Introduction : Grammaticalité et annotations de corpus d'anglais oral – perspectives et problèmes

Sylvie Hancil

Anglophonia
French Journal of English Linguistics

23 | 2017
Nouvelles frontières de la grammaticalité

Introduction : Grammaticalité et annotations de corpus d'anglais oral – perspectives et problèmes

Sylvie Hancil

Édition électronique
URL : https://journals.openedition.org/anglophonia/1162
DOI : 10.4000/anglophonia.1162
ISSN : 2427-0466

Éditeur
Presses universitaires du Midi

Référence électronique
Sylvie Hancil, "Introduction : Grammaticalité et annotations de corpus d'anglais oral – perspectives et problèmes", *Anglophonia* [Online], 23 | 2017, Online since 02 July 2018, connection on 27 May 2024. URL: http://journals.openedition.org/anglophonia/1162 ; DOI: https://doi.org/10.4000/anglophonia.1162

Introduction : Grammaticalité et annotations de corpus d'anglais oral – perspectives et problèmes

Sylvie Hancil

1. Introduction

1 La grammaire se trouve au cœur de la théorisation linguistique depuis des décennies maintenant. En effet, bien que la grammaire soit souvent considérée comme un sujet aride, c'est la grammaire qui offre l'architecture de la pensée et de la communication consciente et explicite, de sorte que toute personne qui se tourne vers la linguistique dans l'espoir de gagner une compréhension plus claire de la façon dont les esprits humains fonctionnent devrait accepter la grammaire comme étant centrale à la discipline.

2 Néanmoins, une raison qui nous pousse à nier la réalité du concept de grammaticalité, comprise comme étant la recevabilité d'un exemple par un locuteur de la langue étudiée, est le simple fait qu'aucun linguiste n'a jamais réussi à établir une frontière entre le grammatical et l'agrammatical, pour une quelconque langue, malgré les nombreuses tentatives pendant maintes décennies. Edward Sapir ([1921] 1963 : 38) a cette citation très connue "all grammars leak", et elle reste vraie encore maintenant, près d'un siècle après. Comme le dit David Graddol (2004) : "No one has ever successfully produced a comprehensive and accurate grammar of any language". Même si l'on date la volonté de développer les grammaires formelles génératives non de Sapir mais, de manière plus réaliste, du temps du livre *Syntactic Structures* de Noam Chomsky (1957), ce livre est apparu plus de cinquante ans en arrière. En réalité, les grammaires les plus complètes élaborées par des linguistes s'avèrent avoir des lacunes dès que leurs prédictions sont mises à l'épreuve des petits exemples de l'usage quotidien.

3 Une des tentatives les plus sérieuses visant à établir une grammaire générative complète de l'anglais fut publiée par Stockwell, Schachter et Partee dans un volume de 854 pages : *Major Syntactic Structures of English* (1973). Ces auteurs ont résumé leurs points de vue sur

la faisabilité de l'exercice en citant dans leur livre un passage d'un grammairien du 17e siècle James Howell (1662 : 80) :

> But the English … having such varieties of incertitudes, changes, and idioms, it cannot be in the compass of human brain to compile an exact regular syntaxis thereof …

4 En commentant la phrase de Graddol mentionnée précédemment, Anne Dalke (2010) compare sa vision de la grammaire avec « the sort of friendly adjustment that happens when we hang out with one another: there is a range of possible behaviors, but no rules ». Des modèles comportementaux mais pas de règles établies ; tel est le résumé de la façon dont nous voyons la grammaire.

5 Après avoir précisé le concept de grammaire, nous étendrons notre étude au domaine de l'oral et en particulier aux problèmes qui peuvent être rencontrés pour l'oral dans le système d'annotations proposé par le programme SUSANNE et appliqué à des échantillons d'anglais écrit britannique du corpus CHRISTINE.

2. Ce qu'on peut dire de la grammaire

6 Assurément, bien que notre position soit susceptible d'être mineure, nous ne sommes en aucun cas les premiers linguistes de ces derniers 50 ans à avoir questionné le concept de « grammaticalité ». Le point crucial de notre discussion est non seulement d'arguer contre le concept de grammaticalité, mais de présenter un panel de découvertes concrètes sur la grammaire qui émergent quand on abandonne ce concept. Le fait que la grammaticalité puisse être un mythe n'implique pas qu'il y a peu à dire sur la grammaire en général au-delà du niveau des études de constructions particulières dans les langues particulières. Il paraîtrait bien souhaitable que la répartition de l'effort en linguistique soit à nouveau rééquilibrée en faveur d'une part plus grande de la recherche sur des langues spécifiques, des dialectes, des familles de langues etc., et une proportion plus petite de la linguistique générale, ce qui ne veut pas dire que celle-ci serait réduite à néant. Il y a beaucoup à dire sur la linguistique en général.

7 Notre façon d'envisager la grammaire en général, cependant, ne fait pas entrer en jeu une nouvelle façon de définir les grammaires des langues humaines. Pendant des décennies, les théoriciens de la grammaire ont discuté du bien-fondé des formalismes comme la théorie X-barre, la grammaire lexico-fonctionnelle, la grammaire des arbres, la grammaire relationnelle, et bien d'autres. Nous n'ajouterons pas une autre candidate à cette liste d'analyses rivales de la grammaire et nous ne prendrons pas parti dans les débats houleux entre les candidates existantes.

8 Nous reconnaissons l'existence de certaines généralisations structurelles que l'on peut faire sur la grammaire de toute langue humaine. Les linguistes ont été conscients pendant bien longtemps que la hiérarchie, à savoir la structure arborescente, est centrale en grammaire. Les mots se regroupent dans des petites unités comme les expressions, les petites unités se regroupent dans des unités plus larges, et ainsi de suite jusqu'à ce qu'on obtienne des unités plus larges ayant une cohérence structurelle, des phrases complexes contenant des propositions subordonnées qui peuvent à leur tour inclure des propositions avec un niveau de subordination plus bas. Quand les constructions grammaticales impliquent de faire bouger des groupes de mots d'une position à une autre, les mots mutés seront typiquement une unité complète dans la structure

arborescente, plutôt que comprenant le dernier mot d'une unité et les nouveaux mots dans l'unité suivante.

9 Nous n'avons pas besoin que la linguistique du siècle dernier nous enseigne la centralité de la structure arborescente en grammaire, bien que ce fût peut-être seulement au 20e siècle que quelqu'un pensa à demander pourquoi ce type abstrait de structure était important. On répondit à cette question en 1962 grâce au psychologue, expert en intelligence artificielle, gagnant du prix « Nobel en science économique Herbert Simon. L'essai qu'il écrivit The Architecture of Complexity » (1962) montra que tout produit complexe d'évolution graduelle, qu'elle soit culturelle ou biologique, est, en vertu du caractère inévitable statistique, fortement susceptible de présenter une structure hiérarchisée. Il est juste plus facile de développer de petites unités et, par la suite, de s'appuyer sur elles pour développer des structures plus larges, que de ne s'appuyer sur aucune structure pour aller directement vers une structure compliquée. Le fait que nous apprenons notre langue maternelle en partant de petites unités vers de plus grandes unités est la raison pour laquelle nous pouvons reconnaître des séquences de mots signifiants en-dessous de la phrase.

10 Les langues humaines sont des institutions culturellement évoluées, de sorte que l'argument de Simon explique pourquoi la structure arborescente est souvent trouvée chez elles. L'idée que la théorie grammaticale devrait dégager des formalismes descriptifs détaillés provient d'une croyance répandue que les langues du monde partagent des propriétés structurelles universelles qui sont bien plus spécifiques qu'une simple tendance à assembler de grandes unités à partir de plus petites unités. Comme le précise Steven Pinker (1995 : 409) :

> The babel of languages no longer appear to vary in arbitrary ways and without limit. One now sees a common design to the machinery underlying the world's language, a Universal Grammar.

11 Les théories formelles permettent de codifier la grammaire supposée universelle : l'idée est que les formalismes corrects offriront une définition pour toute structure qui obéit aux contraintes universelles sur la diversité langagière ; or une langue « non naturelle » hypothétique qui viole les contraintes serait une langue qui n'est pas définissable à l'intérieur des formalismes.

12 Pour notre part, nous ne voyons aucun mécanisme qui puisse contraindre les langues toujours changeantes aux quatre coins de la planète à partager des traits structurels détaillés – et contrairement à ce qu'affirme Pinker, elles ne le font pas en fait. Par conséquent, nous ne sommes pas intéressés par les formalismes. A l'instar de Haspelmath (2010), nous croyons que les linguistes devraient être encouragés à décrire la diversité des structures langagières dans le monde, en utilisant toutes les techniques descriptives adéquates.

13 Les types de questions qu'il nous semblent judicieux de se poser à propos de la grammaire en général sont les suivantes :

- Si la structure grammaticale d'une langue est développée par la communauté qui l'utilise, et acquise par des énonciateurs individuels, selon un modèle non prescrit à l'avance, quel est le degré de complexité de cette structure ?
- Est-ce que ces domaines particuliers de la grammaire sont plus ou moins précisément définis en regard des autres domaines ?
- Est-ce que les réponses à certaines de ces questions diffèrent, par exemple, selon les moyens d'expression de langue écrite ou orale ?

- Pouvons-nous faire des généralisations sur le chemin que prennent les enfants pour atteindre les niveaux de raffinement grammatical atteint par les adultes ?

14 Telles qu'elles sont exprimées, ces questions restent générales, en ce sens qu'elles ne sont pas reliées à des constructions particulières dans des langues individuelles, qui peuvent ne pas avoir d'homologues dans d'autres langues. Elles surgissent pour toutes les langues.

15 Cependant, ces questions ne peuvent être étudiées utilement qu'en se référant à une langue en particulier ou à plusieurs langues. Il est raisonnable d'espérer que les conclusions pour une langue donnée seront des pistes d'exploration pour d'autres langues, mais le degré de similarité de ces langues pour ces questions ne peut émerger que par comparaison avec des études séparées. Puisque ces questions sont plutôt nouvelles, nous avons eu l'occasion de les explorer pour seulement une langue : l'anglais moderne.

16 Ainsi, pour résumer, nous nous proposons de circonscrire la croissance et les limites de la précision grammaticale en anglais.

3. La primauté de la parole

17 La forme de langage la plus biologiquement naturelle est la parole. Les enfants deviennent des énonciateurs parlant couramment leur langue avant même d'avoir appris à lire et écrire, et même en Occident, au 21e siècle, certains adultes sont analphabètes. Il est certain que des langues parlées développées existaient bien avant que tout langage soit réduit à l'écrit, et même aujourd'hui, certaines langues n'ont pas de forme écrite, ou, si elles possèdent des transcriptions orthographiques, celles-ci ne jouent pas de rôle véritable dans la vie de la communauté langagière qui les pratique.

18 Ceci étant dit, il semble particulièrement souhaitable d'étendre notre discussion portant sur précision grammaticale, jusque-là limitée au langage écrit. Dans le domaine de la parole, les problèmes que nous avons considérés ne s'envolent pas, d'autres problèmes surgissent. En tout et pour tout, les linguistes intéressés par la grammaire ont eu tendance à travailler avec du matériel écrit (Linell 2005), ou bien ont inventé des exemples plus stylistiquement comparables à de l'écrit qu'à du langage parlé spontané. Mais la tendance est paradoxale, quand on considère à quel point de nombreux linguistes de ces 50 dernières années ont été influencés par la doctrine selon laquelle la grammaire est innée plutôt qu'acquise dans l'esprit humain. Si l'on veut examiner les aspects du langage qui sont innés plutôt qu'acquis, il y a plus de chance de les trouver dans le domaine biologiquement naturel de la parole que dans le domaine relativement artificiel de l'écrit.

19 Notre propre approche de ces questions provient de travaux consistant à étendre le programme d'annotation arborescente du programme SUSANNE d'anglais écrit pour l'appliquer de façon prévisible à des échantillons de l'anglais parlé spontané. Certains des échantillons annotés qui ont émergé de ce travail ont été publiés, comme le corpus CHRISTINE. Le fichier de documentation CHRISTINE inclut plusieurs sections définissant des ajouts et des modifications des préconisations de SUSANNE nécessaires pour couvrir l'anglais parlé. Comme dans le cas du programme SUSANNE pour l'anglais écrit, le travail qui consiste à développer des recommandations d'annotation satisfaisantes pour l'oral était un processus progressif visant à proposer des règles et enlever des erreurs en expérimentant l'application de ces règles à des échantillons particuliers. Pour le parler

spontané, nous avons travaillé avec du matériel émanant d'un nombre de sources d'anglais oral transcrit ; tout le matériel inclus dans la banque arborescente CHRISTINE telle qu'elle a été publiée est tirée d'une section de parole « d'échantillon démographique » du *British National Corpus*, qui comprend de la langue orale des années 1990 produite par un large échantillon d'énonciateurs répartis de façon équilibrée en termes de région, âge, classe sociale, et genre.

20 Dans ce qui suit, nous examinons certaines difficultés rencontrées dans l'élaboration de recommandations d'annotation modifiées qui ont respecté, pour la grammaire de la parole spontanée, le principe de Jane Edward selon lequel « des exemples similaires devraient être encodés de façons similaires et prévisibles ».

21 Les problèmes rencontrés sont de types différents : certains proviennent sans doute de la nature du langage parlé lui-même, certains, comme les énoncés que les scribes trouvèrent peu clairs, sont liés peut-être davantage à la situation de recherche qu'aux propriétés intrinsèques du langage. Mais l'on ne devrait pas évacuer trop rapidement ces problèmes et les classer comme de purs problèmes pratiques d'enregistrement de données computationnelles, peu pertinents pour la grammaire en tant que capacité intellectuelle humaine. Un individu acquérant sa langue maternelle apprend par lui-même à partir d'exemples, qu'il doit enregistrer et classer mentalement d'une certaine façon afin de construire un ensemble d'informations sur la façon dont le langage fonctionne. Cet individu trouvera sûrement certains mots peu clairs et fera face à des difficultés comparables à celles discutées plus haut. L'examen de difficultés dans l'élaboration d'un programme prévisible pour l'annotation structurelle de la parole est une bonne façon d'identifier et de confronter certaines des difficultés d'un natif cherchant à identifier une grammaire pour sa langue maternelle à partir de l'exposition à des exemples.

4. Le taggage de mots

22 Un aspect fondamental de l'annotation grammaticale est la classification des rôles grammaticaux de mots en contexte – le taggage de mots. Le programme original de SUSANNE a défini un alphabet de plus de 350 tags distincts pour l'anglais écrit, dont la plupart sont applicables à la langue orale, bien que quelques-uns n'aient aucun lien pertinent avec la parole (par exemple, les tags pour les chiffres numéraux, ou les opérateurs mathématiques). La langue orale, aussi, cependant, fait usage de ce que Anne-Brita Stenström (1990) appelle « discourse items », qui ont des fonctions pragmatiques qu'on peut difficilement mettre en parallèle avec l'écrit : par exemple, *well* utilisé comme initiateur d'énoncé. Les items discursifs sont classés dans des catégories qui, dans la plupart des cas, sont aussi clairement distinctives que les classifications applicables à des mots écrits, et l'application du programme CHRISTINE au corpus SUSANNE procure un ensemble de tags pour les items discursifs développés à partir de la classification de Stenström. Cependant, le fait que les items discursifs ne sont pas syntaxiquement intégrés dans des structures plus larges veut dire qu'il y a une faible possibilité de trouver des preuves permettant de résoudre l'ambiguïté du taggage.

23 Ainsi, trois classes d'items discursifs dans CHRISTINE sont les Explétifs (par exemple, *gosh*), les marqueurs de Réponse (par exemple, *ah*), et les marqueurs de Sons Imitatifs (par exemple, *glug glug*). Considérons les extraits suivants (1) et (2) dans lesquels des enfants « jouent aux chevaux », l'un assis sur le dos de l'autre :

24 (1) Enonciateur PS1DV : ... all you can do is <pause> put your belly up and I'll go flying! ... Go on then, put your belly up!
Enonciateur PS1DR : Gung! (KPC.00999-1002)

25 (2) Chuck a chuck a chuck chuck! Ee ee! Go on then. (KPC. 10977)

26 Dans le premier cas, *gung* n'est pas un exemple d'explétif anglais standard, ni une imitation vocale d'une étape quelconque du jeu en cours. Inversement, dans le second cas, *ee* pourrait aussi bien être l'explétif régional du nord exprimant la surprise (les énonciateurs étaient des nordistes), ou une imitation vocale d'un son « hennissant ». Dans de nombreux cas semblables, l'analyste est forcée par le programme actuel de faire des choix arbitraires ; cependant, ces cas explicites de classes d'items discursifs sont trop distincts les uns des autres pour justifier l'élimination d'une approche intuitive en fusionnant les classes en une seule.

27 Certes, tous les mots parlés qui posent des problèmes de taggage ne sont pas des items discursifs. Dans (3) :

28 (3) Ah ah! Diddums! Yeah. (KSU.00396-8)

29 l'énonciateur est ici un jeune homme de 21 ans parlant à un adolescent de 13 ans. Tout énonciateur anglophone aura reconnu le mot *diddums* comme impliquant que l'énonciateur considère le co-énonciateur comme infantile, mais l'intuition ne dit pas comment le mot devrait être taggé (nom ? et si c'est le cas, nom propre ou commun ?) ; les dictionnaires n'aident pas. Nous n'avons trouvé aucune règle de principe pour déterminer ce qu'on doit faire dans de tels cas.

5. Les réparations langagières

30 Probablement la catégorie la plus importante où les standards analytiques grammaticaux développés pour la langue écrite ont besoin d'être étendus pour représenter la structure des énoncés oraux spontanés est le domaine des réparations langagières, où les énonciateurs trouvent leur énoncé peu satisfaisant et le modifient au fil de l'eau. Le système CHRISTINE pour annoter les réparations langagières repose sur le travail de Wim Levelt (1983) et de Peter Howell et Keith Young (1990, 1991). Cette approche a identifié un ensemble qui allait jusqu'à 9 marqueurs de réparations à l'intérieur d'un énoncé réparé, par exemple le point auquel le premier plan grammatical de l'énonciateur est abandonné (« le moment d'interruption »). Cependant, cette approche n'est pas pleinement satisfaisante pour de nombreuses réparations langagières de la vie de tous les jours. D'une part, elle n'est pas suffisamment informée : la notation Levelt/Howell et Young n'offre aucun moyen de montrer comment une séquence locale contenant une réparation s'insère dans l'architecture grammaticale plus vaste de l'énoncé la contenant. À bien des égards, la notation s'avère trop riche ; elle demande à ce que des réparations soient conformes à un modèle canonique dont de nombreuses réparations dévient en pratique.

31 En conséquence, CHRISTINE emploie une version simplifiée de cette notation (Sampson 1995), dans laquelle « le moment d'interruption » dans une réparation langagière est marqué (par un ashtag), mais ne permet pas d'identifier d'autres étapes importantes. Cette approche est opérante pour la majorité des réparations langagières, par exemple (4) et (5) :

32 (4) That's why I said [Ti :o to get ma- ba- #, get you back then] ... (KBJ.00943)

33 (5) I'll have to [VV0v# cha- # change] it (KCA.02828)

34 Dans le premier exemple, *to get ma- ba-* (dans lequel *ma-* et *ba-* sont des mots tronqués), et *get you back then*, sont des tentatives successives de produire une proposition infinitive (Ti) fonctionnant comme complément d'objet (:o) de *said.* Dans le deuxième exemple, le label VV0v# veut dire « structure de réparation comprenant des essais successifs pour prononcer un mot taggé VV0v » (forme de base d'un verbe à emplois transitifs ou intransitifs).

35 Cependant, bien que la notation pour des réparation langagières dans CHRISTINE soit moins informée que le procédé Levelt/Howell et Young, son application systématique n'est pas toujours facile. En effet, les analystes sont obligés de décider si les portions de mots sont en fait des réparations langagières ou bien des constructions bien formées, ce qu'il n'est pas souvent facile de déterminer.

6. Constructions syntaxiquement Markoviennes

36 Un autre type de problèmes est posé par les énoncés que l'on pourrait appeler « syntaxiquement Markovien », dans lequel chaque élément est en cohérence logique avec ce qui précède immédiatement alors que l'énoncé dans sa totalité doit être jugé incohérent, du moins selon les standards de la prose écrite. Les exemples suivants (6) et (7) sont extraits du London-Lund Corpus (les énonciateurs sont respectivement un étudiant en Licence, âgé de 36 ans, décrivant son interview pour une bourse à Oxford et Anthony Wedgwood Benn, député, sur un programme radio) :

37 (6) ... of course I would be willing to um <pause> come into the common-room <pause> and uh <pause> in fact I would like nothing I would like better (S.1.3.0901-3)

38 (7) and what is happening <pause> in Britian today <pause> is ay – demand for an entirely new foreign policy quite different from the cold war policy <pause> is emerging from the Left (S.5.5 0539-45)

39 Dans le premier exemple, *nothing* fonctionne simultanément comme le dernier mot prononcé d'une séquence prévue comme *I would like nothing better* et le premier mot prononcé d'une séquence impliquée comme *there is nothing I would like better.* Dans le deuxième exemple, le syntagme nominal *an entirely new foreign policy* fonctionne à la fois comme complément de la préposition *for*, et comme sujet de *is emerging*. J. Miller et Reinert (1998 :40) citent d'autres exemples de ce type et les appellent « attachement bi-directionnel ».

40 Peut-être que les séquences comme celles-ci devraient être envisagées comme une sorte de « parole réparée », mais si tel est le cas, on ne peut pas les analyser dans les termes précisés dans la section précédente : on ne peut pas identifier un point unique où un plan grammatical est abandonné en faveur d'un autre. Plus important encore, parce que ces structures font entrer en jeu des expressions qui jouent simultanément un rôle grammatical dans la construction précédente et un rôle différent dans la construction qui suit, elles résistent à l'analyse en termes de diagrammes à constituants en forme d'arbre. En construisant le Corpus CHRISTINE, il se trouve que les linguistes ont été forcés d'offrir une notation labellisée, parce que cette approche de la représentation de la grammaire était fondamentale pour le reste du travail au point qu'il semblait impensable de l'abandonner pour ces cas-ci. Mais on est conscient qu'en faisant cela, on représentait les données d'une façon inappropriée.

41 Pour ceux qui envisagent le concept de description grammaticale sérieusement, cette situation est presque scandaleuse. S'il est une chose que nous pensons savoir de la grammaire dans le langage humain, c'est qu'elle impose une structure hiérarchique sur les séquences de mots : des groupes de mots nichés dans des groupes de mots plus inclusifs, que l'on peut dessiner en diagrammes arborescents. Cependant, pour les exemples précités, cette hypothèse est désavouée.

42 Pour l'anglais parlé, il n'est pas même clair que ces cas-ci soient décrits comme des « erreurs de performance ». De tels énoncés sont relativement clairs et ne peuvent être remis en question par des grammairiens professionnels. Un éditeur changera les mots pour une publication écrite, assurément ; mais qu'est-ce qui nous prive du droit de voir de tels faits comme la marque de différences entre les normes culturelles de l'oral et de l'écrit ? Certaines langues ont de très grandes différences entre les grammaires des formes orales et écrites, et aucun linguiste n'en conclurait que les langues orales ne sont « pas des langues réelles », ou « n'ont pas de grammaire ».

43 Les séquences syntaxiquement Markoviennes, par conséquent, sont une illustration extrême du principe selon lequel la grammaire évolue dans des directions imprévisibles. Quelqu'un qui essaierait de spécifier une sorte de série de constructions potentielles que les langues pourraient adopter n'inclurait sûrement pas des structures telles que celles-ci dans sa liste ; cependant, nous voyons qu'en anglais oral, elles peuvent surgir et fonctionner en discours.

7. Des distinction logiques dépendantes du médium écrit

44 Il y a des cas où les distinctions de catégorie grammaticale qui sont très saillantes en anglais écrit semblent bien moins significatives dans la langue orale, de sorte que les maintenir dans le plan d'annotation ne serait pas fidèle à la structure de la parole. Peut-être que la plus importante de ces distinctions est la distinction entre le discours direct et indirect. L'anglais écrit fait grand cas de la distinction explicite entre le discours direct, qui implique un engagement à transmettre précisément les mots exacts de l'énonciateur cité, et le discours indirect qui préserve le sens général de la citation. Le plan d'annotation SUSANNE utilise des catégories qui reflètent cette distinction. Cependant, les signaux les plus pertinents sont des marqueurs orthographiques comme les guillemets, qui n'ont pas d'homologues oraux. Parfois, la distinction peut être faite en anglais oral par l'emploi de pronoms, de formes verbales, de vocatifs, etc. :

45 (8) … he says he hates drama because the teacher takes no notice, he said one week Stuart was hitting me with a stick and the teacher just said calm down you boys … (KD6.03060)

46 Le pronom souligné *he* (plutôt que *I*) implique que le complément de *says* est au discours indirect ; l'emploi de *me* implique que le passage commençant par *one week* est une citation directe, et l'impératif *calm* et le vocatif *you boys* impliquent que l'enseignant est cité directement. Mais en pratique, ces signaux sont fréquemment en conflit l'un avec l'autre plutôt qu'ils ne se renforcent :

47 (9) (énonciateur rapportant sa propre réponse à une objection directement citée) :
I said well that's his hard luck! (KCT.10673)

48 (10) well Billy, Billy says well take that and then he'll come back and then he er gone and pay that (KCJ.01053-5)

49 Dans la première citation, l'item discursif *well* et le temps présent de *is* après le temps passé *said* suggèrent qu'il s'agit du discours direct, mais *his* suggère qu'il s'agit du discours indirect. De même dans la deuxième citation, *well* et l'impératif *take* impliquent l'emploi du discours direct, *he'll* plutôt que *I'll* impliquent le discours indirect. On peut arguer qu'imposer une nette distinction entre discours direct et indirect est une distorsion ; on pourrait, au lieu de cela, dire que le discours utilise une seule construction pour rapporter les énoncés d'autres personnes. D'autre part, la distinction entre ces deux modes est si fondamentale qu'un plan analytique qui manquerait de la reconnaître pourrait être jugé inacceptable.

8. Usage non standard

50 Le discours britannique spontané contient de nombreuses écarts vis-à-vis de l'usage standard que ce soit pour les mots individuels ou les modèles syntaxiques.

51 Dans le cas du taggage, la règle de SUZANNE (Sampson 1995 : §3.67) est que les mots utilisés pour les dialectes non-standard doivent être taggés de la même façon que les mots qui les remplacent pour l'anglais standard. Cette règle est raisonnable dans le contexte de l'anglais écrit, où les formes non standard sont peu usitées, mais il devient rapidement clair, si l'on analyse du discours spontané, que la règle est problématique pour l'analyse du discours qui contient un taux élevé de ces formes. Pour l'annotation de la parole, il a été estimé nécessaire de renverser cette règle particulière ; en général, les mots utilisés dans les fonctions grammaticales non standard se voient attribué les mêmes tags que dans leur emploi standard, bien que les expressions les contenant soient taggées en relation avec leur fonction grammaticale en contexte.

52 Cette révision de la règle tend à être peu problématique pour les pronoms et les déterminants ; soit (11) et (12) :

53 (11) It's a but of fun, it livens up me day (KP4.03497)

54 (12) She told me to have them plums (KCT.10705)

55 Dans ces exemples, les mots soulignés sont taggés comme des pronoms objets, mais les expressions *me day* et *them plums* sont taggés comme des syntagmes nominaux. Il est plus difficile d'appliquer de façon prévisible une telle règle dans le cas des emplois non standard de formes de verbe fort, où le mot utilisé en termes non standard est une tête selon les règles de SUZANNE. Les formes de base standard peuvent être utilisées dans des contextes passés, soit (13) :

56 (13) A man bought a horse and give it to her, now it's won the race (KCJ.01096-9)

57 La solution qui consister à tagger de telles suites comme un groupe de verbal au passé (Vd) est ainsi remise en question parce que souvent, l'anglais non standard omet l'auxiliaire de la construction perfective standard, suggérant que *give* ici pourrait remplacer *given* plutôt que *gave* ; soit (14) et (15) :

58 (14) What I done, I taped it back like that (KCA.02536)

59 (15) What it is, when you got snooker on and just snooker you're quite <pause> content to watch it ... (KCA.02572)

60 Ces formes seront bien familières pour la plupart des lecteurs britanniques, mais elles ont été peu étudiées systématiquement. Edina Eisikovits (1991 : 134) a argué en effet que le système temporel présent dans ces propositions comme *What I done* est le même que celui de l'anglais standard, alors qu'une forme simple *done* est utilisée à la fois pour le temps passé et le participe passé dans le dialecte non standard ; *I done* ici correspondrait à la forme standard *I did.* Cependant, cela semble négliger les cas semblables à l'exemple précédent où *got* correspond explicitement à la forme standard *have got*, voulant dire "have", et non à un temps passé.

61 Il n'est pas souhaitable que l'annotation soit basée sur des analyses grammaticales adaptées à chaque dialecte non standard. Il n'existe pas de frontière rigide entre un dialecte régional et un autre, ou entre un dialecte régional et la langue standard nationale ; peu d'énonciateurs régionaux, assurément, auront une faible expérience de l'anglais standard. Cependant, il n'est pas facile de spécifier des règles stables permettant d'annoter des emplois non standard classés comme des déviations du dialecte standard connu. Le programme CHRISTINE essaie d'introduire de la prédictabilité dans l'analyse de cas comme ceux-ci en reconnaissant un temps supplémentaire en anglais non standard réalisé comme participe passé non précédé d'un auxiliaire, et en statuant que toute forme verbale utilisée dans une structure non standard avec une référence passée sera classée comme un participe passé. Cette approche fonctionne bien dans de nombreux cas, mais elle reste à éprouver pour tous les usages qui surgissent.

62 Passons de la morphologie verbale au niveau de la syntaxe : un exemple d'une construction non standard requérant l'adaptation de l'annotation écrite de l'anglais est fourni par les propositions relatives contenant à la fois un pronom relatif et un syntagme relativisé non effacé, phénomène non connu en anglais standard mais usuel dans des dialectes non standard, soit par exemple (16) :

63 (16) ... bloody Colin who, he borrowed his computer that time, remember? (KD6.03075)

64 Ici l'approche adoptée par le programme CHRISTINE est de traiter le syntagme nominal relativisé (*he*) comme appositionnel au pronom relatif. Pour le cas cité, cela fonctionne ; mais cela ne fonctionnerait pas s'il existait un cas où l'élément relativisé n'est pas le sujet de la proposition relative.

65 Les exemples de ce type soulèvent la question de savoir ce que veut dire spécifier des standards d'annotation grammaticale cohérents applicables à un spectre de dialectes différents, plutôt qu'à un seul dialecte homogène. L'anglais écrit se conforme plus ou moins aux normes de la langue nationale standard, de sorte que la variation dialectale grammaticale est considérée comme marginale et que les standards d'annotation peuvent se permettre de l'ignorer. Dans le contexte de la langue parlée, on ne peut l'ignorer. Il n'en demeure pas moins que l'adoption des standards d'annotation pour des structures variées non prévisibles semble conceptuellement confuse.

9. Différence dialectale vs. erreur de performance

66 D'autres problèmes se révèlent lorsque l'on cherche à déterminer si une expression doit être annotée normalement en rapport au dialecte non standard de l'énonciateur, ou bien selon l'usage standard mais comportant des mots omis analysés en tant qu'erreur de performance. Les énonciateurs omettent souvent des mots nécessaires ; soit, par exemple (17) :

67 (17) There's one thing I don't like <pause> and that's having my photo taken. And it will be hard when we have to photos. (KD2.03102-3)

68 Il semble prudent d'assumer que l'énonciateur avait l'intention de dire quelque chose comme *have to provide photos.* On pourrait penser qu'un processus similaire permet d'expliquer les mots soulignés dans (18) :

69 (18) Oh she was shooting at him at dinner time <shift shouting> Steven <shift> oh god dinner time she was shouting him.

70 où *at* manque : mais ceci est remis en question quand d'autres énonciateurs, dans des exemples séparés, ont produit les énoncés suivants :

71 (19) go in the sitting room until I shout you for tea (KPC.00332)

72 (20) The spelling mistakes only occurred when <pause> I was shouted. (KD2.02798)

73 Ces exemples sont suffisants pour penser que *shout* a un emploi transitif en anglais non standard.

74 Ce problème est particulièrement courant à la fin des énoncés, où l'énoncé est susceptible d'être interprété comme interrompu brusquement avant d'être grammaticalement complet, mais il pourrait être compris comme une élision intentionnelle non standard. Dans (21) :

75 (21) That's right, she said Margaret never goes, I said well we never go for lunch out, we hardly ever really (KE2.08744)

76 les mots *we hardly ever really* n'apparaîtraient pas en anglais standard sans un verbe, par conséquent la séquence serait très probablement interprétée comme un énoncé tronqué d'une proposition comme *we hardly ever really go out to eat at all* ; mais il n'est pas difficile d'imaginer que le dialecte de l'énonciateur pourrait permettre *we hardly ever really* à la place de l'anglais standard *we hardly ever do really.*

77 Il semble inconcevable qu'un plan d'annotation détaillée puisse échouer à distinguer la différence entre dialecte et erreur de performance. Mais les analystes n'auront souvent en pratique aucune base pour appliquer la distinction à des exemples particuliers.

10. Mauvaises transcriptions

78 On ne peut s'attendre à ce que chaque mot d'un échantillon de discours spontané enregistré sur le terrain soit parfaitement transcrit à partir des enregistrements. Notre travail repose sur les transcriptions données par d'autres chercheurs, qui contiennent de nombreux passages jugés « peu clairs » ; la même chose serait sans aucun doute vraie si nous avions choisi de rassembler notre propre matériel. Un système d'annotation structurelle a besoin d'être capable d'assigner une analyse à un passage contenant des segments manquant de clarté ; éliminer un énoncé ou une phrase contenant un seul mot peu clair demanderait à ce qu'on mette trop de données de côté et un choix injustement subjectif serait en fait en faveur d'une collection d'énoncés qui seraient prononcés avec soin avec des propriétés structurelles spéciales. Telles sont les considérations pour le linguiste computationnel ; mais s'il pense à l'acquisition individuelle de sa langue maternelle à travers la simple exposition à des exemples, il doit aussi assurément rencontrer des morceaux de discours où on ne peut identifier les mots et doit par conséquent enregistrer ces énoncés mentalement ainsi que les énoncés parfaitement

clairs – il est impossible que l'apprenant ignore totalement un énoncé s'il contient même la plus brève portion inintelligible.

79 Le programme SUZANNE utilise le symbole Y pour labelliser des nœuds dominant des portions de discours peu claires ou bien des portions qui ne peuvent être assignées une catégorie grammaticale parce qu'elles contiennent des segments peu clairs qui rendent la catégorisation confuse.

80 Ce système n'est pas problématique ; tant que le matériel peu clair en fait consiste en un ou plusieurs constituants grammaticaux. Souvent, cependant, ce n'est pas le cas ; soit (22) :

81 (22) Oh we didn't <unclear> to drink yourselves. (KCT.10833)

82 Ici, il semble que la portion peu claire contenait des mots multiples, commençant avec un mot ou plusieurs mots qui complètent le groupe verbal (V) initié par *didn't* ; et conduisant à ce quela relation des mots *to drink ourselves* à la principale soit différente. Par exemple, si les mots peu clairs étaient *give you anything*, alors *to drink* serait un tagme modifiant à l'intérieur d'un syntagme nominal dont la tête est *anything* ; d'autre part, si la portion peu claire était *expect you*, alors *to drink* serait la tête d'une proposition complément d'objet. Idéalement, un plan d'annotation grammaticale devrait permettre d'indiquer clairement toute la grammaire sans pousser l'analyste à prendre des décisions trop arbitraires.

83 Une grammaire doit permettre des parenthésages labellisés clairs. Cependant, il est très difficile de trouver des conventions notationnelles viables qui éviteraient des choix arbitraires pour des structures auxquelles des mots peu clairs contribuent. Notre meilleure tentative de définition des conventions notationnelles dans ce domaine est un ensemble de règles qui prescrivent, entre autres choses, que le nœud Y dominant une portion inaudible doit être attaché au noeud le plus bas qui domine explicitement au moins le premier mot inaudible, et que les mots clairs suivant une portion inaudible sont attachés à ce nœud Y dominant si les mots clairs peuvent être intégrés à tout constituant inconnu initié dans la portion inaudible.

84 Ces conventions fonctionnent raisonnablement bien et permettent aux analystes de produire des annotations de manière prévisible ; mais elles présentent le désavantage que de nombreuses structures produites sont assurément différentes des structures grammaticales présentées par les mots réellement prononcés. Par exemple, dans l'exemple précédent, le Y au-dessus de la portion peu claire est une fille du verbe dominant *didn't*, *drink yourselves* est placé sous le nœud Y, parce que toute interprétation plausible conduisant à l'incompréhension ferait des derniers mots une partie du tagme initié dans la portion peu claire. Cependant, il est impossible que *to drink yourselves* puisse réellement être une partie d'un groupe verbal commençant avec *didn't*.

85 Si les chercheurs qui travaillent avec ces données sont convaincus qu'une structure arborescente incluant un nœud Y autorise seulement des affirmations limitées portant sur la structure réelle produite par l'énonciateur, ces conventions ne sont pas trompeuses. Mais en même temps, elles ne sont pas satisfaisantes.

11. Conclusion

86 Quand on annote l'anglais écrit, et que l'on s'appuie sur une tradition analytique qui a évolué depuis des siècles, il semble dans l'ensemble que la plupart des décisions

d'annotation sont efficaces ; là où tel exemple particulier est vague et amène à hésiter entre deux catégories, celles-ci tendent à être des sous-catégories d'une même catégorie de niveau supérieur, de sorte qu'une annotation neutre de rechange est disponible.

87 Une façon de résumer les nombreux problèmes soulignés dans cet article est de dire que, en annotant un discours dont les traits structurels particuliers ont eu peu d'influence sur la tradition analytique, les ambiguïtés de classification surgissent constamment, traversant les plans des catégories traditionnelles. Par conséquent, il est souvent difficile de choisir une annotation qui attribue des propriétés spécifiques à un exemple. Contrairement à la langue écrite, il est souvent aussi très difficile de définir des notations de rechange permettant au scribe d'éviter d'attribuer des propriétés pour lesquelles il n'y a pas de preuves, tout en accueillant ce qui est peut-être dit et exprimé avec assurance.

88 Pour de nombreux énonciateurs, les problèmes auxquels font face les linguistes computationnels ne sont pas pertinents. Reste que le langage reflète l'activité humaine, même s'il appartient à une variété qui n'a pas bénéficié de la longue histoire de l'examen philologique que possède l'anglais écrit. La linguistique des récentes décennies a eu tendance à penser que les énoncés avaient des structures grammaticales bien définies, même si des débats théoriques sur l'identité de ces structures ont fleuri. Cependant, cette impression de clarté ne vient pas de la nature de la langue elle-même mais de l'existence d'une tradition de description grammaticale bien établie, qui offre des décisions établies portant sur la structure de certains aspects de la langue, et qui encourage les héritiers de la tradition à négliger des aspects du langage pour lesquels cette dernière n'a pas produit de réponses toutes faites.

89 Prenons un seul exemple : il est évident que l'emploi par les linguistes génératives de la catégorie VP, qui découle de la décision de traiter les structures sujet-verb-complément d'objet comme des structures bipartites plutôt que tripartites, vient du rôle joué par les concepts « sujet » et « prédicat » dans la logique Aristotélienne et médiévale. Il se peut que certains linguistes théoriciens aient proposé une analyse bipartite indépendante de cette tradition, mais si c'est le cas, la plupart des linguistes qui commencent à établir des grammaires de l'anglais en écrivant S → NP VP, VP → V NP (plutôt que S → NP V NP) ne sont pas familiers avec ces arguments.

90 Quand on regarde le discours spontané, on considère un genre de langage pour lequel il existe peu de tradition de description philologique. Et, assurément, quand le discours spontané contient des phénomènes qui ne sont pas identiques à des structures de l'écrit, et que la tradition philologique nous donne aucune réponse toute faite sur la façon de les analyser, nous sommes souvent perdus quand il s'agit de proposer des analyses stables. La tradition philologique est un paradigme Kuhnien qui encourage les linguistes à confiner leur pensée de la grammaire à un domaine limité, dans lequel les questions ont des réponses raisonnablement fournies, et qui rejettent les phénomènes atypiques à l'extérieur du champ de vision des linguistes. La linguistique computationnelle fondée sur l'étude de corpus est une bonne façon de se libérer de ces œillères intellectuelles et de s'exposer à l'anarchie désordonnée du langage.

91 Comme nous l'avons vu, quand on fait cela, même l'anglais écrit s'avère être bien moins défini que le paradigme nous laissait imaginer. L'anglais oral souligne ce point à merveille.

BIBLIOGRAPHIE

CHRISTINE corpus

SUSANNE corpus

Chomsky, Noam. *Syntactic Structures.* Gravenhage: Mouton, 1957.

Dalke, Anne. "All Grammars Leak." 2010. Consulté sur le forum Serendip (serendip.brynmawr.edu/exchange/node/7645) le 14 Janvier 14 2017.

Eisikovits, Edina. "Variation in the Lexical Verb in Inner-Sydney English." *Australian journal of Linguistics* 7: 1-24, 1987. Réimprimé in Trudgill and Chambers (1991 : 120-142).

Graddol, David G. "The Future of Language." *Science* 303 (2004): 1329-1331.

Graves, Norman. "John Miller Dow : educationist and prolific textbook author." *Paradigm* 2 (8), October 2004, <faculty.ed.uiuc.edu/westbury/paradigm/vol2/Graves.doc>

Gross, Maurice. "On the Failure of Generative Grammar." *Language* 55 (1979): 859-885.

Haspelmath, Martin. "Framework-free Grammatical Theory." In Bernd Heine and Heiko Narrog (eds.), *The Oxford Handbook of Grammatical Analysis*. Oxford: Oxford University Press, 2010.

Howell, James. *A New English Grammar, Prescribing as Certain Rules as the Language Will Bear, for Forreners to Learn English*. London: T. Williams, H. Brome, and H. Marsh, 1662.

Howell, Peter and Keith Young. *Speech repairs: report of work conducted October 1st 1989-March 31st 1990*. Department of Psychology, University College London, 1990.

Howell, Peter and Keith Young. "The Use of Prosody in Highligting Alterations in Repairs from Unrestricted Speech." *Quaterly Journal of Experimental Psychology* 43A (1991): 733-758.

Levelt, Willem J. M. "Monitoring and Self-repair in Speech." *Cognition* 14 (1983): 41-104.

Linell, Per. *The Written Language Bias in Linguistics: Its Nature, Origins, and Transformations*. London: Routledge, 2005.

Miller, Jim and Regina Reinert. *Spontaneous Spoken Language: Syntax and Discourse.* Oxford: Clarendon Press, 1998.

Pinker, Steven. *The Language Instinct: The New Science of Language and Mind*. London: Penguin, 1995.

Sampson, Geoffrey. *English for the Computer: the SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press, 1995.

Sapir, Edward. *Language: An Introduction to the Study of Speech.* Reprinted London: Hart-Davis, [1921] 1963.

Simon, Herbert A. "The Architexture of Complexity." *Proceedings of the American Philosophical Society* 106 (1962): 467-482. Réimprimé in Simon, *The Sciences of the Artificial*, 84-118 ; Cambridge, Mass. : MIT Press, 1969.

Stenström, Anna-Brita. "Lexical Items Peculiar to Spoken Discourse." I Jan Svartvik (ed.), *The London-Lund Corpus of Spoken English*. Lund: Lund University Press, 1990.

Stockwell, Robert P., Paul Schachter and Barbara Partee. *The Major Syntactic Structures of English*. New York: Holt, Rinehart and Winston, 1973.

RÉSUMÉS

Réfléchir sur le système d'une langue, c'est réfléchir sur la grammaire qui la régit en tenant compte des exemples recevables et non recevables par le locuteur d'une langue parlée. Cet article défend la position qu'il n'existe pas de grammaticalité puisqu'aucun linguiste n'a jamais réussi à établir une frontière entre le grammatical et l'agrammatical. On se pose la question de savoir ce qu'est la grammaire puis on étend notre champ d'étude au domaine de l'oral, en particulier aux problèmes pour l'annotation de corpus d'anglais oral que sont le taggage des mots, les réparations langagières, les constructions Markoviennes, les distinctions logiques, l'usage non standard, la différence dialectale et les erreurs de performance.

Thinking over the system of a language means that you take into account the grammar that regulates it by relying on the examples that are acceptable and not acceptable for the speaker of the spoken language. This article defends the position of the non existence of grammaticality since no linguist has ever established a frontier between what is grammatical and what is not. The category of grammar is discussed then the study extends to the spoken language with the problems brought about by the annotation of a corpus of spoken English that are word tagging, linguistic repairs, Markovian constructions, logical distinctions, non standard usage, and dialectal difference and performance errors.

INDEX

Mots-clés : grammaticalité, grammaire, langue orale, annotation
Keywords : grammaticality, grammar, spoken language, annotation

AUTEUR

SYLVIE HANCIL

Université de Rouen
hancilfr@yahoo.fr