CONCLUSION

G. Leech (1991, p. 25) souligne le tournant des années actuelles : « Ceux qui travaillent sur corpus électroniques se trouvent soudain dans un univers en pleine expansion. Pendant des années, la linguistique de corpus a été l'obsession d'un petit groupe qui recevait peu de soutien, que ce soit de la linguistique ou de l'informatique. » Ce constat vaut au tout premier chef pour le monde anglo-saxon. Mais si l'on fait le bilan du domaine couvert par les linguistiques de corpus, quelles perspectives s'ouvrent, en particulier pour la francophonie ?

1. BILAN

Face à un domaine riche en travaux d'horizons théoriques et méthodologiques variés – en TALN et en linguistique, nous ne prétendons pas avoir rendu compte des recherches les plus représentatives. Comment, face à un champ en pleine mouvance, en identifier les grandes tendances ? Il aurait fallu un recul dont nous ne disposons pas et qu'à notre avis, on ne peut pas encore prendre. Nous avons plutôt cherché à fournir une typologie de travaux prometteurs. Espérons que cette typologie puisse aussi servir de grille de lecture pour situer d'autres recherches que celles qui ont été directement évoquées.

1.1 Avancées

La robustesse est le maître mot des techniques d'annotation qui sont visées pour les textes tout-venant. On est loin de pouvoir en donner une définition précise. Néanmoins, l'examen des outils disponibles et des corpus annotés le montre : l'étiquetage est relativement bien maîtrisé actuellement, le parsage fruste progresse, même si les tâtonnements dominent encore pour les traitements sémantiques.

Constatons que certaines tâches d'annotation sont progressivement automatisées, avec éventuellement des phases de pré- ou de post-traitement. On commence à mieux cerner ce qui est effectivement automatisable et ce qui ne le sera

probablement jamais. C'est ce que nous avons vu avec l'acquisition terminologique (chapitre II) : la frontière entre le repérage automatique et ce qui relève de compétences humaines peu formalisables se précise.

Il est frappant de constater que certaines de ces avancées reposent sur des techniques somme toute relativement simples. On est étonné par l'écart entre les méthodes utilisées, parfois frustes, et la richesse des résultats, comme l'indique E. Brill (1995, p. 544) : « Les méthodes basées sur les corpus sont souvent capables de réussir tout en ignorant la complexité réelle du langage, en s'appuyant sur le fait que des phénomènes linguistiques complexes peuvent souvent être observés indirectement par le biais de simples épiphénomènes. » C'est le cas pour l'alignement de textes, qui utilise parfois « une corrélation très forte entre la longueur des segments qui sont mis en correspondance traductionnelle » (Isabelle et Amstrong, 1993), que cette longueur soit mesurée en nombre de mots ou en caractères. C'est le cas encore de la production d'ébauches d'entrées de dictionnaires par des méthodes comme celles utilisées par Grefenstette (1994).

Un autre point positif est le recul des illusions en ce qui concerne le traitement automatique de textes tout-venant. Les conditions instutionnelles à réunir, les performances des outils existants ainsi que le coût de l'obtention de corpus annotés sont désormais mieux connus. Les opérations d'évaluation des outils et des ressources qui ont été lancées dans le monde anglo-saxon et qui débutent pour la francophonie (Paroubek *et al.*, 1997) sont salutaires : elles fournissent des états de l'art sectoriels et précis.

L'observation raisonnée de données volumineuses enrichit la pratique linguistique. Elle fournit des données que l'intuition du linguiste aurait refusées (taxées d'inacceptables) ou qu'elle n'aurait pas prévues (variation d'expressions toutes faites et de termes). Elle accroît la précision des descriptions ou les rectifie (en linguistique diachronique par exemple). Elle rend manifeste le poids des différentes règles. Les traitements multidimensionnels permettent de repérer des corrélations inattendues et en tout cas non perceptibles directement entre des phénomènes langagiers relevant de niveaux distincts de l'analyse linguistique.

1.2 Limites

Les ressources pour le français sont encore denrée rare. Il n'existe pas d'équivalents pour le français de ***Brown***, ***LOB*** et de ***BNC***, pour la langue contemporaine, ou d'***Archer***, pour l'histoire de la langue, c'est-à-dire des corpus diversifiés, associant des registres différents et offrant aux linguistes comme aux informaticiens des objets d'étude variés. Il n'existe pas non plus d'étiqueteur-lemmatiseur immédiatement accessible ni d'équivalent français de ***WordNet*** pour l'annotation sémantique. Le risque est que soient baptisés du nom de corpus des rassemblements de textes électroniques disponibles n'offrant pas les mêmes garanties de diversité quant aux types de texte inclus, ce qui biaiserait les études ultérieures.

Une autre limite est celle de l'étanchéité des communautés concernées. Institutionnellement, en France, le TALN et la linguistique[1] relèvent de deux

[1] Il faudrait en outre mentionner le secteur de l'informatique documentaire, dont les recherches sont mal connues en linguistique et en TALN, bien qu'elles soient riches d'enseignement pour le

secteurs disciplinaires aux fonctionnements éloignés : entre ces domaines, les passerelles et les collaborations sont encore fragiles. Les formations autour du traitement automatique du langage, par exemple, relèvent dans l'immédiat d'un secteur ou de l'autre, mais pas d'une convergence des deux.

L'évolution actuelle peut enfin conduire à marginaliser des travaux perçus comme moins directement « utiles ». L'étude diachronique de la langue en fournit un exemple. Mais l'expérimentation de formalismes sophistiqués peut également pâtir du nouveau contexte.

1.3 Questionnements

Du côté linguistique, les travaux que nous avons présentés poussent à examiner, ou à réexaminer sur des bases renouvelées, des phénomènes jusqu'à présent insuffisamment étudiés : place de la ponctuation, structuration globale des textes et grammaires textuelles, articulation langue générale / langues de spécialité, etc.

Du côté informatique, le succès pratique du métissage des traitements à règles et des traitement numériques pose sur le fond la question de modèles qui articulent finement observation et appel à la compétence des locuteurs et à l'expertise des spécialistes.

Une question reste ouverte : quelles généralisations permettent les multiples constats, si fins soient-ils, opérés sur les corpus annotés ?

2. PERSPECTIVES

Sans nous risquer à prédire l'avenir des linguistiques de corpus, nous soulignons à la fois les menaces qui pèsent sur leur développement et les espoirs qui semblent permis. Nous terminons par ce qui nous paraît être les conditions d'une évolution positive du domaine.

2.1 Menaces

Les menaces sont de trois ordres : les retards méthodologiques et techniques dans les moyens d'utiliser des corpus annotés, les dimensions laissées dans l'ombre par les linguistiques de corpus, et enfin des impasses intellectuelles.

Les moyens matériels de calcul ne cessent de progresser. Le versant logiciel des traitements de corpus accuse un retard d'autant plus sensible, ce qui retarde d'autant les expérimentations et partant, les avancées théoriques. On sait mémoriser des corpus et des ressources langagières de plus en plus vastes. Malgré

---

traitement des corpus annotés.

des initiatives de mise en convergence, il n'existe pas encore de chaînes de traitement standard pour ces données. La normalisation commence à devenir effective pour les corpus. Elle ne l'est pas encore pour les programmes correspondants, qui restent la plupart du temps expérimentaux. On est encore assez loin de « stations de travail textuelles » qui permettraient d'articuler des traitements diversifiés sur des corpus : étiquetage, correction interactive, parsage, annotation sémantique, décomptes et modélisation ...

Certaines dimensions restent peu abordées en linguistique de corpus. C'est le cas de la textualité en tant que telle[2]. Même les études de Biber, lorsqu'elles caractérisent les types de texte comme des constellations de traits linguistiques, ne rendent pas compte de l'organisation des textes au delà de la phrase, de l'enchaînement des énoncés. La dimension pragmatique s'efface également, en raison de la primauté accordée à la morpho-syntaxe.

Nous avons déjà cité l'adage de G. Sampson (1994, p. 180) : « la linguistique de corpus prend le langage tel qu'il est. » Le piège serait ... de le laisser tel qu'il est, c'est-à-dire de n'introduire aucun déplacement théorique. La manipulation des corpus annotés est lourde. Le déferlement des données peut aussi dérouter, par son intrication complexe de phénomènes multiples[3]. Tout le langage s'engouffre. Le risque est alors un empirisme linguistique radical[4], à fleur de données et sans recul.

Ceux qui mettent au point traitements et outils peuvent être de leur côté tentés par une certaine commisération pour les études proprement linguistiques. Ces dernières ne se confronteraient jamais au langage « réel ».

2.2 Espoirs

Les recherches dont nous venons de dégager les grands traits renouvellent la dimension empirique et expérimentale de la linguistique, en particulier en ce qui concerne la quantification des faits langagiers.

Pour reprendre les termes de C. Jacquemin, une linguistique véritablement expérimentale est possible. Puisque les corpus et les outils entrent de plus en plus dans le domaine public, les résultats présentés par les recherches sont vérifiables sur les mêmes données ou au contraire amendables par confrontation avec d'autres données. Les faits deviennent un peu plus têtus. Expérimenter, c'est aussi pouvoir construire des modèles, symboliques ou quantitatifs, et les tester sur des données.

Comme l'écrit J. Sinclair (1991, p. 100) : « La langue a l'air assez différente quand on en examine un grand morceau d'un coup. » Les distinctions tranchées s'estompent. Aux différents niveaux de l'analyse linguistique, on peut séparer usuel, exceptionnel et tout à fait improbable. On peut désormais quantifier de

[2] J.-P. Sueur (1982, p. 144) dégage tout de même des pistes et montre des premiers résultats.

[3] C. Filmore et B. Atkins (1994) montrent la complexité de l'analyse du verbe *risk* lorqu'on part, comme eux, de corpus : 1 743 contextes fournis par l'APHB (American Publishing House for the Blind) et de 470 extraits du corpus à la base du dictionnaire COBUILD. Ils comparent les tendances observées dans ces contextes avec le traitement opéré dans dix dictionnaires. Ils insistent sur les choix théoriques comme seuls moyens de s'orienter dans le flux des attestations.

[4] L'expression est de M.-P. Péry-Woodley (1995, p. 216).

nouveaux phénomènes. On peut aussi examiner les corrélations entre des traits linguistiques multiples. Mais il reste à acquérir pour la syntaxe et la sémantique une expérience similaire à celle qui a été développée en analyse statistique du lexique. Elle permettra d'attribuer leur véritable dimension aux résultats obtenus actuellement.

2.3 Conditions

Les linguistiques de corpus se révèleront fructueuses comme domaine de recherche si l'on accepte l'imparfait, c'est-à-dire des ressources toujours « impures », et si s'affirment des collaborations soutenues entre linguistes et informaticiens.

Les corpus annotés comme les outils d'annotation reposent sur des approximations. L'ampleur des moyens à réunir force à des solutions qui, sans être jamais vraiment consensuelles, reposent sur des compromis entre des communautés distinctes et des impératifs techniques multiples. Ces solutions dépendent également de l'usage prévu en aval pour les ressources annotées. Cette imperfection ne constitue pas pour autant un obstacle majeur. Nous l'avons vu, il est souvent possible de « faire des détours » pour isoler les phénomènes visés. Sans doute faut-il aussi abandonner l'horizon, illusoire, de corpus « parfaitement » annotés et d'outils ne faisant pas d'erreur. Pourquoi attendre de la « machine » une cohérence et une perfection que l'annotation manuelle n'atteint pas ?

La collaboration de l'Université de Lancaster et du centre de recherche d'IBM Watson (Black *et al.*, 1993) est exemplaire d'une coopération fructueuse entre les deux communautés concernées au premier chef, la linguistique et le TALN. Les linguistes ont vu leur attention attirée sur des phénomènes souvent conçus comme marginaux et sur la nécessité de les intégrer dans leur description. Les informaticiens ont appris à modéliser des comportements langagiers plus fins que ceux qu'ils traitaient initialement. Les deux communautés ont l'intérêt le plus vif à coopérer. La constitution de vastes corpus finement annotés et la mise au point des outils nécessaires supposent des recherches informatiques importantes et coûteuses. Les linguistes en bénéfieront. Inversement, seuls des travaux poussés en linguistique descriptive permettent de mieux maîtriser les causalités à l'œuvre : influence des types de textes, jeu entre sous-langages et langue générale, poids du temps, etc. Les informaticiens y trouveront matière à améliorer leurs modèles et leurs techniques. Parce que les corpus lui semblent le moyen de constituer les ressources linguistiques nécessaires à des traitements effectifs, le TALN se confronte désormais à toute la complexité du langage. Disposer de corpus annotés renouvelle les méthodes et les objectifs de la linguistique descriptive. Le foisonnement des recherches témoigne de la vigueur du champ. Il y a probablement une chance historique à saisir : celle d'une coopération enfin fructueuse.