple résulter de traitements statistiques sur les marquages de premier niveau ou de l'accueil de traits nouveaux liés à la production ou à la réception.
Cette représentation, dans une base de données, autorise alors toutes formes d'explorations et de mises en relation. Comme le soulignaient Habert et Zweigenbaum (2002), « ces corpus enrichis permettent le test d'hypothèses sophistiquées mais aussi permettent de mettre en évidence des phénomènes et des corrélations inattendues ». C'est la force des bases de données que de permettre ces mises en relation.
Cette représentation est une structure ouverte disponible pour bien d'autres formes d'explorations. Celles-ci peuvent confirmer, compléter ou invalider les règles et les théories existantes. Elles doivent être le moteur pour l'élaboration de nouvelles propositions théoriques. A travers ces allers-retours entre théories, corpus et outils, se met en place une démarche expérimentale et cumulative : des hypothèses sont testées et validées, les résultats trouvés viennent enrichir la base de données et deviennent à leur tour des ressources pour d'autres explorations.

Bien des pistes restent à explorer et offrent d'importantes perspectives de recherche. L'amélioration et l'enrichissement des outils constituent une première piste : affiner l'analyse syntaxique, travailler sur des textes dans leur orthographe d'origine permettraient de fiabiliser davantage les résultats. L'extension des corpus, qui constitue la deuxième grande direction, permettrait de tester de nouvelles hypothèses sur les genres, sur les spécificités d'auteur, sur l'évolution du mètre et du rythme. Il serait intéressant de pouvoir travailler sur l'ensemble du théâtre du XVII^{e} en vers en confrontant les pièces qui ont traversé les siècles avec celles qui ont sombré dans l'oubli, tout comme il serait pertinent de suivre dans le temps l'évolution des modèles métriques. Enfin, il faudrait envisager de pouvoir explorer les influences croisées de la poésie européenne : les formes poétiques, comme le sonnet, ont beaucoup circulé ; les formes métriques dominantes ont été dans beaucoup de pays des formes importées (comme l'hendécasyllabe espagnol qui s'est inspiré du modèle italien). Cela impliquerait au préalable un vaste programme de numérisation de la poésie européenne et deuxièmement la mise au point d'outils équivalents au métromètre pour les autres langues.
C'est là que nous sommes à nouveau confrontés au principe de réalité : force est de constater qu'il n'existe que très peu d'outils d'analyse métrique du vers : le métromètre pour le français et l'outil de Gervas (2000) pour l'espagnol, qui ne traite pas les aspects syntaxiques. La mise au point de ces outils requiert des compétences informatiques importantes, que l'on trouve rarement chez les linguistes. Les associations entre linguistes et informaticiens restent donc incontournables. Si ces cas d'associations sont peu nombreux, c'est sans doute que l'intérêt pour la forme poétique ne concerne que des populations très restreintes. L'engagement d'informaticiens dans des travaux sur le vers relève davantage d'intérêts personnels que de l'intérêt de laboratoires pour qui le vers n'est pas un enjeu de recherche majeur.
La maîtrise des bases de données et la capacité à explorer ces bases sont des compétences indispensables pour les linguistes, plus incontournables encore que la capacité à développer des outils de TAL. Il s'agit bien de savoir mettre en place des processus de capitalisation. Les traits de description qui actuellement enrichissent le corpus doivent pouvoir être conservés, transmis et enrichis par de nouveaux traits. Ces bases de données sur le vers ont en effet vocation à être publiques, partagées et enrichies de manière coopérative.

Je tiens à remercier les deux relecteurs de cet article qui m'ont aidée à progresser dans la réflexion et dans la démonstration.

5. Bibliographie

Beaudouin V. & Yvon F. (1996) : "The Metrometer : a Tool for Analysing French Verse", *Literary & Linguistic Computing,* vol. 11, n°1, p. 23-32.
Beaudouin V. (2002) : *Mètre et rythmes du vers classique. Corneille et Racine*. Paris, Champion, coll. Lettres numériques.
Benvéniste E. (1966) : Les niveaux de l'analyse linguistique. *In* : *Problèmes de linguistique générale*. Paris, Gallimard, p. 118-131.
Billy D. (1984) : "La nomenclature des rimes", *Poétique,* n°57, p. 64-75.
Constant P. (1991) : *Analyse syntaxique par couche*, thèse en informatique de l'ENST, Paris, ENST.
Corneille P. (1629-1674) : *Œuvres théâtrales*, Ed. Ch. Marty Laveaux, Paris, Hachette, 1862.
Cornulier B. de (1982) : *Théorie du vers. Rimbaud, Verlaine, Mallarmé*. Paris, Éditions du Seuil.
Dell F. (1984) : *L'accentuation dans les phrases en français*. Paris, Hermann, 65-122.
Elwert W. T. (1965) : *Traité de versification française. Des origines à nos jours*. Paris, Klincksieck.
Gervás P. (2000) : A Logic Programming Application for the Analysis of Spanish Verse. *First International Conference on Computational Logic, Logic Programming Implementations and Applications stream*, Imperial College, London.
Gouvard J.-M. (1996) : « Le vers français : de la syllabe à l'accent », *Poétique*, n°106, p. 223-247.
Grammont M. (1908, 1965) : *Petit traité de versification française*. Paris, Armand Colin.
Gramont F. de (1876) : *Les vers français et leur prosodie*. Paris.
Habert B. & Zweigenbaum P. (2002) : « Régler les règles », *Traitement automatique des langues*, vol. n°43, n°3, p. 83-106.
Lusson P. (1973) : « Notes préliminaires sur le rythme », *Cahiers de poétique comparée*, vol I - fascicule 1, p. 30-54.
Lusson P. (1998) : « Une méthode d'analyse des rapports texte/musique : application d'une théorie générale du rythme », *Cahiers de poétique comparée, Mezura*, n°13, p. 7-45.
Martinon P. (1905, 1962) : *Dictionnaire méthodique et pratique des rimes françaises*. Précédé de *Versification française (Théorie et pratique) :* Paris, Librairie Larousse.
Meschonnic H. (1982) : *Critique du rythme*. Paris, Editions Verdier, 735 p.
Milner J.-C. & Regnault F. (1987) : *Dire le vers. Court traité à l'intention des acteurs et des amateurs d'alexandrins*. Paris, Seuil.
Milner J.-C. (1974) : « Réflexions sur le fonctionnement du vers français », *Cahiers de poétique comparée*, vol I, fasc. 3, p. pp 2-21.
Milner J.-C. (1987) : « Accent de vers et accent de langue dans l'alexandrin classique », *Cahiers de poétique comparée*, n°15, p. 31-77.
Milner J.-C. (1989) : *Introduction à une science du langage*. Paris, Éditions du Seuil.
Ousaka Y., Yamazaki M. & Miyao M. (1994) : "Automatic Analysis of the Canon in Middle Indo-Aryan by Personal Computer", *Literary and Linguistic Computing,* vol. 9, n°2, p. p. 125-136.
Purnelle G. (2002) : « Les vers semi-libre d'Yves Bonnefoy dans *Ce qui fut sans lumière »*, *Le français moderne*, Tome LXX, n°2, p. 145-168.
Quicherat L. (1838) : *Petit traité de versification française*, Hachette.
Racine J. (1664-1691) : *Œuvres théâtrales*, Ed. Paul Mesnard, Paris, Hachette, 1885.

Reinert M. (1993) : « Les "mondes lexicaux" et leur logique ». *Langage et société,* Paris, Maison des Sciences de l'Homme, n°66, pp. 5-39.
Robey D. (1993) : "Scanning Dante's Divine Comedy: a Computer Based Approach", *Literary and Linguistic Computing,* vol. 8, n°2, p. 81-84.
Roubaud J. (1978, 1988) : *La vieillesse d'Alexandre*. Paris, Éditions Ramsay, (éd. François Maspero 1978).
Roubaud J. (1986) : « DYNASTIE : études sur le vers français, sur l'alexandrin classique, Première partie », *Cahiers de poétique comparée*, n°13, p. 47-109.
Roubaud J. (1988) : « DYNASTIE : études sur le vers français, sur l'alexandrin classique, Deuxième partie », *Cahiers de poétique comparée*, n°16, p. 41-60.
Vaissière J. (1991) : Rhythm, accentuation and final lengthening in French. In: S. Johan, N. Lennart and C. R. (eds), *Music, Language, Speech and Brain*. Wenner-Gren International Symposium Serie. vol. 59, 108-120.
Yvon F. (1996) : *Prononcer par analogie : motivation, formalisation et évaluation*, thèse en informatique de l'ENST.

Valérie Beaudouin
Laboratoire Usages, créativité, ergonomie
France Télécom Recherche et Développement
38-40, rue du Général Leclerc
92794 Issy –les-Moulineaux cedex 9
tel : 01 45 29 62 52
valerie.beaudouin@francetelecom.com