4.1

153 BUSA On m'a demandé d'éclairer d'abord quels sont les matériaux que j'ai sur bandes magnétiques, prêts pour des examens statistiques intelligents, c'est-à-dire qui tiendraient compte de l'histoire et de la typologie des textes. J'ai déjà archivé sur 35 bandes magnétiques (densité 6250) 25 millions d'enregistrements-ligne (enregistrement = *record*).

A) Tous les textes de ST et des autres auteurs font plus de 10,6 millions d'enregistrements-mot, de 150 octets (= *bytes*) chacun. 22 octets sont réservés au mot. Les autres contiennent: la référence et les numérotations progressives; les ponctuations, partagées en préfixes et suffixes; plusieurs codes de typologie, d'homographie et de limites de contexte à gauche et à droite; le code numérique du lemme avec sa classe; le type morphologique de la forme et les dix zônes de ses catégories morphologiques, etc. ...

B) Tous les contextes de la *Concordantia Prima* de ST font 2,5 millions de *records*, de longueur variable jusqu'à un maximum de 450 octets, dans le même ordre que dans la concordance; les premiers 113 octets de chaque enregistrement contiennent tout ce qui est nécessaire pour repérer et classer les mots-clés, le restant contient leur contexte.

C) Des 600.000 contextes de la *Concordantia Prima* des autres auteurs, je garde sur bandes seulement les premiers 113 octets, dont le format est identique à celui de ST.

D) Des 6 millions de trinômes de la *Concordantia Altera* de ST, on garde un enregistrement de 70 octets chacun, qui contient la précision des trois éléments du trinôme, les codes morphologiques et typologiques des mots-clé et sa référence.

E) Les 1.200.000 trinômes de la *Concordantia Altera* des autres auteurs, on les garde enregistrés comme les précédents.

F) Pour chaque lemme, chaque forme et chaque syntagme, on a groupé en trois séries distinctes les totaux d'occurrence dans

chaque ouvrage de ST et des autres, détaillés selon la typologie du discours.

G) Notre système lexical latin contient:

G1) 172.102 enregistrements de 260 octets, qui représentent les formes groupées selon leurs lemmes. Chaque forme est accompagnée par le type, par 10 zônes de codes morphologiques, par des codes d'homographie, de lemmatisation, de concordance etc. et par la précision du lemme; chaque lemme est accompagné par son type, par ses codes morphologiques et par les trois articulations de son thème arrangées pour la flexion automatique;

les deux, formes et lemmes, sont suivis par leurs données quantitatives: la distribution selon la typologie dans les ouvrages de ST et des autres, les totaux d'occurrence dans ST et dans les autres, le nombre d'ouvrages dans lesquels chaque mot a été trouvé, et combien d'occurrences sont des mots-clés dans l'une ou dans l'autre des concordances, etc. ...

G2) Toutes les mêmes formes arrangées en séquence alphabétique (les formes identiques dispersées par lemmes dans G1, on les trouve ici l'une après l'autre);

G3) La liste des 2.742 désinences régulières de la langue latine, codifiées morphologiquement et arrangées soit par alphabet, soit par paradigme, soit par homographie.

H) De 17 ouvrages de ST (800.000 mots en tout) on garde aussi le texte qu'on avait enregistré avant l'édition critique selon laquelle, après, il y a été revu et corrigé.

K) Enfin il y a une série de travaux sur d'autres textes:

K1) Tous les lemmes du Forcellini 1940 avec leurs codes morphologiques, arrangés dans une séquence générale alphabetique, ensuite leur *index a contrario* et enfin leur *index* selon les déclinaisons, les conjugaisons, etc.

K2) Le texte entier de la Vulgate latine du VT et du NT: environ 800.000 mots.

K3) Les textes non-bibliques de Qumran publiés jusqu'en 1965, en hébreu, araméen et nabatéen, et lemmatisés.

K4) Le volume 3 du *Farbenlehre* de Goethe en allemand, lemmatisé.

K5) Les Prolégomènes de Kant en allemand, lemmatisés.

K6) L'opuscule *De Diligendo Deo* de St. Bernard.

K7) Les *Testi Antichi Italiani*, édités par Ugolini.

K8) Le *De consolatione Philosophiae* de Boèce.

K9) Les 15.000 lemmes du *Catholicon* de Jean de Gênes (*Johannes Balbus*).

K10) À mes moments perdus je continue l'enregistrement et le classement morphologique des lemmes du *Thesaurus Linguae Latinae* qui ont déjà paru.

Je laisse tous ces matériaux en testament à qui voudrait les utiliser comme point de départ pour poursuivre la recherche linguistique. Bien entendu, tout est mis à disposition des chercheurs gratuitement (sauf pour les frais de reproduction). Je dois en effet consacrer le temps qui me reste aux recherches lexicographiques sur ST. Mon espoir est que des facultés, même de mathématique, statistique ou informatique, entreprennent des recherches statistiques sur mes matériaux comme sujets de thèses de doctorat. On y trouverait un *input* très vaste, et surtout déjà lemmatisé et classé selon la morphologie, la typologie du discours et la fréquence. Ce qui manque n'est pas la saisie des données, mais l'interprétation statistique proprement dite: tout le travail préparatoire est déjà accompli.

154 AVALLE A nome di tutti ringrazio p. Busa di questa ricapitolazione della sua pluridecennale attività di lessicografo. Come una specie di testamento, mi ha commosso.

155 BUSA La d.ssa Bartoletti chiede qualche precisazione sulla mia registrazione dei lemmi del *Thesaurus Linguae Latinae* (TLL). Per ogni lemma registro:

la specificazione del volume e colonna,

se il lemma ha iniziale maiuscola (nei volumi non dell'*Onomasticon*, presto interrotto),

quel codice che nel TLL dice a inizio voce se se ne riportano tutte le citazioni o no,

nostri codici in zona apposita che specificano se attorno alla parola intera o a parte di essa, vi sono parentesi diacritiche e di che tipo; cioè non intercalo queste parentesi diacritiche tra le lettere della parola come se ne fossero parte, ma le metto a parte come segni tipologici per la lettura della voce; lascio invece entro la parola quei punti che indicano lettere che mancano,

quelle « notizie » con cui il TLL specifica spesso ma non sempre, la morfologia della voce: declinazione, genere, coniugazione, variazioni di desinenze ecc.,

ai lemmi che rimandano ad altro lemma, abbiamo aggiunto a quale lemma rimandano,

per ognuno dei frequenti « lemmi interni » (ce ne siamo definiti la casistica) specifichiamo entro quale lemma ciascuno sia da trovare.

156 Bozzi Come vede p. Busa la ricapitolazione elettronica d'un lemmario del latino post-classico generale?

157 Busa Mi auguro che si realizzi il progetto del prof. Gregory d'un lemmario-repertorio (magari a banca di dati oltre che a stampa) che riunisca quanto di lemmi è stato pubblicato almeno per quella latinità posteriore che io chiamo latino dotto. Vorrei vederlo avviato prima di morire. Penso che il principale ostacolo sia il perfezionismo. L'attenzione consueta in filologia al dettaglio di eccezione, non dovrebbe rappresentare un deterrente per l'elaborazione informatica d'un insieme, il quale dovrebbe mettere in evidenza non solo un caso trovato a caso, ma anche tutti i casi dello stesso tipo e per ogni situazione dare i rapporti quantitativi con tutto il resto. Comunque a me sembrerebbe che si dovrebbero anzitutto registrare i lemmari singoli, uno per uno, senza preoccuparsi di unificarne i formati. Quando molti saranno su nastro magnetico, sarà certamente sempre possibile e facile unificarne i formati a programma sulla base d'un

minimo denominatore comune, con opportuni codici: come quello che il dr. A. Bozzi ha già collaudato a Pisa.

158 BATAILLON M. Muller demande quelles parties de ST on peut dater. Il n'y a aucun ouvrage qu'on puisse dater au jour près. Il y a quelques ouvrages qu'on peut assigner entre limites certaines de date. Les plus sûrs sont l'opuscule sur la théologie grecque et la glose sur Mathieu, datés du pontificat d'Urbain IV. Mais il y a aussi des datations qu'on a été forcé de corriger. Pourtant les trois grandes Sommes sont en séquence: *in Sententiarum, Summa contra Gentiles, Summa Theologiae*, qui font le 40 % de l'oeuvre et sont les plus intéressantes et les moins conditionnées.

159 MULLER Avec une telle série – une dizaine d'unités – et avec l'extraordinaire richesse d'information avec laquelle le p. Busa les a mises sur bande magnétique, on a sûrement la possibilité de chercher des phénomènes qui présentent une évolution régulière, et en tirer des hypothèses pour la datation des autres.

160 BATAILLON Dans les autographes on constate même une evolution de l'écriture et des abréviations de ST ... Par ex. on s'est aperçu que ST n'avait pas écrit *ideae in Deo* mais *ideae nudae* ... Et aussi dans les autographes on apprends que l'indicatif est flou et que les textes ont été normalisés par les copistes et les éditeurs ... Il a écrit pendant un quart de siècle ...

161 MULLER Ce serait surprenant si, pendant un quart de siècle, certains éléments de sa langue n'avaient pas évolué regulièrement en fonction du temps. Il faut les trouver, en cherchant même au hasard ...

162 BURATTO Qui sono di sussidio le tecniche di aggregazione ossia di *cluster analysis* ... In ogni opera si considerano

tutti i suoi aspetti morfologici per vedere quelli che sono fortemente collegati tra di loro ... Questa ridondanza significativa ricercheremo in primo luogo nel comparare testi datati con testi non datati ... con verifiche progressive e magari allargate alla natura del contenuto.

163 BATAILLON Ma ST spesso riprende elementi dai suoi scritti precedenti, come quell'articolo del *De Malo* riportato pari pari dalla *Prima Pars* della *Summa Theologiae*. In quanto ai contenuti, sembra che in teologia eucaristica ST usi *transubstantiatio* più da giovane che da vecchio.

4.2

164 SLOCOVICH Ho iniziato ad analizzare il testo di un poeta inglese del diciottesimo secolo. Su questo testo intendo fare i seguenti conteggi servendomi di un minicalcolatore. – Conto le ricorrenze di ciascuna forma lessicale presente nel testo, scarto anche quelli di elevata ricorrenza (poiché connettivi linguistici). – Costruisco una matrice quadrata avente per lato N, dove N è il numero delle forme lessicali distinte rimaste dopo gli scarti effettuati. – In ciascuna cella di matrice carico un coefficiente ottenuto dall'applicazione di una funzione di decadimento, come ad esempio una esposizione negativa; tale coefficiente dà una misura della correlazione presente nel testo per ciascuna coppia di lessemi. – Dalle informazioni contenute nella matrice, ricavo un grafo di tanti punti quanti gli *n* lessemi, dispongo idealmente il grafo così ricavato in uno spazio di altrettante *n* dimensioni. – Infine conto di rappresentare questo grafo in uno spazio a sole due dimensioni, utilizzando particolari tecniche di riduzione.

165 BUSA L'utilizzazione scientifica dei materiali da me elaborati va in due direzioni. I volumi a stampa sono prevalentemente per uso della lessicografia tomistica, che è mio compito promuovere. I nastri magnetici saranno principalmente per uso di ricerche di statistica linguistica. Resta fermo che i volumi forniscono informazioni necessarie alle ulteriori elaborazioni dei nastri.

In merito al tema propostoci per questa sessione – statistica globale di correlazioni sia fisiche che grammaticali – qualcosa è emerso: per es. che trinomi del tipo di *ad primum ergo* non hanno rilevanza. Che la possono avere binomi del tipo *et ideo*, *et sic*.

Ripeto che sarebbe utile comporre un elenco dei segni di stile da individuare tra consecutività di parole. Mi auguro che tale elenco vengo portato avanti in una tesi di laurea in greco presso il prof. Tarditi dell'Università Cattolica di Milano. È vero, come ha notato il prof. Avalle, che lo stile è un oceano infinito. Non avrebbe quindi senso tentare un elenco esaustivo dei segni. Ma resta sempre attuabile raccogliere da quanto è stato pubblicato di ricerche d'autenticità l'inventario dei segni che vi sono stati presi in esame.

166 AVALLE La metodologia che ho sviluppato sui testi italiani del sec. XIII non potrà essere senz'altro valida per altri periodi. La lirica dei secoli XII e XIII è molto formale ed ha un lessico chiuso ricco di stereotipi. Nel definirne lo stile le frequenze di parole singole non mi sono mai state utili: le etichette linguistiche del linguaggio cortese (*curialis*) come « amore, affetto, gelosia », non variano da un autore all'altro; altrettanto per i temi svolti. I poeti del Duecento sono istituzionalmente ripetitivi. Invece è stato rilevante l'esame della sintassi, anche se la ritengo difficilmente quantizzabile, e l'esame dei connettori, ossia delle congiunzioni coordinanti e subordinanti, rapportate all'ordine delle parole e delle corrispondenti strutture asindetiche. È

qui che i comportamenti linguistici si diversificano. Questi per me sono segni fondamentali nel caso di eventuali peripezie attributive di testi anonimi. La sintassi offriva allora più libertà del lessico. Un grammatico di Tolosa, *Vergilius Maro*, del primo medioevo, in un capitolo del suo trattato di retorica, *De scinderatione fonorum*, ha messo in evidenza le possibilità offerte da una diversa collocazione delle parole, dagli incastri di proposizioni le une dentro le altre, e, al limite, di eventuali parcellizzazioni delle parole stesse. I lirici « pisani » del Duecento, ad esempio, sembrano particolarmente sensibili a questa tecnica (da intendersi, nella fattispecie, un arcaismo), per cui nel caso di componimenti adespoti dove tale tecnica è presente, essa può essere utilizzata (accanto ad altri elementi, fonetici e così via) per localizzarli culturalmente in quell'area. F. de Saussure ha posto l'accento sul problema della divisione in unità lessicali della catena fonica di un enunciato, segnalando l'opportunità di allargare il campo tradizionale della parola sul piano « sintagmatico ». Per esempio « porta-cenere », e, aggiungerò per restare nel settore a me più familiare del lessico dugentesco, certi tipi di congiunzioni subordinanti composte, e così via. Tutte queste sono certamente unità di espressione: ma sono anche unità di parola? Dove vanno a pesare in termini di frequenze? Se i 92.000 lemmi del Forcellini li dovessi trattare come fa il Tommaseo Bellini, essi si molticherebbero vertiginosamente.

167 Tombeur J'ai été très intéressé par ce que vous avez dit. Je crois que vous avez tout à fait raison de souligner ce problème fondamental du lexique imposé. Je voudrais simplement faire une petite remarque. Si j'ai bien compris tout ce que vous avez dit, vous avez souligné que ce qui est important c'est la syntaxe, c'est notamment aussi l'ordre des mots, mais il y a quand même un point de rencontre entre la syntaxe et le lexique. Dans le lexique, il y a deux catégories de mots que l'on distingue avec des étiquettes sans doutes mauvaises, mais nous savons de quoi nous voulons parler: les mots significatifs (les substantifs,

les verbes, les adjectifs et les adverbes qui sont dérivés de ces catégories) et les mots outils. Vous avez dit que dans la syntaxe, ce qui est important, ce sont notamment les conjonctions de coordination et de subordination. Mais celles-ci précisément font également partie du lexique. Donc malgré tout, en ce qui concerne le lexique, vous avez quand même la tranche des mots-outils qui vous révèlent déjà une part de la syntaxe. Étant donné les problèmes posés par l'automatisation des recherches syntaxiques proprement dites, je verrais un deuxième rapport possible entre le lexique et la syntaxe. En interrogeant le vocabulaire d'une nouvelle façon, par groupes de deux mots, de trois mots, etc., on obtient de nouvelles informations de type syntaxique, sans nécessairement devoir faire – ce qui est cependant souhaitable – une analyse syntaxique proprement dite. Une partie du lexique des mots-outils et une nouvelle interrogation du lexique par groupes de mots permettent donc déjà d'atteindre en partie deux points importants que vous avez soulignés.

168 ZAMPOLLI Sono molti i centri che lavorano a trascrivere testi in *machine readable form*. Ma molto spesso tali testi restano senza utilizzazione, al di fuori del progetto che ne ha provocato la trascrizione. Uno dei motivi principali va cercato, a mio avviso, nella mancanza di analisi ai vari livelli. Evidentemente la codificazione delle unità linguistiche che si susseguono nei testi ai vari livelli è lunga e costosa, tanto più se dalle unità lessicali si passa alle unità sintattiche, semantiche, ecc. ... Problemi derivanti dalle varietà delle teorie linguistiche potrebbero essere superati adottando per l'analisi delle unità – per così dire – « neutre » o « preteoriche », e/o adottando un metalinguaggio per descrivere le analisi proposte dalle diverse teorie. Il problema maggiore è quello di creare dei sistemi assistiti dal calcolatore che riducano radicalmente o almeno sensibilmente i tempi e quindi i costi di analisi. Per quanto concerne l'ordine delle parole, temo che le ricerche – di fatto – siano molto scarse. A Pisa stiamo ap-

plicando l'analizzatore ATN per calcolare le probabilità di trascrizione tra i vari stati sintattici, in diversi « stili » dell'italiano.

169 Crespi Reghizzi En attendant qu'un système de représentation syntaxique soit prêt, il faut chercher les ombres de la syntaxe dans les mots outils ou dans les spécifications morphologiques, comme dans les désinences des cas en latin. Bien sûr, l'ambiguïté peut diminuer la valeur de ces ombres. La mésure dans laquelle la présence enchaînée de tels ou tels mots outils est une étiquette univoque de telle ou telle situation syntaxique, doit être d'abord explorée inductivement, phrase par phrase.

170 Zampolli Come ho detto, non conosco programmi capaci di compiere l'analisi sintattica completamente automatica di un testo in lingua naturale. La funzione dei *mots outils* nella delimitazione e nel riconoscimento dei gruppi sintagmatici o di *clauses* è stata sperimentata da molti, per es. per l'inglese e per l'italiano. Ricordo ancora quanto ho detto sui sistemi di esplorazione del contesto locale delle parole omografe nei progetti citati.

171 Voicu Je travaille sur des homélies grecques pseudonymes, assez courtes (presque jamais au-delà de 10.000 mots). Après avoir relevé les formules rhétoriques d'adresse au public, je me suis concentré sur le type « Écoute ce qu'on dit » et ses variations (singulier-pluriel, présent-futur, etc.). Il apparaît ainsi que le même polynôme apparaît dans un groupe défini de textes. Puis j'ai trouvé d'autres polynômes communs au même groupe, pour lesquels j'ai pu vérifier qu'ils n'étaient pas (ou presque pas) attestés dans d'autres tranches du corpus. J'ai abouti ainsi à un ensemble de polynômes exclusifs d'un sous-ensemble de textes, qui me paraît pouvoir définir dans ce cas, extrêmement favorable, un auteur.

172 BUSA En conclusion, sur mes textes, on peut envisager deux directions de travail immédiat. La première est l'élaboration des pourcentages de chaque mot individuel, tranche par tranche, pour en récapituler ensuite l'ensemble. La deuxième direction est de poursuivre une analyse syntaxique du discours de ST qui permettrait de reconnaître par un programme d'ordinateur les corrélations sémantiques des mots. C'est tout.