Hyperbase
Un logiciel pour l'analyse textuelle

Séminaire d'Analyse du texte fantastique :
Littérature et informatique

Ada Myriam Scanu
Università degli Studi di Bologna

Sommaire

Introduction

L'association des termes littérature et informatique ne m'a jamais choquée. Pour écrire un texte on repère du matériel sur le réseau Internet et puis on l'étend à l'aide des logiciels d'édition, comme Word de Microsoft. Toutefois, l'association qu'on me proposait ici était de tout autre niveau. Il s'agissait d'étudier d'abord et de vérifier ensuite si certains logiciels expressément créés pour l'étude littéraire étaient aptes aux buts que l'on se proposait. Je dois avouer ma méfiance, je me disais que jamais une machine ne pourrait se substituer à l'esprit humain et qu'à la limite elle pouvait prendre en charge les mansions les plus ennuyeuses du travail du chercheur et les accélérer. Mais les études que j'ai lues, concernant l'informatique humaniste et l'application de certains logiciels, ont contribué à me faire comprendre qu'ici l'enjeu était fort différent. La création des infologiciels nécessite la collaboration des deux mondes, celui de l'informatique et celui de la documentation, pas de manière simplement complémentaire, mais vraiment structurelle. Les problèmes liés à l'indexation ou à la codification des textes afin de les soumettre à une analyse informatisée, nécessitent une stricte collaboration entre les deux domaines. Le travail du chercheur est soumis aux règles informatiques, mais le logiciel a absolument besoin des directives du chercheur pour pouvoir fonctionner correctement. La structure informatique du logiciel se compose de tags et de codes que difficilement un littéraire pourrait connaître, toutefois sans la présence de ce dernier le programme ne pourrait rendre aucun résultat.

La surprise qu'à créée en moi le fonctionnement de ces logiciels qui ont la capacité de fournir une étonnante quantité de données et de résultats en peu de temps, s'est unie à la surprise pour la facilité avec laquelle on peut commettre des erreurs dans la phase préliminaire du processus qui pourraient fausser les données. Ce manque, s'il en est un - n'importe quel processus automatisé est réglé par l'homme- contribue à unir, dans le cas des analyses quantitatives et statistiques du langage écrit ou oral, la machine et l'homme dans un *unicum* inséparable. Ceci insère dans la démarche analytique l'élément « erreur », dérivant de l'intervention nécessaire de l'homme dans ce

processus. L'ordinateur alors s'humanise, devenant ainsi faillible par définition et se transforme en un « multiplicateur de bêtises »[1].

L'informatique s'humanise, faisant trésor des connaissances humaines et prenant une petite tare indispensable à son humanité, tandis que l'homme s'automatise, s'appuyant de plus en plus sur l'aide des machines et modifiant ses méthodes de recherche. Cette union est bien l'un des buts de cette discipline relativement récente qui s'appelle Informatique Humaniste et qui tend à changer la mentalité des littéraires et des informatiques au biais d'une collaboration féconde qui permet une meilleure compréhension des mécanismes du langage et de l'écriture d'un côté, et un perfectionnement dans les outils informatiques de l'autre. Et naturellement, c'est sur cette union que se basent les recherches sur l'Intelligence Artificielle[2].

Ce que je voudrais montrer dans ce bref travail tient tout à fait de ce principe. Ayant pris un petit recueil de nouvelles fantastiques éditées dans une collection didactique, j'ai voulu voir si une analyse au biais d'un logiciel pouvait être de quelque aide à la compréhension de ces textes et fournir aux lecteurs des indications supplémentaires, surtout concernant la structure des textes et pas seulement leurs aspects linguistiques. Malheureusement, des limites dues à l'utilisation d'une version d'évaluation pour le logiciel considéré, ne m'a pas permis d'effectuer une analyse complète, m'obligeant à couper le corpus en deux bases différentes, et à limiter mes actions.

Un lecteur a, devant un texte, deux possibilités de lecture : soit il s'arrête à l'aspect matériel du texte, donc par exemple au lexique, soit il perce plus en profondeur, cherchant à remonter, à travers des traces et des indices véhiculés par les mots, à la structure du récit.[3] Malgré donc une marge possible d'erreur dans les résultats, en

---

[1] SPINA, S., *Fare i conti con le parole, Introduzione alla linguistica dei corpora*, Perugia, Guerra Edizioni, 2001, p. 14.

[2] À ce propos, cf. les deux définitions qu'en donne Graziella Tonfoni « a) L'Intelligenza Artificiale è quell'area delle Scienze Computazionali che intende esplicitamente scoprire i meccanismi del pensiero, della creatività e del linguaggio, di quegli elementi insomma che fanno degli uomini degli esseri intelligenti. b) L'Intelligenza artificiale è quel settore delle Scienze Computazionali che intende costruire sistemi intelligenti ovvero che esplichino le proprie attività in modo che diremmo "intelligente"», G. TONFONI, *Intelligenza artificiale : automazione dei processi cognitivi e sistemi per la rappresentazione della conoscenza*, in «Lectures», monografie interdisciplinari dirette da R. Campagnoli, V. Carofiglio, Y. Hersant, A. Ponzio, nr. 19, *Automi*, 1986/2, Bari, Edizioni del Sud, pp. 77-95:77.

[3] Cf. la notion de "sintonizzazione" dont parle Graziella Tonfoni dans son article *Intelligenza artificiale memorizzazione e lettura*, in ID, *Sistemi cognitivi complessi, intelligenza artificiale e modelli di organizzazione della conoscenza*, Treviso, Pagus Edizioni, 1991, pp. 59-85:68.

partant de quelques contes, j'ai voulu proposer une ébauche de structure des contes fantastiques qui se base sur une recherche de lien entre quelques mots qui composent le texte et le genre littéraire de ce dernier.

Comme Hoey nous le fait remarquer – ce qui est très vrai dans le cas d'une narration fantastique - « Even the most unlikely of words is found to have secrets »[4].

1. Informatique et Linguistique

1.1 Quelques traits préliminaires

On a vu comment l'informatique pour les humanistes s'appuie sur une stricte collaboration entre le savoir informatique et les connaissances des sciences humaines, et vise à un rapprochement de ces deux domaines qui, en vérité, ne sont pas si éloignés que l'on pourrait penser à première vue. La linguistique est l'une des disciplines qui est unanimement considérée comme l'une des plus proches au monde des ordinateurs. En effet, toutes deux utilisent le langage comme moyen de communication, mais ce langage est aussi l'objet de l'existence même des disciplines, c'est-à-dire leur objet d'étude. Elles se basent sur le traitement de l'information et entrent en jeu quand l'information a déjà été créée. Elles ont comme but d'enquêter la manière avec laquelle des entités différentes communiquent entre elles utilisant des langages structurés et effectuant continuellement des transcodifications de l'une à l'autre pour envoyer des messages[5]. Ces deux disciplines se rapprochent aussi par leur méthodologie, qui se caractérise par ces trois mots d'ordre : repérer, compter, ranger. Le linguiste recherche des informations, compte ses résultats et les distribue à l'intérieur d'un corpus de référence ; de même la machine reçoit des informations qui lui proviennent sous forme d'inputs, les compte pour les comprendre (c'est-à-dire les transforme dans son langage qui se base sur des séquences numériques) et les range afin de pouvoir les comprendre instantanément la fois suivante. Et sur ces trois mots-clefs se basent les instruments que

[4] HOEY, H., *From concordance to text structure, New uses for computer corpora*, in LEWANDOWSKA-TOMASZCZYK, B., MELIA, P.J. (eds.), *PALC'97 Practical Applications in Language Corpora*, Łódź, Łódź University Press, pp. 2-23.
[5] SPINA, S., *Fare i conti con le parole, Introduzione alla linguistica dei corpora*, Perugia, Guerra Edizioni, 2001, p. 5.

la linguistique informatique a pu mettre au point. En particulier, il existe quatre familles principales de logiciels :

1. les gestionnaires de ressources linguistiques, qui créent et mettent à jour notamment les dictionnaires, les lexiques ou les réseaux sémantiques.

2. les moteurs d'indexation et de recherche : le premier analyse le contenu des documents afin de créer divers fichiers d'index pour la mise en œuvre des différents traitements linguistiques ou statistiques, les seconds renvoient aux documents désirés sous forme de recherche, en langage naturel ou booléien.

3. les outils de représentation des connaissances, qui, s'appuyant sur des traitements linguistiques ou statistiques, produisent des cartes de connaissance qui représentent visuellement le contenu informationnel d'un corpus de documents.[6]

4. les analyseurs linguistiques qui décomposent des textes de n'importe quelle typologie et les soumettent à des enquêtes définies. En particulier l'on peut effectuer :

a. une analyse morpholexicale, qui se donne comme objectif l'identification des mots d'un texte. Après le découpage du texte en mots, ceux-ci sont décomposés en morphèmes ; les mots sont lemmatisés, et les lemmes obtenus sont comparés au lexique de l'application afin de trouver la forme canonique correspondante. Chaque terme est alors étiqueté en fonction des données du lexique[7].

b. une analyse syntaxique, qui étudie la syntaxe et la structure grammaticale de la phrase et étiquette les mots sur la base de leurs fonctions.

c. une analyse sémantique, qui recherche à l'intérieur des mots des combinaisons qui identifient un sens au-delà des termes *stricto sensu*.

Les analyses linguistiques peuvent s'effectuer sur un seul texte, ou bien sur plusieurs textes, composant ainsi un corpus textuel. L'emploi des méthodes statistiques pour l'étude de la linguistique a donc aussi le but de mettre en valeur de nouveaux aspects d'un corpus, de permettre des rapprochements immédiats de plusieurs textes et d'en multiplier sensiblement les possibilités de lecture.

---

[6] CHAUMIER, J., DEJAN, M., *Recherche et analyse de l'information textuelle, Tendances des outils linguistiques*,
« Documentariste. Science de l'information », 2003, vol. 40, n. 1, p. 15.
[7] *Ibidem.*

1.2. Corpus linguistiques et l'importance d'un corpus.

La linguistique statistique a comme but général d'identifier les structures mathématico-statistiques propres au langage, considéré en tant que phénomène objectif.[8] Les modèles computationaux et les modèles quantitatifs, finissent par déboucher dans la linguistique nommée *corpus based*. La *corpus linguistics* fonde ses enquêtes sur des recueils de données linguistiques suffisamment représentatives de n'importe quelle typologie et y recherche des phénomènes particuliers afin de garantir une réponse efficace. Un corpus est défini comme

> A collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis[9].

ou encore,

> Computer corpus : a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks[10].

Il existe différents types de corpora, généraux, spécialisés, de référence, d'archives, de textes, d'échantillons. Il y en a de différentes dimensions et en plusieurs langues, ce qui permet d'avancer dans la mise aux points des ainsi dits *cross-corpora*, destinés surtout à la traduction et à l'analyse comparative des langues.

La création d'un corpus est un passage fondamental. Déjà à ce moment-là, le but de la recherche se trouve figé et représente le moment où commence la communication entre homme et machine.

Il est possible de repérer trois phases dans le procédé de création d'un corpus :

---

[8] ROSSINI FAVRETTI, R. (a cura di), *Linguistica e informatica, Corpora, multimedialità e percorsi di apprendimento*, Milano, Bulzoni Editore, 2000, p. 15.
[9] FRANCIS, W.N., *Problems of assembling and computerizing large corp*ora in Johansson S. (ed.), *Computer Corpora in English Language Research*, Bergen, Norwegian Computing Centre for the Humanities, 1982, p. 7.
[10] Expert Advisory Group on Language Standards, *EAGLES Guidelines*, in SPINA, S., *op. cit.*, p. 64.

1. le projet, c'est-à-dire le choix des dimensions (ou du nombre d'occurrences), des textes et de leur typologie et des critères de l'analyse.
2. l'acquisition des données qui dépend de la nécessité de rendre les textes en MRF (*Machine readable form*) à travers la frappe, la scansion, la dictée… Dans ces deux phases la supervision et le contrôle de la part de l'homme sont inévitables pour une réussite possible de la recherche.
3. la codification des données : dans le passage de la forme écrite à la MRF, un texte perd inévitablement quelques caractéristiques, ce qui dérive de la nécessité pour le traitement des textes à l'ordinateur de l'utilisation du code ASCII (*American Standard Code for Information Interchange*) qui représente 128 caractères alphanumériques (256 dans la version agrandie), mais ne tient pas compte de la dimension textuelle dans sa différenciation (les gras, les italiques par exemple). La recherche d'un standard commun pour la codification électronique des textes a favorisé en 1987 la naissance du TEI (*Text Encoding Initiative*)

> The Text Encoding Initiative (TEI) is an international cooperative research effort, the goal of which is to define a set of generic *Guidelines* for the representation of textual materials in electronic form[11].

Dans les *Guidelines* du TEI on remarque la nécessité d'utiliser des langages de marcature comme le SGML (*Standard Generalized Markup Language*) et le XML (*Extensible Markup Language*) pour la codification des textes. L'annotation consiste dans l'application d'étiquettes spéciales (tags) à des sections prédéterminées du texte, afin de les distinguer des autres. Ces langages s'utilisent pour mettre en évidence les éléments graphiques, textuels, phonétiques, syntactiques ou prosodiques du texte qui ne rentrent pas dans la première codification en ASCII. Un exemple d'annotation est représenté par le POS tagging (*Part of Speech Tagging*) qui donne des étiquettes grammaticales aux mots.

[11] TEI *Guidelines*, in SPINA, S., *op. cit.*, p. 81.

> Annotation : the way in which we […] enrich a corpus with all sorts of markup, and as a part of the annotation drive the attempts to provide standards for document structure and formatting, so that we can make sure that our documents conform to a uniform standard[12].

Une fois que le corpus est créé et codifié, il constitue la base de toute étude quantitative et statistique du langage. La linguistique *corpus-based* a depuis longtemps choisi son domaine de recherche préféré : la lexicographie.

Un premier type d'information numérique que l'on fait ressortir au biais de l'interaction entre corpus et logiciel d'analyse est constitué par la fréquence des éléments qui le composent et par la constitution de listes de fréquence. Une distinction dans l'étude des fréquences s'opère entre les occurrences (ou *tokens)* et les formes distinctes (*types*) qui composent le texte. La fréquence dérive justement du rapport entre types et tokens, entre les occurrences totales d'un texte et le nombre de formes distinctes. Plus le rapport types/tokens est bas, plus le vocabulaire du texte est pauvre et répétitif. Toutefois, cette relation n'est pas infaillible, du moment que les mots sont coupés de leur contexte, qu'elle ne tient pas compte de leurs fonctions grammaticales et de la présence de possibles homographes (ou homophones dans le cas de textes oraux).

Une forme d'analyse quantitative des corpus linguistiques est représentée par la concordance, qui est la liste des occurrences d'une ou de plusieurs formes montrées à l'intérieur du contexte du corpus.

> A concordance is a collection of the occurrences of a word-form, each in its own textual environment. Each word-form is indexed, and a reference is given to the place of each occurrence in a text[13].

Le contexte n'est pas considéré du point de vue sémantique, mais linguistique, c'est-à-dire les quelques mots qui précèdent et suivent le mot étudié. Les éléments pris en considération sont les mots (*keywords*) et leur contexte (les quelques mots qui suivent et précèdent). Le résultat est représenté par le KWIC (*Keyword in context*) – qui s'oppose au KWOC (*Keyword out of Context*). Le choix des keywords est naturellement

[12] SINCLAIR, J., *Current Issues in Corpus Linguistics*, in ROSSINI FAVRETTI, R., (a cura di), *Linguistica e informatica, Corpora, multimedialità e percorsi di apprendimento*, Milano, Bulzoni Editore, 2000, p. 29.
[13] SINCLAIR, J., *Corpus, Concordance, Collocation*, Oxford, Oxford University Press, 1991, p. 32.

un passage fondamental, qui dérive d'une connaissance exacte de l'objet recherché. Il existe des programmes spécifiques pour la relève des concordances, comme par exemple Monoconc, ConcApp, Wconcord.

La collocation s'entend comme un enclenchement de deux types de lemmes ou de mots qui se rencontrent de manière fixe et systématique créant ainsi un concept unitaire et précis. Cette notion relève du principe idiomatique, *idiom principle*, identifié par John Sinclair, selon lequel un locuteur a à disposition un nombre d'unités préconstituées composées de parties différentes qui forment un bloc unitaire. Selon une autre définition, la collocation représente « la regolare cooccorenza di due o più parole di solito una vicina all'altra in un enunciato o in enunciati prossimi »[14]. Le point de départ de l'analyse des collocations est l'étude des mots et de leur fréquence. À partir de ces données, on sélectionne les formes où certains mots sont cooccurants et on en calcule la fréquence. On compare ensuite les fréquences obtenues avec celles d'autres lemmes occurrents avec le mot choisi et on donne un calcul statistique du poids de cette cooccurence. D'importance primaire est le choix du terme à analyser et du *span*, c'est-à-dire le nombre de mots à droite et à gauche que l'on considère. Le MI (*Mutual Information*) est le coefficient qui mesure la quantité d'informations que l'on tire de la mesure de l'occurrence de deux mots et indique comment les mots sont strictement liés l'un à l'autre, calculant la probabilité que la présence du mot choisi pèse sur l'occurrence du *collocate.* Le *t-score* est un autre coefficient qui mesure le degré de fiabilité pour pouvoir affirmer qu'il existe une association entre des mots et indique si le lien et la force d'attraction entre deux formes sont justifiés ou non.

Un autre type d'analyse se fait sur les *clusters* qui représentent des segments répétés de mots à l'intérieur d'un texte. Il s'agit d'un concept moins ample que la cooccurence et comprend des mots dont la fonction grammaticale est définie en fonction de l'ensemble de ses membres, et non par les simples parties qui le composent.

Et encore, l'on peut calculer l'index de lisibilité d'un texte qui a son point de départ dans la règle économique de Zipf (1949), et relève des règles pour déterminer le degré de compréhension d'un texte.

Naturellement, ces fonctions d'analyse se basent sur des commandes déterminées par le logiciel qui les contient, mais elles s'appuient à des instruments

[14] BECCARIA, G.L., *Dizionario di linguistica e di filologia, metrica e retorica*, Torino, Einaudi, 1996.

préalablement créés, comme les dictionnaires lexicographiques, terminologiques et les réseaux sémantiques représentés notamment par les thésaurus.

2. De la Lexicographie à l'analyse textuelle

2.1. L'Analyse Textuelle

Quand on analyse un texte, il ne suffit pas de connaître la langue pour le comprendre, mais il faut nécessairement avoir des connaissances préacquises. Ces connaissances, si elles sont appliquées à l'analyse d'un texte préalablement assisté par l'ordinateur, peuvent truquer les résultats. Celui qui les interprète, les lie à son savoir et les données d'objectives deviennent subjectives. Cette accusation de l'infiltration subreptice d'un savoir personnel à l'intérieur de la démarche analytique est sûrement vraie mais, comme on l'a vu au tout début, elle est intérieure au processus de communication entre homme et machine.

> Computer assistance does not bring pure objectivity to text analysis. It is evident that intuition is involved in several stages : which features to study, how delicately to code, how to interpret the findings. It has long been widely recognized that stylistic statistics merely provide quantitative evidence whose significance can be assessed only by experience and common sense[15].

L'acceptation de cette condition permet de justifier l'analyse textuelle assistée par ordinateur qui, de même manière, se base sur des procédés le plus possible objectifs.

If we have to make a mathematical study of an author's style we must fist identify features of his writing which can be precisely quantified : countable events and measurable magnitudes. Such features are not difficult to find. The length of words and sentences and paragraphs can be measured ; it is possible to count the frequency of vocabulary items or syntactic constructions or rhetorical devices.[16]

---

[15] STUBBS, M., *Text and Corpus Analysis, Computer-assisted Studies of Language and Culture*, London, Blackwell Publishers, 1998, p. 154.
[16] KENNY, A., *The Computation of style, An Introduction to Statistics for Student of Literature and Humanities*, New York, Pergamon Press, 1982, p. 15.

Chaque mot, chaque longueur et position ont un sens à l'intérieur d'un texte, et notamment en littérature ; à la machine le rôle de les repérer, au chercheur celui de les interpréter.

2.2. Une approche : Hyperbase

Hyperbase est un logiciel pour le traitement documentaire et statistique des corpus textuels créé par Étienne Brunet. Il est rédigé en langue française, mais il s'applique indifféremment à toute langue qui utilise l'alphabet latin. La version actuelle comprend aussi trois dictionnaires de référence, français, anglais et portugais.

La version que j'ai utilisée ici est celle de l'évaluation de la version 5.4. pour Windows. Ce programme crée dans le disque dur de l'ordinateur une base de données, composée de textes de n'importe quelle typologie, et en permet l'analyse multiple. L'on peut comparer différents textes, mais aussi différentes bases de documents et l'interrogation s'avantage des fonctions hypertextuelles du logiciel : il suffit de cliquer sur un mot pour connaître sa répartition dans le corpus de travail et être conduit dans les passages où le mot se trouve employé. Ces excursions verticales peuvent se faire aussi à partir de l'index, c'est-à-dire de la liste alphabétique à laquelle un menu déroulant donne accès dans la page d'entrée.[17] L'hypertextualité du logiciel se montre aussi dans la possibilité de se connecter à Internet pour comparer les résultats avec d'autres bases, notamment celles dédiées à Rabelais et à Balzac, disponibles à l'adresse http://lolita.unice.fr

Le menu du sommaire, décoré par une belle image naturelle, propose horizontalement les fonctions documentaires (contexte, concordance) et verticalement les fonctions statistiques.

L'utilisateur peut tout d'abord s'appuyer sur une base textuelle déjà créée, comprenant des textes littéraires français, comme *La Marianne, Zadig* ou l'*Émile* et étudier de cette manière les différentes fonctions du logiciel. Ou bien, il peut créer sa propre base. Naturellement cette phase nécessite d'un travail a priori qui est la normalisation et la codification des textes à soumettre.

[17] BRUNET, É., *Hyperbase, Logiciel hypertexte pour le traitement documentaires et statistiques des corpus textuels, Manuel de Référence*, version 5.4. pour Windows (janvier 2002), p. 9.

2.2.1 Créer une base

Tous les textes qu'on souhaite utiliser, doivent être transformés en MRF utilisant le code ASCII, donc format texte seulement. Ceci comporte que les différentes parties du texte doivent être représentées à travers des codes. Pour distinguer les numéros des pages on utilise le signe $, mais si les pages n'ont pas été distinguées, le programme procède au découpage automatique à raison de 200 mots au moins par page, en s'abstenant de couper les paragraphes. Ces derniers sont délimités par le retour du chariot, mais s'ils sont trop longs, Hyperbase se permet de les découper en unités plus petites, en s'abstenant toutefois de couper la phrase. Pour discerner les phrases, le programme détecte la ponctuation forte, principalement le point.[18] Il existe aussi une Option Vers qui permet le traitement de la poésie. Pour ce que concerne le traitement des mots, le programme distingue les accentués des non-accentués, les majuscules et les minuscules et possède un traitement des noms propres qui relève tous ces mots auxquels la majuscule est attachée à leur nature et non due à leur position en début de phrase. La définition des mots est dépendante de la liste établie pour les séparateurs, laquelle outre les blancs, le retour du chariot et la tabulation comprend aussi de nombreux signes de ponctuation. Aux phases de contrôle, de l'importation et de reformatage des données, suit l'indexation, dont l'algorithme de définition est dû à Jean Pierre Anfosso.[19]

La dernière étape du traitement consiste à comparer le corpus traité au corpus littéraire de FRANTEXT qui comprend 117 millions de mots et s'étend sur cinq siècles. On a la possibilité d'utiliser le corpus en entier ou bien de se limiter à une tranche temporelle. Dans le cas d'une langue étrangère, un dictionnaire de référence propre à la langue choisie peut être substitué au fichier existant.

Chaque fois qu'un nouveau texte est introduit dans la base, le logiciel procède à la reconnaissance et à l'analyse des formes, à l'indexation proprement dite, à la création d'un dictionnaire de fréquences alphabétique et hiérarchique, permettant la visualisation en tableau, de plus, il calcule les spécificités externes et internes, les coefficients de corrélation, et encore, il mesure la richesse lexicale, l'accroissement du vocabulaire, la proportion des hapax.

[18] BRUNET, É., *op. cit.*, p. 14.
[19] *Ibid.*, p. 15.

Une fois que la base est créée, l'utilisateur peut exploiter les différentes fonctions d'Hyperbase suivant le but de sa recherche.

2.2.3. Interroger les résultats

Pour ce qui concerne les mots, le chercheur a la possibilité d'en voir la fréquence, répartie dans le texte ou dans le corpus entier. À partir de ce mot, il peut voir la représentation graphique de sa fréquence en le comparant à d'autres mots et il peut en consulter un index hiérarchique ordonné par fréquence. Les fonctions hypertextuelles permettent de voir immédiatement le contexte du mot. À partir de cela on peut rechercher d'autres mots ou encore visualiser les écarts, mis en relief par la couleur rouge, qui signalent les spécificités du texte en reflétant ses caractéristiques par rapport au corpus.

Naturellement, le logiciel propose des outils propres à assumer une exploitation méthodique de la documentation. Le premier de ces outils est Contexte. Sollicitant cette option en rapport au mot, lemme ou expression désirés, l'utilisateur a la possibilité de les voir à l'intérieur des différents textes où ils sont employés. Il faut remarquer qu'Hyperbase effectue une équivalence automatique entre contexte et paragraphe. Toutefois, une spécification de la longueur désirée est possible. Si l'on clique sur un des passages affichés, on a la possibilité de voir la page entière où le mot apparaît. Voilà alors que la structure du logiciel devient claire : il procède en éventail, passant de la petite unité à la page et vice-versa. Naturellement ces allées et retours permettent une vérification ultérieure des résultats et une interrogation plus ample. Dans Contexte, il est aussi possible de considérer les cooccurrences de deux mots dans le même paragraphe et, il suffit d'un autre clic, sur toute la page pour élargir. Une autre fonction documentaire est représentée par Concordance, qui extrait un contexte étroit qui tient en une ligne et montre la forme cherchée, en position centrale[20].

Pour ces deux options, la recherche peut porter sur l'expression, sur le terme ou lemme (verbe à l'infinitif, adjectif et substantif), sur les débuts des mots, sur une chaîne de caractères, sur les fins de mots.[21] S'il l'on veut opérer une distinction entre les

[20] *Ibid*., p. 27
[21] *Ibid*., p. 29.

catégories des mots, il faut utiliser un Hyperbase lemmatiseur, qui permet de créer une liste de mots pour éviter des mélanges indésirables.

2.2.4. Les fonctions graphiques

Le logiciel Hyperbase s'appuie sur la représentation visuelle pour illustrer les fonctions statistiques. On a la possibilité de voir sous forme d'histogrammes les écarts entre la fréquence d'un mot observé dans le texte et la fréquence théorique qu'on aurait dû s'attendre en considérant la proportion du texte par rapport à l'ensemble. L'option double permet de représenter deux séries du même graphique pour laisser visibles deux séries d'écarts. Pour mesurer la force d'attraction mutuelle des deux mots, un calcul de corrélation est établi. Le coefficient de corrélation peut aussi être calculé en comparant, pour chaque mot, les valeurs de l'écart réduit au rang de chaque élément[22]. La recherche thématique, qu'on effectue grâce au programme de repérage automatique d'Hyperbase, permet de retrouver une accointance non seulement entre un mot et un texte, mais entre le texte et tous les mots qui peuvent se trouver dans l'entourage d'un mot, qu'on définit comme pôle.

Les listes aussi possèdent différentes représentations graphiques, en colonnes ou en lignes, et l'on peut les soumettre à deux types d'analyses, factorielle ou arborée.

D'autres outils permettent de voir l'évolution du langage, la richesse du vocabulaire et les distances lexicales d'un texte. C'est le programme factoriel qui calcule les spécificités internes de la langue. Et encore, Hyperbase permet l'approche par phrases-clés, c'est-à-dire des passages caractéristiques par lesquels on tente de donner une idée du contenu d'un texte.

On a vu en passant quels sont les outils que le programme Hyperbase offre pour effectuer des analyses statistiques sur des textes. Ce qu'on relève surtout dans l'utilisation de ce logiciel c'est la possibilité de passer toutes les fois que l'on désire des données au texte et vice-versa, rendant ainsi une double visualisation qui permet immédiatement d'éliminer les résultats faux ou inexacts.

[22] *Ibid.,* p. 42.

3. Une application d'Hyperbase

3.1. Quatre nouvelles fantastiques

Dans un petit recueil édité par Magnard, se trouvent quatre nouvelles définies dans le titre « fantastiques ». Ces textes et leurs auteurs sont connus justement pour avoir cultivé cette veine littéraire. Au bas des pages du recueil, qui s'adresse aux étudiants des lycées, on lit des notes qui relèvent des mots ou des expressions spécifiques du genre littéraire en question, et aident les étudiants à mieux comprendre le texte pendant la lecture. Ce que je me propose ici est de mettre en évidence à travers l'application d'Hyperbase comment un logiciel pourrait être utilisé à l'intérieur d'un parcours didactique pour aider les étudiants à mieux saisir le fonctionnement d'un texte et ses mécanismes et acquérir ainsi une différente vision de la page écrite : non pas un ensemble de mots qui donnent une signification, mais un ensemble spécifique et structuré de mots qui donnent une signification spécifique, relevée en perçant les mystères de la phrase et de sa structure.

3.2. Processus d'analyse

Les quatre nouvelles qui composent le mini-corpus que j'ai créé sont :

1. *La Nuit*, de Guy de Maupassant ;
2. *Deux acteurs pour un rôle* de Théophile Gautier ;
3. *Le cœur révélateur* de E. A. Poe (traduction française par Baudelaire) ;
4. *La perspective Nevski* de Nikolaï Gogol (traduction française par Sylvie Howlett)

Pour des raisons techniques, la version d'évaluation ne permet pas de créer des corpus avec plus de 100.000 occurrences. Cela a déterminé la décision de créer deux bases : la première (Fanta1) incluant les textes de Poe et de Maupassant, et la deuxième ceux de Gautier et de Gogol. Ces associations ont été faites en vertu de la longueur des textes, prenant deux nouvelles de la même longueur, ce qui a impliqué, dès le début une

remarque intéressante : les deux textes plus brefs étaient écrits à la première personne, les deux autres à la troisième.

Pour ce que concerne le corpus créé, il n'a y là aucune volonté d'exhaustivité. Il ne s'agit que d'un corpus d'exemple ou mieux d'une base de documents.

Une deuxième démarche a été la normalisation des documents. J'ai choisi de ne pas signaler les pages, mais de procéder à un découpage automatique selon les paragraphes, dont la répartition de l'original avait été respectée. À l'intérieur des documents, il s'est rendu nécessaire d'opérer des modifications. Le logiciel considère en effet les apostrophes à la même guise des mots, ce qui fausse l'analyse. Les *d', s', j'* acquièrent en effet un rang élevé dans la liste des fréquences. Le fait d'enlever les apostrophes n'a créé aucun problème au niveau par exemple des verbes réfléchis qui, de toute façon, ne sont pas reconnus par le moteur de recherche. Le logiciel ne reconnaît pas la forme *s'empêcher*, mais seulement *s'* ou *empêcher*. La présence d'une éventuelle particule pronominale devient évidente si l'on considère les concordances. Le fait d'enlever les apostrophes peut aider si l'on considère la forme *n'*. En devenant carrément *ne*, elle permet de voir par exemple le poids de la négation dans le texte ou l'usage des formules *ne…que*. Malgré l'ennui de la tâche manuelle, cette première démarche de normalisation oblige à lire le texte attentivement et à y entrer véritablement dedans en le retravaillant.

Une dernière remarque sur la codification des textes concerne l'utilisation du tiret, qui était utilisé indistinctement aussi bien pour signaler les dialogues qu'à l'intérieur des mots ou des locutions qui requièrent l'inversion sujet-verbe. Du moment qu'il fallait les signaler afin d'avoir des résultats corrects, j'ai pensé utiliser une codification différente pour distinguer entre dialogues et monologues. J'ai utilisé *pp* pour signaler les dialogues et *bb* pour mettre en relief les monologues ou bien les pauses que le narrateur fait à l'intérieur du discours qui, elles aussi, signalent en définitive une sorte de monologue. J'ai voulu faire cette distinction surtout en raison des textes pris en considération : *Le Cœur Révélateur* et *La Nuit*. En fait dans les deux textes, les signes soulignent une tension et méritent donc un même traitement que les mots.

3.2.1. Premier corpus : Poe et Maupassant

La première étape de l'analyse devrait consister dans la computation des fréquences des mots. Toutefois, vu que la similarité des deux textes saute aux yeux, on recherchera les spécificités des deux, en essayant de relever dans quel sens ils sont différents et s'ils possèdent ou non une même structure.

En observant les spécificités, l'on remarque ce que reporte la Fig. 1.

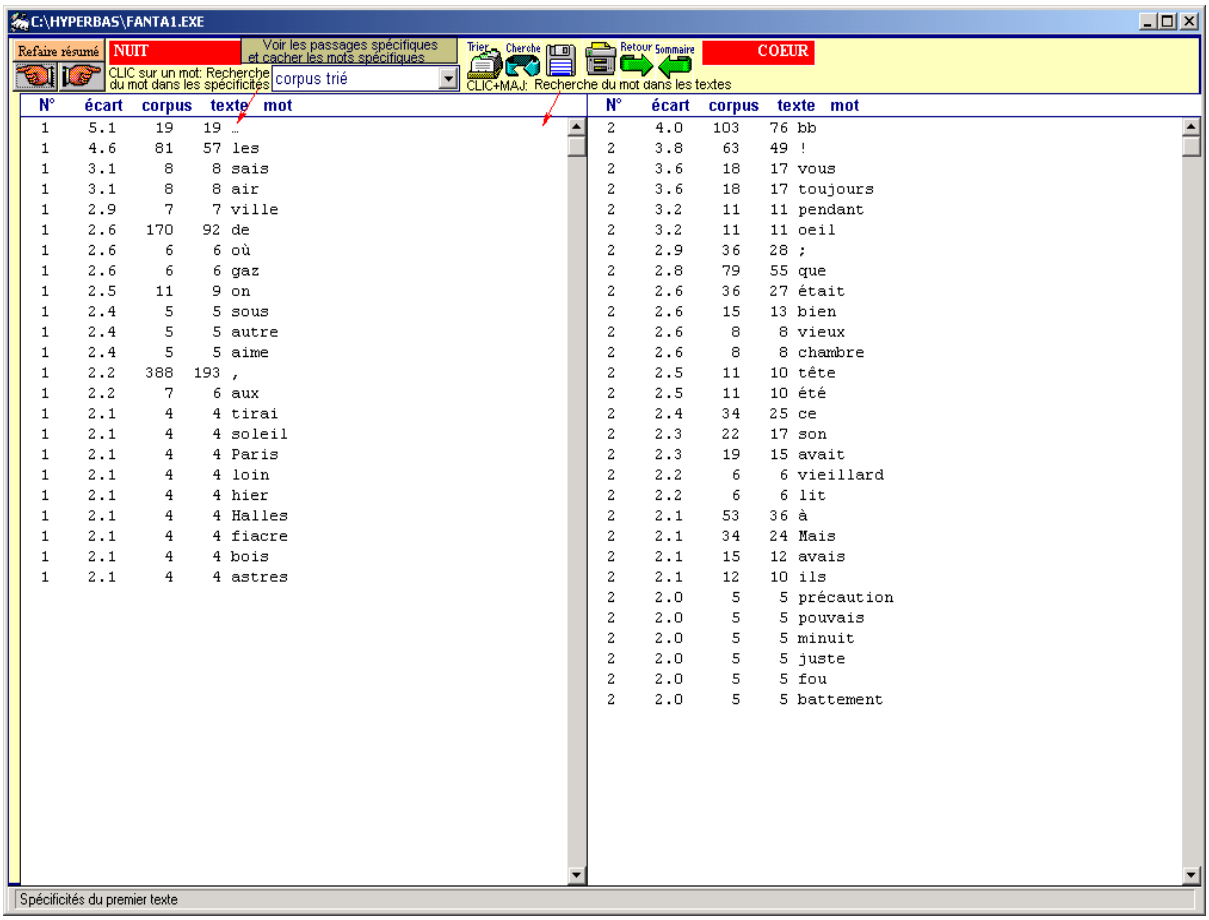

Fig. 1 : Liste des spécificités par ordre hiérarchique

Ces résultats permettent d'établir des oppositions assez marquantes entre les deux nouvelles.

Dans *la Nuit*, prédomine le mouvement : le protagoniste bouge, se trouve à l'extérieur. Les points de suspension, qui sont une prédominante, symbolisent aussi un mouvement, l'action est suspendue pour le lecteur, mais rien n'indique qu'elle s'arrête tout à fait. Remarquable de ce point de vue est aussi l'usage de la virgule (nr. 193), qui renvoie à une suite. Par contre, dans *Cœur Révélateur*, la dimension prédominante est

temporelle. Les adverbes temporels l'emportent, ainsi que les points d'interrogation et le point virgule, qui ont le rôle de permettre au lecteur (et à l'auteur) de faire une pause réflexive ou bien de prendre le temps pour s'étonner. La nouvelle ne s'ouvre donc pas à l'extérieur, mais tous se passe dans une chambre, où il y a des objets et où il y a des corps. Spécifiques de la nouvelle de Poe sont donc aussi des références au corps humain (tête, œil, coeur), groupe que l'on pourrait opposer au terme « air » de Maupassant et à tout le contexte sémantique qui le caractérise.

Ainsi déjà après une première analyse, l'on remarque ces oppositions : temps/espace, dedans/dehors, corps/non-corps. Mais l'élément qui réunit ces couples symboliques, se retrouve dans l'utilisation prédominante du pronom je (et des possessifs *mon* et réfléchis *me*) qui indiquent que même si l'on peut être à l'extérieur, ce qui compte est l'intériorité, le je. Naturellement, le fait de se concentrer sur soi-même aide à se détacher de la réalité et à se renfermer sur une dimension intérieure, qui est aussi réelle, mais subjective. L'utilisation de la première personne du singulier indique donc un basculement de la réalité objective en faveur d'une réalité subjective. L'on remarque tout de suite que le pronom tu est absent. Par contre l'on retrouve *vous*, *il*, *ils* et surtout *elle*.(v. Fig. 2).

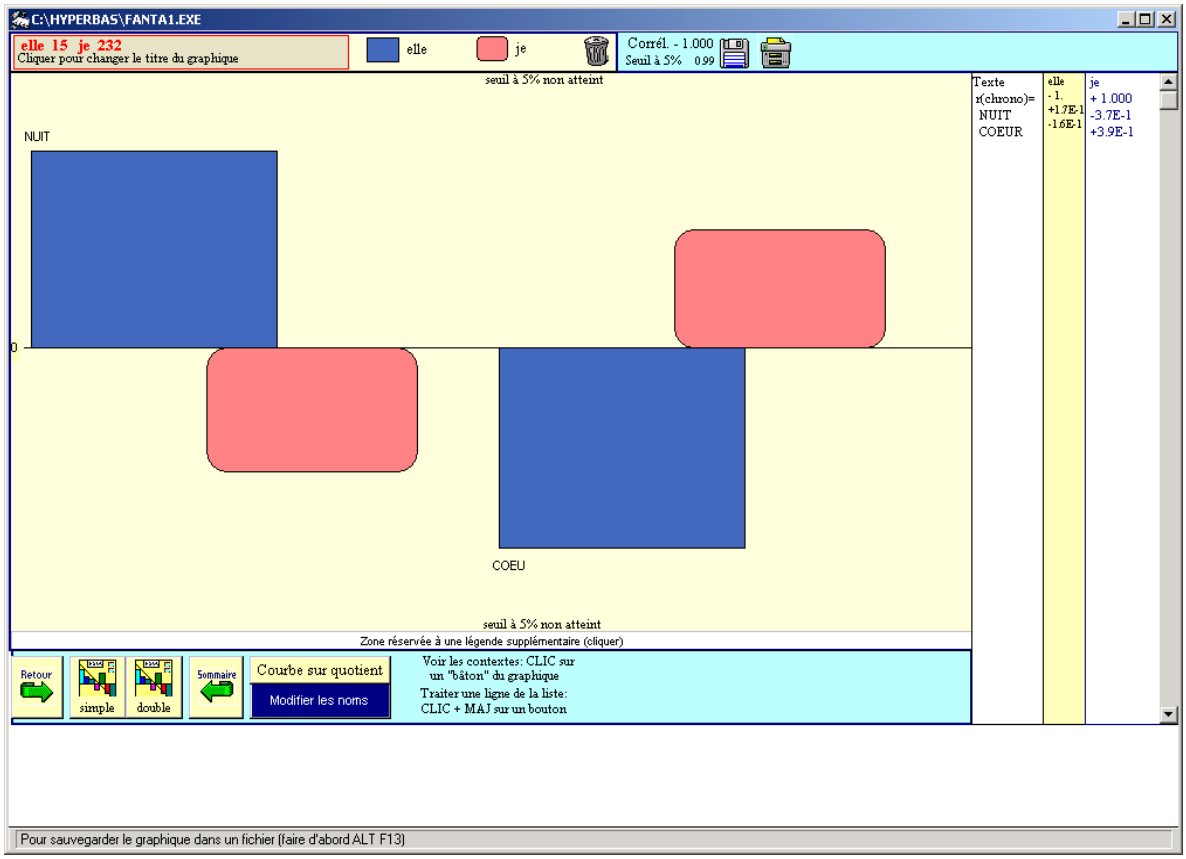

Fig. 2 : Graphique du rapport entre je-elle

Ce dernier pronom indique ce qui s'oppose au « je » qui, dans Maupassant, est la nuit et dans Poe, la maladie (considérée comme hyperacuité des sens). Si l'on regarde les concordances de *elle*, on voit que le pronom se trouve relié, par syntaxe ou sémantique, à des structures qui renvoient au doute : l'interrogation, l'inversion sujet-verbe, les points de suspension, l'utilisation de verbes tels que « paraître ». Ainsi, *elle,* la nuit ou la maladie dans ces deux cas, insinue le doute dans le narrateur, et le pousse à aller outre, à se pousser au-delà dans ces actions. (v. Fig. 3). Ce qui est intéressant de remarquer est, alors, le rôle actif que le sujet *je* occupe dans les récits. Le *je* est presque toujours suivi d'un verbe d'action (aller, aimer, s'habiller, regarder, marcher, sortir, entrer, crier, hurler… dans *la Nuit*, et encore, aimer, procéder, ouvrir, tourner, passer, avouer… dans *Le Cœur Révélateur*). Si la nuit ou bien la folie, concurrent dans le destin du protagoniste, c'est lui qui décide ses actions et il agit de façon consciente. Le rêve, le sommeil ne jouent aucun rôle - au moins qu'ils ne soient insérés a priori dans le texte sans que le lecteur en soit informé - par contre l'on retrouve par exemple des verbes tels *vouloir* ou *chercher à* qui indiquent une précise volonté d'action.

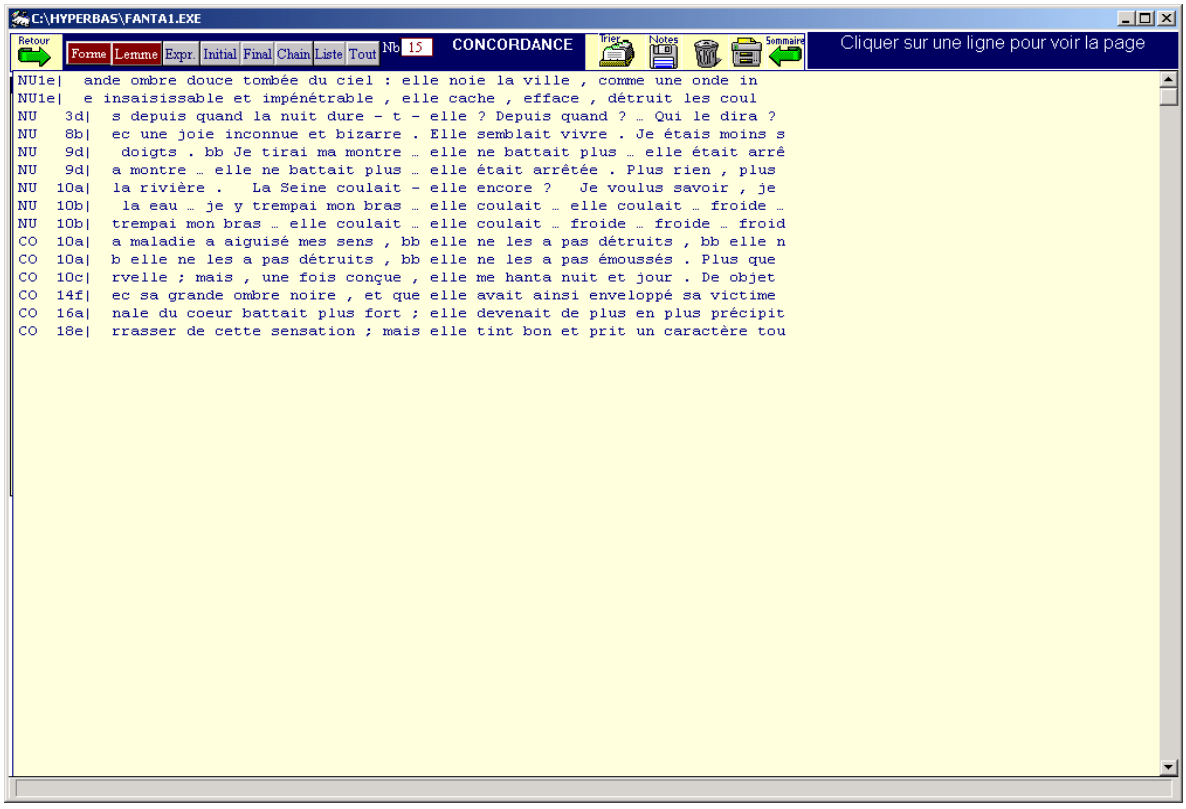

Fig. 3 : Concordance de elle

Encore à remarquer l'utilisation d'autres pronoms tels que *il* et *ils*. (v. Fig. 4), qui sont nettement prédominants dans *Cœur Révélateur*. *Il* est le vieux, donc la victime du protagoniste, mais aussi sa perdition, ainsi que *ils* représente les policiers. Le vieux, dont les seules actions qu'il fait sont dormir, se réveiller, regarder et crier, est un être passif. Pour ce que concerne *Ils*, les policiers, on assiste en plein à leur démarche cognitive : ils s'assirent, ils causèrent, ils restaient, ils entendaient, ils soupçonnaient, ils savaient. À noter à l'intérieur de cette démarche le changement du temps verbal, du passé simple à l'imparfait, ce qui transpose l'action d'une dimension du passé terminé à une dimension de passé continuel et donc toujours présent. Et encore, s'il est important de remarquer ce qui est fréquent dans le texte, il est néanmoins fondamental de mettre en relief ce qui n'est pas courant. Si l'on observe ce même graphique, ce qui nous saute aux yeux, c'est que dans la *Nuit*, *ils* n'apparaît que deux fois, dont une dans la phrase : *ils viendront*. Ce *ils*, jamais introduit auparavant, utilisé avec la forme verbale du futur, reste vague, et pour autant angoissant, mais peut se rapporter à la même fonction que *ils* (les policiers) ont dans la nouvelle de Poe, c'est-à-dire l'élément régulateur qui intervient pour mettre de l'ordre dans le cas d'une effraction des règles ordinaires.

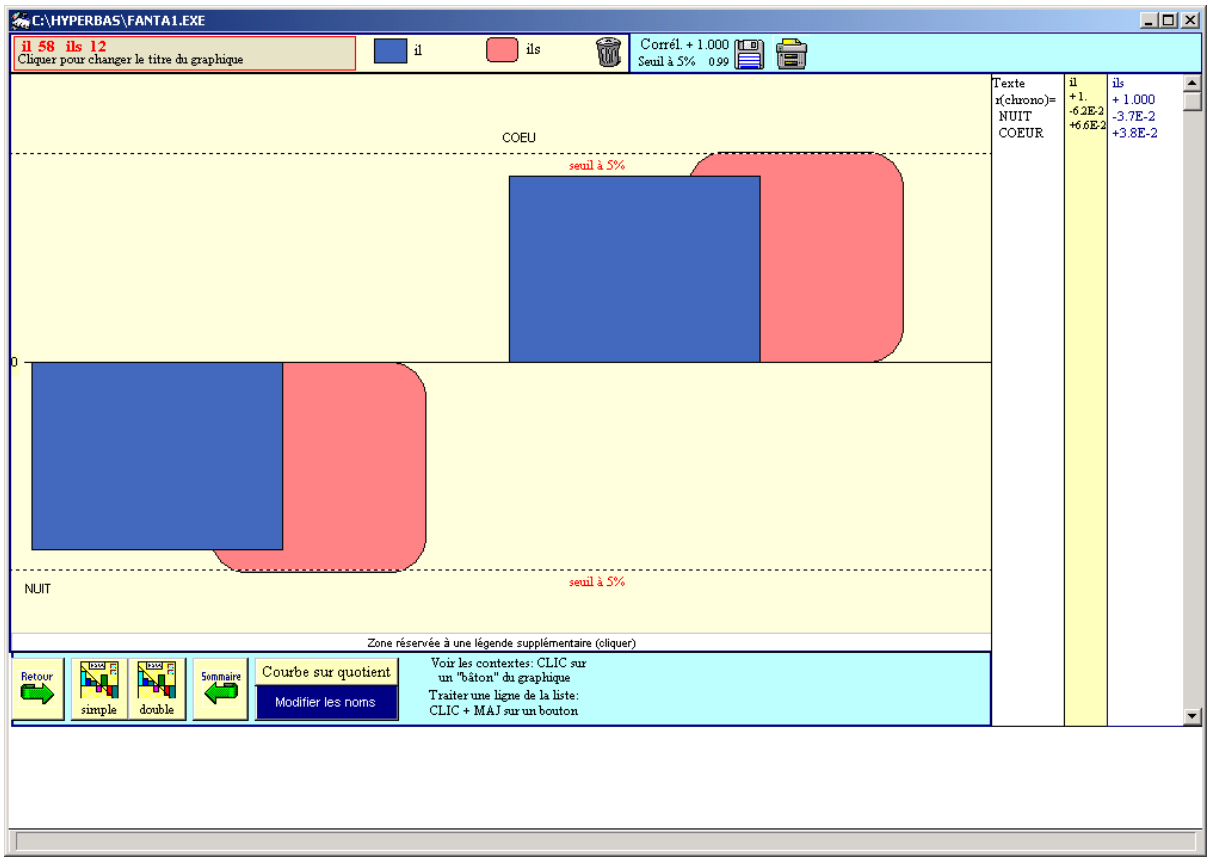

Fig. 4 : Rapport entre il-ils.

L'utilisation du futur nous renvoie à une autre composante essentielle des textes, le temps. Impossible de ne pas remarquer la présence de mots tels que : montre, tic-tac, tintement, horloge, heure, minute, hier, maintenant. Mais s'il n'y a aucune référence au futur (sauf *Ils viendront* qui exprime une attente, les autres signifient plutôt un doute : *Allons aux Halles, là au moins je trouverai de la vie*, *Qui le dira ?*, *Cette croyance s'évanouira*, *Qui le saura ?*), le passé, dans sont alternance entre passé simple et imparfait exprime un temps éternel, toujours passé mais toujours présent et futur aussi. Ainsi s'explique l'angoisse que les deux protagonistes éprouvent envers le temps et le fait qu'ils considèrent l'heure comme le seul point de repère qui leur reste.

L'analyse de l'opposition entre dimension spatiale et dimension temporelle nous a conduit à la conclusion que dans les deux cas il ne s'agit pas d'une succession logique des événements, mais plutôt d'un continuel retour, un cercle spatial et temporel qui envoûte les protagonistes et ne leur permet de s'en sortir qu'au biais d'un événement limite, la mort.

L'opposition dedans/dehors, est symbolisée d'un côté, *Le Cœur Révélateur*, par la chambre et les objets qu'elle contient, et de l'autre côté des noms des villes et des endroits de Paris. Mais cette opposition n'est pas non plus linéaire. En effet, le protagoniste de Maupassant se trouve dehors, mais il est tout seul, personne n'est là, sauf le chiffonnier qui d'ailleurs ne possède pas de montre, et donc a lui aussi perdu son rapport avec la dimension ordonnée de l'univers. La seule présence est donc la nuit. D'un autre côté, le protagoniste du *Coeur Révélateur* est chez lui, mais entouré de présences encombrantes, tels que les objets et tel que le vieillard, qui est en quelque sorte le correspondant masculin de la nuit. Les deux protagonistes partent d'une situation où ils aiment l'un le vieux et l'autre la nuit, mais au cours de l'histoire, tous deux subissent un changement : ce qu'était amour devient haine dans le premier et angoisse dans le second. Un élément destabilisateur que l'on peut remarquer à l'intérieur des récits est le bruit. Dans *Le Cœur Révélateur*, l'on parle « d'un bruit sourd et étouffé qui s'élève du fond d'une âme surchargée d'effroi », dans *La Nuit*, un bruit sourd conduit le protagoniste vers des rues noires et solitaires où il n'y a rien.

C'est le bruit donc l'élément perturbateur qui change la destinée des protagonistes et les emmène vers des contrées inexplorées.

La dernière opposition qu'on avait relevée concerne le rapport entre le corps et l'air. L'élément corporel acquiert du poids dans *Cœur Révélateur* : corps, œil, yeux, cœur, tête, tandis que dans la *Nuit* c'est l'élément aérien qui l'emporte : air, ciel, léger. Cette opposition montre comment le *Cœur Révélateur* nous introduit à l'intérieur d'un processus plus complexe par rapport à celui de la *Nuit*, ou mieux, il nous montre un passage antérieur que le récit de Maupassant ne montre pas. Ce qui pousse le protagoniste de Poe à tuer le vieux est la vision de cet œil de vautour maléfique qui l'effraie. L'œil est par définition l'une des parties corporelles le moins liée au corps, vu la matière qui le compose, et vu la fonction qu'il a dans le système corporel humain. Cet œil de vautour, énorme et grand ouvert, représente donc une menace du moment qu'il introduit dans celui qui le regarde la possibilité qu'il y ait dans le monde des êtres différents qui ne sont pas liés à leur matérialité. Ce que le protagoniste fait, sa tentative de tuer le vieux, représente donc une sorte de défense de l'homme envers l'inconnu, envers ce qui pourrait troubler son univers ordonné. Toutefois, cette tentative d'éliminer la peur n'atteint pas son but et même plus, elle déclenche un mécanisme de peur : le protagoniste a peur d'être découvert et donc confesse le tout, mais avant de tout avouer, la peur qu'il éprouve en cherchant de cacher son action est encore plus forte que la précédente. On ne peut donc pas éliminer la peur, il faut seulement l'accepter. Le protagoniste de la *Nuit*, se trouve dans le deuxième moment de ce processus. Il a accepté une de ses peurs, du moment qu'il avoue aimer la nuit, et il en tire des bénéfices, mais il n'a pas encore surmonté la peur de vivre, d'exister en tant qu'être humain avec ses limites, d'où son angoisse de rester seul, et d'où sa hantise par rapport au temps qui coule. Lui aussi, de même que le protagoniste de la nouvelle de Poe, tente une fausse voie : il se laisse engloutir par la nuit et se laisse mourir de faim et de froid, redevenant de cette manière entièrement humain.

Les deux épilogues ne concluent pas les récits. Dans la *Nuit* on ne sait pas si le protagoniste va mourir ou pas, dans *Cœur Révélateur* le récit se termine sur l'aveu sans raconter les conséquences. Si l'on prévoit une fin régulière, on pourrait supposer la mort du premier et la condamnation du second. Mais on peut envisager d'autres fins tout aussi plausibles. Les récits, à travers leur style qui rend l'angoisse et le suspense, à travers les sujets décrits, sortent de l'ordre reconnu des choses et décrivent des situations exemplaires, dans le sens de particulières. Ceci dit, l'on pourrait penser que le

protagoniste de la *Nuit* a été sauvé par ces « ils » qui devaient venir, et que l'assassin de Poe a été sauvé par le fait qu'il n'y avait plus de cadavre sous le plancher. Ces deux fins, deux simples exemples, font en sorte que la fin ne renvoie à rien d'autre qu'au début. En effet, rien n'a changé par rapport à la situation de départ, et les deux aventures pourraient se passer à nouveau, peut-être avec d'autres personnages. La situation de départ ne change pas dans le sens que la menace, l'origine de cette peur qui meut le récit n'a pas été déracinée, et peut ainsi continuer à hanter les protagonistes.

Voilà alors que cette brève analyse, qui n'a aucune volonté d'exhaustivité, vu la brièveté des textes et leur nombre réduit, peut nous aider à faire ressortir la structure (v. Fig. 5) des deux textes qui, même s'ils le font de façon différente, présentent deux protagonistes dans une même condition. Cette structure se base sur une lecture que l'on définit « responsive reading » et qui privilège quelques éléments à l'intérieur d'un texte, à partir desquelles, par plusieurs détours, il est possible de discerner une interprétation du texte même[23].

Parmi tous ces éléments, ceux que j'ai appelé « douteux » méritent une attention particulière, puisqu'ils relèvent les plus spécifiquement du fantastique. Il s'agit des tous ces indices qui laissent ouverte la porte de l'hésitation entre le caractère réel ou pas du récit, et qui insinuent le doute à l'intérieur d'une situation qui semble finalement normalisée. Ces éléments apparaissent le plus souvent vers la fin du récit et font en sorte de préserver la cohérence dans le point de vue d'une narration fantastique : le fantastique n'est présent que là où s'insinue le doute. Ainsi, de manière paradoxale, ces éléments destabilisateurs donnent cohérence et unité au récit.

[23] Cf. la distinction entre le différentes possibilités de lecture : Scanning, Search Reading, Skimming, Receptive Reading, Responsive Reading in TONFONI, G., *Intelligenza artificiale*, cit., pp.71-72.

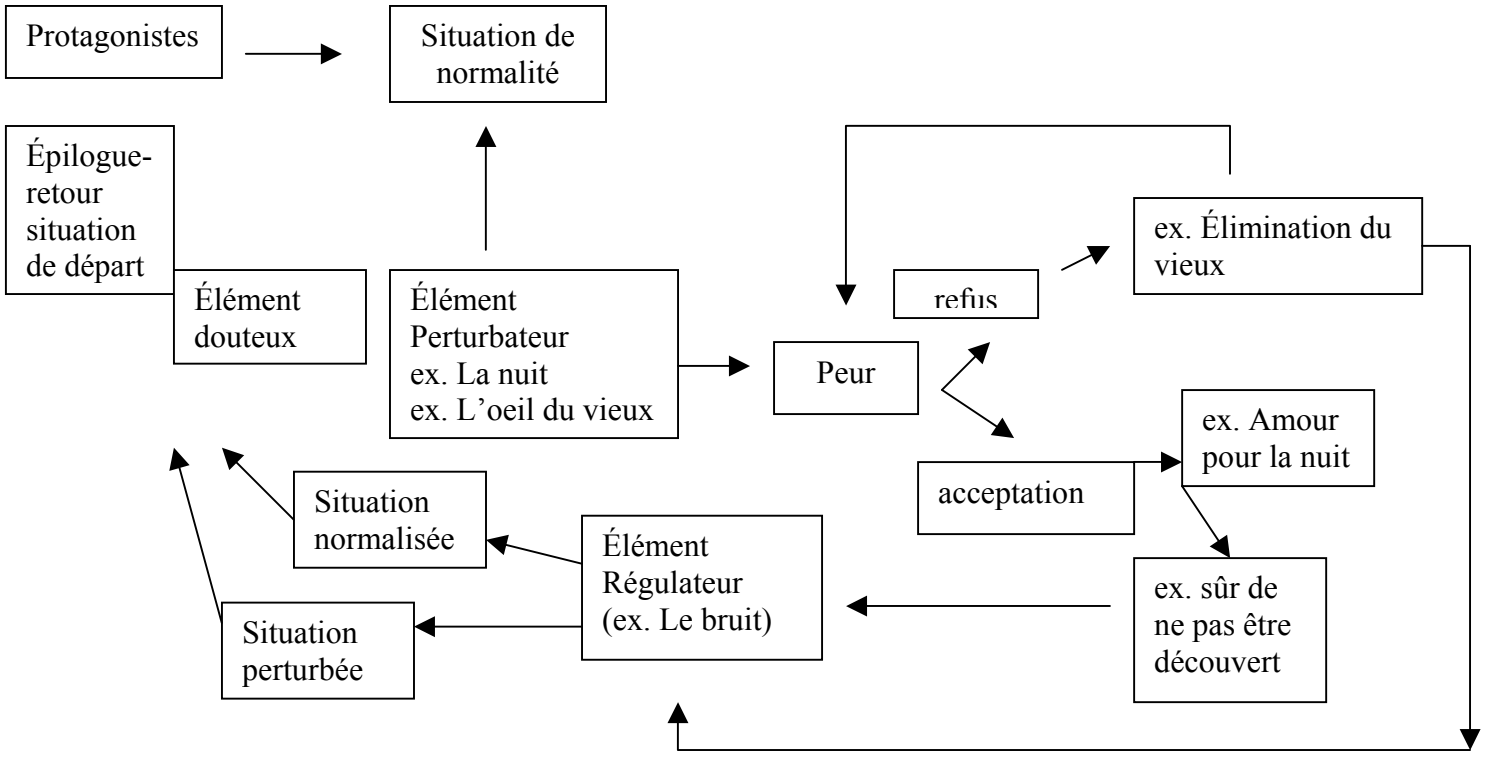

Fig.5 : esquisse d'une structure

3.2.2. Deuxième Corpus : Gautier et Gogol

Le deuxième corpus comprend deux textes assez longs (3812 occurrences pour la nouvelle de Gautier et 15918 mots pour Gogol). Cette différence détermine aussi naturellement que le texte de l'auteur russe est plus riche, et donc le décompte des distances lexicales ou des tranches de mots n'a pas de sens. Ce que je vais faire par rapport à ces deux textes, ce n'est pas, encore une fois, une analyse approfondie, mais plutôt une tentative de faire ressortir leur structure. L'instrument Hyperbase permet maintes possibilités que je n'ai pas exploitées dans mon analyse ; une étude sur si peu de textes n'a pas de raison d'exister, sauf pour faire ressortir non pas une étude linguistique approfondie, mais pour faire en sorte de se servir des données statistiques et documentaires du logiciel afin d'entrevoir une structure et une spécificité qui est propre à tous les textes d'une même typologie, en accord avec les études sémiotiques modernes, qui recherchent la « langue » structurelle justement à partir du discours.

Ce que l'on remarque immédiatement de ces deux textes par rapport aux autres est le fait qu'ils sont écrits à la troisième personne du singulier. Cet élément est important pour ce que concerne la littérature fantastique, car il implique un différent degré de participation du lecteur et différentes manières de modalisation du narrateur. Dans le cas de Gogol, cet aspect est mit en évidence par le biais d'une véritable introduction du « je auctorial » qui intervient le long du récit et met en scène un dialogue avec les lecteurs. Gautier aussi intervient, de manière bien plus subtile à l'intérieur du récit, quand il dit « je ne sais quelle odeur sulfureuse régnait dans la salle ». Ainsi, dans les deux récits, on entrevoit la présence d'un narrateur qui connaît les événements, y a participé et veut les raconter. Pour ce que concerne les protagonistes, la nouvelle de Gautier présente Heinrich, dont le nom propre apparaît 37 fois, tandis que chez Gogol, le protagoniste n'est ni Pigoriov ni Piskariov, mais « elle ». Qui est cette « elle »? La belle débauchée de Piskariov, l'Allemande de Pigoriov ou bien la Perspective Nevski ? Différencier l'utilisation de « elle » à l'intérieur du texte est possible, mais cela n'a pas de sens : « elle » représente l'antagoniste de « il » protagoniste et a la même fonction qu'avait le « elle » chez Poe et Gautier : c'est l'élément qui permet de rentrer dans le fantastique, de découvrir une autre dimension.

Dans le cas de *Deux acteurs pour un rôle*, l'antagoniste n'est pas une « elle », mais un « il » qui est identifié de différentes manières : monsieur, inconnu, homme… Ce personnage, fuyant dans le texte écrit comme dans la réalité, ne représente rien d'autre que le diable. Mais le diable n'intervient qu'au biais de la passion d'Heinrich pour le théâtre qui l'appelle de quelque sorte à intervenir à la première personne. Ainsi l'on peut dire, en observant ces deux textes, et les deux autres précédemment analysés, que ce « elle » ne renvoit pas forcément à un principe féminin – même si le nombre de récits fantastiques qui mettent en scène un héros masculin qui « lutte » contre un être féminin est surprenant - mais symbolise l'altérité, ce qui est différent, ce qui trouble, ce qui voudrait voler l'identité du « il » et le confondre. Ce « elle » est le biais et l'objet de l'intrusion du fantastique.

Dans les nouvelles du deuxième corpus, on trouve un élément qui était absent dans les textes considérés dans Fanta1 : la description, non seulement des personnages, mais aussi des lieux où l'action se passe. La nouvelle de Gautier met en scène des endroits spécifiques où se passe l'action : le jardin impérial, le gasthof et le théâtre.

L'occurrence de ces mots correspond plus ou moins à l'occurrence qu'a la perspective dans le texte de Gogol. (v. Fig. 6, un exemple). C'est dans ces endroits que se passe l'échange entre réalité et fantastique et où tout devient confus : la perspective le met en marche par le biais de sa vue et de sa lumière trompeuse, le gasthof par les vapeurs de l'alcool et le théâtre par la confusion entre rôle et identité. En effet, on pourrait considérer ces lieux comme des coadjuvants du fantastique, dans le sens qu'ils lui permettent de se montrer sans se dévoiler et de valider sa présence en se montrant à plusieurs personnes à la fois. Naturellement, pas tous le perçoivent, mais le sentiment d'étrangeté devient plus universel que celui éprouvé seulement dans l'esprit des deux protagonistes des nouvelles de Poe et de Maupassant.

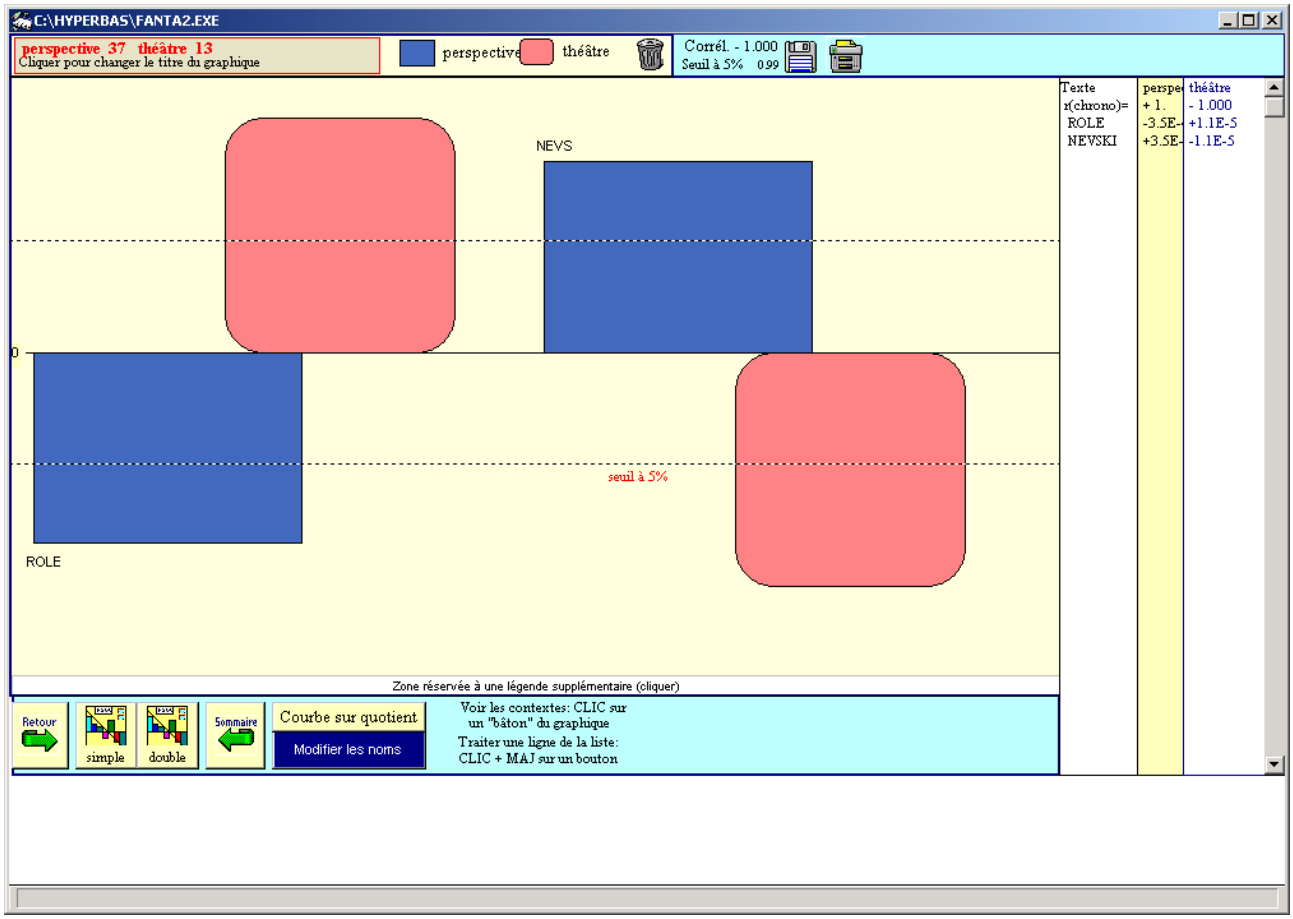

Fig. 6 : Rapport perspective- théâtre

Outre la description des lieux, ces deux nouvelles s'attardent sur la présentation des personnages, beaucoup plus nombreux qu'auparavant. Surtout pour ce que concerne Gogol, l'auteur insiste sur la classe sociale et les vêtements portés, non pas seulement pour une satire sociale – le sentiment de l'ironie est un procédé pour laisser transparaître

le fantastique - mais visant à rendre encore plus réelle la situation. De cette manière, au « je » permettant l'identification dans les deux premiers contes, se substitue une représentation attentive des lieux pour permettre une identification.

Ce que l'on remarque dès maintenant si l'on considère l'ensemble des quatre nouvelles, est qu'elles ont la même structure, sauf que les deux plus longues sont plus articulées. Les éléments perturbateurs se multiplient, des personnages adjuvants rentrent en scène (Kathy par exemple dans le conte de Gautier), l'action se déroule dans plusieurs lieux, les réactions des héros se multiplient. L'on remarque alors ces constantes dans les nouvelles définies fantastiques, c'est-à-dire un héros, une situation de normalité, un ou plusieurs éléments perturbateurs, l'entrée dans une autre dimension où la réalité se mélange avec le fantastique, un élément régulateur qui voudrait ramener la situation à la normalité, un épilogue qui reconduit à une situation similaire à celle du départ – dans le sens que si certaines situations personnelles changent, l'histoire universelle reste la même et le fantastique est toujours aux aguets - et qui laisse ouverte la question de l'explication. Ces éléments peuvent prendre des milliers de formes, se cacher sous plusieurs couches pour ne pas être perçus, utiliser différentes manières de représentation, différents styles, etc., ce qui rend difficile la codification du genre fantastique.

En poursuivant l'analyse des deux textes considérés, l'on pourrait découvrir les particularités du style des deux auteurs, mais aussi les différentes manières qu'ils utilisent pour représenter le fantastique et surtout son mélange avec la réalité. Ces modalisations peuvent être facilement repérées par les logiciels d'analyse. Par contre, faire ressortir une structure est une tâche qui dérive d'une analyse sélective des données et nécessite la démarche ultérieure de l'interprétation. Pour ce qui concerne ces deux nouvelles, la structure ressemble à celle mise en évidence pour les deux autres, mais il faut y ajouter :

Description du lieux et de l'environnement	Les adjuvants du fantastique : femmes, passion, peur…	Les adjuvants du réel : Kathy.

Richesse de style et de vocabulaire qui permet la création d'un rapport entre narrateur et lecteur, de même que la narration à la première personne.

Cette structure est donc sensible d'être agrandie ou réduite par rapport aux nombre d'événements qui occurrent dans le texte. La dénomination des éléments qui composent cette structure est très générale et n'atteint qu'un vague degré d'abstraction. Si ma structure diffère de la netteté des PU, plot units[24], cela est dû principalement au caractère des textes considérés. Un récit fantastique, on l'a souligné maintes fois, dérive de son sujet, mais aussi de sa forme. Un degré d'abstraction trop ample ne saurait suffire à expliquer tous les textes en question (p.ex. la renonce ou le sacrifice ne sont pas propres qu'aux textes fantastiques, même si l'on peut retrouver des PU plus riches de signification telles que la vengeance ou la peur), sans une attention portée sur comment procède le récit. Voilà alors que des éléments tels que les descriptions ou les modalisations acquièrent une grande importance dans le processus de la mise en scène du fantastique, ce que démontre que l'étude d'un texte fantastique ne peut se fonder que sur une interprétation thématique, mais nécessite d'un support pour l'étude linguistique[25].

Conclusion

L'analyse de *Deux acteurs pour un rôle* et de *La Perspective Nevski* pourrait se poursuivre et devenir bien plus détaillée que celle que j'ai faite, faisant ressortir maints éléments que je n'ai pas considérés (comme par exemple le rôle de l'ironie, ou les renvois aux sources littéraires…), la même chose vaut pour les nouvelles de Poe et de Maupassant. Mais le but de cette brève analyse était d'un côté l'étude du logiciel et sa manière de faire face à un texte littéraire, et de l'autre la vérification du fait si ces études pouvaient être, outre que des instruments de recherche, des outils didactiques. On a vu que le but de l'informatique humaniste était de mettre en contact le domaine littéraire et le domaine informatique : l'ordinateur fournit des données qui servent d'inputs afin que l'utilisateur puisse les retravailler et les interpréter en faisant ressortir la structure du texte analysé. Le décompte des mots n'a pas une fin à elle-même, mais

[24] Cf. LEHNERT, W.G., *Plot Unit and Narrative Summarization*, in «Cognitive Science », n.4, pp. 293-331.
[25] À ce propos, une méthode combinatoire entre logiciels d'analyse linguistique et logiciels tels que le TUP (Text Understanding Procedures) crée par Tonfoni et Doyle pourrait être une solution pour une approche plus correcte aux textes fantastiques, cf. TONFONI, G., *Percezione concettuale del testo e processi direzionati di riassunto*, in ID, *Sistemi cognitivi complessi*, cit., pp. 87-102.

tend à englober ces éléments dans un plus grand système qui permet un classement des composants du texte. La démarche cognitive de l'utilisateur est semblable à celle de l'ordinateur : il subit des inputs, les décompte et les interprète. Voilà alors que le processus d'automatisation de l'homme et d'humanisation de la machine, en tant que fournisseur d'information, peut avoir lieu.

D'un point de vue didactique, ces instruments permettent aux utilisateurs de rentrer véritablement dans les textes, aussi bien dans la phase de codification qui permet de remanier les textes que dans le repérage des mots. Toutes les informations que l'on perçoit pendant la lecture, se numérisent à l'aide des logiciels et deviennent des données objectives. Et encore, le fait d'aller outre à une simple lecture et d'interpréter les données pour en faire ressortir la structure, consent d'acquérir un degré d'abstraction ultérieur tendant à la construction d'un système. De cette manière, la définition a priori de quatre nouvelles fantastiques (titre du recueil : *Nouvelles Fantastiques*) peut être vérifiée à l'aide d'instruments statistiques et documentaires. Cette tâche appliquée à des nouvelles fantastiques se révèle bien plus complexe qu'on ne l'a présentée. Le fantastique est un genre qui a rencontré, et rencontre encore, des difficultés de codifications, dues sûrement à la multitudes de formes et de variantes qu'il comporte. Tous les éléments présents à l'intérieur d'un texte fantastique peuvent changer, mais surtout ce qui varie continuellement est le moment où le fantastique s'insinue dans la réalité. Est-ce que ce moment est antérieur au récit ? Est-ce qu'il est décrit ? Est-ce qu'il n'est rien que suggéré mais n'apparaît pas ?

La création d'un corpus de contes fantastiques – les nouveaux logiciels permettent aussi une approche multilangue – pourrait aider dans la codification de ce genre tout en respectant les spécificités de chaque texte et de chaque auteur.

Bibliographie

- AA.VV., *L'analisi del racconto, le strutture della narratività nella prospettiva semiologia che riprende le classiche ricerche di Propp*, Milano, Bompiani, 1969.
- *Analyse d'une base de données documentaires avec Lexica*, http://www.lessphinx-developpement.fr
- BAKER, Mona, FRANCIS, Gill, TOGNINI-BONELLI, Elena (eds.), *Text and Technology, In Honour of John Sinclair*, Philadelphia-Amsterdam, John Benjamins Publishing Company, 1993.
- BANDRY, Anne, *Les mots de Haywood*, in SIAT, Seminario d'informatica applicata all'analisi testuale, Cesenatico 2-5 Ottobre 2003, à paraître.
- BARTHES, Roland, *Éléments de sémiologie*, Paris, Éditions du Seuil, 1964.
- BECCARIA, Gian Luigi., *Dizionario di linguistica e di filologia, metrica e retorica*, Torino, Einaudi, 1996.
- BECUE BERTAUD, Monica, *Apport des méthodes lexicométriques à l'étude d'un texte : Évolution du vocabulaire, coupures thématiques et stratégies discursives*, Dpt. Estádistica e Inv. Operativa, Universitat Politècnica de Catalunya.
- BERTRAND-GASTALDY, Suzanne, PAGOLA, Gracia, *l'Analyse du contenu textuel en vue de la construction de thésaurus et de l'indexation assistées par ordinateur ; applications possibles avec SATO (Système d'Analyse de Textes par Ordinateur)*, http://www.ling.uqam.ca/publications/bibliographie/Db92.htm
- BOLASCO, Sergio, *Meta-Data and Strategies of Textual Data Analysis : Problems and Instruments,* in *Data Science, Classification and Related Methods*, V International Conference of IFCS – Kobe, 27-30 march 1996, Japan, Springer-Verlag, Tokio, 1997.
- BONFANTINI, Massimo A., *Breve corso di semiotica*, Napoli, Edizioni scientifiche Italiane, 2000.
- BRUNET, Étienne, *Hyperbase, Logiciel hypertexte pour le traitement documentaires et statistiques des corpus textuels, Manuel de Référence*,version 5.4. pour Windows (janvier 2002)

- CABALLERO RODRIGUEZ, Maria Rosario, *Using a Concordancer in Literary Studies*, http://218.103.45.154/Concordance/Review/programa.htm
- CHAUMIER, Jacques, DEJAN Martine, *Recherche et analyse de l'information textuelle, Tendances des outils linguistiques*, « Documentariste. Science de l'information », vol. 40, n°1, 2003.
- COUCHOT, Edmond, *La critique face à l'art numérique : une introduction à la question*, « Solaris », Décembre 2000, Janvier 2001, http://bibliofr.info.unicaen.fr/bnum/jelc/Solaris/d07/7couchot.html
- DAOUST, François, DUPUIS Fernande, *Le dispositif linguistique (SATo-CALIBRAGE)*, http://www.ling.uqam.ca/publications/bibliographie/C3dislin.htm
- DAOUST, François, DUPUIS, Fernande, WALLACE Éfoé, *Notes explicatives sur les descriptions lexicales de la BLD*, http://www.ling.uqam.ca/outils/bdl.htm
- DUCHASTEL, Jules,
    - Et ARMONY, Victor, *Un protocole de description de discours politiques*, http://www.ling.uqam.ca/sato/publications/bibliographie/Jul10.htm
    - Et DUPUY, Luc, PAQUIN, Louis-Charles, BEAUCHEMIN Jacques, DAOUST François, *Système d'analyse ce contenu assistée par ordinateur (SACAO)*, http://www.ling.uqam.ca/sato/publications/bibliographie/Jul15.htm
- ECO, Umberto, *Lector in fabula, la cooperazione interpretativa nei testi narrativi*, Milano, Bompiani, 1979.
- GÉLINAS-CHÉBAT, Claire, PRÉFONTAINE, Clémence, DAOUST François, *Exemple d'analyse de documents d'information,* http://www.ling.uqam.ca/sato/publications/bibliographie/C3exem.htm
- GIGLIOZZI, Giuseppe, *Il testo e il computer, manuale di informatica per gli studi letterari*, Milano, Edizioni Scolastiche Bruno Mondatori, 1997.
- HOEY, Michael, *From concordance to text structure . new uses for computer corpora*, in LEWANDOWSKA-TOMASZCZYK, B. / MELIA, P.J. (eds.), *PALC'97 Practical Applications in Language Corpora*, Łódź, Łódź University Press, 2-23.
- *How to us ConcApp*, http://218.103.45.153/pub/concapp/Help/tutorial1.HTM

- http://talana.linguist.jussieu.fr
  http://www.cavi.univ-paris3.fr/lexicometrica/jadt/kadt1998/ouazzani.htm
- JOHNSSON, Stig, OKSEFJELL, Signe, *Corpora and Cross-linguistic Research, Theory, Method and Case Studies*, Amsterdam-Atlanta, GA, 1998.
- KENNY, A., *The Computation of style, An Introduction to Statistics for Student of Literature and Humanities*, New York, Pergamon Press, 1982.
- LAHLOU, Saadi, *Vers une théorie de l'interprétation en analyse statistique des données textuelles*, JADT 1995, *3rd international Conference on Statistical Analysis of Textual Data*, S. Bolasco, L. Lebart, A. Salem (eds.), CISU, Roma, 1995, vol. I, pp. 221-228.
- LELU, Alain, HALLEB, Mohammed, DELPRAT Bruno, *Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-gammes*, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt1998/lelu.htm
- MORDENTI, Raoul, *Informatica e critica dei testi*, Roma, Bulzoni, 2001.
- NOBILI, Claudia Sebastiana, *Il lavoro della scrittura, analisi e retorica del testo*, Milano, R.C.S. Libri S.p.A., 1999.
- NUMERICO, Teresa, VESPIGNANI, Arturo (a cura di), *Informatica per le scienze umanistiche*, Bologna, il Mulino, 2003.
- OLIVER, Andrew, *Retour au Père Goriot, ou ce que nous apprend la statistique*, Département d'Études françaises, Université de Toronto, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/kadt1998/oliver.htm
- ORLANDI, Tito, *Informatica Umanistica*, Roma, La Nuova Scientifica, 1990.
- OUAZZANI, Ilham, *Analyse Statistique des textes littéraires, l'exemple de Driss Chraïbi*, Université de Nice,
- REINERT, Max, *Quel* objet *pour une analyse statistique du discours ? Quelques réflexions à propos de la réponse* Alceste, Sommaire de JADT 1998, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt1998/reinert.htm
- RICCIARDI, Mario (a cura di), *Lingua, letteratura, computer*, Torino, Bollati-Boringhieri, 1996.
- ROSSINI FAVRETTI, Rema (a cura di), *Linguistica e informatica, Corpora, multimedialità e percorsi di apprendimento*,Milano, Bulzoni Editore, 2000.

- SCARLINI, Luca, *La musa inquietante, il computer e l'immaginario contemporaneo*, Milano, Raffaello Cortina Editore, 2001.
- SINCLAIR, John
    - *Corpus, Concordance, Collocation*, Oxford, Oxford University Press, 1991.
    - *Current Issues in Corpus Linguistics*, in ROSSINI FAVRETTI Rema (a cura di), *Linguistica e informatica, Corpora, multimedialità e percorsi di apprendimento*,Milano, Bulzoni Editore, 2000, pp. 29-38.
- SPINA, Stefania, *Fare i conti con le parole, Introduzione alla linguistica dei corpora*, Perugia, Guerra Edizioni, 2001.
- STUBBS, Michael, *Text and Corpus Analysis, Computer-assisted Studies of Language and Culture*, London, Blackwell Publishers, 1998.
- *Thief et Frantext* : http://ancilla.unice.fr/~brunet/pub/THIEF/THIEF2.htm
- TONFONI, Graziella
    - *Didattica del testo, curricolo di formazione linguistica per gli insegnanti*, Teramo, Giunti&Lisciani Editori, 1991.
    - *Intelligenza artificiale : automazione dei processi cognitivi e sistemi per la rappresentazione della conoscenza*, «Lectures», monografie interdisciplinari dirette da R. Campagnoli, V. Carofiglio, Y. Hersant, A. Ponzio, n. 19, *Automi*, 1986/2, Bari, Edizioni del Sud, pp. 77-95.
    - *Paradoxes and censors*, « Semiotica », vol. 60-1/2, Amsterdam, Mouton de Gruyter, 1986, pp. 247-257.
    - *Sistemi cognitivi complessi, intelligenza artificiale e modelli di organizzazione della conoscenza*, Treviso, Pagvs Edizioni, 1991.
- *Tropes : l'analyse du discours haut de gamme à la portée de tous*, http://www.acetic.fr/tropes.htm