argumentative) la valeur logico-sémantique des phrases. Il s'agit là, on le voit, d'un parti-pris opposé à celui de RM, qui cherche à isoler ces relations logiques que je voudrais plonger dans l'activité de langage. Avant qu'on puisse songer à départager les deux directions de recherche, il faut d'abord que chacune soit systématiquement développée le plus loin possible - de façon à faire apparaître ses possibilités et ses implications, l'intelligibilité qu'elle donne, et les hypothèses qu'elle coûte. Par sa clarté, par son caractère explicite, le livre de RM est une précieuse contribution à cette tâche.

Oswald Ducrot
Paris

Bente Maegaard et Ebbe Spang-Hanssen: *La segmentation automatique du français écrit.* Saint-Sulpice-de-Favières: Association Jean-Favard pour le développement de la linguistique quantitative; Paris: diffusion A. Leson, 1978. 122 p. (*Documents de linguistique quantitative, 35.*)

A travail de bénédictin, compte rendu de bénédictin. L'ouvrage dont il sera rendu compte ci-dessous a été fait par trois spécialistes: une linguiste mathématicienne, un grammairien romaniste et un ordinateur. Devant la somme de science, de données, d'observations de tous ordres qu'ils fournissent, c'est une tâche astreignante que de devoir se limiter. Aussi seuls trois points principaux seront pris en considération ici:

1) une description générale de la méthode proposée;
2) un aperçu de quelques problèmes de détails, observations faites en cours de route par les chercheurs, résultats obtenus;
3) une critique des points 1) et 2).

1) *La segmentation automatique du français écrit* décrit comment on peut faire faire par un ordinateur l'analyse d'un texte en propositions. L'analyse des propositions est une discipline qui, à l'échelle «humaine», présente déjà pas mal d'aléas et d'ambiguïtés: conditions *sine qua non* pour qu'il y ait une proposition, délimitation de la principale, de la subordonnée, mots introduisant les subordonnées, limites entre la subordination et la coordination, propositions elliptiques etc. On imagine donc les difficultés à surmonter pour pratiquer une telle analyse sur ordinateur.

L'hypothèse de travail de BM et ESH est qu'il est possible d'identifier les propositions sans faire appel à l'analyse des fonctions. Ceci présuppose que les propositions sont caractérisées par des marqueurs formels qui permettent de les identifier en tant que telles. Si tous les éléments d'une proposition ne peuvent être identifiés par leurs formes, il est pourtant possible d'en isoler certains qui appartiennent soit à des classes plus ou moins fermées, comme les pronoms relatifs, les conjonctions de subordination, etc. (que les auteurs appellent selon la tradition danoise des «introducteurs de propositions subordonnées»), les conjonctions de coordination, les signes de ponctuation, soit à des classes moins restreintes mais définissables

de façon plus ou moins exhaustive comme les verbes. Cet inventaire de classes peut déjà fournir théoriquement le *terminus a quo* et le *terminus ad quem* d'une proposition subordonnée ou d'une principale (définie alors négativement comme proposition non-introduite). Si on ajoute à cela que le programme peut discerner les formes conjuguées en personne des formes non-conjuguées comme l'infinitif ou les participes, on s'approche de façon indéniable des critères exposés dans les manuels de grammaire traditionnelle, c'est-à-dire que pour avoir une subordonnée française, il faut nécessairement qu'il y ait un introducteur et un verbe conjugué en personne, au strict minimum. (La grammaire oublie souvent de nous indiquer qu'il faut nécessairement aussi un sujet exprimé dans la subordonnée, puisque l'impératif est exclu des subordonnées, mais ceci est une autre histoire...). Le programme SEGMENTATION tire le maximum de cette caractéristique des subordonnées françaises. Les auteurs ont joint à leur programme une série d'analyses des formes appartenant à d'autres classes plus ou moins restreintes: ce sont les déterminants (articles, possessifs, démonstratifs etc.), les prépositions, les particules verbales comme les pronoms personnels conjoints, *en*, *y*, la négation etc., qui occupent en quelque sorte une place fixe de satellites autour du verbe. Les auteurs ont en outre tiré parti du fait que les pronoms personnels sujets conjoints indiquent avec sûreté qu'il doit y avoir un verbe conjugué à proximité (alors qu'un syntagme de nature substantivale placé près du verbe n'en est pas nécessairement le sujet).

Le programme SEGMENTATION consiste alors en deux analyses successives:

(a) *l'attribution de symboles de catégories* aux différentes unités du texte (les unités du texte pouvant être des mots ou des signes de ponctuation, comme la virgule ou le point). Chaque unité est donc affectée d'un symbole correspondant à la catégorie à laquelle l'unité appartient. Pour les unités qui n'ont pas été définies à l'avance, comme l'énorme groupe des substantifs et des adjectifs, le programme attribue la valeur N, c'est-à-dire une valeur neutre ou non-marquée.

(b) *la segmentation de la chaîne de symboles* ainsi obtenue en sous-chaînes représentant les principales et les subordonnées. La représentation des propositions est faite de manière à mettre celles-ci à des niveaux différents selon leur nature. Mieux que la description théorique des faits, un exemple pris à la page 16, et légèrement modifié, illustrera cette analyse:

unités de texte	affectées d'abord de la catégorie:	puis «neutralisées» et réduites en:
Sur	P(préposition)	N(non-marqué)
les	D(déterminant)	N
marches	N	N
de	P	N
l'	D	N
escalier	N	N
qui	I(intr. de subord.)	I
tournait	V(verbe conjugué sans pron. suj)	V
et	E(conj. de coord.)	E
qui	I	I
était	V	V
raide	N	N
,	S(séparateur)	S

elle	N
lui	N
demanda	W(verbe conjugué plus pron. suj) W
:	A(arrêt) A

Le programme donne une première représentation:
NNNNNNIVEIVNSWA
puis une analyse en niveaux:
$NNNNNN/_{IV/EIVNS/}WA.$

L'analyse montre bien qu'il y a deux subordonnées au niveau 2 (ici deux relatives séparées par des barres diagonales) enchâssées dans une principale (niveau 1).

Le verbe étant le pivot de la proposition, il a été nécessaire d'établir une procédure exhaustive de reconnaissance des formes verbales. Pour ce faire, deux lexiques ont été retenus: un lexique de racines verbales (4200 verbes différents, soit 5400 racines différentes) et un lexique de désinences verbales (588) qui permettent de reconnaître la plupart des verbes français employés dans les textes courants. L'établissement de ce dictionnaire verbal est intéressant en soi, car il est sensiblement différent de la présentation ordinaire des verbes français, classés en général selon un grand nombre de conjugaisons. Les auteurs se servent d'un système en vigueur au Danemark: celui des «temps primitifs», employés de façon didactique pour définir un verbe (surtout les verbes irréguliers) et pour faire découler des formes primitives d'autres formes du verbe. Les auteurs établissent ainsi pour la cinquantaine de formes non-composées possibles d'un verbe français quelconque, des sous-ensembles ou paradigmes qui permettent un regroupement des formes offrant des similitudes entre elles. Exemple: le paradigme du futur et du conditionnel a comme sous-groupes, trois conjugaisons, et les désinences respectives *-erai*, *-irai*, *-rai*, et *-erais*, *-irais*, *-rais*, pour la première personne du singulier etc. Le dictionnaire des racines est une liste alphabétique de toutes les racines, comprenant chacune un code de compatibilité avec une ou plusieurs désinences à l'intérieur de chaque conjugaison, regroupée sous un paradigme (il y a 6 paradigmes différents). Les désinences, qui sont elles aussi affectées de codes de compatibilité, sont stockées sous forme de structure arborescente pour permettre une comparaison rapide des chaînes de caractères. La reconnaissance d'une forme verbale se fait donc en lisant le mot par la fin et en comparant la finale en question avec les désinences du dictionnaire. Si le segment trouvé correspond bien à une désinence, la procédure s'enquiert alors de savoir si le début du mot est une racine. Par exemple si *-e* est reconnu comme une désinence possible, le programme cherche si *regard-* est une racine permise, etc. Si nous avons passé beaucoup de temps à rendre compte de cette procédure, c'est qu'elle occupe une place centrale dans le présent ouvrage. En effet, la présentation de la méthode fait l'objet de l'introduction et des deux premiers chapitres.

2) Le troisième chapitre traite de la pierre d'achoppement de toute analyse automatique des textes: le problème des homographes. On nous apprend qu'environ un tiers des formes verbales potentielles sont homographes. Et comme la résolution des homographies présuppose l'analyse de la phrase et que celle-ci, à son tour, présuppose celle des homographes, il y a donc un cercle vicieux à éviter. La phrase *la belle porte le voile* est un bon exemple d'homographie à divers degrés. Pour résoudre de tels problèmes, on peut être acculé à devoir attribuer expérimentalement toutes les catégories morphologiques possibles aux mots homographes de la phrase. En un deuxième temps, un programme se basant sur un modèle de la phrase retient les solutions acceptables et rejette les autres (par exemple une solution où ni *porte* ni *voile* ne

seraient un verbe, dans le cas qui nous occupe). Cette solution au problème de l'homographie est une opération longue et coûteuse. La solution retenue par les auteurs du présent ouvrage est différente. Pour les formes susceptibles de causer des problèmes, le dictionnaire des racines verbales a été nanti de codes d'homographie qui signalent quelles sont les formes ambiguës et à quels tests il faut les soumettre pour lever les ambiguïtes. Les tests employés sont les suivants:
1. le *test des particules verbales* qui permet par exemple de délimiter le groupe des verbes qui ont un pronom personnel pour sujet (il s'avère que c'est le cas pour 54% des verbes).
2. le *test du substantif* permet de classer l'homographe comme un substantif, si le mot est précédé d'un déterminant ou d'une préposition. Dans la négative, le programme cherche si l'on peut trouver comme support du substantif une locution figée du type *chercher querelle, avoir cours,* etc.
3. le *test des participes passés en -s et des adjectifs* (*pris* vs *je pris*) celui *des participes passés en -t* (*dit* vs *il dit*) celui *du participe-substantif fait* (*fait* vs *il fait* vs *le fait*), de la *préposition entre* (vs *il entre*) et des *locutions adverbiales en -ons* (*nous tâtons* vs *à tâtons*) affinent encore l'analyse des homographes.

Le quatrième chapitre délimite la catégorie des introducteurs de propositions subordonnées. Une série de problèmes est attachée à ces introducteurs: certains introduisent aussi des principales interrogatives (tel *quand*), d'autres introduisent des propositions elliptiques (*quoique* malade, il est venu). Enfin le mot *que* est tellement ambigu qu'il cause des problèmes à divers niveaux (plus *que*, ne...*que*). Le mot *comme* pose des problèmes similaires. La conjonction *si* est homographe de *si* adverbe. Il a fallu introduire des solutions *ad hoc*, comme celle de traiter les principales interrogatives comme des subordonnées (p. 54) ou bien introduire des tests: tests du successeur verbal, test de *si*, etc. qui examinent le contexte proche pour trouver des indices qu'il s'agit bien d'une proposition.

Le cinquième chapitre examine le statut de *car*. Ce chapitre est, à mon avis, l'un des plus intéressants. Tout en restant d'accord avec la tradition qui voit dans *car* un introducteur de proposition principale, les auteurs ont au départ considéré *car* comme une classe à part, c'est-à-dire, distincte des conjonctions de coordination et des conjonctions de subordination. Le mot *car* est un mot précieux car il indique, de façon univoque, le début d'une proposition; c'est donc un indice sûr pour un programme tel que SEGMENTATION. Les auteurs discutent en détails la position des recherches récentes sur le statut de *car* (par exemple le groupe λ - l, Barbault *et al*). Ces recherches, basées souvent sur des enquêtes d'ordre sémantique, tendraient à voir soit *car* comme une conjonction de subordination, soit *puisque* comme un équivalent de *car*, donc deux conjonctions de coordination. Les résultats auxquels BM et ESH sont arrivés par l'intermédiaire de leurs analyses sur ordinateur tendent à confirmer les vues traditionnelles sur *car*.

Le chapitre 6 décrit l'analyse proprement dite en segments. Les auteurs ont employé une grammaire qui peut aussi être décrite comme un automate ou machine théorique. Cet automate lit un symbole à la fois et passe, en fonction du symbole lu, à un nouvel état ou reste dans l'état premier. La méthode choisie est celle d'un automate «déterministe bidirectionnel», ce qui veut dire,
(1) qu'à chaque symbole lu il y a une seule action à faire et
(2) que l'automate peut lire de gauche à droite, et, revenant en arrière, de droite à gauche. Ce choix a été fait pour des raisons de commodité et pour rendre la grammaire plus compréhensible au lecteur humain. Les auteurs passent alors en revue les quatre types principaux d'automates connus en grammaire formelle: automates finis, automates à pile (pushdown storage), automates linéaires bornés, machines de Turing. Ils expliquent en termes clairs

(même pour les linguistes non-initiés aux arcanes mathématiques) comment fonctionnent ces machines théoriques. L'automate de segmentation est l'équivalent d'un automate linéaire borné, la grammaire de segmentation équivaut, elle, à une grammaire contextuelle (context-sensitive). Enfin, les auteurs évaluent les propriétés linguistiques de leur grammaire. Ils soulignent l'importance de l'ordre linéaire, c'est-à-dire, la tendance générale à rattacher les éléments de la phrase au verbe qui précède, si verbe il y a. Seul l'emploi des conjonctions et des signes de ponctuation modifie le principe de linéarité de la grammaire de segmentation.

Le chapitre 7 évalue les résultats de l'expérience. Malgré l'attitude extrêmement modeste des auteurs, au cours du livre (ils qualifient à plusieurs reprises leur programme de «grossière analyse»), les résultats sont extrêmement prometteurs: au point de vue quantitatif, les phrases analysées ont été segmentées correctement dans 95,8% des cas, chiffre qui peut être monté à 96,8% si l'on fait abstraction des erreurs de niveau. Plus du tiers des phrases du corpus analysé (6 échantillons de 50 pages d'auteurs contemporains) comporte au moins une subordonnée. Cette observation, parmi tant d'autres, est aussi de poids. Au point de vue qualitatif, les auteurs ont observé des différences appréciables entre les échantillons. Ainsi le nombre d'échecs très réduit pour les textes de Sartre et de Cayrol devient considérable pour les textes de Nizan et de de Gaulle. Des figures 9 et 10 (pp. 86-87), on peut tirer une série d'autres observations de nature stylistique sur les prosateurs en question (nombre de phrases par échantillon, nombre de mots par phrase etc.). Ceci prouve bien que les recherches sur ordinateurs peuvent aussi donner des corollaires appréciables pour ce qui est de la stylistique. Les causes d'erreurs dans le cas de la prose de Nizan et de celle de de Gaulle proviennent du plus grand nombre d'énumérations, de l'emploi qu'ils font des incises ainsi que du *que* comparatif.

3) Je pense avoir montré l'importance de l'ouvrage de BM et ESH et les perspectives nombreuses en grammaire, en statistique et en stylistique qu'offre leur travail. Le livre est rédigé dans une langue claire et précise. L'approche des nombreux problèmes traités est faite de façon pédagogique. Les concepts nouveaux sont définis chaque fois (j'ai pourtant quelques réserves à faire quant à l'emploi de *paradigme* vs *conjugaison*, pp. 23, 24 et ss.). Les considérations théoriques sont étayées d'exemples. J'aurais aimé voir citer plus d'exemples au chapitre 4, notamment pour le traitement de *comme*, celui de *que* (p. 57) et le cas de «deux introducteurs potentiels» (p. 59). La liste des formes verbales homographes donnée en appendice a été établie manuellement. (Il me semble que ce travail fastidieux aurait pu être évité en grande partie en compulsant le vol. IV du *Dictionnaire des Fréquences du TLF*, 1971). Je m'étonne de ne pas trouver dans cette liste des homographes aussi fréquents que *été*, *suis*, *maintenant*, et, à un degré moindre, *subit*, *revirent*, etc. J'aurais aimé avoir des éclaircissements sur la délimitation entre V et W (pp. 15 et 16). La figure 1 que j'ai reprise plus haut ne m'explique pas entièrement comment *elle lui demanda* donne NNW puis W seul. Ceci suppose une étape dont il n'a pas été fait mention explicitement. Les auteurs auraient aussi pu porter en appendices les listes exhaustives des mots grammaticaux (en comparant les pages 21 et 42-43, je ne trouve que 127 mots grammaticaux sur 230 de déclarés). De même, les listes des introducteurs de subordonnées et des particules verbales pourraient être utiles pour d'autres chercheurs. Le mot *en*, qui est une particule verbale (cf. diagramme pp. 37 et 39) est-il aussi parmi les mots grammaticaux de haute fréquence? Y a-t-il rapport d'inclusion entre ces deux listes? Mon expérience personnelle avec des listes de fréquences pour d'autres périodes du français me fait penser qu'il faut répondre oui à la première question. Mais comment a-t-on résolu l'homographie riche de *en* pronom et adverbe et de *en* préposition, etc? Enfin, l'ensemble des formes verbales non-composées du français serait de 54 (p. 23). En me frayant un chemin à travers les

paradigmes verbaux des pages 92-97, je n'en trouve toujours que 51 (comme feu le professeur Togeby dans *Fransk Grammatik*, p. 350). Il s'agit peut-être d'une faute de frappe?

Ces remarques finales de détails n'infirment en aucune façon l'impression extrêmement favorable que j'ai du présent ouvrage.

Suzanne Hanon
Odense

Bibliographie

Barbault, M. C., O. Ducrot, J. Dufour, J. Espagnon, C. Israel, D. Manesse (1975): «Car, parce que, puisque». *Revue Romane X*, 2. 248-280.

Caput, J. et J.-P. (1975): *Dictionnaire des verbes français*. Larousse, Paris.

Dictionnaire des fréquences. Vocabulaire littéraire des XIX^e et XX^e siècles. (1971) vol. IV. Table de répartition des homographes. C.N.R.S. - T.L.F. Didier, Nancy.

Faïk, Sully (1978): «*Car, parce que* et *puisque* dans les dictionnaires de fréquence.» *Le français moderne* 46. 143-156.

Hughes, Michel (1972): *Initiation mathématique aux grammaires formelles.* Larousse, Paris.

Maegaard, Bente og Ebbe Spang-Hanssen (1975): «Hvordan kan en datamat finde verberne i en fransk tekst?» *RIDS* 36, 1-31.

Vauquois, Bernard (1975): *La traduction automatique à Grenoble*. Documents de linguistique quantitative no 24. Dunod, Paris.

Conrad Sabourin et John Chandioux: *L'adverbe français: essai de catégorisation.* Saint-Sulpice de Favières, éd. Jean-Favard, 1977, 131 p.

Sabourin et Chandioux ont examiné 1.400 adverbes en *-ment* par rapport à 35 propriétés syntaxiques, sémantiques et morphologiques. Ils ont ensuite eu recours aux méthodes statistiques pour déterminer les relations qui existent entre ces propriétés.

Après une courte présentation de leur travail (p. 11-13), les auteurs rendent compte dans le chapitre 2 (p. 15-20) de leur méthodologie. Dans le chapitre 3 (p. 21-40), ils présentent les tests 1-31. Les tests 32-35, qui sont purement morphologiques, ne nécessitent pas de commentaires. Sabourin et Chandioux exposent dans le chapitre 4 (p. 41-52) les méthodes statistiques, et ils commentent brièvement les résultats. Le reste du livre comprend deux annexes. L'annexe 1 (p. 53-94) donne, en dehors d'une liste des tests utilisés, les résultats de la première partie de leur étude: une matrice qui, à l'aide de + et de ÷, décrit les rapports entre les adverbes et les tests. L'annexe 2 (p. 95-128) indique les corrélations entre les tests.

Les auteurs n'avaient pas, en premier lieu, pour but d'étudier les adverbes, mais ils ont tout simplement choisi cette partie du discours pour illustrer leur méthode. Cela signifie qu'ils n'ont pas commenté les relations, révélées par les méthodes statistiques, qui existent entre les tests. Ils se sont contentés de présenter ces méthodes.