2.1

048 BUSA Hier, le thème – pas toujours la discussion ... – était la statistique globale des mots individuels. Ce matin il faut examiner la statistique des mots groupés en ensembles, c'est-à-dire des ensembles du discours en tant qu'ensembles. Bien sûr les ensembles montent avec une progression concentrique ou, si vous voulez, arborescente: binômes, trinômes, propositions, paragraphes, chapitres, ouvrages, oeuvre.

049 MULLER Une première démarche, très grossière, serait d'examiner le sous-fréquences des sous-ensembles par rapport à l'ensemble. Mais il faut y introduire les notions de l'intervalle entre les occurrences et celle de l'environnement c.-à-d. du contexte.

A Nancy, on a recherché les groupes binaires français qui se rencontrent plus souvent que le hasard le ferait. On y trouve des choses sans intérêt, par ex. que le mot « cheval » est le plus souvent précédé par le mot « le », et beaucoup de verbes par « je », mais on a quand-même ramassé des syntagmes, qu'on trouve déjà dans le dictionnaire après de quelques lemmes.

A St.-Cloud on a étudié les co-occurrences: pour chaque pôle, on regarde 5 mots avant et 5 mots après, et l'on cherche ce qu'il y a de plus fréquent, en dehors des articles et choses pareilles. M.me Dolphin a appliqué cette méthode aux mots « oeil, yeux » dans un « roman fantastique » de Jean Ray et en a tiré des conclusions intéressantes.

Pour rechercher sur les structures syntaxiques, généralement on ne compte pas les mots mais les fonctions comme par ex. sujet, verbe, complément ...

050 HAMESSE Dans l'énorme corpus de S. Thomas que vous avez, père, ne faudrait-il pas comme préalable des distinctions? Hier, on a parlé des citations, de la manière dont les

médiévaux citaient de façon littérale ou non, des réminiscenses etc. Ne faudrait-il pas faire aussi une distinction entre les oeuvres qui ont été rédigés par S. Thomas et les autres? Je connais mal le corpus de S. Thomas, mais je connais un peu celui de S. Bonaventure: plusieurs oeuvres de ce dernier ont été reportées. L'expression, dans ce dernier cas, est-elle de l'auteur ou du reportateur? L'auteur a-t-il revu le texte de la *reportatio*? On ne le sait pas toujours. Ne faudrait-il pas dès lors, au lieu d'étudier immédiatement de l'oeuvre, se comporter différemment et partir d'un échantillon plus limité? On pourrait, par exemple, commencer par l'analyse des opuscules appartenant à un même genre littéraire, et puis élargir peu à peu le corpus. Vouloir tout prender d'emblée, c'est risquer de comparer des choses dissemblables qui ne sont pas comparables.

051 Busa Je ne veux pas partir du tout, mais arriver au tout. Il faut partir de l'analyse de toutes les parties homogènes élémentaires, en commençant par les plus petites et en montant progressivement vers la fusion dans un cadre plus en plus général.

052 Hamesse Si vous prenez deux mots et que vous présentez l'analyse de cette expression dans toutes les oeuvres de S. Thomas, il y a une différence entre les résultats obtenus pour les ouvrages rédigés pas S. Thomas lui-même et pour les oeuvres qui ne sont pas de sa main. Il est difficile de tirer des conclusions générales, lorsqu'on s'occupe de textes rédigés de manière très différente.

053 Busa Jacqueline, est-ce que tu as regardé ma *Concordantia altera*? ...

054 Muller L'homogénéité du texte est une question; une autre question, c'est si l'on ferait la seule analyse des unités lexicales ou bien aussi des chaînes syntaxiques, en faisant abstrac-

tion des mots et en ramenant à la même classe de phrases toutes les phrases avec un même schéma syntaxique bien qu'avec des mots différents.

055 Busa À mon avis, les deux. Comme il s'agit d'énumérer des possibilités de méthode, je pense qu'il faut envisager soit des analyses d'unités lexicales, soit des analyses des types de structures syntaxiques. Evidemment il ne faudrait pas les confondre, ni les traiter en même temps avec les mêmes procédures: « distinguer pour unir » a été le titre d'un livre célèbre.

056 Zampolli Nell'IT c'è ancora del lavoro intermedio di codificazione delle strutture che deve essere compiuto.

057 Berni Canani Pour les groupements de mots on peut tenir ou ne pas tenir compte de la fréquence. Dans le premier cas, les procédures visent à chercher les probabilités conditionnées que chaque mot a d'apparaître précédé ou suivi par tels ou tels autres mots entre des limites pré-établies. Dans le deuxième, à voir à partir des co-occurrences quelles sont les mesures d'association possibles; à en prendre une comme seuil et à lister tous les mots qui sont au-dessus du seuil. L'ensemble de ces mots n'aura pas toujours la même forme: cela servira à établir des classements. Avec les mots qui ont un sens net et défini, on aura de bonnes partitions, c.-à-d. ils seront facilement classifiables. Le seront beaucoup moins les mots qui ont une gamme de significations sans frontières nettes. Cette procédure met en lumière des différences significatives entre mot et mot: les uns associés au-delà d'un certain seuil à peu de mots, les autres à plusieurs.

058 Petöfi In this connection I want to call attention to the third volume of the Frequency Dictionary of Present-Day Swedish (Stockholm 1975). In this dictionary word-chains consisting of *n* elements are treated on the following three levels:

(1) level of word combinations; (2) level of constructions; and (3) level of idioms. The theoretical considerations underlying this distinction are discussed in the Introduction to the dictionary.

059 Busa Dans le groupement des mots, il faut éclaircir deux aspects. Il y a d'abord les corrélations que j'appellerai sémantiques, comme, par exemple, entre un substantif et son adjectif: *vinum bonum* et *bonum vinum* sont sémantiquement une unité d'expression identique. Il y a toute une gamme de corrélations sémantiques: par exemple parmi celles que j'appelle grammaticales ou syntaxiques (c'est-à-dire entre les parties du discours) on a des types comme *bonum vinum* et *ipse dixit* ... Par contre *et ideo* et *sed contra* sont des mots idiomatiquement juxtaposés, mais en corrélation l'un avec l'autre de type différent des précédents. Par ailleurs il y a aussi des co-occurrences disons purement topographiques ou physiques: dans le texte le mot *A*) est suivi par le mot *B*) et celui-ci par le mot *C*). Il s'agit là d'une donnée brute qui fait abstraction de l'existence ou non entre les mots contigus, de quelque corrélation sémantique. Plusieurs experts ont déjà recherché quels rapports il y a entre les deux types de groupements de mot. Nous, nous l'avons fait deux fois, chaque fois sur un échantillon d'un demi-million de mots. Nous avons réparti le texte en tétranômes, trinômes etc.: par exemple avec la phrase *Ad primum ergo sic proceditur*, on avait les trinômes suivants: – *Ad primum / Ad primum ergo / primum ergo sic / ergo sic proceditur / sic proceditur* –. Dans les trinômes chaque mot est pris trois fois, en première position, en deuxième et en troisième, et les ponctuations lourdes y sont incluses. Ensuite on les triait alphabétiquement et on comptait les polynômes identiques. Sur la liste récapitulative des polynômes différents on faisait noter leur fréquence. Enfin on lisait cette liste en cherchant quels polynômes offraient la possibilité d'être en même temps polynômes de corrélation sémantique. Le premier essai a été fait sur des mots bruts, c'est-à-dire pas encore lemmatisés. Les résultats ont été une quantité énorme de plurinômes différents et très, très peu

de plurinômes à haute fréquence, et ceci à un tel point qu'il n'était pas viable d'en extraire les corrélations éventuellement sémantiques. La deuxième fois, sur un autre demi-million de mots, cette fois lemmatisés, on a obtenu quelques résultats, quand, finalement, on a trié les polynômes, non selon leurs formes, mais selon leurs lemmes. Par exemple non plus *actus et potentia* comme trinôme distinct de *actum et potentiam*, mais *actus* (ou *actum* ou *actibus*, etc. ...) suivi de *et*, suivi de *potentia* (ou *potentiam* ou *potentiis*, etc. ...). Cette opération était possible, car, sur mes bandes magnétiques, les différentes formes d'un même lemme ont des codes morphologiques différents, mais un code lemmaire identique. À la fin on s'est retrouvé avec 600 syntagmes d'une fréquence appréciable. Mais – et cela nous sembla significatif – plus de 500 étaient des binômes, seulement quelques douzaines comptaient 3 ou 4 mots. Dans ces deux recherches le travail de base a été celui de la machine. Mais dans d'autres recherches, celles qui portaient sur les corrélations sémantiques, le travail de base a été celui du philologue.

Dans mes séminaires, à l'Université Catholique de Milan et à l'Université Grégorienne de Rome, dans environ 10.000 contextes de ma *Concordantia prima* on a classifié à la main toutes les corrélations *A*) grammaticales *B*) élémentaires (c'est-à-dire à l'intérieur de la même proposition élémentaire) *C*) directes (non pas celles qui sont reliées par un intermédiaire) de chaque mot-clé (*ordo, identitas, metaphora, ens rationis, ratio seminalis* ...). On y fait aussi l'inventaire de la proximité entre le mot-clé et le mot correlé: précédant ou suivant, contigu ou distant, à quelle distance en nombre de mots interposés ... Travail bien sûr long et lourd, mais très fructueux pour la connaissance du discours. On s'est aperçu par exemple que les types de corrélations grammaticales sont peu nombreux; que les prépositions ne sont pas des corrélés, mais des corrélants; que probablement cela est vrai aussi pour le verbe *sum* en fonction de copule. Ce travail devrait nous fournir les renseignements nécessaires pour écrire les programmes qui reconnaissent les corrélations à la machine. Si on y par-

venait, ce serait un grand jour pour notre lexicographie. Ces programmes devraient, par exemple, reconnaître que dans *bonum vinum est bibendum* et dans *bonum est faciendum*, le premier *bonum* est adjectif de *vinum*, et le deuxième substantif sujet de *faciendum*.

Je répète qu'on peut faire maintenant avec ordinateur toute recherche sur les niveaux de sens que chaque mot garde en soi toujours et indépendamment du contexte. Mais quand il s'agit de niveaux de signification qui sont ajoutés ou « disambigués » par l'insertion dans une phrase plutôt que dans une autre, c'est-à-dire ceux qui sont variables en fonction de la syntaxe, il faut avant tout savoir quel est l'élément qui pour le même mot caractérise sa valeur (sens ou fonction) *A*) dans le contexte X et sa valeur *B*) dans le contexte Y. Cet élément pourrait être un code à ajouter à la main: tâche qui est généralement facile, mais qui demande un temps très long. Le mieux serait de découvrir dans le texte quels autres mot ou, plus probablement, quelles structures, quelles situations ou quelles répartitions sont liées à la valeur *A*) plutôt qu'à la valeur *B*) avec un appréciable pourcentage de probabilité. Le traitement informatique de co-occurrences séquentielles des mots dans un texte n'est plus un problème, mais celui des corrélations sémantiques est encore un gros problème ouvert.

060 MULLER Considérer les lemmes et pas les formes est le seul moyen d'obtenir des résultats: les formes ont une telle dispersion! Mais est-ce que vous avez porté vos recherches sur tous les mots, ou seulement sur quelques mots privilégiés? Dans votre réjouissant exemple *bonum vinum est bibendum*, il serait intéressant de savoir quels sont les autres mots qui accompagnent souvent « *vinum* » et à quels substantifs l'adjectif « *bonum* » est fréquemment attaché, etc.

061 BUSA Bien sûr, tous, sans aucune exception. Pour l'IT nous avons accompli trois autres types de recherches de

groupement de mots. Ces recherches ont été précédées par celles des combinaisons séquentielles quaternaires, ternaires et binaires, dont j'ai déjà parlé. Deux types de co-occurrences se trouvent dans la *Concordantia Prima*, le troisième dans la *Altera*. Les deux premiers avaient pour but de repérer, avec pas trop d'exceptions, toutes les corrélations sémantiques contigues. Le troisième avait pour but d'essayer un nouveau type de concordance pour les mots d'une fréquence extrème. *A*) Premier type de groupement: celui des syntagmes (sémantiques), c'est-à-dire des mots-clés composés de plus d'un seul mot, *multi-word-lexical-units*. Voir mon Annexe n. 6.1. Ils sont à peu près 600 et leur liste se trouve au début de chaque volume de la *Concordantia Prima*. Le groupement se fait aux niveaux des lemmes: n'importe quelle forme de *corpus*, suivie de *et*, et suivie de n'importe quelle forme de *anima* constitue le syntagme *corpus et anima*. Par conséquent il fallut accepter que dans quelques rares cas les loix rigides du programme incluent un polynôme qui ne soit pas un syntagme sémantique, mais une séquence physique due au hasard: comme, par exemple, sur 194 contextes de *gratia* + *Deus* il y en a 174 qui contiennent le mot-clé *gratia Dei*, mais on y trouve aussi *infundendo gratiam Deus peccata remittit* et *gratia Deo coniungens*. *B*) J'ai appelé « agglutination » le deuxième type. Dans la *Concordantia Prima*, chaque mot-clé a été trié selon les mots qui lui sont « collés »: c'est-à-dire quand une préposition le précède ou quand le mot qui le suit (sans ponctuation lourde interposée) est un de ces mots qui sont contextualisés seulement dans la *Concordantia Altera*. Ici aussi il fallut accepter quelques cas de contiguïté sans corrélation sémantique. C'est un peu compliqué d'expliquer quels mots sont mots-clés dans la *Concordantia Prima* et lesquels dans la *Concordantia Altera*.

Il y a des mots qui sont mots-clés seulement dans la *Concordantia Altera* et jamais dans la *Prima*; par contre certains mots peuvent être mots-clés dans les deux. Leur distribution est précisée dans le *Systema Lemmatum* du volume 9 des *Indices*. Sont mots-clés seulement dans la *Concordantia Altera*:

Tous les mots invariables, c'est-à-dire prépositions, conjonctions, interjiections, adverbes irréductibles (en effect j'ai lemmatisé comme *casus adverbialis*, « inventé » par moi, autant que possible tous les adverbiaux sous un nom ou sous un verbe: *suaviter* est avec *suavis, diligenter* est avec *diligo, gradatim* avec *gradus* ...);

Tous le mots grammaticaux flexionnés, comme les pronoms, et, sauf ses participes, le verbe *sum*, bien qu'il ne soit pas toujours ni auxiliaire ni grammatical;

Tous les mots « spéciaux », comme les numéraux, les abrégés, les translittérés, les symboles, etc. ... Voir mon Annexe n. 1.5. et 3.;

Quelques autres noms et verbes, comme *ratio, res, possum* (sauf ses participes) *dico* etc. qui entrent fréquemment dans le discours connectif.

Tous ces mots sont contextualisés entièrement et seulement dans la *Concordantia Altera*, mais distingués et spécifiés selon la typologie du discours: on peut y voir très clairement quand ils sont dans une citation, ou dans une référence, ou non.

Tous les autres mots flexionnés sont, en principe, mots-clés dans la *Concordantia Prima*, mais ils le sont dans la *Concordantia Altera*, quand ils appartiennent aux phrases enregistrées comme citations littérales ou comme références, leur typologie y étant clairement spécifiée. Les indéclinables y sont compris, ainsi que plusieurs noms propres non latins, et quelques noms communs (*mane, nephas, opus* ...): en effet il fallut tracer une frontière entre mots invariables et mots indéclinables.

Eh bien, les mots-clés de la *Concordantia Prima* chez ST sont 2,5 millions, ceux de la *Concordantia Altera* sont presque 6 millions. Voir mon Annexe n. 3.6. Le verbe *sum* a plus de 700 pages de concordance ... Selon mon Annexe n. 1.3., dans la *Concordantia Altera* 1.527.803 mots-clés sont classés comme citations littérales et 690.287 comme références, parmi lesquels sont compris les mots grammaticaux. Dans le volume 10 des *Indices*, la table 19 montre que les mots de la classe A, qui sont mots-clés

seulement dans la *Concordantia Altera* ont une fréquence complexive de 4.885.525 occurrences. C) Le troisième type de relevés de co-occurrences est celui de la *Concordantia Altera*. Il est par trinômes, c'est-à-dire par combinaisons ternaires en corrélation topographique, groupées par lemmes et formes. Chaque mot-clé a été d'abord accompagné par les deux mots (ou ponctuation lourde) contigus, le précédent et le suivant. On a eu autant de trinômes qu'il y avait de mots-clés. On les a triés par lemmes et formes: 1 mots-clé, 2 mot suivant, 3 mot précédent, 4 codes typologiques, 5 ponctuations interposées. (On s'est aperçu trop tard qu'il aurait été préférable d'inverser 4 et 5). Dans la concordance chaque trinôme qui diffère du précédent dans au moins une de ces cinq zônes, est suivi de sa référence et des références de tous les trinômes qui lui sont identiques. Mon Annexe au n. 6.2. donne les quantités de leur diversité et de leur identité. Presque 3 millions se vérifient une seule fois ... Et c'est à ces résultats aussi que j'aimerais voir appliquer une analyse statistique poussée.

062 BRUNET Pour ce travail il faudrait d'abord établir le nombre des combinaisons possibles, en prenant la situation la plus favorable c.-à-d. en travaillant sur les lemmes et non sur les formes ... Comparer ce qu'on a observé avec ce qui aurait été possible, c'est de la statistique difficile étant donné le grand nombre de combinaisons et le petit nombre d'occurrences de chacune ... Ce que le hasard n'aurait pû faire est que les combinaisons observées soient si fortement structurées, evidemment par les contraintes sémantiques ... Pour voir s'il n'y a pas de mots pôles, il faudrait utiliser la théorie des graphes et plus tellement la statistique ... Dans les groupes binaires qu'on a élaborés à Nancy, on a effacé les mot-outils (par conséquent « le cheval » n'est plus parmi eux): cela pour faire apparaître des contraintes qui ne soient pas syntaxiques mais plutôt des associations thématiques. Cela peut amener à repérer des clichés rares comme ceux des surréalistes ...

Un chercheur canadien, Conrad Bureau, a aussi dépassé le cadre de mot isolé, pour saisir les syntagmes. Il a établi au départ un découpage assez arbitraire des syntagmes, donnant au mot *syntagme* une définition que tout le monde ne saurait souscrire. Son étude porte sur Proust.

Si on confie à l'ordinateur la recherche des co-occurrences, comme on l'a fait au Laboratoire de lexicologie politique de Saint-Cloud, il faut fixer des limites, par exemple cinq pas en arrière et cinq pas en avant. Mais pourquoi pas les limites naturelles de la phrase?

Je me suis attaqué une seule fois à ce problème dans un roman entier de Giraudoux: j'y cherchais s'il y avait une répartition aléatoire des répétitions c.-à-d. la co-occurrence d'un mot avec lui-même à l'intérieur de la même phrase. Le calcul a été extrèmement difficile ... La phrase la plus longue faisait 168 mots ... Je suis arrivé à la conclusion ... qu'il y a certes principes de rhétorique qui font éviter la répétition dans la même phrase, au moins pour le français (en anglais c'est différent) ... mais que, malgré cette contrainte, il y a des raisons thématiques qui commandent la répétition du même mot dans la même phrase ... Il y a aussi des moyens indirects d'aborder le problème des co-occurrences. Le premier consiste à prendre la syntaxe à un niveau morphologique: les morphèmes qui composent le mot individuel forment un espèce de syntaxe intérieure au mot. Préfixes et suffixes permettent de reconnaître des structures plus larges que celles du mot individuel ... En français la chronologie et plus encore le genre littéraire font un usage très différent de la suffixation ... Un deuxième moyen c'est d'examiner certains mots grammaticaux qui sont des charnières sur lesquelles s'appuie la syntaxe: de ces mots on peut extrapoler à la syntaxe ... Une recherche de même type peut s'exercer sur certaines marques morphologiques comme celles du singulier et du pluriel. ... En français le pluriel descend, le singulier monte ... pourquoi?

063 VOICU Y a-t-il des recherches sur les corrélations dans des langues où la séquence des mots est moins libre qu'en latin? En anglais, par exemple ...

064 ZAMPOLLI È abbastanza chiaro che il tipo di lingua (presenza/assenza di un componente morfologico molto importante, dell'ordine delle parole, ecc.) può influire notevolmente sulle strategie di analisi e sulle caratteristiche computazionali dei *parsers*. Uno degli obbiettivi del nostro progetto per un *parser* dell'italiano è quello di verificare i limiti dell'ATN per il trattamento di lingue come l'italiano.

065 VOICU D'après les statistiques du p. Busa il paraît qu'une séquence de trois mots ou plus, a des probabilités de réalisation assez faibles. Faut-il en déduire qu'elle est utilisable pour la définition du style? À propos des co-occurrences, il y a des corrélations de lemmes qui dépassent les frontières de la phrase, car le même lemme peut se trouver dans des phrases « co-ordonnées » successives.

066 MARTINI Delle due maniere di correlazione menzionate da p. Busa, quella semantica e quella fisica, il calcolatore riconosce solo quest'ultima. Se lavorassimo a mano su un campione e arrivassimo a conoscere con che probabilità un nome col suo aggettivo sono in tale sequenza fisica, avremmo già una risposta probabilistica. Mi chiedo se questo parametro potrà servire per le altre parole dell'universo ...

067 MULLER Je proposais une exploitation différente, car ceux qui ont travaillé sur des probabilités de co-occurrences y ont trouvé même les possibilités les plus impossibles pour la langue. Je proposerais que si par ex. on sait que *et ideo* revient 15.000 fois dans le corpus, on cherche si ces occurrences se répartissent aléatoirement parmi les parties du corpus. Cela est un moyen de relever les différences de style, sans autre hypothèse que l'homogénéité du corpus.

068 BUSA Dans mes bandes magnétiques je possède tous les trinômes détaillés, ouvrage par ouvrage, occurrence par occurrence. Il sera alors possible d'en relever les différences entre les parties du corpus.

069 MULLER Je cite un exemple, bien que ce chercheur n'ait fait aucune utilisation de la statistique. Il y a un texte français anonyme du 17me siècle, « Les lettres d'une Religieuse Portugaise ». Un collègue, il y a 10 ans, a isolé un personnage dont on possède la correspondance. Il a mis en évidence l'identité des certains rythmes de la phrase, en particulier les rythmes quaternaires et ternaires des adjectifs, même sans les compter, et sa démonstration est convaincante. Si vous avez les fréquences de *et*, de *ideo*, et de *et ideo*, controlez si elles sont réparties différemment dans le corpus.

070 HAMESSE Vous parliez des probabilités d'apparition du couple adjectif-substantif. On peut dire qu'en latin la place des mots n'est pas indifférente. En voici un exemple: je devais étudier le concept *voluntas* dans certains textes de S. Bonaventure. J'ai travaillé sur un échantillon d'environ 100.000 mots. J'ai remarqué que dans le cas de *voluntas*, l'adjectif suivait toujours le substantif, sauf dans un cas, lorsqu'il s'agissait de l'adjectif *divinus*: l'expression était toujours *divina voluntas*. Je crois qu'il y a là une intention de l'auteur et qu'il ne s'agit ni d'une probabilité, ni d'un hasard.

Il est assez difficile de déterminer les probabilités d'apparition. En effet, dans l'exemple que je viens de donner, à première vue on peut dire qu'il y a autant d'adjectifs qui précèdent le substantif que d'adjectifs qui le suivent. Mais si on étudie les différents cas, on se rend compte qu'en fait c'est toujours le même adjectif qui précède, tandis que tous les autres suivent. Il faut donc être très prudent.

2.2

071 BATAILLON En ST la réponse à la première objection commence rituellement par *Ad primum ergo dicendum quod*, ce qui était le vocabulaire scolaire de la dispute. Mais dans les Commentaires à Aristote les paragraphes commencent par *Deinde cum dicit* sauf dans une section du *In Metaph.* où ils commencent par *Hic dicit.* Il sérait intéressant d'étudier ce changement d'une formule stéréotypée.

072 TOMAELLO Se guardo la tabella delle voci più frequenti in ST, mi si prospetta di esaminare come queste parole, che sono le più semplici, si aggreghino in binomi e trinomi nelle varie parti del corpus.

073 BUSA Bien sûr. Mais il faut tenir compte de la typologie du texte, comme vient de le dire le Père Bataillon. Quand on dit: *ad primum ergo* c'est le trinôme le plus fréquent, c'est une donnée de fait. Mais si on disait qu'il est un signe du style de ST, cela pourrait être une erreur, car il peut être un stéréotype des formules scolaires.

074 MULLER On aurait en somme un raisonnement complémentaire: si on constate que quelque chose est très fréquente, il faut se demander d'abord si cela est particulier à cette époque: dans ce cas, il faut qu'on le retrouve dans d'autres textes. Sinon, cela serait une marque stylistique. C'est exactement comme cela que nous avons procédé pour le texte du Moyen Age dont on a parlé hier: si le texte est d'un auteur que nous connaissons, alors on y devrait trouver une telle chose avec une telle limite de fréquence.

075 BUSA J'insiste encore que le fond de notre question, c'est d'avoir une liste de « telles choses » à chercher comme signes

du style. On a dit plusieurs fois qu'il faut avant tout vérifier la fiabilité du texte. J'ajoute qu'il faudrait aussi essayer d'estimer au moins en pourcentages approximatifs les quantités relatives des hétérogénéités possibles dans un texte, pour établir un seuil au-dessous duquel leur présence ne causerait que des marges négligeable d'erreur.

076 MULLER On doit voir si par ex. l'association de deux lemmes ou la présence du temps parfait sont constantes partout sans beaucoup d'écart, ou si elles son spécifiques de quelque partie, ... si elles ont une stabilité dans une fréquence élevée. Par expérience sur des textes divers, on a toujours plutôt trop que pas assez d'écarts, à tel point qu'on est généralement obligé de se limiter aux écarts plus gros; si on les transforme en écarts réduits, opération simple – et les statisticiens disent que au delà de 2 c'est déjà intéressant –, on a des cas avec 10 ou 20 écarts types; tout bouge, même chez le même auteur ... Je vous assure que *quod, quam, et* ne seront pas répartis aléatoirement dans le corpus.

077 BERNI CANANI Un autre point de vue. Je prends des textes dispersés, par ex. 10.000 résumés de 2/3 pages chacun. J'y calcule les co-occurrences de quelques milliers de lemmes. J'établis des mesures d'association ... On met une flèche entre un point et un autre point ... Les lemmes seront associés par, disons, 50.000 flèches; dans ce réseau je m'attens à une certaine distribution, des îlots ...

078 ZANELLA Se per questa rete si riuscisse a quantizzare la distanza fra le parole si potrebbero applicare le tecniche statistiche della « analisi dei gruppi » (*cluster analysis*) e determinare con metodo più rigoroso gli eventuali raggruppamenti tipici. In generale l'impiego di variabili quantitative agevolerebbe meglio il ricorso ai metodi statistici. Consideriamo, ad esempio, le lunghezze delle frasi. A queste si potrebbero applicare i metodi effi-

cacissimi dell'analisi dei processi stocastici lineari. Si può ipotizzare che, in un certo autore, la lunghezza della frase sia la manifestazione di un processo stocastico debolmente stazionario, con un suo correlogramma e una sua funzione di ripartizione spettrale. Le corrispondenti caratteristiche dello « spettro » potrebbero risultare tipiche per un determinato autore.

079 BRUNET C'est le problème du rythme. Je l'ai étudié dans un texte très rhétorique, l'Emile de J. J. Rousseau, et ailleurs aussi ... Il y a trois hypothèses possibles. Ou c'est le hasard qui préside à la succession des phrases, les courtes et les longues se succédant dans le désordre. Ou bien la succession est ordonnée, si une phrase courte suit généralement une phrase longue: on peut imaginer un petit Poucet qui poserait un petit caillou, un gros caillou, un petit, un gros, selon un rythme très court mais régulier, et qui ferait exprès de changer, de casser le rythme. On peut imaginer enfin une succession des phrases selon un rythme très ample: une phrase longue, encore une longue, puis une moins longue, la phrase se reduisant petit à petit. Quand on examine cela sur 288.240 mots, on observe que, quoi qu'on étudie, les points ou les autres ponctuations, et même en mélangeant tous les signes de ponctuation, on a toujours un rythme long: la phrase longue appelle la phrase longue, la phrase courte appelle la phrase courte ... Le temps est une constante du langage ... Est-ce-que c'est respiratoire? Je ne le sais, mais c'est certainement rhétorique, au moins chez Rousseau. Dans les derniers deux siècles de littérature française la phrase s'allonge-t-elle ou se raccourcit-elle? Il y a une simplification de la ponctuation: le point et virgule s'effacent de plus en plus. Et le point se multiplie. En conséquence la phrase se raccourcit. Comment interpréter cette observation? On rencontre ici comme toujours le mur de l'interprétation. L'ordinateur permet de s'en approcher, jamais de le contourner.

080 PETÖFI I would like to mention another type of word-configurations. In your lemmata lists there are registered about 3500 « verbs » as lemmas with about 2 million occurrences. I could imagine to elaborate first a so-called « valence lexicon » which would contain the single verbs together with the different argument frames assignable to them (where the arguments are represented in the form of morphological categories). Such a lexicon would represent all possible language contexts of the analysed verbs. By means of this lexicon one could analyse the verb-frames manifested in the single sentences, and would provide a new basis for different kinds of comparisons.

081 BUSA Parmi les formes des verbes il est extrèmement important de séparer les formes strictement verbales des formes participiales qui sont à la fois verbales et nominales. Cette distinction est imposée aussi par les formes composées des verbes latins, qu'on peut reconnaître seulement par la syntaxe et la sémantique des phrases: par exemple, *mortui sunt in bello* et *mortui sunt in coemeterio.*

082 VOICU M. Brunet, est-ce que vous avez examiné aussi le rythme des syllabes?

083 BRUNET Il est difficile de travailler sur les syllabes. Au Laboratoire Informatique pour les Sciences de l'Homme (LISH), une équipe a préparé un dictionnaire avec la division en syllabes. Je renvoie là-dessus à Meisonnier et à Nina Catach. Les syllabes sont des unités de segmentation qu'on est tenté d'utiliser pour apprécier le rythme du mot ou celui de la phrase. Mais on peut parvenir aux mêmes fins en étudiant des unités plus petites, plus sûres et plus faciles à délimiter: les lettres. Ainsi ai-je fait dans une étude sur la longueur du mots et la longueur des phrases. Et j'ai pu observer que depuis 1789 le mot français s'allonge tandis que la phrase se raccourcit.

084 TOMBEUR À la suite de ce que M. Brunet a dit, je veux souligner que nous devons être bien convaincus qu'il faut multiplier les expériences. Ce qui ressort de tout ce qui se dit autour de cette table, c'est qu'il y a un sentiment d'insécurité parce que nous ne sommes nulle part, et ce malgré les travaux extrêmement importants qui ont été faits. Que l'on ne me comprenne pas mal: c'est indirectement un hommage à ces travaux dus notamment à certaines personnes qui sont ici présentes. Mais il ne faudrait pas partir de là dans un sentiment de sécurité, en se fiant là-dessus sans se reposer des questions. Je crois que nous n'avons pas l'outil statistique pour l'étude de la langue. Nous cherchons avec tous les moyens disponibles et il faut continuer de chercher; c'est dans cet esprit-là, conscients de cela, ayant donc une dose d'inquiétude suffisante, que nous ferons du bon travail. Un de mes collaborateurs, Michel Guéret, m'a cité une bonne phrase à ce propos: il faut être suffisamment inquiets tout en étant dans la quiétude. Je crois que travailler cela signifie n'être pas sûr de toute une série de choses, mais poser la question. Saint Augustin dit cela très bien dans son *de Trinitate: Sic quaeramus tamquam inuenturi et sic inueniamus tamquam quaesituri.*

Je voudrais relever une autre phrase de M. Brunet quand il a dit avec un sentiment presque de gêne: vous allez encore me dire que ce sont des évidences. Je dirais dans la même ligne: il n'y a pas d'évidences; je pense qu'il faudrait presque bannir ce mot de notre vocabulaire en ce qui concerne la recherche statistique. Il faut faire l'expérience pour le voir, nous ne le savons pas.

Je voudrais ensuite faire encore une autre remarque. Le Père Busa remarquait la difficulté de distinguer les participes. Je crois qu'il faut dire que si on veut distinguer les participes – en tout cas pour une langue comme celle de saint Thomas –, si on veut entrer dans les distinctions comme *esse* verbe auxiliaire ou verbe plein, on est *dans une sémantique*: on a dépassé toute idée de lemmatisation formelle, morphologique et syntaxique et l'on est dans une lemmatisation sémantique. C'est donc là une

lemmatisation extrêmement discutable, ce qui ne veut pas dire qu'il ne faut pas faire une telle lemmatisation. Ma phrase sur l'inquiétude et sur la quiétude devrait nous y amener. Mais cela veut dire que nous en sommes là à des distinctions très fines, et par conséquent qu'il va y avoir discussion, discussion souvent d'ailleurs avec soi-même. Pour des cas difficiles, on peut en effet penser parfois le lendemain le contraire de la veille.

Enfin quelques petites remarques en ce qui concerne le problème de la ponctuation. Pour tous ces problèmes de ponctuation, il faut toujours se poser la question de l'origine de cette ponctuation. Il y a aura donc une différence de situation radicale à ce sujet entre les textes à transmission directe et ceux à transmission indirecte. Il y a un point sur lequel je me permettrais d'attirer votre attention: quand on va jusqu'à distinguer des ponctuations faibles et dès lors jusqu'à faire des calculs extrêmement détaillés, comme l'a fait Frautschi aux Etats-Unis, par exemple, pour savoir le nombre d'éléments qu'il y a entre chaque élément de ponctuation, se poset-on la question: qui a mis cette ponctuation? Même quand on est au 17e siècle (voir le cas de Descartes soulevé par Crapulli à un des colloques du Lessico Intellettuale Europeo), au 18e siècle, ou au 19e siècle, voire au 20e. Celui qui étude les discours politiques, par exemple, – songez à l'étude des discours de la campagne électorale Giscard d'Estaing-Mitterand faite dans *54.774 mots pour convaincre* – se pose-t-il de telles questions? Ces discours sont publiés et son donc mis en page, avec toute une ponctuation: je suis persuadé que l'on va faire ou que l'on a fait des études sur le nombre d'éléments entre chaque point ou même entre des éléments de ponctuation inférieurs au point.

Pour le latin je voudrais faire part de quelques expériences que nous avons faites au Cetedoc. Pour le domaine du latin antique et médiéval, ce n'est qu'en partant de *1 manuscrit* – il y a autant d'éditions que de manuscrits: un manuscrit est une édition –, en notant la ponctuation de ce manuscrit, que l'on peut observer quelque chose qui risque de correspondre à une réalité, c'est-à-dire une réalité qui n'est pas celle d'un éditeur moderne,

d'un typographe peut-être (voir le problème de Descartes évoqué il y a un instant), qui n'est même pas celle de l'auteur, mais celle du copiste. Cela donne des choses intéressantes parce que l'on a une base assez fiable. La pratique que l'on peut mettre en oeuvre pour l'étude du rythme est une technique tout à fait simple, même élémentaire, qui manifeste le rythme du discours: celle des moyennes mobiles. On prend le nombre d'éléments qu'il y a entre deux éléments de ponctuation forte et on tient compte de ce qui précède et de ce qui suit pour établir une moyenne mobile. Cette technique fait apparaître le moment où le discours se trouve changé. Dans une application faite à un corpus de textes émanant de l'abbaye de Saint-Trond au XIIe siècle, directement à partir de l'examen des manuscrits, il apparaissait clairement que c'était toujours l'introduction dans le discours d'éléments extérieurs qui amenait une rupture de rythme. Ces éléments étaient des citations patristiques en masse ou l'insertion d'un texte de charte, par exemple. Le rythme était coupé par ces éléments: ce n'était plus l'auteur lui-même qui s'exprimait et cela apparaissait clairement avec cette technique des moyennes mobiles.

Par contre – autre expérience – pour les Pères de l'Eglise latins, dont nous traitons systématiquement les oeuvres par ordinateur, nous sommes souvent obligés de recourir à des éditions parfois bien anciennes, spécialement celles reproduites dans la Patrologie latine. Pour les textes de la Patrologie latine, j'oserais affirmer qu'au niveau de la ponctuation, vous ne pouvez rien faire. Nous avons dans ce sens même été puristes lors des enregistrements de textes: ayant peur des mauvaises interprétations possibles, en tout cas quand nous avons faits nos encodages de textes pour le moyen âge, nous n'avons pas repris la ponctuation faible. Mais qu'on reprenne ou non cette ponctuation dans un fichier d'ordinateur, on ne peut absolument rien baser sur cette ponctuation pour l'étude du rythme d'une oeuvre.

De là découle une troisième expérience que nous avons faite, toujours, sur le latin, et nous l'avons exposée lors du congrès parisien célébrant l'année 1274 (l'année de la mort de Thomas et

de Bonaventure ainsi que celle du second concile de Lyon) dans une communication signée par Jacqueline Hamesse, André Stainier et moi-même. Nous avions comparé plusieurs textes de Thomas et de Bonaventure appartenant à des genres littéraires différents. La solution pour le latin réside dans l'étude des « mots ponctuants ». Il faut donc abandonner l'idée de se baser sur la ponctuation éditée. Le latin est une langue très structurée qui possède des mots signes de ponctuation. Nous avions ainsi conçu l'idée d'élaborer un programme de ponctuation automatique. En appliquant un tel programme on constaterait quelles sont les différences entre l'éditeur « x » et l'ordinateur ponctuant le texte. Dans un discours tel que le latin scolastique, la recherche sera d'ailleurs plus facile que dans d'autres textes où l'on ne retrouve pas une argumentation du même type.

Dernière remarque en ce qui concerne les co-occurrences: faut-il s'arrêter ou non à la phrase? Je dépasserais ici l'application au latin et je dirais que cela me paraît être une question ouverte, comme d'ailleurs, j'en ai bien l'impression, la grande majorité des question en statistique linguistique. J'hésite; j'aurais tendance à dire non: il ne faut pas dépasser le contexte de la phrase, mais, sans doute, dans la ligne de ce qui a été souligné tout à l'heure, faut-il refuser les soi-disant évidences, multiplier les expériences et faire les doubles applications. En tout cas, c'est un point qu'il importe d'approfondir. Tenir compte de la phrase, et aller jusqu'à « n » mots avant et « n » mots après le signe de fin de phrase (sans doute la valeur de « n » pourrait-elle se situer entre 10 et 20). Pour la valeur de ce « n », je serais d'ailleurs heureux de connaître vos réactions. De toute façon il faut mener cette expérience d'application de pair avec les autres applications possibles, notamment en tenant compte de la phrase, pour le cas où cette phrase a une signification de base.

085 Bartoletti Ho già sperimentato quanto il Martini sollecitava, sulle *Novellae Constitutiones* di Giustiniano, di circa 500.000 parole, di cui il 60 % in greco e il 40 % in latino. Si

tratta di testi in gran parte bilingui. Ho pubblicato il vocabolario del latino e preparo quello greco. Ovviamente ho rilevato tutte le corrispondenze fraseologiche che mi sono parse interessanti.

086 MARTINI Una doppia versione offre certo una verifica più robusta. Ma individuare correlazioni semantiche con una analisi puramente quantitativa dei grafemi sarebbe impossibile se non ci si potesse basare su qualche regolarità di collocazione ... Si torna al problema di trovare regole di corrispondenza tra correlazione fisica e correlazione semantica ... Se trovassi che l'80 % delle contiguità tra due parole sono semantiche e il 20 % no, avrei già una legge utile ...

087 FARAHIAN Comme le rythme de la phrase chez le même auteur change selon les genres littéraires de ses ouvrages, quelle est la constante qui permettrait de reconnaître le même auteur malgré des rythmes qui changent?

088 MULLER Selon mon expérience, il n'y a rien dans la langue d'un auteur qui corresponde aux empreintes digitales, qui ne changent pas de la naissance à la mort.

089 BRUNET À l'intérieur d'un même auteur il arrive qu'on trouve des écarts plus grands que ceux qu'on trouve entre les époques ou les genres littéraires. Notre esprit est plus puissant que l'ordinateur: il est capable de reconnaître si une page est de Chateaubriand ou de Proust avec peu de chance de se tromper ... L'ordinateur va se tromper presque toujours si on veut à toute force l'obliger à répondre à cette question. Est-ce à dire que le cerveau dispose d'autres armes que l'ordinateur? Pas nécessairement, ce sont peut-être les mêmes moyens, mais plus puissants. Notre sentiment linguistique qui nous fait reconnaître une page de Proust n'est probablement qu'un sentiment statistique. Et sans le savoir ou sans le dire, nous faisons tous les jours de la « statistique globale ».

090 BETTI Riprendo l'argomento con cui proponevo di superare i metodi puramente statistici a favore della ricerca della struttura sottostante il testo condotta con metodi locali. Un esperimento in corso sull'Enciclopedia Einaudi mi ha portato a rappresentare questo testo sotto forma di rete (in questo caso particolare l'individuazione dei nodi della rete è notevolmente semplificata, giacché questi sono in maniera naturale i circa 600 articoli dell'Enciclopedia; qui tuttavia non intendo fornire un algoritmo operativo, ma semplicemente fare una proposta, indicare una direzione). Le connessioni fra nodi sono da interpretare come « rimandi concettuali » da un articolo a un altro, valutati secondo una funzione che tiene conto di alcuni parametri e di possibili autocorrezioni. Come si capisce la messa a punto di questa funzione di valutazione costituisce uno degli argomenti più delicati, giacché valutazioni diverse possono dar luogo a modelli sensibilmente diversi e vale la pena di osservare che le frequenze dei termini danno in questo caso un parametro per la valutazione, ma non l'unico.

La rete, per sua natura, è un oggetto topologico, cioè interessa per le proprietà che sono invarianti rispetto a deformazioni che non tagliano i collegamenti e non alterano i nodi. Da questa rete si possono costruire degli « intorni » di ogni nodo: intendo « intorno » non in senso tecnico, bensì come « zona », « ambiente » in cui si riassume una problematica; un termine più pertinente è quello di « copertura » tematica dell'argomento rappresentato dal nodo stesso. Ad esempio mi aspetto che il nodo *vita* dell'Enciclopedia Einaudi venga ricoperto da *organismo, evoluzione, eredità*, per quello che riguarda i problemi strettamente biologici, ma anche da *angoscia, desiderio, uomo, morte*, se il punto di vista è quello dell'esistenza, o anche da *sistema, comunicazione, organizzazione, intelligenza artificiale*, se il discorso riguarda soltanto il livello cibernetico e il suo grado di plausibilità. Non esiste una sola « copertura » di un argomento. Il problema concreto, oltre la definizione delle funzioni che permettono di rappresentare il testo nella rete, è allora quello di determinare le

possibili coperture tematiche. Questi problemi dipendono dal testo che si ha a disposizione, per la loro soluzione. Quello che rimane fissato è il principio per cui la struttura globale del testo è il prodotto dell'incollamento di strutture definite localmente e che, localmente, presentano un ampio spettro di variabilità.

091 CRESPI REGHIZZI L'analyse des séries temporelles est un traitement mathématique simple et elégant. Il offre des techniques et des algorythmes très performants pour trouver l'harmonique première, la deuxième etc. On pourrait appliquer la même technique de calcul au rythme respiratoire ou rhétorique des phrases ou des mots abstraits; par là on pourrait identifier les caractères idiomatiques d'un auteur. En effet en musique il est bien possible de reconnaître une sonate baroque par le relevé de son rythme.

092 ZANELLA Il dr. Betti, nell'estendere le nozioni proprie della topologia all'analisi di un testo, come valuta l'appartenenza di una parola ad un « intorno » di un'altra?

093 BETTI Intendo il termine topologia in un senso più generale del consueto. Bisogna pensare che i nodi della rete non siano i « punti » dello spazio, ma delle sue parti (ad esempio gli aperti). In tale caso la « metrica » del testo, o altre funzioni di correlazione fra i termini che entrano nella costruzione della rete, sono solo delle componenti, che non individuano completamente la topologia.

094 TOMBEUR Si j'ai bien compris ce qui a été dit, on pourrait baser l'étude du rythme sur autre chose que la ponctuation, ainsi par exemple sur l'alternance des mots abstraits ou des mots concrets. Qu'est ce qu'un mot abstrait? Qu'est-ce qu'un mots concret? Je me souviens d'une intervention de M. Gorcy, le sous-directeur du Trésor de la Langue Française, lors du colloque de Pise organisé par la Fondation Européenne de la Science, le mois

dernier: il nous disait être dans l'impossibilité de faire d'une manière claire la distinction entre mots abstraits et mots concrets. On entrerait donc dans l'arbitraire; or précisément, tout notre appel aux techniques mathématiques correspond à une volonté d'échapper à l'arbitraire.

A titre d'exemple et pour montrer que les choses sont parfois très compliquées, même sans entrer dans la distinction abstrait-concret, je puis mettre sur la table une expérience qui est précisément relative à des problèmes d'authenticité. Il s'agit d'une étude d'un corpus de sonnets portugais du 16e siècle pour lequel on a noté le rythme de chaque vers, en comptant le nombre de pieds, ce qui paraît ne pas poser de problèmes. On se dit que pour un comptage aussi simple, il y aura une entente entre les chercheurs.

Lors du test d'analyse factorielle qui avait été fait en se basant sur cet élément de comptage (le nombre de pieds par vers), il y avait un certain nombre de sonnets qui se comportaient d'une manière différente des autres, qui correspondaient à une typologie différente. Dans la discussion qui a eu lieu lors d'une table ronde organisé pour discuter de ces problèmes, on s'est aperçu que *tous* les cas considérés par celui qui avait fait l'analyse du rythme comme étant un rythme anormal, étaient contestés par un autre chercheur: le premier chercheur se basait sur la graphie, le second sur la prononciation (ainsi, par exemple, *spirito* - *sp(i)rito*). Vous voyez combien il est difficile de sortir de l'arbitraire alors que par la statistique nous voulons échapper à l'arbitraire.

095 Berni Canani Une bonne métrique n'est pas obligatoire, car il y a des chemins parallèles: par ex. en conservant le point de vue topologique de m. Betti on peut chercher des recouvrements, par ex. en italien dans les dictionnaires le mot « abile » est expliqué par d'autres mots comme « idoneo », « capace » etc.: ils n'expliqueront pas tout bien sûr, mais on pourra quand-même les grouper en régions conceptuelles.