DISCUSSIONS CENTREES SUR LE LEL

LE LEL DE GALLARATE

Busa

Vous avez ici, au centre de la salle, les documents qui représentent le LEL de Gallarate. Le premier est la liste de tous les lemmes qu'on a dans le Lexicon Totius Latinitatis de Forcellini. On a entendu par lemme le mot qui est exposé en caractères plus gros comme titre de chaque article. Nous n'avons pas encore recueilli les autres lemmes, qui parfois se trouvent dans le corps des articles: affaire compliquée aussi à cause des renvois qui parfois répètent les lemmes qui se trouvent dans le corps des articles, et parfois ne les répètent pas. Nous y avons ajouté le code de provenance: "L" qui signifie Lexicon et "O" qui signifie Onomasticon; en effet l'Onomasticon est un tiers du Forcellini. On y a aussi ajouté le chiffre qui indique en combien de lignes le lemme est déclaré: information qui s'approche de celle de la fréquence.
Tous ces lemmes, perforés un par fiche, ont été ordonnés alphabétiquement. Il y a 89.609 lemmes dans le Forcellini. Ces lemmes, tels qu'ils ont été perforés, ne sont pas encore codifiés d'une façon pratique pour l'élaboration automatique.

Ensuite vous avez la liste des formes qu'on a trouvées dans les Oeuvres Complètes de St. Thomas d'Aquin. Il y a lá 10.500.000 mots, dont presque un cinquième sont des ouvrages apocryphes ou douteux. Dans 10.500.000 mots on a trouvé à peu près 150.000 formes de mots et 23.000 lemmes. J'ai dit "à peu près", car il y a encore quelques milliers de formes à lemmatiser, c.à.d. celles des formules, dont peut-être on parlera plus tard. Pour comparer avec d'autres situations, comme je viens de dire à M. Delatte, vous savez que nous avons perforé à Gallarate, et qu'on élaboré ici au CNUCE, les textes de Sénèque. Dans Sénèque il y a environ 45.000 lignes de texte avec autant de formes. Dans les 1.700.000 lignes de St. Thomas d'Aquin on a trouvé 150.000 formes. Le texte de St. Thomas est trente fois plus long que le tex-

te de Sénèque, mais les formes y sont seulement trois fois plus nombreuses.

Les formes dans cette liste sont rangées en "rationarium", c'est-à-dire groupées selon les lemmes.

La troisième est la liste des lemmes sous lesquels on a groupé les formes de St. Thomas. Ils sont à présent 23.040, à modifier après correction. Les lemmes y sont formulés de façon qu'on puisse les flexionner automatiquement, car nous avons distingué avec des codes appropriés la constante de la désinence. On a entendu "constante" non comme racine, mais dans un sens purement graphique, telle que nous l'a dit M. Delatte hier soir, c'est-à-dire la section du mot qui reste invariable dans la flexion.

La quatrième est la liste des désinences régulières ou presque régulières, que nous avons dérivées dans la grammaire classique de Schultz. Il y a 2742 désinences, et la liste n'est pas encore complète, je crois.

On a exposé ici les listes avec l'espoir que vous y trouviez au premier coup d'oeil au moins quelques-unes des erreurs qui certainement s'y trouvent. Il y a une loi psychologique à laquelle, je crois, l'humanité obéit depuis l'antiquité: que des erreurs vous échappent. Nous avons lu nos 10.500.000 mots au moins sept fois; nous avons fait trois contrôles complets: vérification à la machine et deux contrôles successifs à la lecture, complets. Après tout cela on y a encore trouvé 1300 erreurs de perforation échappées aux contrôles et ceci veut dire une erreur tous le dix mille mots ou une perforation erronée toutes les 80.000 perforations. Que l'erreur humaine soit inévitable, cela présente une analogie avec la loi de l'entropie en thermodynamique. Les fractutes d'attention dans le travail humain sont inévitables.

Bon, la structure de notre LEL à été commandée par des raison pragmatistes. En effet nous sommes convaincus qu'un LEL parfait et définitif, doit être tel qu'il nous a été décrit hier soir par Liège. Quand on a commencé, nous aussi nous avons examiné la perspective de l'analyse et de la lemmatisation automatiques. Mais à cause des nos quantités on n'a pas osé le faire, bien qu'on fût d'accord que la préparation d'un LEL pour l'analyse et la lemmatisation automatiques permet et impose une connaissance systématique complète de la langue latine. Et nous, qui n'avons pas commencé avec l'analyse automatique, nous y arrivons comme conclusion de notre travail. Je répète: le travail de Liège, à mon avis, ce n'est pas seulement audacieux, mais nécessaire. C'est vraiment une recherche fondamentale et définitive: scientifique au vrai sens du mot.

Vous connaissez la distinction entre index verbaux et index réels. Les index réels sont les index par concepts. La méthode dont nous allons parler

vaut seulement pour les index verbaux, c'est-à-dire pour ceux qui se fondent sur le recensement des mots tels qu'il sont écrits.

Dans le style des concordances il y a deux extrêmes. Des concordances sont de véritables lexiques, c'est-à-dire donnent aussi les groupements catégoriaux, sémantiques et syntaxiques, de chaque mot, comme, je pense, Liège a fait dès le début: c'est la concordance la plus complète, le maximum. De l'autre côté, au minimum, il y a les concordances à formes non lemmatisées. On y considère les mots comme unités graphiques: par exemple, "facies" c'est un seul mot graphiquement, pas encore trié selon l'homographie entre *facies-faciei* et *facio-facere*. Par exemple, notre travail pour l'Université de Padoue sur Sénèque,aboutit à une concordance à formes non lemmatisées. Ces concordances sont seulement des documents pour la recherche: pas d'interprétation, pas de recherche, rien de plus qu'un document. Qu'est-ce-qu'il fallait faire pour St. Thomas d'Aquin?

Ces possibilités extrêmes étaient, toutes les deux, impossibles à cause de la quantité. Dix millions et demi de mots représentent un chiffre trop élevé pour en faire une concordance syntaxique et sémantique. De l'autre côté, se borner à une concordance des formes non lemmatisées, semblait aussi une quantité trop imposante: on y aurait des centaines et des dizaines de milliers de lignes de concordance: impossibles à consulter à cause de leur propre quantité. Par conséquent on a décidé de lemmatiser les formes, c'est-à-dire de les grouper en paradigmes ayant un même lemme, ou titre, avec respect seulement de la morphologie. Comme les dénominations d' "analyse logique", "analyse grammaticale" ont parfois des fluctuations, disons-le ainsi: nous avons lemmatisé chaque mot isolé de tout contexte: chercher tout ce qu'on peut dire d'un mot quand il est tout à fait isolé de tout contexte, nous appelons celá "analyse morphologique", tandis que nous entendons par "analyse syntaxique" tout ce qu'on y ajoute, quand on considère le mot dans son contexte. Par exemple, on a dû renoncer à codifier les mots selon les "parties du discours": surtout la distinction en substantif, adjectif et flexions nominales-verbales n'a pas résisté. On a dû se limiter à grouper et codifier seulement les mots invariables, les mots à flexion purement nominale et les mots à flexion aussi verbale. Par conséquent on a inclus parmi les verbes toute flexion nominale-verbale: *votum, voti,* est parmi les participes de *voveo; peccatum, peccati,* est parmi les participes de pecco; *meritum, meriti; debitum debiti* etc. De même, plusieurs substantifs ont été inclus dans l'adjectif: *bonum, boni* par ex. se trouve dans notre LEL comme forme de *bonus, bona, bonum.*

Et aux usagers notre concordance servira comme document pour approfondir la recherche des catégories et sélectionner quand le mot *peccatum* est employé comme participe et quand il l'est comme substantif; *bonum* adjectif et *bonum* substantif.

Nous avons aussi "inventé" un cas adverbial à inclure parmi les flexions nominales: *dulciter* se touve sous *dulcis, dulce: suaviter* est un cas adverbial de *suavis, suave.* Mai *admodum* est codifié comme un lemme en soi et invariable.

Qu'est-ce qu'on a fait de la polysémie? On a découvert quatre niveaux d'homographie. On s'aperçut d'abord que l'homographie ne couvre pas entièrement tout le domaine de la polysémie. On a laissé la responsabilité à Forcellini de nous indiquer quand il s'agit d'un même lemme avec plusieurs sens ou bien de plusieurs lemmes avec chacun un sens différent: et nous ne disons pas qu'il est toujours cohérent. L'exiger serait comme exiger de préciser au millimètre la limite dans les eaux territoriales des états. Il y a des régions de frontière, où l'on pourrait discuter jusqu'à la fin du monde: aux extrêmes les situations sont claires, mais près des limites les distinctions entre unités sémantiques ne sont pas claires.

L'homographie a quatre niveaux.

1. Homographie entre différentes langues. Gift et Kind en allemand sont une chose et en anglais sont une autre chose. En latin on a par ex. *mane*, qui est l'indéclinable "le matin" et aussi l'impératif de *maneo-manere;* mais dans les commentaires de la Bible on trouve aussi *mane tekel phares,* les mots que pendant le dîner officiel de Balthazar une main invisible a écrits sur la muraille. Phares à son tour est aussi un nom de l'Ancien Testament. Il y a *"pro"*, mais on a trouvé aussi *"pro"* comme translittération du *pro* grec, et non pas le *pro* latin; il y a *peri* de *perio-perire* mais il existe aussi comme translittération du *peri* grec.

2. Homographie entre mots dont les racines ont des sens différents, et dont la coïncidence pourrait on dire, est due tout à fait au hasard: par ex. *facies* de *facies, faciei* et de *facio, facere.* Par exemple encore le numéral romain VI et VI, de *vis, roboris.*

3. Homographie entre mots dont les racines ont un même sens de base mais qui appartiennent à des paradigmes différents. Par ex.: *intellectus* de *intellectus, intellecta, intellectum*, participe de *intelligo, intelligere*, et *intellectus, intellectus* de la quatrième declinaison.

4. Homographie parmi les flexions d'un même paradigme. Par ex.: *amatis* c'est la deuxième personne pluriel de l'indicatif présent actif, mais c'est aussi le datif pluriel du participe passé; *veni* est parfait et impératif, etc. Dans notre LEL nous nous sommes passés de l'homographie du quatrième niveau et on a donné à chaque mot le code de la première flexion qu'il vérifie dans la séquence de son paradigme: ainsi *amatis* est toujours codifié comme deuxième personne pluriel etc., et *rosae* est toujours le génitif singulier. On sait qu'il s'agit de cas vraiment homographes. Mais on a dû eviter de faire le choix entre les différents emplois, pour n'être pas contraint de recourir aux contextes: travail trop long.

C'est à peu près un résumé de la structure de notre dictionnaire.

La procédure était ainsi. On a le dictionnaire de machine en séquence alphabétique sur bande magnétique. On trie les mots du texte en séquence alphabétique. On compare les mots du texte avec les formes du LEL. Si un mot ne se trouve pas dans le dictionnaire de machine, cela peut signifier deux choses: ou c'est un mot à ajouter au dictionnaire de machine, ou c'est une erreur. Par exemple, nous avons trouvé eicicitur: c'était une faute d'impression du texte, pour *eiicitur.*

Au début nous avons sélectionné les mots des deux Sommes, la "Somme Théologique" et la "Somme contra Gentiles": 1.880.000 mots, à peu près. Il y avait 49.000 formes de mots. On a élaboré un premier dictionnaire de machine de ces formes. Après on a sélectionné les mots du *Commentarium in 4 Libros Sententiarum:* un million et demi de mots. Alors on a eu des formes qui existaient déjà dans le dictionnaire et des formes qui n'y existaient pas encore. On a codifié ces dernières formes et on les a ajoutées au dictionnaire. Parmi les formes qui se trouvaient déjà dans le dictionnaire il y en avait dont les codes indiquaient qu'elles étaient une homographie HA (homographie admise a sélectionner avant de poursuivre). L'ordinateur nous a imprimé la concordance de ces formes, pour qu'on puisse faire la sélection à la lecture. Des autres formes qui n'étaient pas homographes, l'ordinateur transférait tous les codes à chaque mot du texte de la même forme. Nous avons sélectionné à la lecture et à la main la concordance des formes HA et nous avons transféré sur bande magnétique les codes de cette sélection, et ainsi la machine a lemmatisé les mots homographes.

Avant de terminer, je dois vous indiquer les trois groupes de codes que nous avons appliqués. Il y a d'abord les codes que nou appellons "codes morphologiques". Nous avons sept colonnes de codes morphologiques, dont

la septième est réservée aux variations graphiques, c'est-à-dire à la polymorphie (par ex. *advenerunt, advenere*)et à d'autres situations (par ex. *fuit, fuitque*, mot sans ou avec enclitique).

Pour ce qui concerne les codes, l'homographie, après la première recherche pour la découvrir, a été partagée en deux groupes: formes dont l'homographie est négligeable et formes dont l'homographie doit être admise à la sélection avant de poursuivre. Exemple de la première: *creator* pourrait être *creator, creatoris*, mais aussi l'impératif passif futur, de *creo, creare*. Alors nous avons indiqué que *creator* pourrait être une forme de *creo, creare*, mais nous l'avons traité comme s'il était toujours de *creator, creatoris*, c'est-à-dire comme univocisé. Prenons un autre exemple: *ibidem*, oiseau égyptien à l'accusatif et *ibidem* adverbe de lieu. Nous avons dit: ibidem, adverbe, univocisé; ibidem, oiseau, négligé. Dans le LEL leur homographie est développée et déclarée, mais la lemmatisation automatique attribue tous les *creator* à *creator, oris* et tous les *ibidem* à l'adverbe. Autrement, la quantité d'homographies à sélectionner était épouvantable. Pour découvrir l'homographie, nous avons procédé ainsi: en lisant la liste des lemmes de Forcellini, on se demandait de combien de lemmes chaque forme pourrait être une des flexions. On est sûr que des homographies nous ont échappé.

Nous avons un peu moins de 1300 formes dont l'homographie est admise.

J'ai déjà une collection des cas où le texte vérifiait le lemme négligé et non le lemme univocisé: parfois on s'est trompé dans le jugement d'improbabilité d'une homographie. Il y a un code qui indique s'il s'agit d'une homographie admise ou négligeable. Il y a aussi, un code qui indique combien d'homographies il y a.

Après on a codifié l'homographie des enclitiques.

Tumet pourrait être de *tumeo*, ou bien *tu-met; divisive* pourrait être de *divisivus, divisivi*, ou bien *divisi-ve*. On a donné un code à ces formes.

Une autre homographie est causée par l'onomastique. Nous nous sommes perdus avant tout à définir quand un nom est un nom propre ou non. A la fin, on a été contraint par la quantité de nos mots à nous limiter aux noms qui sont propres à la personne physique. L'homographie de ces noms propres n'a pas été développée, sauf en certains cas, quand il existait la probabilité d'être vérifiée dans nos textes. Quand cette homographie n'était pas développée, on l'a indiquée avec un code qui dit: Ce mot pourrait être aussi le forme d'un lemme qui se trouve dans l'Onomasticon de Forcellini. (Il est vrai pourtant que les lemmes de l'Onomasticon de Forcellini ne sont pas tou-

jours de véritables noms propres; et qu'il y a aussi des noms propres dans le Lexicon). Ici aussi j'ai déjà une liste de quelques cas où le texte vérifie l'homographie onomastique négligée.

Le troisième groupe de codes sont les codes de concordance: codes opérationnels pour ranger les mots dans les différentes concordances finales. Un de nos gros problèmes est celui de donner une concordance de dimensions raisonnables, bien qu'avec une quantité énorme de mots. Dans les 150.000 formes il y en a un peu moins de 600 qui représentent à peu près la moitié des mots du texte: l'auxiliaire, les conjonctions, pronoms, prépositions, et mots comme *Deus, homo, Christus, humanus, divinus,* etc. . . . Alors nous avons groupé toutes les formes dont la fréquence parmi deux millions de mots est au-dessus de 90; parmi ces formes on a sélectionné toutes les formes qu'on appelle généralement vides ou grammaticales, et qu'on pourrait aussi dire fonctionnelles, comme disent les Anglais.

Un code indique s'il s'agit d'une forme "fonctionnelle", ou "de contenu". De ces dernières on donnera la concordance dans le sens classique du mot. Des premières on donnera une autre espèce de concordance que nous appelons "concordance à syntagmes": groupes de mots à trois, classés par ordre alphabétique.

J'ai terminé.