

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

Fanny RINCK, Agnès TUTIN
LIDILEM, E.A. 609, Grenoble 3.

Introduction

Les passages entre guillemets (PEG) sont représentatifs des difficultés que pose l'analyse des phénomènes polyphoniques, qui résistent souvent à une lecture univoque et à un schéma d'annotation parfois simplificateur. Il est de ce point de vue intéressant de réfléchir aux procédures qui permettent de mieux circonscrire ce phénomène linguistique et d'envisager son annotation. Le recours au contexte linguistique immédiat et plus largement au contexte énonciatif doit aider à déterminer des frontières entre les emplois, mais aussi à tenir compte de leur complexité.

Dans cet article, nous présentons une étude de faisabilité sur l'annotation des fonctions des passages entre guillemets, réalisée dans le cadre d'un projet de plan-pluriannuel formation (PPF) sur les marqueurs linguistiques de la subjectivité et de la polyphonie¹.

Dans un premier temps, en nous appuyant sur la littérature, nous récapitulons les principales valeurs sémantiques véhiculées par les guillemets, en faisant l'hypothèse de configurations relativement stables de leurs emplois. Puis, nous rapportons une méthode d'interannotation, qui permet de tester la faisabilité du projet en cernant les variations d'interprétation et d'ajuster ainsi les consignes. Nous détaillons alors les indices

1. PPF piloté par le LIDILEM (2003-2007) (F. Grossmann et G. Antoniadis) : « Développement et exploitation de ressources linguistiques pour la didactique du français à l'aide d'outils de TAL. Etude des marqueurs linguistiques de la subjectivité et de la polyphonie. »

formels sur lesquels l'annotateur peut en partie s'appuyer et nous affinons la description des valeurs des passages à annoter. Nous montrons enfin la manière dont le codage peut rendre compte de la plurivocité des guillemets comme propriété interprétative de ces signes.

1. Les difficultés d'interprétation des guillemets

Pour paraître accessoires, les guillemets n'en sont pas moins nécessaires à la construction du sens. Ordinairement employés pour indiquer une citation, ils ne se limitent pas à cet usage mais ont ceci de particulier qu'ils bloquent l'interprétation littérale de l'élément qu'ils entourent. Signalant que l'élément guillemeté « ne va pas de soi », ils ne fournissent cependant pas d'autre instruction que cette rupture.

L'usage des guillemets consiste à dire un élément et à signaler, seulement, à son sujet, qu'on est en train de le dire [...] Par rapport à la simple énonciation du signe standard, ce qu'ajoute sa représentation par le moyen de ce signal, c'est une sorte de manque, de creux à combler interprétativement, un « appel de glose » si l'on veut. [...] En fin de compte, il s'agit d'un signe non ambigu, à valeur générale – celle d'une pure opacification – associé en discours à un ensemble non fini d'interprétations. (Catach citée par Authier-Revuz, 1995 :137)

L'hétérogénéité interprétative des guillemets a pour conséquence une absence de description unifiée dans le champ de recherches. La diversité des typologies proposées (Authier, 1981, 1995 ; Fouquier, 1981 ; Martin, 1983 ; Cheong 1985 ; Dahlet, 2003), que nous n'exposerons pas ici dans le détail, montre que trois problématiques linguistiques interviennent.

- **L'autonomie**, *i.e.* la possibilité qu'ont tout signe et tout énoncé de référer à eux-mêmes, dans le sens où les guillemets signalent d'un élément qu'« on est en train de le dire » (cf. *supra*). Cette propriété se manifeste en particulier dans le discours métalinguistique (*le mot* « X »).

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

- La **polyphonie**, *i.e.* la multiplicité des instances de prise en charge énonciative de l'assertion (Bres *et al.*, 2005 ; Perrin, 2006). La plus évidente fonction des guillemets, celle de signalement d'une citation, peut être rattachée à cette problématique : une citation consiste en une rupture au niveau de la source à l'origine du dit. Mais à l'instar de J. Bres et B. Vérine (2002) on peut considérer que la citation n'est qu'une des formes les plus manifestes de la polyphonie et que cette dernière renvoie plus fondamentalement à toute forme de dédoublement énonciatif ; les travaux de J. Authier-Revuz (1995) notamment montrent que c'est en ces termes que peuvent être traités les guillemets qui servent, par exemple, à exprimer sa réserve, à signaler un terme trop technique ou un néologisme, et que nous désignons *infra* comme des guillemets de modalisation.

- La **dénomination** et les différentes facettes de cette activité (catégorisation, construction des objets du discours, mais aussi expression d'un point de vue). Les guillemets peuvent en effet signaler qu'on associe un signifiant à un concept, notamment dans la pratique de définition (*on appelle* « X »)

A travers la complexité des guillemets se manifeste toute la complexité du signe linguistique ; en envisageant le rôle des guillemets en termes d'instructions, à partir des trois problématiques de l'autonomie, de la polyphonie et de la dénomination, nous faisons l'hypothèse de configurations relativement stables d'emplois des guillemets. L'annotation de la valeur des passages entre guillemets représente un objectif dans notre projet d'enrichissement d'un corpus à des fins didactiques, mais elle peut alors aussi constituer une heuristique : elle doit permettre, au-delà du constat de l'hétérogénéité interprétative des guillemets, de mieux caractériser leurs propriétés par l'identification de valeurs stables et de difficultés récurrentes.

En premier lieu, il convient de valider la faisabilité d'une telle entreprise d'annotation de la fonction des guillemets. Nous recourons à cette fin à une méthode d'interannotation ; elle nous semble fournir un outil intéressant pour travailler à l'élaboration d'une procédure d'annotation finalisée par la prise

en compte des variations dans l'interprétation et des indices formels sur lesquels cette dernière peut s'appuyer.

2. Une méthode d'interannotation comme étape de validation du projet

Annoter les PEG dans un corpus de textes variés présuppose un modèle stable. Une phase intermédiaire d'annotation doit permettre de déterminer si un modèle de description linguistique des guillemets est envisageable et c'est dans cette perspective que nous proposons une méthode d'annotation interannotateurs.

2.1. Premier schéma d'annotation

L'expérimentation a consisté à faire annoter indépendamment trois personnes, à partir de la première description du phénomène évoquée *supra*. Les trois problématiques de l'autonymie, de la polyphonie et de la dénomination qui jouent dans les instructions fournies par les guillemets pour l'interprétation du passage qu'ils entourent nous permettent de dégager quatre catégories d'emplois :

- les guillemets signalent le statut exclusivement autonymique du PEG :

[1] Ainsi, pour un sourd, « papa » et « maman » sont pratiquement identiques au niveau de la perception. [LSF]²

- les guillemets, polyphoniques, signalent que le PEG est une citation ; cette catégorie en recouvre deux³ : les citations autonomes au plan syntaxique sont dissociées des citations non autonomes, *i.e.* intégrées à la syntaxe du discours citant :

[2] Il n'y avait que quelques villes clairsemées le long de la côte et le pays était une « immense zone rurale avec pour seules activités l'agriculture et le nomadisme » (Benabdi, 1980: 7). [Filles]

2. Le détail des corpus utilisés apparaît en bibliographie.

3. La distinction entre deux sous-catégories permet d'affiner la description, mais la fonction des guillemets est dans les deux cas celle de citation.

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

- les guillemets, polyphoniques, signalent que le PEG fait l'objet d'un commentaire modalisant (*i.e.* une modalisation, par exemple, une forme de réserve sur l'emploi du mot) :

[3] La langue des signes, langue gestuelle « parlée » par les communautés sourdes. [LSF]

- les guillemets signalent que le PEG est une catégorie dénomminative :

[4] C'est en m'appuyant sur la distinction dégagée par John J. Gumperz (1982) entre « we code » et « they code »⁴ [Rap]

Par ailleurs, il est autorisé de procéder à un double balisage, quand on hésite entre deux balises, ou quand deux valeurs semblent intervenir.

La méthode a été mise en œuvre sur trois textes relevant de l'écrit de recherche en sciences humaines (deux articles et un rapport)⁵, du domaine de la sociolinguistique, soit un corpus de 24 000 mots contenant 309 PEG. Les annotateurs sont des linguistes sensibilisés aux questions d'énonciation⁶, mais un seul peut être considéré comme expert dans ce domaine. Chaque annotateur a effectué le balisage indépendamment en utilisant les valeurs proposées. Un premier document de référence décrivait simplement les différents types d'emploi en proposant quelques exemples, comme nous l'avons fait de manière synthétique ici. Les résultats des annotations ont ensuite été confrontés. L'objectif de la méthode était de voir dans quelle mesure l'on pouvait s'accorder sur des valeurs stables, de caractériser les variations dans l'interprétation et d'affiner ainsi les consignes d'annotation.

2.2. Résultats et analyse

Les résultats de la comparaison interannotateurs sont assez encourageants. Le premier constat que l'on peut faire en observant globalement les résultats au terme de la procédure

4. Cet exemple atteste de l'utilisation des guillemets pour l'introduction de concepts mais nous y revenons plus loin car il correspond aussi à un cas de citation.

5. Cf. détails du corpus en bibliographie.

6. Une étudiante de M2, une doctorante, un maître de conférences.

(cf. tableau 1) est que le degré d'accord entre les trois annotateurs concerne près de 4 balises sur 5. Par ailleurs, un accord interannotateurs minimal est presque toujours obtenu, les cas de triple annotation restant très rares (moins de 1%). Ces résultats valident l'hypothèse qu'il existe des valeurs stables des guillemets ; s'ils s'expliquent en partie par la prédominance de guillemets signalant une citation, facilement repérables, ils suggèrent qu'une annotation du phénomène doit être possible, à condition d'analyser les cas de désaccord.

	Texte Rap (5167 mots)	Texte LSF (13 222 mots)	Texte Filles (4348 mots)	Tous textes
Nombre de PEG	140	125	44	309
Taux d'accord	98 (70%)	109 (86,5%)	36 (82%)	243 (78,5%)
Taux de désaccord	42 (30%)	16 (13,5%)	8 (18%)	66 (21,5%)

Tableau 1 – Calcul du taux d'accord : repérage des balisages (simples ou doubles) identiques

L'imprécision des consignes constitue probablement un facteur entravant une annotation homogène. Cela étant, les variations observées entre annotateurs ne peuvent être imputées à ce seul facteur, car le taux enregistré pour le désaccord dépend d'abord largement du texte annoté : il peut passer du simple au double. Les variations entre annotateurs ne peuvent donc pas être entièrement imputées aux consignes. Elles signalent que la difficulté interprétative se situe au niveau des guillemets eux-mêmes, qui peuvent faire l'objet d'emplois plus ou moins explicites selon les textes.

Par ailleurs, si l'on relève les types d'opposition en observant le nombre de balises attribuées (cf. tableau 2), on remarque que l'attribution d'une balise simple (une valeur) aux guillemets est le plus souvent consensuelle et que la majorité des cas de désaccord intervient lorsque l'attribution d'une seule balise s'oppose à un double balisage intégrant la même balise et

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

une autre (X *versus* X+Y). On note également quelques rares cas de triple choix (X *versus* Y *versus* X+Y).

Type de désaccord	Nombre relevé
balise simple <i>vs</i> balise simple (X <i>vs</i> Y)	18 (27,5%)
balise simple <i>vs</i> balise double (X <i>vs</i> X+Y)	42 (68%)
triple balisage (X <i>vs</i> Y <i>vs</i> X+Y)	3 (4,5%)

Tableau 2 : Repérage des types de désaccords par rapport au nombre de balises attribuées

Les variations entre annotateurs résultent donc de problèmes de frontières entre certaines valeurs. L'examen détaillé des principaux désaccords sur les valeurs employées permet de voir que ces problèmes (cf. tableau 3) se jouent pour beaucoup au niveau des guillemets à valeur polyphonique (citation et modalisation) et de l'ajout ou non d'une dimension polyphonique à des guillemets à valeur de dénomination.

Type de désaccord	Nombre relevé
modalisation <i>vs</i> modalisation + dénomination	13
citation non autonome <i>vs</i> citation non autonome + dénomination	9
dénomination <i>vs</i> modalisation	8
dénomination <i>vs</i> dénomination + citation non autonome	7
citation non autonome <i>vs</i> citation non autonome + modalisation	4

Tableau 3 : Repérage des principaux types de désaccords par rapport aux valeurs attribuées

L'exemple suivant illustre bien les difficultés de l'annotation.

- [5] Ces efforts de formation du rappeur constituent une dépense anticipant un profit symbolique en termes de position sociale et « d'autorité ». [Rap]

Le caractère flottant de l'usage des guillemets se manifeste en premier lieu au niveau des frontières formelles du PEG, puisque l'on pourrait s'attendre à ce que seul « autorité » soit mis entre guillemets.

Par ailleurs, l'interprétation du PEG fait l'objet d'hésitations fortes. On peut considérer que « d'autorité » correspond à une étiquette terminologique du domaine de spécialité (valeur de dénomination), mais la frontière particulière du PEG et l'absence de guillemets pour « position sociale » justifie une analyse qui voit plutôt dans ce terme une réserve de l'auteur (valeur de modalisation).

Cependant, si l'on se réfère au texte pris dans son ensemble, on observe que le même PEG « d'autorité » intervient en amont, où il se présente à la fois comme une dénomination et comme un emprunt à un auteur (valeur de citation non autonome) : *la possession de la parole est légitimée, de même que la position « d'autorité » au sens de Michel de Certeau (1974).*

Cet exemple montre la nécessité d'introduire des directives très claires pour l'annotation.

Il convient en effet de s'interroger sur les critères contextuels, qui seul permettent de trancher entre une valeur de citation et une valeur de modalisation : la citation suppose que le PEG apparaisse comme un énoncé antérieurement produit, et que le locuteur qui en est à l'origine soit mentionné. Or, si l'on se restreint à l'énoncé [5] pris isolément, on conclura à une valeur de modalisation, alors qu'on l'analysera comme une citation si on le met en perspective avec l'occurrence précédente du PEG « d'autorité ». Quelle doit être l'étendue du contexte à prendre en compte ? La mention d'une source (citation) ne se situe pas nécessairement dans la phrase où intervient le PEG, et c'est dans l'ensemble du texte qu'il faut la rechercher⁷.

7. Le contexte d'interprétation des guillemets peut aller jusqu'à impliquer l'intertexte, si l'on en juge, dans le même article, le terme

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

Il importe également de spécifier les cas où on admet une double valeur du PEG. La valeur de dénomination implique souvent une forme de prudence de l'auteur (*ce que j'appellerais « X »*), voire de distance face à un terme consacré (*on appelle « X »*, *« X » est souvent défini comme*, etc.). Cela étant, à côté d'exemples plus nets de distance (du type *on pourrait être tenté d'appeler « X »*), on ne peut systématiquement adjoindre à la valeur de dénomination une valeur de modalisation, bien que les deux problématiques linguistiques soient liées. Si le double balisage est intéressant, en particulier pour une exploitation didactique des annotations, il faut aussi définir de manière stricte les catégories descriptives et introduire dans certains cas une hiérarchisation des valeurs pour les rendre opératoires.

La méthode d'interannotation confirme que les difficultés posées par nos catégories descriptives sont inévitables face à la complexité du phénomène. Dans les cas de désaccord, la discussion qui s'ensuit permet de limiter l'impact de l'imprécision des consignes et de réajuster ces dernières pour aboutir à un consensus sur la plupart des PEG. La procédure d'annotation à élaborer une fois les corrections apportées doit rendre explicites les choix de codage, notamment en détaillant les indices formels sur lesquels peut s'appuyer l'analyse des PEG.

3. L'élaboration d'une procédure d'annotation

3.1. Les indices formels

Pour faciliter la cohérence de l'annotation, il est souhaitable de recourir le plus possible à des indices formels, bien qu'à eux seuls ces éléments soient rarement suffisants pour déterminer la fonction du PEG, en tout cas pour certaines d'entre elles. Les valeurs que nous avons dégagées sont souvent corrélées à des indices lexicaux ou syntaxiques spécifiques, et un inventaire systématique de ces éléments doit bien entendu être intégré dans le guide d'annotation. Une formalisation fine de ces

« distinction » ; on peut comprendre la mise entre guillemets comme une référence à P. Bourdieu, référence implicite que seul le lecteur expert peut reconstruire.

marques devrait par ailleurs permettre d'envisager une automatisation partielle du balisage avec des outils de TAL⁸. Un ensemble de caractéristiques typo-dispositionnelles, syntaxiques et lexicales peuvent être formulées pour chacun des emplois dégagés. Les éléments présentés ici ont été observés sur un corpus de travail de près de 430 000 mots, utilisé dans le cadre de notre projet PPF⁹ et semblent valables quel que soit le genre de texte considéré.

La citation autonome est un emploi facilement caractérisable à partir de purs critères formels. On relève d'ailleurs peu de désaccords mettant en jeu ce type de fonction (cf. Tableau 3). Au plan typo-dispositionnel, une citation autonome peut apparaître dans un paragraphe décroché, comportant plusieurs phrases. La citation autonome apparaît également souvent à la suite des deux points dans le cas du discours direct. Au plan syntaxique, la citation autonome est constituée d'une ou plusieurs phrases, au sens large, puisque certaines d'entre elles peuvent être averbales. Des cas plus complexes apparaissent parfois dans les cas de discours direct, lorsque le discours citant apparaît avant ou après le discours cité dans la même phrase « graphique », ou dans les cas d'incise, comme dans l'exemple suivant, où la citation autonome est « tronçonnée » :

[6] « N'importe, *dit Pencroff*, si je ne vois pas la côte, je la sens... elle est là... là... aussi sûr que nous ne sommes plus à Richmond ! » [ILE MYSTERIEUSE, J. Verne]

Il apparaît donc souvent nécessaire d'effectuer une véritable analyse syntaxique pour isoler la citation autonome. Bien entendu, des indices lexicaux (verbes de parole ou de pensée)

8. Ce travail a été partiellement accompli par C. Hermann (2004) dans son mémoire de maîtrise sur le balisage automatique des citations dans les textes journalistiques. L'annotation automatique du discours direct, isolant le discours cité du discours citant, s'est révélée réalisable dans 70 à 80% des cas.

9. Le corpus comporte 393 000 mots de corpus littéraires (romans du 19^{ème} siècle), 12600 mots d'un essai polémique (*Le droit à la paresse* de Lafargue), et les 24000 mots des écrits de sciences humaines utilisés dans l'évaluation interannotateurs.

*Annoter la polyphonie dans les textes : le cas des passages entre
guillemets*

peuvent introduire ces citations, mais cela n'a rien de systématique.

Les citations non autonomes apparaissent plus complexes à repérer sur le plan formel. Formes mixtes entre le discours direct et le discours indirect, ces séquences sont intégrées syntaxiquement dans le discours enchâssant et, contrairement à la citation autonome, elles ne comprennent pas de marques typographiques ou lexicales propres au discours (typiquement, la ponctuation à valeur énonciative, le « je-ici-maintenant », etc.). Elles peuvent constituer des parties de phrases et sont souvent accompagnées de patrons lexicaux spécifiques ayant trait à la source du discours cité (*selon N, d'après N, N pense que, N dit que*)¹⁰ :

[7] Il allait sur les boulevards, [...], vous *répétant* vingt fois de suite *que* « les cuirassiers blancs de Bismarck avaient été écrasés jusqu'au dernier... » [CONTE DU LUNDI, A. Daudet]

La fonction autonymique est toujours présente dans les fonctions de dénomination ou de modalisation. Nous ne considérerons cependant comme purement autonymiques que les emplois où le signe linguistique n'a pas une fonction référentielle. Sur le plan formel, cette fonction met souvent en jeu des parties du discours variées¹¹ ou des syntagmes, mais rarement des éléments plus longs. On remarque souvent des introducteurs métalinguistiques comme *le mot, l'expression, le verbe X ; un terme comme X ; signifie X*, mais ces marques ne sont pas systématiques (elles ne représentent que 11 occurrences sur 58 éléments balisés, soit un taux de 19%) :

[8] Pencroff avait rayé le mot « impossible » du dictionnaire... [ILE MYSTERIEUSE, J. Verne]

10. Ces formes mixtes sont particulièrement complexes ; nous renvoyons aux contributions de L. Rosier (1997, 1999) sur la « binarité et le continuum » entre le discours direct et indirect et aux descriptions qu'elle fournit de l'extrême variété de leurs combinaisons.

11. Les PEG peuvent même comprendre dans cette fonction des mots grammaticaux, des phonèmes transcrits, des mots étrangers.

Sur le plan syntaxique, les emplois dénominatifs présentent moins de variété que les emplois autonymiques. Ils renvoient également à des segments syntaxiques courts, généralement des noms, des syntagmes nominaux, mais parfois aussi des adjectifs (qui servent dans ce cas de classificateurs). Ils sont souvent introduits par des termes classificateurs comme *le terme / le nom de X, on appelle / nomme X, dit X*). Ici aussi, l'emploi de ces termes est loin d'être régulier puisqu'il n'apparaît que dans 20% des cas.

Contrairement aux emplois précédents, les PEG de modalisation sont difficiles à isoler à l'aide de marques formelles. Ils relèvent de natures syntaxiques variées¹² et sont généralement assez courts¹³. Nous n'avons pas relevé de marques méta-énonciatives comme celles qui sont signalées par J. Authier-Revuz (1984) : *si je puis dire, si vous voulez, passez-moi l'expression, en quelque sorte, si l'on veut, n'ayons pas peur des mots...* qui paraissent relever plus de l'oral que de l'écrit (ces marques n'ont-elles justement pas à l'oral la même fonction que les guillemets à l'écrit ?).

Pour résumer, les marques formelles relèvent de divers paliers (le paragraphe pour les aspects typo-dispositionnels, la syntaxe, le lexique, la ponctuation) et n'apparaissent suffisantes que pour déterminer la plupart des emplois de citations, en particulier pour les citations autonomes. Dans le cas de l'autonymie et de la dénomination, les marques, quand elles sont présentes, sont un auxiliaire précieux pour l'annotateur. En revanche, pour la modalisation, l'analyse devra reposer sur une analyse fine du texte, qui devra prendre en compte à la fois le contexte et le cotexte, ce qui suppose d'en affiner la description comme nous le verrons dans la section suivante.

3.2. Description systématique des différents emplois

Puisque les indices formels n'apparaissent pas suffisants pour déterminer certains emplois comme l'autonymie, la dénomination ou la modalisation, il apparaît donc indispensable

12. Les adjectifs, noms et groupes nominaux prédominent.

13. Même si les phrases ne semblent *a priori* pas exclues, nous n'avons pas rencontré ce cas dans le corpus.

*Annoter la polyphonie dans les textes : le cas des passages entre
guillemets*

d'affiner la description sémantique de ces emplois dans le guide d'annotation afin de permettre aux annotateurs de prendre des décisions cohérentes.

Le cas de l'autonymie est finalement assez simple à résoudre, puisque cette valeur peut être limitée aux purs emplois non référentiels.

Dans les cas de dénomination, une description plus fine passe par un inventaire exhaustif des valeurs possibles, assez variées, à illustrer par des exemples pour les annotateurs. La dénomination, qui associe l'emploi d'un signifiant à un référent ou à un concept spécifique, est employée dans les cas suivants, qui peuvent se superposer :

- l'introduction d'un concept nouveau (*on appelle « oraliste »*) ;
- l'introduction d'un terme spécialisé ou appartenant à une langue étrangère (*le « slang » américain*) ;
- la mention d'un nom propre (*le cheval « Coco »*) ;
- la mention d'un titre (*le roman « Une Vie »*) ;
- la mention d'une façon de parler (terme régional, idiolecte) qui n'a pas valeur de modalisation (*une fille « québécoise »*).

La modalisation est la valeur la plus complexe à attribuer, puisqu'elle présuppose une analyse du contexte énonciatif, en particulier des intentions de l'auteur qui exprime une distance vis-à-vis de l'énoncé et entend la signifier au lecteur.

La distance peut simplement concerner l'inadéquation du signifiant par rapport au message qu'on veut exprimer. Le terme ne relève pas du niveau de langue souhaité ou ne correspond qu'imparfaitement au référent à décrire :

- [9] On peut imaginer la réflexion d'un adolescent confronté à l'incompréhension d'un message émanant de ses idoles, censées avoir partagé les mêmes expériences, les mêmes « galères ». [Rap]

Le PEG peut également avoir une fonction plus polémique en exprimant la non adhésion, voire le rejet de l'auteur par rapport à un énoncé prêté ou repris à d'autres :

- [10] Le langage des rappeurs comme celui des jeunes en mal d'insertion est souvent considéré comme « vulgaire ou incivil ». [Rap]

Ces deux fonctions apparaissent donc assez différentes du point de vue énonciatif, et il aurait peut-être été souhaitable de les dissocier dans l'annotation. Nous y avons cependant renoncé, comme pour les valeurs de dénomination, afin de ne pas multiplier les cas de figure à l'infini.

3.3. *Le codage face à la plurivocité foncière des guillemets*

L'inventaire des marques formelles et la description fine des différents emplois rencontrés ne suffisent pas toujours à interpréter de façon univoque les guillemets. Dans certains cas, les PEG semblent relever de plusieurs emplois simultanément. Dans d'autres, l'interprétation demeure véritablement ambiguë et le choix entre plusieurs valeurs n'est pas décidable.

La spécification du schéma doit permettre d'éviter nombre de difficultés d'annotation, mais le codage ne doit pas faire illusion : s'agissant d'un signe complexe, les guillemets ne peuvent se satisfaire d'une annotation qui gommerait cette complexité, *a fortiori* dans une perspective d'exploitation didactique du corpus enrichi.

Notre schéma d'annotation (cf. l'annexe 1 pour la DTD et l'annexe 2 pour un extrait de texte annoté) intègre donc des précisions d'une part sur les cas de polysémie du PEG, d'autre part sur les cas d'ambiguïté, et vise ainsi à tenir compte de la plurivocité des guillemets, tout en limitant les doubles annotations à des configurations définies.

Les cas de *polysémie* ne sont pas considérés comme tels lorsqu'il est possible d'envisager une valeur primordiale du PEG. Ceux qui restent concernent prioritairement la valeur de modalisation, couplée à la valeur de citation ou à celle de dénomination. Les guillemets peuvent en effet, outre ces valeurs, suggérer un commentaire de l'auteur sur le terme (inadéquat par rapport à ce qu'il désigne, impropre vis-à-vis du lecteur, utilisé par d'autres mais auquel l'auteur refuse de souscrire). Nous décidons de ne procéder à un double balisage que lorsque la valeur de modalisation, co-présente, est marquée

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

par la présence de commentaires méta-énonciatifs (*en quelque sorte, pour ainsi dire*, etc.), ou, fait moins rare dans notre corpus, quand on a affaire à un terme marqué subjectivement, en particulier des noms, des adjectifs ou des adverbes axiologiques¹⁴ (*cet adolescent en « intégration sauvage »*).

Par ailleurs, il s'avère nécessaire d'introduire une catégorie spécifique pour un cas récurrent à l'origine de désaccords, les proverbes (« *Il n'y a pire sourd que celui qui ne veut entendre* »). Proches des citations, leur source n'est cependant pas réellement identifiable, mais il s'agit d'une catégorie trop particulière pour être considérée comme de la modalisation.

Outre les valeurs de modalisation et citation, et de modalisation et dénomination, un dernier cas de polysémie concerne la double valeur de dénomination et de citation. Cette double valeur se manifeste dans des emplois bien circonscrits. Ils sont relativement fréquents dans l'écrit scientifique quand l'auteur utilise un concept et y adhère (dénomination) tout en imputant sa paternité à un autre auteur (citation)¹⁵ :

- [11] C'est en m'appuyant sur la distinction dégagée par John J. Gumperz (1982) entre « we code » et « they code », que j'ai étudié les productions de groupes marseillais, en posant l'hypothèse que le discours rap [...] [Rap]

L'**ambiguïté** concerne quant à elle l'hésitation entre une valeur de citation et de modalisation, non attestée dans notre corpus d'écrits scientifiques, mais que révèle l'exemple suivant, tiré d'un corpus journalistique :

- [12] La loi américaine considère en effet qu'il n'y a pas de différence « substantielle » entre les produits manipulés et leurs homologues naturels (*L'Express*, 18/11/1999)

14. Nous nous référons à l'analyse de la subjectivité langagière menée par C. Kerbrat-Orecchioni (1980).

15. On rencontre également fréquemment cette double valeur dans le texte littéraire : l'expression entre guillemets est attribuée au personnage (citation) et à sa vision du monde, sa conceptualisation (dénomination) : *Dans les jours de pluie, elle restait enfermée en sa chambre à visiter ce qu'elle appelait ses « reliques »* [UNE VIE, G. de Maupassant].

Rien ici ne permet de dire si le terme mis entre guillemets provient de la loi américaine, où s'il s'agit d'un terme dont use l'auteur sans l'assumer complètement.

Face à la plurivocité foncière des guillemets, le codage doit être guidé par des directives strictes, qui intègrent les descriptions fournies ici, mais il n'en demeure pas moins que des difficultés peuvent se poser à l'annotateur. Nous avons finalement opté pour une annotation réalisée par un binôme, qui discute les valeurs attribuées et garde la possibilité de signaler un cas d'ambiguïté lorsque les deux annotateurs n'arrivent pas à se mettre d'accord, en dépit des procédures explicitées dans le guide d'annotation.

Dans l'écrit scientifique, de telles difficultés interviennent quand l'annotateur n'est pas du domaine de spécialité ; il peut hésiter sur l'attribution d'une valeur de dénomination quand celle-ci n'est pas marquée et qu'il ne sait pas si le PEG correspond à un terme consacré dans le champ (par exemple avec *adolescent en « intégration sauvage »* cité plus haut).

Plus largement, les guillemets renvoient à l'implicite des textes, et constituent en ce sens une entrée intéressante pour analyser cette dimension. L'étude menée par J.H. Thaumoux (2002) sur l'usage des guillemets dans la revue politique *Présent* est particulièrement intéressante de ce point de vue, car c'est sur l'implicite des guillemets que repose la dimension polémique du discours de ce journal. Il faudrait alors mettre en perspective ces observations avec les pratiques en usage dans d'autres journaux, et même dans d'autres genres de textes, pour voir dans quelle mesure la plurivocité des guillemets est ou non exploitée, et quels sont ses effets argumentatifs. La tendance à l'ambiguïté, ou à l'inverse à l'explicitation maximale, pourrait ainsi aider à mieux cerner en quoi l'interprétation d'un phénomène local comme les guillemets est déterminée par le fonctionnement global d'un genre.

Au demeurant, l'étude menée sur notre corpus d'écrits de sciences humaines suggère également le poids des différences disciplinaires à ce niveau-là. Dans le domaine de la

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

sociolinguistique, les guillemets sont particulièrement fréquents, et montrent l'importance, dans ce domaine, de la distance critique de l'auteur vis-à-vis de certains termes qui circulent dans le champ social (par exemple, *les « quartiers »*, *le « langage des jeunes »*) ; les guillemets, traces d'une rumeur des discours, révèlent comment le champ se constitue en prise contre le sens commun, tout en intégrant aussi certaines doxa spécifiques, par exemple dans la mise entre guillemets systématique d'un terme comme la « langue », et le refus ainsi signalé de souscrire à cette notion structuraliste sans l'interroger.

Conclusion

L'annotation des guillemets se révèle donc une tâche assez complexe, qui ne peut reposer qu'en partie sur l'observation du contexte immédiat. Si certaines valeurs comme la citation autonome ne prêtent guère à confusion, les emplois de modalisation et de dénomination exigent une parfaite compréhension du texte, du domaine exposé et de ce que l'auteur entend signifier au lecteur. Malgré la complexité de la tâche, l'étude interannoteurs réalisée montre toutefois qu'il n'apparaît pas irréaliste d'envisager l'annotation des PEG. Un repérage systématique des marques formelles et un inventaire détaillé des contextes d'emplois permettent de faciliter la tâche de l'annoteur. Certaines valeurs demeurent toutefois ambiguës et il convient de prévoir dans l'annotation des cas de cumul de valeurs ou d'ambiguïtés en cas de désaccord entre les annoteurs.

Le schéma d'annotation présenté dans cet article a été appliqué à un corpus, essentiellement littéraire, de 430 000 mots dans le cadre de notre projet PPF¹⁶. Les cas de polysémie et d'ambiguïté et leur lien avec la question de l'implicite des textes prennent tout leur sens dans une perspective didactique. On peut ainsi imaginer un travail de décryptage systématique du guillemet et de ses valeurs dans les textes, par exemple dans l'écrit scientifique. Face à l'abondance des guillemets dans les

16. Le corpus sera bientôt interrogeable en ligne sur : <http://ppfesc.free.fr/>.

copies d'étudiants, il pourrait favoriser une meilleure conscientisation des pratiques, et, en cela aussi, le contrôle de leurs effets.

Références bibliographiques

- Authier J. (1981). « Paroles tenues à distance », in B. Conein, J.J. Courtine, F. Gadet, & M. Pêcheux (eds.), *Matérialités discursives*. Lille : Presses Universitaires de Lille, 127-142.
- Authier-Revuz J. (1984). « Hétérogénéité(s) énonciative(s) », *Langages* 73 : 91-151.
- Authier-Revuz J. (1995). *Ces mots qui ne vont pas de soi. Boucles réflexives et non-coïncidences du dire*. Paris : Larousse, 2 tomes.
- Bres J., Haillet P., Mellet S., Nolke H., Rosier L. (eds) (2005). *Dialogisme et polyphonie. Approches linguistiques*. Bruxelles : De Boeck/Duculot.
- Bres J. & Véline B. (2002). « Le bruissement des voix dans le discours : dialogisme et discours rapporté », *Faits de langue* 19 : 159-169.
- Cheong K.S. (1985). *Etude de la construction de valeurs référentielles à travers un marqueur énonciatif : le cas des guillemets*, Thèse de 3^{ème} cycle, Paris : Université Paris VII.
- Dahlet V. (2003). *Ponctuation et énonciation*, Presses Universitaires Créoles : Ibis Rouge éditions.
- Fouquier E. (1981). *Approche de la distance*. Thèse de 3^{ème} cycle, Paris : EHESS.
- Hermann C. (2004). *Repérage et annotation automatique du discours rapporté (discours direct) dans le discours journalistique*, Mémoire de maîtrise de Sciences du Langage, mention Industries de la Langue, Université Stendhal.
- Kerbrat-Orecchioni C. (1980). *L'énonciation. De la subjectivité dans le langage*. Paris : Armand Colin.
- Martin R. (1983). *Pour une logique du sens*. Paris : Presses Universitaires de France.

Annoter la polyphonie dans les textes : le cas des passages entre guillemets

- Perrin L., (éd.) (2006). *Le sens et ses voix. Dialogisme et polyphonie en langue et en discours*, *Recherches linguistique*, 28, Metz : Université Paul-Verlaine.
- Rosier L. (1997). « Entre binarité et continuum. Une nouvelle approche théorique du discours rapporté ? », *Modèles Linguistiques* 38 :7-16.
- Rosier L. (1999). *Le discours rapporté : histoire, théories, pratiques*. Paris/Louvain-La-Neuve : Duculot.

Corpus utilisé pour le test interannotateurs :

- [Filles] : Benrabah M. (1999). « Les filles contre les mères », *LIDIL*, 19.
- [Rap] : Trimaille C. (1999). « Le rap français ou la différence mise en langue », *LIDIL*, 19.
- [LSF] : Millet A. (1990). *La place de la LSF dans l'intégration scolaire des enfants sourds*, Rapport de recherche, Programme 1988 d'action spécifique « Sciences humaines et sociales ».

Annexe 1 : DTD utilisée pour les passages entre guillemets

```
<!--Passages entre guillemets-->

<!-- A chaque passage entre guillemets, on associe
une ou plusieurs fonctions -->
<!-- Il y a plusieurs fonctions quand il y a
désaccord ou ambiguïté irréductible sur l'annotation
-->
<!-- Le segment du passage entre guillemets suit -->

<!ELEMENT PEG (#PCDATA|FCT)*>
<!-- La fonction comprend une ou plusieurs valeurs,
en cas de valeurs multiples (le PEG cumule plusieurs
valeurs). -->
<!-- Ex : il y a à la fois modalisation et citation
non autonome -->

<!ELEMENT FCT (VALEUR+)>

<!-- La valeur est un élément vide auquel on associe
plusieurs valeurs d'attributs -->
<!-- L'attribut TYPE est obligatoire -->
<!ELEMENT VALEUR (EMPTY)>
<!ATTLIST          VALEUR          TYPE
(CITA_NA|CITA_A|AUTONYMIE|DENOMINATION|MODALISATION|
PROVERBE|AUTRE) #REQUIRED>
```

*Annoter la polyphonie dans les textes : le cas des passages entre
guillemets*

¹⁷Annexe 2 : Exemple de passages annotés

Extrait de l'article de Cyril Trimaille, « Le rap français ou la différence mise en langue », dans *LIDIL*, Les parlers urbains n°19, 1999, coordonné par Jacqueline Billiez.

En France, depuis le milieu des années quatre-vingt, les jeunes issus de l'immigration se sont appropriés cet art des mots, l'émancipant du modèle américain. A l'instar de leurs prédécesseurs d'Outre-Atlantique, les rappeurs français prennent la parole en leur nom et en celui d'une partie de la population qu'ils entendent représenter. Il prennent la parole comme on prend les armes, conjurant de la sorte un certain mutisme, pour sortir de la spirale de la relégation ou de la délinquance, qu'illustre bien <SN-Intro-DD> cette phrase </SN-Intro-DD> aux airs d'aphorisme : <PEG><FCT><VALEUR TYPE="CITA_A"/></FCT> " la violence a toujours été la parole du pauvre " </PEG> (Kery James, Rocca, Shurik'n, Hamed Daye, <italique>Animalement vôtre</italique>, Première Classe).

Ces caractéristiques socioculturelle et identitaire, fondamentales du rap français, ont conduit quelques chercheurs à se saisir de l'objet, le faisant entrer de plein droit dans le champ des Sciences Humaines et dans celui de la sociolinguistique. Les quelques sociolinguistes qui se sont intéressés aux textes de rap ont décelé que des <PEG><FCT><VALEUR TYPE="CITA_NA"/><VALEUR TYPE="DENOMINATION"/></FCT> " stratégies identitaires " </PEG> y étaient à l'oeuvre, et qu'elles se manifestaient particulièrement par des choix de langues. L'étude menée s'inscrit dans la lignée de ces travaux qui ont engagé la réflexion. En recherchant les différents niveaux de choix langagiers, en tentant d'inférer leur fonctions, et en appréhendant les discours en jeu comme des <PEG><FCT><VALEUR TYPE="DENOMINATION"/></FCT> " ethnométhodes " </PEG>, constitutives d'un <PEG><FCT><VALEUR TYPE="DENOMINATION"/></FCT> " raisonnement sociologique pratique " </PEG>, j'ai tenté d'approfondir certaines analyses. Cela m'a

17. Comme suggéré par un relecteur, l'annotation pourrait être plus concise en intégrant les différentes valeurs des PEG par un attribut de type IDREFS sur l'élément <PEG>.

FANNY RINCK, AGNES TUTIN

*notamment conduit à étudier les rapports
qu'entretiennent les diverses composantes d'un
répertoire métissé, et particulièrement les variétés
dominantes et celles considérées comme non-légitimes*