En quête d'amitié. Approches méthodologiques pour l'analyse automatisée d'un corpus électronique.

1. Contexte du projet TOUCHER et liens avec la Sator

Par ses principes et ses objectifs, le projet TOUCHER s'inscrit directement dans la lignée de ceux qui ont présidé à la création et au développement de la Sator, car ils découlent de la même observation. Comme on le sait, les travaux de la Sator ont pour objet de rendre compte du « jeu sur la récurrence de topoï narratifs, la plupart repris aux époques précédentes », qui caractérise la littérature jusqu'à la fin du 18e siècle[1]. Afin de mieux comprendre le fonctionnement de l'écriture narrative de cette période, il est donc important de pouvoir identifier ces topoï et d'observer comment ils se recyclent, s'actualisent à travers les périodes et les œuvres. Cela signifie opérer sur un vaste corpus qui s'étend sur plusieurs siècles et surtout qui représente des états de langue différents depuis l'ancien français, tâche irréalisable pour le chercheur isolé. La seule façon de l'aborder est de fonctionner en réseaux de chercheurs spécialistes de chacune des périodes et de textes particuliers[2]. C'est l'optique qui définit la démarche de la Sator depuis l'origine et celle aussi qu'a adoptée TOUCHER. Dès l'origine également, la mise en commun des connaissances et des recherches de chacun a coïncidé avec la prise de conscience qu'elle ne pouvait se réaliser que grâce à l'apport de l'informatique : le fruit du travail de repérage des topoï a abouti à la réalisation du thesaurus informatisé, « www.satorbase.org ». La même conviction est aussi au cœur du projet TOUCHER qui voudrait trouver des éléments de réponse à la question suivante : comment la lecture des textes est elle enrichie par la possibilité de traiter ainsi de vastes corpus de textes littéraires appartenant à une large diachronie, possibilité qui échappe au chercheur individuel dont la compétence se limite aux œuvres de son domaine de recherche ?

Poser la question ainsi, c'est aller au-delà du repérage des configurations narratives topiques pour espérer voir se dégager des principes de leur combinatoire par la mise en perspective de ces situations récurrentes dans un nombre important de textes. C'est en réalité se trouver face au défi que pose l'utilisation de l'informatique pour une authentique démarche de type

[1] http://www.satorbase.org : voir « Historique de la Sator » où l'on trouvera des informations sur la définition du topos narratif.
[2] Jean-Marc Ramos, « La Sator face au Golem. Usages et représentations des instrumentations numériques dans une communauté savante », 2004, p. 643-677.

littéraire. Jusqu'à présent, le caractère algorithmique des outils informatiques a conduit principalement à des résultats présentés sous forme d'index, de concordances, de tableaux de fréquence, etc.[3]. Alors qu'on peut observer des réussites certaines en linguistique et en stylistique, le chercheur intéressé aux questions d'ordre littéraire se trouve face à une masse d'informations pour lesquelles il ne possède pas encore les outils permettant des stratégies de traitement adéquates. Les travaux de la Sator nous semblent pouvoir offrir une possibilité de réponse au défi de l'analyse littéraire informatisée, dans la mesure où identifier des candidats de récurrences de scénarios topiques, mettre en évidence des types de relations entre ces scénarios, constituent des opérations formalisables et donc potentiellement automatisables.

Les conditions semblent propices puisque, parallèlement à l'existence de bases de données qui compilent de l'information de type littéraire, les textes numérisés disponibles sur internet ne cessent de se multiplier. En fait, c'est cette abondance de ressources même qui a suscité la mise en place de TOUCHER. Aux réseaux de textes et de chercheurs propres à l'approche de la Sator, s'ajoutent ceux que constituent les outils informatiques à solliciter. Une première tentative d'interconnexion de ces outils a donné le système de recherche PBLit ou « Polybases littéraires », sur laquelle s'appuie le projet TOUCHER, et qui consiste à interroger plusieurs bases de données hétérogènes en même temps[4]. Il s'agit pour le moment d'un prototype qui permet de solliciter, en plus de Satorbase, le « Thesaurus des motifs merveilleux du roman médiéval » et Hyperliste, bases qui rassemblent des textes littéraires du 12^{e} au 16^{e} siècle et qui ont pour point commun d'être faits d'extraits de textes[5]. En plus de cette possibilité d'interrogation simultanée de plusieurs bases de données, l'autre caractéristique de PBLit, qui augmente ses capacités d'interrogation, est le fait que son moteur de recherche peut mettre en évidence des cooccurrences (la proximité de termes choisis). On peut ainsi non seulement identifier l'environnement textuel d'un terme ou d'une expression, mais également les lieux où se répètent les cooccurrences de plusieurs termes. Cela peut d'autre part se faire à partir de sous-ensembles définis par auteur, œuvre ou niveau de généralité. La capacité d'établir des interconnections entre occurrences qu'offre un outil comme PBLit rend envisageable l'exploitation des corpus textuels en fonction d'unités supérieures au mot ou même au syntagme, notamment par la mise en évidence des modes d'organisation du récit tant au plan

[3] Stephen Ramsay, « Toward an Algorithmic Criticism », 18 (2003), p.1647-174.
[4] http://tapor.mcmaster.ca/pblit/; pour une description plus détaillée de PBLit, voir Madeleine Jeay et Stéphan Sinclair, « L'exploitation des bases de données en littérature : l'approche de PBLit », dans « Por s'onor croistre », 2008, p. 443-455
[5] Le Thesaurus n'est pas encore accessible publiquement ; http://tapor.mcmaster.ca/~hyperliste

syntaxique que sémantique en fonction des transformations de molécules sémiques[6]. Comme nous allons le voir, il va s'agir maintenant de développer PBLiT afin que le moteur de recherche puisse aussi interroger les textes intégraux. Toutefois, avant d'expliciter la démarche de TOUCHER et de faire le point sur les étapes accomplies, il est bon de préciser ses objectifs.

2. Objectifs du projet Toucher

TOUCHER correspond à l'approche actuelle de l'historicité dans les études littéraires qui vise, sur le plan de la poétique textuelle, à concilier diachronie et synchronie, formes et interprétation. Il s'agit de recadrer le texte dans un ensemble globalisant de relations et donc de dépasser la particularité de chaque texte pour travailler sur des corpus (quitte à revenir par la suite à chaque texte pour mieux comprendre les particularités par rapport à l'ensemble). L'objectif principal du projet serait d'offrir aux spécialistes de l'analyse du récit un ensemble d'outils conceptuels et technologiques pour l'exploitation des banques de données et des textes numérisés qui sont à leur disposition. Or la réalisation de cet objectif se heurte à l'obstacle que pose la difficulté d'exploiter les outils informatiques pour une exploitation de type littéraire. Ces limites contrastent avec la diversité et la sophistication des projets d'édition électronique ou d'analyse grammaticale et stylistique. Peut-on aller au-delà de résultats présentés sous forme d'index, de concordances, de tableaux de fréquence ? Surtout comment peut-on passer du stade de l'accumulation de matériel brut que fournit la recherche de termes-clés à celle de son analyse ? Les outils font défaut au moment même où s'amorce le vrai travail littéraire.

Dans l'immédiat, comment cet objectif général se traduit-il dans la mise en œuvre du projet ? Les objectifs spécifiques de TOUCHER peuvent s'énoncer ainsi :

1. Procéder à l'intégration de PBLit à des outils d'analyse textuelle existants, notamment « Voyant » et développer de nouveaux prototypes, conçus pour représenter divers aspects du corpus et les explorer en vue de l'analyse du récit. Le système d'analyse textuelle Voyant, développé par Stéfan Sinclair en collaboration avec Geoffrey Rockwell, est un logiciel en ligne d'analyse de texte qui offre une grande gamme de fonctions analytiques pour étudier la fréquence, la distribution et l'interaction d'unités textuelles (mots, lemmes, phrases,

[6] François Rastier, *Arts et sciences du texte*, 2001, p. 46

etc.[7]). Cela signifie pouvoir, à partir de PBLit, créer un corpus dans Voyant, mais aussi, tout en restant dans la fenêtre de Voyant, être en mesure d'effectuer des recherches dans PBLit.

2. Développer une interface qui permettre divers types de visualisation des tendances topiques à travers notre corpus. L'utilisateur pourrait ajuster les paramètres qui définissent les topoï et le sous-ensemble des textes qui l'intéressent (romans du 18e siècle par les femmes auteurs, par exemple). Les modes de visualisation suivants peuvent être envisagés :

a. de type chronologique

b. visualisation dynamique du réseau des unités topiques et des textes dans lesquels ces unités se trouvent. On pourrait voir quels fragments topiques s'attachent au terme « fuite », par exemple, ou bien la constellation des instances textuelles dans lesquelles se trouve ce fragment topique.

c. visualisation des séquences topiques qui permettrait d'analyser comment les topoï s'enchaînent et s'associent : au-delà du topos individuel, on accède à la grammaire des topoï, à la syntaxe narrative.

d. localisation des topoï dans le dispositif textuel lui-même : étude des lieux stratégiques (incipits, clausules), ruptures, identification des topiques d'ouverture, de dénouement, de transition chapitrale, etc.

3. La démarche et les réalisations d'une première année d'activités

Dans un projet comme TOUCHER, la démarche et la méthodologie, incluant les tâtonnements que le processus implique, font partie des résultats : l'essentiel est en effet de découvrir les chemins qui peuvent conduire à ces résultats, d'élaborer des procédures.

La première opération a été de réunir une équipe de base constituée, sous la coordination de Madeleine Jeay, de quatre spécialistes, un par période : Francis Gingras médiéviste, Hélène Cazes, Daniel Maher et Ugo Dionne, respectivement spécialistes du 16e, 17e et 18e siècles[8]. Stéfan Sinclair est responsable des aspects informatiques du projet. En plus de communiquer par courrier électronique, l'équipe tient des réunions régulières par téléconférence ou en

[7] Voyant Tools (voyant-tools.org) succède à aux systèmes d'analyse HyperPo et de Taporware (pour plus de détails sur sa génétique, voir http://hermeneuti.ca/voyeur/background).

[8] Ces chercheurs sont associés aux universités McMaster (Madeleine Jeay), Montréal (Francis Gingras et Ugo Dionne), Victoria (Hélène Cazes) et Calgary (Daniel Maher).

personne. La page d'accueil du projet est un lieu rassembleur pour des notes de réunion, des consignes techniques et documents d'aide ainsi que d'autres informations pertinentes pour les membres de l'équipe et pour un public plus large (digihum.mcgill.ca/toucher).

Chaque spécialiste a ensuite procédé à un recensement des textes numériques incontournables (ou d'intérêt particulier) pour sa période. Cette liste a été augmentée par le travail d'un assistant qui a fouillé l'internet à la recherche d'éditions numériques de textes de la période que recouvre le projet. Dans l'esprit du réseau du projet (textes, outils, chercheurs en réseau), nous avons choisi de compiler la bibliographie et les textes dans Zotero, un outil collaboratif de gestion – nous avons tous accès et pouvons tous éditer la même bibliographie en ligne ou synchronisée à nos ordinateurs. Zotero est très convivial et peut être utilisé grâce à une interface web, une extension de Firefox ou une application autonome (chaque solution comporte des avantages et des désavantages).

La constitution de la bibliothèque de textes numérisés doit se faire en deux temps. Il ne suffit pas en effet de relier ces textes aux sites sur lesquels ils sont disponibles (Gallica, Google, ABU, etc.), car leur format ou leur mode de présentation ne leur permet pas toujours d'être exploitables par la machine. Il faut donc les préparer à cet effet, ce qui signifie les transposer en format texte, supprimer certains éléments paratextuels qui ne sont pas d'intérêt pour nous, corriger les erreurs de numérisation les plus évidentes.

Parallèlement à cette opération, une première démarche a été entreprise afin de répondre à la question suivante : dans quelle mesure est-il possible d'envisager qu'on puisse faire un balisage automatique des textes qui permettrait d'identifier des situations topiques récurrentes? L'approche que nous avons tentée a été d'étiqueter « manuellement » un certain nombre de textes identifiés comme pertinents pour chaque période dans l'intention de tester par la suite les résultats de ce balisage à la machine. Dans la perspective du colloque de Victoria, nous avons décidé de travailler sur l'amitié. L'idée de départ était d'étiqueter des termes portant sur l'amitié et de les relier à un certain nombre de catégories auxquelles correspond une diversité de termes selon les textes et les périodes. Nous avons donc sélectionné un corpus de textes représentatifs pour chaque période, pour lesquels nous avons créé des concordances du terme « ami » et de ses dérivés. À chacune des occurrences, nous avons déterminé, lorsque c'était possible, des mots-clés correspondant, dans la mesure de nos connaissances, à des situations récurrentes. L'objectif de cette tâche est d'identifier le lexique

qui correspond, dans les textes, à ces mots-clés. À partir de ceux-ci et de l'éventail de termes qui leur sont reliés, il s'agira de permettre ultérieurement à la machine de faire ce travail automatique d'identification, de repérage de situations narratives.

Le corpus d'expérimentation a été déterminé en fonction des critères suivants : la représentativité du texte pour la période et sa richesse en topoi au sens large, c'est-à-dire la récurrences de scénarios, de motifs, de personnages ou de lieux types, etc. Pour le Moyen Âge, les œuvres sélectionnées sont *Amadas et Ydoine*, *Raoul de Cambrai*, *Le Chevalier au lion de Chrétien de Troyes*, *Le Roman de Tristan de Thomas*. Pour le 16e siècle, nous avons choisi *L'Heptaméron* de Marguerite de Navarre et les *Nouvelles récréations et joyeux devis* de Bonaventure des Périers. Les œuvres retenues pour le 17e siècle sont *Epigone* de l'abbé de Pure, *L'Astrée* d'Honoré d'Urfé, *Le Roman comique* de Scarron et *Le Roman bourgeois* de Furetière. Enfin la sélection pour le 18e siècle comporte *Jacques le Fataliste* de Diderot, *Les Liaisons dangereuses* de Choderlos de Laclos, *Gil Blas de Santillane* de Le Sage et *Candide* de Voltaire. L'opération d'étiquetage et de mise en relation des termes identifiés comme pertinents avec les mots-clés, s'est faite à l'aide d'un outil conçu sur mesure dans Voyant. Voyant permettait déjà de téléverser un document numérique et de générer de façon automatique une concordance à partir d'une séquence de recherche ou d'une liste de termes. Ce qui est nouveau c'est la possibilité de catégoriser chaque occurrence et d'indiquer des termes qui contribuent au choix de la catégorie (figure 1).

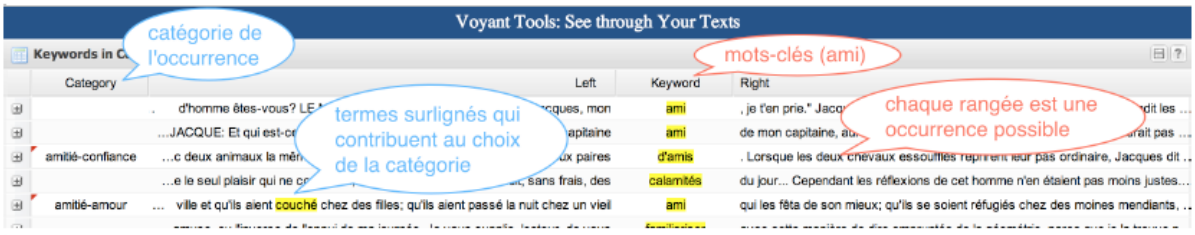

Figure 1. Capture d'écran de l'outil Toucher dans Voyant. Chaque rangée (ligne de texte) représente une occurrence possible du mot-clé (« ami » et ses variantes). Les bulles bleues indiquent des aspects que le chercheur peut modifier : la sélection d'une catégorie pour l'occurrence (ou on peut la laisser vide au besoin) et la possibilité de sélectionner des mots qui contribuent au choix de la catégorie.

À titre d'essai, Stéfan Sinclair et Madeleine Jeay ont identifié trois catégories à partir desquelles procéder à l'étiquetage : amitié / amour ; amitié confiance ; amitié trahison. Nous avons décidé de commencer par une liste très restreinte pour simplifier le choix et la gestion des résultats. Cependant, le passage à l'étiquetage a permis de constater que ces catégories ne

suffisant pas, il faudrait identifier une gamme de situations narratives qui soit représentative sans les multiplier et ainsi compliquer leur gestion. Une première sélection a été faite à partir d'une simple recherche du terme « ami » dans les textes du corpus d'expérimentation. Un ensemble de sept situations s'est dégagé que nous concevions tout d'abord à la façon de catégories :

amitié /amour
confiance / trahison
obtenir amitié
entretenir amitié
aider ami
conséquences de l'amitié
formes variées de l'amitié.

Après discussion au sein de l'équipe sur les limites de toute forme de typologie et donc des catégories qui en découlent, il a été décidé de travailler à partir des situations identifiées, même si certaines d'entre elles ont une portée circonscrite à une période ou un genre, par exemple la relation avec l'ami charnel du récit médiéval. Ainsi les éléments de lexique à étiqueter dans les textes seront mis en relation avec une série de mots clés qui correspondent à ces situations. Voici les mots-clés de base à partir desquels nous avons travaillé[9] :

- amour
- amitié, hostilité
- demander, obtenir, offrir, accorder, renouveler, refuser
- confiance, mériter, douter, prouver, pardonner, trahir
- éprouver / souffrir
- aider, conseiller, consoler, secourir, venger, remercier
- serviteur
- rompre, abandonner, oublier, se souvenir, mort
- animal
- charnel
- lecteur
- maître

Cela représente un bel exemple d'un processus de recherche en évolution, guidé par l'expérimentation concrète : nous avons commencé par trois catégories, ensuite nous avons essayé avec une liste de sept situations, pour enfin terminer avec la notion d'une douzaine de catégories de mots-clés qui peuvent entrer en relation dans une même occurrence. Il reste maintenant à passer à l'étape de l'essai de l'étiquetage automatique ou tout au moins de l'identification de situations similaires à celles qui ont été relevées manuellement sur le corpus d'expérimentation. Toutefois, dès à présent on peut faire un certain nombre

[9] Noter que certains ont pu être ajoutés en fonction des textes.

d'observations, dans l'ensemble de ce corpus, sur les occurrences de configurations narratives relatives à l'amitié, qui n'auraient pas été possibles sans ce premier traitement informatique des textes qui le composent. On observe par exemple l'importance que prend dans trois des textes, avec la question de mériter l'amitié, celle de la confiance et de la trahison. Dans *L'Heptaméron*, l'accent est surtout mis sur le mérite, sur le fait d'accorder son amitié, de la prouver et éventuellement de la renouveler. Dans *Jacques le fataliste* et Les *Liaisons dangereuses*, les problématiques dominantes sont celles de la confiance et de la trahison. Étonnamment, des textes médiévaux comme *Amadas et Ydoine* et *Raoul de Cambrai* leur accordent moins d'importance, le premier plutôt centré sur la souffrance et le second sur les relations entre l'amitié et l'amour. Dans un texte aussi vaste et tentaculaire que *L'Astrée*, l'étiquetage et la mise en évidence des situations relatives à l'amitié grâce à Voyant a permis de mettre en valeur la complexité des intrigues et des cheminements[10]. Ce travail a révélé que l'amitié dans cette œuvre n'est pas simplement ce que l'on recherche en tentant d'obtenir l'affection de quelqu'un, mais quelque chose qui ouvre sur de nombreuses possibilités pour les personnages tout en déterminant en partie le rayonnement de leur action, puisqu'ils se doivent d'honorer et de servir leur amis. Comme on le constate dans *L'Heptaméron*, l'amitié n'y est jamais acquise. Elle doit être constamment méritée, prouvée, renouvelée, ce qui implique de nombreuses obligations morales qui dictent le parcours des personnages et les alliances qui se forment au fil du roman. On voit donc apparaître, grâce à l'opération d'étiquetage, non seulement des réseaux de sens qui permettent de préciser la signification que prend l'amitié dans *L'Astrée*, mais aussi des champs d'action dynamiques et propres à chaque personnage. Ce bref échantillon a suffi pour montrer les possibilités qu'offre la comparaison entre les textes, telle que la permet un outil comme Voyant.

Il vaut la peine d'insister sur la nature hybride de notre méthodologie : nous faisons une lecture interprétative de chaque occurrence d'amitié ce qui permet de les catégoriser et d'observer des phénomènes à l'échelle du texte entier, ainsi que de comparer ces phénomènes à travers un corpus encore plus large. Notre démarche nous permet d'appréhender et d'étudier le fonctionnement d'une thématique (en l'occurrence, l'amitié) à divers niveaux. À titre d'exemple, nous pouvons produire une visualisation de la densité des occurrences d'amitié par catégorie et observer la distribution de certaines catégories (figure 2).

[10] Les observations qui suivent à propos de *L'Astrée* reprennent les conclusions de Renaud Roussel qui a procédé à l'étiquetage de l'œuvre.

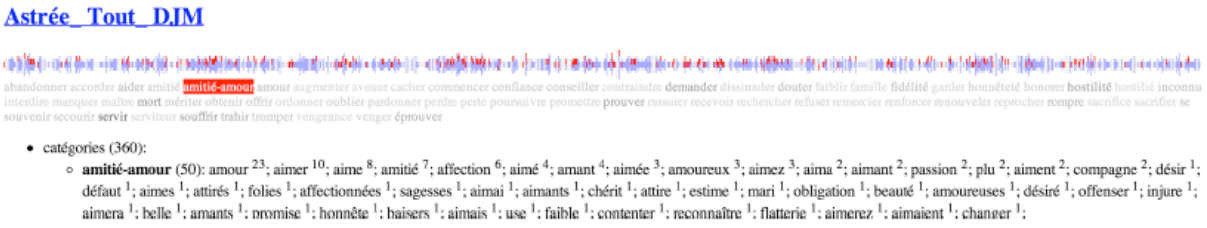

Figure 2 Une visualisation simple des occurrences d'amitié dans *L'Astrée* – les lignes bleues et rouges représentent la distribution des occurrences (une ligne plus longue indique plus d'occurrences) et le rouge indique la catégorie « amitié-amour » qui est surlignée. L'outil permet aussi d'énumérer les mots (avec fréquence) qui ont contribué à la catégorisation.

Tester la démarche d'étiquetage à l'aide de Voyant avec un terme comme « ami » présente plusieurs avantages majeurs qui, dans cette première étape du projet, ont simplifié l'opération. Le premier était de contourner l'obstacle des différents états de langue, puisque du Moyen Âge au 18e siècle, les variantes orthographiques sont mineures, se limitant à la voyelle finale. Par ailleurs, il suffit de ce seul mot pour solliciter les situations narratives relatives à l'amitié. L'inconvénient de ce choix vient du fait que si les occurrences dans les textes renvoient à des situations narratives, elles ne constituent pas par elles-mêmes des événements narratifs ressorts de l'action, comme le sont des événements tels que les scènes de duel, de violence ou les multiples épisodes de tempête et de naufrage. La difficulté qui se présente dans ces cas tient, d'une part, à celle d'identifier à quels termes clés relier ces situations, et de l'autre, à la variabilité orthographique de ces termes. Il est évident que pour ce type de configuration narrative, il sera impossible de travailler à partir d'un seul terme.

Dans tous les cas, la poursuite du travail doit se faire en lien avec PBLit afin de pouvoir 1) interroger l'archive entière des textes disponibles ; 2) mettre en corrélation, grâce à la recherche en cooccurrence, les termes qui se situent dans le même environnement co-textuel, sans être contigus. On pourra par exemple, pour rester dans la topique de l'amitié, repérer dans l'ensemble du corpus les occurrences de « accorder » ou « donner » et « amitié » ou bien d' « amitié / ami » et « amour ».

On le comprend, si l'analyse du vaste corpus que représente la littérature narrative du Moyen Âge au 18e siècle ne peut se faire que de façon collaborative, elle ne peut également s'envisager que par la mise en connectivité de plusieurs outils informatiques. Ainsi, dans le

cadre du projet TOUCHER, on ne peut espérer de résultats que grâce à la circulation entre Zotero, Voyant et PBLit.

Stéfan Sinclair et Madeleine Jeay

Bibliographie

« Historique de la Sator », http://www.satorbase.org. [Voir où l'on trouvera des informations sur la définition du topos narratif].

JEAY, Madeleine et Stéphan SINCLAIR, « L'exploitation des bases de données en littérature : l'approche de PBLit », dans "Por s'onor croistre", *Mélanges de langue et de littérature médiévales offerts à Pierre Kunstmann*, Ottawa, Éditions David, 2008 p. 443-455. [Y. Lepage et C. Milat (éds.)]. [Voir, http://tapor.mcmaster.ca/pblit/. Pour une description plus détaillée de PBLit].

RAMSAY, Stephen, « Toward an Algorithmic Criticism », *Literary and Linguistic Computing*, 18 (2003), p.1647-174.

RAMOS, Jean-Marc, « La Sator face au Golem. Usages et représentations des instrumentations numériques dans une communauté savante », dans Nathalie FERRAND (dir.), *Locus in fabula. La topique de l'espace dans les romans de langue française jusqu'en 1800*, Louvain-Paris, Peeters, 2004, p. 643-677

RASTIER, François, *Arts et sciences du texte*, Paris, Presses universitaires de France, 2001, p. 46

Voyant Tools (voyant-tools.org) succède à aux systèmes d'analyse HyperPo et de Taporware (pour plus de détails sur sa génétique, voir http://hermeneuti.ca/voyeur/background)

Le Thesaurus n'est pas encore accessible publiquement ; http://tapor.mcmaster.ca/~hyperliste