TRANSCRIPTIONS ET TECHNOLOGIES

Beaucoup de débats engagés sur l'utilisation des corpus de langue parlée ont évolué récemment, sous l'influence des nouvelles technologies diffusées par les ressources informatiques. Je voudrais ici rappeler quelques questions importantes issues directement de ces avancées techniques.

Lorsque le terme de *corpus* apparaît en France, vers 1923 (d'après A. Rey[1]), il n'est pas tourné vers les technologies modernes. Il vient du droit ; *corpus juris* désigne depuis fort longtemps le recueil des lois du droit romain. Lorsque, à partir de 1956, Gougenheim, Rivenc et Sauvageot enregistrent des conversations pour élaborer *Le Français fondamental*, nous dirions qu'ils collectaient des *corpus*. Mais eux ne se servaient pas de ce mot. Ils n'utilisaient que le terme d'*enquête*.

Vers les années 1970, le mot *corpus*, doté de sa nouvelle acception importée des États-unis, s'installe dans la nomenclature des linguistes français. Dans cette nouvelle acception, il a une allure technique. « Le mot est pédant », écrivait R. L. Wagner en 1973, »autant que celui de *campus* »[2].

A l'époque actuelle, la diffusion et l'utilisation des corpus de langue parlée ont pris en Europe une grande extension. A quoi servent-ils ? A des fins d'études universitaires sans applications extérieures au domaine : grammaire, recherche lexicale, pragmatique étude des interactions. Mais également à la recherche appliquée, par exemple dans le dialogue homme-machine, les modèles de langage pour entraîner les systèmes d'analyse automatisés, les bases de données lexicales pour des dictionnaires, etc. Les

1. A. Rey, 1992, *Dictionnaire Historique de la Langue française* . Paris: éditions Le Robert.

2. .R.L. Wagner, 1973 *La Grammaire française, volume 2. La grammaire moderne. Voies d'approche. Attitude des grammairiens* . Paris: SEDES, p. 100. Le pluriel latin, *corpora* , adopté par d'autres langues, ne s'implante pas en français; on dira "des corpus".

traitements des *corpus* se sont diversifiés et sont devenus, dans certaines entreprises, hautement technologiques.

> « The area of natural language processing is particularly sensitive to technological advance » (G. Leech, EAGLES 1996)

Je propose de rendre compte, dans les grandes lignes, des informations que diffuse à ce sujet un groupe de chercheurs européens réunis dans un organisme nommé EAGLES (Expert Advisory Group in Language Engineering Standards). Ce groupe s'intéresse depuis quelques années à la standardisation de ce qu'il est convenu d'appeler *la linguistique sur corpus.* J'utiliserai particulièrement les publications de Joachim Llisterri (Universidad Autonoma de Barcelona) pour la transcription[3], de Geoffrey Leech (University of Lancaster) pour les annotations grammaticales[4], et de John Sinclair pour la typologie des textes[5]. Je ferai allusion aux conventions établies par NERC (Network of European Reference Corpora) et TEI (Text Encoding Initiative).

1. LES CONVENTIONS DE TRANSCRIPTION

Le travail sur corpus a amené les linguistes à avoir un respect des données au même titre que dans d'autres domaines de recherche[6]. Lorsqu'il s'agit de représenter par écrit des données orales, ce respect des données exige qu'on fasse des choix (on ne peut pas être fidèle à tous les phénomènes en même temps) et qu'on développe des conventions commodes. Dans la perspective actuelle d'échanges de travail en Europe, ces conventions doivent être largement partagées.

La démarche habituelle consiste à partir, en premier lieu, d'une transcription orthographique (complétée éventuellement par divers systèmes d'annotations). En effet, la plupart des chercheurs qui travaillent actuellement sur les langues parlées tiennent à disposer de grands corpus

3. Llisterri, J. (1994a) Spoken Texts. Draft-Work in Progress, EAGLES Documeeeeent EAG-CS/IR T7.1, October 1994.

4. Leech G. (1994) Morphosyntactic Annotation . Draft- Work in Progress. EAGLES Document EAG-GCS/IR-T3.1.

5. Sinclair J. (1994), Corpus Typology. Draft- Work in Progress. EAGLES Document EAG-CSG/IR-T1.1. J'utiliserai aussi les informations glanées au Colloque de EAGLES qui s'est tenu à Madrid du 6 au 8 janvier 1996.

6. G. Leech, 08/01/96, p. 2).

pour en faire éventuellement des études quantitatives. Ils font rarement, en ce cas, des transcriptions phonétiques. C'est en partie pour une question de temps de travail, bien que certaines transcriptions phonétiques puissent aujourd'hui être faites semi-automatiquement. Mais c'est surtout affaire de spécialisation.

On se sert de l'Alphabet Phonétique International pour les études consacrées spécifiquement au signal sonore (*Speech Research)*. Mais il est très rare de rencontrer des transcriptions de langue parlée (*Spoken Language)* faites systématiquement en API. Cette différence entre *Speech Research*, portant sur l'aspect phonique du langage, et *Spoken Language Research*, (difficile à rendre en français), portant sur l'étude grammaticale, lexicale ou discursive de productions orales, a représenté pendant longtemps une frontière majeure entre deux disciplines. Les études grammaticales et lexicales cherchent à disposer d'échantillons de langage « spontané », stocké en grandes quantités. Les travaux d'ordre phonique portent généralement sur le langage « de laboratoire », représenté par des collections beaucoup plus restreintes. Les avancées technologiques récentes vont sans doute les rapprocher, de l'avis de J. Llisterri, parce que le langage « de laboratoire » pourra absorber davantage de discours dit « spontané ».

> « Even general purpose corpora of impromptu, unrehearsed, unscripted, non elicited informal conversations now seem to arouse some interest in speech research as they can be used as test-beds for speech recognition systems » (Teubert, 1993 :4)[7].

D'autre part, depuis qu'il est possible de faire coïncider la représentation orthographique et une représentation automatique du signal sonore, les techniques sont davantage mises en commun. On peut en effet, au prix d'une certaine dose d'intervention manuelle, transcrire en faisant correspondre une ligne d'écriture orthographique, une ligne de transcription phonétique et une ligne de représentation prosodique. Mais, pour l'instant, comme la mise en place des différents paramètres n'est pas simple, le travail sur la langue parlée n'utilise en général la transcription phonétique que pour un petit nombre d'annotations rajoutées à la transcription orthographique.

Les mots sont donc transcrits comme des unités lexicales graphiques, dont on n'étudie pas le détail des réalisations. Les conventions NERC adoptées en 1992 précisaient que, dans ces transcriptions, on devait suivre les standards orthographiques admis conventionnellement ; on n'utilisait les contractions de mots que dans la mesure où elles figuraient dans un dictionnaire de référence. Les frontières de phrases étaient marquées par un

---

7. W. Teubert (1993) Phonetic/phonemic and prosodic annotation. Final Report NERC-WP!-&È&,. Manneheim: IDS (cité par Llisterri 1994-a:5).

point et une majuscule. On utilisait des guillemets pour les citations et le discours rapporté ; on ne faisait aucun usage de la virgule.

En plus de la transcription orthographique, on a souvent proposé de noter les divers « événements » qui accompagnent la prise de parole[8] comme les allongements, l'accentuation ou divers autres phénomènes phoniques. On a proposé des conventions particulières pour les formes non standard, pour les abréviations, les acronymes, les mots épelés, les coupures dans les énoncés, avec ou sans pause, les amorces, les chevauchements, et quantité d'autres « événements ». On a proposé de noter les unités intonatives, les hauteurs, les intensités, le débit de parole et les pauses. Certains transcripteurs représentent les qualité de la voix, le chantonnement, la voix criée, les passages lus, les passages chantés, les morceaux incompréhensibles ou simplement devinés, etc. Certains ont envisagé de noter systématiquement les éléments non verbaux : gestes, mimiques, regards, attitudes, qualité des rires, voire des toux, bruits divers. Enfin, il a souvent été demandé d'adjoindre à la transcription un descriptif des difficultés qu'on a rencontrées pour la faire. En 1992, J. P. French a publié des propositions visant à unifier ces pratiques[9].

J. Sinclair[10] signalait des abus : les néophytes, lorsqu'ils se lancent dans l'exploitation des corpus de langue parlée, ont tendance à insister sur les aspects les plus bizarres ou les plus anecdotiques. Le résultat est souvent un intérêt assez disproportionné accordé aux éléments comme les onomatopées, les claquements de langue, les raclements de gorge, les rires, toutes choses qu'on aura du mal à intégrer dans une description linguistique. Ces abus avaient peu de conséquences fâcheuses quand ils portaient sur des études limitées. Mais, à partir du moment où l'on établissait de grands corpus, subventionnés par des fonds de recherche, ces abus avaient des conséquences financières très voyantes.

Il y a actuellement de grands débats sur l'intérêt que présentent ces annotations. Certains y voient une procédure d'enrichissement et d'éclaircissement des textes. Mais G. Leech, assez pessimiste, disait, au colloque EAGLES de janvier 1996, qu'on aurait pu aussi bien parler, dans bien des cas, de « pollution » ou « corruption » des textes !

Ces interpolations peuvent être effectivement gênantes, et les chercheurs de EAGLES recommandent maintenant d'avoir toujours à disposition une version du *texte nu*, dépourvue d'annotations. Il se peut que cela crée des

---

8. J. Llisterri recommande particulièrement la publication de J.A. Edwards et M.D. Lampert (eds.), Talking data: Transcription and coding in discourse research . Hillsdale, NJ: Laurence Eerlbaum Ass.

9. French,J.P. (1992) Transcription proposals: multi-level system. NERC-WP 4 50.Bimingham: Universiy of Birmingham.

10. Colloque de Lisbonne, 3/10/95.

difficultés, et que certains textes semblent, en raison du manque d'annotations, peu intelligibles par endroits. Mais cela semble préférable à l'excès inverse, qui consiste à charger les textes, d'emblée, d'une grande masse d'informations. Il arrive, actuellement, que certains chercheurs dépensent une grande énergie à surcharger leurs corpus d'annotations diverses et que d'autres chercheurs en dépensent autant pour effacer cette surcharge.

Un certain nombre des ambitions de fidélité aux phénomènes sonores vont tomber en raison des avancées technologiques, par exemple la possibilité d'avoir, sur un CD-rom, des transcriptions couplées avec les enregistrements. On pourra envisager, par exemple, de vérifier les phénomènes prosodiques d'un passage transcrit orthographiquement, ou les liaisons, sans dépendre d'un système de notation étranger à la recherche en cours.

Au fur et à mesure que progressent ces technologies, le travail spécifique du transcripteur se précise. J. Sinclair demandait que les chercheurs s'entendent sur une sorte de *niveau zéro* de la transcription, faite orthographiquement avec le minimum d'indications supplémentaires[11]. On ménage ainsi, dit-il la facilité de lecture. En fait, les conventions NERC de 1992 proposaient de distinguer quatre niveaux :

- Le premier donnerait une représentation orthographique avec un minimum de ponctuation (en 1996, Sinclair suggère qu'on omette la ponctuation). A ce niveau, aucune marque des interactions, ni même des changements de locuteurs.
- Le niveau deux introduirait les tours de parole des locuteurs et certains éléments non verbaux.
- Le niveau trois aurait des indications d'intonation et d'interactions. Les frontières d'unités tonales et syllabiques seraient indiquées. Seuls des phonéticiens pourraient le faire, et sur de bonnes qualités d'enregistrement.
- Le niveau quatre additionnerait les indications acoustiques et phonétiques, les schémas mélodiques, un tracé du fondamental (avec d'excellents enregistrements).

Il est intéressant de remarquer que, à travers toutes ces discussions et ces mises au point, jamais ne s'est posée, dans le groupe EAGLES, la question des multi-transcriptions telle que l'envisage le GARS, c'est-à-dire la possibilité de proposer plusieurs versions graphiques d'un même passage, comme « il l'apprend / il la prend » ou « elle a marché/ elle marchait ». Dans les études du GARS, l'établissement même du texte est envisagé comme un problème difficile, qui met en jeu autre chose que l'automatisme

11. J. Sinclair dans les recommandations faites au Colloque de Madrid en janvier 1996.

d'une transposition terme à terme, parce qu'une part d'interprétation, phonique ou sémantique, y intervient toujours. C'est pourquoi la transcription y est considérée comme une entreprise qui engage la théorie, et que les chercheurs ne peuvent pas abandonner à des « secrétaires ». Rien de semblable dans le groupe EAGLES, qui n'a pas, pour l'instant, été préoccupé par ce problème.

2. LES ANNOTATIONS

Elles commencent avec les marques de diverses sortes que l'on porte sur le texte transcrit, afin de pouvoir les exploiter ensuite pour une analyse suivie. Tout dépend des buts poursuivis. Les chercheurs intéressés par les interactions notent les événements qui ont accompagné la prise de parole, et particulièrement les attitudes des interlocuteurs, quand il y en a plusieurs : rires, regards, bruits, gestuelles, etc. On a développé, d'autre part, diverses façons de noter l'information grammaticale, en mêlant ce travail à celui de la simple transcription. On porte sur chaque mot un étiquetage morphologique *(tagging)*, qui indique essentiellement à quelle partie du discours il appartient (POS, *Part of Speech Tagging*). On indique aussi, à un autre niveau, les relations syntaxiques (*parsing*).

On peut également porter des annotations sémantiques, surtout lorsqu'on prévoit de faire une application à des dictionnaires, et des annotations discursives, comme par exemple certains réseaux d'anaphore.

Il s'agit donc d'indiquer par des signes conventionnels (« interpolations » à l'intérieur du texte), des catégories d'analyse. Dans un grand nombre de cas, ces catégorisations ont été seulement des indications morphosyntaxiques sur les parties du discours. Mais, comme le reconnaît G. Leech en 1996, l'analyse n'est pas encore très avancée :

> « Morphosyntactic annotation (which has been so far carried out extensively on English, but not on other languages) is at a relatively primitive stage of development » (1994 :4).

Dans certains cas, une partie du travail d'étiquetage peut être fait automatiquement, quitte à corriger ensuite à la main les assez nombreuses erreurs. Mais il faut bien se rendre compte du travail que représente la correction de l'étiquetage pour un corpus de 100 millions de mots. L'expérience des lexicographes, qui ont l'habitude d'indexer les éléments grammaticaux de leurs inventaires, peut servir, mais elle ne suffit pas à donner toutes les indications qu'on jugerait nécessaires pour faire une

analyse syntaxique. Le principe des annotations est le suivant : on accompagne la transcription d'un étiquetage qui forme une sorte de « préanalyse ». On envisage d'appliquer cette préanalyse à un grand nombre de textes, formant des corpus importants, de l'ordre d'un million de mots. En totalisant les étiquetages sur de grandes dimensions, on pourrait obtenir des statistiques de régularités et d'irrégularités grammaticales, qui échappent à la description faite sur de petites dimensions.

Le grand problème préliminaire est de savoir quelle limite minimale on se fixe pour le nombre de ces annotations, et quelle limite maximale, pour autant qu'on puisse en envisager une. On peut imaginer de fournir, comme l'équipe du GARS, des transcriptions qui ne sont accompagnées d'aucune codification grammaticale. C'est l'attitude minimaliste qu'on adopte si l'on pense que la grammaire des productions orales est encore peu connue, qu'il y a des données à dégager et des concepts nouveaux à mettre au point. En ce cas, il serait téméraire de vouloir tout étiqueter, dans la mesure où les analyses traditionnellement acceptées risquent parfois d'être mises en défaut. Ce serait le cas pour certaines parties du discours comme « préposition, conjonction, adverbe », et pour quantité de fonctions syntaxiques, en particulier les notions majeures de coordination et subordination. Dans cette perspective, le plus urgent serait d'assurer en premier lieu les principes de l'analyse et d'envisager seulement par la suite de faire un étiquetage des textes.

Les corpus de langue parlée apportent en effet, pour des entreprises comme celles du GARS, non seulement des données nouvelles, mais un ensemble de délimitations nouvelles des données que l'on avait déjà. Il n'est donc pas toujours facile d'analyser d'emblée la nouveauté.

> « Ce sont souvent les exemples difficiles à classer, ou qui posaient des problèmes au départ, qui permettent par la suite de faire de nouvelles hypothèses et d'avancer dans l'analyse » (Monique Gibier, 1987, Mémoire sur l'accord des participes passés en français parlé, Université de Provence).

On peut au contraire décider que le matériel d'analyse dont disposent les linguistes actuellement est largement suffisant, et qu'on peut adopter une analyse en parties du discours (dotée de quelques modernisations) et une description, au moins « basique » des relations syntaxiques. Cet appareil descriptif, même imparfait, permettrait déjà de passer à une étape statistique rentable.

Le groupe EAGLES, en 1994, propose d'indiquer au moins les principales parties du discours, comme Nom, Verbe, Conjonction :

1 N [nom]	2. V [verbe]	3 AJ [adjectif]
4 PD [pronom, déterminant]	5. AT [article]	6. AV [adverbe]
7. AP [adposition]	8 C [conjonction]	9. NU [numéral]
10. I [interjection]	11 U [unique, non assigné]	12. R [résiduel]
13. PU [ponctuation]		

Une partie de ces désignations sont classiques et admises par tout le monde. D'autres permettent de signaler certaines zones délicates, comme par exemple AP, qui désigne les « adpositions », « prépositions », « circumpositions » ou « postpositions ». Le classement par U. sert à désigner les éléments qui, comme la particule de négation, sont en quelque sorte « uniques en leur genre », et pour lesquels il serait inutile de chercher une assignation à une catégorie :

1. Les marqueurs d'infinitif (anglais *to)*
2. la particule négative (anglais *not, n't)*
3. le marqueur existenciel (anglais *there,* danois *der)*
4. la seconde particule de négation (français *pas)*
5. les éléments anticipateurs (néerlandais *er)*
6. le marqueur de voie médio-passive (portugais *se)*
7. les particules préverbales (grec)

R. s'applique aux éléments qu'on ne peut pas classer dans des parties du discours : mots étrangers, formules.

On peut indiquer également certains de leurs attributs : Genre, Nombre et Cas pour le Nom ; éventuellement l'appartenance à des classes sémantiques comme « noms temporels », « adverbes de manière », etc.

G. Leech énumère ainsi des ajouts optionnels, représentant certaines caractéristiques mises au point dans différents courants de la linguistique contemporaine :

- pour le Nom : comptable ou massif

- pour l'Adverbe :
  classe générale ou adverbe de degré (*très, si, tant)*
  type interrogatif, relatif ou exclamatif (ex : *comment)*

- pour la conjonction :
  simple
  corrélative
  initiale
  non-initiale

Ces caractéristiques reçoivent un codage conventionnel, par exemple,

dans la présentation de G. Leech (1994) :

- nom commun, féminin, pluriel, comptable :
  N122010
- verbe 3ème pers, sing, conjugué, indicatif, passé, actif, principal, non réflexif :
  V3011141101200
- adjectif « général », au comparatif :
  AJ2000000

Il est difficile, même avec tous ces recours, d'atteindre un point ultime de l'analyse, qui serait une sorte de perfection idéale. Leech donne l'exemple du verbe anglais, pour lequel on devrait prévoir une formule maximale assez importante, qui ne couvre pourtant pas le cas de l'infinitif :

V [[-301/ 002] 111/ 000121/ 000130] 0200001

Voici un exemple de cette annotation morphosyntaxique appliquée à une série de pronoms en italien :

PQNS1	Pron. pers. comm. sing. 1	io
PD141001001		
PQNS2	Pron. pers. comm. sing. 2	tu
PD241001001		
PQMS3	Pron. pers. masc. sing. 3	egli
PD311001001		
PQFS3	Pron. pers. femm. sing. 3	ella
PD321001001		
PQNP1	Pron. pers. comm. plur. 1	noi
PD142001001		
PQNS2	Pron. pers. comm. plur. 2	voi
PD242001001		
PQNP3	Pron. pers. comm. plur. 3	loro
PD342001001		

On peut arriver ainsi à des notations plus ou moins raffinées des informations grammaticales. On voit bien que, ici, ce type de codification reproduit, à peu de choses près, les classifications les plus traditionnelles de l'analyse « grammaticale ».

3. EXPLOITATIONS

Ces techniques d'analyse se sont répandues dans quantité de domaines sans que les linguistes s'en aperçoivent, et de façon que les exploitations qu'on en fait échappent à leur contrôle. Dans certains secteurs d'activité, le travail sur la langue parlée implique même d'emblée qu'on fait un étiquetage de tous les mots, ce qui est supposé lancer sur le texte un filet d'analyse dont on pourra de toute façon tirer quelque profit. Un grand nombre de chercheurs d'autres disciplines ont commencé à exploiter les annotations comme des techniques fermement assurées, ce qui peut susciter des problèmes quand il s'agit de textes peu conformes aux normes usuelles. S'ajoute à cela un inconvénient né de la répartition des tâches. Comme il est assez fastidieux de conduire ce travail d'étiquetage morphosyntaxique sur de grandes étendues de textes, les chercheurs délèguent parfois cette tâche, présentée comme élémentaire, à des personnes peu spécialisées. On risque d'aboutir, en ce cas, à une pratique des annotations peu cohérente et peu utile.

Je prendrai un exemple de résultat négatif dans la pratique de certains spécialistes médicaux de troubles du langage (logopédistes de Liège, par exemple[12]), à partir de corpus transcrits et annotés selon des procédés qu'ils croient « scientifiques » et qui, souvent, peuvent nous horrifier, même si le monde médical en fait, somme toute, un usage modéré.

Voici un exemple de l'étiquetage utilisé par des spécialistes des troubles du langage, orthophonistes ou logopédistes (M. F. Granier 1994, *Approches quantitatives des déficits en morpho-syntaxe de patients aphasiques*, Liège, p. 23). Il s'agit de l'enregistrement d'une personne hospitalisée à la suite d'un accident. On lui demande de raconter l'accident. Le texte de la transcription est présenté en deux formats. D'une part la version orthographique, coupée en séquences qui correspondent à des sortes de paragraphes, et en unités qui ressemblent aux « propositions » de la grammaire scolaire. D'autre part un étiquetage morphosyntaxique, sur lequel sont construits des calculs statistiques. Il s'agit, dans ce genre de recherche, de compter les « anomalies », aussi bien les énoncés inachevés que les réalisations non normatives ou les « agrammaticalités ». On compte les anomalies à partir de la version annotée.

A partir du moment où on dispose de la version annotée, on peut se

12. Marie-France Granier, 1994, Approches quantitatives des déficits en morpho-syntaxe de patients aphasiques, Univesité de Liège).

permettre, semble-t-il, d'abandonner la transcription orthographique de départ. Elle est considérée comme une étape préliminaire du travail, à laquelle on peut se référer pour vérification, mais qui n'est plus essentielle. La version « étiquetée » devient la base réelle de l'observation et de la réflexion. Comme elle procède par étiquetage de petits segments, elle ne rend pas compte (du moins pas dans cet état de la technique utilisée) des relations entre les grands constituants de l'énoncé.

Séquence n° 12

C'est à ce moment-là que
proS être prep detN conj ø

qu'on a pris le parti de m'emmener dans un autre hôpital
conj pros aux v deefN gen cliD v prep indef adjq N

et c'est là-bas qu'on m'a trépané très vite etc
coor proS être adv conj proS cliD aux V adv adv (expression)

Séquence n° 14 :

Pour moi c'est comme si j'avais dormi quoi,
prep pro proS être loc conj proS aux V (intj)

sauf que quand je me suis réveillé
loc conj conj proS cli (refl) aux V

je ne savais plus parler quoi
proS aux nég Vinf (intj)

Les annotations cumulent, sur les segments graphiques les plus petits, les indications morphologiques et syntaxiques. Dans *c'est*, (séquence 12), l'auteur analyse comme un pronom (pro) qui est sujet (S). La notation par signe vide ø, (*que* = conj. ø), dans :

c'est à ce moment-là *que* qu'on a pris le parti de m'emmener

est faite pour signaler un inachèvement. L'analyse est la suivante : le locuteur a commencé une subordonnée en *que*, qu'il n'a pas continuée. Cette conjonction qui joue à vide, puisqu'elle n'est pas immédiatement suivie de sa subordonnée, sera comptée, parmi d'autres choses, comme un des indices du « manque de langage » du patient aphasique. Tous les signes « vides » seront comptés, et disponibles pour faire des comparaisons avec d'autres patients.

On voit vite à quel émiettement mène cet étiquetage. Dans la séquence

12, *c'est à ce moment-là que qu'on a pris le parti de m'emmener* on a un luxe de catégorisation morphologique, mais aucun lien n'a pu être marqué entre *c'est* et *que*. La tournure clivée que forme *c'est... que...*, autour de *à ce moment-là,*

c'est à ce moment-là qu'on a pris le parti de m'emmener

ne peut pas être identifiée. La partie analytique ne rend pas compte du rôle de *c'est.. . que*, et de la localisation du complément temporel *à ce moment-là*. Elle traite *c'est...* comme un simple verbe, puisque c'est la donnée morphologique immédiate, sans pouvoir envisager son rôle de support du dispositif clivé.

Il est bien évident que, pour un linguiste minutieux, ce simple travail d'étiquetage doit être complété par une analyse qui tienne compte des grandes relations syntagmatiques. C'est particulièrement délicat quand il s'agit des discours d'aphasiques, dans lesquels les grandes relations sont souvent masquées par des phénomènes superficiels d'hésitations et de bribes inachevées (par exemple *c'est à ce moment-là que qu'on a pris le parti...*). Mais on comprend que, pour des chercheurs qui ne sont pas des linguistes, et qui veulent rendre compte d'un grand nombre de productions, cette approche, même insatisfaisante, leur paraisse être déjà un « bon début ».

Les dirigeants de EAGLES signalent un autre inconvénient, d'un tout autre ordre. Étiqueter un texte en donnant des indications sur les classes de mots est, pour l'essentiel, un travail fait à la main. Pour des textes de grande ampleur, cela revient très cher, au point que G. Leech demandait récemment qu'on ne finance plus ce genre de travail quand il porte sur des textes longs. Les résultats, dit-il, sont souvent assez médiocres, parce qu'on le fait faire par des personnes peu spécialisées, et qu'on paie très mal. Il vaudrait mieux, estime-t-il, qu'on consacre les fonds de recherches à améliorer les possibilités d'analyse automatique des corpus.

4. CONCLUSION

Les avancées technologiques récentes sont en train de changer considérablement le travail linguistique sur les données des langues parlées, aussi bien pour les objectifs que pour les méthodes. Les réflexions faites actuellement au niveau européen sur les transcriptions confortent les choix de transcription « pauvre » qu'avait proposés l'équipe du GARS. Les discussions sur les annotations grammaticales, étalées en pleine lumière,

partagées par des chercheurs de formations diverses, et soumises à des questions de rentabilité financière, ont tout à coup pris une autre allure. La recherche des « bonnes solutions » d'analyse morphosyntaxique en sera sans doute profondément modifiée.

Claire BLANCHE-BENVENISTE