3.1

096 BUSA Je répète que notre thème est: si quelqu'un voulait (pas moi, car je suis trop âgé) attaquer la statistique globale de tout dans tout ST, que devrait-il faire, pas après pas? On a déjà posé des questions: comment couper le texte en tranches? Et quels mots y sont à examiner? On a déjà groupé les ouvrages de ST en 3 parties: ceux dont le discours de base est à lui, ceux où son discours se déroule en commentant le discours d'un autre, et ceux qui ont été enregistrés par des auditeurs. Mais dans ses propres ouvrages il y a différents types de citations et des différences de genre littéraire. En plus, la ponctuation a été introduite et changée au cours des siècles et le père Bataillon nous a alertés en nous informant que, dans les éditions des Commentaires à Aristote pendant la Renaissance, même le texte d'Aristote traduit en latin par Guillaume de Moerbéke, et employé par ST a été remplacé par celui qui avait été traduit par Leonardo Bruni, où *democratia* est devenu *status popularis* et *politia* est devenu *res-publica* ... Tout cela est donc déjà établi: sous peine d'aboutir à des conclusions invalides, il faut que chaque tranche du texte soit homogène.

Dans mon Annexe n. 1.2. vous avez les quantités des mots des ouvrages de ST: 28 unités ont plus de 100.000 et moins de 600.000 mots, mais 9 unités ont moins de 1.000 mots. Il faudra donc examiner dans quelle mesure les dimensions de l'ouvrage affectent les fréquences relatives des mots. Par là on arrive à s'interroger sur les dimensions optimales des tranches. Cela amène à se demander s'il est préférable d'élaborer les tranches naturelles du texte, bien que de grandeur variable ou au contraire les tranches brutes de quantité fixe de mots, par exemple les premiers 10.000, les 10.000 suivants, etc. ... Ou encore est-il préférable d'élaborer les 2 types de tranches pour en comparer les résultats?

097 NICOLODI I temi della semiologia, lessicologia ed ermeneutica van tenuti presenti nel portare avanti la ricerca aperta dall'IT. Se vi è distinzione tra il segnaletico e il semico, anche l'attenzione alla segnaletica ha una sua propria funzione. Si sta parlando in questi giorni della luminosità dei dipinti del Masaccio, come di uno stile senza stile, intuizione di creatività pura. È un luogo comune che Platone sia il più poeta dei filosofi e il più filosofo dei poeti. Però in lui l'anima trascende l'espressione come nei miti. In ST al contrario l'anima entifica e articola l'espressione come luminosità che parte dall'interno. Perciò la nostra ricerca di unità linguistiche punta non tanto verso una unità enciclopedica, quanto verso una unità intensiva, dominatrice creativa delle molteplicità che la esprimono.

098 BOLOGNESI Des traductions il y en a toute une vaste gamme, les unes comme des gloses interlinéaires continues, les autres libres. Les dernières ont une valeur plutôt artistique et littéraire, les premières ont plus de valeur linguistique. Il faut encourager les comparaisons entre traduction et original: on y trouvera des faits intéressants comme calques lexicaux ou sémantiques, néologismes etc. ... Dans des traductions anciennes de textes grecs on a trouvé de quoi confirmer les conjectures de la leçon primitive dans des endroits qui avaient été gâtés par la tradition manuscrite.

099 TOMBEUR En ce qui concerne les tranches, je crois que la méthode probablement la plus fiable, celle par laquelle je commencerais une recherche de ce type, est le découpage par tranches de « x » occurrences. Je m'en réfère à une expérience que nous avons faite, en dehors du domaine médiéval, pour l'oeuvre d'une romancière hollandaise contemporaine, c'est-à-dire pour une oeuvre éditée où la division en tranches (chapitres, livres, etc.) est faite par l'auteur lui-même. Mais nous avons constaté qu'à certains moments il y a des phénomènes que l'on risque d'interpréter très mal, parce que précisément, dans l'oeuvre, il y a

quelque chose qui se produit quand on passe d'une tranche de division du texte à une autre. Des erreurs d'interprétation sont possibles quand il y a une superposition des tranches de division de l'oeuvre sur les tranches établies par « x » occurrences.

Dans le domaine du moyen âge et de l'antiquité, les grandes divisions sont souvent bien établies. Dans ce cas il faut faire l'expérience à l'intérieur des grandes divisions et prendre à l'intérieur de ces divisions des tranches de « x » occurrences. Notre règle générale devrait d'ailleurs être la multiplication des expériences. Cela devrait être notre règle de conduite.

100 BATAILLON Quand ST consacre une question par ex. à une vertu particulière, des mots y sont concentrés, comme par ex. *synesis* ou *gnome*, qui sont absents ailleurs. Il y a dans la série de ses oeuvres des mots qui apparaissent comme « nouveaux » à un certain point. Par ex. après avoir connu une nouvelle version d'Aristote, non seulement il en cite des morceaux, mais dès lors il enrichit son vocabulaire courant avec des mots qu'avant il n'employait pas.

101 TOMBEUR C'est vrai ce que vous dites. Mais ici il s'agit toujours de grands nombres. Le philologue, le philosophe, le théologien est fort frappé, – et il a raison (il doit être très attentif au moment où les phénomènes se présentent) –, quand il y a 1 mot, 2 mots, 3 mots nouveaux qui apparaissent. Mais statistiquement le poids de ce type de phénomène ne va pas être reflété dans un type d'étude comme celui que je donnais en exemple, l'étude du taux d'accroissement du vocabulaire.

D'autre part – et c'est une deuxième chose que je voudrais faire remarquer –, vous dites à un moment donné: l'auteur utilise une citation, il la commente et à ce moment un mot nouveau apparaît. Si l'on est dans un genre littéraire donné, ne croyez-vous pas que cette même situation va se reproduire et que par conséquent, statistiquement, on aura des résultats valables?

Ce phénomène d'apparition d'un mot à tel endroit parce

qu'à un moment un auteur commente Macrobe, par exemple, va se reproduire quand il y aura une autre citation où un autre nouveau mot apparaît qui sera lui aussi repris dans le commentaire qui suit.

Mais là je dirais que cela n'est pas un problème de statistique: c'est simplement un problème d'avoir un texte enregistré avec des références maximales qui puissent précisément vous dire immédiatement quand vous consultez le vocabulaire de ce texte, que tel mot se trouve uniquement dans telle partie de l'oeuvre et que c'est précisément une citation de Macrobe. Ce n'est pas un problème de statistique.

102 Busa Dans mon Annexe au n. 5.4. vous trouvez le résumé des tables 29 et 30, aux pages 987-1202 du volume 10 des *Indices*. Elles montrent dans quel groupe d'ouvrages chaque mot *non-hapax* apparaît pour la première fois. Aux pages 941-986 les tables 27 et 28 font de même pour les *hapax*. Mes bandes magnétiques permettraient de pousser ce relevé jusqu'aux ouvrages individuels à l'intérieur de chaque groupe. Mais bien sûr on n'y dit rien sur l'interprétation du fait brut.

103 Bolognesi Une partie de la terminologie philosophique de l'allemand moderne a été créé par Notker dans les traductions de textes philosophiques latins comme ceux de Boèce.

104 Bataillon Pour le français, des mots son nés avec la traduction de la Politique par Nicole Oresme.

105 Bolognesi Une grande partie de la terminologie grammaticale arménienne d'aujourd'hui a été créé par la traduction en arménien de la grammaire de Denis de Thrace.

106 Grilli Dans le lexique de Cicéron il y a la traduction de mots grecs dont on a perdu l'original: on y peut pas construire grande chose. C'est le contraire quand Cicéron traduit

poiotes par *qualitas*, mot qui après a eu une présence immense dans toutes les langues romanes.

107 VOICU L'enrichissement du vocabulaire peut-être ne modifie-t-il ni les structures de la langue ni celles du style. Au fond, il s'agit d'une sorte de citations.

De l'avis de m. Muller, une analyse statistique globale peut être menée à bien avant même de définir le genre littéraire, la chronologie ou les citations d'un texte. On y trouvera des « bosses », des « creux », des « pics » qui contredisent une distribution régulière. Il y aurait alors trois étapes à franchir: voir tout simplement ce qui se passe dans le texte; y reconnaître les anomalies; les justifier.

108 BRUNET Je suis d'accord avec cette approche candide et innocente. On risque de piétiner très longtemps si l'on cisèle trop d'hypothèses au départ. Les premiers résultats alerteront sur les hypothèses à ajouter. Les données du p. Busa ne sont pas éloignées d'une exploitation statistique innocente. L'IT est plutôt un dictionnaire de fréquences. Pourquoi ne pas partir tout simplement des 179 sous-fréquences des 20.000 lemmes, représentés dans le vol. 1 des *Indices*? Les fréquences seront transformées en écarts réduits et projetées sur un graphique. Si vous aviez mille courbes pour les milles mots les plus fréquents, vous auriez assez d'hypothèses pour vos recherches. Quelles recherches? On a opposé les mots individuels aux mots en chaîne, syntagmatique ou paradigmatique. Il y a beaucoup de principes de groupement et on peut en tirer des résultats plus intéressants qu'avec les mots individuels. Je crois qu'avec les groupements par ex. par catégories grammaticales, vous auriez même des surprises autant que des « évidences ». ... Par ex. dans mon corpus j'ai rencontré des situations bizarres avec « action » et « acte »: parmi les textes il y avait « L'action » de M. Blondel. Vous, p. Busa, vous avez de très beaux matériaux statistiques.

À propos des mots nouveaux je me demande si le langage

des continuateurs de ST est le même que celui de ST ... Il y aura des mots que ST n'a jamais employés et de mots que lui seul a employés ...

109 BUSA Mon Annexe au n. 5.3. informe que dans les textes des autres auteurs on a trouvé 18.306 formes et 2.261 lemmes de la classe A, c'est-à-dire mots communs, qu'on ne trouve pas chez ST. Les tables 27-30 de *Indices* spécifient aussi quels mots de ST n'apparaissent pas chez les autres auteurs.

110 PETÖFI Special attention should be given to words which appear in titles of texts. There will be, of course, an asymmetry in the analysis, if not all texts have a title. In such a case you have to try to analyse first all existing titles, then, in analogy to these, attach a title to those texts, too, which have no title. One should analyse now the words appearing in the titles as to their occurrence in the texts of both texts classes (i.e. the class of texts having a title and the class of texts having been attached a title), and compare the results obtained with respect to both text classes from different angles. In addition, also the associative nets are to be investigated, which have been mentioned by dr. Betti.

111 BERNI CANANI Dans l'analyse des sous-ensembles signalée par m. Brunet, il y aura des précautions à prendre sur les dimensions de chaque boîte où les fréquences sont casées. Car l'IT signale qu'elles ont des grandeurs très différentes. ... Quand on aura à notre disposition des milliers de courbes comme dans un cardiogramme géant, on pourra appliquer les techniques de classification mentionnées par m. Zanella. Par le principe des relations d'équivalence on pourra identifier des mots qui ont des rôles semblables ...

112 ZANELLA Sembra ragionevole pensare ad una ponderazione che tenga conto del fatto che la percentuale riscontrata

per un certo lemma è in un'opera di 5.000 parole piuttosto che di 100.000. Un 10 % in un'opera che « pesi » l'1 % dovrà evidentemente valutarsi in modo diverso del 10 % di una che « pesi » il 20 % (questo accade, in effetti, automaticamente nell'impiego dei consueti *test* statistici). C'è però il problema della « urna » di riferimento. L'urna da cui è estratto un certo vocabolo potrebbe essere tutta la letteratura del periodo per l'argomento esaminato: se in essa il suo peso è del 25 % mentre in un'opera specifica il peso è del 15 % si potrebbe vedere in questo un carattere personale. Ma come ottenere l'urna?

113 Berni Canani Il ne faut pas sous-entendre l'indépendance des mots: un mot dans un texte n'est jamais indépendant des autres.

114 Zanella Dopo aver esaminato parola per parola, non si è escluso di esaminarle tutte assieme.

115 Zampolli Ch. Muller ha scritto varie volte a questo proposito.

116 Muller Oui. À présent il s'agit de la proposition de m. Brunet d'examiner les fréquences des mots individuels dans les 179 sous-ensembles. Comme certains sous-ensembles font moins d'un millième du corpus, il faudra établir jusqu'à quelle fréquence on descend. Je pense qu'il faudrait d'abord diviser l'ensemble en une dizaine de parties pas trop inégales, dont la plus petite soit au moins 5 % du tout. Pour cela on regrouperait, raisonnablement selon genre et date, plusiurs textes dans une même partie. Après, on détaillera, selon plusieurs tranches ou textes, chaque partie. Je ne passerais pas d'emblée de l'ensemble aux 179 ouvrages. Disposant du cadre complet de tous les mots, on commencera par les plus fréquents et on descendra jusqu'aux *hapax*. Mais au-dessous des fréquences moyennes on sera gêné par les ouvrages courts. Pour les *hapax* on pourra calculer seulement la distribution de leur catégorie.

117 BRUNET Je ne puis qu'approuver les remarques de m. Muller. Il faut proceder à des groupements de mots, mais aussi de textes, afin d'atteindre des effectifs suffisants. Mais dans le cas d'un grand corpus comme celui du p. Busa on peut aller assez loin dans le détail des mots et des textes, sans avoir peur de solliciter l'ordinateur. Ainsi pour Giraudoux (700.000 occurrences) tous les mots significatifs ont été calculés en quelques secondes par la loi normale. Bien sûr c'était un IBM d'énorme puissance ... La loi normale n'est pas parfaite, elle est un raccourci, une approximation. La vraie loi dans les données discontinues et avec tirage exhaustif, est une loi hypergéométrique, mais cent fois plus coûteuse. J'ai comparé les deux lois dans mon énorme corpus: dans les grandes nombres elles se rapprochent et se confondent. Dans les millions des mots il ne vaut pas le peine d'employer la loi fine. La différence n'apparût qu'à la troisième décimale.

118 MULLER Quand dans une partie on apprend quels vocables ont une fréquence excédentaire et lesquels ont une occurrence déficitaire, on connaît déjà assez.

119 ZAMPOLLI La mia esperienza conferma quanto Brunet e Muller hanno appena detto. Ricordo anche quanto detto da J. De Kock, di Lovanio, nel corso della Tavola Rotonda recentemente organizzata presso l'Università Cattolica sul tema della statistica linguistica. Mi sembra che la organizzazione dei riferimenti nell'IT permetta lo studio della variazione delle frequenze lessicali non solo tra opera e opera, ma tra sottoinsiemi di varia natura (gruppi di opere; parti omogenee di opere diverse; ecc.).

120 VOICU M. Tombeur, avez-vous des chiffres sur la corrélation entre l'enrichissement du vocabulaire et les dimensions du texte?

121 TOMBEUR Il est très difficile de répondre à cela de manière générale. Je peux vous montrer ici des données tout à

fait particulières à une oeuvre. Il est clair que cela part de 1 et que par conséquent cela augmente très vite, et que très rapidement, à partir d'un certain nombre d'occurrences, l'augmentation est très faible. Faut-il rappeler ici ce que déjà Pierre Guiraud avait mis en lumière et que Zampolli entre autres a rappelé dans le *Lessico di frequenza della lingua italiana contemporanea?* Dans *Les caractères statistiques du vocabulaire*, qui est une oeuvre remarquable, quand on pense aussi à la date de publication, Pierre Guiraud notait que les cent premiers mots représentait déjà un pourcentage extrêmement important de n'importe quel texte, 60 %, et que très vite on arrive aux mille premiers mots de n'importe quel texte qui représentent un pourcentage de 85 %. Par conséquent, ce que Guiraud avait déjà mis en lumière, est déjà une réponse à votre question. Très vite il y a une certaine stagnation puisque, selon Guiraud, au-delà des 4000 premiers mots, un très grand nombre de mots (de 40 à 50.000) ne représentent plus qu'un pourcentage infime de n'importe quel texte (2,5 %). C'est un décalage vraiment énorme. Concrètement, il faut voir cela pour chaque oeuvre et on ne peut pas passer d'un roman contemporain néerlandais à saint Thomas, ni étudier saint Thomas sans tenir compte des genres littéraires en question; quand il s'agit par exemple du genre littéraire de la *quaestio disputata*, il y a toute une forme à laquelle l'auteur obéit, et par conséquent les répétitions de mots, les accroissements du vocabulaire sont dépendants de la nature du genre. Il faut travailler genre par genre. Il n'y a pas de réponse générale. Il faut regarder cas par cas.

J'ai ici simplement une partie des résultats et je ne voudrais pas retenir l'attention avec ces cas. Je peux vous communiquer ces résultats. Remarquez d'abord que cette oeuvre a été lemmatisée et qu'il faut tenir compte du type de lemmatisation retenue, car qu'est-ce que la lemmatisation? Quand le chercheur a-t-il considéré qu'il s'agit d'un mot différent? Qu'a-t-il fait pour les participes? Tout cela est un problème qu'il faut examiner. Je me permets d'y insister parce que c'est le grand danger: il n'y a pas deux personnes ici autour de la table qui pratiquent des ana-

lyses identiques. Quand nous faisons des comparaisons, quand nous disons mot nouveau, vocabulaire nouveau, lemme nouveau, nous avons en fait tous un langage différent. C'est d'ailleurs pour cela que je propose que dorénavant on cesse de parler de lemme et que chaque fois l'on ait la discipline d'utiliser un adjectif qui complète le plus justement possible ce mot lemme pour que l'on sache de quoi il s'agit, de quel type de regroupement.

Dans cet exemple néerlandais – ce n'est pas moi qui ai fait la lemmatisation: c'est une thèse de doctorat défendue à l'Université de Louvain par Marc Geerinck –, si j'ai bon souvenir, c'est une lemmatisation de type morphologico-syntaxique, c'est-à-dire que le chercheur a fait, je pense, la distinction, par exemple, entre le participe verbe et le participe adjectivé. Alors si vous voulez les chiffres: si cela n'ennuie pas trop les autres, je vais vous en donner quelques uns. Dans un des romans, pour les mille premières occurrences, il y a 480 lemmes différents; quand on passe de 1001 à 2000, il y a 525 lemmes différents dont 395 nouveaux; quand on passe de 2001 à 3000, il y a 479 lemmes différents dont 282 nouveaux par rapport aux deux tranches précédentes; de 3001 à 4000, 443 lemmes différents dont 216 nouveaux. Et cela continue en diminuant progressivement. Ainsi entre 34.001 et 35.000, il y a 485 lemmes différents dont 86 sont encore des lemmes nouveaux. Il faut noter que l'oeuvre en question est remarquable au point de vue stylistique, que l'auteur dispose d'un vocabulaire assez étonnant. Cela se reflète dans les chiffres, puisqu'il y a 6692 lemmes différents pour 39.122 occurrences. L'oeuvre citée est d'ailleurs la plus riche du corpus étudié.

122 Busa Dans mon Annexe au n. 5.2. et 5.4. on trouve la progression des « mots nouveaux » dans les groupes des ouvrages. Je précise que, pour nous, les mots sont *hapax*, quand ils sont uniques dans l'oeuvre entière, et non dans une seule ouvrage.

123 Brunet Dans la dernière tranche de mon corpus, des 1945 a 1964, on a trouvé 3798 mots nouveaux: plus que dans

la tranche précedente. Là il y a à considerer l'hétérogénéité des sujets. Il n'y a pas de limites dans le vocabulaire. Les mots naissent tous les jours: ils naissent plus vite qu'ils ne meurent. Il y a là une inflation lexicale qu'on peut appeler aussi créativité lexicale.

124 TOMBEUR Je crois que la situation décrite par M. Brunet est tout à fait différente de celle que je pourrais présenter pour l'antiquité, la patristique ou le moyen âge. Vous êtes dans une situation où la langue maternelle est également la langue culturelle, tandis que pour saint Thomas comme pour tous les gens de cette époque, la langue qu'ils utilisent en écrivant est une *Traditionssprache*, une langue de tradition, pour reprendre la définition du latin médiéval donnée par Richard Meister et considérée par une personalité scientifique comme Christine Mohrmann comme la meilleure qu'on puisse fournir. Par conséquent, dans cette *Traditionssprache* il y a vraiment une attention absolument extraordinaire à continuer, à répéter un vocabulaire; ce vocabulaire ne va se renouveler que dans la mesure où l'école change – on a dit aussi que l'histoire du latin médiéval est l'histoire de l'école médiévale et on le constate bien pour saint Thomas: il y a tout un vocabulaire nouveau qui apparaît parce qu'Aristote est entré dans l'école médiévale. Pour cette introduction de mots nouveaux, nous sommes donc dans une situation radicalement différente par rapport à celle décrite par M. Brunet. Il faut toujours que ceux qui recourent à la statistique fassent de la critique historique en même temps et sachent exactement quelle est la spécificité du document de base des fichiers qui sont les leurs. Comment se sont constitués ces fichiers et de quel contexte historique sortent-ils?

125 ZAMPOLLI Ci sono evidentemente due accezioni dell'espressione « parole nuove »: parole di nuova creazione rispetto a un certo stato di lingua; parole non ancora incontrate nello spoglio delle opere analizzate fino a un certo momento. Per

quanto concerne questa seconda accezione, non mancano certo i dati per varie lingue. In genere l'accrescimento si riduce a un livello pressoché stabile dopo un certo numero di opere. Non è evidentemente il caso di entrare qui nella discussione se il lessico di una lingua sia da considerarsi infinito, aperto, chiuso, ecc. ... I fonemi sono invece, chiaramente, un elenco finito. Nella mia tesi ho rilevato che, all'interno dello stesso « genere letterario », le frequenze dei singoli fonemi si stabilizzano dopo *tranches* di 10.000 fonemi. Il problema delle conseguenze dei diversi criteri di lemmatizzazione adottati è certamente gravissimo. Nel nostro Istituto abbiamo dei progetti per creare dei metodi e degli strumenti che permettano la omogeneizzazione semiautomatica (nei limiti del possibile) di materiali lemmatizzati diversamente in italiano e in latino. Quando varammo il progetto del LIF, scegliemmo criteri e precauzioni statistiche che garantissero la facile comparabilità dei nostri risultati con quelli del progetto analogo che Juilland stava conducendo a Stanford. Purtroppo il confronto tra i due lessici risultò laboriosissimo e incompleto a motivo delle diversità tra i criteri di lemmatizzazione, sia pure non apparenti a livello esplicito, ma di fatto presenti nella lemmatizzazione concreta.

126 MULLER Je me félicite qu'on n'ait pas parlé de richesse lexicale ... Le problème est de savoir si on peut pousser un dépouillement assez loin pour dire qu'on a épuisé le lexique, c.-à-d. le lexique virtuel, qui est dans la tête de celui qui écrit. Il y a trois manières de chercher cette limite. A) Faire un dépouillement très, très grand B) Limiter le dépouillement à des catégories des mots, comme fasait Yule il y a un demi-siècle pour les substantifs C) S'attaquer à un lexique de situation très pauvre.

Quant au A) je crois qu'il n'est pas possible d'épuiser le vocabulaire, même indépendamment de la création des mots nouveaux.

Quant au B), il y a des catégories fermées, mais pas beau-

coup, par ex. les pronoms; mais je ne suis pas sûr que les prépositions les soient. Avec les catégories ouvertes, verbes et noms, on n'arrive pas à l'épouisement. Je connais une recherche sur la catégorie des verbes français semi-auxiliaires, que l'auteur appelle co-verbes, c.-à-d. tous ceux qui sont suivis d'un infinitif, comme « pouvoir, devoir, vouloir »; après 20.000 occurrences, la catégorie n'est pas épuisée.

Quant au C) je me suis martyrisé avec la lemmatisation de 42.000 occurrences d'un bulletin metéorologique: il y a des signes statistiques qu'on approche de l'épuisement, sans l'atteindre.

Les statisticiens s'en occupent à propos des chasseurs de papillons. Les papillonologues savent qu'on ne connaît pas toutes les espèces. On va dans une île où il n'y a pas d'humains mais beaucoup de papillons. On les attrappe et on les met dans des boîtes: même type, même boîte. Aussi longtemps que les boîtes avec un seul exemplaire sont plus nombreuses que les autres, on sait qu'il y a un plus grand nombre d'espèces pas encore trouvées et on réussit à donner une estimation de ce nombre ...

Et bien – cela a été une suprise pour P. Busa – pour tous les corpus de n'importe quelle grandeur, les mots avec fréquence 1 sont plus nombreux que les mots avec fréquence 2, qui sont plus nombreux que les mots à fréquence 3, etc. On n'a pas trouvé d'exceptions jusqu'ici. Le vocabulaire est donc ouvert. Pas tellement à cause des mots nouveaux, mais parce que la différence entre les mots les plus fréquents et les *hapax*, même sur 70 millions de mots, cette différence est telle qu'il y a toujours une fréquence zéro ... Il y a un calcul qui permet d'estimer le nombre de ceux qui sont restés « dans le fond de l'urne ».

127 Zampolli È quello di cui ha parlato ad Ottawa, durante il Coling 76?

128 Muller Le système de Waring oui donne des résultats, mais n'y croyez pas trop. J'ai fait des expériences avec Corneille et des textes de Molière. Corneille contient environ

4000 vocables: pas beaucoup pour une oeuvre de 500.000 mots. Si on prenait le dictionnaire qui a paru peu après la mort de Corneille, on peut y biffer les mots dont je peut dire que, à cause des lois du genre, ils ne seraient jamais entrés dans le texte de Corneille. Je crois qu'on n'arriverait pas très loin de cet chiffre-là. Mais il faut être décourageant: rien à présent ne permet de dire: « Après une certaine quantité du texte on a déjà le lexique entier. Cela ne vaut pas la peine d'aller plus loin ». Cela n'est pas vrai.

Mais il y a aussi des faits plus encourageants. On va publier une thèse sur la richesse lexicale. Le chercheur a pris six auteurs contemporains, dont *a priori* on pouvait dire que le patrimoine lexical est très différent. Le classement par richesse des mots dénoncé par les premiers 500 mots (une page ...) n'a pas été changé en continuant l'analyse.

3.2

129 Busa Il me semble que dans nos séances la statistique globale est comme un sommet vertigineux de montagne, que l'on atteint pour en redescendre immédiatement ...

On n'a pas encore parlé de la segmentation morphématique des mots, qui pourrait être un chapitre du traitement des mots individuels.

À propos des types et des catégories de mots, voir d'abord mon Annexe n. 1.5. A) mots communs, B) noms propres, C) mots speciaux. A) et B) sont de veritables lemmes. Au contraire les C) sont des pseudo-lemmes, c'est-à-dire des mots groupés pour d'autre raisons que l'identité d'un même thème flexionné.

Il y a le problème de la définition du nom propre et des frontières entre propres et communs. Dans mon IT une partie de ces problèmes a été évitée en réunissant dans la classe B) tous les noms propres et leurs dérivés proches, sur la base de

suffixes déterminés, par exemple: *Arius* et *arianus*, *Donatus* et *donatista*: bien sûr on n'a pas distingué, encore, les emplois substantifs ou adjectifs.

On a envisagé qu'il faudra bien différencier les analyses lexicographiques, lexicologiques et statistiques des noms propres de celles des noms communs.

Par conséquent on a accepté l'existence d'homographie entre, par exemple, *aquila*, nom commun et *Aquila*, nom propre chez St. Paul.

Au n. 3. de mon Annexe, d'autres coupures typologiques se rattachent aux formes en tant que formes. Est-ce que dans une statistique globale il serait intéressant de les analyser type par type?

La même question se pose pour les homographes. Mon Annexe au n. 4.2. montre que 57 % des mots de ST sont des « homographes possibles », selon mon système de l'homographie, mais sans y compter l'homographie entre les désinences flexives du même lemme. J'appelle « homographes possibles » les mots du texte qui, abstraits de tout contexte et isolés comme mots disponibles, pourraient être des formes de lemmes différents.

130 Bolognesi Pour les linguistes l'analyse des préfixes, infixes et suffixes, est très intéressante. ... Le nombre des homographes latins devient plus large aussi parce que l'informatique n'enregistre pas la quantité des voyelles latines.

131 Tombeur *Cum timore et tremore.* Je voudrais d'abord poser une petite question au père Busa – peut-être ai-je été distrait il y a un instant –: est-ce que votre question maintenant quant à la statistique globale porte uniquement sur le lot de mots qualifiés de homographes, ou est-ce que votre question porte sur l'ensemble du vocabulaire étant donné qu'à l'intérieur de ce vocabulaire il y a ce lot de homographes?

132 Busa C'est le deuxième point.

133 TOMBEUR Donc, est-ce que l'on peut valablement faire une statistique globale en ayant une catégorie extrêmement importante de mots homographes, puisque celle-ci dépasse la moitié du vocabulaire? J'ai commencé par dire *cum timore et tremore*, parce qu'il y a encore pas mal de points qu'il faudrait préciser, me semble-t-il. Quand vous avez parlé tout à l'heure – et cela intervient dans l'homographie –, de la distinction noms communs-noms propres, jusqu'où allez-vous? Dans une phrase de psaume telle que *nolite tangere christos meos*, est-ce que *christos* est un nom propre? J'ai des tas d'exemples de ce type.

134 BUSA Si je demande à toi, Paul?

135 TOMBEUR Je vous répondrai assez clairement, ce que je n'aurais pas été capable de faire il y a quelques années. En ce qui concerne cette catégorie, noms communs-noms propres, je crois que nous devons « désidéologiser » le vocabulaire – c'est un nouveau mot, celui-là! – et noter que *christus* n'est pas un nom propre: c'est l'*unctus* en grec qui est *sémantiquement* employé pour désigner *Iesus Christus: dominus noster Iesus Christus.* Le mot fonctionne d'ailleurs dans les textes de deux manières différentes. Pour moi *christus*, au niveau d'une lemmatisation de type formel, général et non sémantique, c'est un nom commun. Et par conséquent il en va de même pour *antichristus.* Mais je rencontre d'autres problèmes en sens invers: *leuita*, c'est qulqu'un de Levi: par conséquent, c'est un nom propre et je ne distinguerai donc pas le *leuita quidam*, qui est simplement un ministre du culte.

Est-ce que vous croyez qu'il est possible, à partir du moment où l'on parle de statistique globale (c'est-à-dire à partir du moment où l'on désire faire fonctionner tout le vocabulaire et interroger l'ensemble comme un tout) de tirer quelque résultat qui soit fiable tant qu'il y a cette dynamite que représente l'homographie, présente dans le fichier? Je comprends très bien que vous n'ayez pas résolu l'homographie et que vous vous soyez contenté de la signaler, étant donné l'ampleur de votre travail. Il faut

consacrer souvent bien du temps pour distinguer *peccatorum* venant de *peccator* ou de *peccatum* – encore qu'il s'agisse là de mots de la même famille –, ou *oblitus*, de *obliuiscor* ou de *oblino*. Tant qu'il y a cette homographie non résolue, est-ce que vous croyez que l'on peut faire une statistique?

Si je peux encore dire une chose: je ne suis malheureusement pas d'accord avec ce que nous venez de dire, Père, quand vous notez que l'ambiguïté est la qualité d'un mot isolé. Il me vient quantité d'exemples. Je vais rapidement vous en donner quelques uns; sans doute peut-on trouver des contextes, des situations morphologiques et syntaxiques qui entourent ces mots et qui résolvent les ambiguïtés – cela vaudra surtout si l'on connaît les textes qui se trouvent comme en-dessous de ces contextes, et qui sont de nature à lever l'ambiguïté (pensez à une ambiguïté telle que *uos mundi estis*, qui est levée si le passage du chapitre 13 de saint Jean se trouve en-dessous) –. Prenons quelques lemmes courants pour lesquels bien souvent les situations morphologiques et syntaxiques ne résolvent pas l'ambiguïté. *Leuis* et toute la famille (*leuitas, leuare*, etc.): sans doute la quantité est-elle différente, *e* bref ou *e* long, lequel vient du grec, mais cette quantité n'apparaît pas dans l'écrit. Autre mot ambigu, important dans les textes philosophiques et théologiques: *condicio* et *conditio*: venant de *condicere*, de *condere* ou de *condire*; dans ce cas, l'étymologie est différente, la sémantique est différente, le concept est différent. *Comparare*, ce qui vient de *cum* et *paro* et ce qui vient de *compar*; il en va de même pour toute la famille (*comparatio*, etc.). *Contentus*, de *contendere* ou de *continere*. Pour tous ces vocables, l'environnement immédiat lève-t-il l'ambiguïté? Dans certains cas, oui; dans beaucoup d'autres cas, non. C'est ce que j'appelle la dynamite!

136 BUSA Tu as eu besoin d'ajouter: des contextes immédiats ...

137 BUSA Il y a là des problèmes de lemmatisation à résoudre au moins avec des conventions. Mais il n'est pas possible de se passer de la catégorie des noms propres.

Le mots communs, comme « action » et « arbre » et les mots déictiques comme « ceci » et « cela », sont bien différents sémiologiquement: les premiers signifient un concept que j'ai dans ma pensée; avec les derniers j'indique un objet tout entier. Les noms propres sont plutôt du côté des mots déictiques, proches parents des acronymes comme IBM ou FIAT, « mots-étiquettes » qui se réfèrent directement à un seul objet sans l'intermédiaire d'une représentation mentale, idée de leur définition. Selon leur adjectif de « propre » (*quod uni, quod soli, quod semper*), le nom propre en toute rigueur serait celui qui signifie une réalité individuelle dans le sens intensif de ce mot.

L'homographie empêcherait-elle la validité d'une statistique des mots? Ma réponse est « non », pourvu qu'elle soit faite intelligemment. La raison ultime en est que l'homographie est une ambiguïté de quelques mots isolés de tout contexte. Dans le texte on ne doit pas trouver de mots ambigus, car la syntaxe leur ôte toute ambiguïté, sauf les très rares cas d'expression imparfaite ou de malice trompeuse ou de quelque stylème parfois déliberé.

C'est vrai que dans l'IT il y a beaucoup d'homographes possibles pas encore sélectionnés, mais (sauf pour les 13 formes mentionnées dans mon Annexe n. 4.2. alinéa 3), il s'agit toujours de mots dont la probabilité d'être employés dans d'autres sens est tout à fait négligeable, bien que pour des raisons différentes.

Quand j'ai dit que dans mes textes 57 % des mots sont des homographes, cela ne signifie pas que de ce 57 % des mots on ignore le lemme. Nous avons « désambiguïsé » 800.000 mots en lisant leur contextes l'un après l'autre. Comme je viens de dire, nous avons mis de côté 400.000 mots comme homographes encore à sélectionner. Sur ces 13 formes (*quam, quod, secundum* etc.) dont l'homographie est monumentale, je ne ferai jamais, avant de les trier, une statistique qui impliquerait leurs différences sémantiques et syntaxiques. De presque 4,5 millions de mots, nous

avons jugé négligeable la possibilité qu'ils vérifient l'autre sens signalé comme possible par notre LEL (*Lexicum Electronicum Latinum*).

138 Crespi Reghizzi La lemmatisation est-elle toujours faite à la main ou aussi par un programme?

139 Zampolli Questa domanda chiama in causa non solo il settore degli spogli, ma anche gli altri del *linguistic data processing*. A livello morfologico, e cioè per la riconduzione di diverse flessioni alla forma di base, disponiamo di metodi e strumenti affinati in anni di lavoro, anche per lingue con difficoltà affatto peculiari, quali il tedesco (parole composte, ecc.).

Resta aperto il problema della lemmatizzazione degli omografi. In pratica ancor oggi, quando il dizionario di macchina segnala che una forma può essere omografa, tocca al ricercatore esaminare una per una le occorrenze della forma nei loro contesti per sciogliere la omografia. Si sta cercando, ovviamente, di rendere automatico, o il più possibile automatico, questo scioglimento. Ci sono due approcci fondamentali.

Il primo consiste nel tentare di assegnare all'intera frase una rappresentazione della sua struttura nel quadro di una determinata grammatica. Evidentemente si hanno qui problemi analoghi a quelli dell'analisi automatica delle lingue naturali, per i quali non basta operare a livello sintattico, ma spesso occorre invocare il livello semantico, pragmatico, cognitivo, ecc. ... Tuttavia non conosco un solo sistema di questo tipo capace di trattare un normale testo in linguaggio naturale.

Il secondo approccio sta invece nella esplorazione del contesto immediato, alla ricerca della presenza (o assenza) di categorie o parole specifiche che rendano accettabile (o inaccettabile) solo una delle possibili classificazioni dell'omografo. A titolo di esempio si possono citare i progetti per la codificazione grammaticale automatica: per l'inglese del Brown Corpus (a Providence, Rh.I); per lo spagnolo del nostro Istituto (dove le regole sono ricercate in-

duttivamente per mezzo di una *discovery procedure* che opera su stringhe di testi già lemmatizzati); per il francese all'ILF a Nancy e al LASLA a Liegi; ecc. ...

140 MARINONE Quand par ex. Macrobe parle de *maximi doctorum* le contexte ne suffit pas pour disambiguer *doctorum* entre *doctus* et *doctor*.

141 BUSA Dans des cas pareils le jeu est entre mots apparentés, c'est-à-dire qui ont un même sémanthème.

142 MARINONE Quelles raisons pour relever aussi l'homographie possible et pas seulement celle « réelle »? Par ex., pourquoi s'occuper de *creator* comme d'un possible impératif passif? Un autre cas: les grammairiens latins nous assurent que *marium* comme génitif pluriel de *mare* en outre que de *mas* n'existe pas dans la langue, au moins jusqu'au 4ème siècle. Ne serait-il mieux de l'introduire dans le dictionnaire de machine seulement en cas d'attestation certaine?

143 BUSA Nous avons élaboré le système de l'homographie possible dans notre LEL, mais pas pour l'analyse statistique du texte. L'homographie « réelle », c'est-à-dire celle qui est attestée dans le texte, c'est lorsque une même forme de mot est employée une fois dans un sens, une autre fois dans l'autre: comme ST, qui, par exemple emploie 2 fois *marium*, pluriel de *mare* et une fois *marium* pluriel de *mas*. Mais cela implique que, en amont, parmi les possibilités morphologiques de la langue, il y avait déjà deux lemmes différents qui coïncident dans une forme flexionnelle graphiquement identique. Dans la description du système de la langue il fallait dénoncer ce fait. En aval, on pourra ou l'on devra dénoncer aussi l'autre fait que la possibilité de rencontrer réellement l'un des deux sens dans un texte est très mince, à cause des préclusions sémantiques. Mais possibilité morphologique et impro-

babilité sémantique sont deux données de fait à deux niveaux qu'il faut bien distinguer. Dans mon IT on les a toujours dénoncés pour alerter le lecteur: on ne sait jamais ...

144 Bataillon L'homographie est augmentée aussi par la graphie, qui dépend de l'arbitraire dû aux éditeurs et qui a changé pendant les siècles. Par ex. *ae* et *e* se distribuent différemment au XIII siècle: *equus*, juste, est pareil à *equus*, cheval, mais *aeris* de *aer, aeris* est différent de *eris* de *aes, eris* ...

145 Berni Canani Il faudra attendre peut-être des siècles pour enlever automatiquement d'un texte toutes les ambiguïtés; mais on peut déjà en enlever un nombre suffisant pour continuer la recherche.

146 Zampolli I programmi cui ho accennato prima hanno un rendimento tra l'80 % e il 90 %.

147 Tombeur Il faut bien faire attention quand on parle de pourcentages dans ce domaine: ce pourcentage reflète une quantité mais cela ne révèle rien en ce qui concerne le temps nécessaire pour examiner les homographes. Selon les langues, les 40 %, les 30 %, les 15 %, les 5 % mêmes qui restent peuvent représenter un travail identique.

148 Bozzi Un dizionario automatico va costruito anche in vista di applicazioni statistiche come, per esempio, a riguardo delle varianti grafiche menzionate da p. Bataillon. A differenza dei lessici automatici preparati per le lingue vive, generalmente predisposti per analisi per lo più sincroniche, p. Busa intuì che un primo dizionario di macchina del latino non poteva che essere semplice, uniforme e, per così dire, inerte diacronicamente. Di conseguenza organizzò il suo lemmario su di un principio fondamentale di uniformità morfologica. La sua formidabile esperienza, che i risultati ottenuti per l'*Index Thomisticus* confermano, ha

permesso oggi di progettare, grazie anche allo sviluppo degli strumenti e delle tecnologie a disposizione, lemmari più complessi dotati di sofisticati analizzatori fonetici e morfologici (mi riferisco all'interesse che rivestono le indagini su affissi, suffissi, prefissi ecc., come hanno giustamente fatto notare Brunet e Bolognesi), con evidente vantaggio per le applicazioni a testi di epoca diversa e per gli studi di carattere statistico.

149 TOMBEUR Mais finalement, Père, quelle est la réponse à votre question sur la possibilité d'une analyse statistique globale alors qu'on a quantité d'homographies non résolues présentes dans le fichier? La réponse est-elle oui ou non?

150 BRUNET Il faut pas atteindre une désambiguïsation profonde et complète avant d'analyser statistiquement les textes: on n'en finirait jamais, car de l'homographie strictement dite on tomberait dans la polysémie ...

151 TOMBEUR Cela n'est pas un problème de statistique globale. Il est évident que le père Busa peut faire une statistique parfaitement valable, et qui serait du béton, sur quantité de choses, mais il ne peut se livrer a une statistique globale. Voilà mon opinion. Les exemples que vous prenez, M. Brunet, ne sont pas pris dans une statistique globale. Il y a quantité de mots en latin pour lesquels l'interrogation statistique peut être extrêmement précise parce qu'il n'y a pas d'homographie.

152 MULLER Il y a deux cas d'homographie. L'un quand une forme a deux fonctions grammaticales: par ex. en français « le, le, les » sont à la fois articles et pronoms. Un autre quand par ex. « tu » est pronom personnel et aussi participe de « taire », relativement rare. Le cas comme « tu », je les ai codés. Pour « le, la, les », j'en ai fait la statistique cumulative des deux valeurs ensemble: a été un sacrifice, mais en bonne analogie avec les cas de polysémie, comme par ex. celui du mot « état ».