

# **Analyse statistique ou analyse syntaxique des textes scientifiques**

par

**J. GALLAIS-HAMONNO**

*Université de Metz - FRANCE*

1.1. L'étude des textes scientifiques anglo-saxons que nous effectuons est faite à l'aide de deux programmes informatiques, dont le premier classe les termes par ordre alphabétique, en indiquant leur numéro d'apparition dans le texte, avant de les reclasser par ordre de fréquence croissante, tandis que le deuxième permet l'étude du contexte choisi des termes analysés. Ces programmes ont été mis au point par Monsieur Georges Arkhipoff.

1.2. Le premier programme permet ainsi une étude statistique des termes utilisés par l'auteur. Nous en prendrons pour exemple la liste des termes aux fréquences les plus élevées pour le texte de PIGOU, A.C. "The value of Money"<sup>(1)</sup>.

L'examen de ces termes montre que ce sont d'abord des "mots-outils" : *the* (284 fois), *of* (205), *to* (134), *in* (116), *is* (86), *that* (81), *and* (60), *a* (57).

Les premiers termes qui ne soient pas des mots-outils sont *legal* et *tender*, tous deux utilisés avec la même fréquence : 48 fois. Pigou est en effet connu pour avoir analysé "legal tender money", et cette fréquence élevée d'utilisation de l'expression ne semble donc pas étonnante pour ce texte.

Après *legal* et *tender*, les termes les plus utilisés sont à nouveau des mots outils : *be* (40 fois), *are* et *for* (34), *it*, *this* (33), avant le terme *titles* utilisé 33 fois.

De même que pour *legal* et *tender*, cette utilisation élevée de *titles* n'étonne *a priori* pas, puisqu'elle semble indiquer l'utilisation de l'expression *titles to legal tender*, elle aussi connue pour faire partie des concepts analysés par Pigou.

Après les mots-outils *by* (27 fois) et *as*(24), les termes *resources* utilisés 24 fois et *bank* 23 fois précèdent *demand* utilisé 21 fois et *money* 19 fois, de même que *people*, *proportion* et *their*.

Si nous bornons notre analyse aux seuls termes qui ne soient pas des mots-outils et dont les fréquences soient inférieures aux termes que nous venons de noter, nous trouvons : *keep* (17 fois), *form* (16), *quantity* (16), *actual* (15), *balances*, *equation* et *production* (13), *commodities* (11), *banks* et *wheat* (10).

1.3. En résumé, les termes les plus utilisés dans ce texte, à l'exception des mots-outils sont :

Termes les plus utilisés	Nombre de fois où le terme apparaît dans le texte
legal, tender	48
titles	33
resources	23
bank	23

demand	21
money, people, proportion	19
keep	17
form, quantity	16
actual	15
balances, equation, production	13
commodities	11
banks, wheat	10

L'étude de ces termes montre que, selon son habitude, Pigou analyse dans ce texte *titles to legal tender*.

Toutefois, l'utilisation de termes tels *resources*, *commodities* et *wheat*, auxquels s'ajoute *production* montre que l'analyse de la demande (*demand*) peut ne pas être seulement celle d'une demande de monnaie (*money*). Il y a en effet juxtaposition du rôle des banques (*bank*, *banks*) et de celui de la population (*people*). Et semble-t-il celle-ci a un rôle particulier à jouer, le verbe *keep* ne pouvant s'appliquer qu'à elle. En tenant compte des termes utilisés et que nous n'avons pas encore considérés, il semblerait que l'on puisse dire que dans ce texte Pigou, outre l'étude de *titles to legal tender*, semblerait s'intéresser à une proportion (*proportion*, *quantity*), indiquée par une équation (*equation*), de quelque chose désigné par *actual*.

2.1. Par extrapolation, nous pouvons essayer d'en déduire que Pigou doit analyser dans ce texte, si la population effectue un choix entre *titles to legal tender money* et des biens tangibles (*resources*, *commodities*, *wheat*) que désignerait l'adjectif *actual*, la demande de l'une ou l'autre de ces formes d'investissement ou de placement pouvant se traduire par une équation que l'auteur doit vraisemblablement essayer d'établir dans le texte. L'important pour l'analyse économique étant la *proportion* de leurs avoirs que les agents économiques décident de garder (*keep*) dans l'une ou l'autre de ces formes d'investissement ou de placement; ce qui conditionne en effet la demande de monnaie.

2.2. Par l'analyse qui précède nous voyons que, à partir de la liste des formes les plus utilisées dans un texte, nous pouvons inférer de quoi parle ce texte. Nous n'avons cependant aucune certitude que telle est vraiment l'analyse présentée par l'auteur. La déduction porte sur l'utilisation de tel ou tel terme, auquel nous pouvons assigner un signifié erroné. Il faut en effet tenir compte du fait que le signifié d'un terme peut être modifié par son contexte d'utilisation. Et ceci est particulièrement vrai dans les textes scientifiques. L'étude d'un terme dissocié de son contexte peut, ainsi, empêcher de savoir si ce terme est employé avec son signifié de langue commune, ou avec le signifié propre à la science que l'analyse économique lui assigne.

En outre, il est possible que le terme le plus important pour l'analyse présentée par l'auteur soit très peu utilisé. Il arrive qu'il n'apparaisse qu'une seule fois. En ce cas, l'étude statistique ne peut aider au

décodage du message de l'auteur.

Enfin, l'interprétation effectuée à partir des fréquences d'utilisation des termes est entachée de subjectivité. A partir d'une sélection de mots donnée, nous pouvons les réarranger chacun à notre guise, en fonction de notre imagination ou de nos connaissances de ce que dit en général l'auteur analysé, mais rien ne garantit que notre imagination ne nous égare pas, ou que nous n'enfermions pas l'auteur dans une doctrine que nous croyons être la sienne, et qui peut justement ne pas être celle qui a guidé son choix pour le texte en question.

2.3. C'est pourquoi nous dirons que l'analyse statistique des termes utilisés dans un texte scientifique, permet de se faire une idée de ce dont parle l'auteur dans le texte, sans toutefois garantir que ces conclusions soient celles auxquelles une lecture efficace du texte conduirait.

2.4. Par "lecture efficace" d'un texte scientifique, nous entendons une lecture qui ne soit pas biaisée par les idées du lecteur qui se projette sur le texte qu'il lit, et ne perçoit que ce qu'il veut y voir. Il se trouve en effet que de telles lectures existent, pour les textes scientifiques aussi bien que pour les textes littéraires. Or, ce qui, pour le texte littéraire peut être source de richesse, par la dimension nouvelle que le lecteur ajoute au texte originel (faisant à son tour oeuvre de création), s'avère catastrophique pour le texte scientifique. Ce sont en effet de telles lectures biaisées qui freinent le progrès de la science. Celui-ci s'effectue en effet par l'apport d'une théorie ou d'un point de théorie nouveau à l'analyse connue d'un phénomène donné. Or, par la lecture biaisée, le lecteur ne reconnaît dans le texte que ce qu'il connaît déjà et ne prête pas attention à ce qui est nouveau : il évacue de ce fait tout l'apport scientifique du texte. Il n'y a alors pas progrès de la science, puisque le progrès consiste en l'adoption par d'autres chercheurs de la théorie nouvelle proposée par un auteur donné. C'est alors qu'apparaissent ce que T.S. Kuhn appelle des "paradigmes", sans lesquels la science n'avancerait pas. Tout chercheur isolé ne faisant pas progresser la science si ses recherches ne sont pas adoptées et développées par la communauté scientifique à laquelle il appartient.

2.5. Le problème du décodage adéquat du texte scientifique est donc un problème important, non seulement pour le lecteur, mais pour l'avance scientifique. C'est pourquoi il convient d'éliminer toute source de subjectivité dans la lecture des textes scientifiques.

Or, comme nous venons de le voir, l'extrapolation qui permet de déduire de quoi parle un texte à partir de l'examen des termes les plus utilisés dans le texte, n'écarte pas ce danger d'une interprétation subjective erronée.

Nous en donnerons, pour le texte qui nous sert à illustrer notre propos, l'interprétation du terme *actual*. Ce terme, avons-nous dit, doit s'appliquer aux ressources. Nous donnons au terme cette interprétation car nous savons que la science oppose *real* à *financial*. Par analogie, puisque *real* s'applique à ce qui n'est pas financier, nous pensons que *actual* doit s'y appliquer aussi, et puisque,

dans ce texte, l'on semble opposer de la monnaie à des biens, nous l'avons appliqué à ceux-ci. Toutefois, étant donné que l'auteur parle de *titles to legal tender* il n'est pas impossible que cet adjectif *actual* s'applique également à ces titres, qui ont une forme tangible. Par ailleurs, le terme *proportion*, de même que *demand*, peuvent s'appliquer aussi bien à la monnaie qu'aux biens de consommation. Or, selon qu'on l'applique à l'un ou à l'autre de ces termes, l'analyse présentée par l'auteur est totalement différente. Elle n'appartient même pas à la même branche de la science : c'est en effet ou bien de la demande de monnaie, qui fait partie de l'analyse macroéconomique, ou bien, pour la demande de biens, de la microéconomie. Or, les variables de l'équation analysée seront tout à fait différentes dans l'une ou l'autre branche de la science.

2.6. C'est pourquoi nous dirons que l'étude statistique des termes qui composent un texte scientifique, si elle éclaire sur les thèmes traités dans le texte, n'est toutefois pas fiable, et peut conduire à des interprétations imprécises, voire erronées.

3.1. Nous avons de ce fait pensé à porter notre attention, non plus sur les termes "pleins" du texte, mais sur les mots-outils. Ceux-ci sont les plus fréquents. Et ceci n'est pas vrai pour le seul texte que nous analysons ici. En particulier, l'article défini a toujours été le terme le plus fréquent pour tous les textes que nous avons analysés de façon automatique<sup>(2)</sup>. De même *of* a toujours été le terme utilisé le plus fréquemment après *the*.

Ceci ne pouvait manquer d'attirer notre attention. C'est pourquoi nous sommes passés d'une analyse statistique à l'étude du contexte des termes qui composent le texte, et en particulier l'étude du contexte des mots outils. Pour la commodité de l'analyse, nous avons trouvé que le contexte des cinq mots ou formes qui précèdent et des cinq mots ou formes qui suivent le terme analysé était suffisant pour la plupart des cas. Dans la mesure où nos deux programmes numérotent chaque terme ou forme selon son numéro d'apparition dans le texte, nous avons toujours la possibilité, si ce contexte 5 n'est pas suffisant, de retrouver dans le texte, le contexte plus vaste qui permettra de compléter ce que le contexte 5 a d'insuffisant.

3.2. Avant de présenter notre décodage automatique du message théorique présenté par un texte scientifique, nous devons toutefois rappeler<sup>(3)</sup> la théorie (que nous avons appelée la *valeur relative de la parole*) que notre étude des textes scientifiques anglo-saxons, nous a conduits à proposer.

Dans le texte scientifique, nous distinguons deux choses : le message théorique de l'auteur (que nous appelons le *langage*) et le *discours* qui permet à l'auteur de faire comprendre sa théorie au lecteur.

Tout comme "le langage des abeilles" de Benveniste, le message scientifique consiste en effet en l'énoncé de "coordonnées", représentées par les traits définitoires, que l'auteur assigne au concept. Coordonnées qui, cette fois, ne sont plus données en fonction de la position du "concept" par rapport à l'essaim et au soleil, mais par rapport à d'autres concepts, dont les traits définitoires ont été élucidés

par des analyses antérieures. Ainsi, la science progresse du connu vers l'inconnu, le connu servant d'outil d'analyse pour ce qui est à démontrer.

Toutefois, à la différence des abeilles, l'homme a besoin d'être convaincu que la théorie nouvelle proposée par un auteur est valide. Il faut pour cela que le chercheur qui innove, devienne également pédagogue. Il lui faut d'une part illustrer sa théorie par des exemples ou des références à des théories ou points de théorie connus et admis de tous : c'est ce que nous appelons le *discours scientifique*. Mais il lui faut également utiliser des procédés oratoires qui retiennent l'attention d'un lecteur ou d'un auditeur dont l'esprit peut ne pas être toujours disposé à suivre un raisonnement qui dérange ses connaissances acquises. C'est par ce que nous appelons le *discours pédagogique* que l'auteur d'un texte scientifique conduit son lecteur à travers le développement de l'analyse, grâce à l'utilisation d'une trame argumentative qui souligne pour lui les points importants du texte, et isole les propositions de *langage* sur lesquelles doit porter toute son attention.

Or, ce *discours pédagogique* est facilement repérable dans un texte scientifique, dans la mesure où il consiste en termes ou expressions qui donnent au texte une structure rhétorique qui n'a rien de propre à l'idée présentée dans ce texte. C'est-à-dire que les termes ou expressions qui composent ce *discours pédagogique* pourraient être utilisés tels quels pour présenter un tout autre point de théorie, ou une toute autre théorie scientifique.

Les utilisations de l'article défini dans les propositions de *discours pédagogique* du texte de Pigou permettent de voir comment est construit le texte.

Celui-ci commence en effet par un exemple marqué par l'utilisation de l'article défini 10, qui se prolonge jusqu'en 119 où est introduit un point de vue théorique. Celui-ci est à nouveau suivi par un exemple, introduit en 892, qui va en fait se prolonger jusqu'en 3401 où se retrouve la même expression qu'en 892 : "to clothe the dry bones". C'est donc par des exemples que l'auteur va démontrer son argument théorique, avant de comparer la théorie qu'il propose à la théorie admise à laquelle il fait référence en 3422. La différence entre les deux approches étant annoncée en 3690.

3.3. Afin de savoir interpréter la liste des utilisations de l'article entre l'occurrence 119 et l'occurrence 866, il nous faut tenir compte des règles syntaxiques propres à la langue scientifique anglo-saxonne que nous avons mises en lumière<sup>(3)</sup>.

En effet, dans le texte scientifique anglo-saxon, le *langage* et le *discours scientifique* se distinguent par une utilisation inversée des articles, et notamment de l'article défini et des adjectifs.

Alors que l'article défini sert dans les propositions de *langage* à introduire le concept objet d'analyse (concept sur lequel porte l'analyse présentée par le texte), ainsi que les concepts qui servent d'outils à cette analyse, dans le *discours scientifique*, l'article défini sert à introduire des termes ou expressions

qui désignent des faits observables, des exemples tirés de l'expérience quotidienne, ou bien des théories connues auxquelles il est fait référence afin de mieux illustrer la théorie nouvelle proposée dans le texte.

Par ailleurs, l'adjectif sert dans les propositions de *discours scientifique* à qualifier le terme ou l'expression auquel il se rapporte, c'est-à-dire qu'il sert à porter un jugement sur la chose désignée, alors que dans les propositions de *langage*, l'adjectif sert à modifier le signifié du terme ou de l'expression auquel il se rapporte, c'est-à-dire que le syntagme désigne un concept différent de celui désigné par le seul substantif, ou l'expression, avant l'antéposition de l'adjectif, ou bien, pour les syntagmes qui servent à désigner des concepts formés par un travail en extension, par sa postposition au terme ou à l'expression qu'il modifie.

3.4. L'examen des termes ou expressions introduits par l'article défini, pour les propositions de *langage* de ce texte, montre que Pigou cherche à établir une fonction mathématique : *the demand schedule for titles to (legal tender)*, qui lui permet d'établir une deuxième fonction : *the derived demand schedule for (actual legal tender)*. Ces deux fonctions auront la forme d'une "rectangular hyperbola", et les variables analysées pour les établir sont : *the aggregate value of a given (quantity of wheat)*, *the aggregate quantity of wheat*, *the aggregate number of legal tender*, *the number or units of legal (tender)* et *the value or price (per unit of these titles in terms of wheat)*.

Les utilisations de l'article défini que nous ne relevons pas dans ce résumé de la théorie de l'auteur servent à introduire les outils de la démonstration qui permettent d'expliquer le bien-fondé de l'analyse : *the provision of convenience* et *the provision of security* servent à expliquer le comportement des agents économiques. *The total resources . . . that are enjoyed by the community*, *the proportion of these resources that (it chooses to keep in the form of titles to legal tender)*, sont les éléments qui conditionnent *the choice that the general body of the public exercises*, conduisant à *the proportion that . . . the representative man (chooses to keep in actual legal tender)*.

3.5. Si nous reprenons maintenant l'interprétation que nous avons essayé de donner de ce texte à partir de la liste des termes les plus fréquemment utilisés, nous nous apercevons que notre interprétation de l'utilisation de *actual* était totalement erronée. En effet, *actual* se rapporte à *legal tender*, et non aux ressources. En outre, et c'est bien plus important, l'auteur ne cherche pas du tout à déterminer quelle proportion de leurs avoirs les agents économiques vont garder sous forme monétaire, et quelle proportion ils vont dépenser pour la consommation, comme le laissait envisager l'opposition entre *resources*, *wheat*, *quantity* et *legal tender*. Bien au contraire, l'auteur n'envisage les dépenses de consommation que comme une variable permettant de déterminer le montant de *actual legal tender* opposé aux *titles to legal tender* que vont garder les agents économiques. Toute la démonstration porte sur l'équation qui va permettre de déterminer la proportion de *titles to legal tender* opposée à la proportion de *actual legal tender* nécessaire à satisfaire la demande des agents.

Notre interprétation fondée sur les seuls termes les plus utilisés était donc parfaitement erronée.

Enfin, à l'examen des utilisations de l'article défini dans la dernière partie du texte, nous nous apercevons que notre interprétation de ce qu'il pouvait contenir, à partir des listes de fréquence de mots, était erronée en ce qui concerne le signifié du terme *resources*. En effet, nous lui avons donné un signifié proche de celui de *commodities*, par analogie avec le concept de *natural resources*. Or, nous voyons avec l'utilisation 3716 de l'article défini que Pigou donne à *resources* le signifié de moyens monétaires. La phrase est en effet ' "When people do decide to keep half as much of their resources as before in the form of titles to legal tender" '.

Comme nous l'avions dit lors de l'étude de la liste des termes les plus utilisés dans le texte, le signifié d'un terme dépend de son contexte d'utilisation, et l'extrapolation peut être extrêmement dangereuse, surtout dans un texte scientifique où les termes désignent des concepts, qui servent d'outils à une analyse et reçoivent donc de la part de l'auteur, un signifié qui dépend de l'hypothèse qu'il choisit.