Chapitre 03 : Définition et base de la typologie (texte , typologie)

Introduction

De nombreux travaux en description de « types de texte », dans le cadre des linguistiques de corpus, ont rendu possible une nouvelle forme de description de la variation interne d'une langue et une nouvelle forme de typologie, fondée sur des données quantitatives[1]. Les Enjeux de l'identification et de la description de types de textes sont devenus essentiels pour les « traitements automatiques » des langues, puisqu'ils permettent de concevoir des outils non plus pour une langue, mais pour un type de texte, et d'accroître ainsi leur qualité. Mais ces enjeux sont tout autant essentiels pour la description linguistique elle-même, puisqu'ils font apparaître de nouveaux observables et per- mettent de reconsidérer des questions comme celles de la relation entre système et variation, de la distinction des niveaux de description, ou de la relation entre description « quantitative » et « qualitative ».

Or, la plupart des travaux en typologie textuelle s'appuient sur un seul niveau de description – généralement le niveau morphosyntaxique – en considérant qu'il représente l'ensemble des propriétés empiriques des textes qui sont l'objet de la description. On peut cependant se demander pour quelle raison une norme (un sociolecte ou un idiolecte) devrait être relative à un seul niveau de description. Le choix des critères de classification paraît en partie arbitraire. En outre, ces critères ne permettent pas de sortir de classifications *ad hoc* pour établir de véritables typologies : un seul niveau permet certes le plus sou- vent de mettre au jour des variations et de réaliser des classifications automatiques entièrement relatives à des corpus où la variation est restreinte, mais n'ont pas permis, jusqu'à présent, de fonder des typologies. On peut faire l'hypothèse, à la fois théorique et méthodologique, qu'une norme doit être caractérisée par l'ensemble des niveaux de description et que les phénomènes qui permettraient de la définir peuvent relever de corrélations entre niveaux, plutôt que de phénomènes définis dans un seul niveau. Il est dès lors arbitraire de limiter la description à la prise en compte d'un seul niveau. En d'autres ter- mes : pourquoi fonder des typologies textuelles sur des données qui représentent à l'évidence une part si faible des propriétés empiriques des textes ?

Cette question trouve un écho direct dans l'état de l'art en linguistique de corpus. À la limitation à des typologies « mono-niveau » sur un plan méthodologique correspond une difficulté sur le plan technique à construire les corpus nécessaires à l'investigation des relations entre niveaux.

Alors que de nombreux niveaux de description peuvent aujourd'hui être annotés automatiquement (au moins les niveaux morphologiques, morphosyntaxiques, syntaxiques et lexicaux), que des progrès significatifs à court terme de ces instruments d'annotation sont peu probables, il est difficile de constituer des corpus articulant plu- sieurs niveaux de description. La « sortie » d'un instrument d'annotation est le plus souvent utilisée, sans autre forme d'enrichissement, comme l'objet à décrire. Cela induit une dépendance immense à l'instrument, non pas simplement vis-à-vis de ses erreurs d'annotation, nimême des choix théoriques de son concepteur mais plus profondément vis-à-vis de la nature même de l'objet empirique décrit.

Dans cet article je voudrais explorer les enjeux, pour la linguistique textuelle et les typologies textuelles, des nouveaux observables que per- mettent de construire les linguistiques de corpus.

1. Norme et description quantitative

La notion de norme est utile pour fonder l'utilisation de quantifications dans le cadre des linguistiques de corpus. Elle permet, d'une part, de dépasser l'opposition entre « méthodes quantitatives » et « méthodes qualitatives » et d'autre part d'articuler les moyens de description qu'apportent les linguistiques de corpus à un programme de typologie textuelle.

La norme (Coseriu 1982) rend compte des régularités qui ne relèvent pas du système fonctionnel de la langue (d'oppositions distinctives), mais qui possèdent cependant une systématicité et caractérisent une tradition :

> La norme est un ensemble formalisé de réalisations traditionnelles ; elle comprend ce qui « existe » déjà, ce qui se trouve réalisé dans la tradition linguistique ; le système, par contre, est un ensemble de possibilité de réalisation ; il comprend aussi ce qui n'a pas été réalisé, mais qui est virtuellement existant, ce qui est « possible » […][2].

L'opposition entre la norme et le système est donc une opposition entre la réalisation traditionnelle de l'activité de parler[3] et la systématicité fonctionnelle. La norme rend compte de nombreux phénomènes allant des phénomènes collocatifs[4] jusqu'aux genres et aux discours eux-mêmes[5].

La norme n'est certes pas définie quantitativement, puisqu'elle rend compte d'une systématicité. La quantification n'est pas en elle- même une description si l'on ignore les unités

de mesure et les objets décrits. Dans une analyse des conditions de l'utilisation des quantifica tions dans les sciences, Bachelard montre la nécessité d'articuler données

Quantitatives et objets théoriques préconstruits[6] ; de construire la mesure plutôt que de la prendre comme « intuition directe d'un objet » : « l'objectivité est […] affirmée en-deça de la mesure, en tant que méthode discursive, et non au-delà de la mesure, en tant qu'intuition directe d'un objet. Il faut réfléchir pour mesurer et non pas mesurer pour réfléchir »[7].

Cependant, si la norme n'est pas instituée par un fait de fréquence, elle est intimement liée à une dimension quantitative et des phénomènes quantitatifs sont indispensables à sa description[8]. La norme est en effet ce qui n'existe pas sans une fréquence et une attestation. Cette notion a une affinité profonde avec la dimension quantifiable des phénomènes linguistiques. Ce n'est donc pas la fréquence qui fait la norme mais au contraire la norme qui est nécessaire à l'étude de la fréquence et à l'utilisation de données quantitatives.

Bachelard souligne également les dangers du recours à des précisions quantitatives pour caractériser un objet insuffisamment défini. Des données quantitatives utilisées sans notion scientifiquement construite de l'objet décrit, présentées comme une appréhension immédiate de l'Object, relève du « pittoresque »[9]. Bachelard (1993, p. 255) rapporte par exemple, au XVIII^{e} siècle, les chiffres de Buffon, qui arriva à la conclusion qu'il y avait « 74 832 ans que la Terre avait été détachée du soleil par le choc d'une comète […] Cette prédiction ultra précise du calcul est d'autant plus frappante que les lois physiques qui lui ser- vent de base sont vagues et plus particulières. »

Ces exigences de tout raisonnement scientifique quant à l'usage de données quantitatives doivent être poussées plus loin dans le domaine des sciences humaines – ne serait-ce que du fait du danger, précisément, d'y adopter les normes des sciences « dures »[10]. Contrairement aux sciences « dures » en effet, les points de vue ou les niveaux de descriptions ne convergent pas dans les sciences humaines, et l'objectivité ne peut consister à appréhender un même objet à partir de différents points de vue.

La question de l'utilisation non critique de données quantitatives est particulièrement cruciale pour les typologies textuelles et les classements en types de textes. En effet, les approches émergentistes des typologies textuelles sont fortement exposées aux critiques de Bachelard : on quantifie avec une grande précision des phénomènes (des classes de textes) dont on ne définit pas par ailleurs la nature, et on fait émerger par la seule mesure des objets qui restent arbitraires. Toutes les expériences concordent en effet pour montrer que, quels que soient les

critères et les corpus employés, les textes se regroupent en classes[11] ; mais ces classes ne sont jamais les mêmes d'un corpus à l'autre et ces expériences ne semblent pas faire progresser vers l'identification des objets de la description. En ce sens, l'accumulation de données quan titatives est un obstacle à la commutativité de la description.

La notion de norme est donc essentielle à un programme de typologie textuelle puisqu'elle permet d'inscrire la mesure d'une « fréquence » dans la description d'une systématicité. Elle porte précisément sur les propriétés du « réalisé », de l'attesté, en tant qu'il est distinct d'un système fonctionnel. Dans une perspective de typologie textuelle, il est utile de distinguer différents types de normes pour articuler plusieurs critères typologiques. Rastier (2001) propose pour organiser cette diversité de distinguer notamment le discours, le genre et l'idiolecte. Chaque texte est caractérisé relativement à chacun de ces types de normes. Ces différentes normes déterminent donc autant de dimensions d'un

« Espace des normes » dans lequel tout texte possède des coordonnés, semblables aux dimensions de l'espace variationnel.

Ces propositions permettent de souligner la nécessité de prendre en compte la pluralité et l'interaction des normes dans la description.

En effet, une norme n'est jamais observable hors d'une situation d'inter- action avec d'autres normes. Par exemple si l'idiolecte peut être l'objet d'une large variation au sein du discours littéraire, c'est en partie une caractérisation du discours littéraire en tant que discours : c'est le ni- veau du discours qui autorise la variation des idiolectes et ceux-ci, en tant qu'ils varient, caractérisent le discours littéraire et non les idiolectes eux-mêmes. De la même façon, lorsque Brunet (2004) montre que la variation entre genres surdétermine la variation entre auteurs à l'intérieur du théâtre classique (les textes d'un même genre sont classés en- semble, même s'ils appartiennent à des auteurs différents), il faut peut- être rapporter cette articulation entre idiolecte et genre au discours littéraire, voire au champ générique du théâtre classique. En tout cas, elle ne peut caractériser la relation entre genre et idiolecte dans l'absolu.

Une description d'une norme peut donc gagner à s'inscrire dans le cadre d'une architecture des normes et prendre en compte l'effet des autres normes sur les textes étudiés. En somme, il s'agit de contextualiser les normes.

Finalement, c'est l'opposition entre « méthodes quantitatives » et

« méthodes qualitatives » que la notion de norme permet d'essayer de dépasser. « Quantitatif » et « qualitatif » n'opposent pas des méthodes (sinon dans un sens général, non scientifique mais plutôt seulement technique) mais des types de données. Les méthodes sont bonnes ou

mauvaises en elles-mêmes indépendamment du type de données.

En effet, une norme n'est jamais observable hors d'une situation d'inter- action avec d'autres normes. Par exemple si l'idiolecte peut être l'objet d'une large variation au sein du discours littéraire, c'est en partie une caractérisation du discours littéraire en tant que discours : c'est le ni- veau du discours qui autorise la variation des idiolectes et ceux-ci, en tant qu'ils varient, caractérisent le discours littéraire et non les idiolectes eux-mêmes. De la même façon, lorsque Brunet (2004) montre que la variation entre genres surdétermine la variation entre auteurs à l'intérieur du théâtre classique (les textes d'un même genre sont classés en- semble, même s'ils appartiennent à des auteurs différents), il faut peut- être rapporter cette articulation entre idiolecte et genre au discours littéraire, voire au champ générique du théâtre classique. En tout cas, elle ne peut caractériser la relation entre genre et idiolecte dans l'absolu.

Une description d'une norme peut donc gagner à s'inscrire dans le cadre d'une architecture des normes et prendre en compte l'effet des autres normes sur les textes étudiés. En somme, il s'agit de contextualiser les normes.

Finalement, c'est l'opposition entre « méthodes quantitatives » et « méthodes qualitatives » que la notion de norme permet d'essayer de dépasser. « Quantitatif » et « qualitatif » n'opposent pas des méthodes (sinon dans un sens général, non scientifique mais plutôt seulement technique) mais des types de données. Les méthodes sont bonnes ou mauvaises en elles-mêmes indépendamment du type de données.

2. Normes et pluralité des niveaux de description

Relativement au projet de décrire quantitativement des normes, entendues comme des genres, des discours ou des idiolectes, on peut émettre l'hypothèse que des régularités sont d'autant plus définitoires et caractérisantes pour une norme qu'elles impliquent plusieurs niveaux de description et qu'elles stabilisent la relation entre niveaux de description. En réduisant l'observation à un seul niveau de description, non seulement on réduit drastiquement les régularités observables en se passant de la combinatoire entre niveaux, mais surtout on s'interdit d'observer des régularités qui caractérisent beaucoup plus fortement une norme que celles qui s'inscrivent dans un ensemble déjà pourvu d'une systématicité fonctionnelle. Plus généralement, la distinction de niveaux de description n'a de justification que parce qu'elle permet d'observer la systématicité fonctionnelle. Il n'est donc pas justifié de se situer à

l'intérieur de l'un de ces niveaux pour caractériser la norme ; c'est au contraire par la relation qu'elle établit entre niveaux que la
norme est nécessaire comme objet descriptif.

La nécessité d'accéder aux corrélations entre niveaux de description est aujourd'hui soulignée dans de nombreux secteurs de la discipline. Ainsi, depuis une perspective variationniste, Gadet écrit :

[La] perspective qui [prend] en compte les énoncés selon des principes de différents ordres, devrait renouveler la définition des genres en les montrant comme des faisceaux de paramètres, et non plus des rubriques rhétoriques ou situationnelles héritées de la tradition[12].

Dans le cadre du traitement automatique des langues, Habert et Zweigenbaum soulignent de même que

[Le traitement automatique effectif des langues] enjoint aussi de munir les données attestées d'annotations fines, multiples, permet- tant de progresser vers les régularités sous-jacentes[13].

Cette problématique est également présente dans le domaine des corpus oraux, où des dispositifs et des formats de représentation sont proposés pour permettre l'articulation de niveaux de descriptions et de modalités (Bird et Liberman 2001). Blache *et al.* soulignent ainsi :

From a linguistic standpoint, language and speech analysis are based on studies of distinct research fields, such as phonetics, phonemics, syntax, semantics, pragmatics or gesture studies. […]. The perspective adopted by modern linguistics is a considerably broader one : even though each domain reveals a certain degree of autonomy, it cannot be accounted for independently from its interactions with the other domains. Accordingly, the study of the interaction between the fields appears to be as important as the study of each distinct field[14].

Ou Blanche-Benveniste :

On peut enfin produire des analyses distributionnelles étendues à de vastes contextes, sur lesquels on peut mesurer l'effet des collocations entre prosodie, lexique et grammaire[15].

Enfin, dans le domaine diachronique, Coseriu (2007, chap. IV) souligne cette interdépendance entre niveaux : « […] il y a une intime

solidarité entre le phonétique, le lexical et le grammatical ; ce qui, dans la perspective diachronique, signifie qu'un changement affectant n'im- porte lequel de ces aspects possède des répercussions sur l'ensemble du système »[16].

Pour la description de normes en corpus, l'accès aux corrélations entre niveaux de description permet de proposer de nouveaux observables voire de nouvelles catégories descriptives. En effet, les catégories descriptives de la linguistique textuelle, comme les notions d'isotopies, de rythme, les catégories de l'analyse actantielle, ou la modélisation du contexte, permettent de caractériser le fonctionnement sémantique profond d'un texte. Or, ces catégories descriptives peuvent aujourd'hui être constituées en observables, en corpus, et être utilisées dans une perspective typologique.

Dans Loiseau, 2007 j'ai proposé plusieurs observables inspirés des catégories descriptives de la linguistique textuelle pour caractériser par des phénomènes quantitatifs le fonctionnement d'un discours.

Par exemple, dans l'œuvre de Gilles Deleuze, la prise en compte de la linéarité (la « tactique ») de phénomènes (lexicaux comme grammaticaux) a permis de caractériser le paragraphe comme unité : le début et la fin des paragraphes opposent de manière récurrente et significative des registres, des acceptions de certains lexèmes, et des orientations axio- logiques. Le paragraphe, peu pris en charge et au statut controversé, a pu être caractérisé dans le corpus étudié comme un palier sémantique- ment fort : l'enjeu est important, puisqu'il s'agit là de qualifier de nouvelles unités linguistiques, notamment supra-phrastique, dans le contexte d'une norme.

Un autre observable a essayé de caractériser des textes par leur structure isotopique (par un aspect du fonctionnement de leur contexte). Pour cela j'ai représenté par un graphe les relations de cooccurrences dans un contexte fortement assimilateur (dépendances à la conjonction *ou*). La dépendance à un *ou* est utilisée pour sa « valeur contextuelle » : quelles que soient les acceptions de *ou* dans ses différentes occurrences, le contexte de la dépendance à *ou* est à chaque fois fortement assimilateur et implique l'existence d'afférences sémantiques fortes entre les lexèmes coordonnés. Dans le graphe construit, les nœuds représentent l'ensemble des lexèmes dans la dépendance d'un *ou* et les arêtes relient deux lexèmes qui co-occurrent au moins une fois dans la dépendance à une même occurrence de la conjonction. Par une classification automatique de ce graphe, on obtient des ensembles lexicaux fortement interdéfinis (constitués de noms qui co-occurrent dans les contextes de

la conjonction). Ces ensembles partagent des traits sémantiques très abstraits, dits

dimensionnels (tels que /pluralité/, ou /borné/). Cette abstraction des sèmes isotopants caractérise sans doute le discours philosophique : l'abstraction conceptuelle trouve une correspondance dans des mécanismes d'abstraction sémantique par neutralisation des sèmes domaniaux[17]. Ici, la méthodologie permet de caractériser des textes non pas directement par leur unité mais par la nature des relations contextuelles établies entre certaines unités. Le fonctionnement des contextes locaux est caractérisé à une échelle « globale ».

Enfin, l'ensemble de l'axe diachronique de l'œuvre de Gilles Deleuze a été appréhendé par une analyse factorielle multidimensionnelle re- groupant plusieurs niveaux de description (morphologique, syntaxique, morphosyntaxique, lexicaux, typographique). L'ensemble des niveaux de description contribue à opposer une première période académique[18], une période centrale de l'engagement politique[19], et une dernière période caractérisée par un repli littéraire[20].

Ainsi, des observables articulant méthodes quantitatives et linguistiques textuelles peuvent contribuer à caractériser les textes au-delà de profils morphosyntaxiques et lexicaux. Cela induit une notion de con- textualité entre niveaux de description.

Références bibliographiques

BACHELARD G. [1938] (1993), *La formation de l'esprit scientifique*, Paris, Vrin.

BAAYEN H. (1994), « Derivational productivity and text typology », *Journal of Quantitative Linguistics*, 1, p. 16-34.

BIBER D. (1988), *Variation across speech and writing*, Cambridge, Cambridge University Press.

BIRD S. et LIBERMAN M. (2001), « A formal framework for linguistic annotation », *Speech Communication*, 33-1/2, p. 23-60.

BIRD S. et Simons G. (2003), « Seven Dimensions of Portability for Language Documentation and Description », *Language*, 79, p. 557-582.

BLACHE P., RAUZY S. et FERRÉ G. (2007), « An XML Coding Scheme for Multimodal Corpus Annotation », *Proceedings of Corpus Lin- guistics*.

BLANCHE-BENVENISTE C. (2005), « L'Étude grammaticale des corpus de langue parlée en français », in *Les linguistiques de corpus*, G. Wil- liams (éd.), Rennes, Presses universitaires de Rennes, p.

47-66.

BRUNET E. (2004), « Où l'on mesure la distance entre les distances », *Texto !*, mars 2004 (http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html).

CANGUILHEM G. [1966] (2006), *Le normal et le pathologique*, Paris, PUF.

COSERIU E. [1952] (1982), « Sistema, norma y habla », in *Teoria del lenguaje y lingüistica general*, Madrid, Gredos.

COSERIU E. [1966] (2001), «Vers l'étude des structures lexicales », in *L'homme et son langage* (textes réunis par H. Dupuy-Engelhardt, J.-P. Durafour et F. Rastier), Louvain, Peeters (Bibliothèque de l'Information Grammaticale ; 46), p. 215-252.

COSERIU E. (trad. T. Verjans) [1958] (2007), *Synchronie, diachronie et histoire*, Paris, Texto ! (non paginé).

GADET F. (2003), *La Variation sociale en français*, Paris, Ophrys.

GLESSGEN M.-D. (2007), *Linguistique romane – Domaines et méthodes en linguistique française et romane*, Paris, Armand Colin.

HABERT B. et ZWEIGENBAUM P. (2002), « Régler les règles » *TAL*, 43- 3, p. 83-105.

HAGEGE C. (2005), « De la place réelle de la transitivité, ou la typolo- gie linguistique entre passé et avenir », in *Linguistique typologique*, G. Lazard et C. Moyse-Faurie (éd.) Paris, Presses universitaires du Septentrion, p. 55-69.

HIMMELMANN N., « Documentary and descriptive linguistics », *Linguis- tics*, 36, p. 161-195.

KILGARRIFF A. (2005), « Language is never, ever, ever, random », *Cor- pus Linguistics and Linguistic Theory*, vol. 1, p. 263-276.

LAKS B. (2002), « Le Comparatisme : de la généalogie à la génétique », *Langage*, 146, p. 19-45.