Textologie de système

par

Pavel VASAK

Académie Tchèque

En observant l'évolution de la linguistique mathématique dans les vingt dernières années, on peut distinguer des moments différents, en commençant par la période de l'enthousiasme pour la théorie de l'information et le calcul des fréquences de différents symboles linguistiques et en finissant par la période actuelle de la linguistique computationnelle. Dans son article publié en 1965, Y. Bar Hilel se souvient des années cinquante aux Etats-Unis, où plus ou moins partout on calculait les fréquences : "Nous avons été tous profondément impressionnés par les expériences de Shannon avec les approximations de l'anglais courant et nous avons été persuadés que la langue . . . était un processus de Markov. /. . ./ Tout le monde était convaincu qu'une seule chose manquait à la compréhension totale des processus de communication dans les langues naturelles : les données statistiques sérieuses sur les fréquences relatives des digrammes, des trigrammes etc. et nous avons regretté que le calcul de ces données fût une affaire aussi coûteuse(1).

Il me semble que quelquefois, on se trouve dans une situation comparable. Grâce aux ordinateurs on a rassemblé des quantités immenses de données sans réfléchir sur leur utilité, sur la nécessité de leur interprétation. Rappelons les idées de l'exposé inaugural de R. Busa qui s'est posé la question essentielle : Et après ? Rappelons aussi les paroles de P. Tombeur quand il a parlé de K.O. Donc je pense que le problème ne réside pas seulement dans l'utilisation de méthodes mathématiques plus ou moins développées mais aussi dans le niveau théorique d'où nous observons la réalité. Or, la question se pose : Et avant ? Ce n'est pas la mathématique qui pose dans ce contexte les questions, mais la réalité linguistique, littéraire etc. S'il en était ainsi, les résultats du calcul linguistique de l'ordinateur ne seraient pour nous qu'un texte d'utilité très limitée.

I.- Dans la première partie, je vais présenter quelques idées théoriques, un modèle qui constitue une base pour l'étude du processus auteur - texte - société. La première variante du modèle a été présentée pendant le colloque intitulé "*Lexicologie politique du français moderne*" qui s'est tenu en 1980 à Saint-Cloud.

Je ne parlerai pas de la première phase du processus en question qui est l'auteur. Dans le livre intitulé *Méthodes d'attribution des textes*(2), j'ai présenté les différentes méthodes (linguistique, stylistique, littéraire, mathématique etc.) de sa détermination. Par attribution, j'entends le processus qui consiste à déterminer si un élément X appartient ou non à un ensemble M. L'ensemble M, appelé ensemble de comparaison, peut être formé sur la base de différents critères, comme l'ensemble des auteurs, styles, écoles littéraires, périodes linguistiques etc. Conformément à la structure concrète de l'ensemble M, je distingue l'attribution d'auteur et de non-auteur. L'attribution de non-auteur consiste dans l'attribution du texte X à une langue ou à un style, à une période du temps, à une école littéraire, à une génération, etc. De la même manière, l'attribution d'auteur consiste dans l'attribution du texte X à une certaine personne. Revenons au processus auteur - texte - société.

Méthodologiquement, je pars de la notion cybernétique du système considéré comme l'ensemble M des éléments et l'ensemble R des relations entre ces éléments. L'ensemble R est appelé la structure

du système. Les relations sont de différentes sortes, par exemple matérielles, énergétiques et informatiques. Pour les systèmes littéraires et linguistiques, les relations informatiques sont typiques. Et chaque système exerce une fonction de base qui tient les éléments ensemble. La deuxième notion de base que j'utilise est la textologie. Ce terme a été créé dans les années vingt par le textologue soviétique Boris Tomasevskij pour remplacer la notion de critique des textes(3). La textologie dans son principe contient la conception de système du processus textuel. Généralement, la textologie est considérée comme une science (littéraire et linguistique) qui étudie l'histoire, la genèse et l'attribution du texte dans le cadre des relations qui influencent le texte et ses sources. Le textologue veut reconstruire la totalité du processus textuel qui a historiquement existé. Ce processus répond à la genèse du texte et à sa diffusion communicative, c'est-à-dire à la naissance du texte de l'oeuvre et au processus de sa réception - autrement dit, le changement du texte en oeuvre. C'est seulement après la réception que le texte devient oeuvre; jusqu'à la réception, il n'était que l'oeuvre potentielle.

Pourquoi parle-t-on de système ? Le texte de l'oeuvre n'est pas statique, il est toujours en mouvement, comme les textes de travail dans la phase de la création, les corrections d'imprimerie, les changements exigés par les rédacteurs etc. le prouvent. C'est pourquoi on ne peut pas étudier le processus de la genèse du texte et sa diffusion communicative à partir d'une source textuelle isolée. Il faut se rendre compte de la totalité du processus mentionné, où chaque élément remplit une fonction précise et se trouve en relation avec les autres éléments. Le processus auteur - texte - société doit donc être traité comme une totalité fonctionnelle. Ce sytème est composé de deux sous-systèmes :

1. la genèse du texte : la naissance de l'oeuvre potentielle;
2. la diffusion communicative du texte : la réception de l'oeuvre. Le texte est l'élément dans lequel ces deux sous-systèmes se recouvrent. Il est tout d'abord le résultat du processus de la genèse, ensuite, il est le point de départ de la diffusion communicative.

Quels sont les éléments du système de la genèse du texte ?
Ce sont tous les textes ou plutôt pré-textes que l'on voit naître dans la phase auteur - texte(4). Ils ont la valeur autocommunicative, étant en principe destinés "seulement à l'auteur". Ce sont les différents textes de travail (brouillons, etc.) avec lesquels l'auteur ne veut pas entrer en communication. Cette phase est achevée par le texte linguistique terminé que l'auteur destine à la communication (l'autographe de l'oeuvre)(5). Le système de la genèse est donc décrit par deux ensembles :

1. l'ensemble M des sources textuelles;
2. l'ensemble R des relations entre les sources, qui exprime la transmission de l'information textuelle d'une source dans une autre.

Chaque élément de l'ensemble M est un lieu où l'information transmise par une source donnée est traitée. La relation entre l'information conservée et l'information modifiée est généralement exprimée par le nombre des variantes textuelles entre les textes comparés. L'ensemble R des relations entre les sources, qui exprime la dépendance textuelle, est identique à la structure du système, sa représentation

graphique s'appellant stemma.

Il existe plusieurs méthodes de construction du stemma, voir par exemple les méthodes de typologie utilisées dans la stylistique statistique(6). Dans le livre *Méthodes d'attribution des textes*, j'ai utilisé, premièrement, la modification de la méthode de J. Froger et, deuxièmement, une méthode de la stylistique statistique appelée la taxinomie de Wroclaw, basée sur la notion de distance entre les sources textuelles(7).

II.- Le but de la deuxième partie est l'étude de la diffusion communicative du texte, c'est-à-dire la reconstruction du système de la réception de l'oeuvre. Dans un certain sens, nous voulons reconstituer l'histoire littéraire par un procédé objectif, c'est-à-dire reconstruire les relations littéraires (sociales) qui ont réellement existé. Le texte est situé au début du processus de la diffusion, étant prêt à entrer dans le système social et dans ses contextes différents. Quelles sont les traces de l'entrée du texte dans le système social ? Cette entrée est matérialisée dans les documents de la réception. Chaque document exprime le rapport (l'interaction) entre le texte reçu et l'auteur du document.

On peut distinguer les documents publics et non-publics, ou fixés comme un texte et non-fixés. Par les documents fixés en tant que textes, nous comprenons les discours écrits dans la langue naturelle (par exemple comptes rendus, articles, nécrologie, lettres, documents de la censure etc.). Dans la situation historique concrète, le système de la réception, décrit par deux ensembles, s'est donc formé :

1. l'ensemble M, composé des documents sur la réception du texte de l'oeuvre T;
2. l'ensemble des réactions R qui expriment l'interaction entre le récepteur et le texte reçu.

L'objet de notre recherche est donc l'ensemble des documents M, composé d'éléments qui réagissent à l'oeuvre étudiée. Je laisse à part les critères littéraires pour la construction de l'ensemble M.

Par l'analyse de la langue de ces documents, nous voulons découvrir la structure du système de la diffusion du texte de l'oeuvre étudiée. Comme il a été déjà dit, chaque document est une expression de l'interaction entre le récepteur (l'auteur du document) et le texte reçu. Le niveau de l'interaction (relation) est fixé et concrétisé dans le texte du document, généralement par la dénomination (désignation) linguistique. Nous distinguons deux catégories de dénomination :

a) par le nom propre;

b) par le nom commun.

Par la langue des noms propres, la relation entre l'auteur du texte reçu et son contenu d'un côté, et d'autres auteurs, mouvements littéraires, idéologiques, politiques etc. de l'autre est exprimée. Chaque nom propre cité dans un texte du document porte un sens qui reflète la structure du système de la diffusion du texte de l'oeuvre étudiée. Le nom cité interprète le niveau de l'interaction entre le texte et le récepteur.

Pareillement, la caractéristique par le nom commun-significatif (substantifs, adjectifs, verbes, etc.) de l'auteur et du texte reçu exprime le niveau de l'interaction entre le contenu du texte reçu d'un côté et le contenu des contextes esthétiques, philosophiques, littéraires etc. de l'autre. Le point de départ est toujours le corpus des documents, dont nous avons déjà parlé.

Quant à l'analyse des noms propres, nous partons de la matrice de citation qui est construite d'après deux critères : dénomination et catégories de temps (voir annexe).

L'hypothèse nulle : il n'y a pas de différences dans la dénomination dans le temps.

Ce procédé statistique nous permettra de trouver le ou les moments de temps où la réception de l'oeuvre - exprimée par la citation des noms - s'est changée d'une façon significative. Ce moment est considéré comme un tournant littéraire - évidemment du point de vue de l'oeuvre étudiée.

La phase suivante est la matrice de co-occurrences (voir annexe).

Par la co-occurrence des noms *m, n*, nous comprenons leur citation dans un même document. L'analyse est surtout basée sur les groupes binaires. Pour chaque catégorie de temps utilisée plus haut, nous construisons une matrice de co-occurrences. Les questions suivantes se posent :

1. L'hypothèse nulle : il n'y a pas de différences en co-occurrences dans le temps;
2. Par quelles dénominations l'entrée du texte dans le système social a-t-elle été exprimée et mesurée ?

   Pour cela, nous utilisons actuellement les catégories f_j, f_d, d (voir annexe). J'ai construit le coefficient qui unit ces catégories :

   $$K = \frac{f_d}{d} \cdot f_j$$

   Quant à son interprétation, il unit l'ensemble communicatif des documents de la réception avec leur contenu. Il y a un macro-monde des documents en tant qu'unités communicatives d'un côté et un micro-monde de leur contenu.
3. Le nombre des co-occurrences de chaque dénomination *m, n* exprime le niveau de leur interaction. On ne peut rien dire sur le sens de l'interaction : la dénomination est-elle citée d'une manière positive ou négative ? Donc, finalement, nous trouvons dans chaque ligne de la matrice de co-occurrences, le maximum $\max_1(d)$. Par ce procédé, nous avons trouvé les dénominations en interaction la plus intense.

Sur la base des distances maximales que nous avons trouvées, nous construisons l'arbre (graphe). Il exprime la structure de l'interaction au point de vue de dénomination dans le système de la réception de l'oeuvre en question.

J'ai utilisé ce procédé pour l'étude de la réception d'une oeuvre de la littérature tchèque du siècle passé (K.H. Mácha).

NOTES

(1) Y. Bar Hilel, Kybernetika a lingvistika, In : Kybernetika ve společenských vědách, Praha 1965, 255-266.

(2) P. Vášák, Metody určování autorství, Praha 1980.

(3) B. Tomaševskij, Pisatel i kniga, Léningrad 1928.

(4) Voir aussi le terme "avant-texte" utilisé en France. Pour plus de détails, voir la série Textes et manuscrits, publiée par Louis Hay :

1.- Essais de critique génétiques, Paris 1979;
2.- Flaubert à l'oeuvre, Paris 1980. Voir l'épilogue de L. Hay, La critique génétique : origines et perspectives.

(5) Pour plus de détails, voir P. Vašák, Textologie et modèle de communication, Revue du LASLA, 1979, n. 4, 31-45.

(6) Voir par exemple P.M. Alekseev, Computational linguistics and quantitative typology of text, In : Symposium Computational Linguistics and Related Topics, Tallinn 1980; voir la série Statistika reci i avtomaticeskij analiz teksta, Léningrad 1971, 1973, 1974, 1980, publiée par R.G. Piotrovskij. Voir V.I. Perebejnos et col., Statystyčni parametry styliv, Kiev 1967; Voprosy statističeskoj stilistiki, Kiev 1974.

(7) Voir J. Froger, La critique des textes et son automatisation, Paris 1968; P. Vašák, Tekstologija, sistema, stemma, In : Prague Studies in Mathematical Linguistics 7, Praha 1981, 137-147.

MATRICE DE CITATION

DÉNOMINATION	CATÉGORIES DE TEMPS							
	1	2	3	.	.	.	m	$\sum$
i_1	n_{11}	n_{12}		.	.	.	n_{1m}	$n_{1.}$
i_2	n_{21}			.	.	.	n_{2m}	$n_{2.}$
.								
.								
.								
i_k	n_{k1}	n_{k2}		.	.	.	n_{km}	$n_{k.}$
$\sum$	$n_{.1}$	$n_{.2}$		.	.	.	$n_{.m}$	$n_{..}$

MATRICE DE CO-OCCURENCES

	DÉNOMINATION								
DÉNOMINATION	i_1	i_2	. . . .	i_k	f_j	f_d	d	K	
i_1									
i_2									
.			n_{ij}						
.									
.									
.									
i_k									

n_{ij} .... nombre de documents qui citent en même temps i_i i_j

f_j .... nombre de documents qui contiennent la dénomination i_j

f_d .... nombre de tous les groupes qui contiennent la dénomination i_j

d .... nombre de groupes différents