001 Busa Notre thème est l'interprétation statistique globale d'une oeuvre complète dans le but de trouver, s'il y en a, des caractéristiques personnelles du style. Le sujet du style est inépuisable: il y a bien des sciences qui s'en occupent.

Je dois d'abord préciser l'opposition entre contenu et forme d'un texte. Le contenu est le message conceptuel, ce que l'on veut dire, la forme, c'est l'expression: l'ensemble des mots avec lesquels on le dit. Nous trouvons là un premier sens du mot « forme » opposé au « contenu », mais à l'intérieur même de la forme on trouve une autre opposition entre « forme » et « matière »: forme de la forme, matière de la forme de l'expression d'un texte ...

J'appelle « matière » toute unité verbale, c'est-à-dire le vocabulaire comme masse atomique de mots. Bien sûr je ne veux pas me plonger dans la mer sans fond qu'est la définition du mot « mot ». La forme est avant tout le choix des mots: il y a toujours des raisons pour lesquelles on emploie un mot plutôt qu'un autre. Deuxièmement la forme est la distribution des mots, leur assemblage. Il y aurait beaucoup à dire en profondeur à propos de cette distinction entre « forme et matière » dans l'expression. Ce n'est pas tout simplement l'opposition entre signifiant et signifié, car on trouve cette opposition à l'intérieur du signifiant.

Quand j'étais jeune, des livres célèbres parlaient du style. Auerbach, Croce, Devoto, Spitzer ... parlaient de la créativité de la langue, de sa liberté. L'attention était concentrée sur l'ensemble, pour le savourer en bloc, et sur quelques détails pris comme échantillon.

Pour nous ici, le problème est tout à fait différent: comment formaliser le style pour en faire une enquête quantitative et mettre en évidence, de cette façon, les structures sous-jacentes?

Bien sûr une enquête sur l'ensemble doit, avant tout, examiner tous les détails un par un, pour les interpréter ensuite tous à la fois, globalement.

À ce propos une question, à mon avis, est encore ouverte: est-ce que dans l'ensemble d'un texte il y aurait des signes qui le représenteraient en tant qu'ensemble? Le sens global du texte est déjà formalisé par tous ses mots, mais est-ce que parmi cette foule de mots il y aurait des détails qui par eux-même seraient des signes caractéristiques de l'ensemble, de la même façon que l'on mesure par le pouls la température globale du corps, ou que le tableau de bord revèle les conditions de la voiture?

Notre première session sera consacrée à l'examen statistique des mots individuels. La deuxième aux ensembles progresifs des mots associés. La troisième viendra recycler le sujet de la première, la quatrième le sujet de la deuxième.

002 BATAILLON L'ensemble d'un texte par rapport à quoi? Aux oeuvres du même genre littéraire d'autres auteurs, ou bien aux autres oeuvres du même auteur? Par ex. le Commentaire de St. Thomas sur les Sentences le comparer à celui de St. Bonaventure ou aux Questions Disputées du même St. Thomas?

003 BUSA Mon idée est qu'il y a là une progression des ensembles.

A) Pour certains types de structures, on commence par relever leurs quantités dans chaque proposition du même ouvrage. On les récapitule par paragraphe, ensuite par chapitre, et enfin on en fait la récapitulation finale dans l'ouvrage. Pour d'autres types de structure, par exemple la longueur des phrases ou des paragraphes, la seule récapitulation dans l'ouvrage entier serait possible.

B) On fera ensuite la même analyse pour chacun des autres ouvrages de l'auteur. On en classera les résultats selon les genres littéraires et selon les sujets. On verra si et jusqu'à quelle profondeur ces différences ont varié le style de l'auteur.

C) Enfin si on a fait la même analyse des textes d'autres auteurs du même milieu culturel, du même genre et traitant du même sujet, on pourra mesurer si et jusqu'à quelle profondeur la différence d'auteur affecte le style malgré l'identité du sujet et du genre.

004 VOICU Est-ce que, au delà du texte brut, les structures, telles que vous les avez définies, suffisent pour caractériser le style de façon univoque? Pour en donner une description synthétique et chiffrée au delà des mots du texte?

005 BUSA *Individuum non est ineffabile* voulait dire qu'il n'y a pas de situation qu'on ne peut pas formaliser ... En effet toute expression est déjà formalisation de ce que nous pensons. La question est de savoir si on peut avoir une formalisation de type informatique, qu'on puisse insérer dans un programme d'ordinateur: on peut se demander si la formalisation de la langue parlée est entièrement convertible dans celle de l'informatique. Et au fond votre question est philosophique: est-ce qu'il y a des qualités qui ne pourraient jamais être quantifiées? La question touche le fond même de l'être. Même la theologie trinitaire est passible d'arithmétique: *una natura, duo processiones, tres personae, quattuor relationes, quinque notiones* ... En plus le Verbe de Dieu s'est incarné, quantifié, matérialisé, avec dimensions et poids, avec un tas de structures bio-physio-chimiques, etc.

Mais pour l'emploi informatique on peut réduire la question à peu de mots: est-ce qu'on pourra formaliser le sens global et le style global d'un texte avec des symboles qui soient moins nombreux que les mots du texte entier?

Pour attaquer le sujet de cette première séance, statistique des mots individuels, j'aimerais ouvrir une enquête sur les types

de mots qui, jusqu'à présent, ont été choisis pour des recherches d'identité de style: les mots grammaticaux plutôt que ceux de contenu absolu? Les conjonctions (comme mon ami A. Morton avait essayé, il y à plusieurs années, dans les Epîtres de Saint-Paul), les prépositions, les verbes auxiliaires ou, au contraire, des mots spécifiques, particuliers du *stylus sublimis* ou du *sermo humilis?*

Dans St. Thomas les mots grammaticaux représent 67 % du texte, mais les mots *hapax* 30 % du texte ... Une question préalable se posera: est-ce que nous possédons une typologie des mots, fondée et exhaustive? Sur quels types va-t-on commencer à chercher les caractéristiques des styles? Est-ce que quelqu'un a déjà publié la liste des types des mots et la liste des situations sur lesquelles se sont fondés ceux qui ont fait des recherches d'identité de style?

006 TOMBEUR Il y a quand même quelque chose qui me gêne, père. Je crois qu'il faut un peu revenir en arrière. Vous avez à un moment donné fait une parenthèse quand vous avez opposé matière de la forme et forme de la forme. Vous avez dit: matière de la forme, c'est toute unité verbale, et vous avez ajouté immédiatement: sans entrer dans la discussion de ce que c'est qu'un mot. Eh bien, je crois qu'il nous faut y revenir. Mon sentiment est qu'il faudrait que nous y revenions le plus tôt possible pour partir dans la clarté. Car on ne sait pas répondre à votre question de manière claire, quand vous demandez quels sont les mots que l'on a retenus pour des recherches d'authenticité. Vous citiez le cas des 60 % de mots grammaticaux, mais je suis persuadé que si l'on faisait un tour de table, nous manifesterions bien des divergences dans l'analyse de ces mots. *Quamobrem* ou *quam ob rem*: est-ce un mot ou trois mots? Je vous donne un exemple latin, mais le phénomène est identique dans d'autres langues. Je crois qu'il faudrait d'abord établir quelques définitions pour que nous puissions progresser. Cette discussion doit cependant être abordée pour que nous appelions les mêmes choses

par les mêmes mots. Des considérations de Georges Mounin me viennent spontanément à l'esprit – dans l'introduction de son Dictionnaire de la linguistique – au sujet de la nécessité d'une hygiène intellectuelle: nous manquons de définitions.

007 BERNI CANANI Je poserais le problème de la façon suivante: on me donne un texte dans une langue que je ne connais pas et on me demande de le décrire; que devrais-je faire? Je devrais pouvoir définir en termes de quantité n'importe quelle séquence de symboles, même s'il ne s'agit pas d'un texte à proprement parler.

008 BUSA D'accord avec Berni Canani, j'oppose une définition informatique et tout à fait pragmatique du mot à une définition philosophique ... Je voudrais éviter de faire ici de la philosophie sur le mot « mot ». Pour deux raisons: A) Il faut séparer la définition du mot de la définition des frontières de son unité; B) La philosophie, à son niveau le plus profond que j'appelle celui de l'ontologie, s'occupe exactement de ces mots qui n'ont pas de frontières. La scolastique les appellait les « transcendentaux »: *res, ens, unum, bonum, aliquid, verum*, avec tous leurs synonymes, soit propres, soit métaphoriques, sont des mots dont chacun vaut l'autre: *convertuntur*.

Mais heureusement la définition informatique du mot dans nos langues occidentales est bien claire: une séquence continue de graphèmes séparée des autres par des espaces ou par des signes de ponctuation.

Quand je rencontre *quamobrem*, c'est un mot, quand je rencontre *quam ob rem*, il y en a trois. Le premiers pas de la statistique, c'est de relever avant tout les unités informatiques; c'est seulement dans les phases ultérieures de leurs classifications catégorielles qu'on pourra grouper le syntagme *quam ob rem* avec le mot *quamobrem*.

009 BATAILLON Il faudra pas oublier que dans les éditions des textes anciens, il y a des coupures de mots qui sont dues à l'éditeur et pas à l'auteur. Par ex. le *quoad* de l'édition de la Somme Théologique est separé comme *quo ad* dans les éditions suivantes ...

010 BUSA Ce problème est encore plus grand quand on travaille sur des discours parlés, c'est-à-dire sur des textes phonétiques, où la coupure des mots n'existe pas. Pour ce que vous venez de remarquer, dans mon IT, j'ai été obligé d'attacher un code à certains lemmes qui alerte: « tu dois faire attention, car ce lemme tu pourras le trouver partagé en mots separé »: par exemple, *abinvicem, adinvicem, adinstar, filiusfamilias, respublica.* En effet ils ont été enregistrés, élaborés, et documentés tels qu'on les a trouvés dans le texte, tantôt d'une façon, tantôt de l'autre.

011 MULLER Question: quels sont les mots à choisir pour les recherches d'authenticité? Dans des expériences sur des textes d'anciens français dont l'authenticité n'était pas établie, on a pris en considération les mots les plus fréquents, sans préciser s'ils étaient des mots outils ou des mots pleins, en éliminant ceux qui n'avaient pas une fréquence stable; c. à d. qu'on s'occupait plus de la stabilité que de la fréquence; on divisait le texte en un certain nombre de tranches et on faisait un calcul des variances.

Ce n'est pas la même chose que l'usage. L'usage est une combinaison de fréquence et de répartition, tandis que le calcul des variances est plus conventionnel, et il dépend de la longueur des tranches ...

Dans un roman du Moyen Age, une des versions de Tristan, pour savoir si une des grandes parties était du même auteur que les autres, on a pris le mots les plus fréquents. En effet c'étaient en général des mots grammaticaux. On a verifié leur stabilité plus ou moins grande et ensuite on a cherché s'il y avait des différences entre les grandes parties. La longueur des tranches dé-

pend de la fréquence des mots qu'on étudie. Par tranche on devrait avoir 5/10 occurrences du mot qui nous intéresse. Par conséquent, s'il y a un mot de fréquence 500, on pourrait diviser le texte en 50/100 tranches ...

012 PETÖFI We have discussed in detail, how to interpret the notion « word ». How is it, however, necessary/expedient to interpret the notion « style »? Is it necessary/expedient to consider the style of a person or the properties of a gender – the style of Thomas Aquinas or the properties of the gender(s) to which the texts of Thomas Aquinas can be reckoned – as the object of investigation?

013 BRUNET Le père Busa cherche un peu trop la pierre philosophale: un type de mots ou un indice spécifique, qui pourrait donner une réponse aussi difficile que celle que vous proposez. C'est tout l'ensemble des mots qui peut apporter des conclusions. Je ne crois pas qu'on puisse en privilégier un seul.

J'ai aussi attaqué un problème semblable au vôtre: l'étude de l'ensemble du Trésor de la Langue Française qui fait 150 millions de mots. Dans cet énorme ensemble tous les mots, toutes les catégories, réagissent violemment à la chronologie et au genre, le genre littéraire étant le plus déterminant. Votre travail je le ferais vraiment global, sur tous les mots. Vous nous disiez qu'on cherche les structures latentes. Où est votre point de référence pour les trouver? Comparaisons externes ou internes?

Dans votre corpus il y a différents auteurs et dans St. Thomas même il y a les distinctions chronologiques et celles du genre littéraire.

En faisant globale et interne cette approche, on devra prendre l'ensemble comme la norme, suivant une des méthodologies du prof. Muller.

014 MULLER Chez Corneille j'ai vu varier tous les caractères quantitatifs, y compris le mots les plus fréquents. M. Bru-

net a des expériences semblables. Quand on parle de tranches, surtout pour éliminer les mots instables, on doit présupposer qu'il s'agit de tranches à l'intérieur de chaque oeuvre homogène. Je crois aussi qu'il faut considérer soit l'ensemble soit ce qui varie d'une oeuvre à l'autre et à l'intérieur de chacune.

015 BUSA Je voulais tout simplement commencer par établir les démarches progressives de la recherche, étant donné qu'un ensemble est constitué par ses éléments. Si on commence par les mots les plus fréquents, est-ce que ensuite on doit considérer les *hapax* ou non? Est-ce que quelqu'un a déjà appuyé une interprétation d'authenticité sur les *hapax?*

016 MULLER D'authenticité je ne la pense pas, mais de différences de style, oui. Voulant définir la différence de style de deux textes, je commencerais par les deux bouts, en examinant les mots fréquents et après le nombre des *hapax*. Il est certain qu'il ne sera pas le même dans les deux et pas reparti au hasard dans les différentes parties du même ouvrage: bien intendu, toujours en supposant l'homogeneité du texte.

017 BUSA Dans ST 17,42 % des mots sont en citation littérale. On les a signalés tous avec un code approprié, afin de pouvoir les élaborer comme hétérogènes. En outre nous avons signalé les citations *ad sensum*, dont la typologie est multiple et « effilochée »: nous en avons 6 % dispersés partout, mais nous n'avons pas compté les 51 commentaires, pour ainsi dire interlinéaires, où le discours de ST est brodé sur le discours de l'auteur qu'il commente. Bien sûr ces 51 ouvrages devront être examinés tout à fait à part. Cet examen sera encore plus compliqué pour les *reportationes*: ST parlait, un scribe enregistrait: quels mots sont sortis de la tête de ST, quels mots sont sortis de la tête du scribe? Regardez par exemple dans mes Concordances, quand le mot *stylus* signifie « la plume » et quand il signifie « le style » ...

018 VOICU Pour l'Universit!é de Münster j'ai jadis dépouillé des textes des pères grecs afin d'y repérer des matériaux bibliques. Ceux-ci prennent des formes très différentes: il y a des citations indiquées formellement, à tort ou à raison, en tant que telles: d'autres, tout en étant littérales, ne sont aucunement marquées. Même complexité pour les allusions qui peuvent être des paraphrases, des reprises de mots caractéristiques, voire des démarquages des structures d'un récit (Luc dans les Actes des Apôtres reproduit pour le martyre de St. Étienne le schéma de la passion du Christ). Une question à M. Muller: Quelle est l'intérêt de l'analyse de ces mots fréquents « obligatoires » selon les règles grammaticales de la langue?

019 MULLER Les premiers qui ont fait de la statistique lexicale étaient tellement persuadés que les mots grammaticaux sont imposés par des contraintes idiomatiques qu'ils ont établi *a priori* que ces mots là représentent 50 % du tissu de tout texte: et ils les ont laissé tomber. Je ne crois pas que quelqu'un avait pensé qu'il aurait fait des découvertes en les examinant. Quand on a commencé à transcrire le texte sur cartes perforées, on allait plus vite à copier le texte tout entier qu'à charger la personne qui perforait de la tâche de sauter les 40 ou 60 mots signalés. Quand on a commencé à faire des dépouillements intégraux, on s'est aperçu qu'il y avait là des choses intéressantes. On aurait pu croire à une corrélation stricte entre le nombre des articles (défini et indéfini) et celui des substantifs; en fait c'est beaucoup moins simple; ainsi j'ai pu établir que chez Corneille l'effectif des articles (et d'autres mots grammaticaux) varie d'une part avec le niveau de style (opposition tragédie/comédie), d'autre part avec le temps (les pièces de la jeunesse ont moins d'articles que celles de l'âge mûr, qui en ont moins que celles de la fin); d'où une interprétation à deux niveaux; dans ce cas il me suffit de connaître l'effectif de mots comme « le, de, et » pour savoir, compte tenu de la date, si je suis en présence d'une comédie ou d'une tragédie ... C'est une constatation brute; j'ignore si c'est identique

ou semblable chez d'autres auteurs; mais je crois que la densité en mots grammaticaux varie très probablement avec le niveau de style, et toujours dans ce sens.

020 BRUNET Pour répondre à la demande de M. Voicu, je peux fournir quelques exemples de mots grammaticaux très frequents. Ainsi le XX^e^ siècle n'utilise pas autant que le XIX^e^ l'article indéfini « un et une » en français, comme le prouve ce graphique où l'on suit le déclin de l'article indéfini au cours des 15 décennies. C'est le contraire pour l'article défini. C'est là une surprise pour tout le monde qu'on ne peut récuser, car la statistique nous prouve que le hasard ne peut pas produire des courbes pareilles.

Est-ce que dans les basses fréquences on peut atteindre autant des résultats que dans les hautes fréquences? Dans ce graphique voilà les *hapax* dont vous parliez tout à l'heure. Dans mon corpus, il y en a 21.000 et voici leur progression, qui est bien nette. Si le hasard était le seul responsable, tous les points se situeraient dans la petite bande centrale: or il n'en est rien: les points se situent en delà ou au delà de cette zone, suivant une ligne ascendante qui manifeste l'influence directe et imprévisible de la chronologie ...

021 MARINONE Sottolineo la grande differenza fra un testo antico trasmesso direttamente dall'autore e un testo antico pervenuto tramite la tradizione manoscritta, ove intervengono scelte e interpretazioni sia degli scrivani sia degli editori critici. Si aggiunge il problema della tradizione indiretta, nel cui ambito occorre distinguere fra citazioni, passi paralleli e reminiscenze. Quanto alle citazioni, l'esperienza fatta nel gruppo di ricerca da me diretto per l'analisi di testi tardolatini, ha suggerito il seguente schema: 1) dirette, 2) indirette, 3) tradotte, 4) non definibili nettamente; all'interno di ciascun gruppo vengono ancora individuate: A) citazioni attribuite, B) citazioni non attribuite. Ciascun

tipo è contraddistinto da delimitatori di contesto per l'analisi automatica.

022 ZAMPOLLI Sono d'accordo sul fatto che i cosiddetti *mots grammaticaux* devono essere presi in considerazione nei lavori di linguistica quantitativa. È certo che la loro frequenza – contrariamente a quanto alcuni credevano – varia significativamente in dipendenza di diversi *style-forming factors*. Per es.: la frequenza dell'articolo determinativo varia regolarmente tra i 5 sottoinsiemi del LIF (Lessico di Frequenza dell'Italiano), da un minimo di 5000 occorrenze (circa, testi di cinema) a un massimo di 15.000 (circa, testi scientifici). Analoghe variazioni abbiamo riscontrato fra testi di diversi tipi di pazienti psichiatrici. Il problema del fondo della linguistica quantitativa è oggi – a mio avviso – quello di costruire un modello della produzione di testi capace di inglobare i fattori che, determinando – nel corso della produzione degli enunciati – le scelte tra le diverse possibilità offerte dal sistema linguistico, concorrono alla formazione delle frequenze delle diverse unità linguistiche.

023 BERNI CANANI ... Pour la recherche d'ensemble que le p. Busa envisage sur son corpus, c'est peut-être plus important de chercher les co-variances de tous les lemmes dans tous les textes.

024 GRILLI À propos de citation et de paraphrase, deux exemples montrent les problèmes qui viennent du rapport entre la langue de l'auteur et celle de ses citations: les deux exemples sont un texte dans lequel Cicéron traduit du Platon (Rep. 8,562C) et un texte d'Aulu-Gelle qui traduit du Chrysippe (N.A. 7,2). Le lexique de Cicéron est son propre lexique à lui, mais la tournure de sa syntaxe ressents de la syntaxe du texte grec. Au contraire Aulu-Gelle présente un différent contenu de lexique – ayant dans ce cas abandonné sa position de puriste –, mais la syntaxe est identique à celle qui lui est propre dans le reste de son ouvrage.

025 PACIUSKIEWICZ À propos des remarques de M. Marinone, il faut ajouter qu'au Moyen Age n'existait pas le droit d'auteur (copyright) ...

026 MARINONE À propos de lexique je voudrais apporter un cas très significatif. Un auteur latin du 5ème siècle, Macrobe, cite quelques pages de Plutarque, dont une partie est traduite par lui-même et une autre empruntée à la traduction faite presque trois siècles auparavant par Aulu-Gelle. Eh bien, le vocabulaire technique du language médical est différent. Cela impose une attention diacronique toujours veillante dans l'analyse des textes.

027 PETÖFI I think it is not sufficient to distinguish between words having a definite content and words expressing certain grammatical functions. There also exists a third type: the type of words having a definite rhetoric function.

028 ZAMPOLLI Je me rappelle que nous les avions déjà distingué au commencement.

029 BUSA Dans nos textes nous avons caractérisé de grosses portions comme « discours connectif » que j'aimerais appeler « discours-plateau ». Par exemple *Ad primum sic proceditur. Videtur quod. Respondeo. Dicendum quod. Ad primum ergo dicendum quod* etc. ... etc. ... À mon avis le type connectif est une catégorie du discours, et non pas une catégorie de mots individuels. Même dans le discours connectif il faut distinguer « mots absolus » et « mots fonctionnels » ... Par conséquent dans mon IT la *Concordantia Altera* rassemble non seulement tous les mots grammaticaux, mais aussi quelques mots absolus, comme *dicere, modus, ratio, res*, qui ont parfois des prégnances philosophiques profondes, mais en plus sont extrèmement fréquents dans de nombreux types de discours connectif.

En conclusion de cette première moitié de cette première séance, peut-on dire que jusqu'à présent le type de mots inté-

ressant notre problème sont les mots avec stabilité de fréquence?

On a aussi parlé des types du texte et des types du contexte. Les genres littéraires sont types du texte et les contextes sont à l'intérieur du texte. En effet l'auteur du texte introduit parfois des corps étrangers, c'est-à-dire des citations.

Celles-ci présentent plusieurs degrés (par exemple des citations seraient littérales, même quand elles sont traduites?), les plus faibles étant les citations tacites et les reminiscences, pour lesquelles le problème de la définition des limites s'accroît.

On a parlé d'un deuxième type de contexte qui est le discours connectif.

En outre il y a le fait que dans un texte ancien la copie des manuscrits et l'impression ont pu affecter beaucoup plus que la seule orthographie du texte.

L'interprétation statistique des données quantitatives doit évidemment tenir compte de tout cela, dans la recherche de la paternité du style. En théorie l'analyse statistique devrait aboutir à découvrir quelles variations ont été apportées au texte après sa production par l'auteur. Mais il semble que nous sommes encore confinés dans une recherche préalable: sur quels éléments du texte pourrait-on fonder une telle recherche statistique?

## 1.2

030 Busa Notre problème a deux aspects: l'aspect linguistique qui est la reconnaissance de la typologie du texte, et l'aspect de son analyse statistique. Ce matin on a déjà dit qu'il faut examiner avant tout les mots qui ont une stabilité de fréquence dans des tranches homogènes. Même l'analyse statistique des mots individuels doit être globale, c'est-à-dire qu'elle doit être menée sur tous les mots du texte. J'imagine par là que l'examen des pourcentages offrirait des informations précieuses. Prenons par exemple dans ST deux mots: *ordo, inis* et *connaturalis-*

*itas*. Leur fréquence absolue ne me dit rien, mais le pourcentage de *ordo* chez ST est presque le double de celui des autres auteurs de mon corpus. Par contre le pourcentage de *connaturalis-itas* est légèrement plus bas chez ST que chez les autres. Est-ce qu'il serait possible d'en conclure que le mot *ordo* chez ST est un mot « plus important » et qu'au contraire *connaturalis-itas* chez ST n'avaient pas plus d'« importance » que chez les autres, alors que la théologie d'aujourd'hui les a célébrés?

Dans mon IT j'ai donné plusieurs pourcentages pour chaque mot. Je me demande s'il faudrait aussi couper les gros ouvrages en tranches et dans quelles dimensions, pour relever les variations de leur répartition à l'intérieur des gros ouvrages. Je me demande aussi si à la fin il n'y aurait pas moyen de chercher des pourcentages globaux de tous les pourcentages des mots individuels.

031 MULLER L'expérience a démontré qu'en statistique il faut commencer par de choses très simples et après chercher des choses plus subtiles. Par exemple, si on cherche *spiritus* ou *anima*, mots qui ont plusieurs sens, il faut d'abord compter tout simplement et bêtement leurs occurrences. Après on verra si les différents sens sont répartis différemment. Est-ce-que vous ne pensez pas que la première opération à faire serait de commencer par le vocabulaire le plus fréquent, et de voir comment il se répartit entre les sous-ensembles de votre corpus? Je prends un exemple de votre IT: *vero* est employé 14.400 fois: dans les *Sententiae*, qui font 17 % de votre corpus, je m'attends à 2450 occurrences de *vero* ... Il faudra les compter: on calcule facilement deux seuils (ici environ 2360 et 2540) entre lesquels les variations ne sont pas significatives; ... si on est au-dessus, la fréquence est forte, au-dessous elle est faible. De cette façon, dans chacun des sous-ensembles, on trouve les vocables anormalement fréquents ou anormalement rares. L'interprétation doit ensuite examiner si telle fréquence en telle partie est due aux citations ou au thème etc. Quand on possède un gros corpus enregistré, déjà codé, bien analysé au niveau du lemme et de la morphologie, la seule mé-

thode est de passer l'ensemble du vocabulaire en descendant jusqu'à une certaine fréquence ...

032 BUSA Vous avez pris comme exemple le mot *vero* qui est un homographe: ou adjectif de *verus-a-um*, ou adverbe-conjonction. Est-ce qu'il faut avant tout trier sémantiquement les homographes avant d'en examiner la fréquence?

033 MULLER Quand il s'agit de véritable homographie, oui. Quand il s'agit de variance sémantique, non.

034 MARTINI Ho un'urna contenente tanti segni. Devo descrivere il modo con cui vi sono distribuiti. Il suggerimento primo e ovvio è quello di contare le frequenze delle forme delle parole, frequenze che sarebbero identiche se le stesse parole avessero un altro ordine: parità di frequenze potrei averla da testi diversi. Ma le parole nel testo hanno un certo ordine, che è struttura organizzata e non casuale. Per tale caso la statistica suggerisce l'analisi multivariata, che descrive le associazioni tra le parole, a 2 a 3 a 4 ecc., cioè non le qualità delle unità singole, ma quella delle loro combinazioni. Come diceva il prof. Muller, vi è poi l'approccio di spezzare l'insieme in tanti sottoinsiemi e verificarvi stabilità e co-varianze, onde vedere se le differenze siano casuali o sistematiche. Vi è poi l'analisi classificatoria, che esamina le parole a tipi ...

035 BUSA Per me l'interesse non è « più » sulle parole singole « che » sulle loro associazioni, ma « prima » sulle une e « poi » sulle altre.

036 PETÖFI Is it possible to count the clauses and sentences of a text (i.e. to break up a text into clauses and sentences in an automated way)? if the number of the clauses could be determined, the frequency of the words could be compared with the number of the clauses. (When determining the frequency of the words it would be necessary, however, to solve the pro-

blem of re-pronominalization). If the number of the clauses cannot be determined but that of the sentences could, then the middle length of the sentences measured in words could be calculated, and the parameter which we obtain by dividing the number of the words occurring in the sentence to be characterized by the midle length of the sentences could be regarded as a numerical characteristic of this particular sentence.

037 Busa That is quite an intriguing question. Pour ST il y a autant de possibilités de couper le texte en propositions et phrases qu'il y en a de se fier ou de se méfier de la ponctuation. Nous avons gardé la ponctuation telle qu'elle a été imprimée dans l'édition la plus récente. Elle varie beaucoup d'un éditeur à l'autre, surtout quand on cherche à établir les frontières entre ponctuation lourde et ponctuation légère.

Mais je pense qu'on peut se fonder sur cette ponctuation dans la plus grande partie des cas, pour reconnaître au moins les propositions élémentaires, c'est-à-dire ces unités minimales complètes d'expression, qui pivotent autour d'un verbe.

038 Berni Canani Il y a de très forts rapports entre distribution statistique et classificatoire. Le nombre des livres qui sont classés sous un mot-vedette dans un catalogue de bibliothèque, a une distribution semblable à celle des mots dans un texte: il y a des mot-clés suivis de beaucoup de livres et il y en a qui sont suivis d'un seul livre, comme des *hapax* ...

Mais il y a là un problème: pour former des classes, j'ai besoin que les données soient vraies: par. ex. dans une statistique de la consommation de café, je dois compter que ceux qui disent « J'en prends 3 par jour » disent la vérité. Analoguement pour les mots: je dois être sûr que A signifie A et B signifie B; mais la complémentarité entre polysémie et synonymie empêche une analyse classificatoire directe. En outre, examiner les connexions que chaque mot a avec les autres mots, ce n'est pas encore l'analyse du comportement global ...

039 BIFFI Ho usato l'IT, anche se non per quel migliaio di anni che richiederebbe l'applicazione rigorosa del metodo di p. Busa. L'esempio che ci ha portato di *ordo* e di *connaturalitas* conchiude a un giudizio di « più importanza ». Ma per saperlo, dovrei far attraversare dalla diacronia storica l'unidimensionalità necessaria dell'IT. Il lemma *Trinitas* ricorre meno volte di *causa*: potrei conchiudere che è meno importante? Un uso scientifico dell'IT potrà portare in tempi umani a fondare giudizi di importanza?

040 BUSA Ho voluto solo prospettare una equazione: *ordo* in ST sta a *ordo* negli altri autori del mio corpus due volte di più di quanto *connaturalis* in ST stia a *connaturalis* negli altri autori. Ne ho dedotto una domanda e non ancora un'affermazione: può questo fatto essere interpretato come segno che ST concentrava un particolare interesse su *ordo?*

041 BATAILLON Io dico di no, perché i due gruppi non sono paragonabili. Tra i testi di altri autori mancano nell'IT Commenti alle Sentenze e Questioni Disputate, cioè le parti veramente teologiche. Vi sono eterogeneità cronologiche e di genere letterario. Per queste due parole la comparazione esatta andrebbe fatta nei Commenti alle Sentenze di ST con quelli di suoi contemporanei come S. Bonaventura o Pietro di Tarantasia.

042 BUSA Lei fa molto bene a sottolineare che, nell'interpretare diversità di frequenze presso diversi autori, vanno tenuti presenti il genere letterario, l'argomento e la cronologia. Sarebbe un capitolo di metodologia da sviluppare quello di procedere ulteriormente a definire norme, condizioni, modalità e procedure da seguire nella comparazione filologica-linguistica di testi diversi. In merito, io avanzo solo una domanda: con quali diverse misure vada applicato a diversi tipi di parole quanto Lei ha messo in luce. Infatti accanto a parole come « matrimonio » e « creazione » vi sono altre parole che ricorrono in qualsiasi discorso su qual-

siasi tema: anzittutto le parole grammaticali, ma poi anche alcune voci non grammaticali come uno e molti, tutto e parte, fare ed esser fatto, vero e falso.

In merito poi alla domanda di Mons. Biffi su quanti secoli occorreranno per indurre dall'IT un lessico di ST possiamo estrapolare delle previsioni da come procede il *Thesaurus Linguae Latinae*, il quale ha un materiale testuale di quantità press'a poco identica a quello di ST. In circa ottant'anni è arrivato alla lettera « p ».

043 Betti Voglio sottolineare la necessità di passare dalle quantità con cui opera la statistica ad una rappresentazione qualitativa della struttura presente in un dato testo.

L'esperienza della matematica suggerisce che questa struttura sia indagata localmente (per mezzo ad esempio di « funzioni » definite su parti del testo) pur di essere in grado di controllare l'estensione delle indagini da una parte ad un'altra.

Un altro aspetto che è finora emerso riguarda l'incompatibilità dal punto di vista operativo fra le indagini sul testo di un filologo e di un matematico. Per il primo il testo è un punto di arrivo e l'indagine è la ricostruzione delle circostanze che l'hanno prodotto, mentre il matematico, insensibile a questa problematica, considera il testo come un punto di partenza, un dato primitivo su cui condurre le proprie ricerche di leggi generali, regolarità, invarianti, di tutto quello che rientra genericamente sotto il termine struttura.

044 Fattori Non ho la competenza per discutere se i dati « quantitativi » abbiano da soli un significato ermeneutico e interpretativo. Volevo però sottolineare – sulla base della mia esperienza con il Lessico del *Novum Organum* – che, se è vero che i dati vanno ulteriormente elaborati e interpretati dallo storico della filosofia, dal filosofo ecc., è indubbio che essi, assunti come uno dei tanti « fatti », spesso aprono e impongono – attraverso i loro reciproci rapporti, l'assenza o la presenza di un ter-

mine, la frequenza o la cooccorrenza – ipotesi di ricerca altrimenti impossibili da cogliere o da individuare.

045 CRESPI REGHIZZI Il ne faut pas confondre l'analyse formelle de la forme du texte avec ses conclusions sur le contenu, comme par ex. l'importance du mot *ordo* chez ST. Alors il faut établir lequel est le bout de cette analyse. S'il s'agit de voir la fréquence de l'article dans la langue française du siècle dernier et de celui-ci, on fait de la statistique et on s'arrête là. Mais la pensée est un phénomène très complexe: personne n'est capable de le représenter complet et detaillé dans la memoire d'un ordinateur. Pour résoudre la polisémie d'un texte il faut faire recours au sens des mots. En analogie avec le classement informatique des *patterns* il faut commencer par une *feature extraction*, définir des fonctions pour reconnaître des catégories.

046 ZAMPOLLI Il problema evocato da Betti e Crespi Reghizzi è davvero centrale.

Posso solo ricordare quanto ho già detto a proposito della mancanza di un modello della produzione di testi. Un problema di grande peso per il *linguistic and literary computing* è trovare delle metodologie che, assistendo il ricercatore, riducano i tempi necessari per utilizzare convenientemente la sempre crescente quantità di dati prodotti con gli spogli elettronici. Per quanto affinate, le procedure di spoglio producono essenzialmente gli stessi risultati da 30 anni a questa parte. I primi gridi di allarme e le prime discussioni sulla sproporzione tra dati in *machine readable form* e capacità di utilizzarli da parte del linguista-lessicografo risalgono già al 1972, in particolare alla 2ª *International Summer School* da me organizzata a Pisa (interventi di Quemada, Aitken, Bailey, ecc.) e al convegno di Montréal. Il *Workshop* organizzato lo scorso maggio a Pisa per conto della *European Science Foundation* sul tema *Possibilities and limits of the computer in producing and printing dictionaries* ha riproposto il problema, e suggerito delle vie da seguire. Anche al *Coling 80*

sono stati presentati dei lavori in questo senso; per es. Lara (Messico) ha illustrato una procedura per un primo « sgrossamento » automatico che riduce di circa la metà il numero di contesti da esaminare nelle concordanze, eliminando quelli che sono (probabilmente) ripetizioni di uno stesso *pattern*. Questo problema diventa ancora più pressante con l'avvento dei lettori ottici *omnifonts*. In questa direzione di ricerca, inoltre, mi sembra che potrebbero convergere i due filoni o poli tradizionali dell'*automated language processing* (designati di solito come « modelli computazionali » e *text-processing*). I sistemi che ne risulteranno dovrebbero vedere, secondo me, la costituzione di basi di dati linguistiche, corredate da procedimenti interattivi di analisi, di strumenti linguistico-computazionali (come dizionari e analizzatori), di procedure statistiche di analisi, ecc. Cruciali, e a quanto mi consta ancora tutte da immaginare, le analisi delle esigenze e della tipologia delle operazioni richieste dalle diverse categorie di possibili utenti di una tale base di dati.

047 PETÖFI First: I think, instead of speaking of importance, it would be better to speak of the most frequent topics in the different parts of a text. (In this connection one must pay special attention to the analysis of the negative sentences). Second: as far as pronouns are concerned, it is difficult to distinguish between form and content. I do not know how it is possible to find an interpretable correlation between the number of the pronouns and the number of the nouns. Third: it would be necessary to pay attention to the borderlines of the paragraphs, because they are very seldom by-passed by pronominalization - if at all. In a new paragraph the pronominalized noun is generally repeated.