Achour OUAMARA
Université de Grenoble

COOCCURRENCE ET SYNTAXE: RESEAU LEXICO-SYNTAXIQUE

Exposé du 6 Juin
17 h.30, Salle 5

ADRESSE:
Université des Sciences Sociales(GRAD)
Bât. Sciences Humaines et Mathématiques
47X 38040 GRENOBLE Cedex
FRANCE

Résumé

On peut dire, pour simplifier, que les problématiques ayant pour objet l'étude de la forme lexicale dans l'espace d'un texte ou d'un corpus de textes se distinguent selon qu'elles prennent cette forme isolée (hors contexte) ou reliée à d'autres formes (en contexte).

Cette dichotomie ne recouvre pas exactement celle communément admise entre lexique et syntaxe, puisque les méthodes en contexte englobent aussi les analyses de cooccurrences.

La méthode que nous présentons ici est en contexte. Elle est axée sur l'utilisation des résultats d'une analyse syntaxique. Pour ce faire, nous avons mis à profit une grammaire de reconnaissance du français programmée par P. Plante à l'aide du logiciel **DEREDEC** (1). Cette grammaire permet de calculer certaines relations syntaxiques entre les formes lexicales telles que thème/propos, verbe/complément, déterminant/déterminé.

Le programme que nous avons mis au point fournit un "réseau lexico-syntaxique" qui détermine toutes les formes lexicales liées entre elles relativement à un couple de positions syntaxiques donné.

On peut suggérer que cet objet discursif (réseaux) sera une base appréciable pour rendre compte de l'organisation et de l'évolution d'un discours donné. Sur le plan diachronique, par exemple, la modification d'un réseau pour un même ensemble de formes lexicales peut être la trace d'un fonctionnement discursif insoupçonné. Seule l'analyse détaillée de ces réseaux est susceptible de décider d'une telle conclusion.

(1) cf. **Plante** P. *Le système de programmation DEREDEC*, in *Mots*, numéro 6, Presses de la Fondation Nationale des Sciences Politiques, pp.101-134.

Achour OUAMARA
Université de Grenoble

COOCCURRENCE ET SYNTAXE:RESEAU LEXICO-SYNTAXIQUE

Les problématiques ayant pour objet l'étude du lexique dans l'espace d'un texte ou d'un corpus de textes se distinguent selon qu'elles considèrent la forme lexicale hors de son contexte ou qu'elles la relient à d'autres formes (dans son contexte).

Les méthodes hors contexte

La statistique lexicale hors contexte (dite aussi paradigmatique) procède à un relevé, outre du décompte fréquentiel des formes, d'une sorte de topologie lexicale d'un texte: à chaque forme lexicale est associé un ensemble d'informations telles que sa fréquence, sa catégorie grammaticale, sa localisation dans le corpus, etc.

Certaines méthodes sémantiques procèdent de la même démarche quand elles relient la forme lexicale aux autres formes non pas dans ses réalisations discursives effectives, mais définie par un réseau de significations puisées dans un dictionnaire préétabli à l'exemple d'un thésaurus, une sorte de "mémoire sémantique" (cf. Quillian, 1968). Dans ce cas, c'est l'organisation sémantique d'une expression qui déterminerait celle du contenu. Les formes lexicales sont représentées hors contexte comme des "structures syntagmatiques (...) d'atomes sémantiques" (Paillet, 1974, p.63).

Le postulat sous-jacent qui est au principe des méthodes statistiques et sémantiques dont on vient de parler consiste à subordonner la distribution syntagmatique des formes lexicales sur la chaîne phrastique à celle de leur réalisation hors de leur condition d'emploi.

Les méthodes en contexte

Les méthodes statistiques en contexte (syntagmatiques) introduites principalement par le Laboratoire de l'Ecole de Saint-Cloud avaient ouvert une autre approche des relations entre formes lexicales sans toutefois aborder la dimension proprement syntaxique que nous développerons plus loin. Pour l'équipe de Saint-Cloud, "les cooccurrences constituent (...) la manifestation matérielle

des rapports et des relations diverses qui se nouent dans la chaîne syntagmatique du texte. Celles-ci sont aisément et systématiquement repérables tout le long d'un texte pourvu qu'aient été fixées des conditions de contiguité pour que deux occurrences soient en position de cooccurrences" (Lafon, 1981, p.97).

Du point de vue méthodologique, le choix du contexte des termes à cooccurrer est primordial: contexte immédiat? phrase? paragraphe?

Chronologiquement, l'Ecole de Saint-Cloud s'est intéressé d'abord au calcul des cooccurrences autour d'un mot-pôle (recherche des formes qui le précèdent et qui le suivent), puis son attention s'est portée sur l'étude de la distribution (répartition) le long d'un texte d'une paire de formes lexicales jouant le même rôle l'un par rapport à l'autre, enfin dernièrement (cf. Lafon, 1981) elle repense les méthodes précédentes en s'intéressant, la phrase étant choisie arbitrairement comme unité spatiale du texte, à la liaison entre deux formes à l'intérieur de cette unité: "la mesure de liaison de deux formes est (...) fondée exclusivement sur leur co-présence dans les mêmes phrases" (ibid, p.124) avec un seuil arbitraire fixé sur les probabilités des apparitions retenues. Cette méthodes détermine les couples orientés (F -----> G) et les paires non orientées (F,G) de formes cooccurrentes.

Prenons un exemple emprunté à P. Lafon (ibid. p.97). Ce dernier repère dans un corpus de discours de la CGT une liaison forte entre les formes POUVOIR et PATRONAT: liées cinq fois par la conjonction ET et deux fois co-présentes dans la même phrase. Or, appréhendées dans leur environnement syntaxique élargi au prédicat verbal, les deux liaisons apparaissent comme de nature différente.

En effet:

a) Elles sont liées par ET:

- dans la position "sujet":

	! doivent...	!
	! s'efforcent...	!
POUVOIR et PATRONAT	! tiennent compte...	!
	! poursuivent...	! (1)

- dans la position "objet":

(X) faire reculer POUVOIR et PATRONAT

b) Co-présentes dans la même phrase, en position "sujet", chacune dans une position de déterminant d'un nom:

! une partie du PATRONAT	!	
! avec la complicité(...)	! fait recours aux syndicats(...)	CFT (2)
! des milieux du POUVOIR	!	

où, cette fois-ci, l'apparition du syntagme "CFT" vient, outre perturber l'énumération ordonnée des formes POUVOIR et PATRONAT, puisque leur ordre est ici inversé, mais surtout révéler relative leur *apparente* liaison en les inscrivant dans un réseau beaucoup plus large où s'intègre l'élément "CFT". Ainsi on a:

POUVOIR ---R1---> PATRONAT

et

(PATRONAT---R2---> POUVOIR)

|

⇓ R3

CFT

avec R1= relation *ET*

R2= relation *avec la complicité*

R3= relation *sujet-->objet*

Cet exemple montre, s'il en est besoin, la nécessité d'intégrer dans tout calcul de liaisons entre formes lexicales la dimension syntaxique qui demeure la base incontestable du jeu lexical.

Ceci autorise à dire que deux formes ne sont liées que par rapport à un type de relation déterminée , et que la force de liaison diffère sensiblement d'une relation à une autre.

Il est donc souhaitable, pour rendre compte quelque peu de l'organisation et de l'engendrement d'un discours, d'affiner les méthodes statistiques des cooccurrences en y adjoignant l'étude des liaisons syntaxiques. Conjoindre les deux méthodes (cooccurrence et syntaxe), c'est leur faire jouer le même rôle de complémentarité qu'entretiennent les études des fréquences pures et celles qui visent la répartition des formes lexicales à travers un corpus de textes. C'est aussi inviter à corriger les réseaux d'attirances statistiques par les réseaux d'attirances syntaxiques entre formes lexicales.

Cooccurrence et syntaxe:

Cette distinction entre cooccurrence et syntaxe ne fait que reprendre celle qu'établit R.L. Wagner entre cooccurrence et corrélation: cooccurrence "dénote le fait purement matériel que les signifiants s'y succèdent", tandis que correlation exprime "un fait de syntaxe" (cité in Tournier, 1980, p.185).

Cependant ces corrélations sont approchées à partir de l'analyse des cooccurrences par "induction ou hypothèses" (Ibid.) où les relations entre formes lexicales sont celles d'*oppositions*, d' *associations* et d' *identité*, signalées par J. Dubois dans sa thèse (1962), relations qui concernent la chaîne paradigmatique des formes lexicales. Leur détermination, si elle donne un "abord" du

discours, "rate" souvent l'effet de la syntaxe dans son déploiement syntagmatique, support des liaisons entre formes lexicales.

Ces remarques étant faites, l'objet de cet exposé sera donc de proposer une méthode de dépistage de liaisons syntaxiques entre formes lexicales, et qui évite les deux écueils propres aux méthodes hors contexte, qui sont:

a) l'affectation préalable de contenus sémantiques
aux formes lexicales (liaison sémantique),

b) la primauté du quantitatif sur le relationnel (liaison lexicale).

Dans la méthode présentée ici, le fait syntaxique élimine d'emblée toute référence à l'idée de co-présence entre formes lexicales, puisqu'au niveau syntaxique deux formes peuvent être séparées dans l'espace de la phrase tout en étant très fortement liées. Il reste toutefois que le corpus théorique de la statistique lexicale (hors et en contexte) peut déboucher sur une "statistique syntaxique" (3) dont l'objet sera la liaison entre formes lexicales prises dans leur environnement syntaxique immédiat (phrase, paragraphe, etc.) ou médiat (texte entier).

But de la procédure

Le réseau lexico-syntaxique (4) -qu'il faut distinguer des réseaux sémantique ou lexical- n'est pas un regroupement ou un agrégat de formes lexicales, mais une configuration où les liaisons entre les formes peuvent être ou non orientées.

La procédure ne vise pas, bien sûr, la découverte d'une structure du discours à partir des positions syntaxiques. Le résultat n'est pas offert tout prêt à la lecture pour en tirer des interprétations immédiates. L'intérêt de cette procédure réside dans le fait de proposer à l'analyste un outil qui lui permette d'obtenir un "objet discursif" particulier: celui d'un ensemble de formes lexicales reliées entre elles par un ensemble de relations syntaxiques. L'analyste reste le seul juge de l'interprétation quant aux affinités autres que syntaxiques entre ces formes.

On peut par contre suggérer que cet objet discursif sera une base appréciable pour toute étude de fonctionnement discursif, notamment sur le plan diachronique: l'évolution des relations syntaxiques entre formes lexicales est-elle le support d'un fonctionnement discursif? Seule l'analyse comparée de corpus à des moments différents peut répondre à cette question.

Démarche

Supposons que l'on parte du couple de positions syntaxiques *sujet / objet*. Le programme cherchera toutes les formes lexicales se réalisant en position *sujet* et *objet*. Ce qui constituera la première série de liaisons lexicales. Tout autre couple de positions syntaxiques (exemple: déterminant/déterminé) déclenchera une autre série de liaisons.

Exemple de sortie

pour la relation R1:

sujet1---R1--->objet1

sujet2---R1--->objet2

sujet3---R1--->objet3 .....

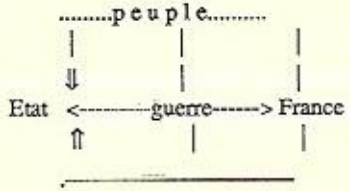
sujetn---R1--->objetn

avec à la place de tout sujet-i et de tout objet-i respectivement les formes en position sujet et objet.

Remarquons que, pour une forme donnée, toutes les formes dépistées qui lui sont liées par la relation R forment une classe d'équivalence au sens harrissien du terme, classe qu'on peut représenter ainsi:

```
                    ! forme-i-1 !
                    ! forme-i-2 !
forme-i------->     ! forme-i-3 !
                     ! ...      !
                     ! ...      !
                    ! forme-i-n !
```

où les formes 1 jusqu'à n constituent la classe d'équivalence "objets" par rapport à *forme-i* qui est dans cet exemple en position "sujet". Par ailleurs la même forme lexicale peut se retrouver à travers le corpus en position *sujet* puis en position *objet*. Ce qui permet de représenter l'ensemble des formes sous l'aspect d'un graphe orienté dont voici un exemple concret, les flèches indiquant le sens sujet--->objet:



c'est-à-dire:

peuple--->France

peuple--->Etat

peuple--->guerre

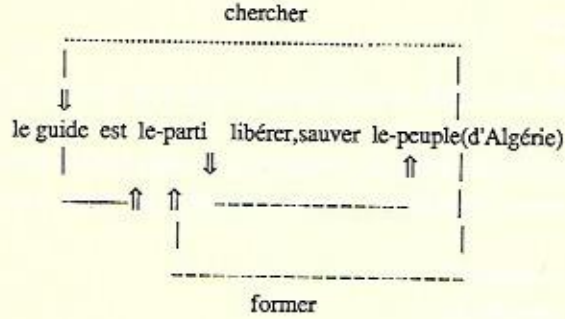
guerre--->France

guerre--->Etat

France--->Etat

Bien entendu, il est possible d'ajouter à ce réseau (sur ces arêtes) le prédicat verbal qui relie chaque fois la position 1 *(sujet)* à la position 2 *(objet)*.

Voici un exemple réduit de ce type de réseau extrait du discours du Parti Communiste algérien (1936-1954), (cf. OUAMARA, 1983):



qu'il faut lire ainsi:

le peuple *cherche* un guide
le guide *est* le parti
le parti ! *libère* ! le peuple
!*sauve* !
le peuple *forme* le parti.

Notons la remarquable cyclicité de ce réseau qui, après substitutions et transitivité, peut formellement être réduit à cet énoncé de base:

le peuple cherche le peuple
ou *le peuple libère le peuple*,

traduction, somme toute, de l'identification du Parti Communiste au peuple algérien.

Mais il s'agit là d'un début d'interprétation hors de notre propos ici.

Réalisation automatique

Le fait de privilégier ici les relations syntaxiques entre formes lexicales suppose, pour organiser celles-ci en réseau, une description syntaxique préalable du discours à analyser.

Pour ce faire, nous avons mis à profit une grammaire de reconnaissance du français, grammaire programmée par P. Plante à l'aide du logiciel DEREDEC dont il est par ailleurs le créateur (Plante, 1983).

Cette grammaire réalise certaines relations syntaxiques entre les formes telles que *thème / propos*, *verbe / complément d'objet*, *déterminant / déterminé*.

Le programme part de cette description syntaxique pour obtenir un réseau lexico-syntaxique où les formes les plus liées syntaxiquement sont représentées dans le réseau avec leurs liaisons (relation syntaxique choisie au départ).

Mais DEREDEC est un langage de programmation qui offre à l'usager la possibilité de programmer sa propre grammaire, d'établir par conséquent les relations qu'il veut analyser entre les formes lexicales. Autrement dit, le réseau est un objet formel qui peut être obtenu avec n'importe quelle grammaire, pourvu qu'elle soit conçue dans ce langage.

NOTES

(1) Cette phrase est réalisée dans le discours de la CGT sous une autre forme: "la politique de démantèlement poursuivie par le pouvoir et le patronat". Nous l'avons donc transformée pour l'exemple.

(2) Cette phrase est réalisée sous sa forme nominalisée: "le recours pour une partie du patronat avec la complicité(...) des milieux du pouvoir aux syndicats maison CFT"

(3) Cf. la première esquisse de ce genre de statistique dans Plante P. "le système de programmation DEREDEC" in Mots n° 6, 1983, Presses de la Fondation Nationale des Sciences politiques, Paris.

(4) Le réseau lexico-syntaxique, quoique les méthodes diffèrent sensiblement, s'apparente à la "grammaire de discours" proposée par Sueur J.P. (1982), plus particulièrement en ce qui concerne les buts visés par les deux procédures: "la grammaire de discours, dit-il, est (...) l'étude systématique de l'intersection entre faits de lexique et faits de syntaxe et d'énonciation" (Ibid., p.148).

REFERENCES

DUBOIS J. (1962): *Le vocabulaire politique et social en France (1869-1872)*, Larousse, Paris

LAFON P. (1981): Analyse lexicométrique et recherche des cooccurrences, in *Mots* n°3, Presses de la Fondation Nationale des Sciences Politiques, Paris

OUAMARA A. (1983): *Quelques procédures automatiques d'analyse linguistique du discours*, thèse doctorat 3ème cycle, Grenoble II

PAILLET J.P. (1974): Problèmes de notation pour l'étude de contenu linguistique, in *Langages*, n° 35, Larousse

PLANTE P. (1983): Le système de programmation DEREDEC, in *Mots* n° 6, Presses de la

Fondation Nationale des Sciences Politiques, Paris

QUILLIAN M.R. (1968): Semantic memory , in *Semantic Information Processing*, Minsky M, Cambridge, M.I.T. Press

SUEUR J.P. (1982): Pour une grammaire du discours. Elaboration d'une méthode; Exemple d'application , in *Mots* n° 5, Presses de la Fondation Nationale des Sciences Politiques, Paris

TOURNIER M. et al. (1978): *Des tracts en mai 68* , Champs libre, Paris

(1980): D'où viennent les fréquences du vocabulaire : la lexicométrie et ses modèles , in *Mots* n° 1, Presses de la Fondation Nationale des Sciences Politiques, Paris