L'analyse thématique par ordinateur : quelques résultats

par

Paul. A. FORTIER
Université du Manitoba - CANADA

Monsieur Colin McConnell a développé pour moi un système d'analyse du contenu thématique des textes littéraires. J'ai déjà décrit ailleurs le système et sa justification théorique; donc il suffit d'en rappeler ici seulement les grandes lignes (1).

L'idée de base c'est une notion introduite dans les milieux littéraires de langue française par Charles Baudelaire (2) lorsqu'il traduisait "The Philosophy of Composition" d'Edgar Poë (3). Cette notion est bien simple : chaque élément, voire chaque mot dans l'oeuvre littéraire doit viser la production de l'effet esthétique de l'oeuvre, rien ne doit être étranger à ce but, rien ne doit être laissé au hasard. Cette notion a été reprise par Paul Valéry (4) et plus récemment par des critiques structuralistes tels que Todorov (5) et Ricardou (6).

On voit facilement comment cette idée s'applique à la création de l'action et des personnages. Mais il y a un problème lorsqu'on veut intégrer les thèmes et l'atmosphère qui est souvent créée par les structures thématiques. Car alors, on semble obligé de comparer ses impressions sur les thèmes aux faits vérifiables de l'action et des personnages. C'est donc pour fournir des renseignements précis et objectifs sur les thèmes, afin de pouvoir les étudier sur un pied d'égalité avec l'action et les personnages que mon système a été développé.

La précision exige qu'on travaille sur des textes lemmatisés (voir Figure 1). Donc on enregistre un texte et on en fait une concordance. Puis, on ajoute la forme de base et la partie du discours à chaque mot-clef de la concordance, afin de réaliser une concordance lemmatisée, et à partir d'elle on construit un fichier-texte contenant la forme de base et la partie du discours de chaque mot dans le texte, suivie d'un numéro indiquant l'endroit où se trouve chaque occurrence du mot.

Une autre partie du système sert à rassembler tous les mots cités comme synonymes d'un mot-thème donné dans onze dictionnaires de synonymes et à écarter les mots n'ayant qu'une relation faible au mot-thème. Les entrées dans ce fichier-thèmes enregistrent également la forme de base et la partie du discours de chaque mot conservé. (Je passe sous silence les multiples étapes de vérification--soit manuelles, soit automatiques-- dans le système qui vise à supprimer les erreurs dans les deux fichiers).

Tout ce travail prépare des données pour le programme final du système, qui fournit des renseignements sur la fréquence et la distribution de tel ou tel thème dans un texte donné. Le système a été développé entre 1971 et 1973. Sa structure reflète mes méthodes lorsque je travaillais à la main sur une concordance non-lemmatisée de *Voyage au bout de la nuit* de Céline. Depuis 1973, j'ai pu utiliser le système pour l'analyse de deux autres romans, *La Jalousie* de Robbe-Grillet et *l'Immoraliste* de Gide.

Voyage au bout de la nuit est un roman de type picaresque à échos philosophiques. Bardamu, le protagoniste s'engage dans l'armée juste avant le début de la première Guerre mondiale, éprouve les conditions guerrières au front et à l'arrière, puis trouve un poste de factorier dans la brousse africaine

où il attrape la fièvre et est vendu au patron d'une galère qui le mène aux Etats-Unis. Rentré en France, Bardamu devient médecin et pratique dans des quartiers pauvres de Paris avant de terminer dans le personnel d'un asile d'aliénés.

Travaillant à la main, j'ai extrait d'une concordance du roman des tables montrant la fréquence et la distribution d'une bonne centaine de thèmes dans ce texte. Quarante de ces thèmes étaient utiles pour l'analyse. Certains autres n'avaient pas assez d'évocations pour être considérés comme importants. D'autres encore ne fournissaient pas de renseignements intéressants; par exemple, le thème de la mort, quoique évoqué presque cinq cents fois dans le roman, se concentre dans la section du texte où Bardamu, qui fait la guerre, est obsédé par le peur de mourir, et dans des chapitres décrivant la mort de certains personnages secondaires. C'est-à-dire que la structuration des thèmes ne révèle que de l'information déjà évidente après un examen de l'action du roman. Cela n'est pas toujours le cas.

Le thème de la violence (voir Figure 2) domine --c'est-à-dire qu'il est évoqué le plus fréquemment-- non pas dans la partie du roman qui décrit la guerre, mais dans les chapitres qui relatent les aventures de Bardamu en Afrique. Cela est assez étonnant car la guerre implique nécessairement la violence tandis que dans la section africaine du roman, ce sont les forces biologiques qui dominent. La violence ne se rattache pas autant à un phénomène historique et limité --la guerre-- qu'aux forces biologiques qu'on ne saurait refuser sans nier notre condition humaine. Cette structuration du thème de la violence explique en partie la forte impression de pessimisme qu'on ressent en lisant *Voyage au bout de la nuit*.

J'ai pu découvrir des groupements de thèmes qui aident également à mieux comprendre le texte. Par exemple, à la guerre, Bardamu décrit d'un ton léger, la décapitation d'un messager et l'éventrement d'un colonel dans une explosion d'obus (2--21) (7). Mais il est dégoûté jusqu'au vomissement en voyant la boucherie improvisée de son régiment (24). J'ai pu déterminer que les thèmes qui dominent l'évocation de la boucherie caractérisent également les descriptions des conditions au front. Donc le dégoût de Bardamu est déclenché par une image de la situation guerrière en général plutôt que par un seul incident.

De même, Bardamu s'affole devant le *Stand des Nations*, une baraque de tir dans une fête foraine (60-61); cela aussi s'explique au niveau thématique. La description de cette baraque résume les thèmes qui caractérisent la guerre dans ce roman --aussi bien au front qu'à l'arrière. Une analyse détaillée des thèmes montre la justification profonde d'une peur que les autres personnages du roman n'interprètent que comme couardise ou bien folie.

Vers la fin du roman, Bardamu embauche une belle infirmière, Sophie. Mais en la décrivant, il montre également un besoin de l'avilir (462-64). Michel Beaujour interprète cette ambivalence comme preuve de "la fondamentale perversité de Céline" (8). Mais si on regarde de près les thèmes, on peut voir que Sophie incarne une vision de la beauté féminine qui avait attiré Bardamu en Amérique --avec des

résultats néfastes pour lui. Une centaine de pages avant la description de Sophie, Bardamu avait déjà constaté les insuffisances de cette vision de la beauté féminine (355-56), montrant qu'elle est reliée aux forces biologiques dont il avait vu les effets néfastes en Afrique et lorsqu'il était médecin à Rancy. Grâce aux renseignements thématiques, j'ai donc pu expliquer le comportement de Bardamu en fonction des structures du roman, sans la nécessité d'avoir recours à d'éventuels complexes de l'auteur.

Même plus, j'ai découvert que chaque section du roman débute par une description qui évoque puissamment le contenu thématique qui sera élaboré par la suite dans le reste de la section (9). Ce procédé ajoute une impression d'inévitabilité au pessimisme que crée le roman et met dans ce texte de 1932 une structure réactionnaire qui ne sera pas exprimée clairement, en ce qui concerne la politique avant 1936.

Je crois que cela suffit pour montrer que l'étude précise des thèmes m'a fourni la matière d'une belle interprétation du roman. Donc avec un système flambant neuf pour faire le gros du travail de recensement, et de grands espoirs, je me suis tourné à l'analyse de *la Jalousie* de Robbe-Grillet, et je me suis cassé le nez.

La Jalousie reflète ce qui se passe dans la tête d'un narrateur invisible pendant qu'il fouille ses souvenirs des circonstances entourant un voyage en ville que sa femme A... a fait avec Franck un voisin un peu trop amical. A cause d'une panne de voiture, les deux ont dû passer la nuit en ville, laissant le narrateur seul à sa plantation. Des incidents précédant et suivant ce voyage sont rappelés plusieurs fois par le narrateur, mais chaque fois, ils sont un peu différents. D'ailleurs, il est impossible d'enchaîner ces incidents pour former un ordre chronologique, car le narrateur les présente de manière contradictoire--tel incident survenant clairement après tel autre puis, à un autre rappel, leur ordre est renversé.

La possibilité de faire des analyses par ordinateur m'a permis de montrer facilement l'importance -- du point de vue de la fréquence-- de certains thèmes comme le corps humain ou l'incertitude. Mais des tables de fréquence et de distribution n'ont pas été très utiles, car ils montraient pour ce roman seulement les lieux de concentration des évocations d'un thème sans indiquer en même temps un certain aspect de l'action ou de la création des personnages. Donc je n'avais rien auquel je puisse comparer la structuration thématique ainsi démontrée.

Autrement dit, le thème du voyage est important là où le narrateur pense à un aspect ou à un autre du voyage en ville qui est l'événement principal du roman, et c'est tout. L'analyse par ordinateur de la distribution des thèmes m'a fourni une preuve du fait que ce texte est structuré thématiquement, mais pas grand'chose d'autre. Heureusement, je possédais une concordance lemmatisée du roman, produite comme étape intermédiaire par le système pour l'analyse thématique. Grâce à elle, je pouvais suivre pas à pas l'élaboration de chaque thème important, identifier les endroits où il dominait, et voir sa relation avec d'autres thèmes. J'ai pu faire une interprétation du roman montrant notamment

comment un très grand nombre des aspects du texte produisent une sorte d'incertitude agaçante, qui correspond en partie à l'émotion de la jalousie (10). Mais j'ai dû constater aussi que pour un texte comme le roman de Robbe-Grillet, mon système d'analyse thématique me fournissait peu de chose qu'une concordance ne m'aurait pas fourni. Puisque le travail nécessaire pour préparer un texte et des fichiers-thème n'avait pas un rendement suffisant, il fallait reconnaître un échec.

Les résultats étaient meilleurs avec un autre texte, *l'Immoraliste* d'André Gide. Ce roman raconte l'histoire d'un homme, Michel, qui tombe malade de la tuberculose pendant son voyage de noces en Afrique du Nord. Grâce aux soins dévoués de sa femme, Marceline, il recouvre sa santé et, guéri, il apprécie davantage la vie. Les deux époux retournent en France, passant par l'Italie, mais peu après Marceline à son tour, tombe malade, et Michel l'emmène en Suisse pour la guérir. Mais il n'y tient pas en place, et les deux époux repartent pour l'Italie, puis retournent en Afrique pour finir à Biskra où Michel avait guéri. Et c'est là où, exténuée des déplacements incessants, Marceline meurt.

Il semblerait qu'il s'agisse tout simplement d'un cas d'égoïsme quasi pathologique. Mais un examen de la distribution des pronoms à la première personne du singulier (voir Figure 3) montre qu'il faut nuancer cette impression première. Puisque le protagoniste est aussi le narrateur, ces pronoms sont assez fréquents. Ce qui est remarquable, c'est la baisse, irrégulière, certes, mais substantielle de l'utilisation de ces pronoms, à mesure que le roman se déroule. Le langage de ce texte indique que le narrateur parle progressivement moins de lui-même --qu'il se révèle de moins en moins égoïste-- comme il raconte son histoire.

Un groupe de thèmes centré sur la maladie suggère le point de départ pour une autre interprétation (voir Figure 4). Les évocations de ce groupe de thèmes sont les plus fréquentes vers le commencement du texte lorsque le narrateur est gravement malade, puis elles diminuent rapidement comme il guérit. Leur fréquence s'élève de nouveau --c'est le cas surtout dans les descriptions de la guérison, de l'Italie, de la Normandie et de Paris-- parce que le narrateur a une forte tendance à associer certains thèmes à la notion de maladie. C'est le cas des thèmes de l'ennui, de la laideur, de la faiblesse, de la tristesse et de l'ordre. Quand ces thèmes prennent de l'importance, la maladie est toujours évoquée également, et la fréquence du groupe maladie va croissant.

Puis, dans la deuxième section parisienne, Marceline souffre d'une fausse couche et de complications qui la rendent gravement malade. Elle va mieux lorsque les deux époux sont en Normandie, puis son état s'empire pendant leurs voyages en Italie et en Afrique jusqu'à ce qu'elle meure. Les évocations de la maladie croissent de nouveau dans les deux dernières sections du roman, qui décrivent ces voyages, mais pas d'une manière régulière comme on s'attendrait. Cette anomalie s'explique en fonction d'un autre groupement thématique.

Autour de la notion de santé s'assemble un groupe thématique comprenant la beauté, le jour, la vie, la chaleur, l'eau, la force et la lumière (voir Figure 5). Les évocations de la santé sont les plus

fréquentes au moment où Michel guérit de sa maladie. Elles reprennent de l'importance lorsqu'il évoque la beauté ensoleillée de sa ferme en Normandie. De même, pendant que Marceline est très malade, Michel sort un beau soir d'hiver à Paris, et la description de cette sortie est reflétée dans la fréquence des thèmes associés à la santé. Les évocations de ce groupe croissent comme Marceline va mieux en Normandie, puis diminuent avec l'arrivée de la mauvaise saison et une rechute de Marceline. Ensuite, pendant les voyages en Italie, comme Marceline devient de plus en plus malade, les évocations de la santé deviennent très fréquentes. C'est que Michel la mène dans les endroits qui pour lui étaient ceux de la santé, qui s'associent à la santé dans son esprit. Cette association est clairement indiquée par une observation assez étourdie qu'il fait : "Pourquoi tousse-t-elle par ce beau temps ?" (465) (11).

Puis, comme l'agonie de Marceline arrive, le groupement thématique qui associait à la santé une certaine sorte de décor, se dissout sous la pression des événements. A la fin du roman, Michel reste sans femme, abasourdi, ne comprenant plus du tout comment il a pu agir comme il a fait, mais ne voyant pas où il a mal agi. Et le lecteur est aussi mystifié que Michel à moins qu'il n'ait prêté attention à la structuration thématique du texte. Je crois que cette interprétation montre assez clairement l'utilité de mon système.

Et que dire en guise de conclusion ? Que mon système fonctionne utilement deux fois sur trois, c'est-à-dire qu'il vaut une note de treize sur vingt, pas fameux ? Eh bien, je crois que non. Il s'agit des points de départ. La base théorique du système c'est la notion de coordination dans l'oeuvre littéraire entre trois éléments : l'action, les personnages, les thèmes. Quand un de ces trois éléments est absent ou très faible --comme l'action dans le nouveau roman et la plupart des poèmes, comme les thèmes dans le roman d'aventure, le roman policier et un grand nombre de pièces-- alors mon système ne peut fournir des résultats utiles. C'est-à-dire qu'il est un outil de précision avec un rayon d'application limité. Cela ne me gêne pas du tout; c'est plutôt le contraire.

REMERCIEMENTS

Ce travail a bénéficié des subventions de recherche suivantes accordées par le Conseil des Arts du Canada : 69-418, S70-0735, S70-1561, S71-1933, S72-1650, W74-0453, S76-0734, 451-77-520. L'Université du Manitoba a fourni une bourse postdoctorale et plusieurs allocations de recherche.

NOTES

(1) Voir Paul A. Fortier et J. Colin McConnell, "Computer-Aided Thematic Analysis of French Prose Fiction", *The Computer and Literary Studies*, éds. A.J. Aitken *et al.* (Edinburgh , Edinburgh University Press, 1973), pp. 167-81; Paul A. Fortier et J. Colin McConnell, "Computer-Aided Analysis of French Prose Fiction : II. Analysis of Texts and Preparation Costs", *The Computer in Literary and Linguistic Studies*, éds. Alan Jones et R.F. Churchhouse (Cardiff , University of Wales Press, 1976), pp. 215-222; Paul A. Fortier et J. Colin McConnell, *THEME : A system for Computer-Aided Theme Searches of French Texts* (Winnipeg , University of Manitoba, 1975).

(2) "Notes nouvelles sur E.A. Poe", *Nouvelles Histoires extraordinaires*, éd. J. Crépet (Paris , Conard, 1933), pp. v-xxiii.

(3) *The Poems and Three Essays on Poetry*, éd. R.B. Johnson (Oxford , Oxford University Press, 1927), pp. 189-202.

(4) *Oeuvres I*, éd. J. Hytier (Bibliothèque de la Pléiade; Paris , Gallimard, 1957), pp. 769-84, 1313-39, 1359-78, 1412-14, 1438-43, 1456-91, 1786-88; *Oeuvres II*, éd. J. Hytier (Bibliothèque de la Pléiade; Paris, Gallimard, 1960), pp. 1207-09.

(5) *Introduction à la littérature fantastique* (Paris , Seuil, 1970), pp. 97-106.

(6) *Pour une théorie du nouveau roman* (Paris, Seuil, 1971), pp. 59-88.

(7) Les numéros entre parenthèses renvoient au texte du roman dans l'Edition de la Pléiade, éd. H. Mondor (Paris , Gallimard, 1962).

(8) Michel Beaujour, "Temps et Substances dans *Voyage au bout de la nuit*", *Cahiers de l'Herne* V (1965), p. 178.

(9) "La Vision prophétique : Un procédé stylistique célinien", *Pour une poétique célinienne*, éd. J.-P. Dauphin (Paris , Lettres Modernes, 1974), pp. 41-55.

(10) *Structures et Communication dans "la Jalousie" d'Alain Robbe-Grillet* (Sherbrooke : Naaman, 1981).

(11) André Gide, *L'Immoraliste* dans *Romans, Récits, Soties, Oeuvres Lyriques*, éds. Y. Davet et J.-J. Thierry (Bibliothèque de la Pléiade; Paris , Gallimard, 1958).

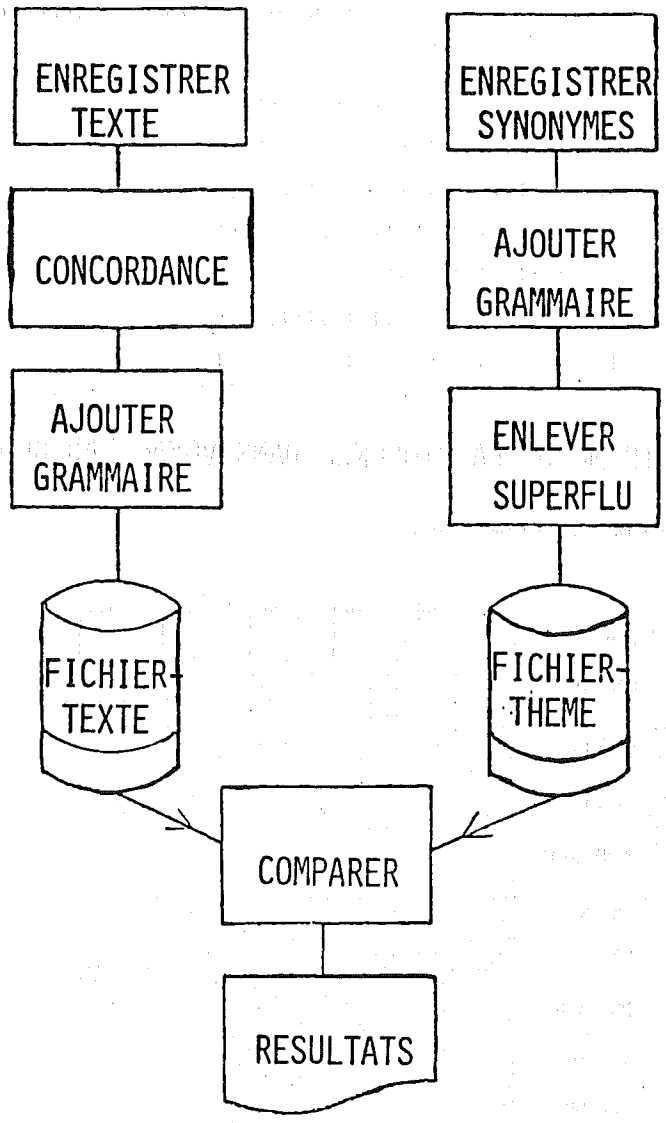

FIGURE 1: LE SYSTEME THEME

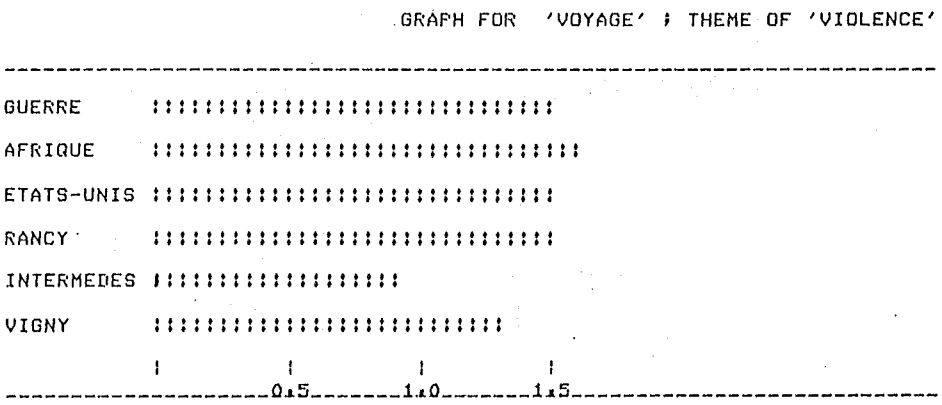

FIGURE 2: LE THEME DE LA VIOLENCE DANS VOYAGE AU BOUT DE LA NUIT

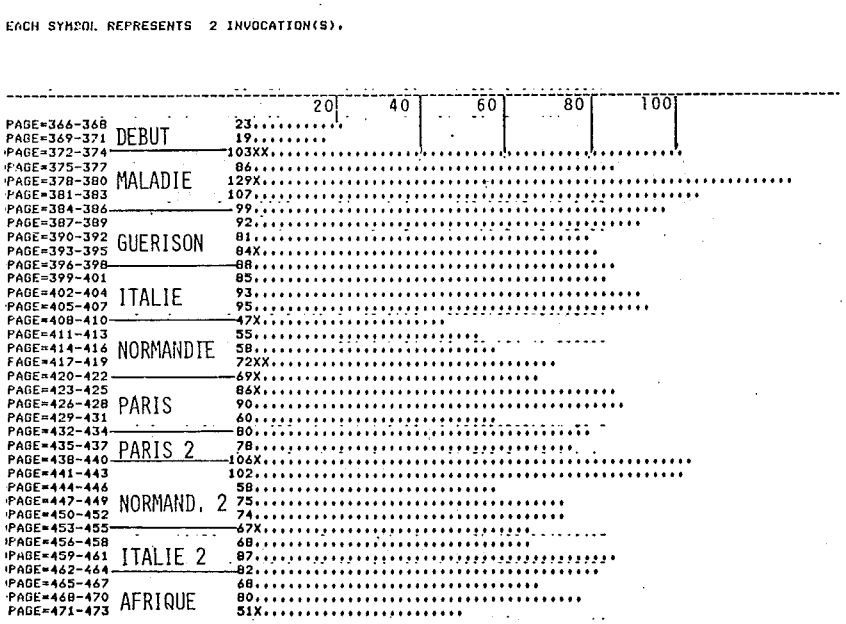

FIGURE 3: PRONOMS A LA PREMIERE PERSONNE DU SINGULIER DANS L'IMMORALISTE

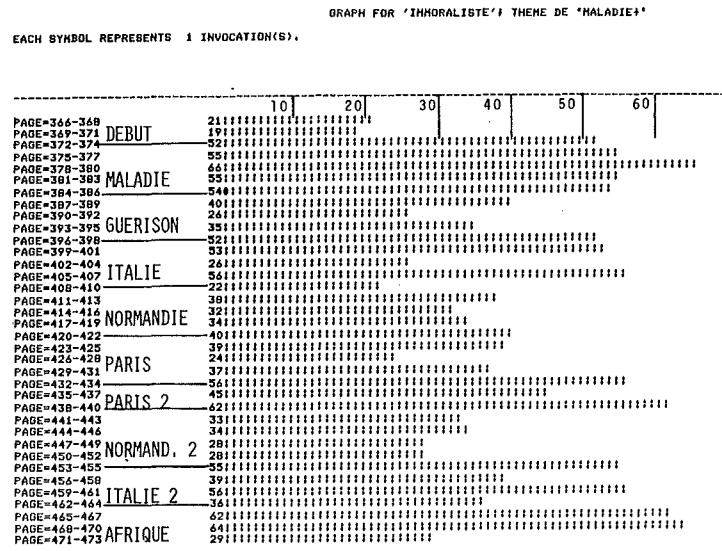

FIGURE 4: LE GROUPE THEMATIQUE "MALADIE" DANS L'IMMORALISTE

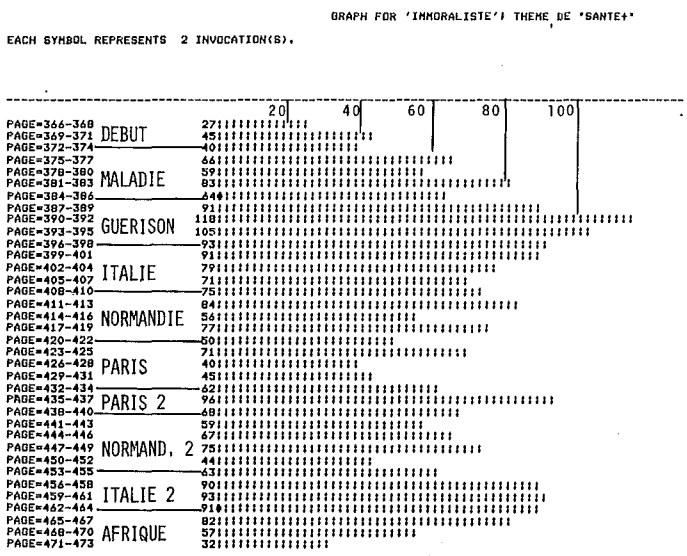

FIGURE 5: LE GROUPE THEMATIQUE "SANTE" DANS L'IMMORALISTE