PROJETS DE DICTIONNAIRES ITALIENS AUTOMATISÉS

Tagliavini

Nous sommes maintenant en train de préparer un dictionnaire de fréquence de l'italien contemporain.

On a établi, pour cela, de dépouiller des textes postérieurs à la dernière guerre mondiale: depuis 1945 jusqu'aujourd'hui. M.lle Umberta Bortolini et M. Antonio Zampolli collaborent avec moi

Les savants qui ont publié des dictionnaires de fréquence d'autres langues, ont d'abord défini une longue liste de titres: puis ils ont choisi au hasard parmi ces mêmes textes, tandis que moi je prefère faire le choix moi-même pour éviter de prendre des textes qui sont trop ressemblants dans leur sujet et pour choisir des textes qui ont des thèmes différents.

Notre échantillon est composé de cinq catégories de textes: cinéma, théâtre, romans, journaux et hebdomadaires, divulgation scientifique. Sans inclure les noms propres et les mots étrangers, il y a là 500.000 mots, 100.000 pour chaque catégorie.

Nous nous attendons à y trouver à peu près 50.000 formes graphiques différentes, qui se grouperont en moins de 20.000 lemmes.

La lemmatisation sera opérée avec la procédure semi automatique employée par le CNUCE, et aux lemmes aussi on appliquera des codes grammaticaux.

Les études de statistique linguistique nous ont donné des formules pour assigner à chaque lemme une "valeur d'emploi", qui est calculée sur la base pas seulement de sa fréquence absolue, mais aussi de sa distribution dans les différents secteurs de l'échantillon.

Il est bien aisé de traduire ces formules en programmes de machine. Nous prévoyons qu'il y aura plus de difficulté avec le choix des lemmes à introduire dans les dictionnaire de fréquence: en effet lequel est le rang de fréquence qu'on doit choisir comme le seuil d'introduction ou d'exclusion des lemmes?

Mais il faut aussi ajouter que ce même corpus de textes italiens contemporains pourra rendre service à d'autres recherches, par ex. phonématiques, grammaticales, syntactiques stylistiques.

Etablir les mots les plus fréquents dans l'italien moderne pourra très bien servir pour des manuels d'italien destinés aux étrangers.

D'ordinaire on a des mots qui se rapportent à des exercices; si un mot est une exception grammaticale, même l'enfant qui fait la deuxième année doit apprendre ce mot et par contre des mots fréquents ne sont pas connus.

Tous ceux qui ont étudié par exemple le Français connaissent le mot *hibou* car c'est une exception quant à la formation du pluriel. Mais combien de personnes qui parlent français ont l'occasion de dire que le mot hibou est un nom très fréquent?

Ou bien, quand on étudie l'anglais, on étudie que le pluriel de *brother* c'est *brothers* ou bien *brethren,* mais cette dernière forme est rare.

. . . . . .

J'ai dit ce matin que nous sommes en train de faire un dictionnaire de fréquence pour l'italien contemporain, et j'ai dit que seulement après cela on pourrait faire quelque chose d'autre. Il était bon que je le dise car ily a peut-être quelqu'un qui a la même idée. L'Institut de linguistique de l'Université de Padoue, en collaboration avec le C.N.U.C.E (et c'est pourquoi je pense que c'est justement l'occasion de le dire ici, au C.N.U.C.E.), est en train d'enregistrer, avec un calculateur, un vocabulaire de moyenne ampleur de la langue italienne.

Chaque lemme sera accompagné d'indications morphologiques, pour permettre la flexion automatique avec une table des désinences. Cela est tout à fait ressemblant à ce que vous êtes en train de faire pour le latin à Liège, avec en plus, ce que vous ne voulez ou ne pouvez probablement pas faire pour une langue morte telle que le latin, des indications de caractère phonétique et spécialement phonologique, en particulier pour la distinction, en syllabe tonique, des "e" ou "o" ouvertes et des "s" sourds ou sonores, des "z" sourds et sonores. Avec la spécification de ces éléments, et d'autres aussi par lesquels l'orthographe italienne n'a pas habituellement un signe de distinction, nous pouvons donner à la machine la possibilité de faire une transcription phonétique.

Avec ces moyens, on pourrait avoir aussi une transcription obtenue à la machine, comme d'ailleurs le Docteur Zampolli a déjà fait, quand il a fait son essai sur la phonologie italienne avec les machines électroniques.

Chaque lemme aura encore d'autres indications, par ex. sur l'étymologie, sur les acceptions diverses surtout en fonction des élaborations statistiques qu'on pourra faire.

Nous pensons alors (mon Institut de Linguistique de Padoue a chargé de ce travail M.lle Bortolini et M.Zampolli, mes anciens élèves, spécialisés ju-

stement dans ce domaine) qu'on pourra utiliser nos résultats comme dictionnaire de machine pour la lemmatisation des textes de la langue italienne.

Le dictionnaire de base pour nous est celui de Migliorini, c'est-à-dire non pas le Migliorini-Cappuccini, mais la nouvelle édition.

Nous avons choisi ce dictionnaire, car c'est le seul, à mon avis, qui donne des indications toujours sûres sur les voyelles ouvertes et fermées, "s" sourds et sonores etc. . .

Notre dictionnaire de machine ne permettra pas seulement d'automatiser les transcriptions phonétiques, mais aussi de calculer les oppositions phonologiques du lexique italien.

Je disais tout à l'heure que M. Zampolli a déjà publié une recherche statique sur la phonologie italienne, obtenue avec des calculateurs IBM.

Mais déjà Troubetzkoy, il y a un quart de siècle, a vu que il est très intéressant d'établir la différence entre données obtenues dans un corpus et données obtenues dans un dictionnaire.

Tollenaire

Vous permettez que je demande un petit renseignement à M.Tagliavini? Vous avez dit que les mots seront pourvus d'indications étymologiques. Est-ce-qu'il est possible de donner quelques précisions?

Tagliavini

Mais oui, certainement pour des études futures. Le dictionnaire de Migliorini d'ailleurs donne des indications étymologiques rès belles, en général bien mises à jour.

Le problème en italien est tout à fait particulier.

Nous avons une grande quantité de mots qui sont d'origine latine et qui ont eu une évolution régulière, une évolution populaire; mais il y a aussi un tas de mots qui sont des latinismes cultivés, qu'on a pris à chaque époque.

On dit: é il latino o viene dal latino? Le mot d'origine latine qui est un mot populaire, doit avoir une autre indication que ceux qui sont des mots savants ou semi-savants.

En général il nous suffira de dire quelle est l'origine la plus proche. Si par exemple un mot vient d'une langue américaine, mais à travers l'espagnol, comme le mot "patata", pour nous il suffira de dire: espagnol, patata. Que patata en espagnol soit d'origine des langues des Caraïbes, c'est autre chose, naturellement.

A ce dictionnaire de Migliorini qui donne des indications assez exactes,

on pourra même changer ou ajouter quelque chose, si on sait par exemple qu'il y a une autre étymologie intéressante. Je pense qu'il sera utile de voir la proportion des mots qui dérivent de l'allemand, du français, de l'anglais, sinon pour faire une statistique à la machine, au moins pour avoir un premier panorama.

Tollenaire

Alors, en ce qui concerne les mots, qu'est-ce qu'il arrive, lorsqu'il y a des mots dont l'étymologie est très douteuse?

Tagliavini

Oui, il y a aussi des mots d'étymologie inconnue. On peut dire: étymologie incertaine. C'est la même chose que fait, par exemple, Meyer en France, pour prendre le dictionnaire de Meyer qui est le plus recommandé pour les jeunes étudiants.

. . . . . .

Zampolli

En ce qui concerne les radicaux, je crois qu'ils seront formulés avec les préverbes; je ne sais pas si nous pouvons faire la séparation des préverbes, mais sans cette élimination, je crois que nous arriverons à 45-50.000 racines à peu près.

. . . . . .

Zampolli

Le Rév. P.Busa nous a demandé si dans le DM (dictionnaire de machine) envisagé par le Prof. Tagliavini et le C.N.U.C.E. l'élément de recherche sera représenté par des thèmes radicaux (au sens défini par M. Delatte) ou par des formes fléchies. Cette question concerne aussi la technique de consultation avec laquelle les mots du texte seront recherchés dans le DM. Je pense que le choix entre ces deux méthodes dépend strictement, soit du type de machine dont on dispose, soit des dimensions des textes à lemmatiser.

Bien qu'un DM organisé par stems ait été employé avec de bons résultats sur bandes magnétiques (par ex. dans le *SLC programming language and system* de *Georgetown University*) il est clair que son rendement optimum est lié à un équipement de disques magnétiques, tandis que pour une grande quantité de input le temps machine de *sorting* est compensé par la vitesse de l'algorythme de consultation en ordre alphabétique: j'ai déjà discuté de ces

aspects techniques dans une récente communication à l' "Accademia dei Lincei". L'optimum serait, je crois, un DM adaptable, selon les circonstances, aux deux méthodes.

Un DM est normalement défini comme un ensemble d'unités lexicales ou articles.chacun consistant en deux parties: l'élément d'entrée (ou titre) et la fonction (ou description). Une fois achevé le travail de définition des unités lexicales et des descriptions relatives, on peut les représenter en vue de la consultation avec un ou, s'il est nécessaire, plusieurs stems accompagnés d'opportunes codifications morphologiques et de tables de désinences. Mais on peut ensuite engendrer aussi, grâce à l'apparat morphologique d'analyse, toutes les flexions possibles. On obtient ainsi deux dictionnaires équivalents qui diffèrent par la technique de recherche. C'est justement ce que nous avons l'intention de faire. D'ailleurs l'explicitation de toutes les formes potentielles n'est pas sans utilité. Elle permet, par exemple, de mettre en évidence automatiquement toutes les homographies possibles et, si l'algorythme de flexion fonctionne même au niveau phonématique, elle permet d'établir sur des bases statistiques l'étude du rendement fonctionnel des oppositions phonologiques, grâce au dépouillement des "paires minimales".

Dans la langue italienne littéraire, il est possible de déterminer une grande partie des correspondances entre phonèmes et graphèmes de l'alphabet historique traditionnel en employant un algorythme relativement simple. Toutefois, dans certains cas il sera nécessaire de recourir à un DM, par exemple pour différencier *e, o* ouvertes ou fermées en position tonique; *s, z* sourdes et sonores; *i, u* voyelles ou semi-voyelles. On doit indiquer la position de l'accent et, en fonction de la phonétique syntaxique, les mots qui peuvent renforcer les consonnes initiales suivantes. Dans la tradition lexicographique italienne les dictionnaires remarquent habituellement ces phénomènes.

La remarque est formellement claire, toutefois il y a plusieurs exceptions pour les semi-voyelles, évidemment à cause de la complexité des problèmes qui s'y rapportent.

Sur la transcription phonématique de toutes les formes, dérivées des lemmes du DM il est possible d'exécuter ces statistiques de fréquence dans le système, qui, à partir de M. Troubetzkoy, sont considérées comme le complément nécessaire aux statistiques de fréquence dans le discours. L'école roumaine de linguistique quantitative a réalisé des exemples excellents d'analyse statistique sur le rendement fonctionnel des différents types de structure phonématique: "Etudier le rapport qui s'établit entre un certain nombre de

phonèmes, leurs possibilités de combinaison et les combinaisons réalisées, revient à établir . . . . *le rendement de la forme*" (N. Saramandu).

Les dictionnaires inverses et les études métriques offrent d'autres motifs pour ajouter au DM la transcription phonétique. On dit souvent qu'en italien l'orthographe et la prononciation sont moins différentes que dans des autres langues. Je pense, toutefois, qu'il serait correct, quand on organise des index inverses et des dictionnaires des rimes, de travailler sur des mots transcrits phonétiquement comme A. Julliand a fait dans son "Dictionnaire inverse de la langue française". M. le Dr. W. Ott, qui est ici présent, étudie la scansion automatique de l'hexamètre latin; au CNUCE est en cours d'exécution un projet analogue sur la poésie chorale grecque. Quant'à l'italien, nous avons à l'étude un projet d'édition automatique, en machine à photocomposition, de la Raccolta Barbi de chants populaires italiens.

Nous avons déjà dressé des dictionnaires de rimes qui diffèrent, dans leur structure, des dictionnaires de rimes traditionnels.

Peut-être faudra-t-il à l'avenir avoir recours à la transcription phonétique, du moins pour les mots ou les syntagmes "in posizione di rima" (A.M. Cirese). C'est notre intention de dégager, à côté des formes canoniaks de la rime de l'assonance et de la consonance, le tissu subtil et varié des correspondances phoniques qu'on a l'intuition que la conscience populaire doit accepter.

Le DM ouvre des perspectives intéressantes à la statistique lexicographique même au dehors de la phonétique. Sur le plan paradigmatique le DM offre des renseignements statistiques nécessaires pour établir des modèles de lexique: "proportions relatives de diverses classes syntagmatiques, structures statistiques des dérivés et des composés, extensions relatives de bases, rendement des formants lexicaux etc." (J. Dubois).

L'étude des suffixes, par exemple, a été conduite statistiquement soit à l'intérieur d'un micro-système, faisant remarquer la proportion des termes affectés d'un suffixe donné, relativement à l'ensemble des termes suffixés:(par exemple le travail de C. Tagliavini et U. Bortolini sur les *"etnici"*) soit pour tout le système d'une communauté, remarquant la disponibilité relative de chaque formant suffixal et la proportion des termes suffixés relativement à l'ensemble du lexique. Pour des motifs analogues on introduira dans le DM les renseignements d'ordre étymologique, dont M. le Prof. Tagliavini a parlé, et ceux d'ordre stylistique. Quant à leur codification nous avons des exemples illustrés dans les projets pour la langue tchèque et la langue hongroise.
Les remarques de style contribuent, entre autres choses, à sélectionner, dans l'ensemble du lexique enregistré sur le DM total, des DM plus ou moins ré-

duits en fonction de la nature des textes à lemmatiser.

Le DM peut et doit être considéré comme un instrument qui rend moins difficile la comparaison des données statistiques obtenues des dépouillements de textes différents.

"L'un des préceptes élémentaires de toute statistique est que l'on n'opère valablement que sur des populations, des individus et des caractères bien définis - a dit Ch. Muller - Aussi peut-il sembler hasardeux de faire des statistiques sur les "mots", puisque les linguistes sont d'accord pour constater qu'aucune définition satisfaisante de cet élément du langage ne peut être proposée".

P. Busa dans son exposé a clairement montré les alternatives possibles dans la lemmatisation d'un texte latin; des choix différents amènent de substancielles variations quantitatives dans les index de fréquence du "latercu-lum formarum" et du "conspectus lemmatum". Pendant la lemmatisation des 18900 mots environ du drame *"Veglia d'Armi"* de Diego Fabbri (élaboré électroniquement en 1959) j'ai pu vérifier ce type de phénomènes pour l'italien: en particulier le nombre des lemmes du texte se révélait augmenté à peu près de 25°/$_0$ si je considérais comme des lemmes différents les mots qui prenaient plusieurs valeurs syntaxiques à la suite des changements de catégorie: par exemple les formes nominales des verbes, les adjectifs substantivés, les substantifs adjectivés, etc.

En ce qui concerne l'italien littéraire, j'ai dressé une liste de ces alternatives dans ma thèse. A. Julliand a discuté, dans les introductions aux respectifs dictionnaires de fréquence, des problèmes que présentent l'espagnol et le roumain.

Pour la langue française on connaît bien l'article de Ch. Muller de '63 "Le mot unité de texte et unité de lexique en statistique lexicologique", dans lequel il distingue deux moments fondamentaux du dépouillement. La première des deux opérations segmente la suite linéaire du texte, en mots-occurrence; leur nombre est symbolisé par N. Avec la deuxième opération, qui correspond à notre lemmatisation, on classifie les "occurrences" et on obtient le nombre des mots différents qui se trouvent dans le texte; ce nombre est symbolisé par V.

Le calcul de V implique celui du nombre des mots employés une fois, deux fois, n fois dans le texte (valeurs $V_1 V_2 \ldots . V_n$). Muller énumère tout d'abord pour le français les cas où l'unité graphique, définie comme une série de lettres ou signes diacritiques ou pro-lettres entre deux espaces, ne corre-

spond pas au mot unité de texte tel que le lexicographe le conçoit.

L'unité graphique peut contenir plusieurs mots; ou bien, au contraire, plusieurs unités graphiques, consécutives ou non dans le texte, peuvent être considérées comme un seul mot. Il n'est pas difficile de retrouver des exemples en italien. En effet en italien contemporain les types d'unité graphique qui, pendant le calcul de N, peuvent être dédoublés en deux ou plusieurs unités de comptage, sont nombreux, même si l'apostrophe sépare d'habitude deux mots, et il ne cause pas d'ambiguïtés à ce point de vue, comme il arrive par exemple en français entre l'apostrophe de *l'agneau* (2 mots) et celui d'*aujourd'hui* (1 mot). Au contraire l'apostrophe remplace souvent des graphèmes différents, et, dans le cas de certains articles, pronoms, prépositions articulées ecc. *(l'anima, l'ozio)* engendre des possibilités d'homographie de fréquence remarquable.

Par exemple dans les index et dans les concordances de l'*Accademia della Crusca*, les articles et pronoms de genre différent *(il, lo, la, le, ecc.)* se présentent comme des lemmes autonomes et par conséquent, par exemple, la forme *l'* doit être distinguée en *lo, la, le ecc.*

Le trait d'union, normalement, n'est pas fréquent dans un texte italien. Il y a des cas comme "*il diretto Roma-Milano*" et "*la partita Torino-Juventus* " et comme "Super-Presidente" et "super-lo", "super-stazione". Dans la langue contemporaine on peut remarquer non seulement la diffusion de suffixes particulièrement fertiles mais on peut remarquer aussi la formation de plusieurs *prefissoidi* (B.Migliorini) qui possèdent la possibilité presque illimitée de former des composés nouveaux (*auto-; foto-; elettro-; galvano-; anti-; para-; mini-;* ecc.) qu'on trouve parfois écrits avec le trait d'union.

Certains savants proposent de les considérer comme des unités autonomes: on peut avoir des situations textuelles telles que "*sei iper o ipoteso?* ". Les remarques de A. Martinet dans son article "Mot, monème, synthème" sont très intéressantes à ce propos.

Certains prépositions articulées ont une fréquence remarquable; si on les analyse comme deux unités différentes (préposition + article) le nombre *N* augmente et les valeurs *Vn* varient sensiblement. Dans les index de l'Accademia della Crusca on assigne les fréquences des prépositions articulées seulement au lemme de la préposition simple correspondante: par exemple une seule forme *all'* est présente (cette forme correspond à *allo; alla; alle*), tandis qu'on a trois formes "*l'* ", lemmatisées respectivement: *lo, la, le.* Dans les index de *Veglia d'armi* et des textes-échantillons pour le dictionnaire de fréquence de l'italien, la fréquence des prépositions articulées est assignée aussi au

lemme de l'article. En ce dernier cas on doit choisir au moins entre trois alternatives: 1) traiter chaque préposition articulée comme une flexion autonome de l'article (Julliand, par exemple, l'a fait pour l'espagnol); 2) cumuler toutes les fréquences des prépositions articulées dans une forme unique qu'on définit *forma dell'articolo unita a preposizione* (mais alors, on se demande ce qu'il faut faire dans certains cas, par exemple: *collo* et *con lo?*); 3) ajouter la fréquence de la préposition articulée à la fréquence des formes correspondantes de l'article (par exemple la fréquence de "*alla*" ajoutée à celle de "*la*").

Les particules enclitiques *(mi, ti, ci,* ecc.) présentent une condition analogue. On néglige souvent l'enclitique et une forme telle que *amami* est tout simplement classée comme une forme différente du lemme *amare.* Certains index, ajoutent aussi une forme *-mi* enclitique au lemme pronominal *mi* (ou, parfois, *io)* tout à fait distincte de la forme proclitique (comme *mi ama*); il est toutefois raisonnable de se demander si en outre, par exemple, les occurrences de *darmi, darti, dargli, darci, darvi, darlo,* ecc. ne devraient pas être résumées dans une forme unique *dar-* ou bien tout bonnement ajoutées à la fréquence de *dar* ou de *dare.*

Les cas dans lesquels on peut estimer que plusieurs unités graphiques ne forment en réalité qu'un seul mot, sont plus complexes. Les formes des verbes composées par l'auxiliaire sont très nombreuses et, surtout dans les combinaisons *essere +participio passato,* il est difficile de décider si l'on a affaire à des formes composées ou bien à des participes autonomes.

Lorsqu'il s'agit de concordances qui ne se proposent pas directement un comptage statistique, les ressources typographiques permettent des solutions de dégagement. Avec la machine à photocomposition, d'habitude on imprime en italique, dans l'exemple, le mot exposant. Sous le lemme du verbe tous les exemples du participe sont résumés dans une forme unique, mais à l'intérieur de l'exemple on imprime aussi en italique l'auxiliaire éventuel. On peut employer le même expédient pour les *locuzioni* composées soit par des mots lexicaux *(segnare il passo, bagnato fradicio, avere fame)* soit par des mots grammaticaux *(se non altro, al di là, al di sopra).* Leur détermination est difficile, et, pour se rendre compte de cette difficulté, il suffit de penser aux discordances et souvent au silence des dictionnaires traditionnels.

Peut-être, les études sur la fréquence d'accouplement de mots dans le discours comme par exemple les *sintagmi* de P. Busa et les "groupes binaires" TLF de Nancy, pourront nous donner des connaissances plus précises sur ces groupes de mots qui, dans un certain sens, "préexistent au discours, sont assemblés en langue". A ce point de vue il est utile de ne pas négliger pendant

les dépouillements le problème des locutions.

Quand j'ai lemmatisé *Veglia d'armi* j'ai dressé à part une liste de locutions, même si seulement supposées, et j'ai transcrit la locution au bas de chaque lemme de tous les mots composants. J'ai rédigé, dans un but expérimental, trois différentes statistiques lexicales: la première sans aucune reconstruction des locutions; la seconde avec la liste des seules locutions considérées comme des lemmes autonomes dans le dictionnaire choisi; la troisième avec toutes les locutions qui me semblaient telles.

Je veux signaler, seulement à titre d'exemple, et, pour ainsi dire, "aneddotticamente", certains des plusieurs points controversés qui nous "tourmentent" dans la reconduction au lemme des unités que le procédé de segmentation a reconnues dans le texte.

En italien contemporain les cas de graphies différentes d'un seul mot sont plutôt rares *(dinanzi, dinnanzi; pronuncia, pronunzia; ricuperare, recuperare;* etc.*)*. Il y a aussi des locutions pour lesquelles on est incertain entre graphie unie et graphie séparée *(per lo più* et *perlopiù; in fine* et *infine)*. On trouve aussi des formes syncopées *(torre* pour *togliere)*; avec un i prosthétique *(ischerzo* et *scherzo)*; aphérétiques *(sto* pour *questo)*; les verbes avec des diphtongues mobiles *(sonare, suonare)*. Il n'est pas toujours aisé de dire laquelle entre les deux graphies est la plus commune, et on doit avoir recours à des conventions qui rendent plus cohérent le choix. Dans la langue ancienne et lorsqu'on dépouille des ouvrages distribués dans un vaste laps de temps, les variantes graphiques sont bien plus remarquables. Les chercheurs de *l'Accademia della Crusca* employent un procédé de "normalizzazione" des variantes. Ils les enregistrent sous la forme du lemme la plus commune aujourd'hui, mais, pour chaque type différent de graphie, ils créent un *"sottolemma"* qui est formulé de façon qu'il représente les variations de graphie. Souvent le lemmatiseur s'engage en réflexions et en consultations d'ordre étymologique pour distinguer entre mots différents et simples variantes, et pour formuler le *sottolemma* et le lemme. Il est probable que la forme se représente dans d'autres textes. Le DM fournit alors la solution décidée la première fois.

Le substantif est enregistré d'habitude dans la forme du singulier.

Plusieurs substantifs qu'on appelle *"mobili"*, du type *"animato"*, ont les deux genres masculin et féminin, auxquels correspondent, mais pas toujours, des formes différentes. Si on lemmatise distinctement *il maestro - la maestra*, et *cavallo-cavalla*, considérera-t-on homographes *il lepre* et *la lepre?*

L'adjectif est enregistré d'habitude dans la forme du masculin singulier. Plusieurs substantifs et adjectifs on les degrés altérés (diminutifs, augmentatifs etc.).

Les dictionnaires distinguent comme des lemmes autonomes *panca* et *panchina*, *tavolo* et *tavolino*, mais non *libro* et *librino*. Certains comparatifs et superlatifs ont une formation exceptionnelle *(massimo, minimo)*; on a aussi des superlatifs de formation régulière qui ont une acception caractéristique *(chiarissimo, Altissimo* etc.*)*. Comme le dit Muller, une régle générale n'existe pas: le lien sémantique devrait être évalué chaque fois dans la conscience des parlants qu'on pourrait sonder objectivement, seulement avec des enquêtes sur le type du différentiel sémantique des psychologues (dont je parlerai après). Cette affirmation vaut surtout pour les formes qui expriment deux ou plusieurs "sens" sans que leur fonction syntaxique en soit modifiée. Il faut décider s'il s'agit d'un seul mot ou de plusieurs, c'est à dire, il faut séparer les cas de polysémie des cas d'homographie. Tous les savants que j'ai connus se trouvent d'accord lorsqu'ils affirment que la seule solution pratiquement réalisable est de suivre un dictionnaire choisi comme arbitre.

Les verbes ont généralement en lemme la forme de l'infini. Il y a, toutefois, des verbes défectifs dont l'infini est très rare ou bien il n'est pas documenté; plusieurs verbes se présentent dans le contexte soit avec la conjugaison pronominale (réfléchie ou intransitive: *bagnarsi, alzarsi, dolersi)* soit en forme non pronominale *(bagnare, alzare, dolere)*.

Les alternatives sont particulièrement nombreuses pour les formes grammaticales. Certains pronoms ont une flexion régulière comme les adjectifs *(questo, codesto, quello, tutto*, etc.*)*; mais les pronoms personnels *(io, tu, lui, ci*, etc*)* peuvent être placés en exposant comme des lemmes autonomes ou bien groupés en séries différentes (par exemple: *io, tu, egli, noi*, etc.; *mi, ti, si, ci* etc.; *io, me, mi, me* etc.) ou encore tout bonnement comme des formes d'un seul lemme, pronom personnel. Les différentes *specie* traditionnelles des pronoms (*che* relatif, interrogatif, etc.) peuvent engendrer autant d'homographes.

Le nombre des mots qui peuvent prendre plusieurs valeurs syntaxiques par le "changement de partie du discours" est très remarquable: du reste pour le latin P. Busa a renoncé à élaborer ce type d'homographie, qui aurait énormément augmenté le pourcentage d'homographes possibles.

Infinitifs substantivés, participes passés adjectivés et substantivés, adjectifs substantivés et substantifs adjectivés, noms propres formés de noms communs *(Palazzo Reale, Casa Bianca)*, pronoms et adjectifs pronominaux, adjectifs et adverbes (ex. *tanto, poco, molto)*, prépositions et adverbes *(su, sotto)*, partitifs et prépositions (ho visto *dei fiori*, uno *dei tanti*) etc.: certains lemmatiseurs examinent comme des nomographes seulement les cas

où les différences de fonction se sont consolidées en différences sémantiques considérables, codifiées dans les dictionnaires comme des lemmes différents (ex. *amante,* mais non *bollente; fatto* e *detto* mais non *stimato, illustrato*). Pour une statistique des parties du discours il serait nécessaire de distinguer toujours, même si plusieurs fois l'analyse du contexte ne résout les doutes, comme souvent il arrive pour l'alternative participes passés-adjectifs. Cette casuistique nous donne l'occasion de remarquer qu'une codification opportune, peut rendre le DM valable opérativement pour des perspectives différentes de lemmatisation.

Supposons que nous voulons enregistrer les homographies dans le DM, avec un code qui les qualifie par catégorie.

Celui qui utilise le DM pourra employer le code pour annuler la marque de certaines homographies. La classification de l'homographie en types plus ou moins hiérarchiquent organisés, semblable à celle que P. Busa nous a proposée, devrait, à mon avis, être codifiée dans le DM, pour permettre des différents niveaux d'analyse du texte. Par exemple, en italien, l'homographie entre flexions d'un même lemme est très étendue (ex: les 3 premières personnes du subjonctif présent; la 3ème personne du singulier et la 2ème du pluriel de l'indicatif avec les formes de l'imperatif; la 2ème personne du pluriel de l'indicatif et le participe passé féminin pluriel *(pensate, stordite* etc.*).* M. le Prof. Tagliavini a distingué ces homographies dans les concordances de la "Divina Commedia" tandis que l'*Accademia della Crusca* les néglige d'habitude. Le DM devrait les enregistrer de façon que l'utilisateur puisse décider s'i les élaborera ou non.

Le DM pourra exercer une fonction de normalisation dans la lemmatisation à deux niveaux: il assure la cohérence des lemmatiseurs avec un dictionnaire choisi comme arbitre dans chaque cas particulier pour lequel une règle générale n'existe pas. Le DM garantirait aussi un comportement uniforme pour tous les mots qui, constituant une classe, ont les mêmes alternatives de lemmatisation.

Nous avons vu sommairement qu'une grande partie des incertitudes dans la lemmatisation se reconduisent à peu de types; le choix de l'une ou de l'autre alternative, puisqu'on n'a pas une définition complète de "mot", présuppose des conventions. Il n'est pas probable que les chercheurs acceptent d'employer les mêmes conventions pour des dépouillements qui ont des perspectives différentes, surtout lorsq'ils ne sont pas directement intéressés au dénombrement et à la statistique lexicale.

Cela n'empêche pas que, si les conventions choisies sont explicitées et

suivies avec la cohérence garantie par le DM, les résultats des dépouillements lexicographiques, après peu de transformations précises, soient comparables statistiquement entre eux.

Le DM peut aider le lexicographe dans la "formulation" du lemme, qui exige un soin particulier dans la publication des index et des concordances, et qui se compose de plusieurs parties. Tout d'abord on doit choisir un mot comme exposant, pour représenter tout le paradigme: les plus grandes difficultés sont posées par les variantes, et par les renvois relatifs.

Souvent le lemme est classifié grammaticalement pour des statistiques morphologiques. Outre cela il y a des indications complémentaires (synonymes, spécifications grammaticales) qui peuvent distinguer les exposants qui sont homographes, dans l'index même ou dans le système.

Il faut être cohérents quand on assigne ces annotations et il n'est pas aisé de se souvenir de tous les lemmes qui les demandent, surtout relativement au système. En outre, les annotations nécessaires peuvent varier pour des textes en états de langue diachroniquement distincts: par exemple, dans la lemmatisation d'un texte contemporain il n'est pas nécessaire de dire que *"compagno"* est un substantif et non un verbe. Ces différenciations peuvent être codifiées et rendues opératives dans le DM.

Aux difficultés de définition d'unité linguistique, il faut ajouter les difficultés de "définir les observables linguistiques" (R. Moreau).

"Le style - a dit P. Guiraud - est un écart qui se définit quantitativement par rapport à une norme." Ch. Muller remarque que "si on veut éprouver les méthodes quantitatives dans l'étude stylistique d'une oeuvre quelconque, la première démarche, et la plus décisive, sera le choix de la norme". Devra-t-on se rapporter à tous les ouvrages publiés par l'auteur, à l'ensemble des ouvrages du même genre littéraire, ou à un corpus de textes différents, de la même époque?

Dès les premières recherches sur les aspects quantitatifs du langage, conduites surtout par des chercheurs non-linguistes, la fréquence d'une unité (phonématique, lexicale etc.) mesurée dans un corpus suffisamment étendu et judicieusement composé est considérée assez généralement comme une bonne estimation de sa probabilité d'utilisation sans d'autres textes.

A la base des recherches sur la fréquence des graphèmes de la part des cryptographes, dactylographes, phoniatres etc. et sur la fréquence des mots pour la didactique de la langue, il y a l'utilisation d'un procédé tout à fait statistique. D'après les caractéristiques mesurées sur un échantillon bien

composé on formule une estimation sur les caractères d'une population non délimitée ou, de toute façon, trop étendue pour être soumise complètement à des textes.

Quand la statistique linguistique devint une discipline, on formula des postulats. P. Guiraud affirma que "la fréquence constitue un des attributs du signe linguistique": G. Herdan proposa, par des formulations différentes, le parallèle entre *univers statistique/échantillon* et *langue/parole* de Saussure. Ces idées ont provoqué de nombreuses réactions sur le plan théorique (N. Chomsky, Morris Halle, N.A.K. Halliday, R. Moreau etc.) et en partie sur le plan du contrôle concret (travaux du C.R.E.D.I.F.) etc. Mais encore aujourd'hui maints chercheurs de statistique linguistique considèrent ces postulats comme les bases théoriques de leur discipline (A.R. Alexandrescu, etc.) selon les prévisions de P. Guiraud. Les résultats des dépouillements semblent justifier l'affirmation de la stabilité de la fréquence d'unités comme les phonèmes, dans des textes différents. Mais en effet il n'est pas facile de définir la langue comme un ensemble bien délimité.

Même si nous ne considérons pas la difficulté de délimiter chronologiquement un tel ensemble, il reste que romans, journaux, revues, essais, pièces etc. constituent autant de sous-ensembles, plus ou moins caractérisés, plus ou moins structurés. Il est tout simplement impossible d'en dépouiller certains, p. ex. la correspondance et les conversations privées.

Faut-il parler de "fréquence" dans la "langue" ou doit-on se borner à des sous-ensembles? et en ce cas, de quelle façon les délimiter?

Le problème se présente, par exemple, lorsqu'on envisage la recherche, dont le prof. Tagliavini a parlé, pour un dictionnaire de fréquence. La littérature spécialisée donne des recommandations et des idées, mais elle ne fournit pas de renseignements quantitatifs pour la composition de l'échantillon.

D'excellents travaux sont munis d'un ensemble élaboré de formules pour l'évaluation des résultats des dépouillements, mais les auteurs ne justifient qu'en partie et d'une façon impressioniste la composition et l'étendue de l'échantillon choisi pour les dépouillements. Pratiquement on est encore aux recommandations de N. Troubetzkoy et aux adjectifs "assez étendu" et "judicieusement composé" des premiers chercheurs non linguistes.

Pour sortir de l'impasse, des élaborations théoriques seront utiles, telles que la division du lexique en mots de "classe fermée ou ergodiques et mots de classe ouverte, non ergodiques" (Moreau). Mais j'estime aussi nécessaire de procéder expérimentalement et de commencer par l'analyse quantitative de morceaux de textes, de textes entiers, d'états de langue strictement limi-

tés, mais bien définis. Ensuite on les comparera entre eux avec des méthodes statistiques convenables, pour vérifier la possibilité de stratifications plus étendues, déduites "a posteriori". Un tel projet ambitieux et improbable pour le passé, apparaît concrètement réalisable à l'état actuel de la *computational linguistics,* qui voit l'établissement de corpus toujours plus étendus.

La bibliothèque électronique du C.N.U.C.E. possède à présent plus de 30 millions de mots et son accroissement est rapide.

Dans cette perspective le DM est l'indispensable point de référence et non seulement sous l'aspect de nomenclature normative.

Pour chaque forme du DM on devrait enregistrer successivement la fréquence d'occurrence en ouvrages dépouillés par chercheurs différents.

J'ignore si on pourra réaliser le futur annoncé par G. Hays dans le récent "Seminar on Computational Linguistics" de Bethesda (USA - 1966): *la typewriter* de l'écrivain branchée, *time-sharing* au *computer;* un DM corrige les premières fautes, résout les abréviations, complète les références bibliographiques, les tirant d'une file spéciale. La machine à photocomposition, sous son contrôle direct, contribue à l'architecture de la page.

Il est pourtant vrai qu'aujourd'hui non seulement les linguistes s'adressent à la *computational linguistics,* mais aussi les savants d'autres disciplines: par exemple au CNUCE s'adressent des psychologues, des psychiatres, des documentalistes, des historiens, des juristes, des musicologues etc. A ces chercheurs les progrès de *l'hardware* et du *software* des machines offrent bien de nouvelles possibilités: l'input est facilité par les textes déjà perforés sur bande par les éditeurs; les langages conversationnels de programmation pour le dialogue direct chercheur-machine permettent l'intervention immédiate dans les procédés d'analyse des textes et dans la simulation des modèles linguistiques; la machine à photocomposition jointe à l'ordinateur, joint, à la variété et à la qualité des caractères, l'exactitude d'impression, tandis que la *type-setting,* à cause des fautes mécaniques très fréquentes dans la chute des caractères portés à la fusion, demande la relecture des épreuves.

Le CNUCE a équipé l'ordinateur IBM 360/30 réservé aux applications linguistiques d'un lecteur de bande, de bandes et de disques, de chaînes d'impression spéciales, grecque et latine, dont chacune a 120 caractères différents. De fait le nombre des caractères imprimables est bien supérieur, puisqu'on peut imprimer plusieurs fois sur une même ligne et pourtant il sera possible d'imprimer au dessus et au dessous et de combiner les caractères.

Au cours de l'année nous croyons pouvoir faire paraître les premiers index et concordances grâce à la machine à photocomposition.

Toutefois, les savants qui se servent de l'ordinateur, doivent compter aussi sur des instruments d'analyse valables, et avant tout sur un DM.

J'ai lu avec beaucoup d'intérêt les considérations des collègues du LASLA sur les difficultés de l'analyse d'une phrase latine et je suis d'accord avec eux. Je pense, toutefois, que lorsqu'on projette un DM italien on doit tenir compte de la condition actuelle des études pour la formalisation de l'analyse syntaxique.

Je ne peux pas comparer ici les différentes techniques employées; les publications dans ce champ sont très nombreuses: par exemple la description, au point de vue algorythmique, de G. Hays-S. Kuno a récapitulé le développement de *l'Art of Automatic Syntactic Analysis* au congrès de Bethesda. Les actes de la 2ème conférence internationale sur le traitement automatique des langues de Grenoble, 1967, donnent une bonne image des projets à l'étude, même dans le monde oriental.

Par exemple, en ce qui concerne l'anglais, il y a des programmes de *parsing;* certains de ces programmes expérimentés sur un vaste corpus de textes, ont été mis à la disposition du public. Je mentionne à titre d'exemple l'*English Parser* de Robinson (1962, 1965) et le *Predictive Analyzer* de Kuno et Oettinger (1963, 1966); tous les deux se servent d'une grammaire anglaise compilée à la perspective de *"context-free phrase structure grammars"*. La version courante du *"Predictive Analyzer"* employée à Harvard sur un IBM 7094, qui se compose à peu près de 3500 règles et d'un dictionnaire de 2500 formes, élabore 70 phrases par minute.

Il y a plusieurs structures qu'on ne peut pas actuellement analyser, ou bien auxquelles on attribue des descriptions correctes au point de vue de la grammaire employée, mais qui ne le sont pas dans la connaissance normale de la langue anglaise.

Plusieurs des analyses qui ne sont pas correctes pourraient être rejetées par le programme si on assigne au dictionnaire des catégories ultérieures, par exemple "verbe qui n'accepte qu'un sujet animé" et le contraire.

Toutefois, pour occasionner de semblables restrictions de compatibilité et de sélectivité entre sujets et prédicats et parmi les constituants, en général il faut de nombreuses catégories et de sous-catégories selon ce que Gardin a démontré, et pour le moment on ne voit pas "la possibilité de représenter systématiquement notre connaissance de l'univers" (S.Kuno). Les problèmes d'ambiguité syntaxique et sémantique sont à la base du pessimisme de beaucoup d'auteurs sur la possibilité d'une "high quality fully automatic mechanical translation" (Oettinger 1963, Dreyfus 1965, Kuno 1966, et surtout

J. Bar-Hillel, dans ses célèbres quatre leçons que j'ai eu la chance d'écouter à Venise en 1962.

Moins encore, donc, on pourra avoir quelques espoirs de réaliser dans un futur immédiat un analyseur syntaxique qui soit à même de satisfaire les exigences des élaborations lexicographiques qui sont encore plus rigoureuses, comme je crois l'avoir démontré dans ma communication aux "Lincei". Sur le plan pratique on pourra, tout au plus, penser à un *parser* qui découvre et qui révèle toutes les structures ambigües, laissant le choix au chercheur.

On ne doit pas sous-estimer les possibilités offertes au linguiste par une *console* et par un *display-scope*, placés dans son bureau en qualité d'extrémités d'un computer central; grâce à la multiprogrammation, l'intervention du linguiste n'arrête pas la machine; de cette façon les préoccupations des collègues du LASLA peuvent être vues sous un jour nouveau. B. Quemada aussi, dans une conférence à Nancy, en 1966, a préconisé l'utilisation de ces possibilités de *l'hardware* et du *software* dans nos recherches.

Dans la production automatique de concordances on demande une analyse de la structure des phrases, non seulement comme instrument pour l'interprétation des homographes, mais aussi pour des concordances particulières dans lesquelles on pourra grouper les exemples, non pas tant par "concordance" de mots exposants que par analogie de construction; c'est justement notre projet dans la recherche sur les chants populaires où il faut mettre au jour les formules qui en constituent le tissu: en l'absence d'un analyseur (on ne doit cependant pas oublier à ce propos les recherches de M. Alinei sur l'identification des phrases en italien) on commencera d'abord à rassembler des groupes binaires, ternaires, n. aires de mots. On doit considérer le développement de l'analyse syntaxique automatique et par conséquent du DM qui en est une partie intégrante non seulement pour ses applications de la science du langage dans le domain de l'automation, mais aussi par rapport au développement intrinsèque de la linguistique. G. Lepsky a éclairé historiquement les rapports de réciprocité entre linguistique mathématique ou computationnelle et linguistique structurale.

T. De Mauro affirme (1968) que le rapport entre l'identification du caractère structurel des phénomènes linguistiques et le procédé de formalisation fut déjà assez évident chez Saussure. I.I. Rezvin (Modeli Jazyka) montre la nécessité d'emploi de la méthode de la simulation dans le cadre des méthodes déductives en linguistique, et l'importance des applications computationnelles pour la théorie des modèles. On pourra employer l'ordinateur pour vérifier des modèles proposés (M. Pacifico. 1967).

Les modèles linguistiques sont classés en modèles "analytiques" (Rezvin) ou "analysant" (G.S. Cejtin) et en modèles inverses "synthétiques (ou "synthétisants"), selon que l'on part de séquences ou phrases données (modèle analytique) ou bien qu'on obtient ces séquences à la suite de certaines opérations (modèle synthétique ou génératif). Les deux types de modèles correspondent à deux orientations possibles de la description linguistique, à savoir: premier sens, des faits de discours au système de la langue, et deuxième sens, du système de la langue aux faits de discours (Rezvin).

Dans la traduction automatique ils sont employés tous les deux, le premier pour la langue source, le second pour la langue cible, tandis que dans la lemmatisation, dans le sens ici discuté, on utilise essentiellement l'analyse. En tout cas, les modèles opèrent avec des éléments ou de constituants auxquels on applique les règles de la grammaire. Au moment du choix de l'unité linguistique plus pratique on pourra soulever beaucoup de discussions: par exemple le morphème, (ou "monème" selon les différentes terminologies), ou bien le "mot"?

Au niveau de l'application, seul le DM peut procurer ces unités fondamentales avec les renseignements relatifs et les fonctions grammaticales, sur le plan de l'analyse, pour réaliser la segmentation, le morcellement de la chaîne parlée ou écrite d'après des critères formels clairs, et pour assigner à chaque unité sa "valence" de combinabilité. Sur le plan de la synthèse pour substancier avec des éléments appropriés les différentes terminaisons de la structure. Au contraire on pourra se demander si, dans les différents types de grammaires et techniques de "parsing" les renseignements qui doivent être reliés à un mot du DM seront les mêmes, ou du moins convertibles avec des règles formelles. Sur le plan pratique, cela signifie se demander s'il est possible de classer les unités lexicales d'un DM sans choisir, en même temps, des modèles et des techniques bien définis qui peuvent exclure les autres. Notre expérience ne nous permet pas encore de donner une réponse avec des faits établis, mais pour une indication positive sur une possible généralité du DM, je renvoie à la discussion qu'on a eu sur ce problème aux congrès de Bethesda et de Grenoble.

Le problème, peut-être, doit être abordé du point de vue d'une codification complète et bien structurée des informations syntaxiques.

Il n'est pas impossible de penser à un DM équivalent à la fusion du lexique et des renseignements de DM différents comme on a fait pour les dictionnaires russes chez Wayne State University.

H. Rubestein distingue trois orientations dans les recherches pour rédiger un DM pour l'analyse sémantique.

La première (*scaling*), dont on peut trouver l'exemple dans le *"Semantic Differential"* di Ch. Osgood, suppose d'obtenir des jugements, selon un ou plusieurs paramètres, de la part de ceux qui parlent sur la signification d'une expression donnée. Au C.N.U.C.E, il y a des psychologues qui ont employé cette méthode pour l'évaluation d'enregistrements cliniques. Toutefois je suis d'accord avec Rubestein qui juge cette méthode la moins convenable à nos buts.

La deuxième méthode *(statistical association* ou analyse distributionnelle*)* a été surtout employée dans *information retrieval.* On en a beaucoup discuté, par exemple dans les leçons du cours NATO à Venise en 1962.

Fondamentalement la méthode demande l'analyse statistique des distributions de deux mots dans le contexte, pour exprimer un jugement sur leur degré de similarité sémantique. Dans un certain sens c'est le programme de l'ordinateur qui constitue un thesaurus à partir d'un certain nombre de phrases de input. Je crois que cette méthode dans notre cas peut représenter surtout un complément ou une vérification d'un DM déjà constitué et tout au plus elle pourra donner une base remarquable de documentation.

En troisième lieu Rubestein examine les méthodes de la "componential analysis". Je ne sais pas s'il se réfère seulement à la théorie de Katz et Fodor ou bien à l'analyse sémique aussi (M. Joos, B. Pottier, A.J. Greimas), à l'analyse en facteurs sémantiques (Ivanov, Aspresjan), à la théorie de Weinrich etc. . . Même dans les différences qui les distinguent, ces théories se rapprochent sur plusieurs points: "les catégories sémantiques de Katz et Fodor coïncident évidemment avec les sèmes" (T. Todorov, Langages, 1966).

Nous pouvons résumer ces principes (de l'analyse sémantique) sous la forme suivante: 1 . . le sens d'un mot n'est pas une unité indivisible mais composée [. . . . .]; 2 . . les mêmes "atomes" de sens (les mêmes sêmes)se retrouvent tout au long du vocabulaire. Le nombre des composants serait inférieur au nombre des mots du lexique et les composants seraient organisés en structures déterminées.

Ces concepts sont utilisables évidemment dans la description sémantique des articles d'un DM, puisqu'ils permettraient la formalisation complète: l'article de J. J. Katz et J. A. Fodor "The structure of a semantic theory" paru dans la revue "Language" a été traduit par les "Cahiers de lexicologie" de Besançon. Les auteurs même préviennent" ≪ contre l'idée d'interpréter la conception d'une théorie sémantique (. . .) soit comme une procédure de dé-

couverte mécanique, soit comme une procédure de décision mécanique quant aux articles des dictionnaires ≫ et ils ne donnent ni listes de catégories sémantiques ni règles de projection. Malgré cela, on a fait des tentatives d'application concrète, par exemple celle de Rubestein. En Italie l'équipe de Parisi-Crisari a présenté un projet d'analyse componentielle au Congrès annuel de la Société italienne de linguistique qui a eu lieu à Rome en septembre dernier. Dans le CLTA de Bucarest (IV, 1967) I. Evssev présente un essai pour le roumain.

A. J. Greimas affirme que "l'analyse sémique du niveau sémiologique du langage apparaît non seulement comme la première tâche de la sémantique structurale, mais elle rendrait de grands services aux différentes démarches de la linguistique appliquée".

Pourtant ces applications présentent un certain intérêt pour la théorie même. "L'analyse sémique -a dit T. Todorov- n'a qu'une dizaine d'années d'existence et de nombreux aspects de la théorie seront sans doute révisés et perfectionnés. Il est souhaitable, en même temps, d'entreprendre une étude semblable du lexique entier d'une langue développée (disons, des 5000 mots les plus fréquents); une telle entreprise sera supérieure à l'échantillonage actuel: [. . .] on pourra se rendre compte des possibilités effectives d'économie dans le langage descriptif, faisant apparaître nombre de problèmes que nous ne percevons pas aujourd'hui".

On peut sûrement affirmer en dehors de tout jugement de valeur sur le plan théorique, que ces possibilités de formalisation de la description peuvent être appliquées concrètement pour enregistrer dans le DM l'ensemble de renseignements sémantiques qu'on trouve dans les dictionnaires traditionnels. D'ailleurs, comme affirme T. De Mauro (1968), ". . . I normali dizionari delle singole lingue, ... documentano ampiamente la possibilitá di analisi in componenti semiotici".

Négligeons pour le moment le problème du langage descriptif; s'agit-il d'un métalangage purement artificiel, relativement plus maniable pour la codification du DM, ou bien d'un lexique descriptif choisi dans le vocabulaire de la langue?

Même si nous envisageons le problème du point de vue pratique, on arrive, comme a écrit, mutatis mutandis, F. De Tollenaere dans son article "Lexicologie alphabétique ou idéologique (1960)" ≪au problème débattu, si le vocabulaire a une structure, et si jamais une description structurale du vocabulaire sera possible≫.

Cependant il y a des considérations théoriques qui pourraient soulever

des difficultés plus graves.

Même si T. DE Mauro affirme que ≪I noemi sono analizzabili in componenti . . . cui diamo il nome di componenti semiotici≫, il distingue toutefois entre ≪termini morfologici i cui noemi siano analizzabili in componenti semiotici dei quali uno almeno pertinente≫ et ≪termini lessicali i cui noemi siano analizzabili in componenti semiotici (. . .) nessuno dei quali é pertinente≫ tandis que ≪é convinzione comune degli studiosi citati (Eringa, Prieto,Lamb, Bendix, Coseriu ecc.) che i componenti semiotici. . . . siano elementi pertinenti, o per lo meno che in ogni noema vi siano almeno alcuni componenti semiotici pertinenti≫.

De là vient que ≪i componenti semiotici dei noemi lessicali non essendo pertinenti, possono essere di numero non finito (. . .) [e] che noi possiamo grazie ai noemi lessicali, assumere un numero infinito di nuovi sensi entro il codice (= langue)≫.

Une telle perspective n'exclurait à mon avis, ni la possibilité ni l'opportunité de l'analyse sémiotique nécessaire pour la compilation d'un DM; plutôt cela nous rend conscients du caractère institutionnellement ouvert de la recherche sémiotique. On ne pourrait pas envisager pour chaque terme la composition d'une grille de ses composants, mais seulement choisir entre eux un faisceau, conduisant la recherche sur le plan des ≪norme sociostatistiche di utilizzazione e realizzazione della infinita possibilitá semiotica della lingua≫.

Je pense que, encore mutatis mutandis, on peut citer pour le problème de l'ensemble d'informations sémantiques du DM, les conclusions de F. De Tollenaere sur la lexicographie idéologique. "Pourquoi ne pas essayer l'outil mis à notre disposition (. . .)? Bien sûr les difficultés surgiront nombreuses, et beaucoup de problèmes resteront à résoudre, mais on peut procéder à tâtons et sur une échelle réduite".