Overdruk uit:

Feestbundel F. de Tollenaere

1977

962/L



# Statistique linguistique et dépouillements automatiques par A. Zampolli, Pisa

## 1. *Quelques données*

L'on mentionnerait ici quelques données extraites de l'étude, à peine commencée, des deux dictionnaires de fréquence de l'italien écrit publiés récemment; le *Lessico di Frequenza della Lingua Italiana Contemporanea* de U. BORTOLINI, C. TAGLIAVINI, A. ZAMPOLLI (1971, nous l'indiquerons avec la lettre *a*) et le *Frequency Dictionary of Italian Words* de A. JULLIAND et A. TRAVERSA (1973, nous l'indiquerons avec la lettre *b*).

Mon but n'est pas de confronter les deux dictionnaires, même si pour la première fois peut-être, pour une même langue nous nous trouvons en présence de deux dictionnaires dont la comparaison pourrait avoir, dans une certaine mesure, quelque valeur[1].

Ces dictionnaires ont en effet en commun quelques importantes caractéristiques méthodologiques:

a) les critères quantitatifs de composition du corpus sont, en général, les mêmes. Aussi bien *a* que *b* en effet sont basés sur un ensemble de 500.000 occurrences réparties en cinq sous-ensembles de 100.000 occurrences chacun.

b) les critères de lemmatisation ont été sensiblement différents dans les deux dictionnaires, à tel point qu'une comparaison directe entre les divers ensembles d'articles n'aurait pas de sens. Il a toutefois été possible d'exécuter une série d'interventions qui, éliminant presque complètement les conséquences des divers critères de lemmatisation, ont permis de réduire les différences dans les limites problement acceptables pour pouvoir établir la comparaison[2].

c) les formules statistiques pour le choix et l'organisation des lemmes ont été les mêmes dans les deux dictionnaires. A chaque lemme du corpus a été assignée, aussi bien dans *a* que dans *b*, la valeur U (usage), avec une formule qui tienne compte de la fréquence globale du lemme dans le corpus et de l'uniformité de sa répartition dans les cinq sous-ensembles. Dans *a* et dans *b* ont été accueillis les quelques 5.000 lemmes avec une plus grande valeur de U.

Les différences concernent certains critères de composition du corpus qui sont aussi bien de type qualitatif, que de type quantitatif.

– *Limites chronologiques*: les textes de *a* sont tous postérieurs à la deuxième guerre mondiale; ils sont en effet compris entre 1947 et 1968. Les textes de *b* appartiennent au contraire à la période de l'entre-deux guerres: le premier texte est de 1920, le plus récent de 1940.

– *Sous-ensembles*: trois sous-ensembles de *a* peuvent être mis en correspondance avec trois sous-ensembles de *b*, du moins selon les étiquettes que leur ont donné leurs auteurs: "théâtre" (de *a*) que l'on indiquera désormais avec le sigle Ta, et "Dramatic Literature: Plays (de *b*), sigle Tb; romans (*a*) sigle Ra, et "Fictional Literature: "Novels and short Stories" (*b*), sigle Rb; journaux (*a*) sigle Ga; "Periodical Literature: Newspapers and Magazines" (*b*) sigle Gb[3]. Les autres ensembles ont au contraire des "étiquettes" indépendantes: "cinéma" (sigle c) et subsidiaires (sigle S) dans *a*, "Essaystic Literature: Essays" (sigle E) et "Technical Literature: Technical and Scientific Writings" (sigle TS) dans *b*.

– *Composition des sous-ensembles*: *a* et *b* ont suivi des critères différents pour composer les sous-ensembles. En premier lieu la composition de *a* est moins variée, plus

homogène. Dans *a* nous trouvons en effet: Ta, 10 textes et 10 auteurs; Ra, 10 textes et 10 auteurs; Ga, 5 noms de journeaux; E, 50 textes et 42 auteurs; les textes TS sont extraits de 33 publications différentes. En outre, tandis que pour *b* on a tiré au sort dans les textes des phrases isolées (en moyenne une phrase par page)[4], de façon à obtenir 100.000 mots par sous-ensemble, pour *a* on a pris dans chaque texte des passages continus. (Pour le cinéma on a pris le texte entier, pour le théâtre et les romans on a pris, dans chacun des 10 textes, un bloc de 10.000 mots. Dans les journaux on a pris des articles entiers de différentes sections; dans les subsidiaires on a pris des blocs de chaque matière.)

La diversité de composition du corpus de *a* et de *b* ressort donc d'après au moins trois facteurs différents. Il s'ensuit que les différences qui résultent d'une éventuelle comparaison entre *a* et *b* peuvent être mises en relation avec, en principe, l'un ou l'autre de ces facteurs ou avec leurs combinaisons. Nous ne nous trouvons donc pas dans une bonne situation pour établir une comparaison.

Une fois tiré au clair le fait que nous ne nous proposons pas de confronter *a* et *b* dans le but de rechercher des explications sur leurs éventuelles différences, examinons quelques données sur leur structure quantitative, en profitant des ressemblances méthodologiques dans leur élaboration.

Appliquons le calcul du $X^2$ aux fréquences des catégories grammaticales pour les 10 couples qui peuvent être formés avec les cinq sous-ensembles de *a*[5]. Les valeurs obtenues sont toutes très hautes, et l'hypothèse d'une différence significative dans la répartition des catégories grammaticales entre les sous-ensembles peut être naturellement maintenue. Par exemple, la valeur qui se trouve au croisement de la première ligne avec la deuxième colonne représente la valeur de $X^2$calculée pour la répartition des catégories grammaticales dans le cinéma et le théâtre.

Si nous considérons maintenant la ligne du cinéma, nous remarquons que les valeurs de $X^2$ augmentent de gauche à droite. Autrement dit, la différence mesurée par $X^2$ entre la distribution des catégories grammaticales dans le cinéma et cette même distribution dans les autres sous-ensembles croît dans l'ordre suivant: C, Ta, Ra, Ga, S. Considérons maintenant la dernière ligne, celle des subsidiaires: la différence entre les subsidiaires et les autres sous-ensembles décroît. Ainsi se dessine donc une séquence possible de sous-ensembles C, Ta, Ra, Ga, S, confirmée régulièrement par les trois autres lignes[6].

Si nous appliquons le calcul de $X^2$ à l'ensemble des lemmes présents dans chaque couple de sous-ensembles[7] (tableau 2), nous trouvons pleinement confirmé l'ordre en question[8]. Le tableau 2 mentionne, pour chaque couple de sous-ensembles, les valeurs de:

$$Z = \sqrt{2X^2} - \sqrt{2V-1}$$

où V représente les degrés de liberté. Ces valeurs sont comparables pour chaque couple car elles sont indépendantes de V qui est différent pour chaque couple.

L'indice de corrélation entre les rangs des lemmes des divers sous-ensembles suit le même ordre[9] (voir tableau 3).

La probabilité selon laquelle les valeurs citées se disposeraient toujours dans cet ordre par pur hasard est tout à fait insignifiante. On doit donc se demander si cette ordonnance des cinq sous-ensembles a un sens.

Essayons d'analyser la donnée globale relative aux catégories grammaticales. Dans le tableau nous mentionnons les fréquences en pour-cent de chaque catégorie grammaticale dans les cinq sous-ensembles disposé selon l'ordre indiqué ci-dessus.

Par rapport à cet ordre, les catégories grammaticales se divisent en trois groupes principaux:

a) les fréquences par substantifs, articles, prépositions, adjectifs augmentent selon cet ordre.
b) les fréquences de verbes, pronoms, adverbes diminuent selon cet ordre.
c) les conjonctions et les numéraux ne suivent pas cet ordre[10].

L'ordre C, Ta, Ra, S, Ga, est substantiellement confirmé par les données relatives aux classes de fréquence (tableau 5)[11], à la moitié, à la moyenne et à la mediane des cinq sous-ensembles (tableau 6)[12].

Recommençons l'examen des mêmes données pour *b*. Si nous ne considérons que trois sous-ensembles avec des étiquettes communes à celles de *a* (théâtre, roman, journaux) nous trouvons confirmé, pour tous les paramètres examinés, la séquence Tb, Rb, Gb, semblable en tous points à l'ordonnance Ta, Ra, Ga.

Nous trouvons par contre quelques oscillations si nous considérons les cinq sous-ensembles. L'ordre est en effet: Tb, Rb, E, TS, Gb en ce qui concerne les catégories grammaticales (tableaux 7 et 10)[13]; Tb, Rb, E, Gb, TS pour l'indice Z et l'indice de corrélation pour les lemmes (tableaux 8 et 9); Tb, TS, Rb, E, Gb pour les données relatives aux classes de fréquence, à la moitié, à la moyenne et à la médiane (tableaux 11 et 12, avec quelques exceptions pour la position de TS[14].

Il saute aux yeux que, si l'on considère ensemble *a* et *b*. Ra et Rb constituent une transition entre deux groupes caractérisés plutôt nettement: C, Ta, Tb, d'une part, Ga, S, E, TS, Gb d'autre part.

Essayons de considérer en détail l'union des 10 sous-ensembles.

La séquence pour $X^2$ sur les catégories grammaticales est représentée par la séquence C, Ta, Tb, Ra, Rb, Ga, E, S, TS, Gb (tableau 13).

Il y a seulement cinq inversions; toutes concentrées l'une à côté de l'autre. Elles se situent en effet: dans la ligne Rb, entre E et Ga; dans la ligne de Ga, entre R, S, TS, Gb; dans la ligne de E entre S et Ts.

Si l'on considère la séquence des fréquences en pour-cent des catégories grammaticales nous avons une confirmation de l'ordre dans la séquence de C à Ga, tandis que nous avons de nombreuses inversions entre E, S, TS et tandis que Gb se situe régulièrement à la fin de la séquence. Si nous substituons aux valeurs de E, S, TS leur moyenne, nous avons une séquence presque totale, avec seulement trois inversions: dans les substantifs et dans les adjectifs entre Ta et Tb, et dans les adverbes entre Rb et Ga (tableau 16)[15].

Si nous passons à Z calculé pour les lemmes nous trouvons la séquence:

C, Ta, Tb, Ra, Rb, Ga, E, Gb, S, TS, où se trouvent sept inversions qui concernent les catégories habituelles: deux sont situées entre Ga et E, deux entre E et Gb, trois entre S et TS (tableau 14).

L'indice de corrélation entre les rangs (tableau 15) montre l'ordre C, Ta, Tb, Ra, Rb, E, Ga, Gb, S, TS avec quelques inversions pour les cinq derniers ensembles de droite.

Comme il resulte évident de mon exposition, les tableaux rapportés ne veulent pas du tout être une interprétation statistique des données.

Ceux-ci peuvent être soumis à plusieurs élaborations statistiques: en particulier à mesurages de la signification de la corrélation entre les variations des différantes catégories, l'importance de ces variations etc.

Il est surtout clair que la classification morphologique adoptée est extrêmmement rudimentaire et qu'elle devrait être détaillée par une réduction en catégories qui devrait tenir compte de plusieurs fonctions linguistiques, soit sintaxiques que sémantiques.

Il ne manque pas exemples de ces élaborations statistiques qui sont indispensables même pour ne pas changer de simples régularités statistiques pour des faits importants

d'un point de vue linguistique (ROSS 1976; GEENS 1976).

On a voulu, ici, simplement rapporter des données indicatives qui illustrent, d'une façon claire et immédiate, les faits quantitatifs pour lesquels la statistique linguistique devrait fournir un modèle explicatif, en profitant aussi de la possibilité que nous offre, aujourd'hui le traitement automatique des langues.

Ces types de données, en effet, comme d'ailleurs ceux qui viennent de l'exploitation des grands corpora des texts en "machine readable form"[16], semblent confirmer l'orientation actuelle à diriger les recherches vers l'accumulation de données sur le comportement quantitatif des unités linguistiques des differents niveaux, et à construir, pour étages successives, un nouveau model explicatif qui en rende compte, dans le but de "améliorer, par les moyens statistiques, notre connaissance du fonctionnement du language humaine, et particulièrement les aspects quantitatifs de ce fonctionnement". (MULLER, 1976, p. 5).

## 2. *Quelques modèles proposées*

On sait que les premières statistiques sur la fréquence dans les textes des unités linguistiques au niveau phonologique et lexical ne sont pas l'oeuvre de linguistes professionnels. Des audiologues et des phonéticiens ont comptabilisé la fréquence des phonèmes ou des groupes phonématiques pour créer des instruments de mesure du d'éficit acoustique ou pour guider des processus de rééducation de dislalies ou dislexies. Des cryptographes ont, pour leur part, comptabilisé la fréquence des graphèmes ou des groupes graphématiques pour les utiliser dans l'élaboration ou la décryptation des codes. Des sténographes ont compté la fréquence des graphèmes, des syllabes, des séquences graphémiques de diverses longueurs, et parfois, des morphèmes, pour optimiser le rendement des systèmes dactylosténographiques.

Les données quantitatives fournies par ces dépouillements attirèrent bien vite l'attention des linguistes. G.K. Zipf utilisa largement les dépouillements du sténographe G.B. ESTOUP (1916) pour l'établissement de certaines de ses "lois", fameuses désormais, qui interpretèrent la distribution de fréquence des unités linguistiques, en particulier les phonèmes et les mots, à la lumière du principe du moindre effort. Mais ce furent surtout les pédagogistes et les psychologues qui, au début de notre siècle, utilisèrent les données statistiques disponibles pour certaines applications pratiques telles que la création d'un dictionnaire de fréquence pour l'enseignement d'une langue, l'étude de l'expansion du vocabulaire enfantin à des âges différents, l'étude du développement de l'apprentissage linguistique etc. De telles applications pressupposaient un modèle théorique, bien qu'incomplet, et rarement, sinon jamais présente de façon explicite, sur lequel il est bon de se pencher brièvement puisqu'il a représenté pendant longtemps les prémisses théoriques des recherches de statistique linguistique. En effet, c'est récemment seulement que les premières analyses des données quantitatives fournies par l'accumulation de corpus de textes en *machine readable form*, ont falsifié ce modèle en montrant que les théories et les soit-disant lois soutenues dans les 13 premiers lustres de notre siècle étaient basées sur un nombre insuffisant d'observations.

Les pédagogistes partaient de la constation que, si l'on s'en tenait aux résultats des dépouillements disponibles, un nombre relativement petit de mots très fréquents semblait constituer la majeure partie du texte tout entier. Il aurait donc été possible de "établir une liste de mots tels que:
les 100 premiers mots couvrent 60% de n'importe quel texte
les 1000 premiers mots couvrent 85% de n'importe quel texte
le reste (40 à 50.000 mots) couvre 2,5% de n'importe quel texte."

On supposait non seulement que ce type de distribution des fréquences lexicales fût

commun à tous les textes mais aussi que les mots de plus haute fréquence fussent les mêmes dans tous les textes. Il en dériverait donc, d'une part, que, pour les identifier, il suffirait de dépouiller un nombre adéquat de textes; d'autre part que, une fois énumérés, il conviendrait de partir de ces mots pour l'enseignement du lexique dans une langue, car en apprenant un nombre limité de mots, par exemple quelques centaines, l'élève serait en condition de comprendre plus de 80% des mots de n'importe quel texte.

D'un point de vue applicatif comme celui de l'enseignement, ce type de considération est certainement valable. On doit en revanche soumettre à la critique le modèle "schéma diurne" construit en axiomatisant ces propositions. Le rapport entre les textes et la langue est assimilé au rapport échantillon–univers statistique.

Les unités du système linguistique seraient caractérisées non seulement par des traits qualificatifs émergeant des oppositions et des relations qui forment la structure du système lui–même, mais aussi par leurs probabilités d'usage respectives. Ces probabilités ne sont pas directement observables comme ne l'est pas, du reste, le système. Cependant, elles seraient traduites par le fait (tenu pour certain) que les unités linguistiques se repeteraient dans les textes parlés et écrits à des fréquences relativement stables. Dans cette perspective, les fréquences observables dans les textes sont considerées comme des "approximations" des probabilités non observables du système.

Cette conception a reçu une formulation théorique explicité par P. Guiraud, même si celle–ci transparaissait déjà d'études précédentes, et fut acceptée et développée en premier lieu par G. Herdan et ensuite par d'autres chercheurs qui reconnaissent en elle le fondement théorique sur lequel la statistique linguistique pose sa propre autonomie en tant que science.

Ces prémisses n'ont cependant pas trouvé, jusqu'à nos jours au moins, un consentement général.

Le fait crucial et déterminant n'est pas, comme l'ont affirmé certains, dans le soi–disant caractère "statique" et "tassonomique" de la langue saussurienne par opposition à "l'aspect créateur" de la compétence chomskyenne, mais bien plutôt dans le fait que les dépouillements de textes qui se sont multipliés grâce à la diffusion des ordinateurs, ont démontré et continuent à démontrer que la fréquence des unités linguistiques, exception faite de cas exceptionnels, n'est pas stable. Les auteurs du *Français Fondamental* ont été les premiers à ébranler cette conception en introduisant la notion de disponibilité et au Congrès de Strasbourg qui s'est tenu en 1964 sur le thème "Statistique et analyse linguistique", R. MOREAU explicitait cette situation en affirmant que "les premiers pas de la statistique appliquée à la linguistique ont précisément consisté à admettre des règles du jeu qui soient simples. C'est ainsi qu'on a énoncé (voir par exemple Guiraud) que la fréquence des mots était constante dans la langue, ce qui supposait donc que l'on pouvait assimiler le choix d'un mot au tirage d'une boule dans une urne dont la composition reste inchangée au cours du temps" (MOREAU, 1966, p. 130).

E. HEGER considère qu'il est indispensable de modifier le modèle de Herdan, en introduisant entre la langue et la parole un troisième niveau, la Σ parole, où là seulement serait possible la quantification. Contre la solution de Herdan "à première vue tout à fait convaincante, il faut faire remarquer d'abord que les possibilités d'occurrences qu'elle postule et qu'elle attribue aux unités de langue n'existent pas indépendamment de facteurs thématiques et stylistiques, c'est–à–dire de facteurs étrangers au système qu'est la langue" HEGER (1969, p. 56–57). Mme Hirscheberg et Ch. Muller ont démontré que la stabilité de la fréquence lexicale est un mythe: les fréquences des mots, à l'exception de cas exceptionnels, ne sont pas stables, mais varient en conséquence du style et du sujet. Une seconde conception de la langue considérée comme univers statistique, conception qui a toujours été présente comme alternative à celle de Guiraud et de Herder.

Dans cette seconde conception, le concept de la probabilité comme caractéristique

intrinsèque de l'unité dans le système n'est pas exprimé. L'univers statistique est plutôt défini comme l'ensemble de tous les textes (parlés et écrits) produits en un certain laps de temps. Les textes du corpus sont donc considérés comme des échantillons extraits d'un ensemble plus vaste de textes, qui constitue la population statistique. La relation entre ces textes échantillons et la population est rendue complexe surtout par le fait que "A satisfactory delimitation of the text population (or populations) is a complicated task because for all practical purposes text populations are open set (uncountable)" (DOLEZEL 1969, p. 22). Mais, même si l'on admet que l'ensemble peut être exactement défini, une grande partie de cet ensemble échappe à toute sorte de mesure: par exemple les lettres, les conversations privées etc... A l'intérieur de cette population statistique, les textes doivent être regroupés en classes ou couches, selon les facteurs qui conditionnent leur production, tels que le style et la situation (au sens large), et qui semblent être responsables des différences que l'on rencontre dans les fréquences des unités linguistiques. On pourra ne considérer un corpus comme représentatif d'une classe homogène de textes que s'il est composé de textes extraits uniquement de cette classe. Si l'on veut au contraire un corpus représentatif de la population dans son ensemble, sa composition devra reproduire la stratification de la population et donc contenir non seulement toutes les catégories de la population mais encore dans la même proportion dans laquelle ces catégories figurent dans la population. Nombreuses sont les difficultés liées à la définition non seulement qualitative mais encore quantitative des diverses classes ou sous-ensembles de textes dans la population. Il suffit de penser aux incroyables diversités de composition du corpus pris comme échantillon par les auteurs des dictionnaires de fréquence, et à la subjectivité et à l'arbitraire de leur choix malgré le sérieux des efforts employés à reproduire dans la composition du corpus échantillon, la complexité de couches de la langue dans son ensemble.

Un autre problème, étroitement lié au précédent, est celui de mettre en relation les classes de texte avec la variation ou la stabilité des diverses caractéristiques quantitatives.

Moreau distingue deux catégories de mots: les mots de *classe ouverte* et les mots de *classe fermée.*

Les mots de classe fermée seraient tous les mots "athématiques", c'est-à-dire, ceux qui nous servent à nous "exprimer à propos des choses plutôt que d'exprimer les choses elles-même". On pourrait classer dans cette catégorie un certain nombre d'adjectifs et de vocables courants, de mots grammaticaux, et quelques noms très généraux: "des termes plus ou moins communs à tous les sujets et à toutes les situations". Leur emploi ne semblant pas varier sensiblement dans les différents centres d'intérêt, l'estimation de leur fréquence, à partir de textes échantillon, ne poserait pas de problèmes.

Le cas des mots de *classe ouverte* ou "thématiques" qui, au contraire présentent des oscillations de fréquence d'un texte à l'autre et souvent entre les parties d'un texte, est différent. Pour évaluer la fréquence de ces mots, le seul moyen qu'il conviendrait d'adopter consisterait à stratifier à priori la langue, à délimiter des centres d'intérêt, à l'intérieur desquels les mots seraient "thématiques".

L'idée de Moreau est précisement que les mots d'usage non stable dans la langue peuvent, au contraire, l'être dans des textes ayant un thème, un centre d'intérêt commun. Il propose donc de stratifier une fois pour toute la langue "à une époque donnée" en un certain nombre de centres d'intérêt, qui pourront être même très nombreux; par exemple: mathématique, physique, chimie, etc. À l'intérieur de chaque couche, on déterminera la fréquence des mots thématiques.

Il faut avant toute chose observer que ce modèle lexical ne répond pas tout à fait aux données de nos dépouillements. Dans les deux dictionnaires de fréquence de l'italien que l'on a cités, on a pu relever des variations de fréquence statistiquement significatives même à l'intérieur des mots que Moreau place parmi ceux de *classe fermée.* Il est

significatif par ailleurs que dans la fréquence des *mots de relation* entre les sous-ensembles dont les deux dictionnaires sont composés, les variations apparaissent étroitement liées à la nature du sous-ensemble, vu qu'elles suivent la même direction pour la plupart des unités qui appartiennent à la même catégorie grammaticale[17].

Dans cette conception il manque avant tout presque entièrement une liaison explicite entre la langue considérée comme système et la langue-population statistique prise au sens d'ensemble de textes. En second lieu, et probablement non indépendamment de cette absence, il manque un inventaire des faits linguistiques dont le comportement statistique peut être décrit en relation avec les classes de textes.

On est donc encore à la recherche d'un modèle quantitatif qui puisse être incorporé dans une théorie de la communication linguistique. Ce modèle devrait caractériser explicitement les variables linguistiques et extralinguistiques qui influencent la production des textes et indiquer comment individualiser les faits linguistiques déterminés quantitativement par ces facteurs. Un tel modèle, comme on l'a dit, n'existe pas, bien que certains chercheurs aient récemment apporté quelques contributions à l'étude des caractéristiques qu'il devrait avoir, des exigences auxquelles il devrait répondre, et des faits qu'il devrait expliquer.

Je mentionne ici le schéma proposé entre autres par DOLEZEL (1969) pour la stylistique et par ROSENGREN (1971) pour la linguistique. Ce schéma se situe dans le cadre de la théorie générativo-transformationnelle et suppose en particulier une théorie de la "performance".

On suppose qu'il existe dans la production d'un texte des mécanismes de "performance" qui veillent au choix et à l'application des règles fournies par la "compétence" et que ces mécanismes incorporent les processus de formation du style. Les fréquences relevées dans les textes sont déterminées globalement par les mécanismes de production du texte. Les mécanismes qui agissent de façon déterministe sont responsables des fréquences stables (en admettant qu'il y en ait: ce qui revient à dire que ces fréquences sont déterminées par le fonctionnement interne de la langue); par contre les mécanismes qui impliquent un choix sont responsables de la variation des fréquences et doivent donc être considérés comme des composants des "style-forming processes". Mais on ne sait rien dire d'autre sur ces mécanismes dont on se limite pratiquement à affirmer l'existence. On ne dit rien non plus des éléments qui mettent en mouvement ces mécanismes ou, pour ainsi dire, de leur input: le contexte de l'acte de communication; évidemment une telle spécification exigerait une théorie de la communication linguistique (V. LYONS, 1966, ch. 2 et 8).

Étant donné toutes ces indéterminations nous pouvons donc négliger les liaisons entre le schéma dont nous parlons et la théorie générativo-transformationnelle, et considérer le programme de travail proposé comme illustration de la nature et de la complexité des problèmes à résoudre.

Comme W. Winter l'avait déjà affirmé dans une communication au IXe Congrès International des linguistes, même selon L. Dolezel et I. Rosengren, dans la production d'un texte, le processus central en ce qui concerne le style est un processus de choix. Le processus sélectif peut grosso modo être ainsi défini: à chaque pas de la production du texte, le locuteur choisit un "mode d'expression" dans un ensemble A d'alternatives offertes par sa compétence linguistique: A = (a, b....n). Dans les cas les plus simples, un ensemble d'alternatives comprend deux membres seulement; par exemple A = (construction active, construction passive). Au moment de produire une proposition le locuteur décide d'utiliser la forme active ou la forme passive. Ce choix se répète chaque fois que l'ensemble est présenté au locuteur, dans notre exemple lors de la production de chaque proposition. Dans d'autres cas, naturellement, l'opération de choix ne se fait pas pour chaque proposition: par exemple le choix entre un ensemble de synonymes ne

se fait que lorsque l'on doit exprimer le "contenu" correspondant à telle série de synonymes.

En tout cas, la répétition de l'opération de choix a pour conséquence la distribution probabiliste des éléments de l'ensemble A: P(a), P(b)...P(n). A chacune des alternatives correspond donc une certaine fréquence d'occurrence dans la population des textes produits.

On suppose que le processus de choix est contrôlé par des facteurs de nature pragmatique. Chaque locuteur compose le texte "In accordance with his individual idea of efficiency, on the one hand, and in accordance with the requirements of the context (in the broadest sense) on the other. Those qualities of the speaker that are relevant for the selection processes (verbal preferences, mental type, stylistic skills, etc.) represent the *subjective factor* of selection. –Context– can be designated as the *objective factor* of selection; it is independent of speaker, though it exercises its influence through him. Context is represented by such factors of verbal communication as the form of language (written or oral), the form of discourse (dialogue or monologue), the genre, the function of the text, and so on" (p. 13).

Les facteurs pragmatiques, subjectifs et objectifs, sont toujours présents lorsque l'on produit un texte. Toutefois, le modèle reconnaît que l'impact de ces facteurs dans les processus de formation du style varie. Il existerait théoriquement trois possibilités fondamentales.

(1) Un locuteur (X), lorsqu'il choisit entre les alternatives possibles, n'est contrôlé que par ses propres préférences et ses habitudes personnelles; l'influence du contexte est éliminée. Ce cas semble plutôt exceptionnel. Empiriquement, ce type pourrait être représenté "by a poet who imposes his highly personal style on all texts he produces, regardless of differences in genres or forms".

(2) Le locuteur X supprime complètement ses préférences personnelles et lorsqu'il choisit dans l'ensemble des alternatives, il est entièrement soumis à l'influence du contexte. Exemple pratique: les documents burocratiques.

(3) Le locuteur X adapte ses choix personnels au contexte et conserve en même temps quelques–unes des caractéristiques individuelles qui le distinguent des autres locuteurs. C'est là probablement le cas du locuteur commun.

Dans le modèle, les deux classes de facteurs (subjectifs et objectifs) sont considérés comme deux forces en compétition entre elles: un même locuteur peut se comporter à chaque fois selon les possibilités (1), (2) et (3), même si (3) peut être considéré comme le comportement qui prévaut. Les processus sélectifs devraient expliquer l'origine et la variété des styles. Nous ne possédons pas de théorie sur la production de l'acte de parole qui incorpore ces processus et variables qui opèrent sur eux en input. Cette constatation est également vraie pour la théorie générativo–transformationnelle de la performance. L'étude des variables qui caractérisent les différences entre les divers styles reste encore presque entièrement à faire. Ces caractéristiques doivent être individualisées et mesurées à divers niveaux: phonématique, morphologique, lexical, syntaxique etc... Quelques–unes de ces mesures ont déjà été étudiées, mais d'habitude en dehors d'un schéma général.

Nous pouvons essayer de voir comment devrait être organisée cette étude. On doit avant tout établir une première tentative de classification des textes basée sur l'existence de facteurs pragmatiques de production. Les facteurs subjectifs et objectifs doivent être déterminés empiriquement, avant de connaître la structure statistique du corpus. Étant donné les textes du corpus TCL, ils seront classés suivant deux dimensions: a) selon les facteurs subjectifs: la classification produit les classes T (Xi), autrement dit les textes T de la classe T (Xi) sont tous produits par un même locuteur Xi; b) selon les facteurs objectifs: la classification produit les classes T (Qi), autrement dit les textes de la classe

T sont tous produits dans le contexte Qi. Les classes (Xi, Qi) représentent donc des ensembles de textes produits par un certain écrivain dans un certain contexte. Chaque classe peut être interprétée comme une "population" et être utilisée pour déterminer les propriétés statistiques des textes.

Examinons les cas théoriquement possibles, à l'intérieur d'un corpus de textes qui appartiennent à des auteurs et à des "contextes" différents.

1) Supposons que certaines caractéristiques se révèlent statistiquement stables dans l'ensemble des textes. Nous pouvons dire que ces caractéristiques sont de nature supra-stylistique. Si ce que plusieurs auteurs soutiennent est vrai (voyez les affirmations de Troubeckoy), la distribution des phonèmes et des graphèmes dans n'importe quel corpus de textes, pourvu que ce soit des textes en prose et qu'ils soient tous de la même époque, en serait un exemple.

2) Supposons que, pour certaines caractéristiques, il soit impossible d'identifier des parties du corpus à l'intérieur desquelles ces caractéristiques seraient statistiquement stables. Ce problème a été très discuté. Nous pourrions rechercher quels facteurs occasionnels, passagers, déterminent la fluctuation. Nous pourrons retenir ces caractéristiques comme sub-stylistiques. Toutefois l'existence de ces caractéristiques est un point intéressant pour la statistique linguistique, qui doit créer un modèle capable "d'expliquer" également ces caractéristiques.

3) Les valeurs d'une caractéristique sont statistiquement stables dans les textes d'un certain locuteur, indépendamment du "contexte", tandis qu'elles varient d'un locuteur à l'autre. Ces caractéristiques peuvent être prises en tant qu'expression du style individuel, c'est-à-dire produites par des facteurs subjectifs: (*S.C*).

4) Une caractéristique varie d'un contexte à l'autre, mais à l'intérieur du contexte elle ne varie pas pour chacun des locuteurs. Dans ce cas la caractéristique peut être prise en tant que résultat de facteurs objectifs (*O.C*).

5) Si une *S.C* fluctue entre des intervalles différents dans des contextes différents, nous pouvons considérer qu'elle est contrôlée par les deux types de facteurs. Nous l'appellerons alors subjective-objective (*S.O-C*).

Les caractéristiques ainsi spécifiées "can be treated as elementary quanta "as distinctive features", of style - that are practically unlimited in their combinatorial possibilities" (DOLEZEL, 1969, p. 21).

Ce schéma ne doit pas amener à croire qu'une caractéristique classée d'une certaine manière (*S.C, O.C, SO-C*, etc.) dans un corpus, doive appartenir à la même classe dans n'importe quel autre corpus. La nature statistique spécifique d'une caractéristique stylistique est déterminée par les facteurs qui la contrôlent. Comme il est possible que des facteurs différents s'alternent au contrôle, il est possible que "the statistical nature of a linguistically identical characteristic is subject to change in various texts (or text clas)" (ibidem, p. 22). Ce point est extrêmement intéressant mais naturellement il complique beaucoup les procédés de découverte des caractéristiques, de leur nature, des facteurs qui les contrôlent.

### 3. *Conclusions*

Comme on peut le voir facilement, ce modèle est surtout le résultat d'un exercice combinatoire sur quelques éléments dont l'existence est postulée. Méthodologiquement, on peut toujours prendre ce modèle, ou un autre modèle équivalent, comme hypothèse de travail et essayer ensuite de le vérifier expérimentalement. Il est essentiel, à mon avis, de constater que: les modèles précédents ont déjà été falsifiés; nous ne disposons pas d'une théorie adéquate ou même nous ne possédons peut-être aucune théorie des processus qui déterminent le choix entre les possibilités du système au

moment de la communication; les propositions actuelles de modèles ne sont guère plus que de simples expressions de l'exigence de construire cette théorie. Dans la situation actuelle il manque donc un modèle suffisamment élaboré qui donne à la statistique linguistique une base théorique adéquate[19]. Pour sortir de cette situation, il faut tenir compte des quelques données de fait acquises jusqu'à présent, qui semblent indiquer l'existence de sous-ensembles relativement homogènes de textes, et semblent adopter cette existence comme hypothèse de travail. Parmi les opérations à accomplir certaines sont particulièrement urgentes. Il faut examiner le comportement statistique du plus grand nombre possible de caractéristiques à tous les niveaux, et pas seulement les caractéristiques traditionnellement étudiées telles que la longueur des mots, la longueur des phrases, les fréquences lexicales. On les avait naturellement choisies surtout pour la simplicité de la procédure de dépouillement. D'autres caractéristiques, par exemple les "patterns" syntaxiques ou les chois entre les synonymes impliquent des définitions linguistiques et des procédures de dépouillement beaucoup plus sophistiquées et laborieuses; c'est pour cela que la statistique linguistique fait directement appel à la linguistique computationnelle. Il faut surtout procéder de manière heuristique: regrouper les textes à priori selon l'homogénéité présumée des facteurs de production, où "homogénéité" veut dire qu'un ou plusieurs facteurs connus sont constants; essayer après les dépouillements de regrouper les textes à postériori en classes homogènes d'après les ressemblances dans la distribution de fréquence des catégories étudiées; confronter les classes de textes ainsi obtenues à posteriori avec les classes de textes déterminées à priori.

La réalisation de ces tâches n'est pas possible sans l'emploi de procédures automatiques de dépouillement des textes, qui sont indispensables pour l'accumulation d'une base suffisante de données; elle impose, aussi, que ces procédures assument quelques caractéristiques particulières, telles que, par ex.:

- La nécessité de créer des instruments pour l'échange des textes enrégistrés en "machine readable form" entre centres différents. En d'autres mots, il est nécessaire d'unir les efforts et les ressources des chercheurs des différents pays pour créer une grande "banque de textes" internationale qui soit accessible à tous les chercheurs. Quelques propositions et réalisations à tel sujet sont décrites en Zampolli (1976).
- La nécessité de définir des critères d'analyse linguistique, qui garantissent la possibilité de confronter, entre eux, les résultats quantitatifs obtenus par les dépouillements de chercheurs différents. Cela implique, par ex., au niveau de statistique lexicale, l'emploie de normes et critères communs de lemmatisation. Ce problème a été longuement et amplement discuté au "Colloque sur l'indexation maximale" de Strasbourg, en avril 1973. Zampolli 1974 B et 1976 montrent quels sont les problèmes linguistique-théoriques et de procédure qui doivent être résolus pour atteindre le but, et ils proposent l'utilisation d'un dictionnaire de machine, opportunément construit, pour chaque langue, comme instrument principal de travail.
- L'opportunité d'étudier et mettre au point des techniques qui permettent la reconnaissance automatique ou semi-automatique des unités linguistiques des différents niveaux; pas seulement phonèmes, lemmes, leur flexion, mais aussi catégories grammaticales, groupes syntaxiques, structures syntaxiques etc.. De on propose immédiatement le recours aux récentes méthodologies produites par la linguistique computationnelle et par l'intelligence artificielle, surtout dans le domaine des "parsers" automatiques. Il semble opportune, d'ailleurs, de penser à des systèmes plus simples, du type décrit par ex. par MILIC (1976); ROSS (1972); LARA (1976); BENTEGARD (1975) etc., à cause des sous-ensembles linguistiques extrêmement bornés traités par ces "parsers", qui ont été construits surtout pour des buts théoriques et d'étude et pas pour être appliqués aux nombreuses structures et phénomènes qui caractérisent les textes

communement soumis à dépouillement dans la recherche lexicographique et statistique.

En conclusion nous pouvons affirmer que la statistique linguistique à la fois, tiré nécessairement ses matériaux des activités de traitement automatique des langues et exercite sur celles-là une fonction de stimulation qui a déjà commencé à porter ses premiers fruits.

*Tableaux*

| | C | Ta | Ra | Ga | S |
|---|---|---|---|---|---|
| C | | 613,993 | 6310,065 | 14676,763 | 18179,802 |
| Ta | 613,993 | | 3572,389 | 10369,345 | 13042,016 |
| Ra | 6310,065 | 3572,389 | | 2665,428 | 4183,161 |
| Ga | 14676,763 | 10369,345 | 2665,428 | | 1143,189 |
| S | 18179,802 | 13042,016 | 4183,161 | 1143,189 | |

Tableau 1
Valeurs de $X^2$ pour la répartition des fréquences des catégories grammaticales pour les cinq sous-ensembles de *a*.

| | C | Ta | Ra | Ga | S |
|---|---|---|---|---|---|
| C | | 43,130 | 133,517 | 202,442 | 252,861 |
| Ta | 43,130 | | 103,673 | 169,561 | 223,501 |
| Ra | 133,517 | 103,673 | | 121,034 | 165,907 |
| Ga | 202,442 | 169,561 | 121,034 | | 136,851 |
| S | 252,861 | 223,501 | 165,907 | 136,851 | |

Tableau 2
Valeurs de Z pour l'ensemble des fréquences des lemmes communs à chaque couple de sous-ensembles de *a*.

| | C | Ta | Ra | Ga | S |
|---|---|---|---|---|---|
| C | | 0,678 | 0,592 | 0,491 | 0,434 |
| Ta | 0,678 | | 0,594 | 0,509 | 0,425 |
| Ra | 0,592 | 0,594 | | 0,440 | 0,437 |
| Ga | 0,491 | 0,509 | 0,440 | | 0,476 |
| S | 0,434 | 0,425 | 0,437 | 0,476 | |

Tableau 3
Valeurs de l'indice de corrélation entre les rangs des lemmes dans les couples de sous-ensembles de *a*.

| | C | Ta | Ra | Ga | S | Ta + Ra + Ga | TOT |
|---|---|---|---|---|---|---|---|
| Subst. | 13,376 | 14,848 | 16,729 | 18,929 | 20,440 | 16,804 | 16,820 |
| Adj. | 4,277 | 4,701 | 4,958 | 6,701 | 7,070 | 5,437 | 5,522 |
| Art. | 8,208 | 10,159 | 13,794 | 16,804 | 19,286 | 13,535 | 13,580 |
| Prép. | 8,624 | 9,443 | 14,048 | 16,735 | 15,374 | 13,354 | 12,786 |
| V. | 24,038 | 21,848 | 17,713 | 14,763 | 14,911 | 18,162 | 18,722 |
| Pron. | 11,627 | 11,435 | 8,502 | 5,161 | 5,419 | 8,414 | 8,481 |
| Adv. | 9,987 | 8,711 | 6,454 | 4,869 | 3,221 | 6,707 | 6,691 |
| Conj. | 5,989 | 5,553 | 6,428 | 5,159 | 5,338 | 5,728 | 5,698 |
| Num. | 0,811 | 0,817 | 0,631 | 0,998 | 0,490 | 0,813 | 0,749 |
| Interj. | 0,778 | 0,773 | 0,131 | 0,020 | 0,017 | 0,315 | 0,350 |
| Aux. | 2,980 | 2,645 | 2,348 | 2,755 | 1,535 | 2,581 | 2,457 |
| Adv. + Prép. | 0,983 | 1,054 | 1,565 | 1,310 | 1,234 | 1,308 | 1,227 |
| Interj. + Subst. | 0,007 | 0,018 | 0,008 | 0,001 | 0,002 | 0,009 | 0,007 |
| Adj. + Pron. | 3,026 | 3,195 | 2,966 | 2,736 | 2,950 | 2,969 | 2,977 |
| Adj. + Adv. | 2,726 | 2,468 | 2,052 | 2,097 | 1,765 | 2,208 | 2,227 |
| Interj. + Adv. | 0,436 | 0,312 | 0,130 | 0,078 | 0,059 | 0,175 | 0,205 |
| Pron. + Adv. | 1,438 | 1,302 | 0,839 | 0,471 | 0,436 | 0,877 | 0,905 |
| Adv. + Conj. | 0,694 | 0,708 | 0,702 | 0,412 | 0,453 | 0,610 | 0,596 |

Tableau 4
Fréquences en pour-cent des catégories grammaticales dans *a*.

| | C | Ta | Ra | Ga | S | TOT |
|---|---|---|---|---|---|---|
| Subst. | 14,67 | 16,31 | 18,71 | 21,86 | 23,05 | 18,92 |
| Adj. | 10,23 | 10,89 | 11,06 | 12,75 | 13,06 | 11,60 |
| Art. | 8,12 | 10,02 | 13,36 | 15,88 | 18,36 | 13,14 |
| Prép. | 8,57 | 9,13 | 13,49 | 16,06 | 14,87 | 14,42 |
| V. | 26,72 | 24,19 | 20,17 | 17,22 | 16,19 | 20,90 |
| Pron. | 12,00 | 11,82 | 8,62 | 5,10 | 5,24 | 8,56 |
| Adv. | 12,69 | 11,22 | 8,23 | 6,06 | 4,24 | 8,49 |
| Conj. | 5,87 | 5,36 | 6,16 | 5,02 | 4,97 | 5,48 |
| Interj. | 1,13 | 1,06 | 0,20 | 0,065 | 0,02 | 0,49 |

Tableau 4a
Fréquence en pour-cent des catégories grammaticales dans le corpus de *a* avant la fusion de l'homographie fonctionnelle.

| | C | Ta | Ra | Ga | S | Ta + Ra + Ga | TOT |
|---|---|---|---|---|---|---|---|
| Subst. | 14,504 | 16,547 | 18,392 | 21,606 | 22,792 | 18,711 | 18,705 |
| Adj. | 4,594 | 5,178 | 6,010 | 7,555 | 8,170 | 6,244 | 6,303 |
| Art. | 7,913 | 9,711 | 12,902 | 15,400 | 17,762 | 12,745 | 12,668 |
| Prép. | 8,441 | 9,293 | 12,437 | 16,021 | 14,837 | 12,945 | 12,437 |
| V. | 23,713 | 21,666 | 17,847 | 14,655 | 14,707 | 18,060 | 18,511 |
| Pron. | 11,282 | 11,030 | 8,028 | 4,791 | 5,036 | 7,954 | 8,030 |
| Adv. | 9,385 | 8,273 | 6,257 | 4,812 | 3,073 | 6,449 | 6,355 |
| Conj. | 6,053 | 6,004 | 5,324 | 4,750 | 4,934 | 5,378 | 5,413 |
| Num. | 0,772 | 0,913 | 0,916 | 0,966 | 1,227 | 0,987 | 0,916 |
| Interj. | 0,917 | 0,935 | 0,185 | 0,037 | 0,021 | 0,386 | 0,418 |
| Aux. | 2,872 | 2,528 | 2,196 | 2,525 | 1,414 | 2,416 | 2,305 |
| Adv. + Interj. | 0,799 | 0,875 | 1,275 | 1,133 | 1,130 | 1,095 | 1,043 |
| Interj. + Subst. | 0,079 | 0,044 | 0,013 | 0,001 | 0,002 | 0,019 | 0,028 |
| Adj. + Pron. | 2,909 | 3,048 | 2,769 | 2,507 | 2,716 | 2,775 | 2,790 |
| Adj. + Adv. | 2,619 | 2,347 | 1,911 | 1,913 | 1,624 | 2,057 | 2,082 |
| Interj. + Adv. | 0,447 | 0,326 | 0,153 | 0,073 | 0,056 | 0,184 | 0,211 |
| Pron. + Adv. | 1,386 | 1,245 | 0,784 | 0,432 | 0,402 | 0,821 | 0,849 |
| Adv. + Conj. | 0,985 | 0,919 | 0,863 | 0,520 | 0,560 | 0,768 | 0,769 |

Tableau 4b
Fréquences en pour-cent des catégories grammaticales dans tout le corpus de *a*.

FRÉQUENCE EN POUR-CENT

| | CLASSE | C | Ta | Ra | S | Ga |
|---|---|---|---|---|---|---|
| 1 | 1-500 | 88,764 | 84,451 | 82,555 | 81,397 | 78,747 |
| 2 | 501-1000 | 5,272 | 6,135 | 6,763 | 8,240 | 8,217 |
| 3 | 1001-1500 | 2,532 | 3,184 | 3,728 | 4,340 | 4,845 |
| 4 | 1501-2000 | 1,481 | 1,956 | 2,466 | 2,547 | 3,080 |
| 5 | 2001-2500 | 0,979 | 1,292 | 1,713 | 1,609 | 2,015 |

FRÉQUENCE MOYENNE

| | CLASSE | C | Ta | Ra | S | Ga |
|---|---|---|---|---|---|---|
| 1 | 1-500 | 175,10 | 168,21 | 159,69 | 154,79 | 148,03 |
| 2 | 501-1000 | 10,40 | 12,07 | 13,48 | 15,67 | 15,44 |
| 3 | 1001-1500 | 51,96 | 62,26 | 7,21 | 8,25 | 9,10 |
| 4 | 1501-2000 | 29,22 | 3,85 | 4,77 | 4,84 | 5,79 |
| 5 | 2001-2500 | 1,93 | 2,54 | 3,31 | 3,06 | 3,78 |

Tableau 5
Classes de fréquences en *a*.

| | | MOITIÉ | | MOYENNE | | MÉDIANE | |
|---|---|---|---|---|---|---|---|
| | | % Sur le total des occurrences | Fréquence moyenne du lemme | % Sur le total des occurrences | Fréquenc moyenne du lemme | % Sur le total des occurrences | Fréquence moyenne du lemme |
| | C | 97,339 | 55,498 | 85,475 | 242,967 | 0,838 | 28,45 |
| I | Ta | 96,701 | 47,781 | 83,044 | 209,589 | 0,753 | 24,70 |
| | Ra | 95,967 | 43,990 | 81,186 | 179,688 | 0,592 | 22,92 |
| PARTIE | S | 96,106 | 48,049 | 79,923 | 171,948 | 0,630 | 24,99 |
| | Ga | 95,409 | 42,421 | 77,953 | 157,234 | 0,567 | 22,23 |
| | C | 2,661 | 1,517 | 14,525 | 4,603 | 99,161 | 28,51 |
| II | Ta | 3,299 | 1,630 | 16,956 | 4,663 | 99,247 | 24,70 |
| | Ra | 4,239 | 1,848 | 18,814 | 4,811 | 99,408 | 22,92 |
| PARTIE | S | 3,894 | 1,946 | 20,077 | 5,678 | 99,370 | 24,99 |
| | Ga | 4,591 | 2,041 | 22,047 | 5,509 | 99,433 | 22,23 |

Tableau 6
Moitié, Moyenne, Médiane dans *a*.

| | Tb | Rb | E | Ts | Gb |
|---|---|---|---|---|---|
| Tb | | 5762,045 | 9183,548 | 12067,696 | 12786,394 |
| Rb | 5762,045 | | 848,290 | 1868,339 | 2284,064 |
| E | 9183,548 | 848,290 | | 725,107 | 1229,993 |
| Ts | 12067,696 | 1868,339 | 725,107 | | 596,42 |
| Gb | 12786,394 | 2284,064 | 1229,993 | 596,42 | |

Tableau 7
Valeurs de $X^2$ pour la répartition des fréquences des catégories grammaticales pour les sous-ensembles de *a*.

| | Tb | Rb | E | Gb | Ts |
|---|---|---|---|---|---|
| Tb | | 121,071 | 151,050 | 193,176 | 226,817 |
| Rb | 121,071 | | 90,879 | 143,737 | 176,524 |
| E | 151,050 | 90,879 | | 78,252 | 128,986 |
| Gb | 193,176 | 143,737 | 78,252 | | 113,152 |
| Ts | 226,817 | 176,524 | 128,986 | 113,152 | |

Tableau 8
Valeurs de Z pour l'ensemble des fréquences des lemmes communs à chaque couple de sous-ensembles de *b*.

| | Tb | Rb | E | Gb | Ts |
|---|---|---|---|---|---|
| Tb | | 0,629 | 0,541 | 0,408 | 0,326 |
| Rb | 0,629 | | 0,522 | 0,319 | 0,305 |
| E | 0,541 | 0,522 | | 0,552 | 0,462 |
| Gb | 0,408 | 0,319 | 0,552 | | 0,533 |
| Ts | 0,326 | 0,305 | 0,462 | 0,533 | |

Tableau 9
Valeurs de l'indice de corrélation entre les rangs de lemmes dans les couples de sous-ensembles de *b*.

| | Tb | Rb | E | Ts | Gb | Tb + Rb + Gb | TOT |
|---|---|---|---|---|---|---|---|
| Subst. | 13,955 | 18,199 | 18,365 | 18,625 | 20,412 | 17,575 | 17,954 |
| Adj. | 4,435 | 6,269 | 8,088 | 7,560 | 7,957 | 6,250 | 6,888 |
| Art. | 10,216 | 16,030 | 16,912 | 17,988 | 18,357 | 14,935 | 15,942 |
| Prép. | 10,541 | 15,458 | 17,149 | 20,165 | 18,396 | 14,864 | 16,347 |
| V. | 21,383 | 16,050 | 12,903 | 12,732 | 12,815 | 16,678 | 15,120 |
| Pron. | 11,368 | 7,045 | 5,746 | 4,266 | 4,289 | 7,508 | 6,516 |
| Adv. | 8,136 | 4,457 | 4,149 | 4,123 | 3,471 | 5,316 | 4,844 |
| Conj. | 7,168 | 6,273 | 6,953 | 5,998 | 5,278 | 6,221 | 6,332 |
| Num. | 0,391 | 0,481 | 0,390 | 0,539 | 0,404 | 0,425 | 0,440 |
| Interj. | 0,499 | 0,076 | 0,030 | 0,003 | 0,011 | 0,191 | 0,121 |
| Aux. | 2,555 | 1,603 | 1,411 | 1,422 | 2,687 | 2,283 | 1,936 |
| Adv. + Prép. | 0,928 | 1,771 | 1,266 | 1,314 | 1,090 | 1,264 | 1,274 |
| Interj. + Subst. | 0,003 | 0,001 | 0,004 | 0,0 | 0,0 | 0,001 | 0,001 |
| Adj. + Pron. | 3,752 | 2,831 | 3,136 | 2,929 | 2,588 | 3,047 | 3,044 |
| Adj. + Adv. | 2,464 | 1,988 | 2,206 | 1,549 | 1,518 | 1,982 | 1,945 |
| Interj. + Adv. | 0,271 | 0,084 | 0,074 | 0,067 | 0,073 | 0,141 | 0,113 |
| Pron. + Adv. | 0,912 | 0,425 | 0,434 | 0,195 | 0,226 | 0,515 | 0,437 |
| Adv. + Conj. | 1,025 | 0,960 | 0,782 | 0,526 | 0,437 | 0,802 | 0,745 |

Tableau 10
Fréquences en pour-cent des catégories grammaticales dans les sous-ensembles de *b*.

FRÉQUENCE EN POUR-CENT

| | CLASSE | Tb | Ts | Rb | E | Gb |
|---|---|---|---|---|---|---|
| 1 | 1–500 | 87,102 | 83,739 | 82,863 | 80,066 | 78,582 |
| 2 | 501–1000 | 5,968 | 8,320 | 7,041 | 7,337 | 8,636 |
| 3 | 1001–1500 | 3,058 | 4,037 | 3,872 | 4,334 | 4,916 |
| 4 | 1501–2000 | 1,751 | 2,091 | 2,491 | 2,904 | 3,105 |
| 5 | 2001–2500 | 1,156 | 1,141 | 1,665 | 1,434 | 2,058 |

FRÉQUENCE EN POUR-CENT

| | CLASSE | Tb | Ts | Rb | E | Gb |
|---|---|---|---|---|---|---|
| 1 | 1–500 | 136,50 | 130,54 | 132,86 | 133,87 | 129,49 |
| 2 | 501–1000 | 9,35 | 12,97 | 11,29 | 12,26 | 14,23 |
| 3 | 1001–1500 | 4,74 | 6,29 | 6,21 | 7,26 | 8,10 |
| 4 | 1501–2000 | 2,74 | 3,26 | 3,99 | 4,85 | 5,11 |
| 5 | 2001–2500 | 1,81 | 1,77 | 1,94 | 3,41 | 3,39 |

Tableau 11
Classes de fréquences dans *b*.

| | | MOITIÉ | | MOYENNE | | MÉDIANE | |
|---|---|---|---|---|---|---|---|
| | | % Sur le total des occurrences | Fréquence moyenne du lemme | % Sur le total des occurrences | Fréquenc moyenne du lemme | % Sur le total des occurrences | Fréquence moyenne du lemme |
| | Tb | 96,633 | 46,143 | 82,545 | 211,379 | 0,859 | 24,06 |
| I | Ts | 96,158 | 49,570 | 80,191 | 166,666 | 0,628 | 25,77 |
| | Rb | 95,666 | 41,086 | 80,162 | 165,760 | 0,815 | 21,48 |
| PARTIE | E | 95,035 | 38,295 | 78,064 | 158,788 | 0,530 | 20,15 |
| | Gb | 95,005 | 39,991 | 77,023 | 145,222 | 0,510 | 21,04 |
| | Tb | 3,365 | 1,620 | 17,455 | 4,636 | 99,141 | 24,06 |
| II | Ts | 3,842 | 1,820 | 19,809 | 5,828 | 99,372 | 25,77 |
| | Rb | 4,334 | 1,861 | 19,838 | 4,755 | 99,385 | 21,48 |
| PARTIE | E | 4,965 | 2,013 | 21,936 | 4,907 | 99,470 | 20,15 |
| | Gb | 4,995 | 2,089 | 22,977 | 5,443 | 99,490 | 21,04 |

Tableau 12
Moitié, Moyenne, Médiane dans *b*.

| | C | Ta | Tb | Ra | Rb | Ga | E | S | Ts | Gb |
|---|---|---|---|---|---|---|---|---|---|---|
| C | | 613,993 | 1244,852 | 6310,065 | 11145,811 | 14676,763 | 16109,802 | 18179,802 | 19319,083 | 20006,249 |
| Ta | 613,993 | | 635,242 | 3572,389 | 7425,966 | 10369,345 | 11644,866 | 13042,016 | 14567,272 | 15044,349 |
| Tb | 1244,852 | 635,242 | | 2957,730 | 5762,045 | 8968,699 | 9183,548 | 11003,662 | 12067,696 | 12786,394 |
| Ra | 6310,065 | 3572,389 | 2957,730 | | 1234,573 | 2665,428 | 3400,450 | 4183,161 | 5064,661 | 5530,797 |
| Rb | 11145,811 | 7425,966 | 5762,045 | 1234,573 | | 1198,743 | 848,290 | 1180,391 | 1868,339 | 2284,064 |
| Ga | 14676,763 | 10369,345 | 8968,698 | 2665,428 | 1198,743 | | 1295,715 | 1143,189 | 1319,847 | 1064,314 |
| E | 16109,802 | 11644,866 | 9183,548 | 3400,450 | 848,290 | 1295,715 | | 964,967 | 725,107 | 1299,933 |
| S | 18179,802 | 13042,016 | 11003,662 | 4183,161 | 1180,391 | 1143,189 | 964,967 | | 1188,65 | 966,931 |
| Ts | 19319,083 | 14567,272 | 12067,696 | 5064,661 | 1868,339 | 1319,847 | 725,107 | 1188,65 | | 596,462 |
| Gb | 20006,249 | 15044,349 | 12786,394 | 5530,797 | 2284,064 | 1064,314 | 1229,933 | 966,931 | 596,462 | |

Tableau 13
Valeurs de $X^2$ pour la répartition des fréquences des catégories grammaticales entre les sous-ensembles de *a* et *b*.

| | C | Ta | Tb | Ra | Rb | Ga | E | Gb | S | Ts |
|---|---|---|---|---|---|---|---|---|---|---|
| C | | 43,130 | 71,779 | 133,517 | 179,629 | 202,442 | 211,129 | 241,776 | 252,861 | 269,396 |
| Ta | 43,130 | | 54,921 | 103,673 | 149,461 | 169,561 | 176,783 | 211,713 | 223,501 | 241,482 |
| Tb | 71,779 | 54,921 | | 96,790 | 121,071 | 163,471 | 151,050 | 193,176 | 200,047 | 226,817 |
| Ra | 133,517 | 103,673 | 96,790 | | 63,101 | 121,034 | 114,623 | 161,165 | 165,907 | 189,795 |
| Rb | 179,629 | 149,461 | 121,071 | 63,101 | | 126,051 | 90,879 | 143,737 | 142,822 | 176,524 |
| Ga | 202,442 | 169,561 | 163,471 | 121,034 | 126,061 | | 88,120 | 84,555 | 136,851 | 131,174 |
| E | 211,129 | 176,783 | 151,050 | 114,623 | 90,879 | 88,120 | | 78,252 | 109,401 | 128,986 |
| Gb | 241,776 | 211,713 | 193,176 | 161,165 | 143,737 | 84,555 | 78,252 | | 121,944 | 113,152 |
| S | 252,861 | 223,501 | 200,041 | 165,907 | 142,822 | 136,851 | 109,401 | 121,944 | | 158,842 |
| Ts | 269,396 | 241,482 | 226,817 | 189,795 | 176,524 | 131,174 | 128,986 | 113,152 | 158,842 | |

Tableau 14
Valeurs de Z pour l'ensemble des fréquences des lemmes communs à chaque couple de sous-ensemble de *a* et de *b*.

| | C | Ta | Tb | Ra | Rb | E | Ga | Gb | S | Ts |
|---|---|---|---|---|---|---|---|---|---|---|
| C | | 0,678 | 0,638 | 0,592 | 0,575 | 0,516 | 0,491 | 0,479 | 0,434 | 0,420 |
| Ta | 0,678 | | 0,640 | 0,594 | 0,593 | 0,549 | 0,509 | 0,499 | 0,425 | 0,436 |
| Tb | 0,638 | 0,640 | | 0,634 | 0,629 | 0,541 | 0,533 | 0,408 | 0,534 | 0,326 |
| Ra | 0,592 | 0,594 | 0,634 | | 0,678 | 0,570 | 0,440 | 0,482 | 0,437 | 0,441 |
| Rb | 0,575 | 0,593 | 0,629 | 0,678 | | 0,522 | 0,505 | 0,319 | 0,541 | 0,305 |
| E | 0,516 | 0,549 | 0,541 | 0,570 | 0,522 | | 0,592 | 0,552 | 0,572 | 0,462 |
| Ga | 0,491 | 0,509 | 0,533 | 0,440 | 0,505 | 0,592 | | 0,631 | 0,476 | 0,547 |
| Gb | 0,479 | 0,499 | 0,408 | 0,482 | 0,319 | 0,552 | 0,631 | | 0,546 | 0,533 |
| S | 0,434 | 0,425 | 0,534 | 0,437 | 0,541 | 0,572 | 0,476 | 0,546 | | 0,487 |
| Ts | 0,420 | 0,436 | 0,326 | 0,441 | 0,305 | 0,462 | 0,547. | 0,533 | 0,487 | |

Tableau 15
Valeurs de l'indice de corrélation entre les rangs des lemmes dans les couples de sous-ensemble de *a* et de *b*.

| | C | Ta | Tb | Ra | Rb | Ga | S + E + Ts | Gb | (T + R + G)a | (T + R + G)b | TOT a | TOT b |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Subst. | 13,376 | 14,848 | 13,955 | 16,729 | 18,199 | 18,929 | 19,143 | 20,412 | 16,804 | 17,575 | 16,820 | 17,954 |
| Adj. | 4,277 | 4,701 | 4,435 | 4,958 | 6,269 | 6,701 | 7,573 | 7,957 | 5,437 | 6,250 | 5,522 | 6,888 |
| Art. | 8,298 | 10,159 | 10,216 | 13,794 | 16,030 | 16,804 | 18,062 | 18,357 | 13,535 | 14,935 | 13,580 | 15,942 |
| Prép. | 8,624 | 9,443 | 10,541 | 14,048 | 15,458 | 16,735 | 17,563 | 18,396 | 13,354 | 14,864 | 12,786 | 16,347 |
| V. | 24,038 | 21,848 | 21,383 | 17,713 | 16,050 | 14,763 | 13,515 | 12,815 | 18,162 | 16,678 | 18,722 | 15,120 |
| Pron. | 1,627 | 1,435 | 1,368 | 8,502 | 7,045 | 5,161 | 5,144 | 4,289 | 8,414 | 7,508 | 8,481 | 6,516 |
| Adv. | 9,987 | 8,711 | 8,136 | 6,454 | 4,457 | 4,869 | 3,831 | 3,471 | 6,707 | 5,316 | 6,691 | 4,844 |
| Conj. | 5,989 | 5,553 | 7,168 | 6,428 | 6,273 | 5,159 | 6,096 | 5,278 | 5,728 | 6,221 | 5,698 | 6,332 |
| Num. | 0,811 | 0,817 | 0,391 | 0,631 | 0,481 | 0,998 | 0,473 | 0,404 | 0,813 | 0,425 | 0,749 | 0,440 |
| Interj. | 0,773 | 0,778 | 0,499 | 0,131 | 0,076 | 0,020 | 0,017 | 0,011 | 0,315 | 0,191 | 0,350 | 0,121 |
| Aux. | 2,980 | 2,645 | 2,555 | 2,348 | 1,603 | 2,755 | 1,456 | 2,687 | 2,581 | 2,283 | 2,457 | 1,936 |
| Adv. + Prép. | 0,983 | 1,054 | 0,928 | 1,565 | 1,771 | 1,310 | 1,271 | 1,090 | 1,308 | 1,264 | 1,227 | 1,274 |
| Interj. + Subst. | 0,007 | 0,018 | 0,003 | 0,008 | 0,001 | 0,001 | 0,002 | 0,0 | 0,009 | 0,001 | 0,007 | 0,001 |
| Adj. + Pron. | 3,026 | 3,195 | 3,752 | 2,966 | 2,831 | 2,736 | 3,005 | 2,588 | 2,969 | 3,047 | 2,977 | 3,044 |
| Adj. + Adv. | 2,726 | 2,468 | 2,464 | 2,052 | 1,988 | 2,097 | 1,840 | 1,518 | 2,208 | 1,982 | 2,227 | 1,945 |
| Interj. + Adv. | 0,436 | 0,312 | 0,271 | 0,130 | 0,084 | 0,078 | 0,066 | 0,073 | 0,175 | 0,141 | 0,205 | 0,113 |
| Pron. + Adv. | 1,438 | 1,302 | 0,912 | 0,839 | 0,425 | 0,471 | 0,355 | 0,226 | 0,877 | 0,515 | 0,905 | 0,437 |
| Adv. + Conj. | 0,694 | 0,708 | 1,025 | 0,702 | 0,960 | 0,412 | 0,587 | 0,437 | 0,610 | 0,802 | 0,596 | 0,745 |

Tableau 16
Fréquences en pour-cent des catégories grammaticales dans les sous-ensembles de *a* et de *b*.

*Notes*

1. Voyez à ce sujet ENGWALL (1974, chap. 3) et LYNE (1972). Le fait est que les dictionnaires de fréquence compilés jusqu'à présent pour la même langue sont en général compilés selon des méthodes différentes. Il n'est pas possible donc de décider si les différences sont à attribuer simplement aux méthodes de compilation, ou bien aux populations statistiques que les dictionnaires de fréquence se proposent de représenter.

Les différences méthodologiques concernent en général:

a) *les critères de composition du corpus* sur le dépouillement duquel est basé le dictionnaire. Les critères sont: 1) quantitatifs: dimensions globales du corpus; nombre des sous-ensembles dans lesquels le corpus est stratifié et leurs proportions relatives; nombre de textes différents dans chaque sous-ensemble; quantité et distribution de mots ou de phrases prises dans chaque texte; 2) qualitatifs: choix et définition des sous-ensembles; choix des textes dans les sous-ensembles.

b) la définition et l'individualisation des unités lexicales: autrement dit, étant donné qu'en général les unités de travail sont les lemmes, les critères de lemmatisation.

c) les formules qui déterminent le choix et l'organisation des unités lexicales (nous nous référerons désormais à ces unités comme à des lemmes) qui doivent être insérées dans le dictionnaire.

Il serait évidemment intéressant de faire varier une seule de ces caractéristiques et d'étudier les conséquences sur la composition du dictionnaire. Dans le cadre problématique que l'on a discuté au paragraphe précédent en particulier, il serait surtout intéressant de faire varier la composition qualitative du corpus.

2. Dans certains cas ces interventions ont été assez simples. Par exemple, dans *a* toutes les occurrences de l'article déterminatif (y compris celles qui se trouvent dans les soi-disant propositions articulées) ont été réunies en un lemme unique *il*, qui présente une fréquence globale de 54752 et occupe le premier rang. En *b*, au contraire, apparaissent trois lemmes distincts, *il*, *la*, *lo*, avec des fréquences respectives de 24333, 23663 et 6821 et les rangs 2, 3 et 9. Des deux solutions qui permettraient un rapprochement:

a) subdiviser la fréquence de *il* dans *a* pour créer trois lemmes distincts, *il*, *la*, *lo*

b) réunir dans *b* la fréquence des trois lemmes *il*, *la*, *lo*, en un seul lemme *il*.

La première est impossible parce que dans *a* apparait, sous *il*, la forme *-ll'* qui comprend les occurrences, dans les propositions articulées, soit de l'article masculin, soit de l'article féminin. (Remarquez au passage que, selon une évidente incohérence *l'* dans *a* a été au contraire distingué en masculin et féminin). On a donc choisi l'intervention (2). Ainsi l'article déterminatif est passé, dans *b* également, au premier rang avec la fréquence de 54817. Si nous calculons l'écart réduit entre les deux fréquences 54752 et 54817 nous trouvons:

$$Z = 0{,}1$$

ce qui indiquerait que la différence entre les deux fréquences a plus de neuf probabilités sur 10 pour être obtenue selon un pur effet du hasard.

Malheureusement dans de nombreux cas la solution n'a pas été aussi simple, et il a fallu fournir un long travail de "detective" sur les nomenclatures des deux dictionnaires, car souvent les critères de lemmatisation déclarés dans les introductions ne sont pas suffisamment explicites et même ils ont été contredits dans l'application. D'ailleurs c'est là une situation assez commune dans les dépouillements (cf. ENGWALL, 1974, chap. 6; KOSTER, 1970, p. 291; LYNE, 1972 et 1973) et seul l'emploi d'un lexique automatique peut y apporter remède. Une description détaillée des interventions exécutées serait trop longue et je vous renvoie à ZAMPOLLI (1974 B). Je me limite à

mentionner une classe d'interventions à laquelle je ferai référence dans la suite de mon exposition. Comme on le sait, un des points les plus controversés dans la lemmatisation, non seulement des textes italiens mais encore de nombreuses autres langues, concerne les "mots qui peuvent prendre une autre valeur syntaxique pour le passage de catégorie grammaticale" (*a*, LV), c'est-à-dire les mots que *b* appelle "syntactic homographs" et qui "are identical occurrence of members of different word classes: the same shapes, often with the same lexical meaning, are distinguished by sentence function" (*b*, XXVI). Pensez par exemple à l'emploi substantivé d'infinitifs, de participes passés et présents à l'emploi adjectival de participes passés et présents aux adjectifs substantivés et aux substantifs en fonction adjectival, aux adjectifs employés adverbialement, à l'emploi prépositionnel et adverbial de certains mots, etc.... . Tandis que *b* a en général distingué avec soin ces "homographes syntaxiques", *a* les a laissé unis sous un même lemme, auquel il a cependant attaché les deux catégories grammaticales. Ainsi nous trouvons par exemple que dans *a* le lemme *su* est accompagné du double sigle Av.PZ., le lemme *questo* du double sigle AG.PR.= Dans ces cas, très nombreux, nous sommes intervenus en réunissant également dans *b* les homographes que *a* n'avait pas distingués. Ainsi s'explique le fait que, dans les tableaux relatifs aux pourcentage des catégories grammaticales qui suivent (relatifs aussi bien à *a* qu'à *b*) apparaissent les fréquences de couples de catégories tels que *adv.* + *prép.*, *adj.* + *pron.* etc.... . Les interventions ont également réduit le nombre de lemmes en *b* de 5014 à 4505. Je renvois à ZAMPOLLI (1974 D) pour plus de détails sur ce point et sur les conséquences dans l'application des formules statistiques aux deux nomenclatures.

3. Cette correspondance a été affirmée dans un colloque qui s'est tenu en 1971 entre A. Juilland, C. Tagliavini, A. Zampolli. Elle doit naturellement être accueillie sous bénéfice d'inventaire et se trouve conditionnée par la subjectivité de la définition des sous-ensembles, qui, comme on l'a dit, attend par principe une vérification et éventuellement une correction à posteriori sur la base des résultats de l'analyse. Toutefois les données de notre dépouillement semblent confirmer, au moins dans une première instance, la correspondance susdite.

4. Pour plus de détails voyez JUILLAND et CHANG RODRIGUEZ (1964, XXVI, XXVII).

5. Nous avons employé le test du $X^2$ comme indice de ressemblance entre les distributions de fréquence des catégories grammaticales d'un couple d'ensembles.

6. Nous reproduisons ci-dessous, pour chaque sous-ensemble, les deux sous-ensembles qui, en ordre décroissant, s'avèrent "les plus semblables", par rapport au $X^2$, sur les catégories grammaticales:

C : Ta, Ra
Ta : Ca, Ra
Ra : Ga, Ta
Ga : Sa, Ra
Sa : Ga, Ra

L'observation la plus immédiate, confirmée par les données que nous exposerons par la suite, est que l'on puisse reconnaître deux blocs, d'un côté C et Ta, de l'autre S et Ga, entre lesquels se situerait Ra, qui constitue pour ainsi dire un intermédiaire entre eux.

Si nous considérons l'ordre de "ressemblance décroissante" (c'est-à-dire $X^2$ croissant) entre les couples

a) (1) C – Ta
(2) Sa – Ga
b) (3) Ra – Ga
(4) Ra – Ta
(5) Ra – S
(6) Ra – C
c) (7) Ga – Ta
(8) S – Ta
(9) Ga – C
(10 S – C

nous voyons qu'aux deux premières places se trouvent les couples qui constituent les deux blocs, suivis (3–6) des couples constitués par Ra avec les éléments des deux blocs, suivis enfin (7–10) par les couples constitués d'éléments de blocs différents.

7. C'est–à–dire que dans le calcul de $X^2$ pour un couple donné n'ont été inclus que les lemmes présents dans chacun des deux sous–ensembles du couple.

8. Tandis que les deux couples "les plus semblables" sont les mêmes que pour les catégories grammaticales (voir la note 21) l'ordre décroissant de ressemblance entre les couples est partiellement différent:
C–Ta, Ra–Ta, Ga–Ra, Ra–C, S–Ga, S–Ra, Ga–Ta, Ga–C, S–Ta, S–C.
Tandis que la ressemblance du bloc C–T reste confirmée, on a essentiellement un éloignement de S de tous les autres sous–ensembles, S restant néanmoins davantage semblable à G.

9. Comme on le sait, l'indice de Spearman mesure la corrélation entre les rangs attribués par deux séquences différentes à la même série de données. Dans notre cas l'indice a été calculé pour les dix couples possibles de sous–ensembles de *a*. Étant donné un couple de sous–ensembles *x* et *y*, le rang en *x* a été attribué de la façon suivante. Les lemmes présents dans au moins un des deux sous–ensembles du couple (c'est–à–dire qui sont présents aussi bien en *x* qu'en *y*, ou seulement en *x*, ou encore seulement en *y*) ont été ordonnés selon l'ordre des fréquences en *x*. Le rang d'un lemme en *x* est son nombre progressif dans cet ordre. Naturellement, comme d'habitude, on a attribué comme rang aux lemmes qui ont la même fréquence la moyenne de leurs nombres progressifs. Les lemmes avec fréquence zéro en *x*, c'est–à–dire qui ne sont présents qu'en *y*, se trouvent à la fin de cette séquence, et ont reçu comme rang en *x* le rang immédiatement successif à celui attribué au lemme le moins fréquent en *x*. On a procédé de la même façon pour attribuer le rang en *y*. (Le traitement des lemmes avec la fréquence m'a été suggéré par Ch. Muller).

La formule pour le calcul de $\rho$ pour un couple de sous–ensembles *x* et *y* est:

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

où *d* est la différence entre le rang d'un lemme en *x* et son rang en *y*, et *n* est le nombre de lemmes présents dans au moins un des deux sous–ensembles du couple.

Quand *n* est très grand, la capacité de signification peut être vérifiée (approximativement) en calculant

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}}$$

Toutes les valeurs se sont révélées positives et largement significatives. Les tableaux 3, 9 et 15 indiquent les valeurs de *t* qui, étant indépendantes de *n*, sont comparables entre elles.

Comme je l'ai déjà fait pour $X^2$, j'utilise *t* non pas pour déterminer s'il existe ou non une corrélation entre les rangs des deux sous-ensembles, mais pour évaluer leur plus grand ou plus petit rapprochement. L'analyse des données confirme amplement ce que l'on a dit dans la note 8.

10. A propos des pourcentages reportés sur le tableau 4, certaines considérations préliminaires s'imposent.

a) Dans le tableau 4, outre les 11 catégories grammaticales (dans l'ordre: substantifs, adjectifs, articles, prépositions, verbes, pronoms, adverbes, conjonctions, numéraux, interjections, auxiliaires) apparaissent sept couples: adverbes + prépositions, interjections + substantifs, adjectifs + pronoms, interjections + adverbes, pronoms + adverbes, adverbes + conjonctions.
La note 5 explique les raisons de ces accouplements, qui doivent être interprétés de la façon suivante: par exemple 0,983 est, dans le sous-ensemble C, le pourcentage de fréquence que l'on obtient en additionnant entre elles les fréquences de tous les lemmes pour lesquels on n'a pas fait de distinction entre l'usage adverbial et l'usage prépositionnel. On remarquera que les catégories les plus touchées sont les adverbes, les pronoms, les adjectifs et les prépositions. Pratiquement le pourcentage de 9,987% d'adverbes à l'intérieur de C est inférieur au pourcentage effectif parce qu'il ne comprend pas les occurrences des adverbes compris dans les couples adv. + prép., + adj., + int., + prop., + conj.
On peut toutefois démontrer que ce fait ne dément pas les considérations sur l'ordre des sous-ensembles. Du fait que la démonstration complète serait trop longue, on suggère de confronter le tableau 4 avec le tableau 4ba, sur lequel sont reportés les pourcentages de fréquences des catégories grammaticales recueillies dans une phase intermédiaire de l'élaboration de *a*, dans laquelle n'étaient pas encore fondus les divers usages grammaticaux des lemmes en question.

b) Les pourcentages reportés en 4a (comme en 12 et 12a) se réfèrent à l'ensemble des occurrences correspondantes aux seuls lemmes acceptés en *a* et en *b*, et non aux deux corpus respectifs. Le tableau 4b reporte les pourcentages sur le corpus entier de *a*. Comme on le voit, les variations entre les tableaux 4a et 4b sont minimes, et consistent, comme il était logique de l'attendre, en une légère augmentation des substantifs (un peu moins de 22%), des adjectifs (un peu moins de 1%), des interjections (0,2% environ), c'est-à-dire des catégories auquelles appartiennent généralement les lemmes qui n'ont pas atteint la valeur de U nécessaire pour être admis dans *a*. Toutefois, on peut aisément constater que ces petites variations n'influent absolument pas sur les considérations relatives à l'ordre des sous-ensembles. Même si les occurrences du corpus qui ne sont pas représentées en *b* (environ 80.000) sont plus nombreuses que celles qui sont exclues de *a* (environ 20.000), il apparait juste de juger que cette considération vaille également pour *b*. On peut alors analyser plus en détail l'ordre des sous-ensembles (C, Ta, Ra, Ga, S) selon lequel, comme on l'a dit:
1) les pourcentages des substantifs, adjectifs, articles et prépositions croissent;
2) les pourcentages des verbes, pronoms, adverbes, interjections et auxiliaires décroissent.

(Il faut dire que des deux catégories qui ne suivent pas cet ordre, les numéraux et les conjonctions, cette dernière peut être décomposée en deux sous-catégories: les conjonctions de coordination qui se comportent comme 1 et les conjonctions de subordination qui se comportent comme 2). On note quatre exceptions sur 45 cases.

Trois d'entre elles concernent l'ordre de Ga et de S, qui se trouve inverti par des prépositions, des verbes, des pronoms. La quatrième exception concerne la fréquence des auxiliaires dans les journaux, qui se situe entre C et Ta. En ce qui concerne les sept couples, deux d'entre eux sont constitués par des catégories de type 2; quatre sont mixtes: deux (adj. + pron. et adj. + adv.) se comportent comme le type 2, les deux autres sont irrégulières, tout comme le septième couple (pron. + adv.). Nous observons en outre que toutes les exceptions tombent si nous considérons les trois blocs dont nous avons parlé dans la note 10. Pour une analyse des possibles interprétations de ce type de régularité: BESEMANN (1925); BODER (1940); SCALISMANN (1948); MILLER (1951); ANTOSCH (1968); MARKWORTH, BELL (1967); LEVISON et alii (1968); BRANDWOOD (1969); MICHAELSON, MORTON (1972); KOLLMANN (1973); CASTROGIOVANNI (1973); DYER (1973) et leur discussion en ZAMPOLLI (1974 B) qui est trop longue pour être reprise ici.

11. Les classes de fréquence sont délimitées en faisant la liste des lemmes en ordre décroissant de fréquence et en les divisant en groupes de 500. Pour les cinq premières classes de fréquence (nous avons ommis les classes successives uniquement en raison du manque de place) le tableau 5 reporte la fréquence moyenne des m lemmes de la classe et leur pourcentage de fréquence par rapport au total des occurrences dans le sous-ensemble respectif. Comme on le voit, par rapport à l'ordre C, Ta, Ra, S, Ga, d la fréquence moyenne et le pourcentage de fréquence de al la première classe décroissent, tandis que celles des classes successives croissent (ainsi que celles des classes qui ne sont pas reportées sur le tableau); la seule inversion se situe entre S et Ga dans la deuxième classe). Dans l'ensemble, le bloc "dramatique" présente une plus grande concentration des occurrences en correspondance avec les lemmes les plus fréquents, mais, en revanche, une richesse inférieure de vocabulaire.

12. Soient les $x$ lemmes d'un sous-ensemble, ordonnés par fréquence décroissante, et soit $y$ le total des occurrences du sous-ensemble:

- La première partie de la moitié se réfère aux x/2 premiers lemmes
- La première partie de la moyenne se réfère aux lemmes qui ont une fréquence supérieure à y/x
- La première partie de la médiane se réfère aux lemmes qui correspondent aux y/2 premières occurrences
- La deuxième partie de la moitié, de la moyenne et de la médiane se réfèrent respectivement aux autres x/2 lemmes, aux lemmes qui ont une fréquence inférieure à y/x et aux lemmes correspondants aux y/2 autres occurrences.

Comme on voit, le pourcentage de fréquence sur le total des occurrences et la fréquence moyenne des lemmes décroissent selon l'ordre C, Ta, Ra, S, Ca dans les premières moitiés et croissent dans les secondes moitiés selon le même ordre.

Les inversions concernent Ra et S (pourcentage et fréquence, moyenne de moitié et médiane dans la première partie et de la médiane dans la seconde partie; pourcentage de la moyenne dans la seconde partie) et S et Ga (fréquences moyennes dans la seconde partie). Ce fait pourrait être interprété comme une indication du caractère pas complètement défini du sous-ensemble "subsidiaires", qui est un mélange de divers "modes d'exposition": description, narrative, exercices, etc.

13. En ce qui concerne les pourcentages de fréquence des catégories grammaticales, les mêmes groupements que ceux que l'on observe en *a* sont valables. Les faits suivants sont cependant à noter. Un maximum en E pour les adjectifs; trois inversions entre Ts et Gb (prépositions, adverbes et pronoms: que l'on se souvienne du phénomène analogue

relevé pour *a* entre S et Ga) et en outre, de façon analogue à *a*, la fréquence des auxiliaires se situe entre celle de Tb et celle de Rb. Les numéraux sont, comme dans *a*, irréguliers, tandis que les conjonctions décroissent, comme le font en *a* les conjonctions de subordination.

14. Dans ce cas également, Tb présente des analogies avec S.

15. Nous pouvons faires certaines observations sur le tableau 16. Avant tout, dans les couples de sous-ensembles de *a* et *b* ayant la même étiquette (T, R, G), le sous-ensemble de *a* se situe à gauche de celui de *b*. Si nous acceptions l'hypothèse qui dérive des discours de Antosch, soit que l'ordre en question reflète un éloignement progressif des caractéristiques du langage parlé, nous pourrions avancer l'interprétation que, avec le temps, les caractéristiques des trois sous-ensembles se sont modifiées et rapprochées de celles du langage parlé. Ce fait est surtout important pour les journaux: tandis que Ga se situe immédiatement à droite des romans et à gauche de E, S, Ts, Gb se situe à droite de ces derniers. En outre, si l'ordre de C, T, R est constant et rigide, E, S et Ts présentent diverses inversions. Dans la perspective de Antosch, ce fait pourrait être interprété comme un indice du fait que dans les trois sous-ensembles opèrent des tendances contradictoires, ou des tendances semblables mais d'intensité diverse. On peut relever, encore plus simplement que la définition à priori de E, S, Ts n'est pas très claire.

16. Pour l'italien, au corpus du a) l'on est en train d'ajouter des textes complementaires de nature diverse: Par exemple l'on a enregistré du "parlé" (75.000 mots divisés en cinq sous-ensembles: conversations dans la famille, conversations chez un coiffeur, conversations telephoniques transmises à la radio, débats cultivés à la television, interviews sportives à la television) et l'on en a tiré les fréquences des phonèmes, des syllabes, des mots (v. MARIONI, 1975). 500.000 mots des principaux journaux italiens sont en train d'être dépouillés, divisés en sous-ensembles du type sport, politique, etc. Pour le français l'on est en train d'étudier le *Dictionnaire de Fréquences* du *Trésor de la Langue Française* de Nancy qui fournit, pour 70.000 vocables, des données quantitatives fondées sur le dépouillement d'un corpus à dominante littéraire, comportant plus de 70 millions de mots, et soumis à plusieurs découpages: par chronologie, par catégories stylistiques, etc. Voir les travaux cités dans MULLER (1976).

17. Par exemple dans a:

| | Cinéma | Théâtre | Romans | Journaux | Subs. |
|---|---|---|---|---|---|
| Il (art.) | 5930 | 7730 | 10892 | 13453 | 16747 |
| Di (prep.) | 3021 | 3434 | 5132 | 7044 | 6770 |
| Da (prep.) | 692 | 729 | 1070 | 1249 | 1400 |
| Fra (prep.) | 37 | 42 | 64 | 108 | 113 |
| In (prep.) | 1014 | 1216 | 2020 | 2420 | 2181 |
| Tra (prep.) | 40 | 48 | 84 | 95 | 122 |

18. En particulier, au niveau lexicale, l'on dispose des travaux de Ch. Muller, qui a formulé très clairement les rapports entre le lexique et le vocabulaire: il réserve le premier de ces termes à la "langue", le deuxième au "discours". Dans cette terminologie les unités qui composent le lexique sont des lexèmes; quand ces unités virtuelles sont actualisées dans le discours, chacune d'elles est un vocable auquel correspond un certain nombre d'occurrences dans le texte. On définit donc comme vocabulaire d'un texte l'ensemble

des vocables présents dans le texte considéré, chacun de ces vocables ayant une fréquence propre. Le vocabulaire d'un texte suppose l'existence d'un lexique, qui dépasse le texte dont le vocabulaire n'est qu'un échantillon. On peut concevoir une série d'ensembles lexicaux, dont chacun contient le suivant; le lexique d'un idiome, dans le sens extensif, avec quelques limites de temps; le lexique d'une langue, considéré synchroniquement; le lexique d'un groupe humain limité; le lexique d'un individu de ce groupe; le lexique de l'individu dans une situation définie du point de vue stylistique et thématique.

La notion de lexique de situation comprend deux types d'éléments: les premiers sont d'ordre stylistique et sont liés à l'interlocuteur et à l'effet que le locuteur veut produire; les seconds sont d'ordre thématique et sont liés à ce que le locuteur veut communiquer, au contenu du message.

"On pose en principe que seule la situation détermine une *probabilité d'emploi* pour chacun des éléments du lexique, probabilité qui se manifeste par la fréquence de l'élément dans le texte. L'idée d'une fréquence ou d'une probabilité "en langue", attachée à chaque lexème indépendamment de tout discours, ne résiste pas à l'examen, et encore moins à l'expérience" (MULLER, 1975).

La situation provoque, à l'intérieur du lexique individuel "un déplacement des probabilités d'emploi"; certaines unités sont exclues du discours et tombent à la probabilité zéro.

"En d'autres termes, le lexique de situation $L_s$ est un sous-ensemble du lexique individuel $L_{in}$, et son effectif $L_s$ est inférieur à celui du lexique individuel, $L_{in}$:

$$L_s \subset L_{in} \qquad L_s < L_{in}$$

Il est clair que la stabilité absolue d'une situation et du lexique qu'elle détermine est une vue théorique. Tout discours d'une certaine étendue entraîne des variations dans le thème, parfois dans le style. Il faut alors considérer que le lexique "en jeu" dans la totalité du discours, $L_d$ est formé par la réunion de plusieurs lexiques de situation: $L_{s1}$, $L_{s2}$,... $L_{sn}$, et que son effectif est égal ou supérieur à chacun de leurs effectifs:

$$L_d = L_{s1} \cup L_{s2} \cup \ldots \cup L_{sn} \qquad L_d \geqslant L_{s1}, L_{s2}, \ldots, L_{sn}."$$

(MULLER, 1975, p. 5)

On peut donc admettre qu'au moment de la rédaction d'un texte ou de la conception d'un énoncé, l'auteur a un certain nombre d'unités de son lexique en jeu, tandis que les autres sont exclues pour des raisons stylistiques et thématiques.

"On admettra donc qu'à tout discours, matérialisé par un texte, est lié un lexique; que le vocabulaire du texte est un sous-ensemble du lexique; que le complément de ce sous-ensemble, formé de tous les lexèmes non actualisés dans le texte, n'est pas vide" (MULLER, 1975, p. 4).

Muller a proposé recemment (MULLER, 1975, 1976), sur la base des résultats obtenus par M. Dolphin dans la modification de la "loi de Waring-Herdan", un modèle qui permettrait une évaluation quantitative acceptable du lexique en jeu, c'est-à-dire, de la partie virtuelle, mais non actualisée, du lexique.

Des élaborations comparables manquent pour les autres niveaux linguistiques, mais il faut rappeler ROSS (1976) pour le début de l'étude d'un modèle pour les catégories grammaticales.

*Bibliographie*

ABRAHAM S., KIEFER F. (1965) – *An algorithmic definition of the morpheme*, in <<Statistical Methods in Linguistics>>, IV, 4–9.
ACCADEMIA DELLA CRUSCA (1967) – *Norme per la schedatura lessicografica*, Firenze.
ACCADEMIA DELLA CRUSCA (1967) – *Concordanze degli Inni Sacri di A. Manzoni*, Firenze.
ACCADEMIA DELLA CRUSCA (1968) – *Novella del Grasso Legnaiuolo. Testo. Frequenze. Concordanze*, Firenze.
ACCADEMIA DELLA CRUSCA (1973) – *Tavola rotonda sui grandi lessici storici*, Firenze.
AITKEN (1972) – *Historical dictionaires, word frequency distributions, and the computer*, dans A. Zampolli (1974 b).
AKADEMIJA NAUK SSSR (1968) – Naučnyj Sovet po Kibernetike, Sekcija Semiotiki, *Statistika reči*, Leningrad, e traduzione italiana: Accademia delle Scienze dell'URSS. Consiglio scientifico per la Cibernetica, Sezione di Semiotica, *Statistica linguistica*, con l'aggiunta di due appendici. Traduzione dal russo di M. Cârstea, Bologna, 1971.
AKHMANOVA O.S. (1969) – *Linguistics and the quantitative approach*, in *Actes du X.e Congrès international des linguistes*, Bucarest.
AKHMANOVA O.S., MEL'CHUK I.A., FRUMKINA R.M., PADUCHEVA E.V. (1963) – *Exact methods in linguistic research*, translated from Russian by D.G. Hays and D.V. Mohr, Santa Monica (California).
ALIPRANDI G. (1940) – *Frequenze dattilografiche*, in <<Bollettino dell'Accademia Italiana di studi stenografici di Padova>>, XV.
ALLEN J. (1973) – *Speech Synthesis from Unrestricted Text*, Report, Research Laboratory of Electronics, MIT, Cambridge (Mass.).
ALLEN S. (1971) – *Frequency Dictionary of Present-Day Swedish*, Stockholm.
ANTOSCH F. (1969) – *The Diagnosis of Literary Style with the Verb-Adjective Ratio* dans Doležel, Bailey (1969), 57–65.
APOSTEL L., MANDELBROT B., MORF A. (1957) – *Logique, langage, et théorie de l' information*, Paris.
ASTRAHAN M. (1970) – *Speech Analysis by clustering or the hyperphoneme method*, AI Memo 124, Stanford (Calif.).
ATAL B.S. (1971) – *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave*, in IEEE Trans. on Computers.
ATTI DEL CONVEGNO (1968) – *Sul tema: L'Automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali (Accademia dei Lincei, Roma*, 1967), Roma.
BACH E. (1968) – *Nouns and Noun Phrases*, in Bach E., Harms R.T. (1968), 91–122.
BAILEY R.W. (1969) – *Statistics and Style: A Historical Survey*, in Doležel, Bailey (1969), 217–236.
BAILEY R.W., DOLEZEL L. (1968) – *An annotated bibliography of statistical stylistic*, Ann Arbor (Mich.).
BAHR J. (1972) – *The lexical Data Bank*, in A. Zampolli (1974 b).
BARONOFSKY S. (1970) – *Some heuristics for automatic detection and resolution of anaphora in discourse*, MA Thesis, Austin (Texas).
BARTOLETTI COLOMBO A.M. (1973) – *Per un vocabolario delle Costituzioni di Giustiniano. Premessa e saggio d' elaborazione*, Firenze.
BELEVITCH V. – *Langage des machines et langage humain*, Bruxelles.
BLOOMFIELD L. (1933) – *Language*, New York.
BOBROW D.G., FRASER I.B. (1969) – *An Augmented State Transition Network Analysis Procedure*, in L.M. Norton, D. Walker (eds.), *Proceedings of the International Joint Conference on Artificial Intelligence*, Bedford (Mass.), 557–567.
BOBROW D.G., KLATT D.H. (1968) – *A Limited Speech Recognition System*, in Proc. AFIPS Fall Joint Computer Conference, Thompson, Washington, D.C., 33, 305–318.
BODER D.P. (1940) – *The Adjective-Verb quotient; a contribution to the psychology of language*, dans <<Psych. Rec.>>, 3.
BOISVERT S., DUGAS A., BELANGER D. (1974) – *Obling: a Tester for Transformational Grammars*, in A. Zampolli (1974 a).
BOLDRINI M. (1958) – *Statistica, teoria e metodi*, Milano.

BOLINGER D. (1965) – *The atomization of meaning*, in <<Language>>, 41, 4.
BOLOCAN G.M. (1961) – *Unele caracteristici ale stilului publicistic al limbii romane literare*, dans <<Studi si cercetări lingvistice>>, XII, 35–71.
BOOTH A.D., CLEAVE J.P., BRANDWOOD L. (1958) – *Mechanical Resolutions of Linguistic Problems*, London.
BORKO H. (1967) (Ed.) – *Automated Language Processing*, New York.
BORTOLINI U., TAGLIAVINI C., ZAMPOLLI A. (1971) – *Lessico di frequenza della lingua italiana contemporanea*, IBM Italia.
BOTHA R.P. (1968) – *The function of the lexicon in transformational generative grammar*, The Hague.
BRANDWOOD L. (1969) – *Plato's Seventh Letter*, in <<Revue>>, IV, 1–25.
BROAD D.J. (1972) – *Basic directions in automatic speech recognition*, in <<Int. Journal of manmachine studies>>, 4, 105–118.
BUCHANAN M.A. (1972) – *A graded Spanish word book*, Toronto (Ont.).
BUSA R. (1968) – *De lexico electronico latino*, Pisa.
BUSA R., ZAMPOLLI A. (1968) – *Centre pour l'Automation de L'Analyse Linguistique (C.A.A.L), Gallarate*, in *Les machines dans la linguistique*, Prague, 25–34.
BUSEMANN A. (1925) – *Die Sprache der Jugende als Ausdruck der Entwicklungsrhythmik*, Jena.
BUSEMANN A. (1948) – *Stil und Charakter: Untersuchungen zur Psycologie der individuellen Redeform*, Meisenheim–Glan.
CARROLL J.B. (1970) – *An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI)*, in <<Computer studies in the humanities and verbal behavior>>, III, 2, 61–65.
CARROL J.B. (1973) – *Towards a Performance Grammar for Core Sentences in Spoken and Written English*, Princeton.
CASTROGIOVANNI P., CERRI S.A., MAFFEI G., PASQUINUCCI P.J., TORRIGIANI G., ZAMPOLLI A. (1968) – *Analisi Linguistica delle risposte al test di Rorschach di schizofrenici e neurotici e dei rispettivi familiari*, in <<Neopsichiatria>>, XXX, IV (1968), 810–837.
CASTROGIOVANNI P., TELARA A. (1973) – *Primi risultati di un' analisi statistica morfologica e lessicale delle risposte al test di Rorschach nelle prospettive di uno studio dei rapporti tra psicopatalogia e linguaggio*, in Zampolli (1973 a), 307–323.
CELCE–MURCIA M. (1972 a) – *Paradigms for Sentence Recognition*, Los Angeles (prepr.).
CELCE–MURCIA M. (1972 b) – *English Comparatives*, Ph.D. Thesis, UCLA, Los Angeles.
CHAFE W.L. (1970) – *Meaning and the Structure of Language*, Chicago.
CHOMSKY N. (1957) – *Syntactic Structures*, The Hague, trad. it. Bari, 1970.
CHOMSKY N. (1958) – Recensione a Belevitch (1956).
CHOMSKY N., MILLER G.A. (1963) – *Finitary Models of Language Users*, in R.D. Luce, R.R. Bush, E. Galanter, *Handbook of Mathematical Psychology*, New York. Trad. it. in Chomsky (1969), *L'analisi formale del linguaggio*, Torino.
CHOMSKY N. (1965) – *Aspects of the Theory of Syntax*, Cambridge (Mass.).
CHOMSKY N. (1966) – *Topics in the Theory of Generative Grammar*, The Hague.
CHOMSKY N. (1971) – *Deep Structure, surface structure, and semantic interpretation*, in D.D. Steinberg, L.A. Jakobovits (1971), 183–216.
CIAMPI C. (1973) – *Les projets de recherche automatique des informations juridiques dans l'Institut pour la documentation juridique du Conseil National des Recherches*, in Zampolli (1973 a), 249–268.
CIRESE A.M. – *Inventaires et répertoires lexicaux, formulaires et métriques des chants populaires italiens*, in Zampolli (1973 a), 209–239.
COHEN M. (1935) – Recensione a ZIPF (1935), in <<BSL>>, XXXVI, 8.
COHEN M. (1967) – *Sur l'histoire de la statistique en linguistique*, dans <<Études de linguistique appliquée>>, 5, 3–8.
COLBY K.M., SCHANK R. (1973) (Eds.) – *Computer models of thought and language*, San Francisco (California).
COLES L.S. (1969) – *Talking with a robot in English*, in D.E. Walker, L.M. Norton (1969) (eds.), *Proceedings of the International Joint Conference on Artificial Intelligence*, Boston.
COLES S.L. (1971) – *Techniques for Information Retrieval Using an inferential Question–Answering System with natural Language Input*, Technical Note, ARI, Menlo Park (Calif.).

COLLINS A., QUILLIAN R.M. (1969) – *Retrieval Time from Semantic Memory*, in <<J. Verb. Learn. and Verb. Behav.>>, 8, 240–47.
COLMERAUER A., KITTREDGE R., STEWART G. (1970) – *Résultats Préliminaires. Project de Traduction Automatique*, Université de Montreal.
CONWAY M.E. (1963) – *Design of a Separable Transition-Diagram Compiler*, in <<CACM>>, VI, 7, 396–408.
CORBEIL J. (1971) – *Les Structures syntaxiques du français moderne*, Paris.
CROCETTI C., Studio statistico della lingua dei giornali, (Thèse non publiée), Pisa, 1976.
DENOOZ J. (1968) – *Programmes d'ordinateur pour la détermination et le traitement statistique des fréquences de phonèmes*, in <<Revue>>, III, 41–53.
DE MAURO T. (1970 A) – *Storia linguistica dell'Italia unita*, Nuova edizione riveduta, aggiornata e ampliata, Bari.
DE MAURO T. (1970 b) – *Introduzione alla Semantica*, Bari.
DE MAURO T., POLICARPI G. (1971) – *Ricerche sulla struttura del periodo italiano*, in SLI, *L'insegnamento dell'italiano in Italia e all'estero*, Roma, 583–694.
*Dictionnaire de fréquences. Vocabulaire littéraire des XIX et XX.e siècles*, Préf. par P. Imbs, Paris, 1971.
DIMITRESCU F. (1973) – *Projet d'un dictionnaire de la langue roumaine du XVI.e siècle*, in Zampolli (1973 a), 41–48.
DI SPARTI A. (1973) – *Analisi sintattica automatica delle lingue naturali*, Palermo.
DOLEŽEL L. (1969) – *A Framework for the statistical Analysis of Style*, in Doležel, Bailey (1969), 10–25.
DOLEŽEL L., BAILEY R.W. (1969) (Eds.) – *Statistics and Style*, New York.
DURO A., ZAMPOLLI A. (1968) – *Analisi lessicali mediante elaboratori ellettronici*, dans *Atti del Convegno sul Tema: L'elaborazione elettronica...* (1968).
DURO A. (1973) – *Elaborations électroniques de textes effectuées par l'Accademia della Crusca, pour la préparation du Dictionnaire Historique de la langue italienne*, dans Zampolli (1973 a), 53–75.
DYER R. (1973) – *The measurement of individual style*, in Zampolli (1973 a), 325–348.
EARLEY J. (1970) – *An efficient Context-free parsing algorithm*, <<CACM>>, XIII, 2, 94–102.
EBELING C.L. (1960) – *Linguistic units*, 's-Gravenhage.
ENGWALL G. (1972) – *A. Juilland, B. Brodin, C. Davidovitch. Frequency dictionary of French words* (recensione), dans <<Studia Neophilologica>>, 44, 2, 445–462.
ENGWALL G. (1974) – *Fréquence et distribution*, Stockholm.
ESTOUP J.B. (1912–1916) – *Gammes sténographiques*. Recueil de textes choisis pour l'acquisition méthodique de la vitesse, précédé d'une introduction, Paris (la prima edizione è del 1907).
FAEDO A. (1973) – *Discorso di Apertura*, dans Zampolli (1973 a), VII–IX.
FEIGENBAUM E.A. (1969) – *Artificial Intelligence: Themes in the second decade*, in *Proceedings of the International Conference on Information Processing*, New York.
FILLMORE C.J. (1968) – *The Case for Case*, in E. Each, R.T. Harms (1968), 1–88.
FRANCIS W.N., RUBIN G.M., SVARTIVIK J. (1969) – *A method of computer-produced graphical representation of dialectal variation in initial fricatives in Southern British English*, Preprint for the 3th ICCL.
FRIEDMAN J. (1969) – *Applications of a Computer System for Transformational Grammar*, Preprint for the 3th ICCL.
FRY D.B. (1973) – *Speech and Language*, preprint, Padova.
GARCIA HOZ V. (1953) – *Vocabulario usual, común y fundamental*, Madrid.
GARDIN J. (1965) – *A Typology of Computer Uses in Anthropology*, in D. Hymes (ed.), *The Use of Computers in Anthropology*, The Hague, 105–117.
GARVIN P.L. (1962) – *Computer Participation in Linguistic Research*, in <<Language>>, XXXVIII, 4, 385–389.
GEENS D., – *On measurement of lexical differences by means of frequency*, Louvain, 1976.
GODEL R. (1948) – *Homonymie et identité*, dans <<Cahiers F. de Saussure>>, VII, 5–15.
GOLDMAN N.M. (1974) – *Sentence Paraphrasing from a Conceptual Base*, in A. Zampolli (1974 a).
GOLOPENTIA-ERETESCU S. (1966) – *La structure phonologique des monosyllabes roumains*, in <<Cahiers de Linguistique théorique et appliquée>>, III, 59–67.

GOUGENHEIM G., MICHEA R., RIVENC P., SAUVAGEOT A. (1964) – *L'élaboration du franc,ais fondamental*, Paris.
GOUGENHEIM G., RIVENC P., M.me HASSAN (1964) – *Le français fondamental*, dans <<Tendances nouvelles en matière de recherche linguistique>>, Strasbourg, 53–72.
GRASSI C. (1973) – *Perspectives de l'emploi de l'élaborateur électronique en géographie linguistique et en dialectologie*, in Zampolli (1973 a), 233–239.
GRICE H.P. (1968) – *Utterer's meaning, sentence–meaning and word meaning*, in <<Foundations of Language>>, IV, 225–242.
GRICE H.P. (1957) – *Meaning*, in <<Philosophical Review>>, LXVI, 377–388.
GRIFFITHS T., PETRICK S. (1965) – *On the relative Efficiencies of Context–free Grammar Recognizers*, in <<CACM>>, VIII, 5, 289–300.
GROSS M. (1972) – *Mathematical Models in Linguistics*, Prentice–Hall.
GUILBERT L. (1967) – *Le dictionnaire du français contemporaine*, <<Cahiers de Lexicologie>>, X, 1, 115–119.
GUIRAUD P. (1954 a) – *Language et communication. Le substrat informationnel de la sémantisation*, <<Bulletin de la Societé de Linguistique de Paris>>, L, 119–133.
GUIRAUD P. (1954 b) – *Les caractères statistiques du vocabulaire*, Paris.
GUIRAUD P. (1960) – *Problèmes et méthodes de la statistique linguistique*, Paris.
GUIRAUD P. (1967) – *Structures étymologiques du Lexique Français*, Paris.
GUIRAUD P., WHATMOUGH J. (1954) – *Bibliographie de la statistique linguistique*, Utrecht.
HALLIDAY M.A.K. (1966) – *The English verbal group: A specimen of a manual of analysis*, London.
HALLIDAY M.A.K. (1970) – *Functional diversity in language as seen from a consideration of modality and mood in English*, in <<Foundations of Language>>, VI, 322–361.
HARKIN D. (1957) – *The history of word counts*, <<Babel>>, III, 113–124.
HARRIS Z.S. (1957) – *Co–occurrence and Transformations in Linguistic Structure*, in <<Language>>, XXXIII, 283–340.
HARRIS Z.S. (1959) – *Computable Syntactic Analysis: The* 1959 *Computer Science Analyzer*, in <<TDAP>>, 15.
HARRIS Z.S. (1961) – *String Analysis of Sentence Structure*, The Hague.
HARRIS Z.S. (1968) – *Mathematical Structures of Language*, New York.
HARRIS Z.S. (1970) – *Two Systems of Grammar: Report and Paraphrase*, in Z.S. Harris, *Papers in Structural and Transformational Linguistics*, Dordrecht.
HAYES C.W. (1968) – *A Transformational–Generative Approach to Style: Samuel Johnson and Edward Gibbon*, <<Language and Style>>, I.
HAYES C.W. (1969) – *A Study in Prose Styles: Edward Gibbon and Ernest Hemingway*, dans L. Doležel, R.W. Bailey (Eds.) (1969), 80–94.
HAYS D.G. (Ed.) (1969) – *Readings in Automatic Language Processing*, New York.
HAYS D.G. (1967) – *Introduction to Computational Linguistics*, New York.
HAYS D.G. (1969) – *Computational Linguistics: Introduction*, in *Encyclopaedia of Linguistics, Information and Control*, Pergamon Press, 49–51.
HEGER K. (1969) – *La sémantique et la dichotomie de langue et parole. Nouvelles contributions à la discussion sur les bases théoriques de la sémasiologie et de la onomasiologie.* <<Travaux de Linguistique et Littérature>>, VII, 47–111.
HEIDORN G.E. (1972) – *Natural Language Inputs to a Simulation Programming System*; Naval Postgraduate School (NPS–55 MD 72101 A), Montery (Calif.).
HEILMANN L. (1962–1963) – *Considerazioni statistico–matematiche e contenuto semantico*, <<Quaderni dell'Istituto di Glottologia>> (Università di Bologna), VII, 34–45.
HENISZ–DOSTERT B., THOMPSON F.B. (1974) – *The REL system and REL English*, in Zampolli (1974 a).
HENNON V.A.C. (1924) – *A French word book based on a count of* 400.000 *running words*, Madison.
HERDAN G. (1956) – *Language as Choice and Chance*, Groningen.
HERDAN G. (1960) – *Type–token mathematics. A textbook of mathematical linguistics*, The Hague.
HERDAN G. (1964 a) – *Quantitative Linguistics*, London.
HERDAN G. (1964 b) – *Quantitative Linguistics or Generative Grammar*, <<Linguistics>>, 4, 56–65.
HERDAN G. (1966) – *The advanced Theory of Language as Choice and Chance*, Berlin–Heidelberg–New York.

HJELMSLEV L. (1953) – *Prolegomena to a Theory of Language*, Bloomington (Ind.).
HJELMSLEV L. (1957) – *Pour une sémantique structurale*, in <<Travaux du Cercle Linguistique de Copenhague>>, XII, 96–112.
HOCHETT C.F. (1955) – *A manual of Phonology*, Bloomington (Ind.).
HOCKETT C.F. (1966) – *The Quantification of Functional Load: A Linguistic Problem*, Santa Monica (California).
HOCKETT C.F. (1967) – *Language, Mathematics, and Linguistics*, The Hague.
HOUSEHOLDER F.W., SOPORTA S. (1967) – *Problems in Lexicography*, Bloomington (Ind.).
HOWE W.G., KRULEE G.K. (1971) – *A logic of English Questions*, RC 3490, IBM, Yorktown Heights.
JACKENDOFF R.S. (1972) – *Semantic Interpretation in Generative Grammar*, Cambridge (Mass.).
JONES L.V., WEPMAN J.M. (1966) – *A Spoken Word Count*, Chicago.
JOSSELSON H.H. (1953) – *The russian word count and frequency analysis*, Detroit.
JUILLAND A., CHANG-RODRIGUEZ E. (1964) – *Frequency Dictionary of Spanish Words*, The Hague.
JUILLAND A., EDWARDS P.M.H., JUILLAND I. (1965) – *Frequency Dictionary of Rumanian Words*, The Hague.
JUILLAND A., BRODIN D., DAVIDOVITCH C. (1970) – *Frequency Dictionary of French Words*, The Hague.
JUILLAND A., TRAVERSA V. (1973) – *Frequency Dictionary of Italian Words*, The Hague.
KAEDING E.W. (1898) – *Häufigkeitswörterbuch der deutschen Sprache*, Steglitz.
KAPLAN R.M. (1971) – *Augmented Transition Networks as Psychological Models of Sentence Comprehension*, in *Second International Joint Conference on Artificial Intelligence*, London, 429–443.
KATZ J.J. (1972) – *Semantic Theory*, Harper and Row, Publishers.
KATZ J.J., FODOR L.A. (1963) – *The structure of a semantic theory*, in <<Language>>, XXXIX, 2, 170–210.
KATZ J.J., POSTAL P.M. (1964) – *An Integrated Theory of Linguistic Descriptions*, Cambridge (Mass.).
KAY M. (1967 a) – *Standards for Encoding Data in a Natural Language*, in <<Computers and the Humanities>>, I, 5, 170–177.
KAY M. (1967 b) – *Experiments with a Powerful Parser*, RM-5452 PR, Rand Corporation, Santa Monica (Calif.).
KASMER A. (1972) – *The Book of Isaiah: Characterization of Authors by Morphological Data Processing*, in <<Revue>>, III, 1–62.
KENNISTON H. (1933) – *A basic list of spanish words and idioms*, Chicago.
KERCKHOFFS A. (1883) – *La cryptographie militaire*, Paris.
KIEFER E. (1964) – *Some aspects of mathematical models in linguistics*, in <<Statistical Methods in Linguistics>>, III, 8–26.
KIEFER F. (1968) – *Mathematical linguistics in Eastern Europe*, New York.
KIMBALL J. (1973) – *Formal Theory of Grammar*, Prentice-Hall.
KNEASE T.M. (1933) – *An Italian Word List from Literary Sources*, Toronto.
KNUTH D.E. (1968) – *Semantics of Context-free Languages*, <<Mathematical Systems Theory>>, II, 2, 127–145.
KÖSTER P. (1970) – *Words and numbers. A quantitative approach to Swift and some understrappers*, in <<Computers and the humanities>>, IV, 5, 289–304.
KOLLMANN E.D. (1973) – *Word Frequencies in Latin Literature*, in <<Revue>>, IV, 1–18.
KRÁMSKÝ J. (1964) – *A Quantitative Phonemic Analysis of Italian Mono-, Di-, and Trisyllabic Words*, in <<Travaux Linguistiques de Prague>>, I, 129–143.
KRÁMSKÝ J. (1969) – *The word as a linguistic unit*, The Hague.
KUČERA H., FRANCIS W.N. (1967) – *Computational Analysis of Present-Day American English*, Providence (Rhode Island).
KUČERA E., MONROE G.K. (1968) – *A Comparative Quantitative Phonology of Russian, Czech and German*, New York.
KUNO S. (1963) – *The Current Grammar for the Multiple-Path English Analyzer*, in *Mathematical Linguistics and Automatic Translation*, Report NSF 8, Cambridge (Mass.).

KUNO S. (1965 a) – *The Predictive Analyzer and a Path Elimination Technique*, in <<CACM>>, VIII, 7, 453–462.
KUNO S. (1965 b) – *A System for Transformational Analysis*, in *Mathematical Linguistics and Automatic Translation*, Report NSF 15, Cambridge (Mass.).
KUNO S. (1967 a) – *A Context-Sensitive Recognition Procedure*, in *Mathematical Linguistics and Automatic Translation*, Report NSF 18, Cambridge (Mass.).
KUNO S. (1967 b) – *Computer Analysis of Natural Languages*, in *Proceedings of Symposium in Applied Mathematics*, vol. XIX, Providence, 52–110.
KUNO S., OETTINGER A.G. (1963) – *Multiple-Path Syntactic Analyzer*, in Popplewell (ed.), *Information Processing* 1962, Amsterdam, 306–312.
LAKOFF G. (1971 a) – *On Generative Semantics*, in D.D. Steinberg, L.A. Jakobovits (1971), 232–296.
LAKOFF G. (1971 b) – *Presupposition and relative well-formedness*, in D.D. Steinberg, L.A. Jakobovits (1971), 329–340.
LAMB S.M. (1961) – *The digital Computer as an Aid in Linguistics*, in <<Language>>, XXXVII, 3, 382–412.
LAMB S.M. (1965) – *Linguistic Data Processing*, in D. Hymes (ed.), *The Use of Computers in Anthropology*, The Hague, 159–188.
LAMB S.M. (1966) – *Outline of Stratificational Grammar*, Washington, D.C.
*Le applicazioni dei Calcolatori alle Scienze morali e alla Letteratura* (1962), Milano.
LEHMANN W.P. (1973) – *On the Design of a Central Archive for Lexicography in English*, dans McDavid, Duckert (1973), 312–317.
LEPSCHY G.C. (1964) – *Note sulla fonematica italiana*, in <<Italia Dialettale>>, XXVII, 53–67.
LEPSCHY G.C. (1966) – *La linguistica strutturale*, Torino.
*Les Machines dans la linguistique* (1968), Prague.
LEVISON M., MORTON A.Q., WINSPEAR A.D. (1968) – *The seventh Letter of Plato*, in <<Mind>>, LXXVII, n. 307, 309–325.
*Lexicologie et Lexicographie françaises et romanes* (1961), CNRS.
*Literary Data Processing* (1964), Yorktown Heights (N.Y.).
LOCKE W.N., BOOTH A.D. (1955) (eds.) – *Machine Translation of Languages*, New York and London.
LYNE A.A. (1972) – *The problem of non-comparability of word-frequency counts. A proposal for a standard decision-procedure to be followed in lemmatizing French*, Communicatione présentée au III Congrès Internationale AILA.
LYNE A.A. (1973) – *L'élaboration des listes de fréquence*, in <<Cahiers de lexicologie>>, II, 23, 83–108.
LYONS J., WALES R.J. (1966) (eds.) – *Psycholinguistics Papers*, Edinburgh.
LYONS J. (1968) – *Introduction to Theoretical Linguistics*, Cambridge. Trad. it., *Introduzione alla Linguistica Teorica*, Bari, 1971.
MAGNO CALDOGNETTO (1973) – *Scopi, metodi e applicazioni della fonetica sperimentale*, in <<Miscellanea II>>, Udine.
MARCKWORT M.L., BELL L.M. (1967) – *Sentence-Length Distribution in the Corpus*, in H. Kučera, W. Nelson Francis (1967).
MARIONI B. – *Studio statistico dell'italiano parlato*, (Thèse non publiée), Pisa, 1975.
MARKOV A.A. (1961) – *Theory of Algorithms*, IMEI, Accademia delle Scienze dell'URSS vol. 42. (Tradotto dell'Israel Program for Scientific Translations).
MARTINET A. (1960) – *Éléments de linguistique générale*, Paris.
MATORÈ G. (1953) – *La méthode in lexicologie*, Paris.
MAXWELL A.E. (1967) – *Analysing Qualitative Data*, London.
MELBY A.K. (1974) – *Junction Grammar and Machine Assisted Translation*, in Zampolli (1974 a).
MENARD N. (1972) – *Mesure de la richesse lexicale. Méthodologie et vérifications expérimentales* (Thèse), Strasbourg, (datt.).
MERIGGI P. (1973) – *Un lexique de l'hittite cuneiforme*, in Zampolli (1973 a), 111–113.
MICHAELSON S., MORTON A.Q. (1972) – *The Spaces in between: A multiple test of authorship for greek writers*, in <<Revue>>, I, 23–77.

MICHÉA R. (1964) – *Les vocabulaires fondamentaux*, dans <<Recherches et techniques nouvelles au service de l'enseignement des langues vivantes>>, Strasbourg, 21–38.
MICHIE D. (1968) (ed.) – *Machine Intelligence* 3, New York.
MIGLIORINI B. (1941) – *La lingua del* 900, Firenze.
MIGLIORINI B. (1951) – *Che cos'é un vocabolario?*, Firenze.
MIGLIORINI B. (1960) – *Storia della lingua italiana*, Firenze.
MIGLIORINI B. (1963) – *La lingua contemporanea*, Firenze.
MIGLIORINI B., BALDELLI I. (1964) – *Breve storia della lingua italiana*, Firenze.
MINSKY N. (1968) (ed.) – *Semantic Information Processing*, Cambridge (Mass.).
MISTRIK J. (1969) – *Frekvencia slov v Slovenčine*, Bratislava.
MOLINA E. (1916) – *Le gamme stenografiche*, in <<Bollettino stenografico italiano>>, XV, 4–8; 13–19; 55–64.
MOREAU R. (1962) – *Au suijet de l'utilisation de la notion de fréquence en linguistique*, <<Cahiers de Lexicologie>>, III, 140–158.
MOREAU R. (1966) – *Intervention de M.R. Moreau*, dans *Statistique et analyse linguistique*, Paris 125–132.
MOUNIN G. (1964) – *La machine à traduire*, The Hague.
MOYNE J.A. (1972) – *Some Grammars and Recognizers for Formal and Natural Languages*, in Julius T. Tou, (Ed.), *Advances in Information System Science*, vol. 5, New York.
MULLER Ch. (1963) – *Le mot, unité de texte et unité de lexique en statistique lexicologique*, <<Travaux de Linguistique et de Littérature>>, 155–173.
MULLER Ch. (1965) – *Fréquence, dispersion et usage. À propos des dictionnaires de fréquence*, <<Cahiers de Lexicologie>>, 7, 2, 32–42.
MULLER Ch. (1967) – *Étude de statistique lexicale: Le vocabulaire du théâtre de Corneille*, Paris.
MULLER Ch. (1968) – *Initiation à la statistique linguistique*, Paris.
MULLER Ch. (1973 a) – *Éléments de statistique linguistique*, dans Zampolli (1973 a), 348–378.
MULLER Ch. (1973 b) – *Initiation aux méthodes de la statistique linguistique*, Paris.
MULLER Ch. (1975) – *Peut-on estimer l'étendue d'un vocabulaire?*, <<Cahiers de Lexicologie>>, II, 3–29.
MULLER Ch. (1976) – *Travaux récents de statistique linguistique*, <<Invited paper>> ou COLING 76, Ottawa.
NEWELL A., BARNETT J., FORGIE J.W., GREEN C., KLATT D., LICKLIDER J.C.R., MUNSON J., REDDY D.R., WOODS W.A. (1973) – *Speech Understanding Systems*, Amsterdam.
NEWELL A., SIMON H.A. (1971) – *Human Problem Solving*, Englewood Cliffs (N.J.).
NOVAK L.A. (1963) – *Nekotorye vosprosy lingvostatistiki i častotnye slovari.* (Alcuni problemi di statistica linguistica e i vocabolari di frequenza), in <<Ministerstvo Prosveščenija MSSR, Belčkij gosudarstvennyj Pedagogičeskij Institut im. A. Russo>>, Učenye Zapiski, vyp. 6, 22–76 (e traduzione italiana nella II appendice del volume *Statistica linguistica*, Bologna, 1971, 377–440).
OHMANN R. (1964) – *Generative Grammars and the Concept of Literary Style*, <<Word>>, XX, 423–439.
PACAK M. (1967) – *Homographs: Classification and Identification*, <<Études de linguistique appliquée>>, V, 88–105.
PALME J. (1971) – *A natural Language Parsing Program for Question Answering*, FOAP Rapport C 8268–11 (64), Stockholm.
PAPP F. (1966) – *Mathematical Linguistics in the Soviet Union*, The Hague.
PETRICK S.R. (1965) – *A Recognition Procedure for Transformational Grammars*, unpubl. Doct. Diss., MIT, Cambridge (Mass.).
PETRICK S.R. (1972) – *Computer-oriented grammars and parsing*, in Center For Applied Linguistics, 22–37.
PETÖFI J.S. (1969) – *On the complex analysis of languages as synchronic systems*, in <<Revue>>, I, 1–18.
PIKE K.L. (1967) – *Language in Relation to a Unified Theory of the Structure of Human Behavior*, The Hague.
PLATH W. (1962) – *Mathematical Linguistics*, in *Trends in European and American Linguistics*, 1930–1960, Utrecht.

PLATH W.J. (1974) – *Transformational Grammar and Transformational Parsing in the REQUEST System*, in Zampolli (1974 a).
QUILLIAN M.R. (1968) – *Semantic Memory*, in M. Minsky (1968), 216–270.
RABEN J. (1969) – *The death of hand made concordance*, New York.
REED D.W. (1949) – *A statistical Approach to quantitative Linguistic Analysis*, <<Word>>, V, 235–247.
REGULA M., JERNEI J. (1965) – *Grammatica italiana descrittiva*, Berna.
REVZIN I.I. (1968) – *Les modèles linguistiques* (Modeli jazyka), Traduit et adapté par Y. Gentilhomme, Paris.
REY-DEBOVE J. (1970) – *Le domaine du dictionnaire*, in <<Langages>>, 19, 3–34.
REY-DEBOVE J. (1971) – *Limites des applications de la linguistique à la lexicographie*, in G.E. Perre, J.L.M. Trim (eds.), *Applications of Linguistics*, Cambridge, 369–375.
RIESBECK C. (1973) – *Computer Analysis of Natural Language in Context*, Ph.D. Thesis, Stanford (Calif.).
RÍHA A., MACHOVÁ S. (1974) – *Computer Testing of Generative Grammar*, in A. Zampolli (1974 a).
ROBINSON I.J. (1962) – *Preliminary Codes and Rules for the Automatic Parsing of English*, RM-3339-PR, Rand Corp., Santa Monica (Calif.).
ROBINSON J.J., MARKS S. (1965) – *PARSE: A System for automatic Analysis of English Text*, RM-4564-PR, Rand Corp., Santa Monica (Calif.).
ROCERIC-ALEXANDRESCU A. (1968) – *Fonostatistica limbii române*, Bucarest.
RODRIGUEZ BOU I. (1952) – *Recuento de vocabulario español*, Rio Piedras (Puerto Rico).
ROSENGREN I. – *The quantitative concept of language and its relation to the structure of frequency dictionaries*, in <<Études de linguistique appliquée>>, I, 1, 103–127.
ROSETTI A. (1947) – *Le Mot. Esquisse d'une Théorie générale*, Copenhague-Bucarest.
ROSIELLO L. (1965) – *Struttura, uso e funzioni della lingua*, Firenze.
ROSS (1976) – *The use of word-class Distribution Data for Stylistics communication* présentée au COLING 76, Ottawa.
RUBENSTEIN M. (1966) – *Directions in Semantic Research*, in *Seminar on Computational Linguistics*, Bethesda, 97–104.
RUMELHART D.E., LINDSAY P.H., NORMAN D.A. (1972) – *A process model for long term memory*, in Tulving, Donaldson, *Organization and Memory*, New York.
RUWET N. (1964) – *La linguistique générale aujourd'hui*, in <<Archives européennes de Sociologie>>, V, 277–310.
RUWET N. (1968) – *Introduction à la grammaire générative*, Paris.
SAGER N. (1972) – *The Sublanguage Method in String Grammars*, in R.W. Ewton Jr., J. Ornstein (eds.), *Studies in Language and Linguistics* (1970–71), El Paso (Texas).
SARAMANDU M. (1966) – *Sur le rendement fonctionnel des types de structures phonématiques en roumain*, in <<Cahiers de linguistique théorique et appliquée>>, III, 147–160.
SCHLISMANN A. (1948) – *Sprach- und Stilanalyse mit einem vereinfachten Aktionsquotienten*, in <<Wiener Zeitschrift für Philosophie, Psychologie, und Pädagogik>>, II, 2, 42 e segg.
SEDELOW S.Y., SEDELOW W.A. jr. (1972) – *Language Research and The Computer*, The University of Kansas, Lawrence (Kansas).
SEILER M. (1964) – *On defining the word*, dans <<Proceedings of the ninth International Congress of Linguists>>, The Hague, 767–770.
SHANNON C.E. (1948) – *A mathematical theory of communication*, in <<Bell System Technical Journal>>, XXVII, 379–423, 623–656.
SIMMONS R. (1970) – *Natural Language Question Answering Systems*: 1969, in <<CACM>>, XIII, 1, 15–30.
SIMMONS R., BRUCE B. (1971) – *Some Relations between Predicate Calculus and Semantic Net Representations of Discourse*, in <<Proceedings of 2nd International Conference on Artificial Intelligence>>, London, 524–530.
SIMMONS R. (1973) – *Mapping English Strings into Meanings*, Techn. Report NL 10, Dept. of Computer Sciences, University of Texas, Austin (Texas).
SMITH R.N. (1972) – *Interactive Lexicon Updating*, in <<Computers and the Humanities>>, VI, 3, 137–145.

SPANG-HANSSEN H. (1956) – *The study of Gaps between Repetitions*, dans M. Halle, *et alii* (eds.), *For Roman Jakobson*, The Hague, 492–502.
SPANG-HANSSEN H. (1963) – *Sentence Length and Statistical Linguistics*, dans *Structures and Quanta: Three Essays on Linguistic Description*, New York, 58–72.
SPANG-HANSSEN H. (1967) – *Fini et infini dans le vocabulaire*, <<Langages>>, 6, 100–105.
ŠTEINFELDT E. (1969) – *Dictionnaire des fréquences des mots dans la langue russe moderne.* Les 2500 mots les plus usuelles à l'usage des professeurs de Russe, Moscou.
*Structure of Language and its Mathematical Aspects* (1961), AMS, Providence (R.I.).
SZANSER A.J. (1969) – *Automatic error-correction in natural languages*, Preprint to the 3rd ICCL.
TAGLIAVINI C. (1968) – *Applicazioni dei calcolatori elettronici all'analisi e alla statistica linguistica*, dans *Atti del Convegno sul tema: L'automazione elettronica e le sue applicazioni scientifiche, tecniche e sociali* (Accademia Nazionale dei Lincei, Roma, 1967), Roma, 111–118.
TESNIÈRE L. (1959) – *Éléments de syntaxe structurale*, Paris.
THORNDIKE E.L. (1932) – *A teacher's word book of the twenty thousand words found most frequently and widely in general reading for children and young people*, New York.
THORNDIKE E.L., LORGE I. (1944) – *The teacher's word book of* 30.000 *words*, New York.
THORNE J.P. (1965) – *Stylistics and Generative Grammars*, <<Journal of Linguistics>>, I, 49–59.
TOGEBY K. (1949) – *Qu'est-ce qu'un mot?*, <<Travaux du Cercle Linguistique de Copenhague>>, V, 97–112.
DE TOLLENAERE F. (1963) – *Nieuwe wegen in de lexicologie*, Amsterdam.
DE TOLLENAERE F. (1973) – *Travaux de l'Institut de Lexicologie Néerlandaise*, dans Zampolli (1973 a), 29–39.
TORRIGIANI G. (1968) – *Problemi metodologici dell'analisi linguistica mediante elaboratori elettronici presso il CNUCE*, Pisa.
VALERIO C. (1893–1896) – *De la Cryptographie*, Paris.
VANDER BEKE G.E. (1929) – *French word book*, New York.
VINAY J.P. (1968) – *Enseignement et apprentissage d'une langue seconde*, in <<Le Langage>>, (Encicl. de la Pléiade, dir. par A. Martinet), Paris, 685–725.
WATTS A.F. (1946) – *The language and mental development of children*, London.
WILKS Y., HERSOVITZ A. (1974) – *An Intelligent Analyzer and Generator for Natural Language*, in Zampolli (1974 a).
WOODS W.A. (1967) – *Context-Sensitive Recognition*, in *Mathematical Linguistics and Automatic Translation*, Report MSF 18, Cambridge (Mass.).
WOODS W.A. (1970 a) – *Context-Sensitive Parsing*, in <<CACM>>, XIII, 7, 437–445.
ZAMPOLLI A. (1960) – *Studi di statistica linguistica eseguiti con impianti IBM* (Tesi di laurea dattiloscritta), Padova.
ZAMPOLLI A. (1967) – *Nota Tecnica*, dans *Raccolta Barbi di Canti Popolari Italiani, Esperimento di Elaborazione Elettronica* E1/RB, Pisa.
ZAMPOLLI A. (1968 a) – *L'elaboratore elettronico negli studi linguistici*, <<Rivista IBM>>, 2, 14–19.
ZAMPOLLI A. (1968 b) – *Recherche statistique sur la composition phonologique de la langue italienne exécutée avec un système IBM, Machines dans la Linguistique*, Prague, 25–34.
ZAMPOLLI A. (1969) – *Due conversazioni sullo stato attuale della linguistica computazionale*, Pisa.
ZAMPOLLI A. (1973 a) – *Linguistica Matematica e Calcolatori. Atti del Convegno e della Prima Scuola Internazionale*, Firenze, 1973.
ZAMPOLLI A. (1973 b) – *La Section Linguistique du CNUCE*, Zampolli (1973 b), 133–199.
ZAMPOLLI A. (1973 c) – *L'automatisation de la recherche lexicologique: état actuel et tendances nouvelles*, <<META>>, XVIII, 1–2, 101–136.
ZAMPOLLI A. (1973 e) – *Humanities Computing in Italy*, <<Computers and the Humanities>>, 7, 6, 343–360.
ZAMPOLLI A. (1974 e) – *Problemi di linguistica applicata computazionale*, Pisa.
ZIPF G.K. (1935) – *The Psycho-biology of Language: an introduction to dynamic phylology*, Cambridge (Mass.). (Citato nell'ediz. 1968).