Représentations géométriques de relations sémantiques

par

Ugo BERNI CANANI

Centro elettronico di Documentazione - ROMA

Nous considérerons, par rapport à un corpus donné de contextes, deux types de graphe :

a) Un graphe dans lequel les sommets représentent des mots (lemmes) et les arêtes une relation paradigmatique : deux sommets sont reliés par une arête si les mots qui leur correspondent peuvent être considérés comme équivalents dans quelque contexte du corpus (synonymes, hyponymes, antonymes, termes appartenant à un même champ paradigmatique).

b) Un graphe bi-partite, où les arêtes représentent une relation syntagmatique simple, de compatibilité ou "application" : deux mots sont reliés par une arête si le deuxième est appliqué au premier dans quelque contexte du corpus; autrement dit, si le couple de mots est un syntagme dans le corpus.

A chaque sommet correspond le sous-graphe formé par les sommets adjacents (et par les sommets qui à leur tour leur sont adjacents, s'il s'agit d'un graphe bi-partite) et par les arêtes qui les relient entre eux. Nous l'appellerons "adjacence" du sommet.

Nous examinerons quelques aspects de ces sous-graphes, en essayant de déterminer des expressions synthétiques de leur structure dans ses différences de complexité, des composantes capables de simuler des unités de sens, une orientation possible des arêtes. Enfin, nous indiquerons quelques façons de représenter géométriquement des graphes de relations sémantiques.

I. Relations paradigmatiques (1)

a) Composantes caractéristiques.

Considérons pour chaque sommet du graphe de relations paradigmatiques l'ensemble des sommets adjacents et celui des arêtes qui les relient entre eux. Les composantes connexes de cet ensemble, que nous pouvons appeler fibres, représentent une première unité de sens, celle qui reflète des acceptions nettement distinctes. Elles correspondent à uneévidence intuitive : si un mot a deux sens nettement distincts, nous nous attendons à ce qu'entre les termes qui lui sont respectivement liés dans les deux sens, il n'y ait pas de liaison.

A l'intérieur de chacune des fibres nous pouvons distinguer une unité de sens plus fine, donnée par les composantes complètes maximales (cliques) : à savoir les groupes de sommets tous reliés entre eux (composantes complètes) n'étant contenus dans aucune composante complète plus grande. Les cliques représentent des nuances, des oscillations à l'intérieur d'un même sens. Fibres et cliques constituent une première approximation de la définition d'unités de sens dans un champ paradigmatique (2),

b) Partitions et recouvrements.

Par rapport à un graphe de 5-6.000 sommets et 15-20.000 arêtes, nous pouvons estimer que l'approximation est satisfaisante (3). Mais, si l'on augmente le nombre de contextes et donc on ajoute de

nouveaux termes et de nouvelles liaisons, nous observerons la fusion de plusieurs cliques ou plusieurs fibres ce qui, dans le deuxième cas, entraîne une confusion possible de différents sens. Ceci nous amène à modifier les conditions de séparation des fibres dans l'adjacence (ensemble des sommets adjacents à un sommet donné) de chaque sommet.

La séparation peut être obtenue substantiellement de deux façons : en opérant, par rapport à chaque sommet, soit une partition (regroupement des sommets adjacents dans des classes disjointes), soit un recouvrement (regroupement dans des classes non nécessairement disjointes) de son adjacence. La nature des données conseillerait d'opter pour des recouvrements, que l'on peut obtenir de toute façon à travers des partitions : si nous remplaçons l'adjacence d'un sommet par le graphe formé par les cliques qui la composent (les sommets représenteront des cliques, les arêtes les intersections non vides entre elles), une partition des sommets de ce graphe équivaut à un recouvrement de l'adjacence remplacée. En conservant ainsi l'organisation en cliques de l'adjacence de chaque sommet nous avons l'avantage d'opérer sur des structures interprétables, moins abstraites que celles qui sont directement produites par les techniques courantes de clustering et de classification automatique que nous devrons employer pour obtenir partitions et recouvrements.

c) Complexité.

Pour essayer d'exprimer l'intuition de la complexité plus ou moins grande de la structure d'un graphe on peut utiliser différentes notions. En voici quelques unes :

- la densité (rapport entre le nombre d'arêtes qui relient les sommets du graphe et le nombre d'arêtes nécessaires pour relier tous les sommets entre eux);
- la connexité, mesurée en fonction du nombre de composantes connexes, du nombre minimal de sommets (ou d'arêtes) dont l'élimination augmente d'une unité le nombre de composantes connexes, ou aussi du nombre de cliques (4);
- la complexité, au sens technique, définie comme le nombre de "spanning trees" (5).

Si l'on applique aux fibres du graphe de relations paradigmatiques, l'une ou l'autre de ces notions, l'on obtiendra une information synthétique, un indice de leur complexité intuitive. Il n'est pas exclu qu'à des mots différents, de par leur catégorie grammaticale ou leurs propriétés sémantiques, correspondent des caractéristiques différentes de complexité.

Une autre voie pouvant amener à une description synthétique de la structure de chaque fibre consiste à représenter sur un arbre les différents niveaux de connexion entre les cliques qui la composent. Si l'on appelle n-connexes deux cliques qui ont en commun au moins n sommets, l'on forme (en partant du plus petit n tel que toutes les cliques soient isolées) les composantes (n-1)-connexes, (n-2)-connexes . . . et ainsi de suite jusqu'à l'unique composante 1-connexe qui épuise toute la fibre. Nous obtenons un arbre qui met en évidence la façon dont une fibre se forme, par agrégation autour des pôles les plus denses. Si l'on néglige les éléments (les cliques) pour ne considérer que les classes (les composantes)

nous avons une simplification des arbres qui, représentés par exemple comme des séquences d'astérisques, correspondant aux pôles, et de parenthèses, indiquant les niveaux de connexion, fournissent une image extrêmement maniable de la structure des fibres.

d) Orientation.

La structure en arbre qui peut être tirée des fibres nous amène à considérer un thème voisin : l'orientation. A chaque graphe peut être donnée une orientation arbitraire des arêtes, qui permet de démontrer des propriétés indépendantes de cette orientation et d'introduire en outre l'homologie (6). Cependant, ce que nous cherchons est une orientation interprétable dans le lexique, et l'on pense tout de suite au rapport de subordination hiérarchique entre des termes. Une arête devrait aller du sommet qui représente un mot de sens plus spécifique vers celui qui représente un mot de sens plus général. On pourrait s'attendre à une diversité de "degré" (nombre d'arêtes incidentes à un sommet) entre des sommets reliés correspondant à des mots d'un niveau différent de généralité. Or, ce n'est pas ce qui arrive : à cause d'un phénomène caractéristique de localisation du lexique (7), même si l'on représente chaque mot d'un corpus avec autant de sommets qu'il a de sens, et donc qu'on élimine ainsi la polysémie du graphe, la différence de degré entre deux sommets adjacents ne reflète pas systématiquement une différence de généralité. Il faut donc rechercher ailleurs un indice de subordination hiérarchique, probablement dans l'articulation interne des fibres de sommets adjacents; et dans cette optique la recherche peut justement commencer en partant de l'attribution d'une orientation arbitraire.

II. Relations syntagmatiques

a) Composantes caractéristiques.

Un graphe de relations syntagmatiques présente des caractéristiques très différentes de celles qui sont associées aux relations paradigmatiques. C'est idéalement un graphe bi-partite qui représente un ensemble A de "mots-argument" (8) et un ensemble F de mots-fonction : les arêtes relient un mot de type F aux mots qui peuvent en être argument ou, vice-versa, un mot de type A aux mots qui peuvent s'y appliquer (par ex. un adjectif ou un verbe appliqués à des substantifs, un adverbe à des verbes).

Pour identifier des unités de sens nous chercherons un analogue des fibres et des cliques. Si un mot de type A (resp. de type F) a plusieurs sens, les mots de type F (resp. A) qui lui sont associés devraient pouvoir être répartis en groupes correspondant à ces sens. Dans le graphe de relations paradigmatiques, les groupes étaient identifiés dans les fibres. Dans un graphe de relations syntagmatiques, puisqu'il s'agit de graphe bi-partite, nous devons recourir à une connexion indirecte : deux mots sont indirectement reliés s'il y a au moins un mot auquel les deux sont reliés. L'analogue des fibres sera constitué par les composantes indirectement connexes de l'adjacence de chaque sommet.

Idéalement, nous devrions trouver pour un mot X autant de composantes indirectement connexes

dans son adjacence qu'il a de sens. Cependant, dans des graphes de dimensions analogues à celles qui pour les relations paradigmatiques permettaient une bonne discrimination des sens à travers les fibres, l'adjacence de chaque sommet sera presque toujours indirectement connexe et nous ne pourrons rien distinguer. Imposer une connexité indirecte plus forte (par ex. au moins deux termes dans l'intersection des adjacences de deux sommets) n'a d'autre résultat que de privilégier les sommets de degré plus élevé. Nous sommes donc obligés de recourir à des critères statistiques avant même de chercher les composantes qui nous intéressent.

Nous demanderons, par exemple, pour les termes indirectement reliés à un terme X, une proximité de X supérieure à un seuil donné, en excluant les autres de l'adjacence indirecte de X. J'ai essayé, sur quelques milliers de syntagmes, différents indices de proximité (9) et ai obtenu les meilleurs résultats avec le coefficient $\frac{1}{\sqrt{x.y}} \cdot \sum_{j \in J} \frac{1}{Aj}$ (10)

qui n'est autre que le terme général de la matrice à diagonaliser dans l'analyse des correspondances quand la matrice des données est binaire (11). Un seuil élevé sélectionne parmi les termes indirectement liés à X presque exclusivement des synonymes, ou des termes équivalents à X par rapport au corpus; en abaissant le seuil on tend à couvrir tout le champ paradigmatique de X dans le corpus, mais au prix de termes qui lui sont étrangers : on retrouve la complémentarité de précision et exhaustivité qui est caractéristique de l'information retrieval.

Dans un graphe bi-partite l'analogue des cliques est constitué par les "étoiles", sous-graphes où chaque sommet d'un type est relié à tous les sommets de l'autre type. Le nombre des étoiles maximales est dans un graphe de relation syntagmatique de loin supérieur à celui des cliques dans un graphe de relations paradigmatiques de dimension et densité analogues. En outre, si celles qui sont très "allongées" (par ex. deux ou trois termes de type A tous reliés à de nombreux termes de type F, ou vice-versa) peuvent correspondre à de petits groupes de synonymes, les autres ne sont généralement pas interprétables. Nous pouvons toutefois obtenir un analogue des cliques en exploitant les indices de proximité déjà adoptés pour les fibres. Si dans le graphe bi-partite nous relions, par une arête, les termes de type A dont la proximité dépasse un seuil donné, puis nous faisons la même chose avec les termes de type F, nous obtenons un graphe dans lequel les cliques, par construction, tendent à identifier des unités minimales de sens (les nouvelles liaisons introduites sur la base de la proximité doivent neutraliser les étoiles déterminées "fortuitement" par les termes de degré plus élevé).

b) Les seuils.

A chaque sommet X du graphe de relations syntagmatiques, nous associons le sous-graphe bi-partite constitué par les sommets adjacents à X et par les sommets à leur tour adjacents à ces derniers; à l'intérieur de ce sous graphe, nous avons déterminé, en exploitant des seuils statistiques, un analogue des fibres (sur les sommets adjacents à X) et un analogue des cliques (dans les étoiles du sous-graphe).

A ce stade, nous devrions considérer des partitions et des recouvrements du sous-graphe qui tiennent compte de ces structures, mais il nous semble plus important d'essayer de préciser le rôle des seuils que nous avons introduits pour filtrer à l'origine une partie du bruit de fond, des coïncidences "aléatoires".

Comparons des entités dotées chacune d'un certain nombre de caractéristiques. Une mesure de la proximité de deux entités pourra être fonction du nombre de caractéristiques communes. Laissons de côté celles qui n'appartiennent qu'à l'une ou à l'autre de ces entités : elles peuvent correspondre, quand les entités sont des mots, à des sens différents du sens qui est éventuellement commun à ces entités. Le nombre de caractéristiques communes n'est cependant pas suffisant : intuitivement, nous voudrions attribuer plus d'importance aux caractéristiques moins fréquentes qu'à celles qui sont plus fréquentes dans la population. Le coefficient choisi pour déterminer les fibres satisfait justement cette exigence.

Supposons maintenant que les entités A et B aient en commun deux caractéristiques (R, S) fortement corrélées entre elles et les entités A et C, deux caractéristiques (U, V) faiblement corrélées. Si les quatre caractéristiques ont la même fréquence et qu'il n'y a pas d'autres caractéristiques communes, considérerons-nous A plus semblable à B ou à C ?

Si nous restons attachés au nombre de caractéristiques communes, nous serons amenés à préférer la seconde hypothèse puisque deux caractéristiques fortement corrélées tendent à se fondre en une seule caractéristique. D'autre part, nous sommes également amenés à penser que, dans le cas d'une forte corrélation entre les caractéristiques, chacune d'elles confirme l'autre, et à attribuer plus d'importance à la coïncidence sur l'une des caractéristiques si elle est confirmée par la coïncidence sur l'autre; ce qui n'arrive pas quand les caractéristiques sont indépendantes. Ce que nous attendons d'entités semblables, et en général d'une classe, c'est qu'elles aient en commun plusieurs groupes de caractéristiques, celles de chaque groupe étant positivement corrélées entre elles mais indépendantes de celles des autres groupes. Et l'analogue des cliques que nous avons déterminé dans le graphe bi-partite répond à cette hypothèse.

c) Orientation.

Dans le graphe de relations syntagmatiques, nous avons supposé qu'une orientation naturelle, induite par les deux ensembles de sommets, était donnée; mais si , dans le syntagme "abandon (du) travail", il nous semble naturel d'étiqueter "travail" comme argument de "abandon", et dans le syntagme "abandon immédiat" aussi naturel de considérer "abandon" comme argument de "immédiat", quelle attitude adopter devant des syntagmes comme "abandon (pour) grève" ? (12)

Si nous considérons les neuf arêtes correspondant aux couples que nous pouvons former avec les termes "abandon", "interruption" et "continuation" d'un côté; "activité", "travail", "grève" de l'autre, nous sommes en présence d'une étoile où les trois premiers termes ont un même rôle par

rapport à "travail" et "activité" (leur objet direct), différent de celui qu'ils ont dans les syntagmes "abandon (pour) grève", "interruption (pour) grève", "continuation (de la) grève". Si nous pouvons regrouper les arêtes du premier groupe en une seule unité de sens ("abandon" est à "travail" et à "activité" comme "interruption" et "continuation"), en revanche, nous ne pouvons pas les réunir aux autres arêtes ("abandon" n'est pas à "grève" comme il est à "travail" ni comme "continuation" est à "travail"). L'orientation "naturelle" est multiple : elle traduit les relations de cas, la complémentation; ses dimensions sont les dimensions de l'analogie.

En réalité, la structure de graphe bi-partite n'est qu'une première approximation vers la description d'un champ de relations syntagmatiques. Les éléments qu'il faut considérer ne sont pas tant les sommets que les arêtes, non pas les mots mais les syntagmes. Nous devons tenir compte du fait qu'un mot varie en sens d'un syntagme à l'autre; il s'agit de variations brusques quand il devient un autre mot, et nous parlons alors de polysémie, mais ce peut être de simples oscillations, des variations continues ou, pour ainsi dire, infinitésimales (pensons par exemple aux contextes de "aspirine").

III. Représentations dans des espaces métriques et des espaces topologiques (13)

On obtient une représentation géométrique directe d'un graphe de N sommets, en le plongeant dans R^N (14). Chaque sommet est identifié avec un vecteur unitaire de la base et chaque arête est représentée par le segment qui unit les sommets qui la déterminent. L'union de ces segments est un espace où la distance entre des points appartenant à des segments disjoints est toujours 2.

Une autre représentation nous est fournie par l'analyse factorielle du graphe (15) qui permet de l'approximer avec un nombre réduit de dimensions. Appliquée aux sous-graphes associés aux sommets d'un graphe de relations paradigmatiques ou syntagmatiques, elle produit une bonne distribution des mots dans des régions différentes, même sur un plan (celui déterminé par les deux premiers axes factoriels). L'examen des projections sur les différents axes et des contributions de chaque sommet à la formation de ces axes nous permet de contrôler, d'un autre point de vue, les composantes caractéristiques identifiées jusqu'à maintenant. Notons en passant que l'analyse factorielle peut être utilisée également pour représenter directement des phrases ou des contextes plus longs (16), mais dans ce cas, les axes et les "clusters" ne devront pas correspondre à des "aires sémantiques", à des régions du champ paradigmatique d'un mot, mais à des thèmes. Une illustration des résultats de cette approche à des fins d'information retrieval et de résumé automatique de textes dépasse toutefois les limites de cette communication. Ce que nous voulons souligner c'est l'importance du choix de la distance : dans l'analyse en composantes principales, la distance entre les variables est déterminée par le coefficient de corrélation, dans l'analyse des correspondances par le "chi carré". Il me semble qu'avec la première, tous les clusters du graphe analysé tendent à être représentés sur chaque axe, tandis que la deuxième tend à associer chaque axe à des clusters différents. Il se peut que d'autres distances adhèrent aux caractéristiques particulières de la distribution des mots (17).

Si nous voulons éviter les problèmes inhérents au choix d'une distance, tout en continuant à chercher une représentation géométrique de nos graphes, nous pouvons recourir au concept d'espace topologique, plus général que celui d'espace métrique. Nous pouvons penser aux complémentaires des cliques comme à une sous-base du système d'ouverts, mais la topologie qui en résulte ne fournit pas une interprétation satisfaisante. L'intersection de deux cliques n'est pas une clique, même si c'est encore une composante complète; si deux cliques ont une intersection non vide, leur union peut être interprétée comme un sens plus large, ou plus étroit, mais l'union de deux cliques disjointes ne correspond pas normalement à un sens. On peut faire des considérations analogues pour les fibres. Les entités que nous réussissons à construire en partant des unités de sens que nous avons définies n'ont plus de rapport avec ces définitions. Nous rencontrerions le même genre de difficultés si nous voulions partir, pour construire la topologie, de la notion d'ordre.

Nous pouvons attribuer ces difficultés au fait que, dans nos données, la proximité (affinité de sens) n'est pas transitive, alors qu'une propriété nécessaire des topologies est justement une sorte de transitivité du voisinage (18). Si nous renonçons à cette propriété, nous avons des notions affaiblies, des prétopologies, dont l'une, attribuée à Fréchet, a été justement appliquée à la description de graphes, à des relations non transitives (19). On peut la construire en définissant un opérateur "adhérence" ou un opérateur "intérieur". Dans le cas de graphes non orientés, non pondérés, et avec une boucle pour chaque sommet, l'opérateur "adhérence" associe à chaque ensemble X de sommets le sur-ensemble constitué par les sommets adjacents à au moins un sommet de X; l'opérateur "intérieur" associe à chaque ensemble X le sous-ensemble constitué par les sommets qui ne sont adjacents qu'aux sommets appartenant à X.

Un ensemble est ouvert s'il coïncide avec son intérieur, fermé s'il coïncide avec son adhérence. La construction est plus proche que les précédentes des entités que nous voulons décrire.

Une approche encore différente est centrée sur une notion abstraite de recouvrement qui nous permet d'imaginer un lexique, ou de toute façon un graphe de relations paradigmatiques, comme un système de recouvrements. Commençons par orienter le graphe en transformant chaque arête en une flèche allant du terme le plus spécifique vers le terme le plus général, quand nous sommes en présence d'une hyponymie; autrement en deux flèches de sens opposé. Puis, complétons le graphe en introduisant de nouveaux sommets et de nouvelles flèches, de façon que pour chaque couple de flèches orientées vers un même sommet, il y ait un autre sommet relié à l'origine des deux flèches. Nous interprétons ce dernier comme une conjonction, qu'il est bien sûr superflu d'introduire lorsque la conjonction de deux termes est lexicalisée, déjà présente dans le graphe.

Pour chaque terme, choisissons maintenant comme recouvrements, des groupes de termes capables d'en exprimer le sens (synonymes, antonymes, hyponymes, termes appartenant autrement à son champ paradigmatique), exactement ce qu'on observe dans les dictionnaires quand les définitions se présentent sous forme de renvois (ex. habile = capable, apte . . .). Les recouvrements n'épuisent pas

nécessairement le sens du terme recouvert; en termes d'ensembles, on dirait que l'ensemble recouvert ne doit pas nécessairement être contenu dans l'union des ensembles recouvrants. Normalement, on peut trouver des contextes non artificiels dans lesquels le terme recouvert ne peut être remplacé sans altération de sens par aucun terme recouvrant. Et même s'il n'en était pas ainsi, nous devrions, de toute façon, réserver à chaque mot une marge de variation, de nouveaux usages possibles (la situation rappelle l'analyse factorielle classique qui, à chaque variable, "recouverte" par les covariances, attribue une marge de variance individuelle).

Si le choix des recouvrements satisfait des conditions qu'il serait trop long de préciser ici (20), on obtient, associée au graphe, une sorte de topologie, inventée par Grothendieck et appelée justement topologie de Grothendieck, ou "site". Avec cette topologie, nous pouvons essayer de fondre en une représentation unique, relations paradigmatiques et relations syntagmatiques.

Considérons chaque sommet du site comme un mot et en même temps comme un contexte, un microcontexte, pour les mots qui, avec lui, peuvent former un syntagme. Regroupons les mots compatibles avec un mot-contexte A en classes dont chacune est formée par des mots équivalents par rapport à A, dans le contexte A, et associons à A l'ensemble F(A) de ces classes. Si l'on fait la même chose pour tous les sommets du site, nous obtenons une famille d'ensembles sur laquelle nous pouvons transporter, en les invertissant, les flèches du site. A une flèche allant du mot-contexte A au mot-contexte B correspondra une flèche (ou fonction) allant de F(B) à F(A) : chaque classe de mots de F(B) sera associée à la classe de F(A) qui contient ces mots. Ainsi, par exemple, le mot "donner", compatible avec le contexte A, sera équivalent dans des contextes plus spécifiques respectivement à "attribuer", "confier", "concéder", etc . . . Des mots comme "attribuer" et "concéder", appartenant à des classes différentes dans un contexte C, peuvent coïncider dans un contexte C' (et dans tous les contextes plus spécifiques). Si l'on généralisait cette construction à des contextes plus larges, chaque mot serait représenté comme l'association, aux mots correspondant aux sommets du site qui lui sont compatibles dans un syntagme, d'ensembles de contextes possibles du syntagme (ou de "conditions de l'énoncé").

Il nous faudrait maintenant parler de pré-faisceaux, éventuellement de faisceaux, sur le site. Mais il est temps de conclure cet exposé. Plutôt que d'approfondir une seule approche du problème, nous avons préféré présenter des points de vue différents, avec des inexactitudes et des approximations qui, nous l'espérons, sont toutefois restées dans des limites tolérables. Nous croyons dans la fertilité de ces points de vue dans le domaine du lexique, mais nous ne pouvons pas ne pas rappeler les mots de Mac Lane : ". . . good general theory does not search for the maximum generality, but for the right generality" (21).

NOTES

1) Dans le même cadre mais avec un traitement différent, voir J.P. Michon et M. Potdevin, *Recherche d'associations paradigmatiques et théorie des graphes*, in *Le français moderne*, 1973, 4.

2) Pour une illustration plus étendue du rôle de ces composantes : U. Berni Canani, *Graphes de relations sémantiques* (Actes du troisième colloque international du Lessico intellettuale europeo, Roma, janvier 1980).

3) Pour des graphes extraits d'un corpus de recherches juridiques : U. Berni Canani, *Fragments of a semantic model* (Actes du Congrès international "Logica, informatica, diritto", Firenze, avril 1981).

4) Différents indices de connexité sont illustrés dans : A. Bellacicco e A. Labella, *Le strutture matematiche dei dati*, Feltrinelli, 1979.

5) Sous-graphes d'un graphe connexe qui en contiennent tous les sommets et qui ne contiennent aucun circuit (voir : R.J. Wilson, *Introduction to graph theory*, Oliver and Boyd, Edinburgh, 1972).

6) Voir : N. Biggs, *Algebraic graph theory*, chap. 4 (Cambridge University Press, 1974) et aussi *Interaction Models* (Cambridge, University Press, 1977).

7) Commenté par M. Alinei dans *La struttura del lessico*, Il Mulino, 1974, p. 33.

8) "fonction" et "argument" dans le sens mathématique habituel.

9) Voir I.C. Lerman, *Les bases de la classification automatique*, Gauthier Villars, 1970.

10) Où x est le nombre des arêtes incidentes au sommet X; y celui des arêtes incidentes au sommet Y; J est l'ensemble des sommets adjacents tant à X qu'à Y; A_j le nombre des arêtes incidentes au sommet j.

11) Cf. L. Lebart, A. Morineau, N. Tabard, *Techniques de la description statistique*, Dunod, 1977, p. 76.

12) Cf. par ex. le choix de Ch. J. Fillmore in *The case for the case* (E. Bach and R. T. Harms eds, *Universals in linguistic theory*, Holt, Rinehart and Winston, 1968, p. 24) et celui de R. Montague in *English as a formal language* (Linguaggi nella società e nella tecnica, Edizioni di Comunità, 1970, pp. 189-224).

13) Espace métrique = ensemble muni d'une distance.
Espace topologique = ensemble muni d'une topologie.

Topologie sur X = famille de sous-ensembles, comprenant X et l'ensemble vide, fermé par rapport à l'intersection finie et à l'union arbitraire. Ces sous-ensembles sont appelés "ouverts", leurs complémentaires "fermés".

14) R^N comme espace vectoriel muni de la norme $x \rightarrow \Sigma |x_n|$. Voir R. et A. Douady, *Algèbre et théories galoisiennes*, vol. 2, p. 45 (Cedic/Fernand Nathan, 1979).

15) Cf. Benzecri, J. P., *L'analyse des données*, Dunod, 1976, vol. 2, p. 244.

16) En partant du nombre d'occurrences et de co-occurrences des mots de la phrase dans le corpus de contextes considéré.

17) Sur lesquelles, cf. Ch. Muller, *Peut-on estimer l'étendue d'un lexique ?* , Cahiers de lexicologie, 1975, 2 et *Statistique lexicale et théorie du lexique*, Cahiers de lexicologie, 1976, 2.

18) Cf. la définition de topologie au moyen de la notion de voisinage (par ex. in J. L. Kelley, *General topology*, Van Nostrand Reinhold, p. 56) et la note suivante.

19) M. Mougeot, G. Duru et J. P. Auray, *La structure productive française*, Economica 1977, pp. 53 et suiv.

20) Voir M. Artin, *Grothendieck Topologies* , Harvard, 1962 et, pour le rôle joué par cette idée au-delà de son domaine d'origine, M. Makkai et G.E. Reyes, *First order categorical logic*, LNM vol. 611, Springer-Verlag, 1977.

21) S. Mac Lane, *Categories for the working mathematician*, p. 103, Springer-Verlag, 1971.