L'alignement multicritères des documents médiévaux

Hatem Ghorbel, Giovanni Coray

École Polytechnique Fédérale de Lausanne, Faculté Informatique et Communications, Laboratoire d'Informatique Théorique, CH 1015 Écublens

André Linden, Olivier Collet, Wagih Azzam

Université de Genève, Faculté des lettres, Département MELA, 3, rue de Candole, CH 1211 Genève 4

ABSTRACT

The aim of text alignment is to establish correspondence relations between subparts of two or more translations or versions of the same document. The majority of the methods in use in the technique of alignment are based on the statistical analysis of word or character frequencies or of string occurrences. In order to improve the efficiency of the process of alignment, other methods have incorporated some structural properties of the documents (e.g. chapters, sections, paragraphs, etc.) as further criteria.

When applying the problem of alignment to parallel versions of medieval French manuscripts produced between the XIIth and the XVIth century, classical approaches have shown their limits due to the considerable variation of the appearance and content of these documents. This is basically caused by, (a) the partial evolution of the language, second, (b) the variation of the style (verse and prose) and (c) the various personal interpretations that could come about when rewriting new versions.

In this article, we adapt the technique of alignment to parallel versions of ancient texts and we propose a multicriteria approach which takes into account, first the similarities at the lexical, morpho-syntaxic and lexico-semantic levels, and second the similarities of the typographic and rhetorical structure of texts.

KEYWORDS : *multicriteria alignment, parallel versions of medieval manuscripts, linguistic similarities, structural similarities.*
RESUME

Le but de l'alignement des textes est la mise en correspondance des sous-parties similaires de deux ou plusieurs traductions ou versions d'un même document. La plupart des méthodes utilisées dans la technique de l'alignement se fondent sur l'analyse statistique des fréquences de mots ou de caractères, ou sur la cooccurrence des chaînes que ceux-ci composent. Afin d'améliorer l'efficacité de ces méthodes, d'autres approches incorporent certaines propriétés linguistiques (morpho-syntaxiques et lexico-sémantiques) et structurelles (marques de chapitres, de sections, etc.) des documents.

Lorsqu'on applique de telles techniques aux versions parallèles des manuscrits en langue française produits entre le XIIème et XVème siècle, celles-ci montrent leurs limites en raison de la variation considérable de l'aspect et du contenu de ces documents. Les causes en sont premièrement, l'évolution de la langue, deuxièmement, les possibilités de transformation stylistique des textes, enfin, les diverses interprétations personnelles auxquelles la réécriture de nouvelles versions peut donner lieu.

Cet article expose les résultats d'une adaptation de la technique d'alignement aux versions parallèles des écrits anciens à partir d'une approche multicritères qui tient compte d'une part de la similitude au niveau lexical, morpho-syntaxique et lexico-sémantique du français de la période médiévale et, d'autre part, de celle que l'on constate sur le plan de la structure typographique et rhétorique des textes.

MOTS-CLES : *alignement multicritères, versions parallèles d'écrits médiévaux, similitude linguistique, similitude structurelle.*

1. Introduction

L'étude des textes anciens et en particulier des documents de la période médiévale offre des perspectives nouvelles dans le domaine du traitement automatique de la langue. L'environnement d'édition réalisé dans le cadre du projet MÉDIÉVAL (Modèle d'Édition Informatisée d'Écrits médiévaux, Visualisés par ALignement) [AZZAM 01] permet ainsi aux étudiants, chercheurs et spécialistes du domaine d'effectuer à différents niveaux une étude comparative des variantes entre quelques versions manuscrites d'un texte français du XIVe siècle, l'*Ovide Moralisé*.

Lorsqu'on applique des techniques d'alignement – soit la mise en correspondance entre segments homologues de textes – aux versions parallèles des retranscriptions en langue française exécutées entre le XIIème et XVIème siècle, les approches classiques, qui reposent dans une large mesure sur des modèles statistiques (corrélation entre le nombre de caractères ou de mots d'un segment donné, etc.), montrent leurs limites en raison de la variation considérable de l'aspect et du contenu de ces documents. Les causes en sont

premièrement, l'évolution de la langue, deuxièmement, les possibilités de transformation stylistique des textes, enfin, les diverses interprétations personnelles auxquelles la réécriture de nouvelles versions peut donner lieu.

En dépit de ce facteur, les copies d'écrits médiévaux partagent plusieurs traits communs, soit sur le plan linguistique (morpho-syntaxique, lexical et sémantique), soit au point de vue structurel et organisationnel (*i.e.* la manière dont les segments sont organisés pour contribuer au sens global du texte). Ces similitudes peuvent donc être exploitées en vue d'une comparaison et, avec la base de données lexicale qui complète l'éditeur, elles procurent l'arrière-fond de la technique d'alignement que nous exposons dans cette article.

L'éditeur a aussi été testé sur d'autres documents, en français contemporain, à partir d'un choix de textes législatifs multilingues. Les méthodes utilisées pour l'alignement des écrits anciens ne s'appliquent donc pas seulement dans un contexte intra-linguistique mais également à des traductions.

2. L'étude comparative des textes médiévaux : un problème nouveau pour l'alignement

Au départ, l'alignement a été conçu comme une aide à la traduction automatique. Toutefois, cette technique a rapidement été utilisée pour d'autre tâches de manipulation et d'exploitation des documents multilingues [CATIZONE 89, BROWN 91, GALE 91, DAGAN 96, VERONIS 00]. Elle s'applique alors aux versions de textes traduits d'une langue à une autre. Aligner deux textes équivaut à mettre en correspondance les parties homologues de ces derniers – paragraphes, phrases, expressions, mots – qui représentent des équivalents directs. Ce problème a été traité à partir de méthodes statistiques qui considèrent essentiellement les textes comme des flux de caractères et de mots [BROWN 91, GALE 91]. La comparaison, avant tout statistique, d'après la longueur des segments analysés, s'avère en effet suffisante dans certains cas de traduction, en particulier avec les langues européennes. Elle donne par exemple de bons résultats lorsqu'elle s'applique à des textes en langage courant tels que les documents législatifs ou techniques [GALE 91, BALLIM 98].

Un alignement fondé sur de telles méthodes révèle toutefois ses limites dès lors qu'il s'agit d'appréhender des phénomènes linguistiques plus complexes et diversifiés, comme la réécriture et la ré-interprétation des textes, anciens ou modernes. Les insuffisances sont particulièrement flagrantes dans le cas des documents médiévaux. Écrits à une époque où la langue n'était pas encore fixée et transmis par des moyens artisanaux (les manuscrits), de tels textes offrent en effet une grande diversité d'aspects et de contenu ainsi qu'une grande quantité de variantes sur le plan linguistique. Un traitement purement statistique de leurs particularités se heurte très vite à des impasses et d'autres approches doivent être choisies, en particulier pour ce qui concerne l'étude de leur langue et de leur structure.

Les documents en ancien français posent donc un problème nouveau pour l'alignement et de fait, très peu de travaux ont abordé le problème de la réécriture dans un contexte intra-linguistique, qu'il s'agisse d'adaptations, de conversions ou d'interprétations [OWEN 98].

3. Le traitement des documents en ancien français

Les textes en ancien et en moyen français ouvrent des perspectives prometteuses pour le traitement électronique des documents et pour la recherche de nouvelles méthodes de comparaison des documents. En effet, les écrits de cette période comportent non seulement de l'information, mais ils reflètent aussi un univers de pensée textuaire différent de celui auquel l'imprimerie nous a habitué. L'environnement informatique de comparaison des manuscrits médiévaux développé dans le cadre du projet MÉDIÉVAL se propose ainsi d'offrir aux médiévistes des outils de navigation dans les différentes versions des textes transcrits en français entre le XIIe et le XVIe siècle. Cette interface pour l'édition et pour l'exploration des documents vernaculaires du moyen âge offre également des outils de comparaison microscopique du contenu textuel qui recourent aux techniques d'alignement des textes.

La conception de cet environnement s'appuie sur un certain nombre d'évidences qui concernent la diffusion traditionnelle des textes anciens :

- Par l'absence systématique d'original jusque vers la fin du XIVe siècle au moins et par une diffusion garantie au moyen de copies manuscrites, multiples et très mouvantes dans leur contenu et leur aspect, les œuvres du moyen âge s'inscrivent dans un espace très particulier, à la fois pluriel, virtuel et matériellement hétérogène, puisque leurs supports associent souvent texte et image ou décoration à travers une série de liens interprétatifs.

- L'imprimé, mode le plus courant de sa publication actuelle, ne permet de rendre compte, au mieux, que du contenu verbal de l'écrit médiéval; il en gomme généralement la structure profonde et le contenu iconographique, et « aplatit » les textes eux-mêmes en les réduisant à une dimension exclusive (restitution d'un état unique et figé; rejet des états variants en notes). Cela prive l'étude linguistique et littéraire du moyen âge d'une part d'information indispensable à la compréhension des textes eux-mêmes, de l'histoire littéraire et de l'histoire de la langue ancienne.

3.1 Le projet MÉDIÉVAL

L'environnement de comparaison des documents manuscrits développé dans le cadre du projet MÉDIÉVAL (Modèle d'Édition Informatisée d'Écrits médiévaux, Visualisés par ALignement) [AZZAM 01] doit ainsi permettre aux spécialistes et aux chercheurs des différents domaines concernés de procéder à des études comparatives sur les origines et sur les propriétés linguistiques des écrits en ancien et en moyen français. Il vise en outre à pallier les difficultés résultant de la diversité de tels matériaux grâce à une navigation dans les différentes versions des œuvres (comparaison macroscopique) ou entre des éléments textuels et structurels similaires (comparaison microscopique).

A l'échelle macroscopique, cette navigation se fonde sur l'extraction des catégories marquées pendant la phase d'annotation philologique. Cette étape, manuelle, offre le moyen d'enrichir le contenu des écrits par des interprétations, tout en conservant leurs attributs originaux.

Sur le plan microscopique, l'alignement des textes est la principale méthode de comparaison entre ces derniers. La nature des documents anciens impose toutefois une mise en œuvre dans un contexte intra-linguistique. L'opération consiste alors à détecter les similitudes linguistiques et structurelles entre les versions comparées.

Avant de procéder à leur traitement informatique, il est nécessaire de convertir les textes médiévaux en format numérique. Cette phase doit non seulement transmettre le contenu verbal de l'écrit, mais aussi conserver ses traits et attributs originaux. Elle doit tenir compte en particulier de l'état de conservation des manuscrits, des caractéristiques de mise en page des textes et de leur iconographie ou de la présence d'ornementations, de la difficulté de segmentation au niveau des lettres et des mots, de la présence d'abréviations, plus ou moins fréquentes, qui ne correspondant pas à des caractères ordinaires et ne peuvent être résolues qu'en fonction d'un certain nombre de critères linguistiques et contextuels; enfin de l'absence de ressources lexicales et grammaticales susceptibles de guider la reconnaissance.

Une transcription manuelle et une annotation des extraits choisis doivent donc être préalablement confiées aux experts du domaine.

3.2 Comparaison des textes par alignement

L'alignement permet la mise en correspondance des parties textuelles de versions parallèles. Par versions (ou documents) parallèles, nous entendons non seulement des traductions diverses, mais aussi des réécritures et des ré-interprétations à un niveau susceptible d'autoriser une comparaison. Appliqué aux textes médiévaux, l'alignement a pour objectif le développement d'un environnement informatique permettant la navigation et le rapprochement entre plusieurs copies ou rédactions d'une œuvre.

Les premières expériences réalisées dans le cadre du projet MÉDIÉVAL ont consisté à adapter les méthodes actuelles d'alignement au corpus sélectionné afin d'évaluer la difficulté et la faisabilité de cette tâche. Différentes adaptations de l'algorithme de Gale et Church ont été appliquées à cette fin aux versions parallèles extraites de quatre manuscrits en vers et d'une adaptation en prose de l'*Ovide moralisé* (XIVe siècle). Ces expériences ont montré, entre autres, qu'un algorithme d'alignement monotone basé sur la longueur des chaînes de caractères reste utilisable tant que la comparaison ne porte que sur des échantillons relativement homogènes (vers-vers). Ceci est dû en partie au fait que les variations qui se produisent au niveau microscopique n'altèrent pas ou ne modifient que peu la taille des segments considérés. Dans le corpus analysé, les inversions de vers sont en outre rares. Au niveau macroscopique, l'alignement doit tenir compte en plus d'éventuelles insertions ou suppressions des vers et de la présence des rubriques (titres de chapitres ou de sections du texte). Puisque celles-ci apparaissent dans le modèle de document comme des éléments distincts des vers eux-mêmes, l'alignement doit opérer la distinction nécessaire afin d'éviter une mise en correspondance fautive. L'algorithme doit donc être sensible au modèle de document et calculer l'alignement en respectant la structure de ce dernier.

3.3 Alignement vers-vers

Un alignement des extraits en vers a été réalisé avec l'aligneur TALCC [BALLIM 96, 98], système qui tient compte de la structure des documents et qui se fonde sur l'algorithme

de Gale et Church dont le critère d'appariement est la longueur des chaînes de caractères. Les essais ont donné de bons résultats (plus de 99% de réussite). A condition de modéliser la variation au niveau microscopique et d'approfondir l'analyse linguistique des documents, il serait même concevable d'affiner la correspondance entre les composants des textes et de l'appliquer aux syntagmes, voire aux mots.

3.4 Alignement prose-vers

Pour l'alignement de ces extraits avec leur équivalent en prose, nous disposons d'une des deux rédactions qui ont été dérivées au XVe siècle de l'*Ovide Moralisé* en vers. Entre ces deux état de texte, on remarque tout d'abord une différence au niveau de la structure typographique. Les rédactions versifiées sont en effet organisées comme une suite de segments de la même taille et de la même structure (couplets d'octosyllabes). Aucune contrainte linguistique ne pèse de manière systématique sur le contenu des vers. L'objectif principal d'une telle disposition est d'assurer la mise en forme poétique – d'un point de vue à la fois musical et rythmique – du contenu verbal. Si la structure des unités de texte est rendue explicite par l'emploi de différents marqueurs au niveau macroscopique (miniatures, initiales et lettrines), l'absence presque complète de ponctuation ne rend pas celle des phrases manifeste. Dans l'adaptation en prose, la structure linguistique est plus évidente. La syntaxe est soumise à des contraintes qui excluent certains agencements. La ponctuation, même si elle est peu fiable, tant à cause des difficultés de lecture des manuscrits que de l'inconsistance de ses règles d'application, permet néanmoins une segmentation approximative du texte.

Pour comparer les extraits en vers et en prose il convient tout d'abord de décider de la granularité de l'alignement, c'est-à-dire de la nature des éléments à mettre en correspondance. Ces segments doivent être compatibles et cohérents afin de déterminer un modèle de substitution du type (M : N) qui met en relation M parties du texte source avec N parties du texte cible. L'absence de segmentation explicite et systématique dans les écrits médiévaux (en dehors du découpage en vers des écrits poétiques qui adoptent ce dispositif) représente une difficulté supplémentaire. Diverses expériences d'alignement, qui tiennent notamment compte de la structure syntaxique des phrases et de celle des paragraphes, ont été menées à bien pour résoudre ce problème. Toutefois, ces essais n'ont pas produit d'issues satisfaisantes.

4. Approche linguistique et structurelle pour l'alignement

La nature des documents étudiés ôte une grande partie de sa pertinence à une approche fondée sur la taille de la phrase. Cette technique n'offre de bons résultats que lorsqu'il s'agit d'aligner des traductions dont la taille des phrases traduites est corrélée, comme c'est le cas avec les langues européennes. Les variations de longueur résultant par exemple d'une différence de style et de schéma formel (vers et prose dans le cas particulier), ou de genre littéraire invitent à la recherche d'autres méthodes fondées plutôt sur le contenu des textes, voire sur l'étude de leurs propriétés linguistiques.

L'approche structurelle reste néanmoins intéressante. En effet, la comparaison à ce niveau permet de diminuer la perplexité du calcul d'alignement. Les modèles existants

doivent cependant être réévalués et enrichis au moyen d'un certain nombre d'apports liés non seulement à la forme mais aussi au contenu sémantique des écrits.

4.1 Approche linguistique

Linguistiquement parlant, le but poursuivi est d'obtenir une meilleure appréhension des critères de comparaison dont l'alignement se nourrit au niveau des mots ou des expressions. L'analyse linguistique permet d'associer deux ou plusieurs entités et de les rattacher à une même base – morphologique, morpho-syntaxique, etc. – ou à deux syntagmes que l'on suppose unis par une même fonctionnalité ou valence sémantique, deux unités signifiantes, etc.

Le cas des documents médiévaux représente sans doute l'une des situations les plus complexes que l'on puisse imaginer. En effet, la tradition écrite de la période ancienne procède d'un état langue dont les traits essentiels sont encore loin d'être fixés par la grammaire. Les usages du français médiéval ne sont pas stables et évoluent sur tous les plans avec une beaucoup plus grande rapidité qu'aujourd'hui. De plus, cette évolution n'est pas géographiquement uniforme.

Les variantes observées entre deux représentants de la tradition documentaire du moyen âge expriment ainsi un ensemble de faits non seulement linguistiques, mais aussi géolinguistiques, culturels et historiques. Cet état de fait impose d'élargir l'enquête à tous les niveaux qui concernent l'étude de l'ancien et du moyen français. Dans l'idéal, l'approche définie au sein du projet MÉDIÉVAL consisterait donc à décrire toutes les occurrences des constituants lexicaux du corpus en explicitant leurs différentes propriétés – grapho-phonétiques et morpho-syntaxiques, mais aussi lexico-sémantiques, en tenant compte des critères chronologiques et géo-linguistiques qui infléchissent ces propriétés –, de sorte à enrichir les processus de segmentation et d'alignement, à en améliorer les résultats et à permettre une navigation comparative entre les différents états de textes. L'extension d'une telle démarche pourrait d'ailleurs aboutir à la création de bases de connaissances du plus haut intérêt dans les domaines grammaticaux et lexico-sémantique et servir de support à de nombreux instruments d'analyse.

Les caractéristiques les plus immédiatement saisissables, à savoir morphologiques (ou morpho-syntaxiques), sont celles qui ont été privilégiées. En dépit des obstacles qui s'opposent à une formalisation de l'ensemble des règles nécessaires à décrire la langue médiévale sous ce point de vue, il est relativement aisé de dégager quelques traits pertinents d'une analyse formelle à l'échelle du mot, tandis que les tentatives d'exploitation des textes à partir d'une analyse de leur structure grammaticale – sans parler de leurs aspects sémantiques ou stylistiques ! – se heurtent quant à elles à des difficultés de très haut niveau, ne serait-ce que pour la description des unités pertinentes du discours. L'analyse morphologique est soumise à quatre critères qui déterminent les variantes de surface des mots :

– critère grapho-phonétique, qui concerne les variations issues de l'évolution des sons de la langue (plan diachronique) ou résultant des combinaisons à l'intérieur d'une suite de phonèmes (plan synchronique);

– critère analogique, qui concerne les variations produites par l'influence de certaines formes sur d'autres, au sein d'un paradigme ou entre deux paradigmes;

– critère orthographique, qui concerne les variations dont l'origine ne peut être imputées à aucune autre cause que l'absence des normes fixes de transcription dans les textes médiévaux;

– critère dialectal, qui concerne les variations relevant des particularités de *scripta* propres à une région.

4.1.1 MMORPH et l'analyse morphologique

A partir de ce classement, l'analyseur morphologique MMORPH [PETITEPIERRE 95] du programme MULTEXT a été mis à contribution pour générer une première base de données lexicales et un ensemble de règles structurelles. MMORPH est un analyseur qui permet d'obtenir de tels résultats à partir d'un formalisme grammatical à deux niveaux, équivalant, sur le plan morphologique, à l'analyse générative transformationnelle classique pour la syntaxe. Le programme établit la correspondance d'une forme de mot avec une entrée d'une base de données qui contient une description morpho-syntaxique de celui-ci. Cette description unit deux composantes : un type et un ensemble de caractéristiques exprimées en termes d'attributs-valeurs. Le type décrit en général la classe ou la catégorie syntaxique (par exemple le nom, le verbe, etc.). D'autres distinctions plus fines – telles que le temps, le mode ou la personne d'une forme verbale, etc. – sont apportées par la structure d'attributs-valeurs (par exemple temps=présent, personne=1, etc.).

La difficulté majeure rencontrée avec MMORPH n'est pas directement liée à ce programme mais plutôt au formalisme adopté : l'analyse à base de règles. En effet, le nombre élevé de cas isolés, d'ambiguïtés et de non-déterminisme dans la morphologie de la langue médiévale rendent très complexe, pour ne pas dire impossible, la conception d'un système de règles exhaustives qui décrirait toutes les entités et leurs caractéristiques. Un autre des problèmes constatés lors de cette étape est la surgénération de formes théoriques plausibles en elles-mêmes mais qui n'existent pas dans le corpus et donc sujettes à provoquer des confusions ou des analyses erronées. Pour pallier ce risque, le lexique d'entrée doit parfois être spécifié sous forme d'entrée lexicale directe, ce qui augmente la complexité et le coût de la tâche.

4.1.2 La base de données MEDIEVLEX

MMORPH n'a donc été utilisé que pour la génération des formes simples des verbes, des noms et des adjectifs. Les cas de variations isolées ont été enregistrés manuellement. L'enrichissement du lexique avec les autres types de mots a été réalisé de la même manière à partir d'une extraction automatique du corpus.

Par ailleurs, MMORPH possède une structure de base de type plat. Elle n'opère donc aucune hypothèse de regroupement des formes qui partagent certaines communautés. Cette caractéristique restreint considérablement les possibilités d'enrichissement à partir de l'ensemble du lexique, tant à cause des difficultés que l'on rencontre au niveau de la construction qu'au point de vue de la manipulation des informations. A l'évidence, il

s'avérerait préférable de regrouper par exemple les formes conjugées d'un verbe à partir de catégories déterminées (mode, temps, personne, etc.), ce qui rendrait l'élaboration moins coûteuse en cas d'introduction manuelle. Il en va de même pour la recherche, l'accès et pour tous les types de réutilisations possibles.

Une des structures apte à remédier – en partie – à ces difficultés est la disposition hiérarchique des données qui autorise une classification des formes du général au spécifique et ainsi, une mise en commun de certaines caractéristiques pour des classes d'entités particulières. Le modèle le plus apte à représenter cette classification est celui des documents structurés. Le langage de marquage XML (Extended Markup Language) permet de décrire l'agencement hiérarchique de la base lexicale d'une façon optimale et aisée à construire.

Une fois le modèle choisi, il reste à résoudre le problème des niveaux hiérarchiques en fonction des traits morphologiques envisagés plus haut. Le regroupement doit être optimal pour minimiser l'effort de construction et pour garantir l'efficacité des automates d'accès, de gestion et de recherche. Plusieurs classifications sont possibles, notamment au niveau des catégories syntaxiques (nom, verbe, adjectif, etc.) et de leurs caractéristiques internes (mode, temps, personne pour les verbes, etc.). Il est également envisageable de créer différentes classes plus abstraites pour regrouper les sous-catégories les plus fines.

Plusieurs travaux de ce type ont abouti à la fabrication de modèles optimaux et efficaces. Citons par exemple la base lexicale du français moderne développée dans à l'intérieur du projet voué à la construction et à l'exploitation d'un large corpus de données lexicale pour le français et l'anglais au Centre National de la Recherche Scientifique (CNRS) [IDE 95, ERJAVEC 00]. En nous inspirant de cette catégorisation fondée sur la structure de traits, nous proposons un modèle spécifique de base lexicale, baptisé MedievLex, qui se présente comme suit :

– * Elements :

  Famille = Catégorie *

  Catégorie = (Morpho + , sens)

  Morpho = Surface +

  Surface = CDATA

– * Attributs :

  Famille : Nom

  Catégorie : Type, Lemme, Groupe, Intransitif, Transitif_direct, Bitransitif

  Morpho : Mode, Temps, Personne, Nombre, Genre, Cas, Rhet

  Surface : Variation

– La notion de la famille repose sur les origines étymologiques des mots.

- Les lemmes sont les formats canoniques du lexique (infinitif pour les verbes, masculin singulier pour les noms à l'exception des substantifs essentiellement féminins, pour les adjectifs, pronoms et déterminants).

- Les attributs « Groupe, Intransitif, Transitif_direct, Bitransitif, Mode, Temps et Personne » se rattachent uniquement aux verbes (la différenciation du rang personnel concerne toutefois aussi les pronoms).

- L'attribut « Cas » concerne les noms, les adjectifs, les pronoms et les déterminants.

- L'attribut « Rhet » concerne uniquement les complémentaires, les propositions et les adverbes. Il décrit la fonction du mot au point de vue rhétorique.

4.2 Approche structurelle

Les documents médiévaux doivent être appréhendés comme un tout. A la dimension écrite – ou imprimée – s'ajoutent les particularités (typo)graphiques des manuscrits : enluminures, initiales, lettrines, pieds-de-mouche, signes diacritiques, qui marquent l'organisation des épisodes, paragraphes et phrases, chacune contribuant à l'organisation structurelle des documents et à la compréhension du contenu verbal. Un alignement de textes anciens doit, en plus de l'analyse morpho-syntaxique, tenir compte de ces différentes propriétés.

Toutefois, un découpage en accord avec la structure typographique des manuscrits n'est pas fiable. D'une part, les signes de ponctuation et autres signes graphiques ou marqueurs de structuration n'expriment pas toujours la même fonction, ils ne figurent pas dans tous les exemplaires, ne répondent pas à la même distribution ou ne possèdent pas la même densité, et ainsi de suite; d'autre part, quand ils s'y trouvent, ils peuvent dans certains cas induire une segmentation incompatible avec la structure linguistique de la phrase. Par conséquent, il faut envisager d'autres paramètres pour vérifier et corriger la disposition des textes, notamment au niveau des phrases et des paragraphes.

Pour intégrer l'ensemble de ces caractéristiques, linguistiques, organisationnelles et typographiques, dans une structure hiérarchique unifiée, susceptible d'une construction automatique, il est nécessaire de recourir à un modèle abstrait. Celui qui a été élaboré au cours du projet MEDIEVAL en vue de l'alignement automatique des écrits anciens tient compte de leur segmentation en phrases à partir des données grammaticales, philologiques et codicologiques. Ce modèle est ensuite enrichi par la description rhétorique du contenu discursif.

4.2.1 Le problème de la segmentation

La segmentation des textes médiévaux en une structure de phrases peut être induite à partir d'un certain nombre d'observations sur des connaissances linguistiques superficielles (morpho-syntaxiques, par exemple), sur leurs contenus lexicaux, ou sur la nature typographique des manuscrits. Ces observations peuvent être recensées à partir de corpus segmentés manuellement qui constituent des données d'entraînement et permettent d'alimenter un modèle de distribution probabiliste dont les paramètres sont ajustés au cours

du processus d'apprentissage afin de déterminer l'existence ou non d'un délimiteur de segment à chaque emplacement du texte. Celui qui est proposé dans le cadre du projet MEDIEVAL est un modèle probabiliste de la famille exponentielle [BEEFERMAN 99] qui permet de représenter au mieux la distribution des limites entre segments dans les textes. Les différents paramètres sont calculés à partir des observations recueillies lors de l'entraînement. Les détails techniques de se modèle ne sont pas l'objectif de cet article. Ils peuvent être consultés dans [GHORBEL 02_2].

Le processus d'apprentissage se fonde sur le recours à un mélange de traits, morpho-syntaxiques (tels que des combinaisons de k catégories lexicales), lexicaux (tels que les combinaisons de k mots) et typographiques (tels que l'existence de majuscules, de signes de ponctuations, de marqueurs supra-segmentaux, etc.).

L'entraînement s'est fait à partir d'un corpus de 1500 mots pour la version en prose et de 1500 mots pour celle en vers. L'évaluation a été accomplie en termes de rappel et de précision sur des *corpora* de test segmentés manuellement en phrases de la même taille que celles du modèle d'apprentissage. Le taux de succès est estimé – en termes de rappel et de précision – à 75%.

La segmentation en phrases des textes médiévaux a nécessité le calcul de deux modèles séparés, un pour la version en prose, un autre pour la rédaction en vers. La segmentation en paragraphes qui nécessite beaucoup plus de données d'entraînement n'a pas été évaluée en raison de la taille limitée des échantillons disponibles. Sur la base de ces découpages, il a été possible de (re)constituer la structure des textes selon le modèle abstrait unifié discuté plus haut. Un des problèmes rencontrés au cours de ce travail est que la taille des segments semble parfois trop large. Dans le but d'affiner la granularité de l'alignement – ces segments étant destinées à jouer le rôle d'unités élémentaires dans le processus de comparaison –, un travail préalable d'analyse syntaxique superficielle visant à détecter des unités syntagmatiques (par exemple, des syntagmes nominaux, verbaux et adjectivaux) a été accompli. Un analyseur syntaxique par coin gauche en Prolog, LHIP [BALLIM 94], a été utilisé à cette fin. Cet essai n'a abouti qu'à des résultats partiels et exigerait de plus amples approfondissements.

4.2.2 Structure rhétorique

D'autres informations contextuelles, d'ordre organisationnel et rhétorique, ont été ajoutées aux critères structurels d'alignement. A cette fin, différentes approches relevant de la théorie du discours, et en particulier la Théorie de l'organisation rhétorique (en anglais : Rhetorical Structure Theory -RST) de Mann et Thompson [MANN 87, 88] ont été mises à contribution. Celle-ci offre des modalités plausibles pour l'amélioration des algorithmes d'alignement des écrits médiévaux. Appliquée à de tels textes, elle offre des résultats prometteurs grâce aux similitudes qui peuvent être détectées en rapport avec la cohérence, la cohésion et le rôle de ces constituants.

La RST procède par segmentation du texte en unités rhétoriques élémentaires. A chaque unité est attribué un rôle ou une fonction. Ce résultat est obtenu principalement par la liaison des parties délimitées au moyen de relations et par leur agrégation en segments composés. La structure globale du texte se présente sous une forme hiérarchique et arborescente, les feuilles constituant les parties élémentaires et chaque nœud, les relations

qui unissent de manière récursive les parties qui y conduisent (*cf.* fig. 1). La nature des relations est de type rhétorique au départ. Toutefois, l'ensemble des relations est laissé ouvert et aucune contrainte ne pèse sur leur nombre ou sur la taxonomie adoptée. La RST définit également la notion de noyau et de satellite pour décrire le rôle d'un élément dans le texte par rapport à un autre. Le segment noyau joue un rôle plus important dans la compréhension du texte que le segment satellite qui revêt une fonction secondaire (reformulation, explicitation, adjonction de détails ou description du noyau).

Afin de structurer les documents selon le modèle RST, une étape d'annotation rhétorique s'avère indispensable. La nature de cette description, qui est par définition une tâche subjective et dépendante de l'interprétation humaine, invite à la recherche de solutions capables d'en accroître l'objectivité et le rendement. Le schéma général d'annotation proposé – RST-Light –, version simplifiée et adaptée de RST, permet d'alléger cette tâche et de réduire à la fois la complexité de l'intervention humaine, l'aléatoire ou les distorsions que comporte cette opération, sans pour autant négliger le concept de base du modèle original. A cet effet, il convient :

- de se limiter à un ensemble fini mais cohérent de relations rhétoriques afin de restreindre l'espace de choix de l'annotateur, d'homogénéiser les résultats et d'éviter les conflits sans pour autant condamner la description à un résultat trivial;

- de ramener les segments élémentaires du modèle aux phrases, conformément à l'étape de la segmentation;

- de représenter chaque paragraphe par un arbre rhétorique indépendant.

L'hypothèse fondamentale du modèle est que des textes parallèles du genre de ceux qui composent le corpus de l'*Ovide moralisé* (traductions, réécritures, conversions, etc.) présentent des structures rhétoriques comparables dans une large mesure. Il suffit donc en principe de décrire cette architecture pour chaque texte puis, sur la base de cette représentation, de considérer les similitudes au niveau de la fonction rhétorique des éléments dans le processus d'alignement.

Tous les aspects – linguistiques, organisationnels (ou rhétoriques) et typographiques – peuvent être intégrés, en respectant la compatibilité de ces éléments entre eux, au sein d'un modèle abstrait fondé sur un schéma RST-Light et adapté à l'alignement des textes médiévaux.

Un annotateur semi-automatique (RhetAnnotate[1]), élaboré d'après RST-Light, a également été mis au point. Il intègre le module de segmentation probabiliste généré par la phase d'apprentissage, un étiqueteur de discours pour le français médiéval, un module d'accès à la base de données MedievLex et un outil d'accès à WordNet. RhetAnnotate permet de proposer parmi l'ensemble des relations rhétoriques possibles celles qui sont les plus vraisemblables. A cette fin, il exploite une liste de mots-clé élaborée au cours d'une étude du corpus, susceptible d'un enrichissement en fonction des besoins et adaptable à

[1] *Cf.* http://lithwww.epfl.ch/~ghorbel/rhetannotate.

d'autres langues ou états de langue. RhetAnnotate permet également de manipuler avec beaucoup de souplesse la structure arborescente des documents (création, modification, suppression des nœuds de relations, des arbres, etc.) tout en contrôlant sa validité.

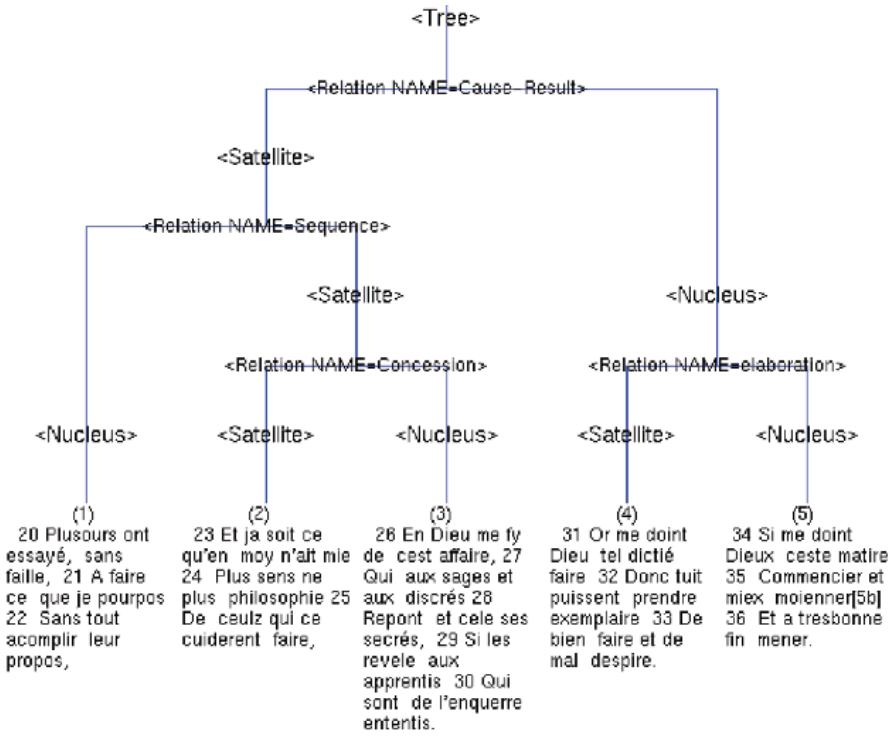

Figure 1 : La structure RST appliquée aux textes médiévaux en vers

Compte tenu de l'interaction complexe de la structure rhétorique avec des notions linguistiques, discursives et pragmatiques profondes, l'automatisation complète de cette procédure se heurte toutefois à des difficultés considérables.

Un espace de description multidimensionnel – structurel, morpho-syntaxique, lexical, sémantique, rhétorique – est ainsi créé, permettant un alignement multicritères des documents médiévaux.

En résumé :

– Au point de vue linguistique, l'intégration de règles morpho-syntaxiques et typographiques propres aux textes élaborés et transcrits à l'époque médiévale permet de préciser le calcul de la similitude des mots apparentés. La similitude de deux mots est donc évaluée en fonction du nombre de caractères qu'ils partagent, de leur appartenance lexicale (à la même famille), de leurs propriétés morpho-syntaxiques et de leur sens. Au niveau des groupes de mots (syntagmes), la similitude lexicale est calculée à partir de la

distribution des mots similaires et de la nature des unités syntaxiques auxquelles ils appartiennent. Au niveau des phrases, la similitude lexicale est calculée d'après la distribution des syntagmes similaires qu'ils partagent.

– Sur le plan organisationnel, la structure rhétorique permet d'établir des règles pour mesurer la similitude entre deux segments, notamment au niveau des phrases et des paragraphes. Ceci peut être effectué grâce à la structure arborescente qui reflète la cohésion et la cohérence entre les parties du texte. Par conséquent, la similitude des phrases est mesurée selon la nature des relations qui les unissent aux phrases voisines et de leur statut (satellite, noyau). La similitude des paragraphes, d'autre part, est évaluée d'après la similitude des chemins parcourus tracés par les arbres et en fonction des relations qui composent ces derniers.

5. Alignement multicritères des textes anciens

Afin de décider si deux éléments participant non plus de versions homogènes (ou relativement proches) mais hétérogènes sont plus ou moins semblables, il convient de déterminer les méthodes de mesure de similitude adéquates. Celles-ci, nous l'avons vu, ne peuvent se ramener au seul critère de la longueur des segments comparés, mais prennent place dans un espace de caractéristiques multidimensionnnel. A partir d'un ensemble de fonctions heuristiques de comparaison qu'on appelle fonctions de similitude, dont chacune fournit une évaluation de la proximité de deux éléments par rapport à un critère donné, la similitude globale est définie comme une combinaison linéaire pondérée de ces calculs.

La définition des valeurs sur lesquelles repose la relation d'alignement entre les éléments de deux textes est donc considérée comme un problème combinatoire où l'on cherche à établir une mise en correspondance qui maximise la similitude globale. Cette mesure se fonde sur deux critères principaux : linguistique et structurel. C'est surtout le premier de ces critères que nous discuterons ici.

5.1 Critères linguistiques pour l'alignement des textes anciens

La similitude entre plusieurs données textuelles est en général évaluée grâce à des méthodes de comparaison au niveau des flux de caractères qui composent les textes. Pour l'essentiel, le rapprochement s'opère à travers des heuristiques statistiques qui découlent soit du nombre de caractères lui-même, soit de la quantité de mots, soit de l'analyse des séquences de caractères ou de mots qui composent le texte. Dans les documents multilingues, la notion de mots apparentés (*cognates* en anglais) est souvent utilisée pour définir des points d'ancrage propres à faciliter l'alignement ou à fournir une heuristique au calcul de la similitude entre deux éléments. Deux mots sont considérés comme apparentés dès que cette évaluation atteint un seuil donné. Ce niveau est fixé expérimentalement et dépend fortement des langues.

Etant donné la forte corrélation que l'on observe entre leurs lexiques, la technique des mots apparentés s'adapte également aux versions parallèles des écrits médiévaux. Toutefois, pour de simples traductions, elle se limite ordinairement aux traits de surface des mots (formes graphiques) et ne prend pas en considération d'autres propriétés linguistiques

qu'ils sont susceptibles de partager en dépit des variations imposées par la syntaxe, par le style du texte, etc. Dans le cas des documents anciens, l'importance des variations graphiques et donc le risque de non reconnaissance entre deux formes similaires, de même que l'intérêt général de ce type d'informations poussent à l'élaboration de fonctions de similitude capables d'affiner la détection des mots apparentés au niveau morpho-syntaxique, lexical et sémantique. Elles permettent d'établir une fonction de similitude globale $\boldsymbol{f_1}$ que nous définissons en fin de cette section.

5.1.1 Similitude lexicale : la notion des mots apparentés

Un paire de mots apparentés est formée d'un vocable dans une langue A et d'un autre vocable dans une langue B analogue au précédent sur le plan formel et sémantique. Ainsi, *thèse* en français et *thesis* en anglais peuvent être considérés comme des mots apparentés. Quand deux termes ne présentent qu'une similitude formelle, comme par exemple *library* (« bibliothèque » en français) et *librairie* (« bookshop » en anglais), on les appelle des faux amis.

Plusieurs heuristiques existent pour détecter les mots apparentés, toutes fondées sur le calcul de distance ou de similitude au niveau graphique. Au nombre des différentes méthodes existantes (SIMARD 92, HOFLAND 98, MCENERY 95, MELAMED 99, etc.), nous avons privilégié le rapport que nous appelons "Dice_1" (*cf.* formule 1), variante du rapport de Dice proposée par SIMARD *et al*. Prenant appui sur les monogrammes plutôt que sur les bigrammes pour calculer la similitude des mots, ce rapport peut aussi être appliqué à la détection des mots apparentés :

$$Dice\_1 = \frac{2*a}{length(w_1)+length(w_2)} \qquad (1)$$

où a représente le nombre de caractères identiques qui se trouvent à la fois dans le mot w_1 et dans w_2, et la fonction length exprime la longueur en termes de chaîne de caractères. D'une manière générale, les rapports de Dice permettent d'inclure les variations graphiques et s'adaptent donc bien aux textes médiévaux dans lesquels ces transformations interviennent surtout localement (permutation de certains caractères ou groupes de caractères pour des raisons phonétiques, morpho-syntaxiques, dialectales ou arbitraires, comme dans les couples *accomplir - acomplir*, *doinst - doint*, *enchaint - enceint*, *repondre - respondre*, etc.). Malgré l'incertitude qui pèse sur les causes propres de chaque alternance et sur leurs contextes d'application, les heuristiques avec la méthode des rapports de Dice procurent de bons résultats.

Afin d'améliorer le rappel dans la détection des mots apparentés, nous avons choisi le rapport de Dice_1 avec un taux de 0.8. Toutefois, la précision s'est avérée insuffisantes. Ce seuil a en effet produit la mise en relation de couples erronés (par exemple *ainsi* et *sains*; *aoure* et *autre*; *beste* et *estre*; *bien* et *rien*; *ceste* et *cesse*; *chaille* et *escaille*, etc.). Avec le seuil de 0.8, la précision de la détection des mots apparentés entre la version en prose et celle en vers peut être estimée à 80%.

Pour remédier à ce problème, nous avons augmenté le seuil à 0.82. Cette élévation a permis d'atteindre une précision de l'ordre de 87%. En revanche, le rappel a diminué de 17% ce qui s'explique par la perte de mots effectivement apparentés dont le rapport de

Dice_1 se situe entre 0.8 et 0.82. Pour éviter la perte des mots apparentés faiblement corrélées (moins de 0.82), nous avons affiné la méthode grâce à l'inclusion de quelques règles de transformations relevant des particularismes grapho-phonétiques, morpho-phonétiques, dialectaux de la langue médiévale ou encore de caractéristiques aléatoires, et

1. Le eu non final peut être suivi de s, x , z, ls, lx, lz et peut se transformer en ex, es, ez, elx, elz, els. Le i se permute avec le y en fin de mot
3. Le t se permute avec le s (ou le z) en fin de mot
4. Le f en fin de mot se permute avec le s (ou le z) avant une consonne
5. Le eu non final se permute avec ou

Figure 2: Liste des règles pour déterminer les mots apparentés

susceptibles de s'appliquer à des contextes simples.

Le tableau 1 offre un échantillon de l'ensemble des mots apparentés détectés à partir de la rédaction en vers et de la version en prose qui forment le corpus après application de ces règles. Les colonnes Mot1 et Mot2 correspondent aux mots apparentés dont le rapport de Dice_1 initial figure dans la colonne Dice_1_A. Mot1_T représente la transformation du Mot1 qui fournit le meilleur rapport de Dice_1 (Dice_1_B) avec Mot2.

Mot1	Règles	Mot1_T	Mot2	Dice_1_A	Dice_1_B
Ainsy	2	Ainsi	ainsi	0.8	1
auctour	5	acteur	aucteur	0.77	0.92
cieulx	5	cielx	ciel	0.8	0.88
mieux	1	mieulx	mieulx	0.8	1
parels	5	pareulx	pareulx	0.77	1
saint	3	sains	sains	0.8	1

Tableau 1: Exemples de recherche de mots apparentés après application des règles graphiques

La mesure de similitude entre les données textuelles de deux segments est calculée en fonction de la fréquence relative des mots apparentés qu'ils partagent. Le rapport de Dice_1 est utilisé pour l'évaluation de cette fréquence. On dénote cette fonction de similitude par le terme anglais de *cognateness*.

5.1.2 Similitude morpho-syntaxique

La fonction de similitude lexicale fondée sur des heuristiques statistiques et sur l'application de quelques règles graphiques ne permet pas d'établir les correspondances susceptibles de se produire entre des segments soumis à une forte variation linguistique. La ré-élaboration d'un texte, le passage d'une structure à une autre ou d'un genre à un autre comme dans le cas de la mise en prose d'un poème ou de la réécriture en vers à partir de la prose, impliquent toute une série de transformations qui affectent aussi bien la structure des phrases que le choix du vocabulaire, fait qui entraîne de plus ou moins grandes divergences sur le plan lexical et morpho-syntaxique (*cf.* fig. 3).

Rédaction en vers	Adaptation en prose	Dice_1
Si com Saincte Eglise vorra / Que je doy **croire** ce qu'il **croirra**	Selon Sainte Eglise en qui je **croy**	0.8 0.72
Combien que li paien **creüssent** / Des Diex et que pluseurs en fussent	Car quoy que les payens **croyent** pluiseurs dieux	0.62
Car nulle meilleur ne **peut** estre	Car meilleur ne **povoit** estre	0.4
Ovides en sa **commençaille**	Ovide en son **commencement**	0.66
Ci **commence** le premier livre d'Ovide Methamorphoses	Comment Ovide au **commencement** de ce livre invocque l'ayde divine	0.8
Pour plus plaire a ceulz qui l'orront / Et maint **profiter** y pourront	La mutacion des fables **prouffitables** au mieulx que possible me sera	0.66
Et Le tiengne en **humilité** / **Membrer li doit** que de vilté / **Soit estrais** et creé de boe	Et se tiegne **humble** et **souvenant** que **estrais est** de vilté	0.71

Figure 3: Variation morpho-syntaxique et lexico-sémantique dans les textes médiévaux

De telles alternances ne peuvent être prises en compte par la fonction de similitude lexicale. Une comparaison efficace requiert l'intervention de modèles de transformation morpho-syntaxiques et lexico-sémantiques. La base de données MedievLex nous a permis de définir d'autres heuristiques de similitude en considérant tout d'abord les traits « lemme » et « famille ».

Le trait « lemme »

Le trait « lemme » permet de définir une forme canonique pour les entrées lexicales. Cette forme est représentée par l'infinitif pour les verbes et par le masculin singulier pour les substantifs (sauf pour les noms essentiellement féminins), les adjectifs, les pronoms et les déterminants. Grâce à cette étape de lemmatisation, il est possible d'établir la correspondance entre les formes conjuguées des verbes (par exemple, *croyent* et *creüssent*)

et entre des dérivés morphologiquement distinct (par exemple, *commencement* et *commençaille*).

Le trait « famille »

Dans la base MedievLex, le trait « famille » découle de l'origine étymologique des entrées lexicales. Il procure en outre une forme canonique pour les différentes catégories syntaxiques dérivées. Grâce à cette nouvelle étape de lemmatisation, plus abstraite que celle proposée dans le cas précédent, il est possible d'opérer des recoupements entre les verbes, les substantifs, les adjectifs et les adverbes partageant la même origine, comme par exemple entre le verbe *profiter* et l'adjectif *prouffitable* ou encore entre le substantif *humilité* et l'adjectif *humble*. Etant donné l'importance des transformations entre les différents états des écrits médiévaux, cette heuristique s'avère très utile.

La similitude morpho-syntaxique des données textuelles de deux segments se mesure d'après la fréquence relative des formes linguistiques canoniques (lemmes et familles) qu'ils partagent. Le rapport de Dice_1 est à nouveau utilisé comme heuristique statistique pour la mesure de cette fréquence. On dénote cette fonction de similitude par le terme d'*allomorphism*.

5.1.3 Similitude sémantique

Outre les similitudes morpho-syntaxiques et lexicales qu'ils expriment, les textes médiévaux révèlent d'autres types de correspondances, notamment sur le plan de la signification des mots et des expressions employées. En effet, si la retransmission des œuvres par voie de copie aussi bien que leur réécriture (sous forme de mises en prose par exemple, ou lors de changements de catégorie littéraire, de structure poétique, etc.) peuvent entraîner des modifications considérables à l'échelle macroscopique – interpolations, déformations, remaniements, suppressions, volontaires ou non –, elle garantit en principe une certaine adéquation entre deux retranscriptions ou entre deux rédactions au point de vue du contenu général et du sens global, du moins pour des parties plus ou moins étendues du texte. Au point de vue lexico-sémantique, les mots peuvent être remplacés par des synonymes, par des syntagmes équivalents ou par des termes ou des expressions destinés à « rajeunir » les textes en remplaçant les vocables ressentis comme archaïques par de nouveaux lexèmes ou des formules plus récentes; les tournures font place à d'autres tournures sémantiquement conformes, etc. Afin de déterminer la correspondance entre ces éléments, d'autres méthodes de comparaison sont nécessaires.

Diverses ressources ont été développées dans le but de représenter les relations sémantiques qui s'expriment dans le vocabulaire des langues modernes (antonymie, hyperonymie, hyponymie, synonymie, etc.). On peut citer entre autres WordNet, [FELLBAUM 98] réseau sémantique qui figure la relation lexicale entre les mots d'une part et la relation sémantique entre les champs notionnels d'autre part. Dans WordNet, les mots contiennent très peu d'information syntaxique, mais ils désignent surtout des concepts. WordNet constitue de fait un thesaurus : sa structure repose sur un ensemble de synonymes (en anglais : *synsets*) qui renferment des mots traduisant un concept donné. En appelant l'un des termes qui correspond à cette notion, l'utilisateur de WordNet peut ainsi trouver d'autres vocables qui actualisent le même concept.

Les ressemblances que l'on constate entre la structure sémantique du lexique médiéval et moderne du français invitent à compléter la base lexicale MedievLex au moyen d'un lien vers la langue contemporaine. Ce lien de redirection – le trait « sens » dans la base – indique la traduction des entrées en français moderne. La mise en regard des mots devient par conséquent une comparaison des synsets de leurs traductions dans WordNet. Deux termes de la langue médiévale, ou deux usages particuliers d'un vocable pour les termes polysémiques, sont considérés comme synonymes si leurs équivalents en français moderne appartiennent au même synset. Le résultat est bien sûr déterminé de façon prépondérante par la qualité des traductions.

Cette méthode concerne uniquement les mots. Afin d'enrichir les connaissances sémantiques sur les expressions et sur les groupes de mots, une base de données terminologiques a été développée. Elle se présente sous la forme d'un simple tableau analogique qui explicite les paires équivalentes (comme dans le cas de *aimer* / *avoir chier*, *occire* / *metre a mort*, *nomer* / *estre appelé* par exemple).

La similitude sémantique des données textuelles de deux segments se mesure en fonction de la fréquence relative des analogies sémantiques qu'ils partagent dans leur lexique. Le rapport de Dice_1 est à nouveau employé comme heuristique statistique pour la mesure de cette fréquence. On dénote cette fonction de similitude par *synonymy*.

5.1.4 Fonction de similitude linguistique

Après avoir calculé trois fonctions de similitude différentes – lexicale, morpho-syntaxique et sémantique –, il ne reste plus qu'à les combiner pour en déduire une première fonction de similitude (fonction de similitude linguistique d'ordre 1). La combinaison choisie est une combinaison linéaire où les coefficients (c_1, c_2 *et* c_3) évaluent l'importance ou la confiance attribuée aux fonctions de similitude (*cognatness, allmorphism et synonymy*).

$$g_1 = c_1 * cognatness + c_2 * allomorphism + c_3 * synonymy$$

L'ordre des mots est également un critère important dans le processus de comparaison. Une suite de n (n= 2) mots similaires dans deux segments constitue un indicateur précieux pour leur mise en correspondance.

Nous proposons ainsi d'autres fonctions de similitudes (d'ordre n) g_n qui reposent sur un modèle n-gram de suites de mots linguistiquement semblables. Un n-gram est une suite de n mots $(m_k^S,...,m_{k+n}^S)$ dans le segment textuel S qui expriment des relations de similitude linguistique – mots apparentés, mots morphosyntaxiquement proches ou synonymes – avec $(m_p^D,...,m_{p+n}^D)$ dans le segment textuel D, c'est-à-dire que m_k^S est similaire à m_p^S , m_{k+1}^S est similaire à m_{p+1}^S et ainsi de suite jusqu'à m_{k+n}^S et m_{p+n}^S. La suite *Quant Dieu ordeneement* ([*eut*] / *Assiegié chascun element*) est par exemple similaire à *Quant dieu eut ordonné* (*les elemens*) si l'on applique à ces segments un modèle de tri-gram en ne considérant que les mots d'une taille supérieure à 3 caractères. La méthode fait apparaître les couples suivants, compte tenu du taux de confiance qui peut être accordé à chaque paire : (*Quant*, *Quant*, c_1); (*Dieu*, *Dieu*, c_1); (*ordeneement*, *ordonné*, c_2). Le poids de ce n-gram c_n est égal au produit des c_i qui correspondent à chaque couple de mots

similaires, c'est-à-dire $c_n=c_1*c_1*c_2$. Avec ce modèle, il est possible de calculer la fréquence des n-grams similaires (n≤2) dans les segments en appliquant le rapport de Dice_n comme suit :

$$g_n(S,D) = \frac{2*\sum c_n}{l_1 + l_2 \_ 2(n-1)} \qquad (2)$$

où $\sum c_n$ correspond à la somme des taux de confiance des n-grams trouvés entre les segments textuels S et D.

En dernière analyse, la fonction de similitude f_1 fondée sur les traits linguistiques énoncés au début de cette section représente une combinaison linéaire des n fonctions de similitude de n-gram entre les éléments des documents S et D. f_1 s'écrit alors sous la forme :

$$f_1 = \Sigma_{i=1}^{n} \mu_i g_i$$

où μ_i est le poids des modèles i-gramme et les g_i tels que (i≤ n) sont calculés selon la formule (2). Dans notre modèle de comparaison, nous nous limitons à l'ordre 3, c'est-à-dire n=3.

5.2 Critères structurels pour l'alignement des textes anciens

Nous chercherons maintenant à définir les différentes fonctions qui permettent de mesurer la similitude entre deux documents parallèles d'après leurs caractéristiques structurelles. Les particularités inhérentes à l'organisation rhétorique du texte, en particulier le statut propre à chaque segment (noyaux et satellites), la nature des relations qui les unissent et la structure qu'ils composent sont surtout pris en considération.

5.2.1 Similitude des chemins de relation

Le chemin de relation d'un segment est la suite des relations rhétoriques que constituent les nœuds appartenant à ce chemin de la racine jusqu'à la feuille terminale. La similitude des chemins de relation des segments se fonde sur la recherche des sous-chaînes de relation d'un chemin P_1 dans P_2 avec une distance minimale. L'écart se définit par rapport aux suppressions, insertions et substitutions constatées dans la comparaison des chaînes de caractères. La recherche des sous-chaînes de relations ayant une distance minimale est équivalente à celle des correspondances approximatives entre les chaînes avec une différence de k (approximate string matching with k difference [LECROQ 95]). Ce problème est traité par la méthode de programmation dynamique où l'on suppose que chaque P est une chaîne finie de relations et que l'espace de calcul est un tableau T de taille m× n (m et n sont les tailles respectives de P_i et P_j). Chaque case du tableau $T_{i,j}$ est remplie comme suit :

$$\min \begin{cases} T_{i-1,j-1} + substitution(x_i, y_j) \\ T_{i,j-1} + insertion(y_j) \\ T_{i-1,j} + deletion(x_i) \end{cases}$$

Les *xi* et y_j représentent les noms des relations à la position respective i et j de P_1 et P_2. Le coût de la substitution est donné par *substitution*(x_i,y_j) et il est calculé en fonction de la similitude des relations rhétoriques. Le coût de l'insertion et de la suppression est fixé à la valeur maximale que la substitution de deux relations peut avoir. Celui de la substitution dépend de la classification des relations rhétoriques (il est moins coûteux de substituer des relations similaires telles que *contraste* et *antithèse* par exemple que de remplacer *contraste* par *comparaison* ou l'inverse, etc.).

Soit par exemple P1=(b,e,c,e) et P2=(e,a,m,s) où a=antithèse, b=arrière-plan, c=contraste, e=élaboration, m=commentaire et s=séquence. La différence minimale approximative de P_2 dans P_1 est obtenue avec l'alignement suivant :

$$\begin{bmatrix} - & e & a & s & m \\ b & e & c & e & - \end{bmatrix}$$

On trouve alors : min_diff(P1,P2)=substitution(e,e) + substitution(a,c) + substitution(m,e) + insertion(s).

5.2.2 Similitude des arbres rhétoriques

Les arbres rhétoriques sont des arbres binaires ordonnés qui décrivent de quelle manière, en termes de cohérence et de cohésion, les segments sont disposés dans le texte. Les nœuds regroupent un ensemble de segments unis par des relations rhétoriques et décrivent leur statut. Les changements majeurs au cours du processus de traduction ou de réécriture interviennent sur le plan linguistique (lexical, morphologique et syntaxique). De telles transformations préservent en général le contenu sémantique et l'on peut admettre qu'au point de vue de la structure rhétorique profonde, les correspondances sont également maintenues.

Les expériences d'annotation et de comparaison rhétoriques ont montré que les modifications qui peuvent atteindre ce niveau concernent pour l'essentiel la segmentation, une phrase pouvant être remplacée par une suite de mots d'ordre supérieur ou inférieur, ce qui se traduit par l'ajout ou par la disparition de segments élémentaires et donc, par l'apparition ou par l'élimination de sous-structures arborescentes, surtout au niveau inférieur de l'arbre.

Le statut des segments de même que le choix des relations peuvent également être affectés lors de la transformation d'un texte. En ce qui concerne la nature des liens rhétoriques, il faut cependant tenir compte de la forte dépendance vis-à-vis de l'appréciation de l'annotateur. Ce critère n'est donc pas déterministe. Tout processus de comparaison entre deux structures doit, par ailleurs, être basé sur des méthodes heuristiques.

Un dernier type de discrépance est induit par l'évaluation subjective de la cohésion des segments que l'annotateur opère au cours du processus d'annotation. Deux segments S_1 et S_2 peuvent apparaître plus cohésifs que S_2 et S_3 d'où la structure $R_2(R_1(S_1,S_2),S_3)$. Après la traduction ou la réécriture du texte, la cohésion peut être inversée et la structure sera perçue sous la forme $R_1(S_1,R_2(S_2,S_3))$. Ce type de problème peut apparaître à toutes les étapes de la

représentation – segments élémentaires et segments composés – et modifier entièrement la hiérarchie des relations à l'intérieur des arbres ou la segmentation entre ces derniers.

La détection des similitudes au niveau de la structure globale des arbres doit par conséquent reposer sur des méthodes heuristiques aptes à qui prendre en compte ces types de modifications éventuelles. Dans les travaux sur cette question, le problème de la comparaison entre les arbres ordonnés est envisagé sous l'angle mathématique et il a été traité surtout de manière analogue à celui de la comparaison entre chaînes. La mise au point de notre méthode s'est inspirée des travaux effectués par [ZHANG 94, 97] dans le domaine de l'édition des arbres ordonnés.

6. Résulats et évaluations

Les expériences d'alignement que nous décrivons ont été effectuées avec l'aligneur automatique MultAlign [GHORBEL 02_1] développé au sein de ce projet. Les résultats que nous présentons concernent l'alignement de la version en prose (Paris, BNF, f. fr. 137) et d'un représentant de la rédaction versifiée (Genève, BPU, fr. 176) de la première partie de l'*Ovide Moralisé* (prologue et début du Livre I). Les métriques d'évaluation utilisées sont celles proposées par [ISABELLE 96] et révisées dans le cadre du projet ARCADE [LANGLAIS 98, 99, VERONIS 00]. L'idée principale est d'utiliser les notions de rappel et de précision employées dans le domaine de la recherche de l'information pour situer un alignement généré automatiquement par rapport à un alignement de référence.

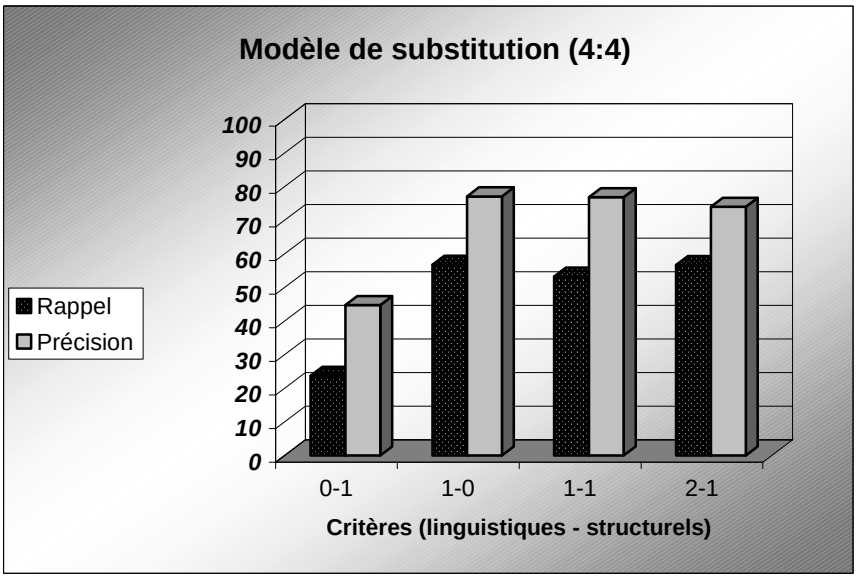

Figure 4: Croissance du taux de rappel avec l'utilisation des critères linguistiques et structurels

Comme on le constate sur l'histogramme de la figure 4, la prise en compte simultanée des critères linguistiques et structurels avec une pondération convenable (illustrée sur l'abscisse de l'histogramme) dans l'alignement avec un modèle de substitution 4:4 permet d'améliorer le rappel de l'alignement, sans pour autant en abaisser la précision.

Dans ces expériences d'alignement, l'emploi du critère linguistique permet d'obtenir un taux de F- mesure de l'ordre de 60%. Ces résultats pourraient être améliorés grâce à un accroissement des ressources linguistiques – telles que la base de données lexicales MedievLex et la base de données terminologiques ou l'interface de WordNet en français –, afin de favoriser la détection de similitudes morpho-syntaxiques et sémantiques.

Enfin, au vu du coût de l'annotation rhétorique, l'application du critère structurel peut être restreinte aux passages difficiles à aligner.

7. Conclusion

Les textes médiévaux offrent une richesse et une diversité exceptionnelles au point de vue formel et sur le plan de leur contenu, mais l'une et l'autre entraînent de nombreuses difficultés pour l'exploration des tels documents et pour la mise en lumière des similitudes et des disparités qui s'expriment entre eux.

L'objectif prioritaire du projet MÉDIÉVAL est de créer un environnement de comparaison des œuvres anciennes et de permettre aux spécialistes et aux chercheurs sur le moyen âge vernaculaire de procéder à des études comparatives sur les origines et sur les propriétés philologiques et linguistiques des textes du XIIe au XVI siècle. Cette interface éditoriale vise donc à faciliter la collation des textes tant au niveau macroscopique que microscopique.

Au point de vue macroscopique et, en particulier, de la structure exégétique des écrits, MÉDIÉVAL offre la possibilité d'enrichir les transcriptions par les expertises et par les annotations des philologues. Ceci est rendu possible grâce à la navigation entre les éditions en versions originales ou reconstituées artificiellement.

Au point de vue microscopique, l'éditeur permet de visualiser l'alignement automatique des éléments mis en correspondance. Cependant, les techniques classiques d'alignement, en particulier les méthodes statistiques telles qu'elles sont appliquées aux textes multilingues, ne réussissent pas à vaincre la complexité de certaines transformations dont les œuvres médiévales nous offrent le témoignage. Pour obtenir un rapprochement pertinent lors de conversions, de réécritures comme dans le cas d'un dérimage ou d'une mise en prose par exemple, ou d'interprétations, il y a lieu d'envisager des approches linguistiques et structurelles, différentes de celles employées pour de simples traductions.

Sur le plan linguistique, des méthodes de calcul de similitudes concernant les aspects morpho-syntaxiques, lexicaux et sémantiques ont été appliquées au corpus étudié. Pour l'analyse structurelle, la théorie de l'organisation rhétorique (RST) a été choisie comme modèle de référence pour comparer les caractéristiques de ces textes. Cette double approche, combinée à une appréciation des composantes typographiques des documents, permet de définir la similitude entre plusieurs versions d'un même écrit dans un espace

multidimensionnel et d'augmenter la précision et la fiabilité dans le processus d'appariement.

Pour résoudre le problème – essentiellement combinatoire – de l'association des parties analogues et de la maximisation de la similitude globale, des algorithmes de programmation dynamique ont été adoptés. La pondération des fonctions de similitude des différents critères est fixée expérimentalement et dépend du genre de document à aligner. De telles fonctions de similitudes entre entités textuelles ou structurelles sont suffisamment générales et paramétrables pour être réutilisées dans d'autres contextes d'applications, comme la constitution de ressources linguistiques, les bases de données terminologiques, les dictionnaires bilingues, les mémoires de traduction ou encore, l'apprentissage des langues assisté par ordinateur.

Ce dernier domaine, pédagogique et éducatif, concerne très directement le projet MÉDIÉVAL, puisque celui-ci a aussi pour but de permettre à des étudiants de découvrir la langue ancienne par la recherche des paradigmes flexionnels et la compréhension des mécanismes de variation, des exemples pratiques de phrases et de conversions au point de vue syntaxique, de transformations sur le plan lexico-sémantique ou encore, des incidences liées à l'époque ou à la provenance du texte, etc. Cet aspect pédagogique peut s'étendre au niveau argumentatif et stylistique. La structure rhétorique des textes parallèles offre ainsi par exemple le moyen d'extraire et de confronter les expressions argumentatives utilisées dans des états de texte différents.

L'annotation rhétorique est à la base de l'alignement structurel. L'inconvénient majeur de cette étape est la part de subjectivité inhérente à une telle opération qui, même avec le modèle RST-Light, reste dépendante de l'exécutant. Il serait donc utile de prolonger la recherche en direction de méthodes automatiques qui puissent contribuer à une vérification de la cohérence de l'annotation et à la correction éventuelle de l'analyse. Ceci peut être réalisé en intégrant des méthodes heuristiques de calcul de la cohésion lexicale entre les segments (en raisonnant sur la sémantique du lexique, par exemple). De telles procédures doivent être capables d'estimer le degré de dépendance entre les segments et de contribuer à la formation des arbres rhétoriques. Compte tenu de l'impact de ces connaissances dans la description structurelle du discours, il serait également intéressant d'intégrer des techniques de résolution des co-références (expressions de type anaphorique).

Loin d'être réservé aux seuls matériaux écrits, l'alignement offre des perspectives prometteuses pour des applications comme le couplage son (parole) / texte (script ou télétexte). L'alignement multimédia est une des techniques efficaces pour la navigation synchronisée dans les documents vidéo et dans la recherche multimodale de l'information, au même titre qu'entre texte et miniatures ou autres éléments iconographiques dans les manuscrits, par exemple. A cet égard, en associant sur la même page des enluminures ou des partitions musicales à des retranscriptions destinées avant tout à être lues, récitées ou chantées, les copistes du moyen âge n'ont certes pas attendu l'avènement des multimédia, dont ils sont en quelque sorte les précurseurs. Aujourd'hui, ces dimensions du texte ancien, que l'imprimé « aplatit » et occulte, peuvent être restituées de manière pertinente grâce à l'ordinateur.

Références

[AZZAM 01] AZZAM W., COLLET O., CORAY G., GHORBEL H., *Rapport final du projet MÉDIÉVAL*, http://lithwww.epfl.ch/~ghorbel/publications.html, 2001.

[BALLIM 96] BALLIM A., *Multext Aligner v. 2.0*, Technical Report, Institut Dalle Molle pour les Études Sémantiques et Cognitives, ISSCO, 1996.

[BALLIM 98] BALLIM A., CORAY G., LINDEN A., VANOIRBEEK C., " The use of automatic alignment on structured multilingual documents", *In* : *Proceedings of the Seventh International Conference on Electronic Publishing*, Saint Malo, 1998, pp. 464-475.

[BALLIM 94] BALLIM A., RUSSEL G., "LHIP : Extended DCGs for configurable robust parsing", *In* : *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, 1994, pp. 501-507.

[BEEFERMAN 99] BEEFERMAN D., BERGER A., LAFFERTY J., "Statistical Models for Text Segmentation", *In* : *Machine Learning*, 34 (1-3), 1999, pp. 177-210.

[BROWN 88] BROWN P., DELLA PIETRA S., DELLA PIETRA V., MERCER R., "A statistical approach to language translation", *In* : *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988, pp. 1-6.

[BROWN 91] BROWN P., LAI J., MERCER R., "Aligning sentences in parallel corpora", *In* : *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 1991, pp. 169-176.

[CATIZONE 89] CATIZONE R., RUSSELL G., WARWICK S., "Deriving translation data from bilingual texts", *In* : ZERNIK U. (Ed.), *Proceedings of the first Lexical Acquisition Workshop*, Detroit, Michigan, 1998.

[DAGAN 96] DAGAN I., "Bilingual word alignment and lexicon construction", *In* : *Tutorial Notes of the 34th Annual meeting of the Association for Computational Linguistic*, Santa Cruz, California, 1996.

[ERJAVEC 00] ERJAVEC T., EVANS R., IDE N., KALIGARIF A., "The CONCEDE model for Lexical Databases", *In* : *Proceedings of the Second International Language Resources and Evaluation Conference*, Paris, European Language Resources Association, 2000.

[FELLBAUM 98] FELLBAUM C., *WordNet, An Electronic Lexical Database*, The MIT Press, 1998.

[GALE 91] GALE W., CHURCH K., "A program for aligning sentences in bilingual corpora", *In* : *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 177-184.

[GHORBEL 01] GHORBEL H., BALLIM A., CORAY G., "Rosetta : Rhetorical and semantic environment for text alignment", *In* : *Proceedings of Corpus Linguistics conference*, Lancaster, pp. 224-233.

[GHORBEL 02_1] GHORBEL H., CORAY G., LINDEN A., "SAM : System for Multi-criteria Text Alignment", *In* : *Proceedings of the International Conference On language Ressources and Evaluation LREC 2002*, Las Palmas, 2002, pp. 404-410.

[GHORBEL 02_2] GHORBEL H., *Alignement Multicritères des Textes : Critères linguistiques et structurels appliqués aux documents médiévaux*. Thèse de Doctorat en Informatique N° 2609, École Polytechnique Fédérale de Lausanne, 2002.

[HOFLAND 98] HOFLAND K., JOHANSSON S., "The Translation Corpus Aligner : A program for automatic alignment of parallel texts", *In* : Johansson S., Oksefjell S. (Eds.), *Corpora and Cross linguistic Research : Theory, Method, and Case Studies*, Amsterdam, Rodopi, 1998, pp. 87-100.

[IDE 95] IDE N., LE MAITRE J., VÉRONIS J., "Outline of a Model for Lexical Databases", *In* : Current Issues in Computational Linguistics. *In Honour of Don Walker, Linguistica Computazionale IX*, Pisa, 1995, pp. 283-320.

[ISABELLE 96] ISABELLE P., SIMARD M., *Propositions pour la représentation et l'évaluation des alignements de textes parallèles*, Rapport Technique, Centre d'innovation en technologies d'information Industrie et Sciences, Canada, 1996.

[KAY 93] KAY M., ROESCHEISEN M., "Text-translation alignment", *In* : *Computational Linguistics*, 19(1), 1993, pp. 121-142.

[KUPIEC 93] KUPIEC J., "An algorithm for finding noun phrase correspondences in bilingual corpora", *In* : *Proceedings of the 31st Annual Meeting of the ACL*, Columbus, Ohio, 1993, pp. 17-22.

[LANGLAIS 98] LANGLAIS P., SIMARD M., VÉRONIS J., ARMSTRONG S., BONHOMME P., DÉBILI F., ISABELLE P., SOUISSI E., THÉRON P., "Arcade : A cooperative research project on parallel text alignment evaluation", *In* : *Proceedings of the First International Conference On Language Resources and Evaluation (LREC)*, Granada, 1998.

[LANGLAIS 99] LANGLAIS P., SIMARD M., VÉRONIS J., *ARCADE Methods and Practical Issues in Evaluating Alignment Techniques*, Université d'Aix-en-Provence, 1999.

[LECROQ 95] LECROQ T., "Experimental results on string-matching algorithms", *In* : *Software Practice and Experience*, 25(7), 1995, pp. 727-765.

[MANN 87] MANN W., THOMPSON S., *Rhetorical Structure Theory : A Theory of Text Organization*, Technical report, Information Science Institute, 1987.

[MANN 88] MANN W., THOMPSON S., "Rhetorical Structure Theory : Toward a functional theory of text organization", *In* : *Text*, 8(3), 1988, pp. 243-281.

[MCENERY] MCENERY A., OAKES P., "Cognate extraction in the Crater project", *In* : *Proceedings of the EACL-SIGDAT workshop*, Dublin, 1995, pp. 77-86.

[MELAMED 99] MELAMED D., "Bitext maps and alignment via pattern recognition", *In* : *Computational Linguistics*, 25(1), 1999, pp. 107-130.

[OWEN 98] OWEN C.B., "Parallel Text Alignment", *In* : *proceedings of the Second European Conference for Digital Libraries ECDL'98*, Heraklion, 1998, pp. 235-259.

[PETITPIERRE 95] PETITPIERRE D., RUSSELL G., *MMORPH- The Multext Morphological Program Version 2.3*, Technical Report, ISSCO, 1995.

[ROMARY 00] ROMARY L., BONHOMME P., "Parallel alignment of structured document", *In* : Véronis J. (Ed.), *Parallel Text Processing*, Dordrecht, Boston, London, Kluwer Academic Publishers, 2000, pp. 201-217.

[SIMARD 92] SIMARD M., FOSTER G., ISABELLE P., "Using cognates to align sentences in bilingual corpora", *In* : *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, 1992, pp. 67-81.

[VAN DER EIJK 93] VAN DER EIJK P., "Automating the acquisition of bilingual terminology", *In* : *Sixth Conference of the European Chapter of the Association of Computational Linguistics*, Utrecht, 1993, pp. 113-119.

[VÉRONIS 00] VÉRONIS J., "Evaluation of parallel text alignment systems. The ARCADE project", *In* : Véronis J. (Ed.), *Parallel Text Processing*, Dordrecht, Boston, London, Kluwer Academic Publishers, 2000, pp. 369-388.

[VERONIS 00] VERONIS J., "Alignement de corpus multilingues", *In* : Pierrel J.-M. (Ed.), *Ingénierie des langues*, Paris, Éditions Hermès, 2000.

[VÉRONIS 00] VÉRONIS J., *Parallel Text Processing : Alignment and Use of translation Corpora*, Dordrecht, Boston, London, Kluwer Academic Publishers, 2000.

[WAGNER 74] WAGNER R.A., FISCHER M.J., “The string-to-string correction problem”, *In* : *Journal of the ACM*, 21(1), 1974, pp. 16-173.
[ZHANG 94] ZHANG K., SHASHA D., WANG J.T.L, “Approximate tree matching in the presence of variable length don’t cares”, *In* : *Journal of Algorithms*, 16(1), 1994, pp. 33-66.
[ZHANG 97] ZHANG K., SHASHA D., *Pattern Matching Algorithms*, Chapter 14 : “Approximate tree pattern matching”, Oxford University Press, 1997.