Recherches sur la vérité. Définition, élimination, déflation

Henri Galinon

Université Paris I - Panthéon Sorbonne, U.F.R. de Philosophie

IHPST

THÈSE

pour obtenir le grade de
Docteur de l'Université Paris I
en Philosophie

Henri Galinon

soutenance prévue
le 25 Septembre 2010

Recherches sur la Vérité

Définition, Élimination, Déflation

Sous la direction du Professeur François Rivenc

Jury :

P. Engel	Professeur de Philosophie à l'Université de Genève
L. Horsten	Professeur de Philosophie à l'Université de Bristol
F. Rivenc	Professeur de Philosophie à l'Université Paris-I
Ph. de Rouilhan	Directeur de Recherche au CNRS
G. Sandu	Professeur à l'Université d'Helsinki

Recherches sur la Vérité

Henri Galinon

8 septembre 2010

Table des matières

Remerciements

Je remercie mon directeur de thèse, François Rivenc, qui m'a soutenu et encouragé avec attention et bienveillance tout au long de ce travail, ainsi que Pascal Engel, Leon Horsten, Gabriel Sandu et Philippe de Rouilhan qui ont accepté de faire partie du jury.

Je remercie également le directeur de l'IHPST, Jacques Dubucs, ainsi que le directeur de l'Ecole Doctorale de Philosophie de Paris-I, Jean Gayon, pour les conditions de travail exceptionnelles qui sont celles des doctorants en philosophie à l'IHPST et à l'Université Paris-I.

Pendant ces années, j'ai énormément appris des nombreux chercheurs avec lesquels j'ai eu la chance d'échanger à l'IHPST et ailleurs. Je remercie Dora Achourioti, Alexandra Arapinis, Mark van Atten, Susanna Berestovoy, Julien Boyer, Serge Bozon, Mikaël Cozic, Isabelle Drouet, Michael Detlefsen, Paul Egré, Volker Halbach, Brian Hill, Neil Kennedy, Jeff Ketland, Hannes Leitgeb, David Nicolas, Fabrice Pataut, Cédric Paternotte, Alejandro Perez Carballo, Carlo Proietti, Francesca Poggiolesi, Paula Quinon, Philippe Schlenker, Benjamin Simmenauer, Benjamin Spector, Marion Vorms, Sean Walsh et Pierre Wagner.

Je dois des remerciements particuliers à ceux qui ont bien voulu relire des parties de ce manuscrit à une étape ou une autre de son élaboration, et dont les commentaires, les suggestions et les conseils ont été décisifs. Je pense bien sûr à nouveau à François Rivenc, mais aussi à Denis Bonnay, qui ont tout deux lu intégralement des versions antérieures de ce travail, à Jacques Dubucs, à Leon Horsten, à Gabriel Sandu et à Philippe de Rouilhan, qui ont également bien voulu en lire des parties à une étape ou une autre. J'ai beaucoup appris d'eux, qu'ils reçoivent ici l'expression de ma gratitude sincère.

Je remercie Denis Bonnay non seulement pour ses suggestions innombrables, ses idées fécondes et ses relectures patientes, mais aussi pour son amical soutien dans

les bons et les mauvais moments qui aura parfois été si nécessaire à cette traversée du temps.

Je remercie enfin Adrienne, Gloria et Mathilde, qui n'auront cessé pendant tout ce temps de m'apprendre autre chose.

Il va sans dire que je suis seul à blâmer si je n'ai pas su faire bon usage des richesses qui m'ont été prodiguées sans compter.

Introduction

> Qu'est-ce que la vérité ? C'est avec cette vieille et fameuse question que l'on pensait pousser à bout les logiciens et que l'on cherchait à les prendre forcément en flagrant délit de verbiage ou à leur faire avouer leur ignorance, et par conséquent la vanité de tout leur art.
>
> Kant, Critique de la Raison Pure
> (Logique transcendentale,
> introduction, III)

0.1 Le projet : une épistémologie de la vérité

La vérité, dit-on, est un idéal de l'enquête scientifique. Mais quelle est la place de la notion même de vérité dans la science, dans notre système d'explication du monde ? S'agit-il d'un concept naïf qui n'appartient qu'à un mode relâché du discours, au discours préscientifique peut-être ? A-t-il un rôle à jouer dans le langage de la science, au même titre que les concepts scientifiques ordinaires de la biologie, de la physique ou des mathématiques ? Et si le concept de vérité a un rôle à jouer, quel est ce rôle ?

Ces questions relèvent à la fois d'une réflexion sur la nature de la vérité et d'une réflexion sur la méthodologie des sciences et de l'explication. D'un côté, s'il est possible de donner un sens scientifiquement acceptable à l'idée classique de la vérité comme *correspondance* entre ce qui est dit ou pensé, et le monde, alors il est permis de penser que la notion de vérité *a* un rôle central à jouer dans l'explication scientifique. En particulier, cette relation de correspondance ne pourrait-elle pas être la pierre de touche d'une explication du fondement objectif des relations sémantiques et intentionnelles qui existent entre un sujet et son environnement, ne serait-elle pas la clé de la naturalisation des explications ordinaires en termes d'attribution de

contenu sémantique à des pensées ou à des émissions verbales ? D'un autre côté, s'il n'est pas possible de donner un sens scientifiquement acceptable à l'idée classique de vérité comme correspondance, quel est le statut de cette idée ? Ne faut-il pas la déclarer suspecte et renvoyer son usage hors du discours de la science ? Confrontés aux difficultés persistantes à expliquer la façon dont la propriété de vérité pourrait être *réduite* à des propriétés « naturelles », et refusant d'autre part l'obscurité d'un discours sur la relation de correspondance qui se résoudrait à être métaphysique, un certain nombre de philosophes naturalistes ont proposé de revoir à la baisse les attentes attachées à une explication de la notion de vérité en réexaminant la signification et la portée épistémologique de la notion classique de vérité comme correspondance. Les thèses, dites « déflationnistes », qui ressortent de cet effort d'analyse, tendent à établir l'idée que cette notion classique de vérité, une fois correctement analysée, si elle est claire et légitime, ne peut pas jouer de rôle essentiel dans les explications.

Ce travail s'inscrit dans un effort philosophique pour préciser et évaluer les analyses déflationnistes de la notion de vérité. Dans la continuité des réflexions sur l'usage de la notion de vérité dans l'explication, la question qui m'intéressera est d'abord, et avant tout, celle de comprendre la nature des emplois de la notion de vérité permis par les lois *a priori* de la vérité - les lois que révèle une analyse conceptuelle adéquate de la notion, puisque cette analyse est possible par hypothèse -, que ce soit dans les sciences empiriques ou dans les domaines qui sont traditionnellement ceux de la connaissance *a priori*, et en particulier les mathématiques. Les *théories de la vérité* dans lesquelles nous cherchons à coucher les lois *a priori* gouvernant notre compréhension de la notion de vérité, quels que soient les détails exacts de ces théories, n'ont bien sûr pas vocation à expliquer à elles seules toutes les vérités mettant en jeu la notion de vérité - qu'il est *vrai* que la Pangée s'est déjà reformée sept fois depuis la naissance de la terre, ou qu'il est *faux* que le facteur passe toujours deux fois -, elles ne cherchent qu'à expliquer la notion de vérité elle-même. Mais une théorie de la vérité, comme n'importe quelle théorie, une théorie mathématique par exemple, peut être mise à contribution dans le contexte de théories plus larges dans lesquelles sont couchées d'autres vérités, d'autres propositions, d'autres lois relatives, non à la vérité, mais à tout autre domaine - de grands principes géologiques ou mathématiques - et, dans ce contexte, jouer un rôle dans ce que les théories globales, qui forment ensemble quelque chose comme notre système du monde, permettent de prouver, d'expliquer, de prédire. Notre objectif ici sera donc d'abord

d'essayer de comprendre ce que sont ces théories de la vérité et la nature du rôle qu'elles confèrent à la notion de vérité dans l'économie de nos moyens d'enquête et de justification.

La thèse que je soutiendrai est que ce rôle de la notion de vérité est un rôle purement *logique*. Néanmoins, ce en quoi consiste le caractère logique d'une notion, d'une part, et la signification même des thèses « déflationnistes » d'autre part, prêtent trop à controverse pour que l'on puisse espérer pouvoir établir le caractère logique de la vérité d'une façon qui ne prête pas elle-même à dispute et que la thèse de la logicité de la vérité suffise à établir le déflationnisme dans ses droits. Avant de chercher à défendre le caractère logique de la notion de vérité, c'est le sens même de l'idée qu'elle n'aurait pas de rôle épistémologique « substantiel » - pour reprendre un adjectif qui revient souvent dans le débat contemporain sur la notion de vérité - qu'il faut éclaircir. Si l'horizon de ce travail est la thèse du caractère logique du prédicat de vérité, son projet général est donc d'essayer de rendre raison des thèses déflationnistes en en précisant le sens et en les confrontant à une analyse détaillée des usages *a priori* de la notion de vérité.

Avant de présenter plus en détail le plan que je suivrai, il me faut dire quelques mots des limites principales de ce projet. La première limite touche à ce que j'ai annoncé être le point d'arrivée de ce travail, à savoir la thèse du caractère logique de la notion de vérité. Pour présenter un intérêt, cette thèse doit être accompagnée de cette autre, qu'il existe une distinction philosophiquement motivée entre les concepts logiques et les concepts non-logiques, et donc aussi entre les vérités logiques et les autres. Cette hypothèse est loin d'être anodine et, sous l'influence durable des travaux de W. O. Quine, c'est peut-être l'une des thèses philosophiques dont le rejet a fait l'objet du plus grand consensus dans la seconde partie du vingtième siècle. Depuis les « Deux dogmes de l'empirisme », la *doxa* philosophique n'est-elle pas acquise à l'idée qu'il n'y a, au mieux, en bonne méthodologie, que des *différences de degré* entre la façon dont les « vérités logiques » et les autres se rapportent au monde ? Pourtant, l'idée que les concepts logiques, et avec eux les lois logiques, ont un caractère spécial, est une idée qui peut se recommander de quelques philosophes et, je crois, de l'intuition commune. Je pourrais citer Kant, Frege, Russell, Wittgenstein, Carnap.[1] Dans la seconde partie du vingtième siècle même, quelques philosophes ont continué à chercher à comprendre la spécificité de la logique, malgré l'anathème quinien, en marge du courant dominant de la philosophie

[1] Je le ferai au chapitre 7.

« naturalisée » du langage et de la logique mais aussi au service de celle-ci. Je pense par exemple aux travaux de Michel Dummett et de Dag Prawitz. Plus récemment encore, l'idée que les lois de la logique auraient un caractère distinctif est revenu sur le devant de la scène philosophique dans le contexte plus large du renouveau des études épistémologiques sur la connaissance *a priori*.[2] Dans le chapitre que je consacrerai tout spécialement à la thèse du caractère logique de la vérité, j'en dirai davantage sur la façon dont je comprends la notion de logicité. Néanmoins, et c'est la première réserve que je voulais formuler relativement à la portée de ce travail, une discussion critique approfondie de ces questions demanderait un travail qu'il n'est pas possible de mener de bout en bout ici. Ce travail est d'abord une étude sur la notion de vérité, et non sur le propre de la logique, et ses intuitions philosophiques principales sont largement développées indépendamment de toute hypothèse sur la nature de la logicité. D'où l'attitude que j'adopterai : je rappellerai et développerai assez largement des arguments importants en faveur du bien-fondé de la distinction entre notions logiques et notions non-logiques, mais je m'en remettrai aussi à la bonne volonté du lecteur en ne me livrant pas à l'examen philosophique approfondi et exhaustif de ces arguments que nécessiterait une étude qui leur serait spécialement consacrée.

Puisque c'est une étude sur la notion de vérité que je présente ici, il me faut avant de commencer mentionner un second point important au titre des présupposés de ce travail. Il *ne* sera *pas* question dans cette thèse des paradoxes de la vérité, sinon de façon incidente.[3] La notion de vérité est notoirement impliquée dans des paradoxes célèbres, dont le moins connu n'est pas le paradoxe du Menteur. Dans une version, le Menteur dit : « Ce que je suis en train de dire n'est pas vrai ». Si c'est vrai, alors ce n'est pas vrai ; mais si ce n'est pas vrai, alors il semble que c'est vrai. Contradiction. Puisque les seuls principes mis en œuvre de façon essentielle dans la dérivation de la contradiction sont les lois de la logique et quelques platitudes qui semblent s'imposer à nous en vertu de la signification même du mot « vrai », il n'est pas certain, après tout, que ces platitudes soient aussi innocentes qu'elles en ont l'air. Le problème de la signification des paradoxes de la vérité et du meilleur moyen d'y remédier (s'il faut y remédier) est une question devenue centrale pour

[2] Je pense en particulier à Boghossian (1996) et Boghossian (2000), ou Wright (2001), Wright (2004*a*). En un mot, les objections quiniennes contre l'idée qu'il y aurait une différence de nature entre les vérités logiques et les autres sont retenues (les deux « parlent du monde »), mais l'idée qu'il existe une spécificité fondamentale de la *connaissance* logique est défendue.

[3] Au chapitre 2, dans le cadre d'une présentation du travail de Tarski, sur la vérité.

notre compréhension de la notion de vérité, mais le traiter demanderait un travail à part. En outre, le diagnostic du mal dont les paradoxes sont le symptôme étant pour le moins difficile à établir, il est de bonne méthodologie de commencer par chercher à comprendre le sens des attributions de vérité dans les cas non-problématiques, afin de disposer de raisonnements solides pour nous guider dans le diagnostic et la résolution des paradoxes dans les contextes où ils apparaissent. Un tel travail doit donc précéder, me semble-t-il, une réflexion sur les paradoxes. En tout état de cause, dans ce qui suit, je ferai donc provision d'une police d'usage du prédicat de vérité me mettant à l'abri de surprises désagréables et décide que, sauf mention contraire, lorsque je parlerai d'attribution de vérité à des énoncés, c'est d'énoncés d'un langage ou d'un fragment de langage ne contenant pas lui-même de prédicat de vérité qu'il s'agira. Comme nous le verrons au chapitre 2 consacré au travail d'A. Tarski, cette police, qu'elle soit ou non satisfaisante au titre de *résolution* des paradoxes, est néanmoins suffisante pour garantir notre sûreté.

Pour conclure cette brève présentation de mon projet et de ses limites, je voudrais revenir sur un dernier point relatif à la situation de ce travail dans le débat philosophique contemporain. Comme je l'ai dit, l'idée selon laquelle le prédicat de vérité est un prédicat logique, telle que nous l'avons esquissée, a à voir avec la doctrine philosophique connue sous le nom de « déflationnisme » et, au chapitre suivant, j'en dirai plus long sur ce que je crois être l'essence et l'intérêt du déflationnisme en matière de vérité. Je crois, je l'ai dit, que les thèses que je défends sont en harmonie avec certaines thèses déflationnistes, en tout cas avec l'esprit qui les anime. Mais d'un autre côté, je voudrais ajouter que les analyses et les thèses présentées ici ne sont nullement conditionnées à la plausibilité des thèses déflationnistes dans leur ensemble, en particulier à l'arrière-fond philosophique physicaliste dans lequel il a émergé comme une thèse contemporaine majeure. Quand bien même la plupart des affirmations ordinairement qualifiées de « déflationnistes » - et c'est un sujet où les querelles de mots ne manquent pas - se trouvaient réfutées, mettant en péril la plausibilité philosophique du déflationnisme « général » et le tableau post-positiviste de la science dont il émane, cela n'aurait que peu d'incidence sur le fond et la forme de ce travail : les hypothèses déflationnistes et naturalistes y sont peu nombreuses, et les arguments qui y sont présentés conserveraient, sous un autre nom, la valeur qu'elles ont dans leur présentation actuelle, sans qu'il soit besoin d'y changer quoi que ce soit d'essentiel.

0.2 Le plan

Cette thèse est composée de sept chapitres. J'ai multiplié les renvois en note entre les différents chapitres pour faciliter autant que possible une lecture non linéaire, mais leur progression a été conçue comme celle d'un argument. On peut distinguer quatre moments dans cet argument. Les chapitres 1 et 2, respectivement consacrés à une présentation des thèses déflationnistes générales et au travail d'Alfred Tarski sur la vérité, constituent, du point de vue de l'économie de l'ensemble, un moment préparatoire. On y présente des éléments conceptuels et historiques permettant l'intelligence des problèmes et des développements qui seront l'objet de la suite du travail. Le second moment de l'argument est constitué des chapitres 3 et 4. On y aborde l'étude proprement dite de la signification épistémologique des « preuves » mettant en jeu la notion de vérité, et en particulier des « preuves de cohérence » dont l'analyse, on le verra, se révélera être d'une importance particulière. L'enjeu est celui de la compréhension du statut de certains principes aléthiques jouant un rôle logique important dans ces preuves, et la clarification d'une distinction entre le rôle explicatif et ce que j'appellerai le « rôle expressif » de la vérité. Dans ces deux chapitres, c'est la pertinence d'une analyse du rôle de certains principes aléthiques dans les preuves de cohérence, faite dans les termes devenus classiques de la distinction entre conservativité et non-conservativité des théories, qui est remise en question. Une nouvelle proposition est faite dans la troisième partie pour tenter de rendre compte des intuitions déflationnistes. Il s'agit alors de montrer que, bien comprises, les preuves de cohérence par la vérité ne témoignent pas d'un usage explicatif de ces principes et, plus généralement, de présenter une analyse enrichie de la portée épistémologique des ressources conceptuelles mises en œuvre dans ces preuves. Avec ce travail d'analyse épistémologique des preuves par la vérité, la thèse du caractère logique du prédicat de vérité devient plausible et même naturelle. C'est cette thèse qui est alors explorée et défendue dans le quatrième et dernier moment de l'argument, au chapitre 7.

Le mouvement général étant annoncé, je présente maintenant brièvement le contenu de chaque chapitre. Au chapitre 1, *La vérité pour quoi faire ?*, je propose de reconstruire ce que je crois être les grandes thèses déflationnistes dans le contexte historico-philosohique qui les a vu naître. L'objet du chapitre n'est pas de faire une histoire du déflationnisme mais simplement de présenter quelques points conceptuels importants. Je soutiens en particulier que l'importance philosophique

du déflationnisme, et son sens historique, se comprennent surtout à la lumière du tournant naturaliste de la philosophie dans la seconde moitié du vingtième siècle et de l'interrogation qui s'en est suivi sur le statut *scientifique* des notions *sémantiques* et *intentionnelles*. En traçant ce dessein d'ensemble, je dégage les deux thèmes centraux de la réflexion qui suivra sur la vérité : d'une part le rôle du prédicat de vérité pour l'expression de certaines généralisations, et d'autre part l'absence de rôle explicatif de la notion de vérité. Le point de départ de notre travail apparaîtra alors plus clairement comme la tentative de comprendre comment ces deux thèmes peuvent se conjuguer.

Dans le second chapitre, *Tarski et la vérité*, je présente la théorie tarskienne de la vérité sous certains de ses aspects historiques, philosophiques et logiques, et discute quelques points que soulève son interprétation dans le contexte qui est celui de cette thèse. D'une part, puisque notre point focal est celui du rôle de la vérité dans les preuves, je prends soin de bien tenir à part le compte des ressources logiques et philosophiques que Tarski mobilise pour accomplir son programme de définition, explicite ou axiomatique, de la notion de vérité, et des ressources auxquelles il a recours pour accomplir un programme plus large d'application de la théorie de la vérité aux questions fondationnelles en mathématique. D'autre part, je soutiens que l'analyse tarskienne de la notion de vérité est compatible avec les thèses déflationnistes. Plus généralement ce chapitre me permet également de mettre en place l'appareil logico-conceptuel et terminologique dont j'aurai besoin dans les chapitres suivants en présentant notamment les notions d'extensions aléthiques minimale et tarskienne d'une théorie et quelques-unes de leurs propriétés logiques.

Au chapitre 3, *Vérité et conservativité*, la signification du rôle de la vérité dans les preuves de cohérence où elle figure commence à être discutée. Je présente une reconstruction rationnelle, puis une critique, de l'argument dit de la « conservativité » opposé au déflationnisme. Je distingue une Thèse de la Conservativité et une Thèse de la Réflexion, et réfute la thèse de la conservativité qui identifie la non-conservativité d'une théorie sur une autre et le pouvoir explicatif des notions spécifiques de la première relativement aux notions communes aux deux théories. J'introduis une première formulation de ma thèse principale à la faveur de la discussion : le rôle des principes aléthiques est plus adéquatement compris comme moyen d'expliciter ce qui est implicite dans notre acceptation de certaines théories que comme introduisant de nouveaux principes d'explication.

Au chapitre 4, *Réflexion et philosophie des mathématiques*, j'examine quelques

travaux de philosophie des mathématiques en relation avec le statut de ce l'on appelle les « principes de réflexion » aléthique et le rôle du concept de vérité dans les preuves, en particulier des preuves de cohérence. À certains égards, il s'agit d'un chapitre transitoire entre le travail critique mené au chapitre 3 et les thèses positives qui seront défendues au chapitre 6. D'un côté, le rappel de quelques réflexions épistémologiques classiques et fondamentales issues de l'analyse des phénomènes d'incomplétude vient en effet étayer le bien-fondé de la critique d'une analyse épistémologique trop rapide des preuves de cohérence en termes de conservativité. D'un autre côté, ces réflexions permettent de préparer du même coup le chemin pour notre propre analyse ultérieure.

Dans *Réflexion et rationalité*, au chapitre 5, je laisse momentanément la notion de vérité de côté mais seulement pour mieux comprendre, par contraste, son utilité et sa singularité. La question abordée, qui s'inscrit dans un effort de réflexion plus large pour tenter de comprendre les liens entre rationalité et certains principes de réflexion en un sens large, est de savoir si un sujet qui est en position de construire une preuve de cohérence en faisant appel à des ressources conceptuelles aléthiques, pourrait en construire une justification par d'autres voies en se dispensant complètement de ces ressources. J'explore cette possibilité en discutant de certains principes de réflexion « épistémiques » et des justifications qu'un agent rationnel possède pour les accepter. J'introduis une notion d'*acceptation* d'une théorie par un agent rationnel, et j'examine la situation d'un sujet acceptant une théorie A en demandant s'il est rationnellement permis à cet agent d'accepter des principes réflexifs du type « *A est acceptable* », « *les règles d'inférence préservent l'acceptabilité* », etc. D'une part, il semble qu'il y ait quelque chose d'*incorrect*, en termes d'action rationnelle, dans le comportement d'un agent acceptant une théorie sans accepter, à la réflexion, ces principes supplémentaires, alors même que, pourtant, le fait de l'acceptation de la théorie ne fonctionne pas comme un *indice* de ce que la théorie est accept*able*. Mais d'autre part, l'ajout de tels principes à une théorie A engendre typiquement une théorie logiquement plus forte que A, dans laquelle il est possible de prouver la cohérence de la théorie A. J'essaie de montrer que, pour *une certaine* notion d'acceptation et de justification, si un agent accepte une théorie, il peut justifier son acceptation de la cohérence de la théorie sans faire appel au concept de vérité.

Au chapitre 6, *Conséquence Réflexive*, je développe ma réponse aux différents problèmes ouverts dans les chapitres précédents par la question de la portée épistémologique des preuves par la vérité. Je présente une distinction entre principes

explicatifs et principes expressifs et je montre que les principes aléthiques auxquels il est fait appel pour mener à bien des preuves comme la preuve de la cohérence appartiennent à la seconde catégorie. La notion importante ici est la notion épistémologique de « conséquence réflexive » d'une théorie. Je défends l'idée, fondée sur une réflexion sur les limites de l'exercice du jugement chez le sujet fini, que les principes de réflexion aléthique sur une théorie donnée - « Tous les théorèmes de la théorie A sont vrais » - sont des « conséquences réflexives » de la théorie, où p est une conséquence réflexive d'un corps de propositions B si et seulement si toute justification qu'il est possible de posséder pour B justifie également p. L'idée générale de « conséquence réflexive » est un thème récurrent de certains écrits logiques et philosophiques apparus dans le sillage des théorèmes d'incomplétude de Gödel pour tenter d'en tirer toutes les leçons épistémologique. Mais l'explication simple que je propose du fait que certains *principes de réflexion* sont des conséquences réflexives de leur théorie de base est nouvelle. Je tire alors les conséquences de cette analyse dans plusieurs directions pour rendre compte de la spécificité des emplois de la notion de vérité.

Au chapitre 7, *Vérité et Logicité*, je propose enfin d'examiner plus précisément, et de défendre, la thèse selon laquelle le concept de vérité s'apparente à un concept logique. Mais qu'est-ce qu'un concept logique, et comment la notion de vérité pourrait-elle être logique ? Je présente deux voies possibles permettant d'élaborer une réponse construite à ces questions : une voie sémantique et une voie « épistémique ». La première caractérise la logique par ses objets (ou son absence d'objet), la seconde par le caractère spécial des justifications que nous avons pour les vérités ou les inférences logiques. Je développe alors l'idée du caractère logique de la notion de vérité selon ces deux axes. D'abord en m'appuyant sur une analyse de la « signification inférentielle » des règles qui gouvernent notre compréhension du prédicat de vérité, telles que règles auront été identifiée au chapitre précédent, dans l'esprit de la sémantique preuve-théorique développée par Dummett et Prawitz. Puis sous l'angle sémantique, motivé par une certaine idée philosophique de la nature des expressions logiques comme expressions qui n'ont pas de « contenu » propre mais dont le rôle est à chercher dans l'articulation du discours, je propose, d'une part, un cadre d'analyse du contenu (extensionnel) des attributions de vérité qui permette de rendre compte des intuitions déflationnistes concernant leurs emplois et, d'autre part, un critère d'évaluation du caractère logique de ces emplois.

Première partie

Préliminaires

Introduction

L'objet de cette partie est de présenter les deux grandes entreprises théoriques concernant la notion de vérité qui sont à l'arrière-plan de ce travail. La première est l'entreprise « déflationniste ». Le déflationnisme est ce que l'on pourrait appeler une *métathéorie* de la vérité. C'est un discours sur les usages et le rôle théorique de la notion de vérité. Une de ses thèses négatives centrales est que la notion de vérité n'est pas un outil explicatif. L'objet du premier chapitre est de présenter les thèses du déflationnisme, en replaçant ce dernier dans son contexte historique et philosophique et en en clarifiant le sens général.

A côté de la question de savoir quelle est l'utilité ou le rôle du prédicat de vérité, il y a le problème d'élaborer une théorie de la vérité, c'est-à-dire de donner un certain nombre de principes permettant d'expliquer la notion de vérité elle-même, éventuellement de la définir. Dans le second chapitre, on s'intéresse à la question de savoir s'il est possible de définir ou de donner une théorie adéquate de la vérité. Ici nous suivrons les traces de Tarski, puisque c'est à lui que revient le mérite d'avoir dégagé les éléments de réponse à cette question. Je présenterai les aspects essentiels de son travail en étant attentif aux ressources théoriques mobilisées dans une explication de la notion de vérité.

L'horizon de ces deux chapitres en partie historiques, est la confrontation entre, d'un côté, une thèse relative au rôle que joue la notion de vérité *dans* les explications (la thèse déflationniste), et de l'autre ce que révèle une analyse *a priori de* la notion de vérité elle-même (Tarski). Si l'exploration de ces rapports est le cœur de ce travail, cette première partie en constitue donc le socle. L'exploration détaillée elle-même ne commencera véritablement que dans les parties qui suivent.

Chapitre 1

La vérité pour quoi faire ?

1.1 Introduction

Qu'est-ce que le « déflationnisme » en matière de vérité ? Sans doute vaut-il mieux parler de déflationnisme*s* ici, tant la variété des thèses que recouvre cette appellation semble devoir forcer le pluriel. Le terme de « déflationnisme » est un barbarisme, mais il est commode et il est désormais entré dans l'usage philosophique francophone. Je l'utiliserai donc librement, ainsi que ses dérivés. Le terme dit aussi assez clairement qu'il s'agit de « dégonfler », de revoir à la baisse, certaines attentes philosophiques ordinairement attachées à une analyse du concept de vérité.

L'analyse déflationniste de la notion de vérité s'inscrit à certains égards dans la tradition philosophique classique où la vérité est comprise comme *correspondance* entre les objets qui sont les véhicules, ou les porteurs de la vérité (ce dont la vérité se dit), et le monde, telle que cette conception s'exprime par exemple dans la définition mémorable d'Aristote :

> Dire de ce qui est, que ce n'est pas, et de ce qui n'est pas, que c'est, est faux, tandis que dire de ce qui est que c'est, et de ce qui n'est pas, que ce n'est pas, est vrai. (*Métaphysique*, Γ)

En particulier, un déflationniste reconnaît qu'en vertu de la signification de l'énoncé ou de la proposition qui est mentionné(e) dans le membre gauche de l'équivalence suivante, et de la signification de la notion de vérité,

L'énoncé « la neige est blanche » (ou : la proposition que la neige est blanche est vraie si et seulement si la neige est blanche.

Et de même de toutes les équivalences formées sur ce modèle, que l'énoncé (la proposition) mentionné(e) à gauche et utilisé (exprimée) à droite soit vrai(e) ou faux (fausse), et quel que soit le domaine de discours, moral, physique, ou autre, auquel il (elle) appartient.

Il est relativement clair que la reconnaissance de la vérité de ces équivalences engage une compréhension de la vérité comme « correspondance ». La correspondance est pour ainsi dire là, décrite sous nos yeux : chaque équivalence explique en effet quelle condition doit être réalisée « dans le monde » pour que le porteur de vérité désigné à gauche ait la propriété d'être vrai. Et cette condition n'est autre que celle présentée par Aristote : pour que ce que dise l'énoncé (ou la proposition) soit vrai, il faut et il suffit que ce qu'il dise soit.[1]

Le déflationnisme n'est donc pas un nihilisme en matière de vérité. Il ne s'agit de mettre en cause ni la légitimité de la notion de vérité, ni les caractéristiques que la tradition classique lui a reconnu. Le déflationniste n'a pas non plus de difficulté avec l'idée que la vérité puisse être un but légitime et central de notre enquête sur le monde. Quelles attentes le déflationniste a-t-il donc à cœur de décevoir ? En premier lieu celle qu'un discours éclairant *sur* cette relation de correspondance soit possible, que cette relation de correspondance puisse être elle-même un authentique objet théorique. C'est cette interprétation du déflationnisme que je voudrais expliquer et développer dans ce chapitre.

Dans une version forte des exigences relatives à la méthodologie de l'explication, celle du naturalisme post-positiviste qui est le terreau du déflationnisme contemporain, c'est le discours scientifique qui est ultimement le lieu où se lit ce qui compte comme explication. Dans ce contexte, la relation de correspondance, si elle pouvait être expliquée, devrait être expliquée comme une relation entre des

[1] Le lecteur remarquera que la reconnaissance de la vérité de ces équivalences exclut ce que l'on regarde ordinairement comme les principales rivales de la conception classique de la vérité. Une conception pragmatique en vertu de laquelle *est vrai ce qu'il est utile de croire*, est incompatible avec les principes précédents, sauf à admettre des principes du type :

> Il est utile de croire que la neige est blanche si et seulement si la neige est blanche.

ce qui revient en pratique à ôter tout intérêt philosophique à la thèse pragmatique. De la même manière, on se convaincra aisément qu'une conception « cohérentiste » de la vérité, en vertu de laquelle *est vrai ce qu'il est cohérent de croire*, ou une conception « épistémique » de la vérité, en vertu de laquelle *est vrai ce que nous sommes justifiés à croire*, ne sont pas compatibles avec la reconnaissance de la vérité des équivalences présentées plus haut. Sur ces derniers points on pourra consulter Lewis (2001). Sur l'importance du rôle des équivalences du type de celles mentionnées ici dans le corps du texte pour une analyse de la notion de vérité, voir le chapitre suivant sur Tarski.

objets identifiables en termes scientifiquement acceptables, par exemple entre un sujet et son environnement, et la question se poserait alors naturellement de savoir si cette relation ne serait pas à son tour le socle explicatif adéquat pour élucider la nature de l'ensemble des propriétés intentionnelles et sémantiques, c'est-à-dire toutes ces propriétés que nous attribuons à des états mentaux ou des items linguistiques et en vertu desquelles ces derniers sont compris comme *représentant* des états de choses, comme *étant à propos* d'une certaine réalité, comme *correspondant* d'une manière ou d'une autre à certains aspects de notre environnement. Dans cette perspective méthodologique, le déflationnisme aléthique apparaît donc à la jonction d'une réflexion sur la notion traditionnelle de vérité comme correspondance et d'une réflexion sur la place dans le discours scientifique des propriétés semblant mettre en jeu cette notion de correspondance, les propriétés sémantiques (vérité, référence, signification) et intentionnelles (des attitudes relatives à des contenus propositionnels). On peut alors formuler le *projet* théorique du déflationnisme de la façon suivante : articuler une thèse cohérente à propos du prédicat de vérité qui permette de désamorcer certaines attentes illégitimes susceptibles de s'épanouir à partir de l'idée de la théorie de la vérité comme correspondance, tout en restant fidèle à l'intuition classique.

Dans l'interprétation forte que nous venons d'en donner, le déflationnisme ne se limite pas à une thèse sur le prédicat de vérité, mais porte plus généralement sur les rapports du langage (ou de la pensée) à l'environnement, autrement dit sur ce *à quoi* le langage se rapporte et comment. Mais toute réflexion sur la notion de vérité n'a pas partie liée, en droit, aux programmes de naturalisation dans leurs différentes versions. On peut vouloir expliquer la notion de vérité pour elle-même, pour ainsi dire, dans une certaine tradition du travail philosophique compris comme analyse conceptuelle, et en *se donnant*, par exemple, la notion de *proposition*, c'est-à-dire d'un objet qui d'une manière ou d'une autre, est *à propos* du monde, *représente* des états de choses, est exprimé par des énoncés, est l'objet de nos états intentionnels, est individué par ses conditions de vérité. Or il est possible qu'une réflexion sur la notion de vérité conduite dans ce second cadre théorique rejoigne la première sur certains points : on peut penser que la « meilleure » analyse conceptuelle de la notion de vérité est une analyse « déflationniste », en un sens ou un autre, sans pour autant souscrire à un déflationnisme généralisé aux notions sémantiques ou à une thèse méthodologique relative à l'étude des rapports de la pensée au monde. On peut n'avoir aucune objection à admettre une ontologie d'entités intensionnelles

comme les propositions, prises comme les porteurs ou les véhicules de la vérité, tout en pensant que la notion de vérité, *étant donné la notion de proposition*, est passible d'une analyse triviale qui rende compte au plus près de l'idée classique de vérité comme correspondance. Il faut donc distinguer entre ce que j'appellerai, faute de mieux, un déflationnisme aléthique « local » et un déflationnisme sémantique « global », ou « général ». Si l'on tient que le véritable enjeu d'une analyse de la notion de vérité est la compréhension des relations du langage ou de la pensée au monde, on sera sans doute porté à parler respectivement de déflationnisme *faible* et de déflationnisme *fort*.[2]

Dans la suite de ce chapitre, je me propose de reconstruire plus en détail une position déflationniste globale forte, en suivant un chemin historico-philosophique depuis les *étonnements fondateurs* du déflationnisme jusqu'à cette position articulée générale, et représentative du déflationnisme contemporain, qui celle de Hartry Field[3]. J'avertis le lecteur qu'il ne s'agit pas ici de faire une histoire du déflationnisme,[4] mais seulement d'identifer chez quelques auteurs choisis quelques-unes des thèses qui se sont avérées importantes pour l'élaboration des conceptions contemporaines. En fait de renoncement à l'exhaustivité, je ne m'arrêterai ici qu'à

[2]Je note tout de même, sans m'y attarder, qu'il existe d'autres motivations pour le déflationnisme aléthique que les programmes naturalistes et de déflation ontologique. Dans une certaine tradition pragmatique, par exemple, qui passe par le second Wittgenstein et dont le représentant contemporain le plus éminent est peut-être Robert Brandom, les attributions de signification (et d'attitudes propositionnelles) ont un caractère irréductiblement normatif, et comprendre une norme c'est comprendre quelque chose comme les règles du jeu gouvernant une certaine pratique sociale. Si les choses signifient ce qu'elles signifient, ce n'est pas en vertu de ce qu'elles « représentent », mais en vertu d'un certain rôle qui leur est conféré dans cette pratique. Une explication de ce que c'est pour un sujet d'avoir une croyance ayant tel ou tel contenu doit rendre compte non du caractère représentationnel de ce contenu, mais de la portée normative de l'état intentionnel en question : les conditions sous lesquelles il est permis d'attribuer un état mental donné, et les inférences ou les actions qu'il est permis de faire dans cet état mental. (Pour ce qui est du contenu d'un énoncé, c'est la portée normative de l'acte d'assertion qui sera typiquement prise en compte, c'est-à-dire, ce à quoi une telle assertion *engage* le locuteur. Voir par exemple Brandom (1994) p.143 *sqq* pour un développement). A nouveau il s'agit donc de nier que les notions sémantiques comme celles de vérité, référence, conditions de vérité, associées à l'idée d'un contenu représentationnel, sont des notions centrales d'un point de vue descriptif et explicatif, tant pour comprendre la notion de signification que pour comprendre les états intentionnels, et plus généralement les attributions de propriétés sémantiques. Dans ce contexte philosophique, les attributions de vérité à des énoncés ou des états mentaux recevront typiquement une interprétation déflationniste dans laquelle l'intuition de la correspondance n'a aucun rôle théorique.

[3]Les contributions les plus importantes de Hartry Field sur ce sujet sont réunies dans Field (2001).

[4]Histoire qui, à ma connaissance, reste à faire, mais qui promet d'être fastidieuse par ses redondances.

quatre auteurs : G. Frege, F. Ramsey, W. Quine, avant, donc, d'en venir à H. Field. En outre, je n'affirme pas que ces auteurs soient déflationnistes, même si certains le sont ; tous ont néanmoins nourri par leurs remarques les réflexions déflationnistes sur la vérité, et c'est à ce titre que j'en parle ici. Un des grands absents de cette liste est A. Tarski, dont l'influence sur la philosophie de la vérité de Quine et Field est décisive : le chapitre suivant lui sera entièrement consacré.

1.2 La transparence des usages prédicatif de la vérité : Frege

On sera peut-être surpris de voir figurer ici le nom de Gottlob Frege. Frege ne fut-il pas le premier à introduire l'idée que les conditions de vérité d'une assertion permettent d'élucider la notion de *sens* ? C'est vrai, et pourtant les quelques remarques publiées de Frege sur les usages prédicatifs de la vérité constituent bien quelque chose comme les premiers jalons d'une réflexion déflationniste.

La transparence des attributions de vérité. Ces remarques se trouvent principalement dans deux textes : « Sens et dénotation » (1892) et au début des *Recherches logiques* (1918). Le passage de « Sens et dénotation » qui m'intéresse directement apparaît dans le cours du développement de la thèse selon laquelle le Vrai est un objet, la référence de certaines propositions. À un certain point de son argument, Frege entreprend de procéder négativement, pour ainsi dire, par l'examen des emplois prédicatifs de « vrai » :

> On pourrait être tenté de voir dans le rapport de la pensée au vrai, non pas celui du sens à la référence, mais celui du sujet au prédicat. On pourrait dire à cet effet « la pensée que 5 est un nombre premier est vraie ». À regarder la chose de plus près, il apparaît qu'on a en fait rien dit de plus que dans la proposition « 5 est un nombre premier ». Dans les deux cas, l'affirmation de la vérité réside dans la forme de la proposition affirmative. Par suite, pour peu que l'affirmation n'ait pas sa force habituelle, par exemple dans la bouche d'un acteur sur scène, la proposition « la pensée que 5 est un nombre premier est vraie » ne contient jamais qu'une pensée, la même que le simple énoncé « 5 est un nombre premier ». Il faut donc admettre que le rapport de la pensée au vrai ne peut être comparé à celui du sujet au prédicat. (Frege, « Sens et

dénotation », in FREGE (1971*a*), p.110-111. Traduction modifiée.)

On le voit, il y a déjà dans ce bref passage deux thèses qui seront essentielles au développement du déflationnisme. La première est la thèse de la transparence de la signification de « vrai » : la pensée exprimée par un énoncé et celle exprimée par l'énoncé qui lui attribue la vérité sont identiques.[5] La seconde thèse tient dans cette idée que, malgré leur grammaire superficiellement prédicative, les attributions de vérité ne sont pas des attributions de propriétés.[6]

Ces deux thèses se trouvent confirmées dans les *Recherches logiques*, plus tardives, plus précisément dans la recherche intitulée « La pensée » (1918). Les lignes qui ouvrent cette recherche comptent peut-être parmi les plus connues de Frege :

> De même que le terme « beau » renvoie à l'esthétique et le terme « bon » à l'éthique, le terme « vrai » renvoie à la logique. Certes, toutes les sciences ont la vérité pour but, mais la logique s'en occupe d'une toute autre manière encore [...]. Découvrir des vérités est la tâche de toutes les sciences, mais c'est à la logique qu'il appartient de découvrir les lois de l'être vrai. (FREGE (1971*b*), p.170)

Il serait pourtant erroné d'interpréter l'expression « découvrir les lois de l'être vrai » comme un renoncement à l'analyse antérieure de la notion de vérité. C'est ce qu'indique clairement la suite de l'exposé et en particulier les remarques de Frege dont l'objet n'est plus la logique, mais la vérité elle-même. En effet, lorsque Frege, un peu plus loin, entreprend de « dessiner grossièrement les contours de ce qu'[il] entendr[a] par vrai dans la suite du texte » (*ibid.* p.171), nous retrouvons la tonalité négative des remarques précédentes.[7]

[5] Remarque : la thèse de la « transparence » de la signification de « vrai » *ne* signifie *pas* que « vrai » *n*'a *pas* de signification. Si le mot « vrai » n'avait pas de signification, il faudrait sans doute dire que les énoncés qui attribuent la vérité à d'autres énoncés sont eux-même dénués de signification, ce qui est absurde. M'inspirant de la terminologie mise en place par D. Kaplan pour discuter de la sémantique des démonstratifs, je dirais volontiers que la thèse de la transparence de la vérité n'est pas la thèse que le prédicat de vérité serait dénué de *signification linguistique*, mais seulement qu'il n'a pas de *contenu*. *Cf.* KAPLAN (1989). Pour une tentative de donner un sens précis à une idée de ce genre en vue d'analyser les prédications de vérité, voir les travaux de Dorothy Grover dans GROVER (1992).

[6] Pour Frege, en 1890, si la vérité n'est pas un attribut, ce doit « donc » être un objet, mais c'est une conséquence que la postérité ne retiendra pas. Il en reste néanmoins une trace dans la sémantique du calcul propositionnel héritée de l'Ecole Polonaise. À propos de l'influence de Frege sur les recherches logiques en Pologne au début du vingtième siècle, et sur ce point en particulier, on pourra consulter WOLENSKI (2009), §3.3.

[7] Du reste, concernant la première phrase de « La pensée » que j'ai citée, son interprétation doit se faire en gardant à l'esprit certains développements des *Ecrits Posthumes* de Frege. Le passage

Ainsi, un peu après avoir précisé qu'il appelle *pensée* « ce dont on peut demander s'il est vrai ou faux » [8], Frege continue :

> Au demeurant, il y a tout lieu de penser que nous ne pouvons pas reconnaître qu'une chose a une certaine propriété, sans en même temps estimer vraie la pensée que cette chose a cette propriété. Ainsi à toute propriété d'une chose est liée une propriété d'une pensée, à savoir celle d'être vraie. Il vaut aussi de remarquer que la proposition « je sens une odeur de violette » a même contenu que la proposition « il est vrai que je sens une odeur de violette ». Il semblerait que rien n'est ajouté à la pensée quand je lui attribue la propriété d'être vraie. Et pourtant n'est-ce pas un progrès d'importance quand, après une longue hésitation et des recherches pénibles, le savant peut dire enfin « ce que je présumais est vrai » ? La dénotation du mot « vrai » semble unique en son genre. Serait-ce que nous ayons affaire à quelque chose qui ne peut nullement être appelé propriété dans le sens usuel ? (Frege, « La pensée », *in* Frege (1971*a*), p.174)

On retrouve dans la première partie de ce passage de 1918 le thème de 1892, celui de l'identité de la pensée exprimée par un énoncé et de la pensée exprimée par l'énoncé qui lui attribue la vérité. Et une conclusion très proche sur le fond de celle que je soulignais tout à l'heure, quoique plus prudente, peut-être même hésitante : il ne s'agit plus d'affirmer que la vérité n'est pas une propriété, mais seulement qu'elle semble n'être pas une propriété « dans le sens usuel ». Un certain trouble semble naître de ce que, d'un côté, il y a identité apparente de sens de « p »et de « Vrai(p) », mais que, d'un autre, c'est « un progrès d'importance quand, après une longue hésitation et des recherches pénibles, le savant peut dire enfin « ce que

suivant de *Mes intuitions logiques fondamentales* (1915) est particulièrement clair :

> Ainsi le mot « vrai » semble rendre possible l'impossible, notamment en faisant apparaître ce qui apparaît comme la force assertive comme étant une contribution à la pensée. Et cette tentative, quoiqu'elle n'atteigne pas son but, ou bien plutôt justement parce qu'elle ne l'atteint pas, indique la nature particulière de la logique, et celle-ci apparaît dès lors comme essentiellement distincte de l'esthétique et de l'éthique. En effet, le mot « beau » indique bien réellement l'essence de l'esthétique, comme le mot « bien » celle de l'éthique, tandis que le mot « vrai » ne fait en réalité qu'une tentative malheureuse pour indiquer celle de la logique, dans la mesure où ce qui est réellement en question ne se rapporte pas du tout au mot « vrai » mais à la force assertive avec laquelle une phrase est prononcée. (Frege (1999), p.298.)

Nous revenons sur la relation entre force assertive et vérité un peu plus bas.

[8] *Op. cit.* p.173.

je présumais est vrai » ». Y a-t-il un véritable problème ici, la seconde remarque rendrait elle insoutenable la première ? Comment faut-il comprendre la réserve que semble susciter chez Frege la remarque à propos du savant ? Pour comprendre ce qui est en jeu, il faut revenir aux remarques de Frege sur l'assertion.

Vérité et assertion. Frege avait pris soin, dès la *Begriffsschrift*, de distinguer entre la simple présentation d'un contenu propositionnel et l'affirmation ou l'assertion de ce contenu propositionnel.[9] Un contenu propositionnel peut apparaître dans une phrase déclarative en étant affirmé, comme dans

Le chat est malade.

ou sans l'être, comme ce même contenu tel qu'il figure dans la première partie de la phrase suivante :

Si le chat est malade, j'irai chez le vétérinaire.

Et de même pour tout contenu. C'est ce que l'on appelle aujourd'hui *le point de Frege*. Il semble exister quelque chose comme une tentation linguistique qui nous porte régulièrement à oublier ce point et l'on entend parfois que la force de l'assertion est indiquée par la prédication de la vérité à un contenu propositionnel. Pourtant il n'en est rien : je peux douter que la neige soit blanche ou douter identiquement qu'il soit vrai que la neige est blanche, et des attributions de vérité peuvent apparaître dans des énoncés composés sans être affirmées, comme dans

S'il est vrai que le chat est malade, alors j'irai chez le vétérinaire.

Le point de Frege est justement qu'aucune expression du langage ordinaire ne peut véhiculer par son contenu une indication de force assertorique : quel que soit le contenu exprimé par un énoncé, celui-ci peut encore être affirmé, supposé, mis en doute etc.[10,11]

[9] Le jugement étant, pour Frege, la contrepartie « mentale » de l'assertion.

[10] Voir Geach (1965) pour une défense généralisée de ce point. Smith (2000) rappelle opportunément que Frege a tenu jusqu'à la fin de sa vie ce point comme l'une de ses découvertes philosophiques les plus importantes. Ainsi, dans une brève note posthume d'une dizaine de lignes intitulée « Que puis-je considérer comme le résultat de mon travail » datée de 1906, Frege en consacre deux à ce point. Il écrit

> ... cependant j'aurais dû à vrai dire mentionner auparavant le signe d'assertion. La dissociation de la force assertorique et du prédicat... (Frege (1999) p.219)

[11] Frege lui-même semble bien avoir cru un moment, au début de l'*Idéographie* (Frege 2000, §§

Cela étant précisé, revenons à présent à la remarque à propos du savant citée plus haut, et au trouble qu'elle semble jeter, aux yeux de Frege, sur ses conclusions antérieures. Quel est le « progrès » accompli par le savant ? C'est simplement le suivant : il est maintenant en position d'*affirmer* ce qu'il ne pouvait auparavant que *présumer*. Les conditions sous lesquelles il est permis de présumer étant bien moins strictes que celles sous lesquelles il est permis d'affirmer (affirmer demande des *preuves* ou des justifications, et les découvrir est le travail du savant !), on voit clairement en quoi le savant qui peut « dire 'ce que je présumais est vrai' » a accompli un progrès, pour peu que nous comprenions que ce dont nous parle Frege ici, et il faut que nous corrigions de nous-même, c'est en fait d'un savant qui peut non seulement « dire », mais *affirmer* que ce qu'il présumait est vrai (sinon où serait le progrès ?). Or, maintenant, en quoi le progrès que constitue la découverte de preuves est-il de nature à justifier une réserve relativement à l'affirmation que la vérité n'est pas une propriété ? Sous une certaine interprétation moderne[12], le cas de la phrase prononcée par le savant constitue un joli exemple d'usage anaphorique du prédicat de vérité dans un contexte où plusieurs attitudes d'un sujet relativement à un certain contenu sont en jeu, contextes qui sont typiquement propices aux illusions sur la signification de « vrai ». Ici un contenu de pensée identifié comme objet (passé) d'une certaine attitude, « ce que je *présumais* », est repris énonciativement (ré-énoncé, pour ainsi dire) de façon indirecte par un mécanisme de type anaphorique grâce au prédicat de vérité, mais en étant cette fois *asserté* ou affirmé. Le caractère assertorique de cette énonciation indique que l'attitude du savant vis-à-vis de ce contenu a changé : il *juge* vrai, désormais, ce qu'il ne faisait que présumer vrai auparavant.

Ce genre d'analyse de « vrai », qui donne au prédicat une fonction linguistique sans lui attribuer de contenu à proprement parler, me semble appelée (anachroniquement) par différentes remarques de Frege. Je pense à certaines remarques déjà citées, ou encore à cet autre passage de « Mes intuitions logiques fondamentales » :

> On reconnaît ainsi que l'assertion ne réside pas dans le mot « vrai », mais dans la forme assertive avec laquelle la phrase est prononcée. On

2-3), que la *barre de jugement*, ce signe qu'il avait utilisé dans son symbolisme pour marquer la force de l'assertion, pouvait se concevoir comme un *prédicat* général, tombant ainsi en somme dans l'erreur qu'il avait décelée.

[12] Je pense ici à l'interprétation prophrastique des emplois des locutions aléthiques, telle qu'elle est défendue en particulier par Dorothy Grover depuis Grover, Camp et Belnap (1975). Voir les articles réunis dans Grover (1992). L'analyse que donne Ramsey de la vérité en termes de prophrases est discutée dans la section suivante de ce chapitre.

> pourrait dès lors être d'avis que le mot « vrai » ne possède simplement aucun sens. Mais la phrase dans laquelle le mot « vrai » est prédicat ne possèderait alors aucun sens. On peut seulement dire : le mot « vrai » a un sens qui n'apporte rien au sens de la phrase entière dont il est le prédicat. (Frege (1999), p. 297-298).

L'idée que le prédicat de vérité a un sens mais que ce sens ne contribue pas au sens des phrases qui ont la forme de prédication de vérité a quelque chose de troublant. Pourtant la conclusion de Frege ne s'impose plus, dès lors que l'on a renoncé à la prémisse selon laquelle, si « vrai » n'a pas de sens, alors les phrases où il figure n'ont pas de sens non plus. Et un moyen de renoncer à cette prémisse sans absurdité est de penser la sémantique des locutions aléthiques sur le modèle de la sémantique des termes indexicaux, en distinguant signification linguistique et contenu d'une expression.[13] C'est une intuition de ce genre qui anime aujourd'hui le travail de Dorothy Grover et que j'ai mise en œuvre pour rendre compte du « progrès accompli » dont témoignait l'affirmation que Frege prêtait tout à l'heure à un savant de circonstance.

Reste à savoir si ce que Frege attendait d'une analyse de la notion de vérité aurait jamais pu se résoudre entièrement en ce qui n'est qu'une analyse linguistique de la fonction du prédicat « est vrai ». On peut penser que non, le problème philosophique de la nature de la vérité demeurant derrière la résolution linguistique du problème, ou au contraire penser que les réserves que finit par émettre Frege sur ce qui ressemble à une théorie « redondance » de la vérité, théorie qu'il semble par ailleurs disposé à promouvoir, ne sont pas bien fondées. En l'absence d'indices textuels déterminants dans les écrits tardifs de Frege, il me semble qu'une certaine indécision demeure au bout du compte dans l'intention philosophique de Frege relativement à ce que nous appellerions aujourd'hui le problème de la nature, ou de l'absence de nature, de la vérité.

Conclusion. Pour conclure, quoique nul ne fût sans doute plus éloigné que Frege des préoccupations qui motivent, je l'ai dit, la plupart des thèses déflationnistes contemporaines, il reste que ses remarques sur la vérité constituent une première articulation de deux thèses qui sont au cœur de toute réflexion déflationniste sur la vérité :

– La thèse de la transparence de la signification de « vrai » : la pensée exprimée par un énoncé et celle exprimée par l'énoncé qui lui attribue la vérité sont

[13] Voir la note 7 dans ce chapitre.

identiques.

– « vrai » n'est pas un authentique prédicat, ou la vérité n'est pas une propriété « dans le sens usuel ».

De la première thèse on peut dire qu'elle traverse toutes les versions des doctrines déflationnistes jusqu'à aujourd'hui. Quant à la seconde, si l'idée positive que la vérité est la *dénotation* des propositions n'a pas été suivie[14], l'idée toute négative que la vérité n'est pas une propriété ou que « vrai » n'est pas un « authentique » prédicat a eu une certaine postérité. Je devrais dire « les idées », car « n'être pas un prédicat », comme « n'être pas une propriété », ont fini par recouvrir des thèses assez différentes. La thèse, relativement faible, que la vérité n'est pas une propriété ordinaire, quoiqu'il puisse s'agir d'une propriété de quelque sorte, est une constante du déflationnisme, toute la difficulté étant d'expliquer ce que cette propriété a de « spécial ». J'en ai déjà parlé, j'y reviendrai. La thèse plus forte selon laquelle « vrai », malgré sa grammaire, n'est pas un prédicat, ou que la vérité n'est pas une propriété du tout, a été soutenue par des voies très différentes par Strawson (1949)[15], et par les tenants de la théorie « prophrastique » de la vérité, en particulier dans Grover, Camp et Belnap (1975), théorie reprise par exemple dans Brandom (1994).[16]

1.3 La légèreté de la vérité et le poids des conditions de vérité : F. Ramsey

Pour prolonger les remarques de Frege la transparence des usages prédicatifs de la vérité sur la vérité, je voudrais consacrer maintenant cette section au travail de Frank Ramsey sur la vérité. Le nom de Ramsey est parfois associé à la théorie de la « vérité redondance », mais Ramsey était-il déflationniste ? C'est dans le manuscrit de son projet *Sur la vérité*, rédigé entre 1927 et 1929 et publié dans Ramsey (1991), que l'on trouve les aperçus les plus approfondis sur la pensée de Ramsey à propos du concept de vérité. Le texte de Ramsey n'a pas seulement valeur de document historique, il fournit aussi quelques unes des meilleures clarifications sur les enjeux

[14] Voir la note 8 dans ce chapitre.

[15] Je mentionne en passant que si Strawson soutient que l'usage central des attributions de vérité est pragmatiquement lié à certains actes particuliers par lesquels il s'agit pour un locuteur d'*endosser* un contenu propositionnel affirmé par un autre, sa position, à y regarder de près, ne semble pas contredire pas le point de Frege.

[16] En conséquence, pour ces auteurs, le *contenu* propositionnel des énoncés dont la forme est celle d'une attribution de vérité est tout entier *hérité* des propositions (exprimées par les énoncés) auxquel(le)s elle est attribuée.

d'une analyse conceptuelle de la vérité, clarifications qu'il faut garder à l'esprit dans toute discussion du déflationnisme.[17] Pour cette raison, c'est à lui que je me référerai principalement (Ramsey (1991)), le considérant comme un développement, ou un éclaircissement des quelques remarques sur la vérité que l'on trouve dans le texte publié de *Facts and propositions* (Ramsey (1927)).

La réflexion de Ramsey sur la vérité s'articule en deux temps. Il y a d'une part un développement systématique de l'idée de redondance pour une certaine notion *propositionnelle* de vérité, développement qui constitue, comme chez Frege, une forme de déflationnisme local. Mais Ramsey ne s'en tient pas à cette analyse. Il insiste au contraire sur le fait qu'une analyse complète de la notion de vérité n'est pas séparable d'une analyse des attributions de référence propositionnelle ou, comme je dirais parfois, de conditions de vérité.[18]

Le problème de la vérité. La première chose à faire pour comprendre la théorie de Ramsey, c'est de clarifier la nature des porteurs de vérité. Cette question est développée sur sept paragraphes couvrant à peu près les pages 7 et 8 de l'édition Resher-Majer du manuscrit. Après avoir rejeté les énoncés et les propositions « objectives », les premiers parce que la vérité doit se dire de ce qui a un sens, ce qui est signifié par un énoncé plutôt que l'énoncé lui-même, les secondes parce qu'il ne veut pas présupposer d'emblée l'existence d'entités aussi problématiques que des propositions objectives, Ramsey discute le choix de certains états mentaux comme candidats au titre de porteurs de vérité :

> Notre tâche, donc, est d'élucider les termes *vrai* et *faux* appliqués aux états mentaux, et comme typiques des états qui nous intéressent nous pouvons prendre pour le moment les croyances. Maintenant, qu'il soit ou non philosophiquement correct de dire qu'elles ont les propositions comme objets, les croyances ont sans aucun doute une caractéristique que j'appellerai *référence propositionnelle*. Une croyance est nécessairement une croyance qu'une chose ou une autre est telle et telle, par exemple que la terre est plate ; et c'est cet aspect, d'être une croyance « que la terre est plate », que je propose d'appeler sa référence propositionnelle. (Ramsey (1991), chap.1, p. 7)

[17] Grover, Camp et Belnap (1975) est une tentative moderne d'explication du concept de vérité qui se réclame ouvertement de Ramsey. Les écrits de Ramsey sur la vérité alors disponibles se réduisaient néanmoins à quelques remarques dans *Facts and propositions*. Rivenc (1998*a*) réexamine de façon critique la filiation supposée à la lumière des notes de Ramsey publiées en 1991.

[18] Voir la section 5 de ce chapitre.

En outre les porteurs de vérité doivent non seulement avoir une référence propositionnelle mais encore un caractère affirmatif :

> [...] Je peux espérer qu'il fera beau demain, me demander s'il fera beau demain, et finalement croire qu'il fera beau demain. Ces trois états ont la même référence propositionnelle, mais seule la croyance peut être dite vraie ou fausse. Nous n'appelons pas les souhaits, les désirs ou les interrogations vrais, non parce ce qu'ils n'ont pas de référence propositionnelle, mais parce qu'ils n'ont pas ce que l'on pourrait appeler un caractère affirmatif ou assertif, cet élément qui est présent dans la pensée que, mais absent lorsqu'on se demande si. (Ramsey (1991), p. 8)

Les porteurs de vérité sont donc des objets qui n'ont pas de nom dans le langage courant, et Ramsey propose un emploi technique de « croyance » [*belief*] et « jugement » [*judgments*] :

> Les états mentaux qui nous concernent, ceux, nommément, avec une référence propositionnelle et un certain degré de caractère affirmatif, n'ont malheureusement pas de nom dans le langage ordinaire. Il n'y a pas de terme applicable à tout le domaine allant de la simple conjecture au savoir certain, et je propose de pallier cette lacune en utilisant les termes *croyance* et *jugement* comme synonymes pour couvrir tout le domaine des états en question, et non dans leur sens ordinaire plus étroit. (Ramsey (1991), p. 8)

Mais si la vérité est ici définie comme une propriété de certains états mentaux ayant une référence propositionnelle, avec le problème de la définition de la vérité vient le problème de comprendre ce que sont ces relations qui existent entre la pensée (ou le langage) et ce *à propos* de quoi nous pensons et parlons, telles que nous les identifions *via* la référence propositionnelle. Il y a donc en fait *deux* problèmes derrière le problème de définir la signification du mot « vrai » : le problème de définir la vérité *étant donné* la notion de référence propositionnelle, et le problème de définir la référence propositionnelle elle-même.

Déflationnisme local et théorie prophrastique. La réponse de Ramsey au premier problème[19] est simple : une croyance est vraie si c'est une croyance que p et p, fausse si c'est une croyance que p et non p. Du point de vue qui est le nôtre aujourd'hui, l'esprit, sinon la lettre et la grammaire, de cette réponse, rappelle immanquablement

[19] Ramsey ne prête pas attention aux paradoxes ici.

ce qui sera le point de départ de Tarski, le schéma-T.[20] Mais à la différence de Tarski, Ramsey voit là déjà proprement une *définition* de la vérité. Cette définition, remarque Ramsey, a quelque chose de surprenant dans sa formulation :

> Nous pouvons symboliser toute croyance que p, où « p » est un énoncé variable [...]. Nous pouvons alors dire qu'une croyance est vraie si c'est une croyance que p, et p. Cette définition sonne d'abord bizarrement parce que nous ne comprenons pas tout de suite que « p » est un *énoncé* variable et doit être regardé comme contenant un verbe (Ramsey (1991), p. 9, traduit dans Rivenc (1998*a*)).

Ce que remarque Ramsey ici, c'est qu'il est possible de *définir* la vérité si l'on permet l'usage, non pas simplement de variables d'énoncés au sens de variables dont les valeurs possibles sont des énoncés, mais de *variables en position d'énoncé*, c'est-à-dire de *prophrases*, dont le rôle est supposé être analogue au rôle pronominal des variables ordinaires mais pour une autre catégorie d'expressions, celle des phrases.[21] Et puisque les attributions de vérité sont paraphrasables sans prédicat de vérité à l'aide de prophrases, on peut parler de théorie *prophrastique* de la vérité. Ramsey ne voit aucune difficulté de principe dans cette idée de variables occupant des positions d'énoncés. Qu'il ne semble pas possible d'écrire une telle définition dans le langage ordinaire n'est pour lui qu'un problème « artificiel », un simple effet d'une lacune de certains moyens d'expression, nommément la pauvreté du langage ordinaire en « prophrases » :

> [...] cette particularité du langage donne naissance à des problèmes artificifiels quant à la nature de la vérité, problèmes qui disparaissent dès qu'ils sont exprimés dans le symbolisme logique, où nous pouvons rendre « ce qu'il croit est vrai » par « si p est ce qu'il croit, alors p » (p.10. Traduit dans Rivenc (1998*a*))

[20] A cet égard, la proximité est plus flagrante encore dans le passage suivant, qui semble préfigurer la convention-T (voir chapitre suivant) :

> Et quelle que puisse être la définition complète [de la vérité], elle doit préserver la connexion évidente entre vérité et référence [propositionnelle], qu'une croyance « que p » est vraie si et seulement si p.(Ramsey (1991), p. 13-14)

[21] En logique du premier ordre ordinaire, la quantification est objectuelle : en particulier, les seules variables sont des pronoms, occupant dans les énoncés des positions nominales. Il se peut que parmi les valeurs possibles des variables d'un langage ordinaire du premier ordre figurent des énoncés, comme en témoigne par exemple la phrase vraie « Tous les énoncés du français commencent par une majuscule », mais cela ne fait pas pour autant de ces variables des prophrases. Les variables prophrastiques occupent des *positions* d'énoncés.

D'un point de vue simplement « technique », on peut rendre cette définition au moyen de quantificateurs substitutionnels, en admettant des expressions appartenant à la catégorie des énoncés dans des classes de substitutions appropriées[22] :

$$\Pi p \text{ (la croyance que } p \text{ est vraie si et seulement si } p)$$

Mais la question philosophique de la légitimité et du statut de cette quantification substitutionnelle, n'admet pas de réponse qui aille de soi.[23] Laissant là pour le moment la question du statut de la quantification substitutionnelle,[24] j'irai directement aux conclusions de Ramsey. Tout d'abord, Ramsey note que la vérité définie de cette façon rend compte de la conception classique de vérité comme correspondance avec les faits :

> On peut remarquer que notre définition qu'une croyance est vraie si c'est « une croyance que p » et p, mais fausse si c'est « une croyance que p » et non-p est, en substance, celle d'Aristote, qui ne considérant que les deux formes « A est » et « A n'est pas » déclarait : « dire de ce qui est, que ce n'est pas, et de ce qui n'est pas, que c'est, est faux, tandis que dire de ce qui est que c'est, et de ce qui n'est pas, que ce n'est pas, est vrai ».
>
> Quoique nous n'ayons pas encore utilisé le mot « correspondance » notre théorie sera probablement appelée une théorie de la correspondance. Car si A est B nous pouvons parler selon l'usage commun du fait que A est B et dire qu'il correspond à la croyance que A est B de telle manière que si A n'est pas B il n'y a aucun fait qui lui corresponde. (Ramsey (1991), p. 11).

Mais, déjà pour Ramsey, cette idée de la vérité comme correspondance appelle une forme de déflationnisme, dans l'économie conceptuelle dont il semble prêt à se satisfaire pour expliquer cette correspondance. En effet, Ramsey remarque un peu plus loin qu'il n'y a sans doute pas d'explication uniforme possible de cette relation qui advient entre une pensée et le monde quand cette pensée est vraie, et doute de la pertinence d'une description de cette relation en termes de « croyance » et de « fait » :

[22] Pour une discussion plus approfondie, je renvoie à l'article classique de S. A. Kripke (1976).

[23] Ce serait un sujet à part entière que de discuter ce point. Je renvoie à nouveau à Rivenc (1998*a*) pour une analyse détaillée et des références supplémentaires.

[24] J'en dirai à nouveau un mot dans la section suivante, p. 47.

> Et l'on peut être sceptique sur le fait qu'il y ait une simple relation de correspondance applicable à tous les cas ou même s'il est toujours correct de décrire cette relation comme existante [holding] entre la « croyance que p » et le « fait que p ».[25](Ramsey (1991), p. 11)

En fait, et c'est là un mouvement authentiquement « déflationniste », les attentes philosophiques du discours sur la correspondance sont revues à la baisse. Si correspondance avec des faits il y a, alors la théorie « prophrastique » en rend compte sobrement, et sinon, la théorie prophrastique vaut encore et le parler en termes de faits n'en est qu'une glose inadéquate :

> La vérité, disons-nous, est quand un homme croit que A est B et A est B, qu'une telle occurrence puisse ou non être adéquatement décrite comme une correspondance entre deux faits ; l'incapacité à la décrire en termes de correspondance ne montre pas qu'elle n'advienne jamais et que ce n'est pas ce que nous voulons dire par la vérité. (p. 12)

A ce point, Ramsey a donné sa réponse au premier des deux problèmes que nous mentionnions : celui de la définition de la vérité étant donné la notion de référence propositionnelle. Ce qu'il y a dire à ce niveau d'analyse est essentiellement, d'une part, que le contenu du mot « vrai » est en substance épuisé par les platitudes du type « une croyance que A est vraie si et seulement si A », d'autre part que ce contenu est définissable dans un langage logique suffisamment riche en prophrases et enfin que ces platitudes saisissent l'intuition de la vérité comme correspondance, qui est la conception classique, aristotélicienne, de la vérité.

Le problème de la référence propositionnelle. Mais à ce stade, Ramsey est clair sur ce point, il reste le second problème que nous mentionnions plus haut, et nous n'avons fait qu'effleurer le véritable problème de la vérité, ou ce que nous entendons ordinairement par là :

> Même si nous sommes d'accord que le problème est de définir la vérité au sens d'expliquer sa signification [*N.d.t : et non de donner un critère de vérité*], ce problème peut revêtir deux complexions très différentes selon

[25]Entre ce passage et le précédent, il y a ce passage important :

> Mais nous ne pouvons pas décrire la nature de cette correspondance tant que nous ne connaissons pas l'analyse de la référence propositionnelle de « croire que A est B ». C'est seulement quand nous connaissons la structure de la croyance que nous pouvons dire quel genre de correspondance unit les croyances vrais et les faits.

Je reviens sur ce point dans la suite.

> le genre de définition dont nous sommes prêts à nous satisfaire. Notre définition est donnée en terme de référence propositionnelle, que nous prenons comme un terme déjà compris. Mais on peut penser que cette notion de référence propositionnelle a elle-même besoin d'être analysée et définie, et qu'une définition de la vérité dans les termes d'une notion si obscure ne représente qu'un petit progrès, si elle représente un progrès. Si une croyance est identifiée à ce que M. John pensait à 10 heures ce matin, et si nous demandons ce qui est signifié en appelant vraie la croyance ainsi identifée, pour appliquer la réponse que nous avons obtenue jusqu'à présent nous avons besoin de savoir de quoi la croyance de M. John était une « croyance que » ; par exemple, nous disons que si c'était une croyance que la terre est plate, alors elle était vraie si la terre est plate. Mais nombreux sont ceux qui pensent que c'est là seulement éviter la partie la plus difficile et la plus intéressante du problème, qui est de trouver comment et dans quel sens ces images ou ces idées dans l'esprit de M. John à 10 heures constituent ou expriment « la croyance que la terre est ronde ». La vérité, dira-t-on, consiste en une relation entre des idées et la réalité, et l'utilisation sans analyse du terme de référence propositionnelle annule et évite les problèmes réels que soulève cette relation. Nous devons admettre que cette charge est juste et qu'une explication de la vérité qui accepte la notion de référence propositionnelle sans analyse ne peut être considérée comme complète. Car toutes les difficultés liées à cette notion sont vraiment impliquées dans la vérité qui dépend d'elle ; si par exemple « référence propositionnelle » a une signification très différente en relation avec des genres différents de croyances (comme beaucoup le pensent), alors une ambiguïté similaire est latente dans « vrai » également et il est évident que notre idée de la vérité ne sera pas claire tant que ce problème et les problèmes similaires ne seront pas résolus. (Ramsey (1991), p. 13-14)

Les problèmes attachés à notre compréhension de la vérité comme « correspondance », sont donc plus profondément les problèmes liés à la nature des attitudes propositionnelles auxquelles nous attribuons la vérité : en quoi cela consiste-t-il pour un état mental, d'être une croyance que p ? Quel genre de propriétés doit avoir un individu pour être dans un état de croyance *que la neige est blanche* ? (Que la

proposition soit vraie ou fausse). Quel genre de relation doit-il entretenir avec son environnement ? Dans *Facts and Propositions*, Ramsey disait exactement la même chose, à savoir que la difficulté porte sur la notion de proposition :

> [...] il est clair que le problème ne porte pas sur la nature du jugement de la vérité et de la fausseté, mais concerne la nature du jugement ou de l'assertion, car ce qu'il est difficile d'analyser dans la formulation donnée plus haut, c'est « Il affirme aRb » (Ramsey (1927))

Pour conclure, il est clair d'une part que le problème fondamental de la vérité est identifié comme étant celui de la référence propositionnelle, et d'autre part que le problème de la définition de « vérité », *étant donné* une explication de la notion de référence propositionnelle, est trivial. Dans « On truth », Ramsey ne fournit pas la solution au problème de la référence propositionnelle. Mais pour triviale que soit l'analyse de la vérité *étant donné les propositions*, elle n'en a pas moins son intérêt. Ramsey résume ainsi la situation :

> ... quoique la réduction de la vérité à la référence propositionnelle ne soit qu'une petite partie, et la plus facile, de son analyse, ce n'est pas une partie que nous pouvons nous permettre de négliger. Car non seulement est-il de toute façon essentiel de réaliser que le problème se divise en deux parties, la réduction de la vérité à la référence et l'analyse de la référence elle-même, et d'être clair sur la partie du problème à laquelle nous nous attaquons à un moment donné, mais encore à de nombreuses fins est-ce seulement la première partie, et la plus facile, de la solution qui est requise ; nous sommes souvent concernés non par les croyances ou jugements comme occurrences à des instants particuliers dans les esprits d'homme particuliers mais, par exemple, par *la* croyance ou le jugement « tous les hommes sont mortels » ; dans ce cas la seule définition de la vérité dont nous pouvons avoir possiblement besoin est celle en terme de référence propositionnelle, qui est présupposée dans la notion même *du* jugement « tous les hommes sont mortels » ; car lorsque nous parlons *du* jugement « tous les hommes sont mortels » ce à quoi nous avons vraiment à faire est n'importe quel jugement particulier, dans n'importe quelle occasion particulière, qui a cette référence propositionnelle, qui est le jugement « que tous les hommes sont mortels ». Ainsi, quoiqu'il faille faire face à ces difficultés psychologiques impliquées dans cette notion

> de référence pour tout traitement complet de la vérité, il est bon de commencer avec une définition qui est suffisante pour un grand nombre de fins et ne dépend que des considérations les plus simples.
>
> Et quelle que puisse être la définition complète, elle doit préserver la connexion évidente entre vérité et référence, qu'une croyance « que p » est vraie si et seulement si p. On peut ridiculiser ceci comme un formalisme trivial, mais comme on ne peut le contredire sans absurdité, il fournit comme test simple pour toutes recherches plus profondes, qu'elles doivent satisfaire ce truisme évident. (Ramsey (1991), p. 13-14)

On pourrait reformuler le point de Ramsey de façon à rendre plus apparente sa proximité avec ce qui sera la pensée de Quine en distinguant les deux situations dont parle Ramsey (aux lignes 10, 11, et 12 du passage cité), comme des situations dans lesquelles nous n'avons pas (premier cas) et nous avons (second cas) accès aux *conditions d'individuation* des croyances en terme de leur contenu propositionnel. Dans le second cas, celui où la référence propositionnelle des jugements est donnée, l'analyse prophrastique est suffisante pour expliquer les attributions de vérité. Comme, selon Ramsey, bon nombre d'usages de la notion de vérité se font dans de tels contextes, l'analyse précédente, qui suffit à éclaircir complètement ces usages, est utile. D'un autre côté, l'analyse prophrastique de la vérité ne suffit pas à expliquer ce que c'est qu'une croyance vraie en général, dès que le contenu de croyance n'est plus spécifié en termes de référence propositionnelle, parce qu'à nouveau l'analyse prophrastique suppose que l'on parle non seulement de croyance, mais de croyance ayant une référence propositionnelle donnée, une croyance *que* p : une croyance est vraie si et seulement si c'est une croyance que p, et p.

Au bout du compte, Ramsey était-il déflationniste ? An sens fort de « déflationniste », au sens du déflationnisme sémantique général dont nous avons parlé dans la première partie de ce chapitre, rien ne permet de l'affirmer. S'il prend soin de montrer que l'on peut distinguer deux problèmes derrière le problème de la définition de la vérité, Ramsey ne nous donne que peu d'indications sur le second, le problème que l'on pourrait qualifier de « sémantique » ou de « problème de l'intentionalité », en un sens large qui ne se réduit pas aux expressions d'un langage, mais qui concerne plus généralement la question de savoir comment certaines choses, des mots, des énoncés, des états mentaux, etc., peuvent être *à propos* de quelque chose, avoir la référence propositionnelle qu'ils ont. En tout état de cause, il est difficile de

conclure sur cette question.[26] Mais Ramsey était certainement déflationniste au sens faible du déflationnisme local, dans son analyse conceptuelle de la notion de vérité des propositions. Certes, Ramsey considérait que la vérité est une propriété, et il en cherchait une définition.[27] Il n'en reste pas moins vrai, pourtant, qu'en tant que prédicat *des propositions*, la vérité n'est pas beaucoup plus qu'un outil logique. Un des mérites de Ramsey est d'avoir clarifié la distinction entre la simple déflation de la vérité et le problème plus général de la déflation de la référence propositionnelle, c'est-à-dire des attributions de conditions de vérité, de contenus sémantiques et intentionnels, auxquelles nous nous livrons pour décrire et expliquer certains aspects des rapports entre le langage, ou la pensée d'un sujet, et le monde. Un autre mérite, concernant plus spécialement le problème de la vérité, est non seulement d'avoir montré que le prédicat de vérité est, en un sens, transparent, dans les cas d'attributions de type

« la neige est blanche » est vraie

où le prédicat de vérité est éliminable sans perte de sens, mais encore d'avoir donné un sens précis à cette thèse de la transparence, en montrant comment paraphraser les énoncés contenant des attributions de vérité en éliminant ces dernières à l'aide de *prophrases*, y compris dans les cas où elles ne sont pas éliminables par une paraphrase en langage ordinaire, comme dans les contextes de généralité ou d'attributions dites *aveugles*.[28] Mais si Ramsey, avant Tarski, a pu donner une *définition* de la vérité, cette définition ne sera pas retenue par la postérité comme une authentique *explication* de la notion de vérité, le langage de l'*explicans*, avec ses variables prophrastiques, ayant lui-même besoin d'être clarifié.

1.4 Penser les relations du langage au monde dans l'empirisme radical de Quine

Quine est en général considéré comme le père du déflationnisme contemporain. Mais il est surtout, parmi les auteurs que je considère ici, le premier a défendre

[26] On pourra consulter Loar (1980) pour une tentative d'interprétation.

[27] Nous suivons Rivenc (1998*a*) sur ce point, contre Grover, Camp et Belnap (1975).

[28] Une attribution de vérité à un contenu propositionnel est aveugle lorsque le contenu propositionnel auquel la vérité est attribuée est présenté par une description, par exemple dans :

La dernière phrase écrite par Platon est vraie.

un déflationnisme sémantique général. Je vais présenter la pensée de Quine sur la vérité en l'articulant autour de trois thèmes, l'explication de la correspondance, l'immanence de la vérité, et la fonction logico-linguisitique du prédicat de vérité, en montrant comment cette articulation permet de prolonger mais aussi de radicaliser les réflexions de Ramsey sur la vérité.[29]

Correspondance. Quine présente clairement dans certains textes une motivation « correspondantiste » pour une forme de déflationnisme en matière de vérité. En fait, la conception décitationnelle est souvent présentée par Quine comme ce qu'il reste de la conception de la vérité comme correspondance une fois cette dernière débarrassée de la phraséologie philosophiquement suspecte dans laquelle elle est ordinairement plongée. L'opération chirurgicale est bien décrite dans ce passage de *Quiddités* :

> Qu'est-ce qui du côté des énoncés vrais est supposé correspondre à quoi du côté de la réalité ? Si nous cherchons une correspondance mot à mot, nous nous retrouvons à créer dans la réalité un supplément d'objets abstraits fabriqués aux fins de la correspondance. Ou peut-être nous décidons-nous pour une correspondance des énoncés complets à des *faits* : un énoncé est vrai si et seulement s'il rapporte un fait. Mais là encore nous avons fabriqué une substance pour une doctrine vide. Le monde est plein de choses, reliées de façon variées, mais que sont, en plus de tout cela, les faits ? Ils sont projetés à partir des énoncés vrais pour les fins de la correspondance.
>
> Mais réfléchissons à cette manœuvre un moment. La vérité de « la neige est blanche » est due, nous dit-on, au fait que la neige est blanche. L'énoncé vrai « la neige est blanche » correspond au fait que la neige est blanche. L'énoncé « la neige est blanche » est vrai si, et seulement si, c'est un fait que la neige est blanche. [...] La combinaison « c'est un fait que » est vide et peut être éliminée ; « C'est un fait que la neige est blanche » se réduit à « la neige est blanche ». Notre explication de la vérité en termes de faits se réduit maintenant à ceci : « la neige est blanche » est vrai si, et seulement si, la neige est blanche (Quine (1992), p. 213)

[29] La philosophie de la vérité de Quine est principalement développée dans Quine (1953*b*), Quine (1970), Quine (1960*b*), Quine (1990), et Quine (1992).

D'une part, donc, le projet d'une théorie de vérité correspondance doit faire face à des difficultés méthodologiques majeures et, d'autre part, un tel projet *n*'est *pas nécessaire*. L'intuition de la vérité comme correspondance, au moins dans le cas d'un énoncé particulier, est *complètement* expliquée avec toute la clarté voulue par les biconditionnels construits sur le modèle suivant :

« La neige est blanche » est vrai si et seulement si la neige est blanche.

Nous avons là, nous dit Quine[30], le cœur de l'idée de vérité comme correspondance : cet énoncé donne bien les conditions qui doivent être réalisées dans le monde pour qu'une certaine entité linguistique soit vraie.[31] Mais Quine, et c'est en cela que l'on peut parler de déflationnisme, renonce à expliquer cette relation de correspondance en vertu de laquelle ce biconditionnel particulier, et ceux qui lui ressemblent, sont vrais. Ce n'est pas seulement que la relation de correspondance elle-même n'est pas théorisée[32], mais qu'elle n'a pas à l'être, et qu'elle ne peut pas l'être. C'est pourquoi, même si le déflationnisme n'est que la correspondance bien comprise, il s'oppose au projet d'une *théorie de la correspondance*, d'un discours *sur* la correspondance qui est au cœur des tentatives modernes d'élucidation de la notion classique de vérité, qu'il s'agisse du discours métaphysique visant à articuler la relation de « *truth-making* », ou du discours malheureux parce qu'impossible sur le langage comme « image » du monde, mais aussi, nous le verrons au paragraphe suivant, du discours qui prétendrait parler scientifiquement de cette relation de correspondance.

*Immanence.*On n'aura pas touché au cœur de la philosophie de la vérité de Quine tant qu'on ne l'aura pas resituée dans le contexte théorique plus général d'où elle émerge. On sait les conséquences sémantiques que Quine tire de la thèse du holisme de la confirmation[33] : il n'y a rien de tel qu'une notion empiriquement acceptable de signification d'un énoncé pris isolément, sauf peut-être dans le cas de quelques énoncés d'observations. Avec le holisme sémantique, c'est la notion de référence propositionnelle d'un énoncé qui est rejetée hors du langage de la science.

[30]Reprenant ici les analyses de Tarski sur lesquelles je reviendrai au chapitre suivant, plutôt que celles Ramsey.

[31]Paul Alston parle, à raison, de conception « réaliste » de la vérité. Voir Alston (2001) pour une défense originale contre les théories dites « épistémiques » de la vérité.

[32]Il n'y a pas de lois générales d'une relation de correspondance R entre les mots et les choses.

[33] Cette thèse que Quine emprunte à Duhem et qu'il résume parfois en rappelant que les théories affrontent *en bloc* « le tribunal de l'expérience ». Pour une étude une présentation critique et plus détaillée je me permets de renvoyer le lecteur aux nombreuses études dont la pensée de Quine fait l'objet. En français, le lecteur pourra consulter Gochet (1992), Laugier (2002). Voir aussi plus récemment Rivenc (2008) pour une discussion critique.

Dans *Word and Object*[34], c'est une réflexion fondée sur l'expérience de pensée de la traduction radicale et la reconstruction rationnelle de la genèse du langage chez l'enfant qui viennent renforcer les conclusions précédentes. L'accent est désormais porté sur la génèse de la *référence* et son inscrutabilité, mais en dernière analyse les arguments de Quine restent les mêmes. Pour interpréter un locuteur étranger, il faut lui attribuer une théorie du monde, un ensemble complexe de croyances qui sont supposées motiver chez le locuteur l'affirmation de certains énoncés dans certaines circonstances observables. Mais il y a une infinité de théories que l'on peut attribuer à un locuteur de façon compatible avec les circonstances observables, qui sont autant de fondements de schèmes de traduction de leur langage dans le nôtre. Cette réflexion motive dans un premier temps les thèses de la *sous-détermination* de la traduction par les données observables. Mais Quine va plus loin et un dernier pas est franchi avec le slogan : « pas d'entité sans identité ». Parce qu'il n'est pas possible de donner de conditions d'individuation acceptables pour les propositions en termes de dispositions au comportement verbal, Quine conclut que les propositions, au sens où nous entendons ordinairement ce mot, n'ont pas droit de cité dans le langage de la science. Tout au plus s'agit-il de *façon de parler*.

Les thèses d'indétermination, et l'éliminativisme relativement à la notion de référence propositionnelle qui en est la conséquence, sont décisives pour la philosophie de la vérité de Quine. Puisque, à en croire Quine, la notion de proposition n'a pas de légitimité scientifique, les porteurs de la vérité ne peuvent évidemment pas être des propositions ; logiquement ce sont donc des *énoncés*, dont les conditions d'individuation sont tenues pour moins problématiques. Mais attention, c'est bien d'énoncés *interprétés* qu'il s'agit ici, l'attribution de la vérité à un simple énoncé formel n'ayant aucun sens.[35] En reprenant maintenant à rebours le raisonnement de Quine, demandons-nous quelles sont les conditions d'individuation d'un énoncé interprété. L'individuation syntaxique ne peut être qu'un substitut dans certains contextes particuliers, puisque des énoncés syntaxiquement identiques peuvent compter comme des énoncés interprétés différents selon qu'ils sont vus comme des énoncés d'un langage donné ou d'un autre. Par ailleurs ces énoncés ne sont pas individués par les propositions qu'ils expriment, puisqu'il n'y a rien de tel. De même,

[34] Mais déjà dans *Meaning and Linguisitics* §5 *in* Quine (1953*a*).

[35] Plus précisément, les porteurs de la vérité sont supposés être ce que Quine appelle les « énoncés éternels », qui ne contiennent pas de déictiques, indexicaux, etc. Par ailleurs, il faut peut-être préciser que Quine n'a jamais nié que les énoncés, en un sens relâché, ont un sens !

l'individuation des énoncés interprétés par l'existence d'une bonne traduction qui corrèle un énoncé à un autre n'a pas beaucoup de sens, puisque ce qui constitue une bonne traduction est largement fonction de facteurs pragmatiques, et qu'à peu près n'importe quel énoncé peut, dans un certain contexte, être une bonne traduction de n'importe quel autre énoncé. Non, il n'y a de condition d'individuation satisfaisante des énoncés interprétés (*via* l'individuation syntaxique), que dans *mon* langage, mon idiolecte individuel ou peut-être le langage de ma communauté linguistique. Corrélativement, les attributions de vérité elles-mêmes n'ont d'abord de sens que dans mon langage, dans le contexte de la « théorie » ambiante totale du monde qui est la mienne et les attributions de vérité à des énoncés d'un autre langage que le mien n'ont de sens que relativement à *une traduction* de ces énoncés dans mon langage. La vérité est donc une notion essentiellement *interne* pour Quine, ce qu'il appelle lui-même une notion *immanente*.[36] C'est au paragraphe 6 de *Word and Object* que ce point apparaît peut-être avec le plus de clarté :

> C'est plutôt lorsque nous nous replaçons au cœur d'une théorie réellement existante, au moins hypothétiquement acceptée, que nous pouvons parler et que nous parlons avec sens de tel ou tel énoncé comme vrai. S'il y a du sens à appliquer « vrai » à un énoncé, c'est à un énoncé couché dans les termes d'une théorie donnée, et vu de l'intérieur de cette théorie, complète, avec la réalité posée par cette théorie. (Quine (1960*a*), §6)[37]

D'une part, ces remarques éclairent la portée de ce que Quine présentait tout à l'heure comme la façon adéquate de présenter la relation de correspondance. S'il n'y a rien *de général* à dire sur la relation entre les énoncés vrais et le monde, indépendamment du langage dans lequel cet énoncé est formulé, ce n'est pas seulement que la notion de *fait* est inutile. C'est aussi qu'il ne peut rien y avoir de tel que des faits, la notion de fait elle-même se trouvant souffrir d'indétermination : pas plus qu'il n'y a de conditions d'individuation claire des propositions, il n'y a de condition d'individuation claire des faits. C'est ce à quoi il fallait s'attendre, puisque les faits ne sont, comme les propositions, que de simples projections de *nos* énoncés dans le monde. Mais nous sommes passés d'une thèse d'*inutilité* à une thèse d'*inanité* du concept de fait.

[36] Voir par exemple Resnik (1990) pour un développement plus complet sur le caractère immanent de la notion de vérité.

[37] Voir aussi Quine (1970), chap. 1 et 2.

D'autre part, si la vérité est une notion immanente à notre propre langage, elle n'est pas une propriété robuste sur laquelle nous pourrions fonder une explication scientifique de la relation de « correspondance » qui doit exister entre ce que dit ou pense un sujet et certains aspects de son environnements physique. Si le caractère immanent de la vérité, n'implique pas qu'il ne serait pas possible de faire appel à la vérité *dans son propre langage* dans un processus *d'interprétation* d'un langage étranger,[38] il implique que la notion de vérité n'a sûrement aucun rôle à jouer dans l'explication scientifique, en troisième personne, des relations du langage et du monde, pas plus que les conditions de vérité des énoncés.[39]

Décitation. Mais à quoi sert donc le prédicat de vérité ? La contribution de Quine à cette question revêt une importance décisive pour les tentatives contemporaines de formulation d'une position déflationniste articulée.[40] L'apport de Quine est d'avoir identifié l'utilité des attributions de vérité dans l'expression de certaines *généralisations*. Pour comprendre ce rôle, il faut prêter attention à deux choses.

D'une part il y a, selon Quine, des circonstances dans lesquelles « bien qu'ayant en tête une réalité d'ordre non linguistique, nous devons procéder indirectement et parler des énoncés. »[41] Pourquoi cette « montée sémantique » est-elle nécessaire ? Elle répond, nous dit Quine, au besoin d'exprimer certaines *généralisations* :

> Nous pouvons généraliser sur « Tom est mortel », « Richard est mortel », etc., sans parler de vérité ou d'énoncés. Nous pouvons dire : « Tous les

[38] Ce que le linguiste-interprète imaginaire de Quine tente bien de faire c'est une théorie de la vérité du langage étranger en ce sens, formulant des hypothèses du type :

Une traduction de « Gavagai » est vraie-dans-mon-langage si et seulement si il y a là un lapin.

ou

Une traduction de « Gavagai » est vraie-dans-mon-langage si et seulement si il y a là des parties non-détachées de lapin.

[39] Je serais tenté d'ajouter que la notion de vérité n'est pas mise à contribution pour expliquer les relations du langage et du monde dans la philosophie de Davidson non plus. Le processus interprétatif conduit en première personne par l'interprète, ne constitue pas une *explication* des relations du langage du locuteur et du monde au sens d'une d'explication scientifique fondée sur la formulation de lois objectives et universelles. Interprétation n'est pas explication, et s'il y a « explication » en un certain sens, c'est d'une explication « projective » qu'il s'agit. Pour l'expression par Davidson d'une certaine sympathie philosophique envers les conceptions déflationnistes de la vérité, on se reportera à Davidson (1996).

[40] Sur ce point, le développement de référence se trouve dans Quine (1970). Voir aussi Quine (1990).

[41] Quine (1970), p. 11.

> hommes sont mortels ».[...] Quand par ailleurs nous voulons généraliser sur « Tom est mortel ou Tom n'est pas mortel », « La neige est blanche ou la neige est non blanche », et ainsi de suite, nous nous élevons jusqu'à parler de vérité et d'énoncés en disant : « Tout énoncé de la forme 'p ou non p' est vrai », ou « Toute disjonction d'un énoncé avec sa disjonction est vraie ». Ce qui nous contraint à cette montée sémantique, ce n'est pas que « Tom est mortel ou Tom n'est pas mortel » porterait de quelque façon sur des énoncés, tandis que « Tom est mortel » et « Tom est Tom » porterait sur Tom. Ces trois énoncés portent tous sur Tom. [...] (QUINE (2008), p.22)

Qu'est-ce, alors, qui nous contraint à cette montée ? La généralisation sur « Tous les hommes sont mortels » peut s'écrire

> « x est mortel pour tous les hommes x » - i.e. tous les objets x d'une espèce qui est telle que « Tom » est le nom de l'un d'entre eux. Or que serait l'interprétation analogue de la généralisation de « Tom est mortel ou Tom est non mortel » ? Elle s'écrirait « p ou non p pour tous les objets p d'une espèce qui est telle que des énoncés en sont des noms ». Mais les énoncés ne sont pas des noms, et cette interprétation est incohérente, car elle emploie « p » à la fois dans des places qui appellent des membres de phrase et dans une place qui appelle un substantif. Donc, pour parvenir à l'assertion générale que nous cherchons, nous montons d'une marche et nous parlons des énoncés : « Tout *énoncé* de la forme 'p ou non p' est vrai » (*ibid.*)

Il est vrai que cette nécessité n'en est une, note Quine, que parce que, justement, les énoncés ne sont pas des noms, et pas des noms de propositions. On pourrait, certes, vouloir prendre la décision théorique d'analyser les énoncés comme des noms de propositions, mais la quantification *objectuelle* sur des variables propositionnelles reviendrait en fait à admettre l'existence de propositions, et Quine a donné ses raisons de ne pas le faire. Le prédicat de vérité permet de nous en tenir à cette décision méthodologique. L'on pourrait également, souvenons-nous de Ramsey, formuler la généralité visée sans prédicat de vérité, grâce à la quantification *substitutionnelle* :

$$\begin{array}{c}\text{Pour tout } p,\ p \text{ ou non } p\\ (\Pi p \quad (p \vee \neg p))\end{array}$$

Mais si cette fois on ne suppose pas que les énoncés sont des noms de propositions, on a profondément changé notre façon de comprendre la quantification. Or, pour Quine, la quantification se comprend d'abord objectuellement, toute variable étant essentiellement un pronom, dont le domaine des valeurs possibles forme en dernière instance le compte de nos engagements ontologiques. On peut bien sûr définir la quantification substitutionnelle, expliquer sa sémantique, mais justement tout est là : c'est un artifice et elle a besoin d'être expliquée.[42] Pour cette raison même, d'une part la vérité et la montée sémantique ont bien leur utilité pour exprimer la généralité visée plus haut (en restant dans un registre philosophiquement transparent du discours, avec sa quantification « naturelle ») et d'autre part c'est bien plutôt la quantification objectuelle et le prédicat de vérité qui doivent justement nous permettre d'expliquer la quantification substitutionnelle. Et, pour la même raison, on comprend que pour Quine une *définition* de la vérité en termes substitutionnels, à la Ramsey, si elle est possible, ne peut atteindre son but en tant qu'*explication* de la notion de vérité. Ainsi, de définissable, et donc éliminable, qu'elle était pour Ramsey, la vérité est au contraire devenue indispensable pour Quine.[43] Dans la suite de ce travail nous suivrons Quine sur ce point en admettant que la quantification substitutionnelle est seconde par rapport à la quantification objectuelle et qu'elle doit être expliquée en termes de quantification objectuelle et de la notion de vérité.

Mais revenons à Quine et au rôle du prédicat de vérité. Quel est le rôle exact du prédicat de vérité dans l'expression de ces généralités ? Si le besoin de généralité nous conduit à une montée sémantique, le rôle du prédicat de vérité est « de nous ramener sur terre », pour reprendre l'expression de Rivenc (2008). Comment ? Comme on le voit sur les énoncés du type

« La neige est blanche » est vrai si et seulement si la neige est blanche

l'attribution de la vérité à un énoncé cité a un effet *décitationnel*. Décitationnel, parce que l'attribution de vérité à l'énoncé cité est « équivalente » à l'affirmation de cet énoncé lui-même.[44] Ce cas particulier d'attribution de la vérité à un énoncé cité n'est qu'une illustration, particulièrement claire, du fait que le rôle plus général

[42] Voir à nouveau S. A. Kripke (1976).

[43] Nous revenons au chapitre suivant sur la possibilité de définir la vérité dans un langage avec des quantificateurs ordinaires.

[44] Nous retrouvons là, bien entendu, une variation sur un thème déjà présent chez Frege ou Ramsey, à la différence que, cette fois, c'est aux *énoncés*(-de-mon-langage) que la vérité est attribuée.

de la vérité est « l'annulation de la référence au langage »[45] dans notre discours sur le monde. Et c'est dans les contextes d'expression de la généralité comme ceux que nous mentionnions pour commencer que cet effet rend le prédicat de vérité indispensable :

> Nous pouvons affirmer un énoncé singulier simplement en l'énonçant, sans nous aider de guillemets ni d'un prédicat de vérité ; mais si nous voulons affirmer une collection infinie d'énoncés que nous ne pouvons délimiter qu'en parlant de ses membres, nous sommes bien contents de trouver le prédicat de vérité. Il nous le faut pour rétablir l'effet qu'a la référence à des objets, lorsque, en vue de quelques généralisations, nous avons recours à la montée sémantique. (Quine (2008), p.25.)

La situation du prédicat de vérité est donc assez singulière. Bien qu'indispensable, son rôle semble se réduire à fournir une aide pour l'expression de certaines généralités, en coordination avec une quantification ordinaire sur les énoncés.[46]

Conclusion. Quine était-il déflationniste ? [47] Sans doute ne souscrirait-il pas à la plupart des slogans déflationnistes contemporains. Considérons par exemple l'idée que la vérité ne serait pas une propriété, ou ne serait pas une propriété « naturelle », ou « substantielle » ? La thèse de l'immanence, l'insistance de Quine sur la transparence du prédicat de vérité lorsqu'il est attribué à un énoncé cité, son insistance aussi sur le rôle du prédicat de vérité pour l'expression de certaines généralisations, ces thèses pourraient donner lieu à une tentative de caractérisation de ce genre : après tout il semble bien qu'il y ait quelque chose de spécial avec la propriété de vérité. Mais, bien entendu, ce genre de caractérisation n'a aucune place dans la

[45] Quine, ibid. p.24.

[46] Mentionnons que l'analyse standard des attributions de vérité « aveugles » à un énoncé identifié par une description les réduit à des énoncés quantifiés. On peut ignorer quelle fût la dernière phrase écrite par Pascal, et pourtant vouloir, et même être justifié, à affirmer :

La dernière phrase écrite par Pascal est vraie

Cette phrase peut être analysée comme la forme que doit prendre la généralisation sur :

– Si la dernière phrase écrite par Pascal est « la neige est blanche », alors la neige est blanche
– Si la dernière phrase écrite par Pascal est « la neige est noire » alors la neige est noire
– Si la dernière phrase écrite par Pascal est ...
– etc.

[47] Certains auteurs en ont douté, voir par exemple au début de Ketland (1999), ou de façon plus développée dans Rivenc (2008), chap.1.

philosophie « extensionaliste » de Quine, pas plus que les notions de propriétés et d'attributs eux-mêmes. Quant à l'idée que les énoncés du type de

« La neige est blanche » est vraie si et seulement si la neige est blanche

gouverneraient la signification du prédicat de vérité, seraient analytiques ou exprimeraient des « équivalences cognitives », elle est tout aussi étrangère à la pensée de Quine. La notion d' « équivalence cognitive » n'a certainement pas les titres béhaviouristes pour figurer dans une explication quinienne, et l'on sait le doute que Quine a jeté sur la notion d'analyticité. C'est l'idée même qu'un ensemble d'énoncés (une disposition à les accepter) pourrait gouverner la signification (notre compréhension) d'un terme qui n'a pas de sens immédiat ici. Quant à la thèse selon laquelle « tout ce qu'il y aurait à dire sur la vérité » suit de l'ensemble des équivalences-T, c'est-à-dire de l'ensemble des énoncés du type de celui cité à l'instant (nous y reviendrons), d'une part c'est une thèse vague, et d'autre part, surtout, Quine suit Tarski ici[48], et souscrit à l'entreprise d'une recherche de *définition* de la vérité.[49]

L'idée que Quine ne souscrirait à la lettre à aucun des slogans déflationnistes pourrait accréditer la thèse d'un Quine étranger au déflationnisme. Et pourtant Quine est, j'ai envie de dire *évidemment*, déflationniste. Il faut en effet, je crois, garder à l'esprit que la plupart des thèses déflationnistes contemporaines ont un statut *problématique* et non catégorique. Il s'agit surtout de *tentatives* pour donner une caractérisation de la notion de vérité qui soit en harmonie avec le genre d'affirmations que le déflationniste entend faire relativement aux usages ou aux entreprises scientifiquement illégitimes mettant en jeu le concept de vérité : la recherche d'une

[48] Voir chapitre suivant.

[49] Comme le note Rivenc (2008), Quine a insisté à la fois (1) sur la légitimité d'une recherche d'une définition de la vérité et, paradoxalement, (2) sur l'impossibilité de mener complètement à bien cette recherche. Ces faits parlent-ils contre la thèse d'un Quine déflationniste ? Je fais une remarque simplement sur le point (2). Dans une théorie ne contenant, pour toute constante logique, que des connecteurs vérifonctionnels, les quantificateurs ne peuvent pas être définis explicitement. Plus généralement, si une expression est supposée introduire un surcroît de pouvoir expressif dans un langage, il faut certainement qu'elle ne soit pas définissable dans ce langage, puisque sa définissabilité montrerait que tout ce qui est exprimable avec elle est exprimable sans elle. Or précisément, Quine insiste sur le gain expressif qui résulte de l'introduction d'un prédicat de vérité aux langages qu'il a en vue, les langages enrégimentés dans la logique du premier ordre. L'indéfinissabilité de la vérité dans ces langages ne devrait donc pas être une surprise. Donc, sauf à contester que reconnaître le pouvoir expressif d'un concept soit compatible avec une vélleité de déflationnisme, il ne semble pas y avoir ici de raisons définitives de penser que l'indéfinissabilité de la vérité pourrait étayer l'idée que Quine n'est pas déflationniste.

théorie de la correspondance elle-même, l'idée que les attributions de contenus intentionnels et sémantiques, c'est-à-dire de conditions de vérité à des croyances ou à des émissions verbales, pourrait jouer un rôle dans le langage de l'explication scientifique, causale, des comportements. Or Quine, et c'est le sens même de son analyse de la vérité, est du côté du déflationnisme ici. Il faut donc dire, je crois, que Quine *est* déflationniste, même s'il est un déflationniste prudent, ou plutôt protégé des errances de la recherche d'une caractérisation du statut de la notion de vérité par la sobriété de sa méthodologie et l'austérité du langage dans lequel il a circonscrit le domaine de ce qui est scientifiquement dicible.

Quoi qu'il en soit de ce dernier point, les réflexions de Quine sur la vérité constituent sans aucun doute l'héritage majeur sur lequel s'est construit le déflationnisme contemporain, au cœur duquel on retrouve les quatre thèses suivantes :

1. La vérité est une notion interne/immanente.
2. Les équivalences-T expliquent de façon complètement claire le sens des attributions de vérité à des énoncés individuels, et cette explication rend compte de l'intuition de la vérité comme correspondance.
3. Le prédicat de vérité est un outil de décitation des énoncés cités. Plus généralement, il permet d'annuler la référence au langage dans les contextes de montée sémantique.
4. Le prédicat de vérité permet d'exprimer des ensembles infinis d'énoncés, des généralisations.

1.5 Le déflationnisme de H. Field

Pour conclure cette brève incursion dans l'histoire du déflationisme, je voudrais enfin présenter les grandes lignes du travail de Hartry Field, qui est, avec Paul Horwich, une figure majeure du déflationnisme contemporain.[50] Le travail de Har-

[50] Comme je l'ai dit, je ne suis guidé ici par aucun souci d'exhaustivité. En dépit de certaines différences, les réflexions de Horwich et Field sur la vérité sont proches, et pour cette raison même j'ai choisi de me concentrer sur un seul de ces deux auteurs. Plusieurs raisons m'incitent à me concentrer sur le travail de Hartry Field : Field s'inscrit explicitement dans la filiation de Quine ; son déflationnisme est pensé d'emblée comme un déflationnisme sémantique général (au sens que nous avons donné à ce terme), et Horwich travaille avec une notion idiosyncrasique de proposition. Les travaux de Paul Horwich sur le déflationnisme sont répartis essentiellement en deux volumes Horwich (1998*a*) et Horwich (1998*b*). Les travaux principaux de Hartry Field sur le déflationnisme sont réunis dans Field (2001). Pour une brève mais instructive comparaison, on pourra consulter Field (1992).

try Field s'inscrit, comme celui de Quine, dans la postérité critique de l'empirisme logique. À la fois physicaliste et se réclamant du naturalisme méthodologique, les positions de Field se rapprochent de celles de Quine tout en s'en éloignant sur des points importants : Field, pourrait-on dire, c'est Quine sans le béhaviourisme, sans l'extensionnalisme[51], et avec une théorie computationnelle de l'esprit. Ces raffinements ou libéralisations des critères de scientificité permettent à Field d'avancer un peu plus loin sur la voie d'un déflationnisme sémantique global en donnant des comptes rendus apparentés des rôles respectifs des attributions de vérité et de conditions de vérité dans les explications.

Vérité et conditions de vérité Il y a un lien entre la notion de vérité et ce que, suivant Ramsey, j'appellerai la *référence propositionnelle* en jeu dans les attributions de contenu, ou de signification, à un porteur de vérité. En effet, dire d'une pensée qu'elle a la signification qu'elle a, qu'elle est une pensée *que la neige est blanche*, ou d'un énoncé qu'il est un énoncé signifiant *que l'herbe est verte*, ce n'est rien d'autre que lui attribuer certaines *conditions de vérité*. Avoir une pensée que p (ou affirmer que p) c'est avoir une pensée (faire une affirmation) qui est vraie si et seulement si p. Dans une autre terminologie, on dira également qu'attribuer des conditions de vérité c'est attribuer un contenu *représentationnel* à un porteur de vérité, dire ce *à propos* de quoi il est. Il existe donc un lien conceptuel étroit entre attributions de vérité et attributions de contenu représentationnel (ou sémantique, intentionnel, propositionnel). Pour Field, et c'est également ce qu'aurait dit Ramsey, un déflationniste qui se contenterait d'analyser les attributions de vérité à des contenus propositionnels individués en termes de conditions de vérité n'aurait fait que la moitié du chemin. Ce que Field conteste, c'est l'idée que non seulement les attributions de vérité, mais aussi de conditions de vérité, ou plus généralement de propriétés sémantiques, puissent fonctionner dans des explications méthodologiquement impeccables *en vertu* de leur propriétés représentationnelles.

Le contexte historico-philosophique. On ne comprendra pas le développement historique et le succès des thèses déflationnistes si on ne les replace pas dans le contexte qui était celui du dernier tiers du vingtième siècle, un moment où une partie considérable de l'attention philosophique était concentrée sur le problème d'expliquer, d'une façon qui soit compatible avec les thèses empiristes et naturalistes, les notions de signification (en philosophie du langage) et de contenu propositionnel

[51] En fait d'abandon de l'extensionnalisme, Field tient les notions de causalité (physique) et de propriétés pour des notions scientifiquement acceptables.

(en philosophie de l'esprit).

Le point de départ philosophique, pour le dire brièvement, est le suivant. Le discours pré-scientifique de la psychologie ordinaire fait un usage abondant d'explications *intentionnelles* (et sémantiques) pour rendre compte d'événements que l'on peut décrire en termes purement physiques. J'explique qu'un individu s'est mis à courir derrière une automobile en lui attribuant des états mentaux complexes, par exemple qu'il *désirait* être à l'heure à son rendez-vous, qu'il *croyait que* l'automobile était un taxi libre, qu'il avait une croyance *vraie* relativement à la localisation de la station, etc. De même, pour expliquer un certain nombre de phénomènes observables et descriptibles en termes non-intentionnels, nous supposons d'ordinaire qu'il existe une communication entre individus et que les véhicules principaux de cette communication sont des significations attachées à des émissions verbales, écrites, etc. La question est de savoir si les explications mettant en jeu des propriétés sémantiques peuvent survivre aux exigences d'une méthodologie scientifique de l'explication : les notions sémantiques et les notions de la psychologie intentionnelle sont-elles *naturalisables* ? Les notions sémantiques peuvent-elles figurer dans des lois de la nature ? Les explications sémantiques sont-elles compatibles avec le matérialisme ? Les propriétés sémantiques ont-elles une efficacité causale ? [52] Pour Quine, et pour les notions apparentées à celle de signification, nous l'avons vu, la réponse est non : la notion de signification (et donc de contenu intentionnel propositionnel) n'a pas sa place dans le langage de la science. Mais l'émergence dans les années soixante et soixante-dix de théories de l'esprit plus sophistiquées que le béhaviourisme quinien permet de réouvrir le débat à nouveaux frais.

De nombreux auteurs ont insisté sur la valeur des explications sémantiques et intentionnelles : ces explications, les seules que nous possédions la plupart du temps pour comprendre ou prédire un comportement, ont suffisamment prouvé leur efficacité. Ce constat de succès devrait nous guider dans notre recherche d'une théorie de l'esprit : cette dernière doit être compatible avec l'efficacité de la théorie intention-

[52] La question de savoir à quelle condition une propriété peut compter comme naturelle est sujette à dispute. Les néo-positivistes demandaient que ces propriétés soient réductibles à des propriétés physiques ou observables. Mais ce physicalisme strict n'a plus beaucoup de partisans. Une version faible est simplement qu'une notion compte comme naturelle si elle figure dans des lois causales établies selon les canons de la méthodologie scientifique, condition en général assortie d'une clause de compatibilité avec la vérité du matérialisme, par exemple sous la forme d'une exigence de survenance sur des propriétés physiques. Voir par exemple Fodor (1974) pour une défense d'un matérialisme non réductionniste, et Kim (1989) pour une discussion critique de cette position. Pour une présentation générale des débats et des notions en jeu, on pourra consulter Kim (1996).

nelle naïve et doit permette de réduire, en un sens plus ou moins fort, le vocabulaire intentionnel au vocabulaire physique, de sorte qu'une attribution de contenu intentionnel ou sémantique doive fonctionner, bon an mal an, comme une explication causale, indiquant l'existence d'une relation « naturelle » (au sens d'espèce naturelle) entre un individu et son environnement.[53] Ces considérations, en combinaison avec d'autres portant sur les propriétés de systématicité et de productivité de l'activité de l'esprit, ont conduit une partie de la communauté philosophique à adopter une théorie computationnelle de l'esprit et l'hypothèse d'un langage de la pensée. Selon ce modèle, en un mot, le cerveau peut être décrit comme un ordinateur traitant certains *items* syntaxiques, les *représentations mentales*, dont les rôles causaux engendrent des changements d'états appropriés. Ce qu'il faut noter maintenant, c'est qu'une théorie de la vérité comme « correspondance », si elle était formulée dans des termes méthodologiquement acceptables, s'intégrerait naturellement à la réalisation de ce programme de naturalisation. Comprise comme une relation causalement déterminée entre des représentations mentales et l'environnement d'un sujet, elle ouvrirait la voie à une naturalisation des relations qui semblent être décrites par les attributions de notions sémantiques en général, qu'il s'agisse de vérité ou de contenu vériconditionnel. Une théorie de la vérité comme correspondance a donc parfois été vue comme le premier pas vers la légitimation d'un discours scientifique en termes intentionnels ou représentationnels, qui permettrait de fonder le pouvoir explicatif des attributions de notions sémantiques par leur réduction (il peut s'agir d'une réduction en un sens relâché, une réduction fonctionnelle par exemple) à des propriétés physiques ou dont l'efficacité causale n'est pas problématique. Si avoir une croyance vraie c'est être dans une certaine relation causale avec son environnement (relation que la théorie de la vérité correspondance doit expliquer), pourquoi les attributions de vérités ne seraient-elles pas des explications naturelles *bona fide*, au bout du compte ? Si j'explique que Pierre est à l'heure parce qu'il avait une croyance vraie (relative à l'heure du rendez-vous, le temps que lui prendrait le trajet etc.), est-ce que je ne suis pas véritablement en train de donner une explication *causale* de sa présence au lieu de rendez-vous, à l'heure dite ?[54] A l'opposé de ce programme,

[53] Un des plus célèbres défenseurs de cette méthodologie a sans doute été Jerry Fodor. Voir par exemple Fodor (1987). Pour un panorama critique des principales tentatives de naturalisation des notions sémantiques, on pourra consulter Loewer (1997).

[54] Il ne faut pas confondre l'idée que la vérité est une propriété causalement efficace et l'idée selon laquelle la vérité de l'énoncé (la pensée etc) « la neige est blanche » est *causée* par la blancheur de la neige. Voir Douven et Hindriks (2005) pour une discussion critique de cette dernière idée.

l'autre option théorique semblait devoir être une position *éliminativiste* en matière d'attribution de vérité et de contenu, position qui vouerait le discours sémantique et intentionnel au non-sens.[55] Selon cette seconde conception, les notions sémantiques et intentionnelles sont non-dénotantes, et les explications correspondantes sont simplement fausses. Les seules explications à caractère scientifique relatives à des états mentaux sont, ou seront, des explications neuro-psychologiques, et il n'y a aucune raison de penser que les explications intentionnelles puissent être réduites à ces dernières. Un des mérites du déflationnisme, je vais l'esquisser brièvement, est d'ouvrir une troisième voie dans ce débat.

De la réduction à la déflation. Le socle de la réflexion de Field sur la vérité et, plus généralement, sur les propriétés sémantiques et intentionnelles, est donc profondément ancré dans ses engagements physicalistes et le paysage philosophique que je viens d'esquisser à grands traits. Si ces engagements n'ont pas varié, ses réflexions sur les notions sémantiques ont connu une inflexion importante dans les années quatre-vingt. D'abord ardent défenseur d'une théorie « robuste » de la vérité comme correspondance, cherchant à réduire cette notion de correspondance au moyen de la « théorie » causale de la référence alors naissante, Field a par la suite embrassé le paradigme déflationniste, au point d'en devenir le représentant le plus remarquable. Pour comprendre ce changement, il faut revenir à la menace que constituent les notions sémantiques pour le physicalisme. Le physicalisme est faux si les deux conditions suivantes sont vraies : 1. Les notions sémantiques jouent un rôle dans les explications causales. 2. Ces notions ne sont pas réductibles à des propriétés physiques.

Le changement intervenu dans le cours du développement philosophique de Field correspond à un renversement dans l'appréciation de la vérité de ces deux conditions. Field pensait dans les années soixante-dix que les notions sémantiques et intentionnelles jouaient un rôle explicatif important et qu'il serait possible de dire pourquoi en étudiant les liens de causalité existant entre l'occurrence de certaines représentations mentales et certains événements de l'environnement des locuteurs. En un mot, la réduction des notions sémantiques devait s'opérer par une analyse du contenu des représentations mentales (linguistiques) en termes de leur rôle conceptuel « étendu », lequel est donné non seulement par la corrélation de leurs occurrences aux occurrences d'autres représentations mentales associées (leur « rôle

[55] Après Quine, et de façon plus radicale encore, c'est le genre de position défendue par Paul et Patricia Churchland, voir par exemple Churchland (1984), Churchland et Churchland (1998).

inférentiel »), mais également par la corrélation de leurs occurrences à des occurrences de certains événements ou objets de l'environnement distant.[56] Le fait pour un sujet S d'avoir une croyance que p est alors analysé de la façon suivante : S a une croyance que p si et seulement si X est une représentation mentale que S croit*, et X signifie que p. *Croit** est une relation entre un sujet et une entité syntaxique qui ne pose pas de problème particulier à un physicaliste[57] tandis que « signifie que p » doit être analysé en termes de rôles conceptuels étendus, dans l'esprit de ce qui a été esquissé plus haut.[58]

Au milieu des années quatre-vingt,[59] puis de façon plus affirmée dans les décennies qui ont suivi, Field renverse les choses : d'une part, il ne croit plus à la promesse de la théorie causale de la référence[60], ni plus généralement à la possibilité de réduire les explications par attribution de vérité ou de conditions de vérité à des explications causales, mais d'autre part, et surtout, il ne croit plus non plus qu'*une telle*

[56]La signification du terme de Mentalais « vache » est constituée par la corrélation entre l'apparition d'une vache dans l'environnement d'un locuteur et l'occurrence du terme en question dans son cerveau. Différentes tentatives de développement d'une sémantique naturalisée des représentations mentales existent, parmi lesquelles les plus proches de ce qui vient d'être esquissé sont celles défendues dans Dretske (1981) et Fodor (1987). Voir à nouveau Loewer (1997) pour des références supplémentaires et un bilan critique de ces tentatives. Pour une discussion de la sémantique des rôles conceptuels et ses relations avec la sémantique vériconditionnelle, voir Harman (1982), Loar (1982). Greenberg et Harman (2006) est un bilan plus récent. Field (1977) contient également une discussion de cette question. Voir enfin le classique Fodor (1975) pour un argument en faveur de l'hypothèse d'un langage de la pensée (le « Mentalais »).

[57]L'idée répandue étant que l'on peut distinguer fonctionnellement l'attitude de croyance en termes de rôle causal dans une théorie des états intentionnels. Typiquement, la croyance et le désir sont tenus pour causer ensemble certaines actions, et on peut regarder une théorie bayesienne de la rationalité comme une théorie, certes hautement idéalisée, de ces relations.

[58]La réduction de la vérité des énoncés à la dénotation des termes singuliers *et généraux*, et de la dénotation à une relation causale est présentée dans Field (1972). Field (1978) replace ce travail dans le contexte du problème de la naturalisation des explications intentionnelles, *via* ce que Field appelle le *problème de Brentano* : donner une explication naturaliste des états de type croyance que p, désir que p, etc., étant donné qu'ils sont de prime abord des relations entre des organismes et des entités intentionnelles comme les propositions.

[59]Le tournant est pris dans Field (1986).

[60]Le problème de la référence aux objets abstraits semble avoir été déterminant dans le changement de position de Field. En une phrase, le problème est qu'il semble légitime de dire que « 5 » dénote le nombre 5, mais qu'on ne voit pas bien comment rendre compte en termes causaux de cet usage du terme « dénoter », l'objet dénoté par « 5 » n'étant pas situé dans l'espace et le temps. Une seconde source de difficulté est liée à l'indétermination de la référence, dans sa reconstruction fieldienne. Field (1972) s'était déjà donné beaucoup de peine pour trouver une utilité aux termes sémantiques dans la méthodologie physicaliste en dépit de ce qu'aucun fait physique (au sens large, incluant des faits de causalité) ne semble pouvoir permettre de trancher entre, disons pour simplifier, la thèse que « lapin », « chat », etc. tels que je les utilise réfèrent respectivement à l'ensemble des lapins, des chats etc. et la thèse que « lapin », « chat », etc., tels que je les utilise réfèrent à l'ensemble des parties non-détachées de lapin, à l'ensemble des parties non-détachées de chats, etc.

réduction soit nécessaire d'un point de vue physicaliste. Pourquoi ? Parce que les notions sémantiques sont désormais comprises comme ne jouant dans les explications qu'un rôle secondaire bien particulier : d'une part le prédicat de vérité (ou la relation de dénotation) n'a, quand il apparaît dans les explications, que le rôle de permettre l'expression d'une certaine généralité, et non de désigner une certaine propriété causalement efficace, d'autre part les explications sémantiques et intentionnelles, qui fonctionnent par attributions de contenu ou de signification, c'est-à-dire de « conditions de vérité » si l'on veut, à des croyances, à des énoncés, sont des explications d'un genre très particulier, que Field conçoit comme des explications *projectives* (nous allons y revenir). Or, comme l'ont noté Stephen Leeds et Hilary Putnam[61], si le physicalisme requiert des propriétés auxquelles un rôle *causal* est prêté dans les explications qu'elles soient expliquées en termes physiques, cette exigence ne porte *que* sur ces propriétés-là. Ce changement radical de stratégie de Field, c'est le passage d'une tentative de *réduction* à une tentative de *déflation* des notions sémantiques.

Field sur les attributions de vérité et de conditions de vérité. Je vais maintenant présenter plus en détail la seconde position de Field sur les attributions de vérité et de conditions de vérité. L'analyse que donne Field de la notion de vérité est très proche en esprit de celle donnée par Quine ; je serai donc bref. Selon Field, le prédicat de vérité est tel que les deux membres des équivalences-T du type

« La neige est blanche » est vrai si et seulement si la neige est blanche

sont *cognitivement équivalents*. La notion d'équivalence cognitive n'est pas complètement claire mais, *grosso modo*, elle signifie que les rôles que jouent les deux énoncés dans la pensée sont interchangeables. En second lieu, et sur ce point sa théorie de l'esprit éloigne Field de Quine, les porteurs de vérité sont les énoncés *individués par leur rôle conceptuel*, ou « computationnel », et *non pas individués orthographiquement*.[62] Cette idée permet à Field d'adopter une conception « linguistique » des attributions de contenu, selon laquelle dire :

[61] Voir Leeds (1978) et Putnam (1978*a*).

[62] Par conséquent, la meilleure traduction du biconditionnel précédent en anglais serait sans doute :

« Snow is white » is true if and only if snow is white

et non :

« La neige est blanche » is true if and only if snow is white

S signifie que la neige est blanche

c'est dire quelque chose comme :

S signifie la même chose que
l'énoncé « la neige est blanche » (tel que je le comprends).

L'idée principale est qu'une attribution de signification par un locuteur à un certain énoncé ne relie pas un énoncé et un objet indépendant de l'esprit (une proposition), mais seulement un énoncé à un énoncé[63] du langage du locuteur qui fait cette attribution de signification. De même, l'attribution de contenu de croyance

Pierre croit que la neige est blanche

doit se comprendre comme signifiant que Pierre croit un énoncé qui signifie la même chose que l'énoncé « la neige est blanche » tel que je le comprends. Moyennant cette analyse, Field peut réhabiliter la formulation propositionnelle, souvent jugée plus naturelle, des équivalences du type :

Il est vrai que p si et seulement si p

chacune étant est en fait équivalente à quelque chose du type :

« p » est $\text{vrai}_{\text{dans-mon-langage}}$ si et seulement si p

Discuter en profondeur des mérites et des faiblesses de cette théorie linguistique des attributions de contenus me mènerait trop loin.[64] Je vais simplement illustrer

On pourrait vouloir utiliser des guillemets particuliers pour indiquer que c'est cette individuation des énoncés qui est visée, réservant les guillemets usuels de citation pour la formation de noms d'individus syntaxiques. Cette fidélité à Quine me paraît un peu servile. Je doute que, dans l'usage courant, les énoncés entre guillemets de citations soient des noms d'individus syntaxiques. Est-il vraiment incorrect d'écrire :

> Dans son discours de réception du prix Nobel de la Paix, Barack Obama a dit, *je cite* : « On a vu naître la conception d'une « juste guerre », ce qui laissait à penser que la guerre n'était justifiée que lorsque certaines conditions étaient remplies : si on s'y résolvait en dernier recours, ou en cas de légitime défense ; si la force employée était proportionnelle ; et si, chaque fois que possible, on épargnait les populations civiles. » ?

Il existe un important champ de recherche sur les différents usages des guillemets de citation. Pour une introduction et des références, voir Cappelen, H, et Lepore, E, « Quotation », The Stanford Encyclopedia of Philosophy (Winter 2009 Edition), Edward N. Zalta (ed.), URL = *http ://plato.stanford.edu/*.

[63] A nouveau : dont le type est individué par son rôle conceptuel.

[64] Voir Field (2001), chap. 5, pour une présentation plus approfondie.

par un exemple la façon dont elle soutient la thèse de Field relative au statut des explications sémantiques/intentionnelles.

Tout d'abord, concernant les attributions simples de vérité, Field commence par remarquer, à la suite d'autres déflationnistes et des réflexions de Quine sur leur rôle expressif, que leur occurrence dans les explications semble n'avoir pour fonction *que* l'expression d'une certaine généralité, même dans les contextes philosophiques où le terme semble de prime abord véhiculer une charge métaphysique importante. Un bon exemple est la définition du réalisme métaphysique, ou disons une définition possible :

> Il se peut que des énoncés de notre langage soient vrais et que nous n'ayons jamais aucune raison de les croire.

L'impression que l'emploi du mot « vrai », ici, véhicule une charge métaphysique importante ne serait en fait qu'une de ces illusions qu'une analyse critique du langage peut dissiper. Or cette analyse révèle que nous n'avons affaire ici qu'à une illustration typique d'un usage du prédicat de vérité dans un contexte de montée sémantique rendu nécessaire par un besoin de généralité, comme nous en avons déjà vu à la section précédente : si le nombre d'énoncés de notre langage était fini, nous n'aurions nul besoin du prédicat de vérité ici, une simple disjonction ferait l'affaire.[65] Dans la formulation de la thèse du réalisme, l'usage du prédicat de vérité vient simplement répondre à un besoin logique.

Quid maintenant des explications dans lesquelles sont attribuées des conditions de vérité, par exemple à des états mentaux comme dans les explications intentionnelles ? Ces explications sont, remarque Field, tout à fait particulières. Elles ont en effet un caractère *projectif*, au sens où elles semblent faire référence au sujet qui produit l'explication. Qu'est-ce qu'une explication projective ? Supposons, pour reprendre un exemple de Field, que j'explique les symptômes d'un malade en disant : « Je parie qu'il a le virus que j'ai eu la semaine dernière ». Cette explication est une explication causale, si l'on veut, au sens où elle assigne une cause à des symptômes : un certain virus. Mais cette explication ne satisfait pas les critères méthodologiques de l'explication scientifique, laquelle interdit la référence à la première personne pour

[65] Dans le cas limite où notre langage ne comporte qu'un énoncé, disons « la neige est blanche », la thèse du réalisme se reformule simplement comme :

> Il se peut que la neige soit blanche et que nous n'ayons jamais aucune raison de le croire.

Dans cette reformulation, la notion de vérité n'apparaît pas.

l'identification de ses objets. Le problème est que mon ignorance, mon incapacité à identifier le virus autrement que par référence à moi-même, m'interdit de fournir une explication « authentique ». Plutôt qu'une explication, j'ai fourni quelque chose qui s'apparente à un *schéma d'explication*, un schéma qui reste à compléter. En quoi consisterait ce complément ? Simplement en une identification du virus en question et une description de ses propriétés physico-physiologiques expliquant causalement les symptômes observés. Au bout du compte, lorsque cette description est à ma disposition, je suis en mesure de proposer une explication authentique des symptômes observés qui ne fasse aucun détour par la référence à mon état la semaine passée. L'idée de Field est que les attributions d'états intentionnels fonctionnent sur le même modèle. Lorsque nous expliquons des comportements, qu'ils soient humains ou animaux, en donnant des raisons (« Félix est arrivé parce qu'il croyait que sa pâtée était servie », etc.) nous cherchons à donner un sens au comportement de l'agent *de notre point de vue*.[66] Il s'agit donc, plutôt que d'explications causales à proprement parler, de schémas d'explications causales. Et l'idée de Field est qu'une fois ces schémas complétés et transformés en d'authentiques explications d'où le point de vue de celui qui explique a disparu, les notions intentionnelles et sémantiques auront également disparu.

Mais les explications causales complètes que recouvrent ces schémas d'explications que sont les attributions d'états intentionnels ne sont nullement uniformes et recouvrent des réalités très différentes. Lorsque j'attribue au chat Félix la croyance *que sa pâtée est prête* pour expliquer sa venue, je n'entends pas que les mécanismes causaux sous-jacents sont analogues à ceux qui sont à l'œuvre derrière l'explication de la la venue de Marie par référence à sa croyance *que le train partait à huit heures*. On peut reconstruire ce que serait, selon Field, l'explication complète dans ce dernier cas, afin d'illustrer le processus de disparition de la référence propositionnelle. Lorsque nous disons :

Marie est venue parce qu'elle croyait que le train partait à huit heures

il faut en fait comprendre l'attribution de contenu selon le modèle donné plus haut :

Marie est venue parce qu'elle croit un énoncé dont le rôle fonctionnel est similaire au rôle que joue chez moi l'énoncé « le train part à huit heures ».

[66] Voir Field (2005), p.111.

La véritable explication de la venue de Marie s'obtient alors par une description directe du rôle fonctionnel d'une certaine représentation mentale, plus précisément un énoncé de Mentalais, dans la tête de Marie. C'est la description de ce rôle fonctionnel qui prend en charge les rapports entre les occurrences de certains événements mentaux dans la tête de Marie et l'environnement. La possiblité d'une description du rôle fonctionnel d'un certain énoncé de Mentalais est une hypothèse largement théorique. Pour donner une idée de ce dont il est question, une telle description donnera à la fois la corrélation causale (sous une forme statistique par exemple) entre l'occurrence de certains mots ou certains énoncés « dans la tête » et l'occurrence de certains événements extérieurs (entre l'énoncé « il pleut » et la présence de pluie par exemple, ou entre l'occurrence du mot « vache » et la présence d'une vache), les corrélations entre l'occurrence de cet énoncé et l'occurrence consécutive ou simultanée d'autres énoncés, et des corrélations entre occurrences de cet événement et des réponses motrices ou cérébrales spécifiques. Le pari est donc que, si nous avions une description complète du rôle fonctionnel de l'énoncé « Le train part à huit heures » chez Marie, nous aurions une explication causale, non intentionnelle, non sémantique, de sa venue.[67] La conséquence de tout cela c'est aussi que, puisque les idiomes intentionnels et sémantiques n'ont pas de rôle réel dans les explications causales complétées, la réduction de ces propriétés n'est pas nécessaire du point de vue physicaliste. Au bout du compte, et nous retrouvons un motif que nous avons déjà développé dans le cas spécial du prédicat de vérité, il ne faudrait pas conclure que l'idiome sémantique/intentionnel est *inutile*. En fait il est *indispensable* lorsque, comme c'est si souvent le cas, nous voulons donner une explication (causale) du comportement d'un individu, d'un animal, d'une machine peut-être, tout en ignorant le détail des mécanismes causaux sous-jacents à ce comportement. Mais aller plus loin dans cette présentation serait aller trop loin.[68]

[67] A nouveau, quant aux mécanismes causaux spécifiques vers l'existence desquels pointent les attributions d'états mentaux propositionnels, il ne faut pas supposer qu'il y ait uniformité des processus recouverts par l'attribution d'un contenu intentionnel donné dans différentes circonstances. Si c'était le cas, alors nous aurions l'indication d'une *réduction* de ce dernier aux premiers, ce qui est précisément ce dont Field rejette la possibilité. Parce que les explications intentionnelles sont projectives, il n'y a aucune raison de postuler une uniformité des mécanismes causaux spécifiques vers l'existence desquels pointent différentes occurences d'attributions d'un contenu donné.

[68] En particulier, il faudrait vérifier qu'en effet la science n'a pas besoin du concept substantiel de vérité-correspondance. La linguistique est sans doute la seule science où la vérité joue un rôle proéminent dans le discours théorique. Reste qu'il n'est pas sûr que les notions de vérité et de référence dont le linguiste a besoin soient des notions dénotant des relations causalement déterminées de correspondance entre des items linguistiques et le monde. Pietroski (2005), par

Conclusion. Parmi les auteurs que nous avons rencontrés, Field est le seul à revendiquer l'appellation « déflationniste ». Et au cœur du déflationnisme de Field, il y a l'articulation deux thèmes. Le premier thème est celui de l'absence de rôle explicatif des notions sémantiques. Le second est celui du rôle des notions sémantiques pour l'expression de la généralité. En particulier, le prédicat de vérité est indispensable pour l'expression de certaines généralisations, mais s'il peut figurer à ce titre dans les explications, il n'y figure qu'à ce titre : c'est parce qu'il est compris comme un outil logique, que son rôle dans les explications peut être revu à la baisse. C'est un rôle qui ne fait pas de ces notions des notions authentiquement explicatives dans la méthodologie physicaliste.[69]

exemple, cite Chomsky :

> Pour autant qu'on sache, il n'est pas plus raisonnable de chercher une chose-dans-le-monde désignée [*picked out*] par le mot « rivière », ou « arbre » ou « eau » ou « Boston » que de chercher une collection de mouvements de molécules désignée [picked out] par la première syllabe ou dernière consonne du mot « Boston ». Avec suffisamment d'héroisme, on peut défendre ces thèses, mais elles semblent n'avoir pas de sens du tout. (cité dans Pietroski (2005))

Pour une réflexion critique plus générale sur le paradigme des explications en termes de conditions de vérité en linguistique, inspirée des variations chomskyennes sur ce thème, nous renvoyons le lecteur à Pietroski (2005) et Stainton (2006).

[69]Cette façon de voir les choses soulève incidemment la question de savoir ce que l'on doit entendre par le terme « sémantique ». Nous avons d'une part ces notions auxquelles j'ai réservé, suivant une certaine tradition, le nom de « notions sémantiques » : la vérité, la référence, la signification, etc. La sémantique, dans l'esprit qui préside à cette terminologie, est la science qui s'occupe de comprendre les relations du langage au « monde » *et* dont les concepts principaux sont les notions « sémantiques » que je viens de mentionner. Pour un déflationniste comme Hartry Field, ces concepts tels qu'on les conçoit traditionnellement sont inappropriés pour comprendre les relations du langage (ou de la pensée) au monde. Ce point parfois source de confusions. L'erreur est de croire que le déflationniste ne peut expliquer que le langage « parle du monde », ou qu'il aurait renoncé à l'étude des relations des expressions au monde, ou encore à expliquer l'idée de « contenu » d'un énoncé, d'une pensée. Mais, nous venons de le voir, c'est le contraire qui est vrai. Parce que le déflationniste prend au sérieux la tâche d'expliquer ces relations et ces notions, il procède à un inventaire critique des outils conceptuels dont nous disposons pour les étudier. Cette réflexion méthodologique ne conduit pas le déflationniste à nier ou à renoncer au projet d'une sémantique, entendu en un sens large, mais c'est une sémantique dont les concepts fondamentaux ne sont plus ce que j'ai appelé les « notions sémantiques ». Ce sont les notions de « relation d'indication », de « corrélations », de « dépendance asymétrique », de « rôles inférentiels » etc., qui doivent permettre de comprendre les mécanismes causaux associés à l'usage du langage et de la pensée dans ses formes supérieures. En ce sens, et si le projet de la sémantique n'est autre que l'étude de la façon dont la pensée et le langage se rapportent au monde, de ce que c'est que « parler du monde », il y a bien une « sémantique » déflationniste. Pour éviter les confusions, néanmoins, et cette mise au point faite, je m'en tiendrai ici et dans la suite à ma première terminologie : je conserverai le terme de « sémantique » pour la science dont les concepts principaux sont les concepts ordinairement décrits comme « sémantiques ».

1.6 Conclusion

J'ai essayé dans ce qui précède de dégager les grandes lignes d'une histoire stylisée[70] du déflationnisme. Cette histoire commence avec les remarques isolées de Frege sur la transparence des usages prédicatifs de la notion de vérité : attribuer la vérité à un énoncé ne semble rien ajouter à ce qui est exprimé par cet énoncé. Frege conclut que la vérité n'est pas une propriété « au sens usuel ». Ces remarques contiennent en germe ce que j'ai appelé au début de ce chapitre un déflationnisme *local*. Mais quelle est alors l'utilité du prédicat de vérité ? C'est à Ramsey qu'il revient d'avoir fait le lien entre les emplois du prédicat de vérité et l'expression d'une certaine forme de généralité, en montrant qu'il était possible de paraphraser en tout contexte les emplois du mot "vrai" en termes de quantificateurs et de variables occupant des positions d'énoncés. Ramsey embrasse une analyse déflationniste propositionnelle locale des emplois du prédicat de vérité, mais insiste dans le même temps sur le fait qu'une analyse complète de la notion de vérité doit également en passer par une explication de ce qui est en jeu dans les attributions de contenus de pensée, et plus précisément une explication de la référence propositionnelle. C'est là, en dernière instance, que se mesure selon Ramsey la « profondeur » de ce qui est dit à travers les attributions de vérité.

Mais les analyses de ce premier déflationnisme se trouvent ensuite transfigurées et radicalisées sous l'influence de Quine. Si Quine élabore à son tour un déflationnisme local amendé et précisé, où le rôle du prédicat de vérité pour l'expression de certaines généralisations est désormais mis en avant comme son utilité principale, il franchit un pas supplémentaire en adoptant l'idée que la notion de vérité est avant tout une notion *interne* et en adoptant parallèlement une attitude éliminativiste relativement à la référence propositionnelle et aux attributions de contenus. Or c'est désormais la place d'une notion dans notre meilleure explication du monde (le système des sciences naturelles) qui borne le discours philosophique légitime sur cette notion, rendant obsolète la distinction entre ce qui peut se dire et ce qui, sans pouvoir se dire, pourrait néanmoins se montrer : s'il n'y pas de place dans notre meilleure théorie du monde pour le discours sur la relation de correspondance, il n'y a pas de sens à insister sur l'idée que cette correspondance est bien « là ». Ce mouvement vers l'élaboration d'un déflationnisme sémantique général trouve son expression contemporaine dans les travaux de Hartry Field. Field reprend les

[70] Au sens où l'on parle de « faits stylisés » dans les sciences naturelles.

analyses déflationnistes locales de Quine relativement au prédicat de vérité en les radicalisant : non seulement le prédicat de vérité permet l'expression de certaines généralisations, mais il affirme explicitement que c'est là son seul rôle. Revenant sur l'éliminativisme strict de Quine en matière d'attribution de contenu, Field en propose une analyse en harmonie avec celle du prédicat de vérité : les attributions de contenu sémantique, comprises comme attribution de conditions de vérité à des énoncés, son légitimes et leur rôle est analogue à celui des attributions de vérité. Toutes les deux sont comprises comme des moyens d'expressions jouant un rôle logique dans des situations d'insuffisance épistémique : lorsque nous voulons exprimer notre accord avec un énoncé dont nous ignorons le contenu, lorsque nous voulons exprimer une infinité d'énoncés, lorsque nous voulons pointer vers une explication causale mais que nous ne sommes pas en situation d'identifier les mécanismes causaux en présence. Le déflationnisme émerge alors dans le débat contemporain comme une position originale dans le spectre des tentatives de naturalisation ou d'élimination des notions sémantiques : les attributions de notions sémantiques ont un sens, elles sont importantes et même indispensables dans de nombreuses situations ; mais ces attributions ne sont pas méthodologiquement comparables à des attributions de propriétés naturelles, ou causales-explicatives.[71]

C'est cette idée originale du déflationnisme que je retiens comme sa caractérisation centrale : défendre une position déflationniste relativement à une notion X, c'est défendre l'idée à la fois qu'elle est indispensable et que son rôle n'est pas central d'un point de vue explicatif.[72]

[71] Bien entendu, si les canons de la méthodologie naturaliste devaient conduire à la découverte de lois de la psychologie intentionnelle, les notions intentionnelles se retrouveraient *ipso facto* sur un pied d'égalité avec les espèces naturelles, et par ricochet également les notions sémantiques, comme la vérité, qui permettent de rendre compte des contenus intentionnels ; le déflationnisme s'en trouverait réfuté. C'est un test pour le déflationnisme que Field accepte. Voir Loewer (2005) et la réponse de Field dans Field (2005).

[72] Cette brève plongée dans le déflationnisme montre également qu'il faut être prudent lorsque, comme on le fait souvent, on caractérise le déflationnisme comme l'idée que la vérité n'est pas une propriété « substantielle ». Car ce que l'on veut dire par là est d'emblée, je l'ai dit en commençant, problématique. On l'a vu et on le reverra, il y a bien des sens dans lesquels on peut dire que la notion de vérité *est* « substantielle », et qui ne posent pas de problème à un déflationniste comme Hartry Field. En premier lieu, la vérité, nous l'avons vu, est substantielle au sens où elle est *indispensable* et *indéfinissable*. Elle est substantielle encore au sens où elle est *irréductible*, même en un sens de « réduction » moins exigeant que celui de définition : irréductible à des notions physiques, et dans un autre registre philosophique, nous le verrons, irréductible à la notion de « prouvabilité », comme on le verra au chapitre suivant. Nous pourrions mentionner aussi que la vérité pourrait être substantielle en tant que *norme* du discours, qui semble unique : demander la vérité, ce n'est pas seulement demander ce qu'il est justifié, ce qu'il est rationnel, ce qu'il est utile, ou ce qu'il est

Dans le cas de la notion de vérité, le seul qui retiendra mon attention ici, le déflationnisme apparaît donc comme la conjonction de deux thèses : la thèse selon laquelle la notion de vérité n'« a pas de pouvoir explicatif » et la thèse selon laquelle l'utilité du prédicat de vérité est de « permettre l'expression de certaines *généralisations* » (en association avec d'autres ressources logico-linguistiques comme les quantificateurs et des outils de description du langage). La suite de ce travail peut être vue comme un examen et un prolongement des thèses déflationnistes hors du domaine des sciences empiriques, dans le domaine de la connaissance *a priori*, mais sans préjuger toutefois que ces deux domaines de la connaissance soient indépendants. Plus précisément, mon projet est d'examiner quelques difficultés posées par ces thèses. Sur quel critère doit-on distinguer les usages « expressifs » et les usages « explicatifs » de la vérité ? Tracer une telle frontière est-il seulement possible ? Et de quels usages de la notion de vérité faut-il rendre compte exactement pour rendre compte de son « utilité dans l'expression de certaines généralisations » ? Faut-il, par exemple, qu'à partir d'une théorie de la notion de vérité, ou d'une définition de la notion de vérité, nous soyons en mesure de *prouver* certaines généralisations ? Si oui, lesquelles ? Avec quels moyens auxiliaires ? Si non pourquoi ? Et à supposer que les thèses précédentes puissent recevoir un contenu suffisamment précis, est-il certain que ces thèses soient seulement compatibles entre elles ? N'est-il pas possible, sans attendre le verdict des sciences empiriques, de donner une réfutation *a priori* des thèses déflationnistes ? Est-il seulement possible d'expliquer ou de définir la notion de vérité d'une façon qui permette de rendre compte de son usage dans « l'expression des généralisations » sans légitimer par là-même des emplois « explicatifs » de la notion de vérité ? C'est la tâche d'une *théorie de la vérité* que d'articuler les ressources théoriques qui rendent cet usage du prédicat de vérité possible, et qui sait quel rôle une telle théorie *a priori* ne permettrait pas à la notion de vérité de jouer dans les preuves, les explications ou les justifications, empiriques ou non, par exemple quand cette théorie est mise à contribution dans des contextes théoriques plus large, parmi d'autres théories physiques, biologiques, mathématiques ? Il y a là potentiellement un espace pour une

cohérent de croire. (Ce qui laisse la question intacte de savoir si la « vérité » joue réellement le moindre rôle normatif dans l'activité sociale ou cognitive. A-t-on besoin de faire appel à cette norme pour expliquer, par exemple, la pratique de l'assertion ? Sur ce point délicat, les avis sont partagés. Voir par exemple le débat entre Price (1998) et McGrath (2003), ou encore les remarques de Crispin Wright dans Wright (1999). Sur la question générale du rôle normatif de la la vérité voir par exemple Engel (2001) et Horwich (2006).)

réduction à l'absurde des thèses déflationnistes, et c'est cet espace que nous nous proposons d'explorer.

En résumé, nous cherchons à savoir s'il y a une explication de la notion de vérité, une explication de la distinction entre pouvoir expressif et pouvoir explicatif, et une explication de la classe des usages de la notion de vérité que le déflationniste juge essentiels, qui soient plausibles, conformes aux intuitions déflationnistes que nous avons mises en avant, et telles que l'affirmation que le prédicat de vérité saisit l'intuition classique de la correspondance, permet d'exprimer des généralisations *et* n'est pas une notion explicative, ne soit pas réfutable *a priori*. Le premier pas dans cette direction passe par une présentation du travail de Tarski sur la vérité.

Chapitre 2

Tarski et la vérité

> Je dis toujours la vérité. Pas toute, parce que toute la dire, c'est impossible, matériellement. Les mots y manquent. C'est même impossible parce que la vérité tient au réel.
>
> Lacan, Télévision, ouverture

Au chapitre précédent, j'ai présenté les grands axes d'une réflexion déflationniste sur la notion de vérité. D'un certain point de vue, on pourrait dire qu'un trait marquant de ces réflexions résidait dans ce que l'on pourrait appeler leur caractère *métaphilosophique* : il ne s'agissait pas tant de défendre une notion de vérité (la notion de vérité correspondance) ou de statuer sur la signification du mot « vrai », que de statuer sur l'importance et le rôle de la notion de vérité. A ce trait correspondait le fait qu'une partie des enjeux philosophiques de la discussion se trouvait déportée du problème de l'explication de la notion de vérité vers celui du rôle de la notion de vérité dans l'explication. D'un autre côté, l'évaluation des thèses déflationnistes doit au bout du compte en passer par une confrontation entre cette explication de la notion de vérité que le déflationniste suppose donnée et les (méta-)conclusions philosophiques, proprement déflationnistes, qu'il en tire. Pour préparer cette confrontation, il nous faut donc à présent revenir sur le problème de la clarification de la notion de vérité elle-même. C'est l'objet de ce chapitre.

Comment clarifier le sens de la notion de vérité ? Une telle clarification prendra idéalement la forme d'une *définition* explicite du prédicat « est vrai » ou, à défaut, d'une *théorie* dans laquelle sont couchés les principes essentiels de la vérité, c'est-à-dire les lois permettant de rendre compte des usages de la notion de vérité qui

sont consubstantiels à notre compréhension de la notion. Il est difficile de surestimer l'importance et l'influence du travail du logicien polonais Alfred Tarski dans l'accomplissement de cette tâche. Non seulement le travail de Tarski a durablement et généralement marqué la réflexion philosophique sur la notion de vérité au vingtième siècle mais, dans la perspective plus spécifique qui est la nôtre, c'est l'explication tarskienne qui sert de socle logico-philosophique à la plupart des réflexions déflationnistes sur la vérité. Le but de ce chapitre est donc triple. Il s'agit d'abord de présenter le travail de Tarski de façon détaillée, en prêtant attention au détail technique et aux acquis philosophiques. Il s'agit aussi d'introduire des notions, des notations et des objets théoriques dont nous aurons besoin tout au long de ce travail, et de préparer ainsi la mise en perspective des résultats de Tarski et du problème qui nous avons identifié comme notre problème principal au chapitre précédent, à savoir celui de la compatibilité d'une explication satisfaisante de la notion de vérité et de la thèse de l'absence de « rôle explicatif » de la notion de vérité. Enfin, il s'agit de discuter la question de savoir si le travail de Tarski lui-même est intrinsèquement porteur de conclusions déflationnistes. Le plan du chapitre est le suivant. Dans une première partie, je motive le projet tarskien d'une *définition* de la vérité et en précise les conditions. Dans une seconde partie, je décris la méthode introduite par Tarski pour mener à bien ce projet et explique en détail comment définir la vérité dans un cas particulier, avant d'introduire les théories axiomatiques de la vérité. Dans une troisième partie, je commence à examiner, après en avoir clarifié le sens, la question de savoir ce que l'on peut prouver, ou expliquer, lorsque l'on a les ressources pour définir ou expliquer la notion de vérité. En conclusion, je discute des implications philosophiques de la définition tarskienne.

2.1 Définir la vérité ?

2.1.1 Le contexte philosophique

Lorsque Tarski publie son travail séminal sur la vérité, au début des années trente,[1] le concept de vérité est un sujet de préoccupation philosophique dans le

[1]Une première version de la monographie principale de Tarski sur la notion de vérité a paru en 1933 en polonais sous le titre « Pojecie prawdy w jezykach nauk dedukcyjnych ». Le texte a ensuite été traduit en allemand et augmenté en 1935, et cette seconde version a paru en 1936 sous le titre « Der Wahrheitsbegriff in den formalisierten Sprachen » dans *Studia Philosophica*, 1 : 261–405. Nous nous référerons ici à la traduction anglaise de ce dernier, approuvée par Tarski, et

cercle des avant-gardes philosophiques d'Europe centrale, où il est parfois considéré avec soupçon, en particulier dans le Cercle de Vienne où se joue un partie du destin de la philosophie au vingtième siècle. On peut identifier trois sources de préoccupation relativement à la notion de vérité.

La première est liée à l'absence d'une analyse conceptuelle univoque de la notion intuitive de vérité. Une certaine confusion s'ensuit de ce qu'à côté de l'usage classique de la notion de vérité comme *adaequatio rei et intellectus*[2], sont venues s'adjoindre des conceptions utilitaristes[3] et surtout des conceptions épistémiques.[4] La multiplication des explications du concept vérité et les débats philosophiques auxquels ces conceptions donnent lieu apparaissent naturellement comme le symptôme d'une obscurité inhérente à la nature de ce concept.

La conception classique de la vérité, et en particulier ce qui se présenterait avec Tarski comme son avatar contemporain, la conception sémantique de la vérité, devait elle-même rencontrer un obstacle dans l'attitude de défiance adoptée par

parue dans Tarski (1983) sous le titre « The Concept of Truth in Formalized Languages ».

[2]La formule est celle de Thomas d'Aquin. On la trouve dans la *Somme Théologique* (I.21.2) :

> ... veritas consistit in adaequatione intellectus et rei

et dans le livre *De veritate*, où Thomas d'Aquin introduit la formule en l'attribuant, à tort d'après les spécialistes, à un certain Isaac Israel, mais la rattache explicitement à la célèbre définition du livre Γ de la *Métaphysique* d'Aristote. Voici le passage *in extenso*, dans la traduction des moines de l'Abbaye Sainte-Madeleine du Barroux :

> Ensuite on définit la vérité d'après ce en quoi la notion de vrai s'accomplit formellement ; et en ce sens, Isaac dit : « La vérité est adéquation de la réalité et de l'intelligence » ; et Anselme, au livre sur la Vérité : « La vérité est une rectitude que l'esprit seul peut percevoir » – en effet, cette rectitude a le sens d'une certaine adéquation –, et le Philosophe dit au quatrième livre de la Métaphysique que nous disons le vrai en le définissant, quand ce qui est, est dit être, ou ce qui n'est pas, n'être pas.

[3]Qu'on pourrait résumer très grossièrement par le slogan « Est vrai ce qu'il est utile de croire ». Cette inspiration fut surtout défendue par les pragmatistes américains, par exemple Pierce, James, et Dewey. Voir Damnjanovic et Candlish (2007) pour une panorama.

[4]Que l'on pourrait rassembler sous le slogan « Est vrai ce qu'il est cohérent ou acceptable de croire ». Une approche épistémique de ce genre était défendue par certains membres du Cercle de Vienne, en particulier Neurath :

> Du point de vue de la terminologie, il [Neurath] pense que l'on devrait réserver l'usage du terme « vrai » pour cette Encyclopédie qui a été choisie parmi les nombreuses encyclopédies cohérentes qui sont contrôlées par les énoncés protocolaires, en sorte que chaque conséquence de cette Encyclopédie et chaque nouvel énoncé qui y est accepté serait appelé « vrai » et tout énoncé qui la contredirait serait appelée « fausse ». (Neurath (1936) p.400 cité dans Mancosu (2008*b*))

Sur l'existence et l'état des débats internes au cercle de Vienne sur la notion de vérité avant et au moment de leur introduction au travail de Tarski (principalement au Congrès de Paris de 1935), voir Mancosu (2008*b*).

les positivistes envers toute notion que l'on pourrait soupçonner d'être entachée de conceptions « métaphysiques ».[5] En effet, pour se prémunir de tout glissement du discours vers le non-sens, les membres du Cercle de Vienne avaient souhaité d'ancrer fermement le discours dans le réel en n'accordant de signification « cognitive » qu'au discours *empiriquement vérifiable* selon des protocoles sévèrement définis. Or, dans une telle perspective, une tentative pour réhabiliter le concept de vérité, le concept classique de vérité comme correspondance, pouvait paraître suspecte, à la faveur de confusions auxquelles une telle tentative ne pouvait manquer de donner lieu. Neurath ne s'est sans doute jamais véritablement détaché de l'idée qu'il y avait quelque chose de dangereux dans l'idée même d'une telle réhabilitation. Dans une lettre qu'il adressera à Carnap en 1943, bien après, donc, que ce dernier ait embrassé les idées de Tarski, il écrira encore :

> La Scolastique a engendré le Brentanoisme, Brentano a donné naissance à Twardowski, Twardowski a donné naissance à Kotarbinski, Lukasiewicz (vous connaissez ses relations directes avec la Néo-scolastique en Pologne), les deux ont ensemble donné naissance à Tarski etc., et maintenant ils sont les grands-pères de NOTRE Carnap aussi ; de cette façon Thomas d'Aquin entre par une autre porte à Chicago[6]... (Lettre du 15 janvier 1943, citée dans Mancosu (2008*b*))

Carnap, commentant plus tard la réception des idées de Tarski au Congrès de Paris en 1935 dans son *Autobiographie intellectuelle*, livrera le diagnostic suivant :

> A ma surprise, il y eut une opposition véhémente même du côté de nos amis en philosophie... Neurath croyait que le concept sémantique de vérité ne pouvait pas être réconcilié avec un point de vue strictement empiriste et anti-métaphysique... J'ai montré que ces objections étaient fondées sur une mécompréhension du concept sémantique de vérité, l'incapacité à distinguer entre ce concept et des concepts comme ceux de certitude, connaissance de la vérité, vérification complète et autres.(Carnap (1963), p.61)

La résistance fut donc bien réelle, et il apparaît ainsi que les nombreuses remarques de clarification à propos de la portée de la conception sémantique de la vérité faites

[5] Voir par exemple le *Manifeste du Cercle de Vienne* in Soulez (1985).

[6] NdT : A cette époque Carnap avait émigré aux Etats-Unis et enseignait à l'Université de Chicago.

dans Tarski (1944), en particulier les garanties qu'il insiste à donner du fait que ses conceptions ne contredisent en rien les postulats méthodologiques du physicalisme le plus strict, ces remarques sont pour une part motivées par sa correspondance avec Neurath dans les années trente.[7]

Ces doutes et ces confusions, néanmoins, étaient surtout ceux du positivisme, et ne touchaient pas l'Ecole Polonaise, héritière de la tradition sémantique de Bolzano, et de Frege.[8] Au-delà des Ecoles, pourtant, une troisième cause d'inquiétude touchait les philosophes indistinctement : je veux parler des paradoxes de la vérité. La notion de vérité est connue depuis l'antiquité pour les paradoxes auxquels son usage incontrôlé donne lieu. Diogène Laërce attribue à Eubulides de Milet le plus célèbre d'entre eux, le *pseudomenon*.[9] On peut rendre le paradoxe de la façon suivante. Appelons λ l'énoncé :

L'énoncé λ n'est pas vrai

Le paradoxe apparaît lorsque l'on s'interroge sur la vérité de l'énoncé. S'il est vrai, alors il n'est pas vrai, puisque c'est ce qu'il dit. Donc il n'est pas vrai. Mais c'est ce qu'il dit, il est donc vrai après tout. Contradiction.

Les variations anciennes et modernes autour de ce paradoxe sont innombrables, chacune permettant de préciser un peu plus ce qu'il est et ce qu'il n'est pas.[10] En tout état de cause, si les principes nécessaires à la dérivation de l'antinomie sont profondément ancrés dans notre compréhension du sens du prédicat de vérité, la question se pose de savoir si cette notion n'est pas tout simplement incohérente, et donc illégitime.

[7] Voir les extraits de la correspondance cités dans Mancosu (2008*b*) et Mancosu (2009).

[8] Coffa (1991) affirme, en substance, que si la confusion entre recherche d'une définition et recherche d'un critère est à la source de l'incapacité des positivistes à penser les notions sémantiques, cette confusion est un héritage Kantien auquel la tradition sémantique, de Bolzano à l'Ecole Polonaise, avait échappé. Tirant le bilan de la rencontre de Carnap avec l'Ecole Polonaise, Coffa écrit :

> Le bénéfice philosophique immédiat [de la reconnaissance de la doctrine de Tarski par Carnap] ne fut pas beaucoup plus que la reconnaissance de la vieille distinction Frégéenne entre le contenu d'un énoncé [statement] et son assertion.[...] Carnap avait atteint le point où Bolzano avait commencé.(Coffa (1991), p.372-373)

[9] Diogène Laërce, *Vies, doctrines et sentences des philosophes illustres*, II, 108.

[10] Nous ne les passerons pas en revue. Pour une introduction, le lecteur pourra consulter Quine (1966). Field (2008) est une étude récente logiquement et philosophiquement approfondie sur les paradoxes de la vérité.

Cette brève présentation du contexte philosophique indique clairement quelles tâches une réflexion sur la notion classique de vérité doit accomplir selon Tarski : 1. Il faut d'abord présenter une analyse claire de la notion de vérité en vue. 2. Il faut ensuite montrer que l'emploi de cette notion est conforme aux canons méthodologiques du langage de la science. 3. Il faut enfin montrer qu'il est possible de faire un usage cohérent de cette notion. On peut distinguer dans le travail de Tarski deux étapes : une analyse philosophique de la notion de vérité, et la recherche d'une définition explicite de la vérité. La première étape accomplit la première des trois tâches mentionnées, et donne lieu à la formulation d'un *critère d'adéquation*, c'est-à-dire d'un critère qui doit permettre d'assurer, s'il est satisfait, que c'est la notion de vérité visée qui est saisie par la définition. La définition elle-même permet ensuite d'accomplir les seconde et troisième tâches : si la théorie dans laquelle la définition est conduite satisfait à toutes les exigences méthodologiques que l'on peut souhaiter d'exiger (cohérence, absence de notion dont le sens serait douteux), alors, puisque la définition opère une réduction de la notion définie et des principes qui la gouvernent aux notions et aux lois de cette théorie, la notion définie est de ce fait même lavée de tout soupçon. Tarski doit donc montrer que la définition de la vérité est possible dans une théorie de ce type.

2.1.2 Porteurs de vérité et langages formalisés

Avant de définir la vérité, avons-nous dit, il est indispensable de préciser la notion de vérité qui est en jeu. À titre de première clarification, il est sans doute nécessaire de préciser d'abord quelle est la nature des porteurs, ou des véhicules, de la vérité. Les porteurs de vérité sont, en première analyse, chez Tarski, les énoncés d'un langage identifiés selon leur type, c'est-à-dire non pas les objets concrets que sont les inscriptions particulières de ces énoncés, mais les objets abstraits qui constituent la forme générale de ces inscriptions.[11] Ce choix des énoncés a peut-être, à première vue, de quoi de surprendre : après tout, n'est-ce pas d'abord *ce qui est dit*, le contenu d'un énoncé, qui est susceptible d'être vrai ou faux ? Les objets « naturels » pour les recherches sémantiques, en particulier pour la définition de la vérité, ne sont-ils pas plutôt quelque chose comme les propositions objectives de Bolzano, les pensées

[11] Tarski (1983), p.156, note 1, Tarski indique la possibilité de concevoir les énoncés comme étant plutôt des inscriptions physiques. En renonçant à suivre cette voie, il renonce également à une forme de nominalisme strict qui était celle de Lesniewski.

de Frege, ou les jugements dont parlait Ramsey à la même époque et que nous avons évoqués au chapitre précédent ?[12] Sur ce point, Tarski n'innove pas et suit une partie de l'Ecole Polonaise de cette époque.[13] Mais surtout, il faut ajouter que les énoncés dont il est question ici, Tarski y insiste, sont des énoncés *interprétés*[14], c'est-à-dire simplement doués de sens, d'un langage dont *la structure est exactement spécifiée*.

Que signifient ces deux conditions ? Commençons par la première, la condition que les énoncés sont doués de sens. Tarski est clair sur le fait que le problème de définir la vérité pour un langage n'a de sens que pour un langage interprété.[15] Les langages pour lesquels Tarski entend définir la vérité doivent donc être absolument distingués des langages formels qui intéressaient Hilbert. Pour caractériser sa position, Tarski renvoie au « formalisme intuitionniste » Lesniewski[16], expression sans doute malheureuse (marquant en somme la double opposition méthodologique à Hilbert et à Brouwer) par laquelle il décrivait sa propre décision théorique de prendre comme constituants des théories des énoncés formalisés et interprétés. Lesniewski écrivait :

> N'ayant aucune prédilection pour les « jeux mathématiques variés » qui consistent à écrire selon une règle conventionnelle ou une autre

[12] Voir Chapitre 2 section 2.3

[13] Sur l'influence décisive de Brentano sur l'Ecole Polonaise et le développement de la sémantique, en particulier relativement aux choix des porteurs de vérité, voir Rojszczak (2002) et Rojszczak (2005). Sur le nominalisme durable de Tarski et ses afffinités avec le nominalisme qui fut d'abord celui de Quine, on pourra également consulter Mancosu (2008*a*). On trouvera un exposé synthétique du contexte historique du développement de la sémantique scientifique dans l'Ecole de Lvov-Varsovie dans Wolenski (2009).

[14] Quand je parle d'énoncé *interprété*, ici, ce n'est pas au sens moderne, modèle-théorique, d'énoncé formel auquel à été donné une « interprétation » ensembliste, mais en un sens plus relâché d'énoncé doué de sens. En termes modernes, on pourrait définir un langage interprété simplement comme un couple $(\mathcal{L}, \mathcal{M})$ où $\mathcal{L}$ est un langage *formel* et $\mathcal{M}$ une $\mathcal{L}$-structure ; mais la notion de vérité dans une structure d'interprétation n'apparaîtra que plus tard, dans Tarski et Vaught (1957).

[15] Voir par exemple dans Tarski (1935) :

> Il faut sans doute encore ajouter que nous ne sommes pas intéressé ici par les langages « formels » pour les sciences en aucun sens particulier du mot « formel », nommément les sciences dans lesquelles aucun sens matériel n'est attaché aux signes des expressions. Pour ces sciences le problème discuté ici [i.e. définir la vérité] n'est pas pertinent, il n'a pas même de sens. Nous attribuerons toujours une signification très concrète et, pour nous, intelligible, aux signes qui apparaissent dans le langage que nous allons considérer. (Tarski (1983), p.166)

[16] Dont Tarski fut le seul étudiant. Ce qui, pour l'anecdote, faisait dire à Lesniewski que 100% de ses étudiants étaient des génies.(Source : Jan Wolenski, cité dans Betti (2008))

> des formules plus ou moins imagées qui n'ont pas besoin d'avoir de signification et même - comme préfèrent peut-être quelques « joueurs mathématiciens » - doivent être dénuées de signification, je n'aurais pas pris la peine de systématiser et de vérifier souvent si scrupuleusement les directives de mon système, si je n'avais imputé à ses thèses une certaine signification spécifique et complètement déterminée, en vertu de laquelle ses axiomes, définitions et directives finales [...] ont pour moi une validité intuitive irrésistible.(Lesniewski (1929), cité dans Wolenski (2009), p.49)

Un même énoncé, considéré du point de vue syntaxique, pouvant avoir des significations différentes dans différents langages, la définition de la vérité des énoncés sera donnée relativement à un langage interprété. Ce que Tarski se propose de montrer, c'est comment définir le prédicat « vrai-dans-L », pour un langage L donné.

Néanmoins, pour qu'une telle définition soit possible, un certain nombre de conditions doivent être satisfaites par le langage L pour lequel on cherche à définir la vérité. La condition de *spécification exacte de la structure du langage* est la première d'entre elles. Que faut-il entendre par là ? Tarski [17] impose deux types de contraintes.

1. D'une part, la *grammaire* du langage doit être exactement spécifiée, c'est-à-dire la classe des formes syntaxiques qui ont une signification dans le langage en question, en particulier la classe des termes, des formules et des énoncés.
2. D'autre part, Tarski demande que soient formulées les conditions sous lesquelles un énoncé du langage peut être *asserté*, c'est-à-dire quels sont les axiomes et les règles d'inférence pour ce langage.

La seconde condition peut étonner : nous parlions de langage, et voilà qu'il s'agit maintenant de spécifier des axiomes et des règles d'inférence, autrement dit une théorie. S'agit-il donc de la même chose ? Non, mais Tarski explique son choix par les applications qu'il a en vue :

> ... les langages formalisés ont jusqu'à présent été construits exclusivement dans le but d'étudier *les sciences déductives* formalisées sur la base de tels langages. Le langage et la science grandissent ensemble dans un tout unique, si bien que nous parlons du langage d'une science déductive

[17] Tarski (1935) p.166 , Tarski 1944 p. 346.

> formalisée particulière, et non de tel ou tel langage formalisé.(Tarski (1983) note 1 p.166).

Que, dans la définition des langages à structure exactement spécifiée donnée par Tarski, des conditions qui ressortissent à l'individuation de théories se surimposent à des conditions ordinaires d'individuation d'un langage ne doit pas néanmoins nous empêcher de nous donner des moyens de garder trace de cette distinction pour la suite ; je propose donc de maintenir l'emploi que fait Tarski du terme *langage* et de parler de *langage** pour désigner un langage *stricto sensu*, en tant que distinct d'une théorie. En analogie avec la notion de langage exactement spécifié, nous dirons que la structure d'un langage* est exactement spécifiée lorsqu'il satisfait simplement la première des deux conditions données ci-dessus. Cette précaution nous permettra de clarifier certains points conceptuels par la suite.[18]

Enfin, Tarski définit la notion de *langage formalisé* : les langages formalisés sont ces langages pour la spécification de la structure desquels il n'est fait référence qu'à la forme des expressions. C'est pour ces langages, plus particulièrement, qu'il se propose de définir la vérité. Ceci exclut donc les langages contenant des termes indexicaux ou des déictiques et plus généralement des expressions dont la signification varie avec le contexte. Dans un langage formalisé, « *le sens de chaque expression est déterminé uniquement par sa forme* ».[19] Une définition analogue vaudrait pour les *langages* formalisés*.

Le domaine d'application de la méthode de définition de la vérité que Tarski envisage sera donc limité à l'ensemble des langages formalisés, mais sans toutefois préjuger du domaine de discours de ces langages qui peuvent, en principe, être le langage de sciences purement déductives aussi bien qu'empiriques.[20] Il est à noter également que c'est cette contrainte de la spécification exacte de la structure qui disqualifiait aux yeux de Tarski d'une définition de la vérité relativement au langage

[18] Notons que dans Tarski (1969), la notion de langage formalisé a été épurée et il n'est plus du tout question des règles d'assertation associées au langage pour lequel il s'agit de définir la vérité.

[19] Tarski (1983), p.166. Je souligne.

[20] Tarski (1944) mentionne la physique théorique (p. 347), ou des approximations de fragments des langages "parlés" dont la structure aurait pu être suffisamment spécifiée ; la possibilité d'une contribution substantielle de la sémantique à la méthodologie des sciences déductives et empiriques, développée dans Tarski (1944) (p. 366 et suivantes), par exemple en vue de clarifier la notion d'acceptabilité d'une théorie scientifique quelconque, implique qu'il soit possible en principe d'appliquer le concept sémantique de vérité à des langages scientifiques quelconques pourvu qu'ils satisfassent aux conditions d'application de la méthode de définition que présente Tarski.

ordinaire ou à des fragments importants de langage ordinaire.[21,22,23] En effet, selon Tarski,

> notre langage de tous les jours n'est certainement pas un langage avec une structure exactement spécifiée. Nous ne savons pas précisément quelles sont les expressions qui sont des énoncés, et à un moindre degré encore quels sont les énoncés qui doivent être pris pour assertables.(Tarski (1944), p.349)

2.1.3 Adéquation

La nature des porteurs de vérité étant précisée, revenons au projet d'une définition de la vérité elle-même, tel que le conçoit Tarski. Il est crucial de noter que ce projet ne vise pas une définition simplement stipulative d'un nouveau concept sous un ancien vocable mais une définition qui capture la signification du terme « vrai », et plus spécialement celle qu'il a dans la conception classique de la vérité comme correspondance avec la réalité et que Tarski appellera, dans la version qu'il en donne, la *conception sémantique* de la vérité. Cette conception classique de la vérité, Tarski en identifie une expression particulièrement claire chez Aristote[24] :

> Dire de ce qui est que ce n'est pas, ou dire de ce qui n'est pas que c'est, est faux, tandis que dire de ce qui est que c'est, ou de ce qui n'est pas que ce n'est pas, est vrai (Aristote, Métaphysique Γ)

C'est cette notion, et celle-là seulement, qu'il s'agit de définir rigoureusement.

Lorsque l'on se propose de définir une expression ayant déjà une signification, il faut se donner un critère de succès de cette définition, un critère qui nous garantisse, s'il est satisfait, que la définition saisit bien la notion visée. Pour parvenir à ce critère, Tarski part de l'analyse conceptuelle de la notion classique de vérité, la notion en jeu dans la définition aristotélicienne, qu'il a apprise de Kotarbinski.[25] La remarque

[21] Y compris au sens de langage* ordinaire.

[22] Tarski (1944), p. 349 est clair sur ce point et corrige Tarski (1935) qui attribuait le problème au caractère *universel* du langage ordinaire (*cf.* Tarski (1983) p.160). Voir Rouilhan (1984) et Rouilhan 1998 pour une description de ce changement.

[23] Il est naturel de penser que Tarski aurait pu être conduit à nuancer son jugement sur ce point sous l'influence des développements ultérieurs en linguistique, en particulier après la parution de Montague (1970). Je ne peux que spéculer ici, n'ayant pas trouvé de remarques écrites dans ce sens.

[24] Voir Tarski (1983) p.155, Tarski (1944), p. 343, et Tarski (1969) p. 63.

[25] Sur l'état des recherches sur la vérité dans l'Ecole Polonaise avant Tarski, voir Wolenski et Murawski (2008).

philosophique à la fois simple et cruciale que Tarski met en avant, est que la notion classique de vérité, appliquée à un énoncé, « La neige est blanche » pour reprendre l'exemple de Tarski, est *entièrement expliquée* par l'équivalence suivante :

« La neige est blanche » est vrai si et seulement si la neige est blanche

Nul recours à la notion de fait, d'état de chose ou de correspondance, l'explication de ce en quoi consiste la vérité pour un énoncé est, pour reprendre l'expression de Quine, aussi claire que l'énoncé lui-même. L'intuition de la correspondance est clairement présente dans ce biconditionnel : celui-ci affirme qu'une certaine expression linguistique a une certaine propriété, la vérité, si et seulement si un certain état de chose est réalisé. Nous avons supposé dans cet exemple que le langage auquel appartient l'énoncé dont la vérité était expliquée était identique au langage dans lequel se tenait l'explication. Mais il ne s'agit que d'une simplification. Rien n'interdit que cette définition partielle de la vérité pour un énoncé d'un langage donné ne soit conduite dans un langage entièrement différent. Si nous appelons *langage-objet* le langage auquel appartient l'énoncé dont la vérité est définie, et *métalangage* le langage *dans lequel* la vérité est définie, les explications de ce en quoi consiste la vérité pour un énoncé donné du langage-objet sont fournies les instances du schéma-T, c'est-à-dire du schéma suivant :

X est vrai si, et seulement si, p

où « X » est mis pour le nom d'un énoncé du langage du langage-objet, et la lettre « p » pour une traduction de cet énoncé dans le métalangage.[26] Après avoir montré que les équivalences-T constituent une explication complète de ce en quoi consiste la vérité de chaque énoncé pris individuellement, Tarski est en mesure de formuler le critère d'adéquation recherché. C'est à la convention-T[27] qu'il revient de le formuler [28] :

Convention-T Une définition formellement correcte du symbole « Vr », formulée dans le métalangage, sera appelée une *définition adéquate de la vérité*, si elle a les conséquences suivantes :

[26] Remarquons que les *instances* de ce schéma, les équivalences-T individuelles, sont des énoncés du (méta)langage dans lequel est définie la vérité, tandis que le schéma-T lui-même n'en fait pas forcément partie.

[27] Tarski (1983), p.186-187.

[28] Les deux conditions suivantes n'ont pas la même importance, la seconde condition étant jugée secondaire par Tarski lui-même, et parfois omise. Par exemple dans Tarski (1944), p. 344.

1. Tous les énoncés qui sont obtenus à partir de l'expression « x est vrai si et seulement si p » par substitution au symbole « x » d'un nom d'un énoncé du langage en question et au symbole « p » d'une expression qui est une traduction de cet énoncé dans le métalangage.
2. L'énoncé « Pour tout x, si x est vrai, alors x est un énoncé (du langage-objet) »

La convention-T ici n'est pas formulée dans le métalangage (le langage de la définition elle-même), mais dans un méta-métalangage, et porte sur toute définition à formuler dans le métalangage d'un langage-objet quelconque pour lequel il s'agit de définir la vérité. Ainsi, tandis que la définition de la vérité est relative à un langage-objet particulier (on définit vrai-pour-L, L étant donné), dans la condition d'adéquation le langage L est une *variable*. Je voudrais m'arrêter brièvement sur deux points particuliers soulevés par la convention-T : la question de la traduction et celle du statut des équivalences-T.

Traduction La convention-T implique que, pour prouver qu'une définition de vrai-pour-L dans une métathéorie ML est adéquate, il faudra montrer, éventuellement dans une méta-métathéorie, que l'on peut prouver dans la méta-théorie toutes les équivalences-T ; et pour montrer qu'un énoncé est une équivalence-T (une instance du schéma-T) il faut montrer qu'il est de la forme $Vr(s) \leftrightarrow p$, où « s » est un nom d'énoncé et « p » un énoncé, *et* que p est une *traduction* de l'enoncé s mentionné à gauche.[29] La théorie dans laquelle est formulée la condition d'adéquation doit donc contenir une notion de *traduction* d'un langage dans un autre. Or qu'est-ce que cette notion de traduction ? Tarski n'en dit pas grand-chose. Toute fonction des énoncés du langage-objet dans l'ensemble des énoncés du métalangage est-elle une (fonction de) traduction ou bien la fonction de traduction doit-elle obéir à des contraintes particulières ? Et si une théorie des fonctions de traduction est nécessaire, est-elle

[29]En fait, étant donné un langage de base pour lequel on cherche à définir la vérité, on peut imaginer de se donner dans la méta-théorie (dans laquelle on cherche à définir la vérité) la fonction de traduction requise. Il suffit pour cela d'avoir construit notre syntaxe de façon à pouvoir y définir les notions morphologiques et syntaxiques non seulement du langage objet mais encore du métalangage lui-même. Dans le cas simple où le métalangage est une extension du langage-objet, on pourra ainsi définir (dans la métathéorie) la traduction *homographique* simplement comme la fonction f des énoncés de L dans les énoncés de ML telle que$f(x) = x$. (Dans le cas non-homographique, il pourra être utile de définir d'abord récursivement une fonction de traduction des termes du langage-objet dans les termes du méta-langage et des prédicats primitifs du langage-objet dans les formules ouvertes du méta-langage.)

une théorie purement mathématique, au sens où elle imposerait uniquement des contraintes structurelles à la fonction en question, ou bien pourrait-il s'agir d'une théorie scientifique en un sens plus large ?

Il est naturel de penser que, si Tarski avait eu à développer une théorie de la traduction, sa méthode eût été analogue à celle qu'il emploie pour définir la vérité. Il définirait explicitement, non pas la notion de traduction d'un langage dans un autre en général, mais au cas par cas définirait de façon stipulative (dans la métathéorie) une fonction de traduction $trad_{L:ML}$ des énoncés de L dans ceux de ML. Pour l'entreprise bien circonscrite d'une définition de la vérité pour un langage dans un métalangage, il est suffisant de raisonner à fonction de traduction donnée dans le métalangage en question. La question de l'*adéquation* de cette fonction $trad_{L:ML}$ pourrait alors être posée, à charge pour nous d'y répondre dans une autre métathéorie, relativement à une nouvelle fonction de traduction $Trad_{L+ML:MML}$ prenant ses arguments parmi les énoncés de L et de ML et ses valeurs parmi les énoncés de ce nouveau langage plus englobant, MML, une suggestion assez naturelle étant alors de poser que $trad_{L:ML}$ est adéquate si et seulement si pour tout énoncé x de L et y de ML, si $trad_{L:ML}(x) = y$, l'énoncé y est bien une traduction de x, au sens cette fois de $Trad$, la fonction de traduction définie dans la méta-méta-métathéorie.[30]

Il me semble néanmoins que la façon d'aborder la question de la traduction que nous venons d'attribuer à Tarski de façon hypothétique n'est pas la seule acceptable du point de vue méthodologique, y compris du point de vue de ce que Tarski aurait jugé être une méthodologie acceptable. Au lieu de stipuler, dans chaque théorie, une fonction de traduction arbitraire, et de renvoyer la question de son adéquation à une autre théorie, relativement à une autre fonction de traduction, si bien que la question de l'adéquation d'une fonction de traduction n'a jamais de sens que relativement à une autre fonction de traduction donnée, on peut imaginer de formuler une théorie générale axiomatisée, TRAD, de la notion de traduction d'un langage dans un autre, à l'effet qu'une fonction des énoncés d'un langage dans un autre est une fonction de traduction si et seulement si elle satisfait les axiomes de TRAD. Dans cette perspective, on ne raisonne plus sur l'adéquation d'une fonction de traduction *à fonction de traduction donnée dans une métathéorie*, mais l'on cherche des critères *généraux* d'adéquation. À quoi ces critères pourraient-ils ressembler ? Ces critères devraient, d'une manière ou d'une autre, saisir le fait que les langages

[30] De façon plus générale on pourrait demander simplement que $Trad_{L+ML:MML}(x) = Trad_{L+ML:MML}(y)$.

dont nous parlons sont des langages doués de sens, interpétés (ou sinon l'idée même de traduction n'a pas de sens). Or qu'est-ce qu'un langage interprété ? Il y a deux façons d'aborder cette question. La première est de considérer un langage interprété comme un objet mathématique abstrait identifié à une syntaxe *et* une sémantique. Mais si cette sémantique est une sémantique vériconditionnelle, ne risque-t-on pas de tourner en rond ? Avoir besoin d'une notion de vérité bien définie pour définir la notion d'adéquation d'une traduction pourrait poser problème quand ce que l'on cherche sont les conditions qui doit satisfaire une fonction de traduction pour expliquer en toute généralité ce en quoi consiste l'adéquation d'une définition de la vérité. La seconde approche est de considérer un langage interprété comme un objet individué par des conditions structurelles (sa morphologie, sa syntaxe, susceptible d'une étude purement mathématique) *et* des conditions empiriques typiquement liées à la façon dont des locuteurs utilisent ces langages. Une théorie générale de la traduction est alors envisageable qui ne fasse appel à aucune notion sémantique primitive et dont les axiomes doivent permettre de spécifier à quelles conditions une fonction des énoncés d'un langage dans les énoncés d'un autre langage peut être considérée comme une fonction de *traduction*. Typiquement TRAD comportera des axiomes du type : si $\phi' \in ML$ est une traduction de $\phi \in L$, pour tout locuteur S' de ML et tout locuteur S de L, l'ensemble des circonstances observables dans lesquelles S' donne son assentiment à ϕ' est identique à l'ensemble des circonstances observables dans lesquelles S donne son assentiment à S. On sait que, selon Quine, une théorie de ce genre, si elle est formulée en des termes méthodologiquement acceptables d'un point de vue empiriste, doit sous-déterminer la traduction, au sens où il existera toujours un grand nombre de fonctions de traduction d'un langage dans un autre satisfaisant les axiomes de TRAD. Reste que si la notion de traduction entre langages empiriquement donnés, ainsi théorisée, n'est pas purement mathématique, ce n'est pas non plus une notion dont l'explication fait appel à des notions sémantiques primitives et, en ce sens, elle est conforme aux exigences méthodologiques présentées par Tarski, comme nous le verrons en précisant ce que sont ces exigences tout à l'heure.

Le statut des équivalences-T dans la théorie tarskienne Quel est le statut modal des équivalences-T dans la théorie tarskienne ? La démarche de Tarski est une démarche d'explication, au sens de Carnap, de la notion de vérité, au cours de laquelle une notion préthéorique, la notion ordinaire de vérité, est remplacée par la notion précise de vérité qui est définie. La convention-T ne statue pas elle-même sur

les ressources théoriques permises pour mener la définition à bien, et desquelles les équivalences-T doivent être des conséquences. Mais puisque le projet de Tarski est d'introduire par *définition* une notion précise de vérité, et que dans une définition le *definiendum* est introduit comme une abréviation du *définiens* (les deux sont donc logiquement équivalents), pour comprendre quel statut échoit aux équivalences-T [31], il suffit de se demander à quels moyens il est nécessaire de faire appel pour dériver les équivalences-T de cette définition. Or dans les cas où la définition explicite de la vérité pour un langage peut être menée à bien, comme nous le verrons plus loin[32], Tarski sera en mesure de montrer que les équivalences-T sont en fait conséquences de la définition *via* des principes théoriques minimaux : essentiellement des principes de logique et de syntaxe. Les équivalences-T apparaissent alors comme des équivalences "quasi"-logiques, selon l'expression de Rouilhan et Bozon (2006). Par conséquent, pour la notion de vérité introduite par Tarski, le membre gauche et le membre droit d'une équivalence-T expriment la même proposition (ou "quasiment").[33]

2.1.4 L' « idéologie » de la définition

Nous venons de voir en quoi consistait la *condition d'adéquation* imposée par Tarski à toute définition de la vérité. Il faut à présent ajouter un dernier *desideratum* de Tarski : la définition de la vérité pour un langage doit être donnée dans des termes entièrement non-sémantiques.

> En particulier, nous désirons que *les termes sémantiques* (référant au langage-objet) *ne soient introduits dans le métalangage que par définition.* Car si ce postulat est satisfait, la définition de la vérité, ou de tout autre concept sémantique, accomplira ce que nous attendons intuitivement d'une définition ; c'est-à-dire qu'elle expliquera la signification du terme défini en des termes dont la signification est entièrement claire et sans équivoque. (Tarski (1944), p.351)

[31] Comprise comme utilisant la notion définie de vérité, et non la notion ordinaire.

[32] Le lecteur peut consulter la section 3.2.1.

[33] Au moins dans le sens logique du terme « proposition », où deux énoncés expriment la même proposition si et seulement si ils sont logiquement équivalents. Mais l'on sait que cette notion de proposition n'est pas *a priori* la même celle d'objet des attitudes mentales propositionnelles. Sur la signification de la déductibilité des équivalences-T dans la syntaxe et la logique, voir également les remarques de Rouilhan et Bozon (2006).

Puisqu'une *définition* adéquate explicite de la vérité dans une théorie vaut preuve que l'usage de la notion n'*introduit* pas de contradiction dans la théorie en question, alors si cette théorie ne contient, comme Tarski le souhaite, que des termes clairs et sans équivoque, et que nous n'avons pas de doute relativement à sa cohérence, la définition de la vérité pour un langage donné dans cette théorie doit dissiper tous nos doutes relativement à la possibilité d'un usage adéquat cohérent du prédicat de vérité pour ce langage.

A ce point, nous voyons déjà que le projet de définition de la vérité pour un langage, tel qu'il est présenté par Tarski, et s'il peut être mené à bien, doit permettre de surmonter les trois problèmes faisant obstacle à la reconnaissance de la légitimité de la notion classique de vérité que nous avons mentionnés pour commencer : elle doit permettre de clarifier la notion classique de vérité, de montrer qu'un usage cohérent de cette notion est possible, et enfin que cet usage est métaphysiquement neutre, conforme à la méthodologie scientifique, la définition pouvant être entièrement conduite dans une théorie dont tous les termes sont clairs et sans équivoque.

2.1.5 Langages sémantiquement clos

Il n'est pas suffisant que le langage pour lequel on veut définir la vérité ait une structure exactement spécifiée pour que la définition soit possible. C'est ce que montre l'antinomie du Menteur. En effet, considérons à nouveau l'énoncé M :

$$M \text{ n'est pas vrai}$$

et supposons maintenant que nous puissions donner en français une définition adéquate de vrai (-en-français). À partir de cette définition, en vertu de la convention-T, nous devrions être en mesure de dériver l'équivalence-T correspondant à M, à savoir :

$$M \text{ est vrai si et seulement si } M \text{ n'est pas vrai}$$

qui est une contradiction patente. De quelles hypothèses la dérivation de cette antinomie dépend-elle crucialement ? Tarski en identifie trois :

1. avoir travaillé avec un langage sémantiquement clos, par quoi Tarski entend un langage possédant
   a) un nom de chacune de ses expressions,

b) un prédicat de vérité pour ses énoncés,

c) et tel que toutes les équivalences-T qui déterminent l'usage adéquat de « vrai » y sont assertables.

2. avoir utilisé les lois de la logique classique

3. avoir fait une hypothèse empirique relativement à la dénotation d'une certaine expression, nommément « λ ».

Tarski remarque que l'hypothèse empirique selon laquelle M dénote l'énoncé "M n'est pas vrai" est en fait éliminable, c'est-à-dire qu'il est possible de reconduire le paradoxe sans elle.[34] La preuve repose également sur l'usage des lois de la logique classique. Mais renoncer à ces lois est un prix que Tarski n'est pas prêt à payer.

Parce que, comme le note Tarski, l'hypothèse (3)*Ici l'hypothèse empirique est que M dénote l'énoncé "M n'est pas vrai" .* est en fait éliminable, il n'y a pas de raison de s'y attarder.[35] Par ailleurs Tarski n'est pas prêt à abandonner les lois de la logiques classiques, ce qui élimine (2). Reste donc (1) : le langage sémantiquement clos. Tarski (1944) conclut :

> En conséquence, nous décidons *de ne pas utiliser de langage sémantiquement clos* au sens que nous venons de préciser.(Tarski (1944), p.349)

Les langages (classiques) sémantiquement clos étant inconsistants, un langage (classique) ne peut pas contenir un prédicat adéquat de vérité pour lui-même, et par conséquent, *a fortiori*, il n'est pas possible de *définir* adéquatement la vérité pour un langage L dans L lui-même, puisqu'alors L devrait contenir les moyens de décrire ses expressions, un prédicat de vérité, et permettre de dériver toutes les équivalences-T, autrement dit être sémantiquement clos. Par conséquent il est en fait *nécessaire*, pour que la définition de la vérité pour L soit possible, de conduire cette définition dans un méta-langage distinct du langage-objet.

Nous avons vu que l'instanciation du schéma-T par l'énoncé du Menteur donnait lieu à une contradiction. Et que par conséquent un langage ne pouvait pas même *contenir* de prédicat de vérité adéquat pour lui-même. D'un autre côté, étant

[34]Si le langage est celui de l'arithmétique A, contenant un prédicat, primitif ou défini, Vr, on peut en fait prouver qu'il existe un énoncé λ du langage de l'arithmétique (éventuellement augmenté du prédicat primitif) tel que $A \vdash \lambda \leftrightarrow Vr(\ulcorner\lambda\urcorner)$. C'est une conséquence du lemme de diagonalisation que Carnap avait extrait des preuves d'incomplétude de Gödel.

[35]Si le langage est celui de l'arithmétique A, contenant un prédicat, primitif ou défini, Vr, on peut en fait prouver qu'il existe un énoncé λ du langage de l'arithmétique (éventuellement augmenté du prédicat primitif) tel que $A \vdash \lambda \leftrightarrow Vr(\ulcorner\lambda\urcorner)$. C'est une conséquence du lemme de diagonalisation que Carnap avait extrait des preuves d'incomplétude de Gödel.

donné un langage L contenant un prédicat primitif de vérité, toute instanciation du schéma-T par un énoncé de L contenant lui-même le prédicat de vérité ne semble pas devoir permettre de dériver une contradiction. On peut penser par exemple que l'équivalence-T suivante :

> « L'énoncé « la neige est blanche » est vrai » est vrai si et seulement si l'énoncé « la neige est blanche » est vrai

n'est pas problématique.[36] Si L un langage contenant un prédicat primitif Vr, et Σ est un ensemble d'énoncés de L, appelons Vr Σ-adéquat pour L si toutes les instances du schéma-T correspondant aux énoncés de Σ sont assertables dans L. On sait que si (L est classique et) cohérent Vr n'est pas En_L-adéquat, où En_L désigne l'ensemble des énoncés de L. C'est ce que Tarski vient de montrer. Mais la question se posait de savoir pour quels ensembles d'énoncés Σ il était possible que L contienne un prédicat de vérité Σ-adéquat. De façon rétrospectivement étonnante peut-être, Tarski n'a pas cherché à caractériser cette classe d'énoncés, alors que comprendre leurs propriétés s'est révélé être une des tâches principales des théoriciens contemporains de la vérité.[37] Mais rappelons que Tarski cherchait d'abord une définition explicite de la vérité. Or montrer qu'un langage classique cohérent peut contenir un prédicat primitif de vérité Σ-adéquat pour un fragment Σ étendant proprement le fragment non aléthique du langage, et montrer encore que dans certains cas l'extension de la syntaxe par les équivalences-T correspondantes est conservative[38], et que par conséquent une théorie cohérente et adéquate de la vérité pour un langage dans ce langage lui-même est jusqu'à un certain point possible, tout ceci n'aurait pas permis d'accomplir ce qu'une *définition explicite* accomplit *de surcroît*, à savoir une *explication* de la notion de vérité en termes non sémantiques.

[36] Et elle ne l'est pas, au sens où en ajoutant cette équivalence-T à l'ensemble des équivalences-T obtenues en instanciant le schéma-T par des énoncés du fragment de L ne contenant pas de prédicat de vérité, on obtient une théorie cohérente.

[37] On peut voir cette question comme étant le centre des recherches de S. Kripke (1975), McGee (1991), pour ne prendre que deux exemples. Voir aussi pour un traitement entièrement classique de la notion d'énoncé fondé l'article de Leitgeb (2005). Sur les ensembles maximalement consistants d'équivalences-T, voir McGee (1992). Plus récemment Cieslinsky (2007) a cherché à identifier les ensembles maximalement consistants d'équivalences-T étendant *conservativement* la syntaxe.

[38] Satisfaisant ainsi l'un des critères classiques de la méthodologie de la définition (de la définition axiomatique, en l'espèce).

2.2 Définir la vérité

Les termes du problème sont maintenant clairement définis, et la question est la suivante :

Question 1. *Est-il possible de donner une définition* adéquate *de la vérité pour un langage* formalisé *L, dans une métathéorie dont tous les termes, axiomes et règles d'inférences sont non-problématiques, et en particulier ne contiennent pas de termes sémantiques ?*

Et ce que j'appellerai ici *le cœur* de la réponse de Tarski est, dans le vocabulaire qu'il adoptera en 1944 :

Réponse 1.

C'est possible si, et seulement si, il existe un langage essentiellement plus riche que L.[39]

Mais que signifie au juste cette condition de « richesse essentielle » ? Le texte de « La conception sémantique de la vérité »(1944) n'est pas très explicite sur ce point et il est naturel de nous tourner vers la monographie de 1935 pour trouver une réponse exacte à cette question. Or lorsque l'on cherche dans le *Concept de vérité* une formulation de ce qui semble être le résultat principal de Tarski, on ne trouve rien de tel que la condition nécessaire et suffisante formulée à l'instant. On trouve bien deux ensembles de conclusions, l'un dans le corps principal du texte, et l'autre dans le postscript ajouté en 1935, qui semblent avoir un rapport étroit avec la thèse de la richesse essentielle, mais d'une part ces thèses ne mentionnent pas la notion de « richesse essentielle », et d'autre part les conclusions du postcript semblent contredire les conclusions antérieures. Ce point mérite donc une petite explication.

Commençons par les conclusions du postscript :

> A. Pour tout langage formalisé une définition formellement correcte et matériellement adéquate d'énoncé vrai peut être construite dans un

[39] Tarski écrit, par exemple :

> *... la condition de « richesse essentielle » du méta-langage apparaît être, non seulement nécessaire, mais également suffisante pour la construction d'une définition satisfaisante de la vérité ; ...*(Tarski (1944), p.352, souligné par Tarski)

> métalangage avec la seule aide d'expressions logiques générales, d'expressions du langage lui-même, et de termes de la morphologie du langage - mais à la condition que le métalangage possède un ordre plus élevé que le langage qui fait l'objet de l'investigation.
>
> B. Si l'ordre du métalangage est au plus égal à celui du langage lui-même, une telle définition ne peut pas être construite.(Tarski (1983), p. 273)

Il semble que soient données ici les conditions nécessaires et suffisantes d'une réponse positive à la question 1, mais il y a plus : la thèse A est une réponse directe, inconditionnelle, à la question 1. En effet, la condition qui apparaît à la fin de la longue phrase constituant le point A porte sur le type de métalangage dans lequel la définition sera possible, non sur l'existence d'un tel métalangage qui, elle, est affirmée. Mais évidemment, c'est une question de présentation, et l'on peut décomposer A en deux sous-thèses qui lui sont conjointement équivalentes :

- A.1 S'il existe un métalangage d'ordre plus élevé que L alors il est possible de donner la définition recherchée de la vérité pour L dans ce métalangage.[40]
- A.2. Un tel métalangage existe toujours.

Quant à la thèse B, elle affirme que l'on ne peut pas conduire cette définition dans un métalangage qui ne serait pas « d'ordre supérieur » à l'ordre du langage-objet, ce qui est en somme la réciproque de la thèse A.1. Autrement dit, on peut reformuler le couple de conclusions de Tarski par le couple équivalent :

- C : Il est possible de définir la vérité pour un langage L si et seulement si il existe un métalangage d'ordre supérieur.[41]
- D : Pour tout langage L, il existe un métalangage d'ordre supérieur.

Maintenant, la thèse C ressemble à s'y méprendre à ce que nous avons appelé le cœur de la réponse de Tarski à la question 1, à ceci près qu'il n'est pas question

[40] Une version un peu plus proche du texte de Tarski serait :

> S'il existe un langage d'ordre supérieur à L alors il est possible de définir vrai-dans-L dans ce métalangage à l'aide d'expressions logiques générales, d'expressions de L et de termes de la morphologie du langage

[41] Une version un peu plus proche du texte de Tarski serait :

> S'il existe un langage d'ordre supérieur à L alors : il est possible de définir vrai-dans-L dans un métalangage ML à l'aide d'expressions logiques générales, d'expressions de L et de termes de la morphologie du langage si et seulement si ML est d'ordre supérieur à L

de « plus grande richesse essentielle » mais « d'ordre supérieur ». On serait donc tenté de conclure que par « plus grande richesse essentielle », ce que Tarski veut dire c'est « d'ordre supérieur », en un sens à préciser. Ce qu'il faut entendre par langage « d'ordre supérieur » à un autre, au sens utilisé ici, Tarski l'explique dans le postscript, et le point est clarifié dans Rouilhan (1998). Pour faire court, l'idée générale est la suivante : L est d'ordre plus élevé que L' s'il y a dans L des variables qui prennent pour *valeur* le *parcours* des variables de L'. Dans le cas particulier où L et L' sont des langages basés sur la théorie des types, la différence d'ordre des langages est manifestée dans les différences syntaxiques entre ces langages : L possède des variables d'ordre syntaxique plus élevé que l'ordre de toute variable dans L'. Mais si L et L' sont des langages avec des variables d'« ordre indéterminé », comme c'est le cas des langages de « Zermelo et de ses successeurs », alors il n'y a pas de marqueur syntaxique de l'ordre des langages. La différence d'ordre entre deux langages, néanmoins, a deux contreparties claires. La première en terme de traduction d'un langage dans un autre : d'une part les quantifications non-bornées des énoncés du langage d'ordre le plus petit ont pour traduction dans le langage d'ordre le plus grand des énoncés dont tous les quantificateurs sont bornés et, d'autre part, il n'est pas possible de traduire les formules quantifiées du langage d'ordre le plus grand dans le langage d'ordre plus petit.[42] Une autre façon de distinguer deux langages d'ordres différents est de comparer les théories correctes dans ces langages : L est d'ordre supérieur à L' si et seulement si il y a dans L des théories T correctes essentiellement plus riches que toute théorie correcte T' de L', où une théorie T est essentiellement plus riche qu'une théorie T' ssi T admet

> un nouveau postulat, dont le contenu peut être décrit de la façon suivante : il y a un ensemble X dont les éléments fournissent un modèle pour le système de postulats originaux de $[T']$. (Tarski (1939), p.110)

Maintenant que nous avons clarifié la notion d'ordre d'un langage, tournons-nous vers les conclusions que présente Tarski dans le corps du *Concept de vérité*, avant 1935. Les deux conclusions qui nous intéressent sont les suivantes :

[42] Si bien que même dans le cas où le vocabulaire du métalangage contient le vocabulaire du langage-objet, la traduction d'une formule du langage-objet dans le métalangage n'est pas en général donnée par la transcription homographique, c'est-à-dire par l'énoncé syntaxiquement identique de l'autre langage. L'ignorance de ce point a parfois donné lieu à des incompréhensions et des confusions, par exemple dans De Vidi et Salomon (1999). Voir la mise au point de Ray (2005) avec les grandes lignes duquel nous sommes d'accord, et surtout Rouilhan (1998), auquel nous sommes redevables, de près ou de loin, pour l'ensemble de cette section.

> A$^-$. Pour tout langage formalisé d'ordre fini, une définition formellement correcte et matériellement adéquate de l'énoncé vrai peut être construite dans le métalangage, en faisant usage seulement d'expressions d'un genre logique général, d'expressions du langage lui-même ainsi que de termes appartenant à la morphologie du langage, c'est-à-dire des noms des expressions linguistiques et des relations structurales qui existent entre elles.
>
> B$^-$. Pour les langages formalisés d'ordre infini la construction d'une telle définition est impossible.(Tarski (1983), p.265)

A cette époque des conclusions, et pour des raisons dans lesquelles nous n'entrerons pas ici[43], Tarski n'a en vue que les langages basés sur la théorie des types, et c'est sous cette hypothèse qu'il faut comprendre les conclusions énoncées. Ces conclusions reviennent bien à donner une condition nécessaire et suffisante de la possibilité d'une définition de la vérité pour un langage L - nommément : il est possible de construire la définition recherchée de vrai-dans-L ssi L est d'ordre fini - mais l'origine de la condition n'est pas transparente. Pour faciliter la comparaison avec les conclusions ultérieures de Tarski, on ferait bien de reformuler celles-ci de la façon suivante :

> Supposons que tout langage soit basé sur la théorie des types :
>
> C^-. Il est possible de définir vrai-dans-L ssi il existe un langage d'ordre supérieur à L.
>
> D^-. Il existe un langage d'ordre supérieur à L si et seulement si L est d'ordre fini.

La thèse C$^-$ apparaît clairement à présent comme la contrepartie, dans le cas spécial des langages fondés sur la théorie des types, de ce que nous avons appelé le « cœur » de la réponse de Tarski à la question 1, et qui correspond à la thèse C des dernières conclusions, celles du Postscript telles que je les ai reformulées. C'est d'ailleurs quelque chose comme C^-, et rien d'autre, qui est présenté par Tarski comme le résultat principal de son travail dans le résumé de son adresse au Congrès de Paris de 1935 :

> Il me semble que ce problème [de la réduction de la sémantique à la morphologie au sens large] peut maintenant être regardé comme définitivement clos. Il apparaît être étroitement connecté à la théorie des types logiques.

[43] Voir les justification de Tarski au § 4 (Tarski (1983), p.215-221) et à nouveau les commentaires de Rouilhan (1998) pour une clarification.

> Le résultat principal pertinent pour cette question peut être formulé de la façon suivante :
>
> > *Il est possible de construire dans le métalangage des définitions méthodologiquement correctes et matériellement adéquates des concepts sémantiques si et seulement si le métalangage est équipé avec des variables de type logique supérieur à toutes les variables du langage qui est le sujet de l'investigation*
>
> (« L'établissement de la sémantique scientifique », dans Tarski (1983), p.406. Souligné par Tarski.)

Et de fait, Tarski notera souvent dans les reprises plus tardives de son travail que, dans ce cas particulier des langages fondés sur la théorie des types, on peut comprendre l'expression « être essentiellement plus riche que » comme signifiant « avoir des variables de type logique supérieur à toutes les variables de ».

Il semble que nous soyons arrivé au bout de nos peines. En nous tournant vers le *Concept de vérité*, nous avons finalement retrouvé le résultat dont les affirmations en termes de « plus grande richesse essentielle » étaient la glose, et ces résultats semblent indiquer que ce qu'il faut comprendre par là c'est « être d'ordre supérieur », au sens large indiqué par Tarski.[44]

Mais est-ce bien le dernier mot sur la question ? Pas tout à fait sans doute. La définition de l'ordre d'un langage devrait encore être élargie et sa portée clarifiée, pour pouvoir s'appliquer à d'autres classes de langages que ceux auxquels nous avons eu à faire jusqu'à présent. C'est ce que suggère par exemple la note 16 de Tarski (1944), où Tarski introduit ce qui semble être encore une nouvelle interprétation possible de la relation « être essentiellement plus riche » :

> Pour définir récursivement la notion de satisfaction, nous devons appliquer une certaine forme de définition récursive qui n'est pas admise dans le métalangage. *Par conséquent la « richesse essentielle » du métalangage peut simplement consister à admettre ce type de définition.* D'un autre côté, une méthode générale est connue qui rend possible d'éliminer toutes les définitions récursives et de les remplacer par des définitions normales, explicites.[45] Si nous appliquons cette méthode à la définition de la satisfaction, nous voyons que nous devons soit introduire dans le métalangage

[44] Et reconstruit, plutôt que repris, ici.

[45] N.d.T. Si « *Sat* » est un symbole de relation entre individu défini récursivement par un énoncé $\phi(Sat)$, on obtient une définition explicite de *Sat* en écrivant : $Sat(x, y)$ si et seulement $\forall R\phi(R)$,

> des variables de type logique supérieur à celles qui apparaissent dans le langage-objet ; ou de postuler axiomatiquement dans le métalangage l'existence de classes plus compréhensives que toutes celles dont l'existence peut être établie dans le langage-objet.(Tarski (1944), p.353, n.16. Je souligne.)

Que penser de la nouvelle condition énonçant que L peut être essentiellement plus riche que L' par le simple fait d'admettre un type de définition récursive que L' n'admet pas ? Discuter ce point nous amènerait trop loin, et nous préférons renvoyer le lecteur qui souhaiterait en savoir davantage sur les définitions inductives à la littérature spécialisée sur cette question[46] et nous laisserons ici de côté la tâche d'éclaircir la façon dont on doit comprendre l'expression « admettre un certain type de définition » pour un langage interprété. Plus généralement, si la relation « être essentiellement plus riche que » a été clarifiée pour les cas où les langages considérés sont des langages ordinaires avec des variables typées ou des variables d'ordre indéterminé, le problème de déterminer en toute généralité à quelles conditions nécessaires et suffisantes un langage est « essentiellement plus riche » qu'un autre est un problème ouvert.

Nous manquerions quelque chose d'important si, avant d'entrer dans la démonstration que Tarski donne de ce que nous avons appelé « le cœur de sa réponse », nous ne nous arrêtions pas à la réponse, ou plutôt aux réponses, que Tarski donne à la question principale elle-même. Sur ce point, nous ne pouvons faire mieux que suivre Rouilhan (1998). J'ai rappelé qu'il y avait deux époques des conclusions de Tarski, avant et après le postcript au *Concept de vérité*, et j'ai reformulé ces conclusions pour faire ressortir, dans chaque cas, ce qu'il appellera parfois son « résultat principal », C^- et sa version améliorée C. Restent les thèses D^- et D. La thèse D^-, avant 1935, est simplement une conséquence de l'hypothèse que tout langage est fondé sur la théorie des types : comme dans un tel langage le type de toute variable est fini, il n'y a que pour les langages d'ordre fini, c'est-à-dire ne contenant qu'un nombre fini de types de variable, que peut exister un langage d'ordre supérieur.

où R est un variable de même type que la relation Sat. Dans un langage ensembliste on définirait Sat comme « le plus petit ensemble » de couples satisfaisant $\phi(R)$ (s'il existe). Voir aussi plus loin section 3.2.2.2

[46] Voir par exemple McGee (1991) sur les aspects sémantiques des définitions par induction, ou le classique Moschovakis (2008). Pour la présentation de langages enrichis par des opérateurs spécifiques permettant les définitions par induction, nous renvoyons à l'introduction à Bucholz et al. (1981).

Beaucoup plus problématique est la question de la justification de la thèse D. Ce que montre Philippe de Rouilhan est que 1. la thèse D est une thèse philosophique (la négation de l'existence d'un langage universel) et qu'aucun argument purement logique donné par Tarski ne peut la justifier. 2. Que l'adhésion de Tarski à la thèse D après 1935 manifeste un changement philosophique par rapport au Tarski d'avant 1935. 3. Que la véritable nature de ce changement lui échappa. Tarski crut sans doute que, de la thèse D^- à la thèse D, il n'y avait rien d'autre que l'amélioration d'un résultat, rendue possible par la considération de nouveaux genres de langages. Mais la thèse D n'a rien à voir avec le genre de cadre logique, théorie des types, des ensembles, ou autre, dans lesquels nous pouvons choisir de coucher notre théorie la plus générale du monde. En particulier, il était ouvert à Tarski de considérer le langage de la théorie des ensembles, avec ses variables d'ordre indéterminé, comme un langage universel, un langage dans lequel les quantificateurs parcourent *tout ce qui est*. Et pour le langage de la théorie des ensembles ainsi conçu, variables d'ordre indéterminé ou non, la définition explicite de la vérité reste impossible, puisque cette dernière suppose l'existence d'un langage « essentiellement plus riche ».[47]

2.2.1 La démonstration du « résultat principal »

Nous avons vu que la réponse de Tarski à la question 1 était : c'est possible si, et seulement si, il existe un langage essentiellement plus riche que L. Nous avons expliqué la signification de cette condition, il reste maintenant à démontrer le résultat. La démonstration des deux implications n'est pas symétrique :

1. Si un tel métalangage existe, alors Tarski montrera comment construire une théorie de ce (méta-)langage et y définir adéquatement vrai-dans-L.
2. Réciproquement, s'il est possible de définir adéquatement la vérité pour L dans une théorie ML, Tarski montre que le langage de ML doit être essentiellement plus riche que L.

La démonstration du point 2. repose sur la formalisation de l'antinomie du Menteur, et je ne m'y attarderai pas. Dans « La conception sémantique de la vérité », Tarski résume les choses ainsi :

> Si la condition de « richesse essentielle » n'est pas satisfaite, on peut

[47] La raison en apparaîtra plus loin, section 3.2.2.2.

> usuellement[48] montrer qu'une interprétation du méta-langage dans le langage-objet est possible ; c'est-à-dire qu'à chaque terme du métalangage peut être corrélé un terme bien déterminé du langage-objet de sorte que les énoncés assertables d'un langage sont corrélés à des énoncés assertables de l'autre. Il résulte de cette interprétation que l'hypothèse qu'une définition satisfaisante de la vérité a été formulée dans le métalangage implique la possibilité de reconstruire dans ce langage l'antinomie du Menteur ; et ceci nous force à rejeter cette hypothèse.(Tarski (1944), p.351)

Pour une explication plus détaillée de ce résultat célèbre, je renvoie le lecteur au *Concept de vérité*, et plus précisément au Théorème I et à sa démonstration, dont c'est l'objet.[49]

Plus intéressante pour nous est la démonstration du point 1, c'est-à-dire la méthode de construction d'une définition de la vérité quand les conditions de sa possibilité sont réunies, dans une métathéorie d'un certain métalangage. C'est à présenter cette méthode que je vais consacrer la suite de cette section et la suivante. Etant donné un langage L, pour construire la métathéorie dans laquelle conduire la définition, nous dit Tarski, les ingrédients principaux sont les suivants :[50]

Le métalangage, supposé essentiellement plus riche L, doit contenir :

1. des expressions logiques-ensemblistes générales en quantité suffisante ;
2. des énoncés ayant la même signification que les énoncés de $\mathcal{L}$ (pour pouvoir exprimer les équivalences-T) ;
3. des expressions spécifiques permettant de décrire la morphologie/syntaxe de $\mathcal{L}$ (éventuellement à un codage près).

La théorie correspondante (la métathéorie) suffisante pour conduire la définition est obtenue avec trois groupes d'axiomes :

1. Un appareil logique général ;
2. Des axiomes ayant la même signification, ou sont plus forts, que ceux de la science étudiée ;
3. Les axiomes permettant de développer la théorie de la morphologie/syntaxe d'un langage.

[48] N.d.T : C'est-à-dire quand ce langage-objet contient « suffisamment d'arithmétique » selon la formule vague devenue usuelle.

[49] Tarski (1983), p. 247-251.

[50] Tarski (1983), p.211.

J'appelerai ces trois groupes d'axiomes les *axiomes logiques*, les *axiomes théoriques* et la *syntaxe*, respectivement. Pour montrer comment procède la définition à partir de ces éléments, la meilleure chose à faire est sans doute de prendre un exemple concret de construction. Le choix du langage de l'arithmétique est motivé par le fait que nous aurons besoin d'avoir des notations clairement fixées dans ce cas particulier pour la suite de ce travail.

2.2.2 Esquisse d'une définition de la vérité pour le langage de l'arithmétique

Suivant Tarski, les langages-objets pour lesquels je vais esquisser une définition de la vérité sont en fait des théories-objets.[51] L'arithmétique de Robinson et l'arithmétique de Peano sont deux théories arithmétiques partageant un langage commun (« langage » au sens moderne de ce mot ici), ce que l'on appellerait aujourd'hui le langage de l'arithmétique. Les symboles primitifs de ce langage, outre les symboles logiques et le symbole d'égalité, sont un symbole de fonction unaire S qui représente la *fonction successeur*, deux symboles de fonctions binaires, « + » et « . », représentant l'addition et la multiplication respectivement, et une constante d'individu « 0 », dénotant l'entier 0. Si elles partagent le même vocabulaire, et ont le même domaine de discours, celui des entiers naturels, ces deux théories n'ont pas la même force logique, l'arithmétique de Peano, notée PA, étant plus riche que l'arithmétique de Robinson, notée Q. J'en viens aux axiomes détaillés.

Arithmétique de Robinson (Q) :

Les axiomes de Q sont les suivants[52] :

1. $\forall x \neg(0 = Sx)$
   (0 n'est le successeur d'aucun nombre)
2. $\forall x \forall y(Sx = Sy \rightarrow x = y$
   (Si deux nombres ont le même successeur, alors ces nombres sont égaux)
3. $\forall x(\neg(x = 0) \rightarrow \exists y(x = Sy))$
   (Tout entier non nul est le successeur d'un nombre)

[51] Quitte à vérifier plus tard que la *définition* tarskienne de la vérité ne dépend nullement de la théorie en question mais bien seulement du langage dans lequel est couchée cette théorie.

[52] Je donne entre parenthèses une paraphrase de chaque axiome en langage ordinaire non, ou non seulement, pour faciliter la lecture, mais également pour rappeler que c'est d'axiomes interprétés qu'il s'agit ici.

4. $\forall x(x + 0 = x)$
   (La somme de 0 et d'un nombre est égale à ce nombre)
5. $\forall x \forall y(x + Sy = S(x + y))$
   (La somme d'un nombre et du successeur d'un autre nombre est égale au successeur de la somme de ces deux nombres)
6. $\forall x(x.0 = 0)$
   (Le produit d'un nombre et de 0 est nul)
7. $\forall x \forall y(x.S(y) = (x.y) + x)$
   (Le produit d'un nombre et du successeur d'un autre nombre est égal à la somme du produit des deux nombres et du premier nombre.)

Arithmétique de Peano (PA) :

Les axiomes de PA sont les suivants :

1. Les axiomes de Q
2. Un *schéma* d'axiomes d'induction :
   $[\phi(0) \wedge (\forall n(\phi(n) \rightarrow \phi(Sn)))] \rightarrow \forall x \phi(x)$
   où ϕ est n'importe quelle formule de $\mathcal{L}_{PA}$
   (Si 0 est ϕ et que, pour tout entier n, si n est ϕ alors le successeur de n est ϕ, alors tous les entiers sont ϕ)

2.2.2.1 Syntaxe

Nous avons maintenant besoin d'une théorie permettant de décrire la syntaxe de Q. Nous noterons Syn_Q cette théorie et nous l'appellerons « la syntaxe de Q ». Il convient de noter que ce que nous appelons, suivant Tarski, « la syntaxe », recouvre en fait deux choses bien distinctes du point de vue qui est le nôtre aujourd'hui : d'une part une description de la grammaire de ce que nous appellerions aujourd'hui le *langage* de Q, c'est-à-dire la morphologie du langage de Q, et d'autre part une description de la « syntaxe » de la *théorie* Q, ce que nous appellerions une théorie de la démonstration pour Q. Nous allons décrire la structure de Syn_Q en respectant cette distinction et en omettant les détails.[53]

Morphologie de L_Q Nous noterons L_{Syn_Q} le (méta)-langage de notre (méta)-théorie Syn_Q. Que doit contenir ce langage de la syntaxe ?

[53] Ces détails se trouvent dans Tarski (1935), §3.

Le langage de Q contient des symboles logiques (parenthèses, connecteurs, quantificateurs, variables, symbole d'identité) et des symboles non-logiques (0, le symbole de la fonction successeur, les symboles d'addition et multiplication). Pour décrire la grammaire de L_Q nous avons besoin de *nommer* ces différents symboles dans le métalangage. Le métalangage contiendra donc, au titre des symboles non-logiques, des constantes permettant de désigner les symboles de L_Q. Nous désignerons par v_k la k-ième variable de L_Q, et nous supposerons ici que les noms des autres symboles de L_Q sont obtenus simplement en surlignant ces symboles : le symbole « $\overline{S}$ » désignera le symbole de fonction « S », « $\neg$ » le symbole de la négation, etc.

Outre les constantes permettant de nommer les symboles de L_Q, notre métalangage contient des symboles pour les notions dont nous nous proposons de faire la théorie. Faire la morphologie de L_Q c'est d'abord dire quelles *suites* de symboles sont des *expressions* de L_Q. Pour ce faire, le métalangage contiendra donc, outre les symboles logiques et les noms des symboles de L_Q, un symbole d'opération binaire $*$ dénotant l'opération de concaténation de symboles de L_Q par laquelle des symboles sont attachés les uns aux autres pour former des suites de symboles. Il contiendra également un prédicat Exp dénotant l'ensemble des expressions de L_Q. Il apparaît que ces notions sont suffisantes pour pouvoir définir par la suite d'autres catégories d'expressions de L_Q, comme la catégorie des termes, des énoncés, etc.

La théorie de $*$ et Exp est gouvernée par cinq axiomes seulement :

1. Un axiome affirmant, de chacune des constantes logiques ou non-logiques de L_Q que ce sont des expressions, et que ces expressions sont distinctes.
2. Un axiome affirmant que les variables sont des expressions : pour tout entier naturel k, v_k est une expression distincte des expressions de (1) et $v_k \neq v_{k'}$ si $k \neq k'$.
3. Un axiome affirmant que $x * y$ est une expression si et seulement si x et y sont des expressions et que cette expression est distincte des expressions dans (1) et (2).
4. Un axiome à l'effet que si x, y, z, t sont des expressions, alors $x * y = z * t$ si et seulement si l'une des conditions suivantes est satisfaite :
   - $x = z$ et $y = t$
   - il existe $u \in Exp$ tel que : $x = z * u$ et $t = u * y$
   - il existe $u \in Exp$ tel que : $z = x * u$ et $y = u * t$

5. L'axiome principal (Axiome 5, « Principe d'induction », Tarski 1983, p.173) est l'axiome du second ordre suivant : la classe des expressions (Exp) du langage est la plus petite classe contenant les variables v_1,..., v_k, etc., les parenthèses $\overline{(}$ et $\overline{)}$, les constantes logiques $\equiv$, $\overline{\neg}$, $\overline{\wedge}$,$\overline{\forall}$,..., les constantes non-logiques $\overline{S}$, $\overline{+}$, $\overline{\cdot}$, $\overline{0}$ et close par l'opération de concaténation.[54]

Dans cette théorie il est possible de définir toutes les notions de la morphologie de L_Q.

La classe des termes de L_Q, $Term$, est définie récursivement de la façon suivante :

Définition 1 ($Term$)**.** *L'ensemble $Term$ est le plus petit ensemble tel que :*

1. $\overline{0} \in Term$,
2. *pour tout entier naturel k, $v_k \in Term$,*
3. *et si $u, v \in Term$ alors*
   - $\overline{S} * \overline{(} * u * \overline{)} \in Term$,
   - $\overline{(} * u * \overline{+} * v * \overline{)} \in Term$,
   - $\overline{(} * u * \overline{\cdot} * v * \overline{)} \in Term$.

La classe $Form$ des formules de L_Q pourrait alors être définie récursivement de la façon suivante :

Définition 2 ($Form$)**.** *$Form$ est le plus petit ensemble tel que :*

1. *Si $t_1 \in Term$ et $t_2 \in Term$ alors $t_1 * \equiv * t_2 \in Form$,*
2. *si $u \in Form$ et $v \in Form$ alors*
   - $\overline{\neg} * u \in Form$,
   - $\overline{(} * u * \overline{\wedge} * v * \overline{)} \in Form$,
   - $\overline{(} * u * \overline{\vee} * v * \overline{)} \in Form$,
   - $\overline{(} * u * \overline{\Rightarrow} * v * \overline{)} \in Form$,
3. *Si $v \in Form$ et v_k est une variable*
   - $\overline{\forall} * v_k * v \in Form$,
   - $\overline{\exists} * v_k * v \in Form$.

[54] On pourrait formuler la syntaxe dans un théorie du premier ordre et sans vocabulaire ensembliste (on a en revanche besoin d'arithmétique, comme le montre par exemple la définition de Var). L'axiome prinicipal doit alors être formulé comme un schéma d'axiomes affirmant en substance que si une formule ouverte à une variable libre est vraie des variables, des parenthèses, des constantes logiques et non logiques et que, si elle est vraie de x et de y, alors est vraie de $x * y$, alors elle est vraie de toute expression.

Nous pourrions également définir les notions de termes et de formules clos et ouverts, de substitution d'une variable dans une formule, d'énoncé de L_Q, etc. Nous laissons les détails de côté et renvoyons au texte de Tarski lui-même.

Théorie de la démonstration pour Q Nous avons décrit la grammaire de L_Q, il reste à présent à décrire la théorie Q. Décrire la théorie Q, c'est donner une description de ce qui compte comme preuve dans Q. Il s'agit donc de définir la notion d'axiome de Q et celle de théorème de Q. Il est facile de voir que ces notions peuvent être définies explicitement dans la morphologie que nous avons esquissée. À titre d'exemple, la définition de l'ensemble Ax_Q des axiomes de Q aura la forme suivante :

Définition 3 (Ax_Q (esquisse))**.** $x \in Ax_Q$ *ssi*

1. $(x = \overline{\forall} * v_1 * \overline{\neg} * \overline{(} * v_1 * \equiv * \overline{S} * \overline{0} * \overline{)} \quad \vee$
2. $(x = \overline{\forall} * v_1 * \overline{\forall} * v_2 * \overline{(} * \overline{S} * v_1 * \equiv * \overline{S} * v_2 * \overline{\Rightarrow} * v_1 * \equiv * v_2 * \overline{)} \quad \vee \ldots$
3. *etc.*

De même, avec les seules ressources du langage de la syntaxe, il est possible de définir l'opération de substitution d'un terme à une variable dans une formule, l'ensemble $Cn(X)$ des conséquences (nous dirions « déductives » ou « syntaxiques ») d'un ensemble X de formules, et l'ensemble Pr_Q des énoncés prouvables dans Q, c'est-à-dire des théorèmes de Q.[55]

C'est cet ensemble d'axiomes et de définitions, portant aussi bien sur le langage de Q à proprement parler que sur la théorie Q, que nous appelons la syntaxe de Q et que nous notons Syn_Q.[56]

[55] Voir Tarski (1983), p.181. Supposons pour simplifier les choses que Q ne contienne qu'une unique règle d'inférence, le *modus ponens*. On pourrait alors définir Pr_Q comme le plus « petit » prédicat P satisfaisant :

1. $\forall x\ (Ax_Q(x) \rightarrow P(x))$
2. $\forall x[\exists y, z(En(y) \wedge En(z) \wedge z = y * \overline{\Rightarrow} * x \wedge P(y) \wedge P(z)) \rightarrow Pr(x)]$

Si toutefois nous ne disposons pas des moyens expressifs (théorie des ensembles ou logique du second ordre) permettant d'exprimer le fait que Pr_Q est le plus petit prédicat (ou ensemble) satisfaisant les axiomes précédents (ou des axiomes analogues obtenus en remplacant $P(x)$ par $\in p$) et de transformer ainsi ces axiomes en définition explicite, nous devons nous rabattre sur un système d'axiomes. On ajouterait alors aux axiomes précédents le schéma d'axiome (d' « induction ») suivant :

$$\forall x(Ax_Q(x) \rightarrow \Phi(x)) \wedge \forall x \forall y(\Phi(y) \wedge \Phi(y * \overline{\Rightarrow} * x) \rightarrow \Phi(x)) \rightarrow \forall x(Pr_Q(x) \rightarrow \Phi(x))$$

[56] Il est à noter que la théorie Syn_Q est structurellement identique à une théorie de l'arithmétique,

2.2.2.2 Définition de la vérité

Nous en venons maintenant à la définition de la vérité proprement dite. Si notre langage-objet ne contenait qu'un nombre fini d'énoncés $x_1 \cdots, x_n$ la définition suivante, constituerait une définition adéquate[57] :

$$x \in Vrai \text{ ssi } (x = x_1 \wedge \phi_1) \vee \cdots \vee (x = x_n \wedge \phi_n),$$

où "ϕ_1",$\cdots$,"ϕ_n" doivent être remplacés par les traductions des énoncés désignés par "x_1",$\cdots$,"x_n" respectivement.

Néanmoins, si le langage pour lequel nous souhaitons définir la vérité contient une infinité d'énoncés, une telle définition n'est pas possible. Les langages formalisés ayant une grammaire récursive, l'idée d'une définition inductive de la vérité « se présente d'elle-même » [58], mais se heurte immédiatement à une difficulté : la vérité d'un énoncé n'est pas définissable en terme de la vérité de ses sous-formules, ces sous-formules n'étant pas en général des énoncés, et donc n'étant pas susceptibles d'être vraies ou fausses. Comment définir la vérité de l'énoncé $\forall x \phi(x)$ en fonction de ϕ, « x » étant libre dans ϕ ? Parce que l'ensemble des *formules* est, lui, défini inductivement, Tarski propose de définir d'abord une notion plus générale que celle de vérité d'un énoncé, celle de *satisfaction* d'une formule par une assignation d'objets aux variables du langage. Intuitivement la notion de satisfaction d'un prédicat par un objet est assez claire : on dira que le prédicat *Rouge* est satisfait par un objet o si et seulement si o est rouge. Plus généralement on dira que la formule $\phi(x_1, \cdots, x_n)$ (où $x_1, \cdots, x_n$ sont les seules variables libres) est satisfaite par une assignation σ si et seulement si les objets $o_1, \cdots, o_n$ assignés par σ aux variables $x_1, \cdots, x_n$ sont dans la relation dénotée par l'expression ϕ de L, et où une assignation f est une fonction qui à toute variable de L associe un (unique) objet de l'univers de discours de L, les entiers dans le cas qui nous intéresse.[59]

Pour définir la satisfaction pour notre langage L_Q contenant des symboles de fonction, nous définissons d'abord une troisième notion, celle de dénotation d'un

comme noté par Tarski. C'est ce qui permet d'interpréter la syntaxe d'une théorie dans l'arithmétique (à la façon de Gödel (1986)), ou l'arithmétique dans la syntaxe (c'est le chemin suivi par Quine (1950)).

[57] Tarski 1983, p.188.

[58] C'est l'expression de Tarski. Voir Tarski (1983).

[59] On pourrait définir une assignation comme une fonction qui associe à toute suite *finie* de variables un objet de l'univers de discours de L. Comme les suites finies sont des objets qui peuvent être représentés dans l'arithmétique, ce point montre que la nécessité que notre métathéorie puisse définir les fonctions d'assignations n'est pas une demande très forte.

terme pour une assignation. La dénotation d'un terme de L pour une assignation f est définie par induction :

Définition 4 (Dénotation d'un terme)**.** *1.* $Den(v_k, f) = f(v_k)$

2. si $t_1, t_2 \in Term$*, si* $Den(t_1, f) = x_1$ *et* $Den(t_2, f) = x_2$ *alors*
- $Den(\overline{S} * \overline{(} * t_1 * \overline{)}, f) = S(x_1)$,
- $Den(\overline{(} * t_1 * \overline{+} * t_2 * \overline{)}, f) = x_1 + x_2$,
- $Den(\overline{(} * t_1 * \dot{-} * t_2 * \overline{)}, f) = x_1.x_2$

La définition récursive de la satisfaction d'une formule x par une assignation f s'obtient alors de la façon suivante :

Définition 5 (Satisfaction d'une formule)**.** $Sat(x, f)$ *ssi* x *est une formule,* f *une assignation, et :*
- *il existe* $t_1, t_2 \in Term$*,* $x = t_1 * \equiv * t_2$*, et* $Den(t_1, f) = Den(t_2, f)$*,*
- *ou il existe* y*,* $x = \neg * y$ *et* $\neg Sat(y, f)$
- *ou il existe* y, z *tels que* $x = \overline{(} * y * \overline{\vee} * z * \overline{)}$ *et* $Sat(y, f)$ *ou* $Sat(z, f)$
- *ou ...(clauses attendues pour les autres connecteurs)*
- *il existe une formule* y *et une variable* v *tels que* $x = \overline{\forall} * v * y$ *et pour toute assignation f' identique à f sauf peut-être en* v *:* $Sat(y, f')$*.*

Pour transformer cette « définition » récursive en définition explicite, il est nécessaire que le métalangage soit « essentiellement plus riche » que le langage-objet. On procède alors de la façon suivante. Appelons $\phi(Sat)$ la formule précedente. On obtient la définition explicite de la satisfaction en écrivant :

$$Sat(x, y) \text{ ssi } \forall R(\phi(R) \rightarrow R(x, y))$$

ou bien, en langage ensembliste,

$$Sat(x, y) \;\; ssi \;\; \forall z(\phi(z) \rightarrow (x, y) \in z)$$

où $\phi(z)$ est la formule précédente dans laquelle on a remplacé le prédicat Sat par l'appartenance à z. Il y a d'autres possiblités mais la morale est la même. Considérons la seconde forme, ensembliste. Elle définit Sat comme le plus petit ensemble satisfaisant ϕ. Mais cet ensemble existe-t-il ? Tout dépend. Si le langage-objet est universel au sens où ses quantificateurs parcourent tout ce qui est, cet ensemble

n'existe pas.[60] Donc il n'y a rien de tel que "le plus petit ensemble" satisfaisant les conditions requises, et nous avons échoué à transformer la définition inductive en définition explicite.

A partir de la définition de la satisfaction, il est facile de définir la notion de vérité, de la façon suivante :

Définition 6 (Vérité)**.** *x est un énoncé vrai de L_Q ssi il est satisfait par toute assignation.*[61]

Dans le cas du langage de l'arithmétique, comme nous l'avons noté tout à l'heure, la définition peut être obtenue de façon beaucoup plus simple, sans le détour par la notion de satisfaction en tirant parti du fait que tous les entiers ont un nom. L'idée est simplement que, dans un tel langage, un énoncé universel est vrai si et seulement si toute ses instances de substitution le sont.[62]

Définition 7. *La notion de dénotation d'un terme clos de L_Q est d'abord simplifiée :*

- $Den(\overline{0}) = 0$
- *si t_1, t_2 sont des termes clos, et si $Den(t_1) = x_1$ et $Den(t_2) = x_2$ alors*
  - $Den(\overline{S} * \overline{(} * t_1 * \overline{)}) = S(x_1)$,
  - $Den(t_1 * \overline{+} * t_2) = x_1 + x_2$,
  - $Den(t_1 * \overline{\cdot} * t_2) = x_1.x_2$

Puis la notion de vérité d'un énoncé :

Définition 8. *L'énoncé x est vrai si et seulement si :*

- *il existe t_1, t_2 des termes clos tels que, $x = t_1 * \overline{\equiv} * t_2$, et $Den(t_1) = Den(t_2)$,*
- *ou il existe un énoncé y tel que $x = \overline{\neg} * y$ et $\neg Vr(y)$*
- *ou il existe des énoncés y et z tels que $x = \overline{(} * y * \overline{\vee} * z * \overline{)}$ et $Vr(y)$ ou $Vr(z)$*
- *ou ...(clauses attendues pour les autres connecteurs)*
- *il existe une formule y à une variable libre v telle que $x = \overline{\forall} * v * y$ et pour tout énoncé y' obtenu à partir de y en substituant à v un terme clos t, $Vr(y')$.*

[60] Pour s'en convaincre : la classe des couples dont le premier élément est $\overline{x = x}$ et qui appartiennent à Sat n'est pas un ensemble, $\overline{x = x}$ est satisfaite par toute assignation, et que la classe de toutes les assignations n'est pas un ensemble.

[61] Tarski (1983), p.195

[62] Pour un traitement systématique dans cet esprit, voir McGee (1991).

Nous sommes donc parvenus à la conclusion de la construction avec la définition de la vérité. La définition est-elle adéquate ? On pourrait montrer qu'elle l'est en raisonnant par induction sur la complexité des formules dans une méta-métathéorie.[63]

2.3 Conséquences de la définition

Tout d'abord la définition *à la Tarski* d'un prédicat de vérité offre bien ce que Tarski en attend. Elle est formellement correcte et satisfait la convention-T. Quand elle est possible, la définition montre donc qu'un usage cohérent du concept de vérité est possible. On se demandera peut-être, *quid* de l'énoncé du Menteur alors ? Après tout l'énoncé du Menteur est un énoncé bien formé du métalangage.[64] Il apparaît que l'énoncé du Menteur est un *théorème* de la métathéorie : $MT \vdash \neg Vr(\ulcorner\lambda\urcorner)$. La raison en est simple : puisque λ n'est pas un énoncé du langage-objet, et que dans la métathéorie il est possible de prouver à partir de la définition de la vérité que tout énoncé vrai est un énoncé du langage-objet (deuxième clause de la convention-T), on peut prouver dans la métathéorie que λ n'est pas vrai.[65] La cohérence de la métathéorie est néanmoins préservée, tout simplement parce que l'équivalence-T correspondante n'est pas, elle, un théorème de la métathéorie, i.e. $MT \nvdash Vr(\ulcorner\lambda\urcorner) \leftrightarrow \neg Vr(\ulcorner\lambda\urcorner)$.[66]

Il faut ensuite noter que la définition de Tarski ne donne pas de *critère* de décision pour la vérité dans le langage-objet.[67] Certes, nous avons une définition de l'ensemble des énoncés vrais du langage-objet dans la métathéorie, mais cette définition ne nous permet pas en général de *décider*, dans la métathéorie, de la vérité ou de la fausseté de tous les énoncés du langage-objet, loin s'en faut. Nous ne

[63] Tarski (1983), p.195. Dans la section 3.5.1 on montrera sur un exemple comment dériver une équivalence-T particulière dans la méta-théorie. La preuve générale dans la méta-métathéorie montre que le mécanisme simple à l'œuvre dans cette preuve est généralisable.

[64] Nous simplifions un peu les choses ici. Il faudrait préciser ce que nous appelons l'énoncé du Menteur. Les points essentiels sont les suivants : 1. On peut étendre la syntaxe dans notre métalangage ML de façon à pouvoir y décrire la morphologie de ML, et non seulement celle de L. 2. Dans cette théorie, d'après le lemme de diagonalisation, il existe un énoncé λ tel que $MT \vdash \lambda \leftrightarrow \neg Vr(\ulcorner\lambda\urcorner)$. 3. Cet énoncé λ appartient au méta-langage et non au langage-objet : sinon on aurait $MT \vdash Vr(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$, d'après 2., MT serait incohérente.

[65] $MT \vdash \forall x(Vr(x) \rightarrow x \in Form_L)$. Donc par contraposition, si l'on a une preuve qu'un énoncé λ n'est pas un énoncé de L, on peut prouver dans ML que λ n'est pas vrai.

[66] Un point analogue était déjà noté en passant dans Tarski (1939). On pourra consulter Ketland (2000) pour une présentation détaillée récente. (Ketland semble ignorer le texte de Tarski (1939).)

[67] Voir sur ce point Rivenc (2001).

pourrons trancher que pour ces énoncés de la théorie-objet qui sont prouvables ou réfutables... dans la métathéorie (à la traduction près) ! Une chose est à noter cependant. Puisque le métalangage est « essentiellement plus riche » que le langage objet, il offre des moyens expressifs que le langage-objet n'offre pas, par exemple des quantificateurs et des variables d'un type logique plus élevé que ceux disponibles dans le langage-objet. Par conséquent la possibilité est ouverte, dans la métathéorie, de formuler des principes gouvernant les notions du langage-objet (ou leur traduction) essentiellement plus riches que ceux qui étaient disponibles dans la théorie-objet. Dans ces conditions, nous serons souvent capables, dans la métathéorie, de trancher la question de la vérité d'énoncés laissés indécidés par la théorie-objet, *en renforçant nos axiomes* à l'aide des nouveaux moyens expressifs à notre disposition. Nous aurons l'occasion, tout au long des chapitres suivants, de revenir sur le sens épistémologique de ce processus de construction d'une métathéorie au terme duquel nous sommes en position de décider davantage d'énoncés que dans la théorie-objet.

Quoi qu'il en soit de ce dernier point, il est néanmoins possible de prouver dans la métathéorie un certain nombre de faits intéressants concernant le langage-objet et la théorie-objet, lorsqu'on a défini la vérité dans une méta-théorie construite selon les indications de Tarski :[68]

1. *Principe de non contradiction pour le langage-objet.* Pour tout énoncé x, $\neg(x \in Vr)$ ou $\neg(\neg * x \in Vr)$
2. *Principe du tiers-exclu pour le langage-objet.* Pour tout énoncé x, $x \in Vr$ ou $\neg * x \in Vr$[69]
3. *Les inférences dans la théorie-objet préservent la vérité.* Si $X \subseteq Vr$ alors $Cn(X) \subseteq Vr$.
4. *Les énoncés prouvables dans la théorie-objet sont vrais.* Si A est la théorie-objet on a : $Pr_A \subseteq Vr$ ou $\forall x\ (Pr_A(x) \rightarrow Vr(x))$
5. L'ensemble Pr_A des énoncés prouvables dans A est cohérent
6. Dans le cas où la théorie-objet A satisfait les conditions du premier théorème d'incomplétude de Gödel[70] alors de plus :

[68] Tarski (1983), p.197-199.

[69] On parle parfois de principe de bivalence ici, pour distinguer le principe sémantique du principe logique. Je m'en tiendrai à la terminologie de Tarski, non seulement par souci de fidélité au texte de Tarski, mais également pour une raison qui apparaîtra au chapitre 7.

[70] C'est-à-dire si A est une théorie récursivement axiomatisée dans laquelle il est possible d'interpréter la théorie Q.

Il y a un énoncé arithmétique vrai non prouvable dans Q (notre théorie objet ici). $Vr \nsubseteq Pr_A$.[71]

Il importe de noter que ces conséquences de la définition ne sont pas toutes de même nature. Les deux premières propositions (Non Contradiction et Tiers-exclu) sont seulement à propos du *langage* de la théorie objet, et affirment que l'ensemble des énoncés vrais de ce langage est cohérent et complet ; j'appelerai ces énoncés les *généralisations logiques*, parce qu'on peut les voir comme de simples formulations, à l'aide du prédicat de vérité, de principes purement logiques[72]. Les propositions suivantes en revanche concernent la *théorie*-objet, et je les appellerai, faute d'un meilleur terme, et quand le contexte sera suffisamment clair, les *généralisations théoriques*.[73] A cette distinction correspond une distinction dans les ressources de la méta-théorie qui sont nécessaires à leur preuve. Les généralisations logiques sont dérivables de la syntaxe et des axiomes logiques de la méta-théorie, tandis que la preuve des généralisations théoriques doit également faire appel à la troisième classe d'axiomes, ceux que nous avons appelés[74] les axiomes théoriques.[75]

2.4 Théories axiomatiques de la vérité

Nous avons vu que la définition explicite de la vérité pour un langage n'était pas toujours possible.[76] Néanmoins, même pour les langages ce genre, un usage cohérent

[71] « Preuve » : Tous les énoncés prouvables de Q sont vrais. L'ensemble des énoncés prouvables de Q (Pr_Q) est récursivement énumérable. L'ensemble des énoncés vrais n'est pas arithmétiquement définissable, donc n'est pas récursivement énumérable. Pr_Q est donc un sous-ensemble strict de l'ensemble des énoncés vrais de L_Q.

[72] Souvenons-nous des remarques de Quine sur le rôle nécessaire du prédicat de vérité dans l'expression des généralités logiques, chap. 2, section 4. Nous reviendrons à nouveau sur ce point au chapitre 7.

[73] J'essaierai d'employer cette terminologie le moins souvent possible, mais elle aura tout de même son utilité par la suite. La dénomination de "généralisation théorique" est quelque peu discutable, en outre, dans la mesure où j'ai appelé "axiomes théoriques" les axiomes de la méta-théorie tarskienne qui sont des traductions ou sont plus forts que ceux de la théorie-objet. Or les généralisations théoriques dont il est question à présent ne portent pas en général sur les mêmes objets que ces axiomes, mais concernent la syntaxe des preuves de la théorie-objet. Mais voir la remarque suivante pour une justification.

[74] Voir section 3.2.1.

[75] Nous revenons plus en détail sur les ressources méta-théoriques variées qui doivent être mobilisées pour dériver ces différentes prospositions dans les sections 3.4 et 3.5.

[76] Elle ne l'est pas si le langage-objet est un langage typé (au sens de la théorie des types) d'ordre infini. Mais elle ne le sera pas non plus, et là nous nous éloignons de Tarski, si avec Quine on refuse l'usage de la logique du second-ordre et d'ordres supérieurs et que l'on s'intéresse simplement à la vérité pour un langage capable d'exprimer la *généralité absolue*, c'est-à-dire dont les variables

du concept de vérité est rendu possible

> en incluant ce concept dans le système des concepts primitifs du métalangage et en déterminant ses propriétés fondamentales au moyen de la méthode axiomatique.(Tarski (1983), p.266)

Autrement dit, si A est une théorie formulée dans un langage pour lequel il n'est pas possible de définir la vérité, nous pouvons néanmoins construire ce que j'appellerai une *extension aléthique* de A comportant, outre les axiomes de A (ou des axiomes ayant même signification ou plus forts que ceux de A), une théorie axiomatisée de la syntaxe de A, et un système d'axiomes gouvernant l'usage d'un prédicat de vérité donné comme primitif. On dira que l'extension aléthique $T(A)$ d'une théorie A est *adéquate* pour un langage L si et seulement si les équivalences-T décrites dans la première clause de la convention-T sont des conséquences déductibles de $T(A)$.

Bien entendu, nous retrouvons ici un résultat négatif : si L contient un prédicat de vérité à titre primitif, il n'y a pas de théorie A formulée dans le langage L et telle que : (1) A est assez forte pour décrire sa propre syntaxe, (2) le prédicat de vérité serait adéquat dans A pour L lui-même. Sinon, A serait sémantiquement close, et par conséquent incohérente. La question est donc : étant donné un langage L pour lequel on ne peut pas définir la vérité, peut-on donner une théorie de la vérité adéquate pour L ? La réponse est oui, et la théorie est, bien entendu, formulée dans un langage qui étend strictement L. Dans la monographie de 1935, Tarski envisage deux types d'extensions aléthiques axiomatiques d'une théorie de base A, que j'appellerai extension aléthique minimale et maximale respectivement :

1. ***M(A)*, l'extension minimale de *A*** : $A+ Syn_A$ + eq-T. On obtient l'extension aléthique minimale de la théorie A, que je noterai $M(A)$, en augmentant notre théorie de base A d'un prédicat primitif « Vr » gouverné par l'ensemble infini des équivalences-T (correspondant au langage de A), de l'axiome $Vr(x) \rightarrow X \in En$ (que nous omettrons systématiquement sans dommage) et de la syntaxe du langage de A. Quand nous parlerons de $M(A)$ nous entendrons donc la théorie dont les axiomes sont les suivants :

   a) Les axiomes de A ;

parcourent tout ce qui est. Si les variables parcourent tout ce qui est, elles parcourent tous les ensembles, et nous avons vu plus haut (suite à la définition 5) pourquoi la définition n'est pas possible dans ce cas. Voir plus bas la conclusion de la première partie de ce travail.

b) Les axiomes de Syn_A ;

c) Toutes les instances du schéma suivant

$$Vr(\ulcorner\phi\urcorner) \leftrightarrow \phi \quad (\text{ pour } \phi \in L_A)\,;$$

d) (+ les nouvelles instances des schéma de A et Syn_A obtenues par substitution aux lettres schématiques de formules du langage au vocabulaire étendu de la nouvelle théorie, contenant Vr.)

2. ***$M_\omega(A)$*, l'extension maximale de *A*** : A+Syn_A + eq-T + ω-règle. L'extension maximale de A s'obtient à partir de l'extension minimale par ajout d'une nouvelle règle d'inférence d'un caractère particulier, que Tarski appelle règle *d'induction infinie*. Ce qui rend cette règle particulière, c'est qu'il s'agit d'une règle infinitaire. On peut la décrire de la façon suivante :

> Si $\phi(x)$ une formule à une variable libre, et que tous les énoncés qui s'obtiennent à partir de cette formule en substituant à la variable x le nom d'une expression du langage objet sont des théorèmes de la métathéorie, alors l'énoncé « $\forall$ x, si x est une expression alors $\phi(x)$ » est un théorème.[77]

L'intérêt de l'extension minimale est qu'il est possible de prouver qu'elle est cohérente (relativement à $A+Syn_A$). C'est le contenu du théorème III dans Tarski (1935). En fait, à y regarder de près, la preuve du théorème, donnée par Tarski, dit plus que cela. On peut en effet en déduire que l'extension de $A + Syn_A$ par les équivalences-T est une extension *conservative*, ce qui signifie que l'ajout des équivalences-T ne permet de dériver aucun *nouveau* théorème du langage ne contenant pas le prédicat de vérité.[78] Dans l'extension minimale, le prédicat de vérité est donc adéquat et et son usage n'introduit aucune contradiction dans le langage. Mais pour Tarski, cette extension aléthique est cependant insatisfaisante à deux égards au moins. Tout d'abord, elle ne permet pas de démontrer les principes de non contradiction, du tiers-exclu, etc., ce que nous avons appelé les "généralisations logiques", mais seulement les instances de ces propositions. Par exemple :

[77] Voir (Tarski (1983), p.259)

[78] Preuve rapide : toute dérivation dans la métathéorie ne peut avoir parmi ses prémisses qu'un nombre fini d'équivalences-T. Mais la vérité pour un nombre fini d'énoncés est définissable explicitement dans $A + Syn_A$ (Voir section ?). On peut donc prouver la conclusion de toute preuve de $A + Syn_A + eq - T$ dans $A + Syn_A$. Pour la version de Tarski, voir Tarski (1983), p.256.

« La neige est blanche » est vrai ou « La neige n'est pas blanche » est vrai

peut être dérivé de la façon suivante :

1. « La neige est blanche » est vrai si et seulement si la neige est blanche [Instance du Schéma-T]
2. « La neige n'est pas blanche » est vrai si et seulement si la neige n'est pas blanche [idem]
3. La neige est blanche ou la neige n'est pas blanche [logique propositionelle]
4. « La neige est blanche » est vrai ou « La neige n'est pas blanche » est vrai [1, 2, 3 et logique]

Quant à ce que nous avons appelé les généralisations théoriques, *il n'est même pas vrai que toutes leurs instances sont prouvables dans l'extension aléthique minimale de la théorie de départ*. Pour le voir, supposons que notre théorie-objet est l'arithmétique de Peano PA, et que notre théorie de la syntaxe soit équivalente à PA. Et considérons à titre d'exemple le principe selon lequel

Si un énoncé est prouvable dans PA alors il est vrai

dont nous avons vu tout à l'heure qu'il était dérivable dans la métathéorie dans laquelle Tarski montre comment définir le prédicat de vérité pour L_{PA}.[79] Soit maintenant S un énoncé indécidable de PA, et l'instance suivante du principe général ci-dessus :

Si S est prouvable dans PA alors S est vrai

On pourrait prouver que cette instance n'est pas prouvable l'extension aléthique minimale de PA.[80]

D'un autre côté, ces propositions générales sont correctes ; par conséquent la théorie minimale de la vérité proposée ne semble pas complète.[81] La seconde insuffisance relevée par Tarski est que la théorie ne détermine pas de manière univoque

[79] Puisqu'il se trouve que pour ce langage particulier de l'arithmétique de Peano en premier ordre une définition explicite est en fait possible dans une métathéorie.

[80] D'un côté, si S avait *en fait* été prouvable dans Q, alors nous aurions pu prouver à la fois que S est prouvable dans Q et que S est vrai ; d'un autre côté l'impossiblité de prouver le conditionnel quand S n'est pas prouvable est une conséquence simple du Théorème de Löb. Ce dernier affirme en effet que si le prédicat de prouvabilité satisfait certaines conditions « naturelles », alors : $PA \vdash Pr_{PA}(\ulcorner \phi \urcorner) \rightarrow \phi$ ssi $PA \vdash \phi$. (Pour une présentation du théorème de Löb, et le détail des hypothèses que doit satisfaire le prédicat de prouvabilité, voir par exemple Smorynski (1977).)

[81] Non pas au sens logique que ce terme peut prendre, mais au sens simplement d'une théorie qui échoue à saisir certains principes dont on juge qu'elle devrait rendre compte.

l'extension de « Vr » au suivant : en introduisant un prédicat « Vr' » gouverné par des axiomes analogues à ceux gouvernant « Vr », il n'est pas possible de prouver que « Vr=Vr' ».[82] Par conséquent on ne peut pas, selon Tarski, considérer que les axiomes définissent (même implicitement) une notion, faute en somme, de satisfaire une condition minimale de toute définition, à savoir l'unicité.

La seconde théorie, celle que j'ai appelée l'extension aléthique maximale de A, présente l'avantage de compter parmi ses théorèmes toutes les généralisations qui manquent à la précédente. Elle est catégorique en « Vr » au sens où elle en fixe l'extension. Mais cette théorie soulève au moins deux difficultés : la première est de savoir si l'usage d'une telle règle a réellement sa place « dans les limites de la conception existante de la méthode déductive ».[83] La seconde diffuculté évoquée par Tarski, plus grave si l'on ose dire, est qu'il n'est pas en mesure de prouver que la métathéorie ainsi obtenue est cohérente.[84] Or le problème de la vérité était d'abord d'indiquer sous quelles conditions un usage cohérent du concept est possible. La considération de la théorie maximale ne donne aucune garantie dans ce sens et est écartée. Tarski s'en tient donc, pour les langages tels qu'il n'en existe pas d'essentiellement plus riches et pour lesquels une définition n'est pas envisagable, à leur extension aléthique minimale.

Il y a pourtant une troisième théorie que Tarski ne mentionne pas, et dont la considération semblait devoir aller de soi. Il s'agit de l'extension aléthique de la théorie-objet A par une théorie de la syntaxe de A et d'axiomes gouvernant le prédicat primitif « Sat » gouverné par les « axiomes récursifs » extraits directement de la définition de la satisfaction donnée plus haut. Nous appellerons la théorie constituée de ces clauses récursives la théorie récursive de la vérité. Et parce que la considération de l'extension d'une théorie de base par ces clauses tarskiennes occupera souvent notre attention par la suite, il sera utile du lui donner nom : $T(A)$ désignera le genre de théorie qui résulte de l'extension d'une théorie A par une théorie de la syntaxe et les axiomes récursifs de la vérité ou de la satisfaction.[85]

[82] Tarski (1983), p.258. C'est la catégoricité du système qui intéresse Tarski, mais il note que dans le cas présent cette dernière est équivalente à la condition proposée ici.

[83] Tarski (1983), p.260.

[84] Tarski (1983), p.261-262.

[85] Par conséquent, également, quand c'est nécessaire, avec une théorie de la dénotation (présentée plus haut). Les propriétés sémantiques et preuves théoriques de ce genre d'extension ont souvent été étudiées dans la littérature logique post-tarskienne. Pour des exemples et des références voir Feferman (1991), Kaye (1991), Kotlarski (1991), Halbach (1999*a*) ou encore, dans un contexte plus philosophique Parsons (1983), p.213. C'est aussi bien sûr le genre de « théorie de vérité »

T(A)*, l'extension tarskienne de *A : Les axiomes de l'extension aléthique tarskienne de A, notée $T(A)$, sont les suivants :

1. Les axiomes de A
2. Les axiomes de Syn_A
3. Les axiomes récursifs pour la vérité dans L_A
4. (+ le cas échéant, les nouvelles instances des schémas de A et de Syn_A obtenues par substitution aux lettres schématiques de formules du vocabulaire étendu de $T(A)$.)[86]

A titre d'exemple, $T(ZF)$ serait ainsi axiomatisée de la façon suivante :

Définition 9 ($T(ZF)$)**.** *Le langage est celui de la théorie des ensembles et de la syntaxe augmenté d'un symbole de relation binaire $Sat_{L_\in}$ (nous omettrons l'indice en pratique). Le symbole s dénotera une assignation, c'est-à-dire une fonction de l'ensemble des variables de $L_\in$. La théorie $T(ZF)$ est constituée des axiomes suivants :*

- *Les axiomes de ZF (avec schémas étendus)*
- *Les axiomes de Syn_{ZF}* [87]

Et les axiomes gouvernant la relation de satisfaction proprement dite :

1. $Sat(x, y) \rightarrow$ *x est une formle de L_{ZF} et y une assignation*
2. $Sat(\ulcorner x = y \urcorner, s) \leftrightarrow s(x) = s(y)$
3. $Sat(\ulcorner x \in y \urcorner, s) \leftrightarrow s(x) \in s(y)$
4. $Sat(\ulcorner \neg\phi \urcorner, s) \leftrightarrow \neg Sat(\phi, s)$
5. $Sat(\ulcorner \phi \vee \psi \urcorner, s) \leftrightarrow Sat(\phi, s) \vee Sat(\psi, s)$
6. $Sat(\ulcorner \phi \wedge \psi \urcorner, s) \leftrightarrow Sat(\phi, s) \wedge Sat(\psi, s)$
7. $Sat(\ulcorner \forall x\phi \urcorner, s) \leftrightarrow \forall a Sat(\phi, s[x : a])$*, où* $s[x : a]$ *désigne l'assignation identique à s sauf peut-être en x, où* $s[x : a](x) = a$.

dont la construction était comprise comme la pièce essentielle du processus d'interprétation radicale dans le programme sémantique de Davidson (Voir les essais dans Davidson (1984)). La notation « $T(PA)$ » pour désigner l'extension de PA selon les modalités que je viens de décrire, est devenue plus ou moins standard dans la littérature logico-philosophique.

[86] Cette clause est importante en particulier dans le cas qui retiendra notre attention dans les chapitres suivants et où A et Syn_A sont des théories formulées en *premier ordre*. On suppose également que les schémas d'axiomes *logiques* et le domaine d'application des règles d'inférences sont étendus aux formules du nouveau vocabulaire.

[87] Sur le modèle qui a été donné précédemment et avec les schémas étendus.

Le prédicat de vérité pourrait ensuite être introduit de la façon habituelle :

Définition 10. *ϕ étant un énoncé et s une assignation :* $Vr(\phi) \leftrightarrow \forall s\, Sat(\phi, s)$

Les extensions aléthiques de type $T(A)$, qui utilisent les clauses récursives de la définition tarskienne à titre d'axiomes gouvernant un prédicat primitif, présentent de bonnes propriétés qui les distinguent des précédentes. Tout d'abord dans l'extension tarskienne d'une théorie le prédicat de vérité est bien adéquat au sens de Tarski. Qui plus est, les axiomes purement aléthiques sont en nombre *fini*, ce que n'est pas le cas avec la théorie minimale, et si A et Syn_A sont finiment axiomatisés, l'extension aléthique tarskienne de A est donc également finiment axiomatisée. Surtout, le pouvoir déductif des axiomes récursifs est bien supérieur à celui de la théorie minimale. En particulier, en présence de la syntaxe, elle permet de prouver ces généralisations dont Tarski pensait qu'il était important qu'elles soient conséquences d'une théorie de la vérité : en particulier les lois du tiers-exclu et de non-contradiction, et plus généralement ce que nous avons appelé les généralisations logiques.

L'extension aléthique tarskienne d'une théorie A en est également une théorie réflexive, au sens suivant.[88] Etant donnée une théorie de base A, on aura [89] :

$$T(A) \vdash \forall x(Th_A(x) \rightarrow Vr(x))^{90}$$

(Il y a dans $T(A)$ une démonstration de « Tous les théorèmes de A sont vrais »)

Et il sera par conséquent possible de dériver dans $T(A)$ la cohérence de la théorie-objet A, points que nous avons identifiés également comme des désiderata formulés

[88] Ici j'emprunte la terminologie à Ketland (1999). Voir chapitre suivant.

[89] Il faudra pour cela que le vocabulaire sémantique (Sat, Vr, etc.) puisse apparaître dans les instances des schémas d'axiomes de la théorie objet et de la théorie de la syntaxe, en sorte que par exemple la formalisation du principe suivant soit un théorème de la syntaxe :

si les axiomes de A sont vrais et si, si les prémisses d'une règle d'inférence de A sont vraies alors sa conclusion est vraie, alors tous les théorèmes de A sont vrais.

C'est le rôle de la clause 4. dans notre présentation de $T(A)$ que d'assurer qu'il n'y a pas d'ambiguïté sur ce point. C'est aussi une clause que Tarski regardait comme éminemment naturelle, et qui allait sans dire. Voir la note de bas de page dans Tarski (1983) p. 237. Toutefois il est utile d'être précis sur ce point parce qu'il aura une importance provisoire au chapitre 6.

[90] Je m'autorise un abus de notation habituel ici, en utilisant l'énoncé à droite du symbole de relation binaire "$\vdash$" comme un nom de lui-même. Par ailleurs, le principe affirmant "Tous les théorèmes de A sont vrais" est ce que l'on appelle parfois dans la littérature un principe de réfléxion (généralisé). La qualification "réflexive" à propos de la théorie tarskienne est justifiée par le fait que les extensions aléthiques tarskiennes permettent de dériver ce principe de réflexion sur la théorie de base.

par Tarski.[91]

Enfin, une telle théorie détermine univoquement l'extension du prédicat « Vr » au sens présenté plus haut.[92] En un sens cette axiomatisation fixe bien l'extension de *Sat* et peut être regardée comme une « définition » implicite. Cependant il ne s'agit ni d'une définition éliminative ni d'une définition non-créative. Cette théorie est donc remarquable en ceci qu'elle présente toutes les caractéristiques attendues d'une théorie de la vérité telles que formulées par Tarski.[93] C'est parce que ce type d'extension est décalqué de la définition de la vérité donnée par Tarski, et qu'elle satisfait aux conditions qu'il a lui-même formulées, qu'il est justifié[94] de parler à leur propos d'extensions aléthiques *tarskiennes* d'une théorie donnée, alors que paradoxalement Tarski n'en a pas évoqué la possiblité.

La question se pose de savoir pourquoi Tarski n'a pas envisagé cette théorie axiomatique, théorie qu'il avait, pour ainsi dire, sous les yeux. La raison est sans doute la suivante. À l'époque où Tarski travaillait dans le cadre de la théorie des types, les langages pour lesquels la définition de la vérité était démontrée impossible étaient les langages d'ordre infini, c'est-à-dire dont l'ordre (syntaxique) des variables n'était pas borné. Mais pour ces langages la théorie axiomatique de la satisfaction n'est pas davantage possible que la définition explicite de la vérité, le problème étant que la fonction d'assignation devrait avoir un ordre supérieur à tous les ordres des variables du langage, sans parler de l'ordre du prédicat de satisfaction lui-même, qui aurait dû être encore supérieur.[95] Or, d'un autre côté, lorsqu'en 1935 Tarski abandonne la théorie des catégories sémantiques et adopte les variables d'ordres indéterminées, tandis qu'une telle théorie axiomatique devenait possible, cette théorie devenait aussi, à la faveur d'un changement d'humeur philosophique, totalement inutile : puisque, dans le nouveau paradigme, il n'y avait plus de langage universel et que la définition explicite devenait donc, pour Tarski, toujours possible.

[91] Par contraste avec la théorie minimale, si la théorie tarskienne est une façon de prendre le « *produit logique* » (Tarski) des équivalences-T, l'ensemble des conséquences de l'extension aléthique tarskienne qui ne sont pas des conséquences d'une extension minimale ne se réduit nullement à des généralisations dont toutes les instances pourraient être prouvées dans l'extension aléthique minimale.

[92] Etant donné une copie des axiomes de *Sat* dans laquelle le prédicat *Sat* est remplacé par le prédicat Sat', il sera possible de prouver dans la réunion des deux et la syntaxe que $\forall x, y(Sat(x, y) \leftrightarrow Sat'(x, y))$.Voir McGee (1991), chap.3.

[93] Voir aussi Quine (1990), pour une discussion.

[94] Et admis dans la littérature philosphique.

[95] Il faudrait nuancer ici. Ph. de Rouilhan m'a suggéré qu'elle est en effet possible comme théorie récursive systématiquement ambiguë quant aux types.

Que se fût-il passé si Tarski, converti à la théorie des ensembles mais resté universaliste, avait reconnu l'impossibilité de définir la vérité pour certains langages ? Il n'aurait pu que s'arrêter à considérer sérieusement une telle théorie axiomatique. Mais il faut nous souvenir que le but de Tarski était d'abord de montrer qu'un usage légitime du concept de vérité était possible. Or, pour Tarski, ceci signifiait donner une preuve de cohérence. La définition explicite, quand elle est possible, nous assure que le prédicat de vérité défini n'introduit pas de contradiction dans la métathéorie. Dans le cas de langages pour lesquels la définition n'est pas possible, la théorie minimale de la vérité permet d'introduire dans la métathéorie un prédicat de vérité primitif qui satisfait trivialement la convention-T, *et l'on peut prouver que cette théorie n'introduit pas de contradiction.* Mais précisément à cause de sa richesse déductive, il n'est pas possible de prouver la cohérence, relativement à une théorie A, de son extension aléthique $T(A)$.[96]

2.5 Variations sur le thème de la métathéorie

Dans cette section, tout en restant fidèle à Tarski, je voudrais m'écarter un peu de la façon dont, dans le *Concept de Vérité*, il guide le lecteur vers la définition de la vérité. Le but est une double clarification. Il s'agit d'une part de voir comment reconduire la méthode de définition de Tarski en abandonnant ce qui était le point de départ de sa stratégie d'exposition, à savoir la donnée d'une *théorie*-objet, dont Tarski semble de surcroît suggèrer (par l'exemple) qu'elle se choisit de préférence dans l'édifice de la science constituée [97], en sorte qu'elle soit tout à fait généralement admise pour vraie, et en nous donnant seulement au départ un langage au sens strict,

[96] $T(A)$ prouvant la cohérence de la théorie A. À nouveau, pour que ceci soit vrai, il est essentiel de comprendre $T(A)$ comme incluant la clause 4. Pour Tarski, nous l'avons dit, l'attitude naturelle était d'étendre les schémas de façon systématique, ce qui correspond à notre façon ordinaire de comprendre les schémas. En général, ce que nous entendons faire en acceptant un schéma c'est signer un blanc-seing : en accepter toute instance possible, de tout langage que nous aurons reconnu comme sensé. Pour illustrer sur un exemple, puisque ZF contient non seulement des axiomes, mais également des schémas d'axiomes (le schéma de compréhension, le schéma de séparation), $T(ZF)$ n'est pas seulement la liste des axiomes de ZF, Syn_{ZF} et Sat_{ZF}, mais il faut encore compter parmi les axiomes de nouvelles instances du schéma de compréhension, par exemple :

$$\forall x \exists y \forall z (z \in y \leftrightarrow z \in x \wedge \text{ z est un énoncé et z est vrai})$$

qui ne figuraient pas dans la liste initiale des axiomes de ZF.

[97] Mathématique (l'exemple traité en détail par Tarski est celui du calcul des classes), physique théorique (Tarski suggère que le langage de la physique théorique de son temps peut être considéré comme un langage formalisé, que par conséquent sa méthode de définition de la vérité s'y applique).

ce que j'ai suggéré d'appeler un langage* pour éviter la confusion avec l'emploi que Tarski fait parfois du mot. On peut se demander pourquoi Tarski, dont le but annoncé est de définir la vérité dans un langage*, commence l'exposé de sa méthode en se donnant une théorie vraie. Est-il essentiel de se donner une théorie ? Si l'on s'en donne une, faut-il en outre qu'elle soit vraie, ou que nous la croyions telle, pour que le projet ait le sens que Tarski veut lui donner ? A strictement parler, la réponse à ces questions est négative, mais il apparaîtra en fait que Tarski poursuit dans le *Concept de vérité* deux programmes, un *programme large* et un *programme étroit*. Le programme étroit est celui annoncé : montrer comment définir la vérité pour un langage* lorsque c'est possible. Le programme large est d'illustrer la fécondité de ses recherches sur la vérité dans les études fondationnelles. Le second bénéfice attendu de cette mise au clair, sera de dégager clairement les rôles respectifs des différents axiomes de la méta-théorie tarskienne telle que Tarski la présente pour la preuve des différents résultats dont nous avons vu qu'ils étaient dérivables de cette méta-théorie. La méta-théorie présentée par Tarski est en effet riche en axiomes de différentes natures, et il importe pour la suite de ce travail de comprendre ce qui suit de quoi.

2.5.1 Définition de la vérité pour un langage*

Supposons donc donné un langage* interprété L^* satisfaisant aux conditions présentées en 3.1, et supposons que nous voulions définir vrai-dans-L^* en suivant les traces de Tarski. Cette définition doit être conduite dans un métalangage ML^* essentiellement plus riche que L^*. Que doit contenir le métalangage ? Les conditions que Tarski impose au métalangage et que nous avons décrites à la section 3.2.1. sont certainement toujours nécessaires à la définition adéquate de vrai-dans-L^* : le vocabulaire logique doit bien entendu y figurer, le vocabulaire de la syntaxe également (pour développer la syntaxe de L^*), et le vocabulaire du langage-objet sera nécessaire pour formuler les équivalences-T qui devront être dérivables de la définition. Voilà pour le métalangage. Reste que cette définition doit être conduite dans une *théorie*. Il était naturel de parler de *méta*-théorie, à propos de la théorie dans laquelle devait être construite la définition, lorsque nous avions d'abord une théorie qui était l'objet de nos investigations. Maintenant qu'il n'y a plus de théorie-objet, parler de méta-théorie est moins naturel. Néanmoins, je vais m'en tenir à l'usage de Tarski, pour éviter une certaine confusion. Il y a en effet un risque de

confusion du fait que rien n'interdit que dans notre « méta-théorie » soient *décrites* certaines théories, la syntaxe d'axiomes et de règles possibles ou existantes, sans toutefois que ces règles et axiomes ne soient *assertés* dans la méta-théorie. Dans la syntaxe du langage de la théorie des ensembles on pourrait, comme nous l'avons fait dans le cas du langage de l'arithmétique, décrire la théorie ZF en définissant l'ensemble l'ensemble de ses axiomes et de ses théorèmes, sans pour autant que la syntaxe n'affirme aucun de ces axiomes. Ce serait simplement une situation dans laquelle, dans la théorie que nous employons, ou dont nous faisons usage, d'autres (ou la même) théories sont mentionnées ou plutôt, ici, décrites. Mais dans une situation comme celle-là, il est assez naturel de conserver la terminologie tarskienne et de parler de théorie-objet à propos de la théorie qui est décrite (ZF, dans notre exemple) dans la théorie que nous employons (la syntaxe de ZF dans l'exemple), laquelle est naturellement conçue comme la méta-théorie.

Cette remarque étant faite, revenons à la définition de vrai-dans-L^*. Quelles sont donc les conditions à imposer à la *métathéorie* pour que cette définition soit possible ? Tarski intoduisait trois classes d'axiomes dans la méta-théorie :

1. les axiomes logiques généraux ;
2. des axiomes ayant la même signification ou sont plus forts, que ceux de la science étudiée (ce que nous avons appelé les *axiomes théoriques*) ;
3. les axiomes gouvernant la morphologie du langage-objet.

On voit bien pourquoi la méta-théorie doit contenir des axiomes logiques généraux et les axiomes gouvernant les notions morphologiques. Mais exiger la présence de la seconde classe d'axiome n'a plus de sens dans le contexte qui est le nôtre à présent, et où il n'y a tout simplement plus de « science étudiée ». Faut-il néanmoins admettre dans la métathéorie des axiomes gouvernant l'usage des expressions du métalangage qui appartiennent au langage-objet, ou en sont des traductions ? Si oui, remarquons que nous serions confronté à l'obligation de faire un choix *arbitraire*. Supposons en effet que le langage pour lequel nous voulons définir la vérité soit le langage de l'arithmétique, présenté au début de la section 3.2.2. Il y a de nombreuses théories arithmétiques, quand bien même nous restreindrions notre attention aux théories vraies, naturelles, récursivement axiomatisables, dont Q et PA sont deux exemples ; et je ne vois aucune raison de principe de préférer l'une à l'autre. Et si nous cherchons la théorie « la plus forte » avec à l'esprit l'idée de contraindre autant que possible l'usage des expressions en conformité avec leur signification, nous en serons

pour nos frais. C'est une des leçons du premier théorème d'incomplétude : s'il y a une théorie arithmétique « la plus forte », cette théorie n'est pas récursivement axiomatisable, et si l'on s'en tient aux théories récursivement axiomatisables, il n'y a rien de tel que « la théorie la plus forte ». Mais dès que l'on abandonne l'idée que certains axiomes de la méta-théorie doivent rendre compte de la signification des expressions du langage-objet[98], il me semble que tout ensemble d'axiomes relatifs aux notions du langage-objet est acceptable dans la méta-théorie, y compris, à la limite, l'ensemble vide.

Que manque-t-il à notre métathéorie pour qu'y soit définissable vrai-dans-L^* ? Il ne manque rien. Supposons par exemple que le langage pour lequel nous souhaitons définir la vérité soit le langage de l'arithmétique.[99] Dans la syntaxe, nous définissons l'ensemble des expressions de L^*, l'ensemble de ses termes, et l'ensemble de ses énoncés, exactement comme nous l'avons fait dans la section 3.2.2.1., partie « Morphologie de L_Q ». Désormais, puisqu'il n'y pas de « science étudiée », nous sommes dispensés de la description des axiomes, théorèmes, etc. de cette science. Ces définitions morphologiques étant données, et laissant de côté quelques détails, nous pouvons définir la notion de dénotation des termes de L^*, exactement comme nous l'avons vu tout à l'heure, section 3.2.2.2., et de même pour la définition de la satisfaction et de la vérité. Pour terminer, on pourrait montrer que la définition obtenue est adéquate, l'absence, dans la méta-théorie, d'axiomes relatifs au langage-objet, ne constituant nullement un obstacle à la dérivation des équivalences-T à partir de la définition.

Pour illustrer le point, je propose de manipuler quelques exemples. Supposons que, ayant peut-être en tête l'arithmétique de Peano (en premier ordre), nous voulions définir la vérité pour le langage* de l'arithmétique. Supposons également que, dans la syntaxe, non seulement nous ayons décrit la morphologie de ce langage, mais que nous ayons introduit également par définition une description des axiomes et des règles de PA (et pourquoi pas aussi d'autres théories, peu importe). Et supposons enfin que les axiomes de notre métathéorie relatifs aux notions du langage-objet (les axiomes arithmétiques en l'espèce) se réduisent à ce seul axiome :

Théorie triviale A_0 0 n'est le successeur d'aucun nombre. (*i.e.* $\forall x \neg(0 = Sx)$)

[98] Et que l'on reconnaît que c'est à la fonction de traduction du langage-objet dans le méta-langage que revient cette tâche. Voir section "Adéquation".

[99] Il faut garder à l'esprit que ce choix est arbitraire et que la même chose serait possible avec les langages de la mécanique hamiltonienne ou tout autre langage formalisable.

Cette théorie est évidemment beaucoup plus faible que PA, mais elle n'en est pas moins une théorie, et une théorie vraie de surcroît. Pour faciliter les choses, j'ai fait en sorte que le langage de cette théorie soit le *même* que celui de PA[100], mais il n'y a là rien d'essentiel.

Dans cette métathéorie, nous l'avons dit, il est possible d'écrire les clauses récursives de la satisfaction pour le langage de PA que nous avons présentées en détail tout à l'heure. Si la logique de métathéorie est la logique du second-ordre, supposons-le, nous pouvons alors transformer ces clauses en définition explicite de la satisfaction et de la vérité. Jusqu'ici, il n'a été fait aucun usage des axiomes arithmétiques de la métathéorie, pas plus que des axiomes de la théorie-objet. Pour ce qui est de la dérivation des équivalences-T, nous l'avons dit, elle ne pose pas de problème. À titre d'exemple voyons comment dériver l'équivalence-T suivante :

$$Vr(\ulcorner 0 = S0 \vee S0 = SS0\urcorner) \leftrightarrow (0 = S0 \vee S0 = SS0)^{101}$$

Il suffit d'appliquer la définition de la vérité :

D'une part, d'après la clause de la définition récursive de la vérité concernant la conjonction,

$$Vr(\ulcorner 0 = S0 \vee S0 = SS0\urcorner) \leftrightarrow Vr(\ulcorner 0 = S0\urcorner) \vee Vr(\ulcorner S0 = SS0\urcorner)$$

et d'autre part, d'après les clauses initiales[102]

$$Vr(\ulcorner 0 = S0\urcorner) \leftrightarrow 0 = S0 \text{ et}$$
$$Vr(\ulcorner S0 = SS0\urcorner) \leftrightarrow S0 = SS0$$

Donc par substitution des équivalents,

$$Vr(\ulcorner 0 = S0 \vee S0 = SS0\urcorner) \leftrightarrow 0 = S0 \vee S0 = SS0$$

Il n'y a rien de particulier avec cet exemple : la dérivation des équivalences-T à partir de la définition de la vérité dans la métathéorie ne requiert que des moyens logico-syntaxiques très modestes. Dans la même veine, on pourrait également montrer que pour dériver les lois aléthiques comme la bivalence ou le principe de non-contradiction, les axiomes de la métathéorie correspondant à la théorie-objet ne

[100] Et supposé que la fonction de traduction du langage-objet dans le méta-langage est la traduction homographique.

[101] Il faudrait écrire :
$Vr(\overline{0} * \equiv * \overline{S} * \overline{(} * \overline{0} * \overline{)} * \overline{\vee} * \overline{S} * \overline{(} * \overline{0} * \overline{)} * \equiv * \overline{S} * \overline{(} * \overline{S} * \overline{(} * \overline{0} * \overline{)} * \overline{)}) \leftrightarrow (0 = S(0) \vee S(0) = S(S(0))$

[102] Je simplifie très légèrement, il faudrait passer également par les axiomes de la dénotation ici.

sont pas plus nécessaires qu'ils ne le sont pour dériver les équivalences-T. Enfin, les *définitions* de l'ensemble des axiomes et des théorèmes de la théorie-objet dans la syntaxe, et plus généralement toute la partie de la syntaxe qui relève de l'étude de la notion de preuve dans la théorie-objet, ne jouent aucun rôle dans la définition de la vérité pour le langage-objet. En lieu et place de la théorie A_0, nous aurions tout aussi bien pu considérer la théorie vide, une théorie plus forte que PA, ou une théorie arithmétique fausse ou contredisant PA, rien de ce que nous venons de dire n'aurait à être modifié.[103]

2.5.2 Le programme large

Cette reconstruction doit éveiller une question : pourquoi au juste la métathéorie devait-elle, dans le programme de définition de Tarski, contenir « des axiomes ayant même signification » ou « plus forts », que ceux de la science étudiée ? Que l'on entende « langage » en un sens large incluant la donnée d'une grammaire *et* d'axiomes et de règles, ou au sens étroit, cela ne change rien à l'affaire : définir la vérité pour un langage ce n'est jamais en fait que définir la vérité pour un langage*, éventuellement le langage* d'un langage au sens large, et la possibilité de construire une définition adéquate de la vérité pour un *langage* L dans une « métathéorie » ne dépend *nullement* de la présence dans la « métathéorie » d'axiomes plus forts, ou ayant même signification que les axiomes de la « science étudiée ». Nous venons de le voir, et il est facile de s'en convaincre en reprenant pas à pas la construction de Tarski[104], ce que j'ai appelé les « axiomes théoriques de la métathéorie ne jouent *aucun* rôle dans la définition de la vérité de Tarski, pas plus que pour la *dérivation des équivalences-T* à partir de cette définition. Pourquoi, alors, Tarski impose-t-il ces axiomes théoriques ? C'est qu'il doit poursuivre une autre fin que la seule construction d'une définition adéquate. En fait, il souhaite également illustrer la fécondité théorique d'une définition de la vérité en se donnant un cadre naturel d'application, parmi d'autres possibles.

Quelle est la différence entre le type de méta-théorie que nous avons considéré pour construire une définition adéquate de la vérité-pour-un-langage*, et la métathéorie que se donne Tarski pour définir la vérité-dans-un-langage ? Pour simplifier la discussion, considérons à nouveau le cas particulier d'une théorie arithmétique,

[103] Voir section suivante.
[104] Ici, section 3.2

disons l'arithmétique de Robinson, Q. Le langage* pour lequel nous considérons la définition de la vérité est le langage* de Q, tandis que le *langage* pour lequel Tarski considère le projet de définition est, en somme, Q elle-même. Les deux types de métathéories diffèrent sur deux points. D'une part, dans la méta-théorie tarskienne, la syntaxe du langage-objet contient, en plus de la théorie du langage*-objet, un certain nombre de définitions caractérisant Q en tant que théorie. D'autre part, la méta-théorie tarskienne est supposée contenir les axiomes de Q elle-même, ou des axiomes ayant même signification que ceux de Q, tandis que dans la méta-théorie du premier genre, aucune hypothèse particulière n'est faite sur les axiomes éventuels relatifs aux expressions du langage-objet. Puisque les théories syntaxiques ne diffèrent qu'à des définitions près, on peut considérer la méta-théorie tarskienne comme un simple cas particulier du premier genre de métathéorie, sur laquelle on impose cette contrainte que les axiomes qui sont *décrits* dans la « la syntaxe » (dans la métathéorie), soient également *assertés* dans la métathéorie.

A quoi ces axiomes et ces définitions servent-ils donc ? En un sens la réponse est simple et l'on peut noter deux choses :

1. La *présence* de ces axiomes dans la métathéorie permet de *décider* de la vérité ou de la fausseté de certains énoncés du *langage-objet* dans la métathéorie, et
2. le fait de requérir que ces axiomes soient (consistants avec et) *au moins aussi fort* que ceux de la théorie-objet permet de prouver dans la métathéorie l'affirmation que *tous les théorèmes de la théorie-objet sont vrais* (et par voie de conséquence que la théorie-objet est cohérente).

Voyons comment les choses fonctionnent en détail sur quelques exemples très simples. Supposons que notre théorie-objet, celle qui est décrite dans la syntaxe, soit PA, et que les axiomes de PA figurent aussi au titre d'axiomes de notre métathéorie. Nous avons vu qu'à partir de la définition de la vérité et avec simplement un peu de logique et de syntaxe, il était possible de prouver :

$$Vr(\ulcorner 0 = S0 \vee S0 = SS0 \urcorner) \leftrightarrow (0 = S0 \vee S0 = SS0)$$

Or dans notre métathéorie on peut également prouver :

$$\neg(0 = S0 \vee S0 = SS0)$$

parce que notre métathéorie contient PA et que cet énoncé, on peut s'en convaincre facilement, est un théorème de PA. Il suit immédiatement que dans notre métathéorie

il est possible de prouver que cet énoncé, celui mentionné à gauche de l'équivalence-T précédente, n'est pas vrai, c'est-à-dire :

$$\neg Vr(\ulcorner 0 = S0 \vee S0 = SS0 \urcorner)$$

Si, changeant d'hypothèse, nous supposons à présent que la partie arithmétique de notre métathéorie n'est composée que de la théorie triviale A_0 présentée plus haut, alors l'énoncé $\ulcorner \neg(0 = S0 \vee S0 = SS0) \urcorner$ du langage-objet, bien qu'un théorème de la *théorie-objet*, *n*'est *pas* déclaré vrai par la métathéorie, pas plus d'ailleurs que sa négation. Poussons nos hypothèses encore un peu plus loin et nous verrons que si l'arithmétique de la métathéorie *contredisait* l'arithmétique de la théorie-objet, alors on pourrait prouver dans la métathéorie que *la théorie-objet est fausse*, c'est-à-dire que tous ses théorèmes ne sont pas vrais, et cela, bien entendu, que la théorie-objet soit *en fait* vraie (et que l'arithmétique de la métathéorie soit donc fausse) ou qu'elle soit fausse.

Nous avons indiqué ce que permettaient de prouver les axiomes de la métathéorie qui « correspondent » à une théorie-objet donnée. Endosser ces axiomes dans la métathéorie, comme Tarski nous enjoint de le faire dans son exposition, est sans utilité pour le projet étroit de la définition adéquate de la vérité pour un langage, pas plus que n'est utile, du reste, leur description (définition) dans la syntaxe. La métathéorie ainsi étendue, néanmoins, permet à Tarski de montrer l'utilité des concepts sémantiques pour les recherches fondationnelles.

Nous sommes au début des années trente, Hilbert avait demandé des preuves « réelles » de cohérence pour toutes les mathématiques « idéales », et Gödel venait de montrer qu'une telle demande ne pouvait pas être satisfaite, l'arithmétique primitive récursive ne pouvant pas prouver sa propre cohérence. Or la construction d'une métathéorie à la Tarski fournit une méthode et un cadre pour prouver la cohérence de la théorie-objet. Pour prouver dans la métathéorie toutes les "généralisations théoriques" (les affirmations 3, 4, 5 décrites au début de la section 3.3), il *faut* que les axiomes théoriques correspondants soient assertés dans la métathéorie. Bien entendu, la portée épistémologique de la preuve de cohérence ainsi obtenue est limitée.[105] Comme le notait Tarski (1944) :

> Ainsi la théorie de la vérité fournit une méthode générale pour les preuves de cohérence pour les disciplines mathématiques formalisées.

[105] Et il ne s'agit de toute façon en aucun cas de ressusciter un programme fondationnel du type du programme de Hilbert.

> Il est facile de voir, néanmoins, qu'une preuve de cohérence obtenue de cette façon ne peut avoir de valeur intuitive - i.e. nous convaincre, ou renforcer notre croyance, que la discipline considérée est en fait cohérente - que dans le cas où nous réussissons à définir la vérité en termes d'un métalangage qui ne contient pas le langage-objet comme partie [...]. Car dans ce cas seulement les hypothèses déductives du méta-langage peuvent être intuitivement plus simples et plus évidentes que celles du langage-objet - même si la condition de « richesse essentielle » est formellement satisfaite.(Tarski (1944), p.354 n.18)

Mais si la métathéorie tarskienne fournit un cadre naturel pour prouver la cohérence d'une théorie-objet donnée, elle permet surtout à Tarski de mener à bien une clarification philosophique majeure, celle de la distinction entre les notions de prouvabilité formelle et de vérité d'un énoncé. En effet, Tarski est capable de montrer dans la métathéorie que l'ensemble des énoncés vrais du langage-objet est un ensemble complet d'énoncés[106] et, d'autre part, en faisant *usage* des axiomes et des règles correspondant à ceux de la théorie-objet, que tous les axiomes de la théorie-objet sont vrais et que ses règles préservent la vérité. Or, à l'aide des résultats de Gödel, il est également possible de montrer que si la théorie-objet est suffisamment riche, il existe des énoncés de son langage qu'elle ne peut prouver ni réfuter. Il suit qu'il existe des énoncés vrais non prouvables.[107] Cette clarification constituait en elle-même un authentique progrès philosophique à un moment où l'on pouvait douter que la notion de vérité puisse être distinguée de la notion de prouvabilité, particulièrement en mathématiques.

Appendice Je résume dans le tableau suivant les points techniques principaux évoqués dans cette section. Soit A une théorie d'un langage L_A.
Nous noterons

1. $MT(A)$ la méta-théorie tarskienne comportant :

[106] Au sens où pour tout énoncé ϕ de ce langage, ϕ ou sa négation est vrai.

[107] On pourra remarquer que pour établir ce résultat, Tarski aurait pu là encore se passer d'adopter dans la métathéorie les axiomes correspondant à la théorie-objet. La métathéorie réduite suffit à prouver que l'ensemble des énoncés vrais est complet, comme nous l'avons vu, or l'ensemble des énoncés formellement démontrable dans une théorie T n'est pas complet (si la théorie est cohérente et récursivement énumérable, du moins), donc il y a des énoncés vrais non prouvables dans T. Le passage de Tarski par la preuve, dans la métathéorie, que tous les théorèmes de la théorie-objet sont vrais, est un détour non nécessaire ici (alors qu'il est nécessaire pour prouver la cohérence de la théorie-objet).

   a) des axiomes logiques essentiellement plus riches que ceux de A,
   b) la syntaxe de A et
   c) les axiomes de A elle-même (ou une traduction, ou des axiomes plus forts que ceux de A).
2. MT, la théorie précédente privée des axiomes de A, et comportant donc uniquement :
   a) Les axiomes logiques essentiellement plus riches que ceux de A et
   b) la syntaxe de A.
3. $MT(\neg A)$ une théorie comportant :
   a) des axiomes logiques essentiellement plus riches que ceux de A
   b) la syntaxe de A et
   c) des axiomes qui contredisent ceux de la théorie A.[108]

La situation est alors la suivante :

	MT	$MT(A)$	$MT(\neg A)$
Définissabilité de la vérité-dans-L_A	oui	oui	oui
Généralisations Logiques :			
Prouve la loi de la bivalence	oui	oui	oui
Prouve la loi de non-contradiction	oui	oui	oui
Généralisations théoriques :			
Prouve « Tous les théorèmes de A sont vrais »	non	oui	non
Prouve « A est cohérente »	non	oui	non
Prouve « Un théorème de A n'est pas vrai »	non	non	oui
Prouve « A n'est pas cohérente »	non	non	non

TAB. 2.1: Variations sur le thème de la métathéorie : tableau récapitulatif

Je complète maintenant le tableau précédent par une comparaison avec les théories *axiomatiques* de la vérité vues dans la section 3.4 . Dans les théories que nous considérons maintenant le prédicat de vérité est donc donné comme primitif, gouverné par un certain nombre d'axiomes, et l'appareil logique disponible *n'* est *pas* essentiellement plus riche que celui de la théorie-objet considérée. À nouveau A sera une théorie d'un langage L_A.
Je note

[108] Pour fixer les idées, on peut prendre pour A la théorie PA, et supposer que $MT(\neg A)$ est simplement $MT + \exists x\ 0 = S(x)$.

1. $T(A)$ l'extension aléthique tarskienne de A composée [109] :
   a) des axiomes de A
   b) de la syntaxe de A
   c) des axiomes récursifs de la vérité.

2. $T^{-}(A)$ la théorie $T(A)$ avec cette restriction qu'il n'est pas permis aux schémas de A et de la syntaxe de A d'être instanciés par des formules du langage étendu de $T(A)$ (contenant le prédicat de vérité).

3. $M(A)$, l'extension aléthique minimale de A, composée
   a) des axiomes de A,
   b) de la syntaxe de A, et
   c) des équivalences-T (correspondant au langage L_A).

On a alors :

	$T(A)$	$T^{-}(A)$	$M(A)$
Définissabilité de la vérité-dans-L_A	Non	Non	Non
Adéquation du prédicat de vérité	Oui	Oui	Oui
Généralisations Logiques :			
Prouve la loi de la bivalence	oui	oui	non
Prouve la loi de non-contradiction	oui	oui	non
Généralisations théoriques :			
Prouve « Tous les théorèmes de A sont vrais »	oui	non	non
Prouve « A est cohérente »	oui	non	non
Instances :			
Prouve toutes les instances des généralisations logiques	oui	oui	oui
Prouve toutes les instances des généralisations théoriques	oui	oui	non

Tab. 2.2: Variations sur le thème des extensions aléthiques : tableau récapitulatif

2.6 Conclusion : Tarski était-il déflationniste ?

La définition de la vérité par Tarski accomplit-elle ce que les philosophes attendent d'une explication de la notion de vérité ?

[109] Voir section 3.4.

Dans un article influent, Hartry Field[110] remarque que la définition de Tarski ne constitue pas à proprement parler une réduction de la notion de vérité à des notions non-sémantiques au sens attendu par un physicaliste. Si la définition de Tarski permet de réduire la vérité des énoncés en général à la vérité des énoncés atomiques [111], le caractère irréductiblement énumératif et disjonctif des clauses de base de la définition, celles relatives à la dénotation des termes primitifs et à la vérité des énoncés atomiques, fait de cette explication une explication incomplète. Ce que Field demande à une explication complète, c'est qu'elle présente une « réduction » des propriétés sémantiques à des propriétés physiques. Et selon lui, la réduction opérée par la définition de Tarski n'est pas entièrement satisfaisante d'un point de vue méthodologique tant que la propriété de dénotation (au sens large) n'a pas elle-même été réduite. Or la réduction d'une propriété implique davantage, selon Field, que l'énumération de son extension, alors même que celle-ci peut en constituer une définition explicite extensionnelle. Pour expliquer véritablement la notion de dénotation en termes non sémantiques, selon Field, il faudrait que l'explication fournie soit générale, c'est-à-dire qu'elle spécifie non seulement l'extension de l'expression « dénote » dans le cas de langages particuliers, mais qu'elle nous dise en quoi cela consiste en général pour une expression x de dénoter un objet y, avec x et y variables.[112] La définition de Tarski ne donne aucune indication sur la possiblité de réduire les propriétés sémantiques au sens où Field l'entend.

La critique de Field fait écho à celle que Max Black adressait dès 1949 à la définition de Tarski en tant qu'explication philosophique de la notion de vérité.[113] L'objection de Max Black, en un mot, est la suivante : la définition de Tarski ne donne pas ce que l'on attend d'une explication de la notion classique de vérité parce qu'il ne définit pas vrai-dans-L pour L *variable*. Ce n'est pas que le champ d'application de la méthode de définition de Tarski serait trop étroit : Tarski montre bien comment pour tout langage-objet L (d'un certain type, mais ce n'est pas le problème), on peut construire un métalangage ML dans lequel on peut définir adéquatement la vérité pour L (relativement à une traduction de L dans ML). Ce

[110] Field (1972).

[111] Et donc à la dénotation, au sens large où les termes dénotent des objets et les prédicats des propriétés.

[112] Pour les enjeux philosophiques de la réduction des notions sémantiques et le changement de position de Field relativemen aux attentes d'une définition de la vérité, voir chapitre 2, en particulier section 5.

[113] Black (1949) p.104.

que Tarski ne donne pas, en revanche, c'est un métalangage et une définition dans ce métalangage de la vérité pour L, avec L variable.

Il semble que, dans un cas comme dans l'autre, fondamentalement, l'attente philosophique déçue est celle d'une théorie générale de la notion de correspondance associée à la notion classique de vérité. Ce qui trouble dans la définition de Tarski, au bout du compte, ce n'est pas tant qu'elle soit possible (quand elle l'est), que ce dont elle se passe ; que l'on puisse définir la vérité pour un langage sans rien dire de la *nature* du lien entre les expressions d'un langage et ce à quoi ces expressions renvoient, sans expliquer ce que c'est pour une expression d'un langage interprété d'avoir la signification qu'elle a. Dans la définition tarskienne, la signification des énoncés du langage-objet est simplement stipulée par une fonction de traduction de ces énoncés dans le métalangage, traduction relativement à laquelle seulement la question de l'adéquation peut être posée. La question de savoir ce qu'est une traduction correcte d'un langage déjà interprété dans un autre, non pas relativement à une tierce traduction qui serait stipulée par ailleurs, mais en général, cette question n'est jamais abordée.[114]

Il y a certainement des choses que la théorie de la vérité de Tarski n'explique pas, avait un jour ironisé Tarski.[115] Parmi ces choses, pourtant, il en est que de nombreux philosophes classent volontiers au titre des buts d'une élucidation philosophique de la notion de vérité, comprise comme partie intégrante d'une tâche plus large, celle de comprendre en quoi cela consiste, pour un langage, de « parler » du monde, pour un énoncé, une croyance, d'être *à propos* de ce à propos de quoi il est, pour comprendre sa signification. N'était-ce pas cela la promesse de la Sémantique ? Si la « sobriété » de sa définition ne convainc pas Tarski d'inadéquation, cela pourrait en revanche permettre d'inscrire légitimement Tarski dans la généalogie du déflationnisme. Nous retrouvons un thème central du déflationnisme : saisir l'intuition de la correspondance ou, mieux, la signification de « vrai » dans son acception classique, sans théorie d'une quelconque relation de correspondance. Qu'une explication qui n'accomplirait que cela soit une explication *suffisante* de la vérité, c'est

[114] Voir nos remarques en 2.1.3.1. Par ailleurs, il n'y a aucune raison de penser qu'une théorie de la traduction devrait au bout du compte permettre de réduire les notions sémantiques à des notions physiques, au sens où Field (1972) parle de réduction. Une théorie de l'*interprétation* pourrait faire le travail, qui n'expliquerait pas la correspondance entre le langage et le monde en termes purement physiques, mais les conditions sous lesquelles un énoncé d'un langage peut être compris comme une traduction correcte (éventuellement parmi plusieurs possibles) d'un énoncé d'un autre langage.

[115] Tarski (1944), p. 345.

ce qui motive la convention-T. Qu'une telle explication puisse être donnée, c'est ce que montre la définition de la vérité-pour-L dans une théorie de la syntaxe de L, sans aucune notion sémantique primitive.[116]

[116]Nous n'avons pas abordé ici la question de savoir si, et comment, la théorie tarskienne de la vérité pouvait être mise à contribution dans une entreprise d'explication de ce en quoi cela consiste d'interpréter le discours d'un locuteur, au sens où Davidson l'envisageait. Traiter ce sujet nous amènerait trop loin. Pour une clarification des relations entre le travail de Tarski et celui de Davidson je renvoie le lecteur aux travaux de Davidson lui-même. Voir Davidson (1984). Les études de Rivenc (1998*b*), Heck (1997), et Lewis (1975) sont également très utiles. J'ajoute que je pense pour ma part qu'il n'y pas de contradiction entre la conception déflationniste de la vérité et le rôle qu'assigne Davidson à la vérité dans le processus d'interprétation. Voir Davidson (1996) pour une discussion en sympathie avec les positions déflationnistes.

Conclusion de la première partie

Au chapitre 2, nous avons présenté notre interprétation du déflationnisme. Tel que nous le comprenons, le déflationnisme émerge à l'intersection d'un naturalisme philosophique et d'un certain nombre d'intuitions et de remarques à propos de la notion de vérité. Par son refus des entités propositionnelles, le naturalisme donne une place doublement importante au prédicat de vérité : il devient d'une part indispensable à l'expression de certaines généralisations et d'autre part il devient un outil privilégié dans le discours sur les phénomènes sémantiques et intentionnels. Mais ce dernier rôle pose problème au naturaliste : est-il raisonnable de penser que la vérité d'un énoncé qui serait cru, affirmé etc. est vraiment ce qui peut expliquer *causalement* le comportement d'un agent, en vertu d'une certaine « relation de correspondance », qu'il faudrait alors expliquer par le menu ?[117] Il y a là une tension dans le naturalisme. Nous interprétons le déflationnisme comme une tentative de résoudre cette tension en donnant un sens à l'idée que le prédicat de vérité n'est pas une notion « explicative » mais « expressive », que nous voulons contribuer à éclaircir. Cette idée à son tour est soutenue par l'intuition de la transparence des attributions de vérité, qui voit une profonde communauté de sens entre ce qui est exprimé par un énoncé et l'attribution de la vérité à cet énoncé mis entre guillemets. Dans le domaine des explications des phénomènes intentionnels et sémantiques, elle est également soutenu par le fait, d'une part, que les usages que nous faisons de la vérité pour rendre compte de la signification d'énoncés d'autres locuteurs peuvent être compris comme essentiellement interprétatifs, et non explicatifs, essentiellement un outil linguistique pour construire contextuellement des traductions entre des langages et mon langage, et d'autre part que l'explication des relations causales

[117] Les arguments de Quine sur l'inscrutabilité de la référence valent *a fortiori* pour la vérité

bien réelles qui existent entre l'emploi du langage par un locuteur, ou ses processus mentaux de haut-niveau, et son environnement, ne semblent pas devoir exiger le recours à la notion de vérité comme elle ne permet pas en retour d'en rendre compte et de la réduire.

Au chapitre 3, en introduisant le travail de Tarski, nous avons présenté les ressources théoriques nécessaires à définition de la notion de vérité. Il y a une distinction qui n'est pas tout à fait anodine et qu'il faut noter à présent plus clairement entre ce que Tarski juge être des ressources méthodologiquement acceptables pour mener à bien une définition de la vérité, et les contraintes méthodologiques issues du fil de réflexion précédent. Si les propositions ne sont pas acceptables du point de vue naturaliste, les attributs ne le sont pas non plus. Et par conséquent, si Quine a raison et que la quantification est essentiellement objectuelle, parce que c'est là que se tient nécessairement, en dernière instance, le compte de « ce qui est », la logique du second ordre et les logiques d'ordres supérieurs ne sont pas méthodologiquement bien fondées.[118] Si maintenant les seuls langages bien fondés sont les langages du premier ordre, il ne peut être question de tenir pour une explication de la vérité pour un tel langage une définition construite dans un langage du deuxième ordre. Mais Tarski a montré que la définition de la vérité dans un langage n'était possible que dans un métalangage « essentiellement plus riche », ce qui implique, si l'on s'en tient maintenant aux langages du premier ordre, que le *parcours* tout entier des variables du langage-objet doit être lui-même la *valeur* possible d'une variable du métalangage.[119] Par conséquent, sous les hypothèses mentionnées, il n'est pas possible de définir la vérité pour un langage dans lequel « pour tout » signifie vraiment *pour tout*.[120]

[118] La seule façon de comprendre la logique du second ordre, pour Quine, est de l'interpréter comme de la théorie des ensembles « déguisée » en logique à la faveur d'une confusion entre deux choses radicalement distinctes que sont la prédication et la relation d'appartenance. Pour une défense de la logique du second ordre contre Quine, on pourra consulter Boolos (1984) et Rouilhan (2002).

[119] Je renvoie le lecteur à la présentation de la définition de la vérité dans le chapitre sur Tarski.

[120] Pour des raisons philosophiques on peut penser qu'il *faut* qu'un tel langage existe en droit, puisque qu'au fond c'est le seul dans lequel l'on saurait de quoi l'on parle. Inversement si un tel langage n'existe pas, alors le discours ontologique général, le projet de départ de toute métaphysique telle que traditionnellement conçue, n'est pas possible, puisqu'alors on ne pourrait jamais parler simplement *en toute généralité* de ce qui est. Donc toute métaphysique imagine qu'elle parle un tel langage. L'autre branche de l'alternative est, bien sûr qu'un tel langage n'existe pas. Posée sous l'angle de savoir si le *langage ordinaire* peut ou non exprimer parfois la généralité absolue, la question peut aussi s'enrichir de considérations linguistiques et non seulement philosophiques, et c'est aussi dans cette perspective que la question est discutée dans l'intéressant volume édité

Si l'on pense qu'il existe un langage exprimant la généralité absolue, et *a fortiori* si l'on pense que le langage ordinaire, ou encore le squelette logique du langage enrégimenté de la science, sont de ce genre, on est en droit de penser que les contextes dans lesquels la vérité est définissable perdent de leur importance théorique, et qu'inversement l'importance des théories axiomatiques de la vérité se trouve confortée. En outre les théories axiomatiques, si elles ne permettent pas de *réduire* la notion de vérité, peuvent nous permettre d'essayer de comprendre le rôle que joue cette notion primitive, irréductible, de vérité, en articulant un certain nombre de principes *a priori* qui en gouverne la compréhension et l'emploi. Dans la perspective qui est la nôtre, ces théories viendront donc désormais au premier plan, avec en vue la clarification du « pouvoir explicatif » de la vérité, l'analyse des emplois qui en sont rendus possibles par ces principes *a priori* qui font partie de notre compréhension de la notion de vérité : ce que l'on peut apprendre, déduire, expliquer grâce à eux.

Pour conclure en reprenant le fil directeur de ce travail, ces deux chapitres soulèvent ensemble un certain nombre de questions auxquelles le déflationniste est confronté. Quelle explication de la vérité (quelle théorie de la vérité) le déflationniste doit-il accepter ? Permettent-elles toutes également de rendre compte des usages de la vérité qu'il juge essentiel ? Ces théories sont-elles compatibles avec l'idée déflationniste que le prédicat de vérité n'a pas de pouvoir explicatif ? Il faut à présent préciser ces questions et y répondre. C'est ce que je ferai dans les deux parties suivantes.

par Rayo et Uzquiano (2006). Pour des considérations apparentées, penchants cette fois du côté non-absolutiste, on pourra aussi consulter Hintikka (1996). Voir aussi la mise en perspective philosophique au début de Rivenc (1993).

Deuxième partie

Conservativité et vérité

Introduction

Au chapitre 2, nous avons vu qu'il était possible dans une extension aléthique d'une théorie donnée, en mobilisant certaines lois *a priori* de la vérité, de prouver la cohérence de la théorie. Cette preuve de cohérence prend un relief épistémologique particulier à la lumière des phénomènes d'incomplétude. On sait en effet depuis le second théorème d'incomplétude de Gödel [121] qu'il n'est pas possible de dériver dans une théorie formalisée, si elle est cohérente et capable de décrire sa propre syntaxe, l'énoncé de cette théorie qui en exprime la cohérence.[122] Le fait qu'en ajoutant à une telle théorie des principes aléthiques fondamentaux nous soyons en mesure de dériver l'énoncé en question nous dit-il quelque chose sur la notion de vérité ? Et si oui, que nous dit-il ?

Dans cette partie nous donnerons un sens précis à ce problème et ferons un premier effort pour y répondre. Il existe deux types de leçons philosophiques qui ont été tirées des résultats d'incomplétude, et les deux chapitres peuvent être être vus comme façons correspondantes d'aborder le problème qui nous occupe. D'un côté les résultats d'incomplétude ont été vus d'abord comme une réfutation du programme de Hilbert visant à justifier l'usage des mathématiques idéales en en donnant une preuve de cohérence dans les mathématiques finitaires[123], et plus généralement comme des résultats antiréductionnistes forts.[124] D'un autre côté, les phénomènes d'incomplétude ont aussi marqué par ce qu'ils semblent nous dire de

[121] textciteGodel :1931

[122] Lorsque l'on parle de dérivation ici, il peut s'agir éventuellement d'une dérivation en une ligne. Autrement dit, la théorie ne peut pas même de façon cohérente *postuler* sa propre cohérence.

[123] C'est du moins ce qu'a retenu l'histoire populaire. Dans le détail, la portée des théorèmes d'incomplétude pour le programme de Hilbert est moins claire, en partie parce que n'est pas complètement clair ce qu'il faut entendre par l'ensembles des méthodes finitaires. Pour une critique approfondie on pourra consulter les travaux de Michaël Detlefsen (en particulier Detlefsen (1979), Detlefsen (1986) et Detlefsen (1990)).

[124] Voir par exemple Shapiro (1983) ou Burgess et Rosen (1997).

l'inadéquation des moyens de preuves formalisés dans une théorie à la représentation de nos capacités d'inférence ou de déduction. Ainsi, à la lecture du second théorème d'incomplétude, on peut soit être frappé de ce que l'énoncé de la cohérence d'une théorie ne puisse être dérivé que dans une théorie logiquement plus forte, ou être frappé au contraire de l'évidence avec laquelle la vérité de cet énoncé indémontrable s'impose à nous, ou pour le dire autrement ce que des théories de force logique différentes puissent néanmoins sembler se tenir sur un pied d'égalité épistémique.

Dans le premier chapitre de cette partie, je présente en détail, puis je désamorce, un argument contre les thèses déflationnistes fondé sur les phénomènes de non-conservativité auxquels donnent lieu la mobilisation de certaines lois *a priori* de la vérité dans certains contexte. Ma conclusion est que la non-conservativité des extensions aléthiques d'une théorie sur cette théorie ne suffit pas à trancher contre les thèses déflationnistes. La signification épistémologique de ces phénomènes, à regarder de près les mécanismes de leur apparition dans le cas qui nous occupe, est en fait ambiguë : je suggèrerai que l'on peut y voir la manifestation du pouvoir explicatif des principes aléthiques mobilisés, et le caractère « substantiel » de la notion de vérité, ou bien au contraire une manifestation seulement du « pouvoir expressif » de la notion de vérité, les principes aléthiques étant compris comme des outils permettant d'*expliciter* ce dont l'acceptation était *implicite*. Dans le chapitre suivant je montre que cette idée est en harmonie avec le second type de leçons des théorèmes d'incomplétude mentionnées au paragraphe précédent. Je m'appuie sur cette littérature pour défendre cette intuition et commencer à lui donner un sens précis.

Chapitre 3

Vérité et conservativité

Au chapitre précédent, sur les traces de Tarski, nous avons vu comment définir la vérité pour un langage lorsqu'une telle définition est possible. Lorsque cette définition n'est pas possible, nous avons vu que deux familles de principes se présentaient naturellement au titre d'axiomatisation de la notion de vérité. Il y avait d'une part la *théorie minimale* de la vérité, et d'autre part la *théorie tarskienne*, ou récursive, de la vérité, toutes deux répondant au critère d'adéquation formulé par Tarski. Par ailleurs, au chapitre 2, j'ai souligné que le projet déflationniste était porté par la double intuition de la *transparence* de la notion de vérité et de son *indispensabilité*, et qu'il devait être compris comme une tentative pour soutenir la thèse que la notion de vérité n'est pas une *notion explicative*, ou une notion qui ne joue pas de rôle « substantiel » dans les explications, mais est un outil « expressif », dont le rôle est simplement de permettre d'exprimer certaines généralisations qui ne peuvent être exprimées sans lui. Le problème que nous avons soulevé à la fin du chapitre 2 était alors le suivant : une théorie d'un prédicat adéquat de vérité qui permet de rendre compte des usages de la notion qu'un déflationniste juge essentiels[1] est-elle seulement compatible avec la thèse déflationniste que la notion de vérité n'est pas une notion explicative ? Il est clair que toute tentative de réponse à cette question devra d'abord formuler deux propositions visant à en préciser les termes. Il faudra d'abord proposer une explication, au sens de Carnap, de la distinction entre « notion explicative » et « notion purement expressive » et, d'autre part, préciser ce que sont ces « usages » de la notion de vérité dont une théorie de la vérité doit rendre compte pour être compatible avec le rôle qui lui est assigné par le déflationniste. Toutes les

[1] En particulier, donc, son usage comme moyen d'expression de certaines généralisations.

parties s'accordent sur le fait qu'une théorie de la vérité doit être adéquate au sens de Tarski, et donc permettre d'asserter toutes les équivalences-T. La question est de savoir si une théorie de la vérité doit ou non rendre compte d'*autres* usages, et si oui lesquels.[2] Je propose la terminologie suivante afin de faciliter la discussion ultérieure :

Problème de la stabilité : La thèse de l'absence de rôle explicatif du prédicat de vérité est-elle compatible avec la thèse concernant son rôle expressif ?

C'est le problème principal, celui de la cohérence interne du déflationnisme.

Le Problème de la frontière : Proposer une explication de la distinction entre « notion explicative » et « notion purement expressive ».

C'est la première condition d'une réponse au Problème de la stabilité. Une bonne réponse au Problème de la frontière doit respecter les intuitions déflationnistes fondamentales si elle doit être utilisée pour en étudier la cohérence interne.

Le Problème de l'usage : De quels usages de la notion de vérité une théorie de la vérité doit-elle rendre compte ?

On peut voir ce problème comme celui de la recherche d'un nouveau critère d'adéquation, que l'on pourrait appeler un critère d'adéquation* pour le distinguer du critère de Tarski. Le critère d'adéquation de Tarski assure que le prédicat de vérité théorisé saisit la notion de vérité comme correspondance, en assurant l'assertabilité des équivalences-T. L'adéquation* doit garantir un certain nombre d'autres usages de la notion de vérité jugés essentiels. Le Problème de l'usage est de dire ce qu'ils sont.

Dans ce chapitre, je propose de reconstruire une réponse globale à ces questions présentée par Stewart Shapiro et Jeffrey Ketland sous la forme d'un argument opposé au deflationnisme.[3] Au cœur de cet argument il y a, d'une part, les phénomènes gödéliens d'incomplétude, en particulier le fait que, pour le dire vite, pour toute théorie formalisable suffisamment riche, il existe un énoncé du langage de cette théorie qui « exprime » la cohérence de cette théorie et qui n'est pas prouvable dans cette théorie. Et d'autre part, les preuves de cohérence d'une théorie que l'on peut conduire dans certaines extensions aléthiques de ces théories et que nous avons

[2] En particulier, il s'agit de savoir de quoi exactement doit rendre compte une théorie qui rend compte de l'usage du prédicat de vérité dans « l'expression de généralisations ».

[3] Cet argument, connu sous le nom d'« argument de la conservativité » dans la littérature a été développé indépendamment par les deux auteurs à peu près à la même époque dans Shapiro (1998*b*) et Ketland (1999), dont c'est le thème principal.

évoquées au chapitre précédent.[4] Le raisonnement matriciel qui inspire à Shapiro et Ketland leur tentative de réfutation des thèses déflationnistes est bien exposé dans ce passage de Shapiro (1998*b*) :

> Retournons à notre théorie arithmétique A et à son énoncé de Gödel G (ou Coh).[5] Supposons qu'un professeur de logique asserte que G est vrai, et qu'un étudiant désemparé demande une explication. L'étudiant croit l'affirmation du professeur selon laquelle G est vrai, mais il veut qu'on lui montre pourquoi cet énoncé est vrai. L'étudiant veut quelque chose comme une preuve convaincante ou une preuve explicative. La réponse naturelle est de faire remarquer que tous les axiomes de A sont vrais et que les règles d'inférence préservent la vérité. Il suit que « 0=1 » n'est pas un théorème et donc que A est cohérente. L'énoncé de Gödel est équivalent à la cohérence de A. Il me semble que cette version informelle de la dérivation de Coh et G est une bonne *explication* s'il en est. L'argument montre pourquoi G est vrai. Faire remarquer que Coh et G sont (après tout) des conséquences logiques ou sémantiques de la théorie originale A ne sera d'aucune aide au déflationniste parce que c'est le fait qu'il s'agissait d'expliquer. Notre étudiant veut savoir pourquoi G est vrai - ou pourquoi G est une conséquence - et le passage par la notion de vérité fournit cette explication.(Shapiro (1998*b*), p.505).

Le premier théorème d'incomplétude implique que, *si* la théorie arithmétique A est cohérente[6], elle ne peut prouver un certain énoncé G[7] équivalent à un énoncé « exprimant » la cohérence de A[8]. Mais alors, sous l'hypothèse du théorème, A *est* cohérente, et aucune explication n'est nécessaire, sinon une explication du fait

[4] Voir chapitre précédent, 3.5.

[5] N.d.T. : « Coh » est un énoncé du langage de A qui exprime la cohérence de A. Nous reviendrons sur ces points dans la suite.

[6] Et satisfait les autres hypothèses du théorème, bien entendu : A doit être une théorie récursivement axiomatisée suffisamment riche pour que l'arithmétique de Robinson (et donc sa syntaxe élémentaire) y soit interprétable.

[7] G est un point fixe de la formule « $\neg Pr(x)$ » affirmant sa propre non prouvabilité, autrement dit une formule telle que

$$A \vdash G \leftrightarrow \neg Pr(\ulcorner G \urcorner).$$

[8] L'énoncé de la cohérence en question, noté Coh_A par Shapiro est l'énoncé affirmant qu'il n'existe pas de preuve de l'énoncé « 0=1 » dans A, i.e.

$$\forall x \neg Pr_A(x, \ulcorner 0 = 1 \urcorner).$$

que l'énoncé Coh_A exprime bien la cohérence de A et que G est équivalent à la cohérence en question. Mais ce n'est pas de cela que Shapiro nous parle. Ce que l'étudiant demande, ce n'est pas si, à supposer que A est cohérente, G est vraie ; ce qu'il demande, c'est si G est vraie (ou si A est cohérente), *tout court*. Le théorème d'incomplétude affirme que si la théorie arithmétique A est cohérente, elle ne peut prouver sa cohérence[9], mais ne dit pas si A est cohérente. *Mais les axiomes de A sont vrais*, ce que l'étudiant accepte parce qu'il accepte les axiomes de A. *Donc tous ses théorèmes sont vrais. Donc A est cohérente.*

Ce bref scénario contient le noyau des réponses de Shapiro et Ketland aux Problèmes de la frontière et de l'usage. Parce que le raisonnement précédent est naturel, une théorie de la vérité adéquate* doit permettre d'en rendre compte. Or l'énoncé de la cohérence de A est un énoncé du langage de A qui n'était pas dérivable de A seulement ; on a donc fait un usage explicatif de la notion de vérité. Dans la suite, j'aurai plus d'une fois l'occasion de me référer au raisonnement précédent, que j'appellerai la *preuve de la cohérence par la vérité*.

Le plan du chapitre sera le suivant. Dans une première partie, je propose une reconstruction rationnelle de l'argument général développé par Shapiro et Ketland. Cette exposition détaillée est utile non seulement pour la lisibilité du chapitre, mais aussi parce qu'une clarification de l'argument me paraît requise et constitue un travail philosophique en soi. Dans la seconde partie j'attaque l'argument seulement sur ce que j'appellerai la *thèse de la conservativité*.[10] Je soutiens que les phénomènes de non-conservativité observés sur les théories de la vérité[11] sont compatibles avec les thèses déflationnistes. Pour cela je m'appuie sur un examen détaillé des conditions qui donnent lieu à des phénomènes de non-conservativité des théories de la vérité.

Cet énoncé est un énoncé du langage de A si l'on suppose que A contient sa propre syntaxe. Le prédicat binaire « $Pr_A(x, y)$ » signifie que (la suite finie de signes qui est la valeur de la variable) x est une démonstration dans A (de la suite finie de signe constituant l'énoncé qui est la valeur de la variable) y. Etant donné une démonstration formelle x dans A et un énoncé y, il est facile de vérifier si oui ou non x est une démonstration de y, et $Pr_A(x, y)$ est en fait une relation décidable.

[9] A strictement parler, cette interprétation demanderait à être nuancée, et il faudrait parler d'*un certain type* d'énoncés exprimant la cohérence de la théorie. En particulier, c'est un fait bien connu des logiciens qu'il existe des énoncés dont on peut dire qu'ils « expriment la cohérence » d'une théorie A, et qui sont prouvables dans A. Il existe un débat relatif à la signification épistémologique de ces preuves. Je laisserai ce point de côté ici. Voir par exemple Resnik (1989) pour une discussion.

[10] Mais je reviendrai sur d'autres aspects dans des chapitres ultérieurs, en particulier au chapitre 7

[11] Plus précisément : la non-conservativité des extensions aléthiques de certaines théories sur ces théories.

3.1 L'argument de la Conservativité

Dans la mesure où les arguments défendus par Shapiro et Ketland sont essentiellement les mêmes, je me référerai indifféremment à l'un et à l'autre.[12] La forme générale de l'argument est la suivante :

Argument Principal :

1. Une théorie adéquate* de la vérité doit être réflexive.
2. Une théorie déflationniste de la vérité doit être conservative.
3. Une théorie réflexive de la vérité n'est pas conservative.
4. Donc les théories déflationnistes de la vérité ne sont pas adéquates*.

Les deux premières prémisses de l'argument sont des thèses philosophiques. J'appellerai la première la Thèse de la Réflexion et la seconde la Thèse de la Conservativité. La troisième prémisse est un fait de logique. Je prends les trois points dans cet ordre.

3.1.1 La Thèse de la Réflexion

La Thèse de la Réflexion est la réponse partielle de Shapiro et Ketland au Problème de l'usage, à la question de savoir ce dont une théorie de la vérité *doit* rendre compte. Nous avons vu au chapitre précédent qu'une des raisons pour laquelle Tarski jugeait insuffisante la théorie minimale de la vérité était qu'il n'est pas possible dans l'extension minimale d'une théorie de dériver la moindre généralisation concernant la vérité. En particulier, dans une extension aléthique minimale, on ne peut *prouver* que les *instances* des généralisations logiques[13], mais non ces généralisations elles-mêmes. Cette remarque pourrait naturellement donner lieu à une nouvelle condition d'adéquation pour les théories de la vérité, sous la forme d'une « Thèse de la généralisation », quelque chose comme

Thèse de la Généralisation :

> Une théorie de la vérité doit permettre de prouver les généralisations dont les instances sont toutes des conséquences de la condition d'adéquation de Tarski.

[12] Je remercie Jeffrey Ketland, dont les remarques m'ont aidé à clarifier le sens exact de l'argument de la conservativité.

[13] Voir chapitre 3, section 2.3.

Mais ce n'est pas la thèse que retiennent Shapiro et Ketland. La condition d'adéquation qu'ils proposent est plus forte.[14] L'idée est qu'une théorie de la vérité doit encore être *réflexive* :

Thèse de la Réflexion :

> Dans une extension aléthique d'une théorie A, il doit être possible de prouver que tous les théorèmes de A sont vrais.[15]

Nous avons noté au chapitre précédent que Tarski[16] avait reconnu dans la dérivabilité du principe de réflexion d'une théorie dans sa métathéorie une conséquence bienvenue, lui permettant d'accomplir son programme large : montrer la fécondité de l'application de la théorie de la vérité dans les études fondationnelles. Mais ni Tarski, ni Shapiro, ni Ketland n'ont proposé de justification détaillée de la Thèse de la Réflexion. La réflexivité est également mentionnée par Leitgeb dans la liste des conditions qu'une théorie de la vérité devrait satisfaire, comme « un point qui ne prête pas à controverse ».[17] Mais est-ce bien sûr ? Pourquoi l'extension d'une théorie A par une théorie du concept de vérité (et la syntaxe) *devrait*-elle en droit permettre de dériver la vérité de la théorie de départ ? Quel est le sens de cette exigence ?

Remarquons à titre liminaire que cette Thèse de la Réflexion ne doit pas être confondue avec une autre, plus faible et souvent citée dans la littérature, selon laquelle si P est *un théorème* d'une théorie A, il doit être prouvable dans une extension aléthique de A que P est vrai.[18] Cette condition faible de réflexion revient simplement à exiger que l'inférence d'un énoncé à l'énoncé lui attribuant la vérité soit permise, *pour tout énoncé individuel*.[19] Dans le cas qui nous intéresse, il s'agit de

[14] Que la Thèse de la Réflexion est plus forte que la Thèse de la généralisation a déjà été montré au chapitre 3, section 3.4.

[15] Les énoncés du type « Tous les théorèmes de A sont vrais » sont connus dans la littérature logique comme des *Principes de Réflexion*, dont il existe plusieurs variétés. Nous reprendrons le terme pour désigner plus spécialement les énoncés de la forme que nous venons d'indiquer, en précisant parfois « Principe de Réflexion sur A » pour indiquer sur quelle théorie il s'agit de « réfléchir ». Par analogie, je parlerai aussi parfois de « Principe de Réflexion sur les axiomes de A » pour désigner l'énoncé « Tous les axiomes de A sont vrais ». Quoique nous reprenions ici le terme de « Principe » pour nous conformer à l'usage établi, il sera beaucoup question dans ce qui suit de la façon dont il est possible de dériver ce principe...

[16] Voir chapitre 3, section ? et Tarski (1983), p.236.

[17] Leitgeb (2007), p.278.

[18] Voir par exemple McGee (1991), Friedman et Sheard (1987).

[19] Dans ce travail, rappelons que nous nous intéressons seulement aux prédicats de vérité ne prenant pour arguments que des énoncés ne contenant pas eux-mêmes de prédicats de vérité. Avec cette restriction, la réflexivité faible est trivialement satisfaite comme conséquence de la convention-T.

la possibilité de dériver de A (la théorie), l'énoncé qui affirme que *tous les théorèmes sont vrais* (modulo une théorie de la vérité).

L'argument principal en faveur de la thèse de la réflexion est un argument de *fidélité à l'usage ordinaire*.

L'argument de la Réflexion :

> C'est un fait, pourrait-on soutenir, que si un sujet accepte une théorie A et qu'il possède un concept de vérité, alors il est en position de conclure que tous les théorèmes de A sont vrais.[20] C'est ce qu'est supposé illustrer (entre autres) le petit scénario de Shapiro présenté plus haut : un sujet accepte la théorie A, possède un concept de vérité, et il est capable d'inférer que tous les axiomes, puis tous les théorèmes de A sont vrais. L'idée est donc que si ce scénario représente correctement l'usage ordinaire de la notion de vérité, la fidélité aux attributions ordinaires de vérité commande que d'une théorie A et d'une théorie de la vérité pour le langage de A nous puissions dériver « Tous les théorèmes de A sont vrais ».

De même que la convention-T était pour Tarski une condition d'adéquation permettant de garantir qu'une théorie de la vérité qui la satisfait saisit bien la notion classique de vérité comme correspondance, l'usage crucial rendant compte de la signification de « vrai » étant identifié alors à l'assertabilité des équivalences-T, de même ici un usage jugé central du concept de vérité dans le raisonnement vient fonder la condition d'adéquation*.

Avant de conclure sur la Thèse de la Réflexion, notons que Ketland voit dans cette dernière une extension de l'intuition « d'équivalence » entre un contenu propositionnel et l'attribution de vérité à ce contenu, qui devrait par conséquent être acceptable d'un point de vue déflationniste : non seulement une théorie de la vérité (avec un peu de syntaxe) doit avoir pour conséquence qu'un énoncé et l'attribution de la vérité à cet énoncé sont équivalents, mais encore qu'une *théorie* et l'attribution de la vérité à cette théorie soient équivalents. La convention-T formule l'exigence de la première équivalence, mais comment formuler la seconde ? Les théories étant des ensembles infinis d'énoncés, il n'est pas possible d'exprimer directement avec des moyens finis le principe d'équivalence en question, et nous devons faire un détour.[21] La proposition de Ketland, telle que je la comprends, est qu'étant donné une théorie

[20] On suppose que A elle-même contient sa propre syntaxe.

[21] Tandis que l'équivalence de l'énoncé « la neige est blanche » et de l'énoncé qui lui attribue la vérité est saisi par l'équivalence-T :

de la vérité dans L_A et des principes syntaxiques appropriés, nous devrions être capables, en présence d'une théorie A, de *déduire*, « Tous les théorèmes de A sont vrais », *i.e.* le Principe de Réflexion pour A. Et, à partir du Principe de Réflexion pour A, être capables de déduire les théorèmes de A eux-mêmes. Comme on peut facilement s'en convaincre, la seconde partie de cette condition d'équivalence est automatiquement satisfaite si notre théorie satisfait la convention-T[22], tandis que la première partie est équivalente à la thèse de la Réflexion.

On peut représenter ce mouvement des énoncés vers les théories sur un schéma, avec les notations suivantes : « A » désigne une théorie quelconque (mais contenant une théorie de sa propre syntaxe), les « a_i » désignent les théorèmes de A, « p » désigne un énoncé quelconque, et « T » désigne une théorie de la vérité au sens d'un ensemble d'axiomes gouvernant le predicat « Vrai », jointe une théorie de la syntaxe de A. $Ref_A.$ désigne l'énoncé « Tous les théorèmes de A sont vrais ». En tête des encadrés figurent les contraintes qui sont supposées peser sur T, et en dessous les résultats de dérivabilité que l'on obtient sous ses contraintes :

Convention-T (énoncés) : $Vr(\ulcorner p \urcorner) \vdash_T p$ et $p \vdash_T Vr(\ulcorner p \urcorner)$	→	Convention-T (théories) : $Ref_A. \vdash_T a_i$ et $A \vdash_T Vr(\ulcorner a_i \urcorner)$	→	Convention-T + Réflexivité (théories) : $Ref_A. \vdash_T a_i$ et $\mathbf{A \vdash_T Ref_A.}$

TAB. 3.1: De la convention-T à la Thèse de la Réflexion

« La neige est blanche » est vrai si et seulement si la neige est blanche

comment saisir avec des moyens finitaires l'équivalence d'une théorie A et de l'énoncé qui lui attribue la vérité :

La théorie A est vraie si et seulement si $\cdots ? \cdots$

[22]On suppose toujours que A contient sa propre syntaxe. On peut montrer que si σ est un théorème de A, alors $A \vdash Th_A(\ulcorner \sigma \urcorner)$. Donc si σ est un théorème de A, à partir de

$$\forall x(Th_A(x) \rightarrow Vr(x))$$

et de

$$Th_A(\ulcorner \sigma \urcorner)$$

on peut déduire $Vr(\ulcorner \sigma \urcorner)$. Et de là, avec l'axiome approprié de la théorie minimale de la vérité, on peut dériver σ lui-même.

Notons qu'une conséquence de la Thèse de la réflexion est que la théorie minimale de la vérité n'est pas adéquate* en ce nouveau sens étendu, alors que la théorie tarskienne l'est, d'après les résultats que nous avons déjà mentionnés au chapitre 3. Indépendamment de l'argument contre le déflationnisme, l'Argument de la Réflexion, s'il était correct, serait donc un argument contre la thèse l'idée que la théorie minimale est une explication satisfaisante de la notion de vérité.

3.1.2 La thèse de la conservativité

De même que la Thèse de la Réflexion est une réponse au Problème de l'usage, la seconde hypothèse de l'argument principal est une réponse au Problème de la frontière. La proposition de Ketland et Shapiro peut être formulée de façon suivante :

Thèse de la conservativité :

> La vérité est une notion explicative s'il existe une théorie A, telle que l'extension aléthique de A étend *non-conservativement* A.[23]

La notion de conservativité étant définie comme suit :

Définition 11 (Conservativité)**.** *Soit T une théorie formulée dans un langage L, et T' une extension de T formulée dans un langage L' tel que $L \subset L'$. T' est conservative sur T si et seulement si, pour tout énoncé ϕ du langage L, s'il existe une preuve de ϕ dans T', il existe une preuve de ϕ dans T.*[24]

[23] Où par « extension aléthique de A » il faut entendre la théorie A augmentée d'une théorie de la vérité adéquate pour rendre compte de tous les usages du terme « vrai » jugés essentiels, c'est-à-dire une extension aléthique adéquate*.

[24] La conservativité est relation entre théories, non une propriété des « notions » ou des prédicats, et parler de la conservativité n'a de sens que relativement à une théorie donnée. La distinction conservativité/non-conservativité proposée par Shapiro et Ketland comme réponse au Problème de la frontière doit donc d'abord être comprise comme s'appliquant à des théories et comme donnant une réponse *relative* à un ensemble théorique donné. Le passage d'une thèse relative au caractère « substantiel » d'une théorie relativement à une autre à la thèse concernant la notion de vérité peut être simplement décomposé de la façon suivante :

1. Une théorie B est « substantielle » relativement à une théorie A si et seulement si $A + B$ est une extension non conservative de A.
2. Une théorie B est « substantielle » si, et seulement si, il existe une théorie A relativement à laquelle elle est substantielle.
3. La vérité est une notion explicative si la théorie qui rend compte de ses usages essentiels est « substantielle ».

Pourquoi la Thèse de la Conservativité est-elle de prime abord plausible ? C'est que si, pour toute théorie A, l'extension aléthique de A est conservative sur A, un déflationniste est justifié à affirmer que le pouvoir explicatif de la vérité est nul : un tel résultat montrerait que si un fait que l'on peut décrire sans avoir le concept de vérité et uniquement dans les termes qui sont ceux du langage d'une théorie A, peut être expliqué à partir de la réunion de la théorie A et de la théorie de la vérité pour L_A,[25] il peut aussi être expliqué[26] à partir des seuls principes non aléthiques, ceux de A. Supposons maintenant qu'il existe une théorie A telle que l'extension aléthique de A étend non-conservativement A : alors il y a un fait qui peut être décrit dans le vocabulaire de L_A par un énoncé ϕ et qui possède une explication dans la théorie étendue, mais qui ne peut être expliqué à partir des seuls principes couchés dans la théorie A elle-même. La possibilité d'expliquer un fait non-sémantique à partir de principes aléthiques, alors que ce fait est autrement inexplicable, fournirait un contre-exemple à la thèse déflationniste d'après laquelle la vérité n'a pas de pouvoir d'explication.[27]

[25] A nouveau : on suppose que A contient sa propre syntaxe.

[26] Nous suivons pour l'usage de Shapiro et Ketland où sont assimilés *explication* et déduction aux fins de l'argumentation.

[27] Il y a une petite difficulté dans la formulation des thèses de conservativité que nous devons à présent mentionner, ne serait-ce que pour la laisser de côté par la suite. À strictement parler, la thèse de la conservativité pour la vérité n'est probablement pas celle que Shapiro et Ketland avaient à l'esprit. La raison en est que, lue à la lettre, la thèse implique trivialement que la vérité est substantielle, et ce pour de mauvaises raisons. Ce point a été soulevé par V. Halbach (voir Halbach (2001)) : toute théorie de la vérité pour un langage L donné doit prouver les équivalences-T pour ce langage et, à partir des équivalences-T et de l'ensemble des énoncés valides de ce langage, l'existence de deux individus peut être déduite. (Les dénotations du nom d'un énoncé et de sa négation doivent en effet être distinctes, sous peine de contradiction. Il faut par ailleurs accepter les instances des schémas d'axiomes pour l'identité formulées dans le langage étendu, contenant le prédicat de vérité. Cela doit aller de soi si on regarde l'identité comme une relation logique.) Par conséquent l'ensemble des équivalences-T lui-même n'étend pas conservativement certaines théories très faibles comme l'ensemble des énoncés valides associés à ce langage. Mais il est clair que la non-conservativité sur l'ensemble des énoncés valides est un phénomène parasite relativement à la question qui nous occupe dans la mesure où, pour qu'une théorie de la vérité ait le moindre sens, il nous faut avoir admis auparavant une ontologie de porteurs de vérité. Puisque nous admettons, à la suite de Tarski, que les porteurs de la vérité sont des énoncés formalisés, il est naturel d'admettre, si la conservativité doit nous apprendre quelque chose, que nous ne devons nous y intéresser qu'en relation avec des théories de base qui ne sont pas déjà elle-même étendues non-conservativement par des théories élémentaires de la syntaxe de leur langage. Une convention raisonnable est donc de supposer que les théories de base dont il est question ici sont suffisamment riches pour coder leur syntaxe ou interpréter l'arithmétique de Robinson Q. Nous supposons à partir de maintenant que cette convention a été faite et qu'elle est implicite dans la thèse de la conservativité de la vérité.

Comme nous l'avons vu au chapitre 2, l'extension aléthique minimale d'une théorie est justement une extension conservative de cette théorie.[28] Mais la théorie minimale de la vérité n'est pas réflexive, autrement dit n'est pas adéquate* si la Thèse de la Réflexion est correcte.

3.1.3 La situation logique

A ce point, nous sommes presque à nos fins. Pour conclure l'argument, Shapiro et Ketland ajoutent que c'est une question de logique qu'une théorie réflexive de la vérité étend de façon non conservative certaines théories. Considérons une théorie A « contenant » sa propre syntaxe, éventuellement *via* un codage (l'arithmétique de Peano en premier ordre en est bon exemple) ; la cohérence d'une telle théorie peut s'exprimer par un énoncé de son langage,[29] et Gödel a prouvé de plus qu'un tel énoncé n'était pas prouvable dans la théorie elle-même (second théorème d'incomplétude). Mais il est facile de voir que dans une extension aléthique réflexive de A, on peut prouver la cohérence de la théorie de base, comme le montrait l'argument informel présenté en introduction. De façon un peu plus détaillée l'argument est le suivant (comme dans l'exemple de Shapiro on suppose, pour fixer les idées, que A est une théorie arithmétique) :

1. L'extension aléthique de A prouve « Tous les théorèmes de A sont vrais » (Par Réflexivité),
2. Or « $\neg(0 = 1)$ » est un théorème de A,
3. et l'on peut prouver dans A l'énoncé « $\ulcorner\neg(0 = 1)\urcorner$ est un théorème de A ». (Par 2 et une propriété de la syntaxe, qui fait partie de A par hypothèse. On peut en effet montrer que si σ est un théorème de A, alors « σ est un théorème de A » est un théorème de la syntaxe.)
4. Donc « $Vr(\ulcorner\neg(0 = 1)\urcorner)$ » est un théorème de l'extension aléthique de A. (Par 1 et 3)

[28] Tarski (1983), §5, théorème III. Plus exactement, le théorème III énonce la cohérence relative de l'extension d'une théorie par les équivalences-T relativement à cette théorie elle-même. Néanmoins, la preuve que donne Tarski prouve en fait davantage, à savoir, la conservativité de cette extension. Voir ici le chapitre 3, section 4.

[29] Eventuellement *via* codage. Le sens exact de l'expression « la cohérence d'une telle théorie peut s'exprimer par un énoncé de son langage (*via* codage si besoin) » sera reconsidéré plus tard.

5. Donc « $\neg Vr(\ulcorner 0 = 1 \urcorner)$ » est un théorème de l'extension aléthique de A.
   (Par 4 et adéquation du prédicat de vérité au sens de Tarski[30])
6. Donc « $\ulcorner 0 = 1 \urcorner$ n'est pas un théorème de A » est un théorème de l'extension aléthique de A.
   (Par 1 et 5)
7. Autrement dit on peut prouver dans l'extension aléthique de A que A est cohérente.

Il s'ensuit que les théories adéquates* de la vérité ne sont pas conservatives, et par conséquent que le déflationnisme est faux. Tel est l'argument opposé au déflationniste.

Avertissement : Dans la suite du chapitre j'adopterai le cadre de discussion de Shapiro et Ketland

- 1. En limitant mon attention aux théories *axiomatiques* de la vérité.
- 2. En supposant que la théorie-objet A dont nous discutons les extensions aléthiques *contient sa propre syntaxe*.
- 3. En supposant que toutes les théories dont nous parlons (théorie-objet, syntaxe, extension aléthique) sont des théories du *premier ordre*.

3.2 La thèse de la conservativité

Je pense que l'argument présenté dans la section précédente est incorrect à plusieurs titres. En particulier, je crois que la Thèse de la Réflexion [31] est incorrecte, ou du moins obscure ou ambiguë. J'y reviendrai au chapitre 7. Dans la suite de ce chapitre, cependant, mon intention est de ne pas m'arrêter à la Thèse de la Réflexion mais de me concentrer seulement sur la Thèse de la Conservativité comme réponse au Problème de la frontière. Cette thèse me semble incorrecte, tout spécialement l'application qu'elle trouve dans cette idée que la preuve de l'énoncé de la cohérence d'une théorie A(formulé dans le langage de A) à partir des lois de la vérité (et

[30]

$$\dfrac{\dfrac{Vr(\ulcorner \neg(0=1) \urcorner)}{\neg(0=1)} \quad \dfrac{[Vr(\ulcorner 0=1 \urcorner)]^1}{0=1}}{\dfrac{\bot}{\neg Vr(\ulcorner 0=1 \urcorner)}\,{}^1}$$

[31] Et, par suite, la thèse en vertu de laquelle, dans une extension aléthique de A, la cohérence de A doit, en un sens ou un autre, être prouvable.

de A elle-même) réfuterait le déflationnisme. Ce que je me propose de faire est donc d'accorder momentanément[32] la Thèse de la Réflexion, et de soutenir que la situation logique mise en avant par Shapiro est Ketland illustre un cas de non-conservativité *épistémologiquement neutre*, un phénomène sur lequel on ne peut fonder aucun « argument » en faveur de l'idée que le prédicat de vérité joue un rôle *explicatif*.

3.2.1 La (non-) conservativité et son interprétation

La notion de conservativité a une histoire philosophique chargée au vingtième siècle, les résultats de conservativité jouant en effet un rôle important dans différents programmes *réductionnistes*, au sens large. De quoi s'agit-il ? Supposons un domaine d'entités D régi par des lois T formulées un langage L, et supposons que ces entités, ces notions, et ces lois ont, pour une raison ou une autre, une certaine priorité épistémologique sur les autres dans le corps de la science. Il peut s'agir par exemple du domaine empirique des objets physiques observables, des notions observationnelles, et de ce que les positivistes logiques appelaient des lois empiriques, ou bien, en mathématique, du domaine des entiers, des notions de l'arithmétique finitaires et des lois d'une arithmétique relativement faible comme l'arithmétique primitive récursive. On suppose que ce domaine, ces notions, ces lois jouissent d'une intelligibilité particulière, qui confère un degré de certitude particulier aux théories en question. D'un autre côté, dans notre effort pour comprendre le réel, nous devons souvent nous aventurer loin de ce domaine de discours élémentaire et des entités « réelles » dans lequel nous nous savons assurés dans nos fondements, posant des entités qui n'appartiennent pas au domaine D, formant des notions qui n'appartiennent pas au langage L, et formulant des lois qui ne sont pas dans T. Par définition, l'extension du discours de la science qui en résulte ne jouit pas des bonnes propriétés épistémologiques qui sont celles du domaine base, au sens où l'intelligibilité de ses notions et l'évidence de ses lois sont plus douteuses que ne le sont celles du domaine élémentaire de base. Si la connaissance du domaine étendu est plus douteuse, c'est en général que le caractère spécial des nouveaux objets introduits soulève des questions relativement à l'*accès* épistémologique que nous avons à ces objets : les objets hautement théoriques de la physique nucléaire, par exemple, ne sont pas connus de la même façon que les objets directement observables, et nous n'avons pas pour

[32] Qui ne dit mot consent.

les objets mathématiques « infinitaires » et leurs lois abstraites les mêmes moyens de vérification que ceux qui sont disponibles dans le domaine restreint des objets mathématiques finis et de leurs lois élémentaires. D'où le projet de s'assurer que le domaine *étendu* de la science ne met pas en péril l'édifice entier, en cherchant à *fonder* ce dernier sur le domaine élémentaire.

Pour mener à bien ce projet, deux voies se présentent naturellement. La première est une tentative de réduction au sens fort, qui s'apparente à une élimination : on veut d'une part *définir* toutes les notions scientifiques en termes des notions purement élémentaires et, d'autre part, *déduire* toutes les lois des lois élémentaires. En général, néanmoins, on le sait, une telle réduction est tout simplement impossible.[33] Il y a néanmoins une seconde voie, que l'on pourrait qualifier de réductionnisme *faible* : montrer que la partie de la théorie étendue (avec son domaine étendu, ses nouvelles notions et ses nouvelles lois) qui n'est pas élémentaire est, du moins en principe, *dispensable* dans l'entreprise de connaissance du domaine « réel » ou « élémentaire ». L'idée est que, si, par les méthodes étendues, il est possible de prouver quelque chose relevant du domaine « réel », c'est-à-dire formulable dans les termes non-problématiques élémentaires, alors il doit être possible de le prouver également par les méthodes élémentaires uniquement, c'est-à-dire en ne nous engageant pas sur la vérité d'autres lois que celles à la vérité desquelles nous avons un accès privilégié, le passage par le domaine non élémentaire n'étant en somme qu'un raccourci pratique. Or ceci n'est autre qu'un résultat de conservativité des méthodes « étendues », « théoriques », « idéales » sur les méthodes élémentaires ou « réelles ».[34,35] De façon converse, les résultats de *non-conservativité* peuvent donc être vus comme des résultats d'*indispensabilité*. Si les lois qui gouvernent une no-

[33] C'est un point classique sur lequel il n'est pas utile de revenir ici. Voir par exemple la discussion au début de l'article de Quine « Epistemology naturalized » dans Quine (1969)).

[34] On peut aussi interpréter les résultats de conservativité comme des résultats de *cohérence* des méthodes. Si T est conservative sur T', alors il n'y a rien à craindre d'une utilisation de T comme auxilliaire pour découvrir des vérités relativement au domaine de T', $T + T'$ étant cohérente avec toute extension (cohérente) possible de T' par des vérités du langage de T'.

[35] Le lecteur se demandera peut-être si ce réductionnisme est vraiment plus *faible* que l'autre. La réponse est Oui. Pour prendre un exemple purement mathématique, si A est une théorie du premier ordre, si B est un ensemble d'axiomes constiuant une définition inductive positive d'une notion P n'appartenant pas au langage de A, alors en général on peut montrer que $A + B$ est une extension conservative de A, tandis que P n'est pas en général définissable explicitement dans A. Sur la conservativité des extensions par définition axiomatique inductive positive, voir par exemple Moschovakis (2008). Une autre illustration est fournie par les considérations du chapitre 2. Nous avons vu d'une part que l'extension aléthique minimale d'une théorie A étendait conservativiement A, alors qu'on ne pas définir adéquatement un prédicat de vérité pour A dans A.

tion « théorique » ou « idéale » nous permettent d'expliquer des phénomènes réels que les méthodes élémentaires ne permettent pas d'expliquer, alors cette notion et ces lois sont indispensables. On ne peut pas les considérer simplement comme des artefacts dénués de signification dont l'usage pourrait être justifié au motif qu'il ne s'agit que de raccourcis ou de fiction utiles, et il faut au contraire admettre que ces notions et ces lois ont un contenu propre, et qu'il nous faut les intégrer au rang de constituants primitifs de notre description de l'univers.

Si l'on retrouve ici le tableau conceptuel qui sous-tend une partie de l'élaboration et de la discussion du programme de Hilbert[36] ou les discussions du nominalisme mathématique sous le tour qu'elles ont pris après l'impulsion des arguments dits « d'indispensabilité » de Quine et Putnam[37], de la réponse de Hartry Field[38], et de la discussion de cette réponse par Shapiro[39], on peut s'interroger sur le sens des arguments de conservativité dans le contexte du débat autour du déflationnisme aléthique. Il y a une différence importante entre ces contextes théoriques classiques où les arguments de (non)-conservativité ont eu cours et le contexte de la discussion du déflationnisme. En effet, le déflationniste n'entend pas prouver que le prédicat de vérité est *dispensable*[40], puisqu'au contraire, nous l'avons vu, il tient que le prédicat de vérité *a* un rôle indispensable dans l'expression de certaines généralisations. Le déflationniste ne tient pas non plus que les généralisations mettant en jeu le concept de vérité seraient plus *douteuses* que certaines généralisations « élémentaires », et il ne prétend pas qu'il faille *réduire*, ou *fonder*, ces lois sur des lois plus assurées. Au contraire, pour le déflationniste, la vérité est un concept tout à fait élémentaire et non-problématique, quelque chose comme un concept *logique*. Si quelque chose est problématique dans la position déflationniste, et c'est ce que nous avons identifié comme le véritable enjeu philosophique ici, c'est la possibilité de faire une distinction épistémologiquement intéressante entre les concepts ayant un rôle « purement expressif » et les autres, les concepts ayant un rôle « explicatif ». Il me semble par

[36]Voir par exemple la présentation de Smorynski (1977) ou, en français, le dernier chapitre de Blanché et Dubucs (1997). Pour une discussion critique minutieuse des interprétations du programme de Hilbert en termes de conservativité, voir Detlefsen (1990).

[37]Voir par exemple Putnam (1971), Quine (1976*a*).

[38] Field (1980)

[39]Shapiro (1983).On trouvera une étude approfondie des arguments d'indispensabilité contre le nominalisme et des discussions subséquentes dans Burgess et Rosen (1997).

[40]Le terme est rare en français, mais d'après le Littré son usage est attesté dès le XIVè siècle. Voyant combien son usage est pratique (mis en opposition à *indispensable*), j'espère que le lecteur ne verra pas d'inconvénients à ce que je m'y adonne.

conséquent qu'en replaçant la notion de conservativité au cœur du débat sur le déflationnisme, Shapiro et Ketland nient simplement qu'une telle distinction soit possible, et ils ne font alors que réaffirmer l'indispensabilité de la vérité, ou bien qu'ils ont opéré un déplacement conceptuel relativement à l'usage philosophique classique de la notion de conservativité. C'est sans doute la seconde affirmation qui est vraie, et l'on peut décrire ce déplacement conceptuel de la façon suivante : non seulement les résultats de (non-) conservativité permettent de statuer sur le caractère dispensable ou indispensable des principes aléthiques, mais ils disent aussi quelque chose de plus fin sur le *statut épistémologique* des principes en question, quelque chose qui permet de disqualifier la position déflationniste en montrant que le concept de vérité a un rôle « explicatif ». Les phénomènes gödéliens sont alors mis à contribution : une théorie contenant sa propre syntaxe ne peut pas prouver (comprendre : expliquer) l'énoncé (standard) exprimant sa cohérence, mais il est possible de prouver (comprendre : d'expliquer) la vérité de cet énoncé dès que l'on est muni d'un concept de vérité satisfaisant. Le fondement de mon objection à l'argument par la non-conservativité se trouve là, dans l'idée que les résultats de (non-)conservativité *ne* sont *pas* assez fins pour permettre de trancher la question de savoir si les lois qui gouvernent ce concept jouent le rôle d'authentiques principes explicatifs.

Pour illustrer l'idée que la signification épistémologique des phénomènes de non-conservativité n'est pas toujours transparente, et avant d'en venir à la discussion de la vérité proprement dite, je voudrais commencer par en donner deux exemples assez différents mais mettant tous les deux en jeu des expressions logiques.

(Exemple 1) Le premier exemple illustre les difficultés que rencontre l'interprétation des résultats de (non-)conservativité dans le contexte d'une discussion sur les constantes logiques. Dans son livre *Logical basis of Metaphysics*,[41] Michael Dummett discute le phénomène suivant : les règles d'introduction et d'élimination de la disjonction étendent non-conservativement une logique possédant une certaine disjonction non-standard que nous allons noter $\vee^*$. L'usage de cette dernière est gouverné par les règles d'introduction habituelles de la disjonction, mais sa règle d'élimination contient une restriction relativement à la règle d'élimination « naturelle » de la disjonction standard. En déduction naturelle (présentée en séquents), la règle d'élimination est la suivante :

[41] Dummett (1991), p.290.

$$\frac{\Gamma \vdash A \vee^* B \qquad A \vdash \mathrm{C} \qquad B \vdash \mathrm{C}}{\Gamma \vdash C} \vee\text{-elim}$$

Ici, A et B sont des formules, tandis que Γ est un multiensemble de formules. La règle d'élimination pour le connecteur $\vee^*$ ne permet l'élimination que lorsque, dans les deux prémisses mineures, C a été inféré en un pas d'une seule formule, A ou B respectivement. Supposons que notre logique de départ ne contienne que les règles gouvernant la conjonction $\wedge$, notons-la $L_\wedge$. On peut montrer que les règles standard de la disjonction étendent conservativement $L_\wedge$, mais l'on peut également montrer que les règles gouvernant la disjonction tronquée étendent conservativement $L_\wedge$. Néanmoins, et c'est là que les choses deviennent intéressantes, $L_{\vee,\vee^*,\wedge}$ est une extension *non conservative* de $L_{\vee^*,\wedge}$.[42] Le phénomène est assez général : certaines règles/théories, qui étendent de façon conservative un bon nombre d'autres théories, donnent parfois lieu à des extensions non-conservatives dans des circonstances qui, pour une raison ou une autre, semblent « spéciales ». Ici, une notion logique ($\vee$), dont les règles forment ordinairement des extensions conservatives, échoue soudain au test de la conservativité dans des conditions très particulières, sur la base d'une théorie en quelque sorte incomplète de la disjonction, parce que la disjonction entre dans des interactions particulières avec ce connecteur étrange du langage de base.[43] Mais que dit de la disjonction la non-conservativité observée dans ces conditions spécifiques ? L'interprétation épistémologique devra sans doute être prudente.

(Exemple 2) Mon second exemple, en rapport avec l'arithmétique de Peano, ajoute au problème d'interprétation posé par les constantes logiques un problème d'interprétation des phénomènes de non conservativité lorsque ceux-ci apparaissent en lien avec l'extension de *schémas* dont les instances ont été restreintes dans un premier temps, pour des raisons artificielles ou contingentes. Le simple fait logique qui retient mon attention est le suivant :

PA est une extension *non conservative* de la théorie PA_{At},

où PA_{At} désigne une théorie identique à PA, à ceci près que le schéma d'induction est restreint aux formules atomiques, c'est-à-dire aux formules ne contenant pas de

[42] Pour le voir, noter que dans $L_{\vee^*,\wedge}$ la distributivité n'est pas démontrable. Mais dans $L_{\vee,\vee^*,\wedge}$, la distributivité de $\vee$ sur $\wedge$ se « transfère » à $\vee^*$

[43] Il est également bien connu des logiciens que les règles classiques d'introduction et d'élimination des connecteurs forment une extension non-conservative du système fondé sur leurs analogues intuitionistes. La signification de ce fait étant débattue, l'exemple présenté ici me satisfait davantage.

connecteurs ou de quantificateurs.[44]

Ce fait illustre le principe suivant : lever une restriction sur le langage dans lequel peuvent être choisies les formules qui instancient le schéma d'induction a pour effet de renforcer la théorie. D'un autre côté, le langage libéralisé n'est que le langage engendré à partir du langage arithmétique par la seule adjonction des constantes logiques ordinaires. Les logiques gouvernant les connecteurs et quantificateurs n'ont pas changé dans le passage de PA_{At} à PA, on a simplement permis à des prédicats plus complexes d'instancier le schéma d'induction, c'est-à-dire plus seulement les prédicats primitifs du langage, mais également tous les prédicats complexes qu'il est possible de former en composant les prédicats primitifs grâce aux expressions logiques. En un sens, en passant de PA_{At} à PA on a simplement enrichi le pouvoir expressif du langage dans lequel peuvent être choisies les formules qui instancient les lettres schématiques du schéma d'induction. Quelle conclusion doit-on en tirer ? Je ne crois pas que nous puissions en conclure, par exemple, que c'est notre compréhension des concepts logiques et des lois logiques qui nous a permis d'*apprendre* quoi que ce soit de nouveau à propos des entiers ici.[45] Au contraire, si nous tenons que les constantes logiques sont des exemples paradigmatiques d'outils « purement expressifs », au motif, par exemple, que leurs lois ont de fortes propriétés de conservativité[46], alors le résultat mentionné ne semble pas de nature à nous faire changer d'avis sur ce point.[47] Mais un outil expressif, c'est précisément ce pour quoi les déflationnistes tiennent le prédicat de vérité. Pourquoi alors ne pas envisager la possibilité que les résultats de non-conservativité rapportés par Shapiro et Ketland

[44] Voir par exemple Hàjek et Pudlàk (1993), Théorème 1.26 p.220 et Théorème 1.29 p.221.

[45] Comme nous le verrons bientôt plus en détail, ajouter simplement les clauses tarskiennes pour la vérité à PA ne produit qu'une extension conservative de PA : la non-conservativité n'apparaît que si de nouvelles instances du schéma d'induction, mettant en jeu le vocabulaire aléthique, sont ajoutées à PA. Stewart Shapiro, qui est parfaitement au fait de la situation technique, écrit pourtant :

> A la lumière du requisit de conservativité, si le déflationniste s'en tient à la notion de conséquence en premier ordre, alors il ne peut pas même accepter A'' [ie : l'extension aléthique tarskienne de PA avec schéma d'induction étendu] comme explication partielle de la vérité arithmétique. Si la vérité est admise dans le principe d'induction, il est contre-intuitif de maintenir que la notion est fine [thin] et non-substantielle. Comment la notion de vérité arithmétique peut-elle être fine, si en l'invoquant nous pouvons *apprendre* davantage à propos des entiers naturels ?(Shapiro (1998*b*), p.499, je souligne)

[46] En général. Voir exemple précédent pour une réserve.

[47] Nous renvoyons le lecteur que ne convainc pas d'emblée l'idée que les constantes logiques ne sont pas « subsantielles » au sens qui est en jeu ici aux deux chapitres suivants de cette thèse.

sont l'effet d'un phénomène analogue à celui que nous venons d'observer avec les constantes logiques ?[48]

L'interprétation à donner aux phénomènes de non-conservativité, on le voit, ne va pas toujours de soi, et le diable est souvent dans les détails. Je voudrais maintenant soutenir que l'argument de Shapiro et Ketland, formulé en terme de non-conservativité, n'a pas la force que ces auteurs lui prêtent, en montrant que les phénomènes de non-conservativité sur lesquels ils attirent notre attention sont intuitivement compatibles avec l'hypothèse que le prédicat de vérité est simplement un outil expressif. Pour y parvenir, je me propose de prêter attention au statut épistémologique et logique des *schémas d'axiomes* dans ces théories. Une fois que les ambiguïtés de ce statut auront été levées, il deviendra plausible que la non-conservativité des théories de la vérité sur des théories comme l'arithmétique de Peano est un phénomène inoffensif pour la conception déflationniste. Ma conclusion sera que la thèse de la conservativité n'est pas une réponse adéquate au Problème de la frontière.

3.2.2 Non-conservativité et vérité

Une façon d'énoncer le point que je m'apprête à faire est de dire que l'affirmation de cohérence dérivée dans la métathéorie contenant les principes aléthiques[49] ne peut pas être traduite dans le langage de la théorie de base, *à moins que nous n'en sachions plus sur le domaine attendu de notre théorie de base que ce que nous en avions dit pour commencer (dans la théorie de base)*. Pour le dire autrement : pour constater un phénomène de non-conservativité significatif , nous devons ad-

[48]Remarquons que, contre cette idée, on peut tout simplement vouloir nier qu'il soit possible d'établir une distinction entre ce que nous avons appelé les « outils expressifs » et les « outils explicatifs » du langage, et soutenir, dans une veine quinienne, que tous nos concepts ont indissociablement une fonction explicative et expressive. Pour Quine en effet, le holisme de la confirmation induit un holisme sémantique, et une manifestation de ce holisme est l'impossibilité de distinguer une classe privilégiée d'expressions sur la base de leur absence de contribution spécifique à la signification empirique. Mais ce mouvement théorique n'est pas, je crois, dans l'esprit de l'argument de la conservativité, et une critique du déflationnisme sur cette ligne argumentative aurait requis un tout autre angle d'attaque. Enfin j'ajouterai que, quand bien même il n'y aurait qu'un continuum de notions entre les concepts dont la fonction est plutôt expressive (« tout « , « et », « = ») et ceux dont la fonction est plus directement explicative (« être un électron », « être un cardinal supercompact »), un déflationniste pourrait bien se satisfaire de l'idée que, dans ce spectre, la vérité trouve sa place du côté des connecteurs logiques et des quantificateurs.

[49]Ce que nous appelons de façon imprécise la métathéorie peut varier selon le cadre de travail dans lequel nous nous plaçons. Le point important est qu'il s'agit de la théorie qui contient les lois gouvernant le concept de vérité.

mettre que nous avons laissé implicite dans notre formulation de la théorie de base A un principe de preuve que nous savons correct, et que nous avons explicité ce principe dans le cours de notre dérivation subséquente de la cohérence de A. Cette distinction entre ce qui est implicite et ce qui est explicite dans la formulation d'une théorie a une contrepartie précise. La clé est ici de distinguer entre les schémas d'axiomes compris comme *listes* et les schémas d'axiomes compris comme *règles*. Les schémas compris comme listes dans la formulation d'une théorie ne sont qu'un artifice métalinguistique pour formuler finiment un ensemble infini, *mais bien défini*, d'axiomes, à savoir les instances du schéma obtenues en remplaçant les lettres schématiques qui y figurent par des formules d'un vocabulaire approprié, en général entendu comme le vocabulaire ambiant de la théorie. Les schémas compris comme *règles* sont des *schémas ouverts* (*open-ended schemas*).[50] Ils doivent être compris comme engendrant sans fin de nouveaux axiomes à mesure que le langage environnant s'enrichit de nouveaux moyens d'expression, y compris des moyens d'expression dont nous n'avions pas idée lorsque nous avons formulé la théorie pour la première fois. Le point important est que lorsqu'une théorie est formulée avec des schémas ouverts, il existe un fossé entre leur portée épistémologique réelle et leur pouvoir logique (formel), entre ce à quoi ils nous engagent et l'ensemble de leurs conséquences logiques dans un environnement théorique donné.

Considérons le cas de PA, l'arithmétique de Peano en premier ordre dont les axiomes ont été présenté au chapitre 2. Il y a en somme deux théories de l'arithmétique de Peano en premier ordre : celle que, faute de mieux, on pourrait appeler la théorie formelle standard, dont la liste des axiomes est exactement spécifiée, et puis il y a la théorie schématique dans la formulation de laquelle le schéma d'induction est compris comme étant un schéma ouvert. Notre remarque, grossièrement formulée, est que lorsque nous dérivons la cohérence de PA dans notre métathéorie contenant les principes aléthiques, nous devons étendre le schéma d'induction de l'arithmétique de Peano au nouveau vocabulaire contenant le prédicat de vérité, ou faire quelque chose d'équivalent. Pour distinguer les deux formulations de l'arithmétique de Peano en premier ordre, je noterai PA l'axiomatisation ordinaire et $PA(S)$ l'axiomatisation schématique, celle où le schéma est compris comme étant ouvert.[51] Avec ces

[50] Je ne sais pas quel auteur a le premier fait cette distinction en ces termes. Elle est en tout cas devenue classique dans la littérature. Voir par exemple Burgess et Rosen (1997) pour une discussion dans un contexte philosophique, Feferman (1991) pour une discussion en termes logiques.

[51] Voir Feferman (1991) pour un système de logique schématique. Dans le système de Feferman les schémas ne sont plus des outils métathéoriques de représentation d'une théorie-objet mais

notations, nous avons donc d'un côté le fait sur lequel s'appuie l'argument de la conservativité contre le déflationnisme :

Fait 1. *L'extension aléthique tarskienne de $PA(S)$ est une extension non-conservative de $PA(S)$.*

mais d'un autre côté nous avons le fait logique suivant :

Fait 2. *L'extension aléthique tarskienne de PA étend conservativement PA.*[52]

Rappelons-nous de ce que nous avons vu au chapitre précédent. J'avais fait remarquer que l'extension tarskienne $T(A)$ d'une théorie A permettait de prouver la cohérence de A. L'énoncé de la cohérence de la théorie A était un énoncé syntaxique que Tarski dérivait dans la métathéorie. Mais que se passe-t-il si A elle-même est une théorie syntaxique ? Alors l'énoncé de la cohérence de A est un énoncé du langage de la théorie de base. Mais cet énoncé n'est pas prouvable dans A, d'où maintenant la non-conservativité. Mais, il y a un « mais ». La dérivation de l'énoncé de la cohérence dans la métathéorie utilise une instance du schéma d'induction de la théorie syntaxique dans lequel figure le prédicat de vérité. C'est ce qui est visible dans la dérivation informelle de la cohérence de A donnée en début de chapitre, dont un passage crucial est :

1. ...
2. Tous les axiomes de A sont vrais ;
3. Toutes les règles d'inférences préservent la vérité ;
4. *Donc* tous les théorèmes de A sont vrais
   (À partir de 2 et 3, par une instance du schéma d'induction sur les théorèmes

deviennent des formules bien formées de la théorie objet elle-même, les règles logiques ordinaires étant augmentées de règles spécifiques de substitution. Ce que Feferman essaie de faire est donc précisément de formaliser la notion de schéma ouvert.

[52] Voir par exemple Halbach (1999*a*), Th. 3.1, p.359, pour une preuve « syntaxique ». Les preuves modèle-théoriques sont connues depuis longtemps. Voir par exemple Kotlarski (1991), Kaye (1991). D'un point de vue technique, il est à noter que, malgré le résultat de conservativité, il *n*'est *pas* vrai que tout modèle de PA peut être étendu à un modèle de ce que nous avons appelé l'extension aléthique tarskienne de PA (i.e. $PA+$ les axiomes tarskiens). McGee (2006) en conclut que, malgré la conservativité « syntaxique », les axiomes tarskiens ne peuvent être véritablement conçus comme une simple *définition* implicite de la vérité, et propose une modification. Nous laissons ces subtilités de côté ici.

de A mettant en jeu une propriété[53], Vr, qui n'était pas définissable dans L_A).[54]

5. ...

L'instance cruciale du schéma d'induction n'est pas une instance du schéma d'induction de la théorie de base. Par conséquent, pour dériver l'énoncé de la cohérence dans la métathéorie, il faut non seulement ajouter à A les axiomes récursifs de la vérité proprement dits, mais aussi étendre le schéma d'induction de A (ce qui pour Tarski, nous l'avons dit, était un geste naturel qui ne méritait pas qu'on s'y arrête). Le Fait 2 vient préciser que, sans l'élargissement du schéma, la seule extension de A par les axiomes récursifs de la vérité est en fait *conservative* sur A.

Pour prouver la cohérence de A, nous devons donc en passer par l'*explicitation* d'un savoir (que certains principes de preuves dans A sont corrects), qui n'était au mieux qu'implicite dans notre formulation de A. Mais alors la valeur de l'argument de la conservativité contre le déflationisme s'en trouve considérablement diminuée : si, prêtant attention au fait qu'il est nécessaire, pour observer un phénomène de non-conservativité, de déployer plus avant des connaissances que nous avons concernant les lois qui gouvernent le domaine de A et que nous avions laissées implicites dans notre première formulation de la théorie, alors il devient très naturel d'interpréter le phénomène de non-conservativité comme un signe certain du *pouvoir expressif* de la vérité. La vérité nous a permis d'*exprimer*, ou de *formuler* une théorie plus forte des faits arithmétiques que celle que nous étions en mesure de formuler auparavant, et si les lois tarskiennes ont permis de *définir* axiomatiquement ce prédicat de vérité, elle n'ont pas joué dans le processus le rôle de principes d'explication.[55] Si nous nous souvenons du fait que, du point de vue déflationniste, le pouvoir expressif de la vérité est sa véritable raison d'être, cette conclusion, loin d'être problématique, devrait au contraire le conforter.

[53] En extension.

[54] Voir Chapitre 3, section 3.2.2.1, pour le schéma en question.

[55] Voir note 65 p.165 pour une comparaison avec ce qui ce passe quand le prédicat de vérité est introduit par la théorie minimale.

3.3 Déflationnisme et conservativité : quelqu'un a-t-il changé de sujet ?

Nous venons de voir qu'une condition nécessaire de la non-conservativité est le caractère schématique « ouvert » de la théorie arithmétique de base. Et j'ai indiqué qu'une fois prêté à cette condition l'attention qu'elle mérite, les conclusions anti-déflationnistes ne s'imposaient plus. Je veux maintenant renforcer mon argument en en précisant le sens, ce que je me propose faire à l'aide d'une petite expérience de pensée, puis en confrontant mes conclusions à la notion d'explication que Shapiro et Ketland semblent avoir en vue. Je reviens sur les conclusions générales par la suite.

3.3.1 Préambule

Imaginons qu'une civilisation très semblable à la notre se soit développée à nôtre insu sur une autre planète. Les terri*a*ns (c'est ainsi qu'ils se nomment et que nous les nommerons) sont identiques à nous à de nombreux égards, mais en mathématiques ils en sont venus à reconnaître, au-delà de l'arithmétique, l'intérêt d'un domaine connexe, l'arithm*a*tique. Pour une raison qui nous échappe, ces nouveaux pythagoriciens pensent que les n*a*mbres naturels sont les blocs élémentaires qui constituent l'univers et, pour cette raison même, leur étude leur importe au plus au point. Une axiomatisation partielle de l'arithm*a*tique est donnée par $PA \bigcup \{\neg con_{PA}\}$, où $\neg con_{PA}$ désigne la négation de l'énoncé *du langage de* PA exprimant la cohérence de PA sous le codage standard de Gödel (cet énoncé est vrai dans $\mathbb{N}$ si et seulement si PA n'est pas cohérent). Des axiomes supplémentaires ont été proposés pour renforcer l'axiomatisation de l'arithm*a*tique, mais ils sont actuellement très controversés et nous les laisserons donc de côté. Pour en rester à l'ethnologie, il doit encore être remarqué que les terri*a*ns n'utilisent la plupart du temps que la partie de l'arithm*a*tique axiomatisée par PA, même lorsqu'il s'agit pour eux d'étudier les n*a*mbres naturels ; de plus ils font usage des mêmes conventions et formalismes que nous en matière de logique et, plus admirable encore, ils utilisent eux-mêmes le nom « PA », mais dans leur cas pour désigner aussi l'axiomatisation partielle de l'arithm*a*tique qui est formellement identique à la nôtre. En fait, dans leur langage, « PA », les numéraux (« 0 », « 1 », ...) et les symboles d'opérations « successeur », « + », « . » sont ambigus ; ils désignent tantôt une axiomatisation de l'arithmétique, des nombres entiers et les opérations bien connues sur leur domaine, tantôt, respec-

tivement, une axiomatisation partielle de l'arithm*a*tique, les éléments du segment initial des n*a*mbres naturels, et les opérations moins familières qui leurs sont associés ; dans la pratique des terri*a*ns, néanmoins, cette ambiguïté ne pose pas de problème, le contexte rendant clair ce dont il s'agit. *Nous* écrirons parfois « PA^* » pour désigner le second système et le distinguer de (notre) PA. PA^* et PA sont donc formellement identiques mais intentionellement différents, le premier parlant des n*a*mbres, le second devant être interprété comme parlant des nombres. Il nous faut enfin rapporter pour finir que ces gens ont également deux théorèmes de G*ä*del. Ces théorèmes sont, je dois l'admettre, tout aussi bons que les théorèmes de Gödel. Ils les formulent en général de la façon suivante :

Théorème 1 (Premier théorème)**.** *Si T est une théorie cohérente, récursivement énumérable, et suffisamment riche, alors T est incomplète.*

Théorème 2 (Second théorème)**.** [56] *Si PA est cohérente, alors :*

$$PA \nvdash \forall x \neg Pr_{PA}(x, \ulcorner 0 = 1 \urcorner) \tag{3.2}$$

Tout ceci est parfaitement standard chez les terri*a*ns. Le second théorème avait reçu une attention particulière parce que la question de savoir si $\neg con_{PA^*}$ était ou non indépendante de PA^* était longtemps restée ouverte. Ceci étant dit, bien entendu, quand ils nous entendent affirmer que les théorèmes de G-d-l montrent qu'il y a « des énoncés vrais indécidables dans PA », ces gens sont d'accord avec nous, mais ils ne sont pas d'accord sur le fait que con_{PA} en fait partie ![57]

3.3.2 Le dialogue

Un jour, « quelqu'un de chez nous », appelons-le Pierre, qui ne connaît pas grand chose aux terri*a*ns, engage une conversation sur le pouvoir explicatif de la vérité avec

[56] Bien entendu, nous traduirions le second théorème de G*ä*del de la façon suivante, si nous le souhaitions :

Théorème 3 (Second théorème, traduit)**.** *Si PA^* est cohérente, alors :*

$$PA^* \nvdash \forall x \neg Pr_{PA^*}(x, \ulcorner 0 = 1 \urcorner) \tag{3.1}$$

Je laisse au lecteur le soin de juger de l'intérêt d'une telle traduction.

[57] Il n'y a rien de profond ici : c'est simplement que $\mathbb{N}$ n'est pas l'interprétation saillante de PA dans les contextes conversationnels terri*a*ns.

l'un d'eux. Voici une relation de la conversation[58] :

Pierre - Croyez-vous que les axiomes de PA soient vrais ?

L'Etranger - Oui, je crois que les axiomes de PA^* sont vrais.

Pierre - Et croyez-vous que les règles d'inférences préservent la vérité ?

L'Etranger - Je le crois en effet.

Pierre - Vous croyez donc que tous les théorèmes de PA sont vrais ?

L'Etranger - Oui.[59]

Pierre, tout à coup excité - Puisque PA prouve que $0 \neq 1$, vous croyez que c'est vrai, donc que $0 = 1$ n'est pas vrai, et par conséquent vous croyez donc que PA ne prouve pas que $0 = 1$, autrement dit vous croyez que PA est cohérent.

L'étranger - Tout à fait.

Pierre - Êtes-vous d'accord par conséquent qu'une bonne théorie de la vérité pour le langage de PA doit avoir pour conséquence la cohérence de PA ?

L'Etranger - En effet, c'est une bonne chose qu'une théorie de la vérité permette de rendre compte de ce raisonnement.

Pierre - Vous savez, la théorie de la vérité de Tarski pour le langage de PA fait précisément cela.

L'Etranger (sincèrement) - Oui, c'est assurément un beau travail que T*a*rski nous a laissé là.

Pierre - Mais voyez : PA ne prouve pas la cohérence de PA, tandis que la théorie augmentée des principes généraux qui gouvernent le prédicat de vérité, cette théorie dis-je, *prouve* la cohérence de PA. La vérité a un pouvoir d'explication, elle permet d'expliquer de nouveaux faits dont PA ne peut rendre compte, des faits qui sont exprimables dans le langage de PA. Mon acceptation de PA ne m'engage pas logiquement à accepter con_{PA}, mais une fois reconnue la vérité de PA, je suis forcé d'accepter « con_{PA} ». Il y a un fait purement arithmétique qui est expliqué par l'attribution de vérité à PA.[60]

L'Etranger (comprenant la situation) - Mais la cohérence de PA n'est

[58] J'ai essayé de désambiguïser les occurrences de « PA » en utilisant « PA^* » quand c'était nécessaire.

[59] L'étranger croit donc que tous les théorèmes de PA^* sont vrais...

[60] Cette attribution est une conséquence du caractère réflexif de la théorie tarskienne. Voir le début du chapitre section 1.

pas un fait arithm*a*tique !
Pierre -Comment cela ?
L'Etranger - J'admets sans difficulté que votre raisonnement est un raisonnement arithmétiquement correct, mais ce n'est pas un raisonnement arithm*a*tique. Je m'explique. Tout d'abord il est faux que l'énoncé con_{PA} exprime la cohérence de PA dans l'arithm*a*tique. Deuxièmement, vous ne pouvez pas, en arithm*a*tique, raisonner inductivement sur la vérité comme vous l'avez fait. C'est heureux, car la *négation* de con_{PA} est un énoncé arithm*a*tique vrai ! Par conséquent, je m'accorde avec vous qu'à partir d'une théorie A, de la théorie de la vérité-dans-$\mathcal{L}_A$ et d'une théorie arithmétique suffisamment forte, la cohérence de la théorie A s'en suit. Mais aucun énoncé indécidable de la théorie de base, même « exprimant » la cohérence de la théorie, ne suit logiquement de A et de sa théorie de la vérité seulement. La théorie de la vérité, par elle-même, n'offre pas de base pour établir quelque nouveau fait dans le vocabulaire de la théorie de base, comme le montre du reste notre malentendu.

3.3.3 Tirer les leçons

Avant d'en venir à la signification de cette fable, il me faut clarifier un ou deux points. Pour commencer, il faut souligner que si la vérité de PA et la vérité de PA^* sont deux faits bien distincts, en revanche la cohérence de PA et la cohérence de PA^* sont un seul et même fait : les deux théories sont formellement identiques, et la propriété de cohérence est une propriété des systèmes formels. De plus, l'énoncé con_{PA} du langage $\mathcal{L}_{PA}$ (morphologiquement identique à $\mathcal{L}_{PA^*}$) exprime la cohérence de PA^* (i.e. la cohérence de PA). En quel sens ? D'abord et avant tout au sens où, modulo un codage dans PA de la syntaxe de PA^* (i.e. la syntaxe de PA), il est vrai *dans* $\mathbb{N}$ si et seulement si PA^* est cohérent. Mais ce n'est bien entendu *pas* le cas que con_{PA} est vrai dans l'arithm*a*tique si et seulement si PA^* est cohérente. Le fait que con_{PA} soit vrai ou non dans l'arithm*a*tique[61] n'a rien à voir avec la cohérence de PA^*.

Un second point à présent. Au cours de son argument Pierre raisonne par induc-

[61] Ce que l'on pourrait exprimer, dans le langage modèle-théorique, en disant qu'il n'est pas vrai dans le modèle attendu des axiomes de PA^*, lequel est simplement un modèle non-standard de l'arithmétique.

tion de la façon suivante : les axiomes sont vrais, les règles préservent la vérité, donc tous les théorèmes sont vrais. Cette inférence est correcte mais, comme le remarque l'Etranger, c'est une inférence arithmétique, mais non arithm*a*tique. La raison en est que cette induction emploie un vocabulaire qui n'appartient pas au langage de PA (il contient « vrai »), et ces instances du schéma d'induction ne sont pas correctes en arithm*a*tique, même si l'axiomatisation partielle PA^* n'en dit rien. Pour prouver la cohérence de PA^*, pour produire l'argument inductif qui permettra de conclure à la cohérence de PA^*, nous faisons appel dans la métathéorie, (celle où sont couchés nos principes aléthiques) à un ensemble d'axiomes d'une théorie de la syntaxe de PA^* (i.e. la syntaxe de PA), et ces axiomes contiennent un schéma d'induction s'appliquant au vocabulaire étendu de la métathéorie, contenant le prédicat de vérité. Des variations sont possibles sur la définition précise du cadre du travail métathéorique (axiomes pour la vérité v.s. définition explicite dans une logique enrichie, vocabulaire proprement syntaxique ou syntaxe codée dans un vocabulaire arithmétique, etc. voir section suivante), mais dans tous les cas il y a un énoncé du métalangage, appelons-le CON_{PA}, qui exprime la cohérence de PA^* au sens où il est vrai dans le modèle attendu de la métathéorie si, et seulement si, PA^* est cohérent. Et *cet* énoncé est prouvable dans la métathéorie.

Quel est le mécanisme à l'œuvre dans cet fable ? Nous avons une situation dans laquelle deux individus croient parler d'une même théorie. Même après en avoir donné tous les axiomes (et par conséquent tous les théorèmes), à la faveur d'une homonymie extraordinaire, la confusion est entretenue. Les deux individus ont par ailleurs la même théorie du concept de vérité. Ils s'accordent que la théorie dont ils pensent qu'elle est l'objet de la conversation est vraie, et que de cela ils peuvent conclure que cette théorie est cohérente. Pourtant l'Etranger n'est pas en mesure d'en conclure quoi que ce soit de neuf relativement au domaine d'objets qui était celui de sa théorie de base. Tous deux ont expliqué la cohérence de la théorie de base, mais seul Pierre a expliqué un fait relevant du domaine de sa théorie de base. Mais comment Pierre peut-il savoir qu'il a par là même expliqué un nouveau fait arithmétique ? Il y a là quelque chose à expliquer, puisque l'Etranger n'est pas en droit de faire de même. Il y a donc des principes de preuve arithmétique que Pierre a engagés dans le cours de l'argument et qui n'avaient pas été précisés pour commencer, principes que l'Etranger n'admet pas comme principes de preuve pour les faits relevant du domaine qu'il a en vue.

Pour dire les choses un peu plus précisément, supposons que j'accepte PA comme

vraie, et que j'ai une théorie de la vérité et de la syntaxe. Que je comprenne les énoncés de PA comme de l'arithmétique, de l'arithm*a*tique, ou n'importe quoi d'autre, dans la métathéorie je serai en mesure de prouver que PA est cohérent.[62] Les choses sont comme elles doivent être ici, puisque la cohérence de PA n'a rien à voir avec la façon dont nous comprenons PA, mais a à voir seulement avec ses caractéristiques formelles. Donc l'affirmation supplémentaire que nous avons ce faisant dérivé une vérité relevant du domaine qui est celui de la théorie de base, autrement dit l'arithmétique dans le cas de l'argument de Shapiro et Ketland, cette affirmation ne peut être soutenue que par un argument à l'effet que notre métathéorie est correcte en tant que théorie arithmétique : souvenons-nous que la métathéorie n'était pas correcte en tant que théorie arithm*a*tique ! Mais comment savons-nous que notre métathéorie est arithmétiquement correcte ? Rien dans notre théorie de base, PA, ne le garantit. Clairement, notre reconnaissance que la métathéorie est correcte relativement à l'interprétation attendue de la théorie de base, est une conséquence de notre reconnaissance que quelque système plus fort que PA est arithmétiquement correct, et non l'inverse ! En d'autres termes, l'affirmation que la métathéorie est non-conservative sur PA mais qu'elle est arithmétiquement correcte, revient tout simplement à *poser* de nouveaux axiomes pour l'arithmétique, en plus de ceux présents dans PA. Ce n'est pas la vérité, semble-t-il donc, qui fait le travail de non-conservativité, mais des connaissances d'arithmétique qui nous permettent de reconnaître comme arithmétiquement correcte la métathéorie. La théorie de la vérité par elle-même ne permet pas d' *apprendre* ou d'*expliquer* quelque chose de nouveau à propos des entiers, en dépit de la non-conservativité, parce que pour observer le phénomène en question, nous devons à un certain point étendre notre théorie arithmétique de départ, et d'une façon qui n'est dictée ni par notre compréhension du concept de vérité lui-même, ni par les principes formulés dans théorie de base. Que l'on regarde cette extension comme une pure décision, comme pourrait le faire un conventionnaliste en matière d'arithmétique, ou qu'on la regarde comme dictée par la volonté de formuler une connaissance restée jusque-là implicite, cela ne change rien au fond : au bout du compte, la non-conservativité ne fait que refléter la décision que nous avons prise à un moment de renforcer nos axiomes, ou avérer la possibilité que nous avons eu à un moment de le faire. C'est cette décision que Pierre a prise subrepticement en utilisant une instance du schéma d'induction formulée à l'aide du prédicat de vérité pour prouver la cohérence de PA.

[62] En *interprétant* PA dans ma métathéorie.

3.3.4 Epistémologie de la vérité et épistémologie des termes théoriques

La notion d'explication et de pouvoir explicatif auxquels Shapiro et Ketland se réfèrent par défaut semble être simplement celle du modèle déductif-nomologique pour les sciences *empiriques*. Dans le modèle déductif-nomologique de l'explication scientifique hérité des positivistes, un concept a une valeur explicative (les concepts théoriques), si les lois qui le gouvernent impliquent « davantage » que simplement ce qui suit logiquement de ce que nous savons déjà (les observations déjà faites). Dans le cas de Ketland[63] au moins, c'est le modèle de l'explication auquel il se réfère explicitement. La non-conservativité de la théorie de la vérité donne ainsi lieu à une comparaison entre le statut épistémologique du concept de vérité gouverné par les lois tarskiennes et celui du concept de champ magnétique gouverné par les équations de Maxwell en physique. La volonté de vouloir maintenir une interprétation déflationniste de la vérité malgré le phénomène de non-conservativité est comparée à l'entêtement instrumentaliste devant le succès de la théorie des champs. Cet entêtement, qui lui paraît méthodologiquement douteux, est illustré de la façon suivante :

> A l'origine, nous n'avions aucune idée de ce que la situation serait, mais en utilisant la théorie abstraite, nous avons fait une prédiction qui a par la suite été vérifiée. Ceci accroît notre confiance dans la correction des lois abstraites (les équations différentielles reliant le champ magnétique B au courant I) utilisées pour obtenir la prédiction. Il serait tout à fait *ad hoc* de répudier l'explication hautement fructueuse donnée par la loi abstraite et de proposer à la place l'« explication » la plus faible possible de la situation. (Ketland (2005), p.86-87)

Autrement dit, le raisonnement de Ketland semble être le suivant : si la théorie ordinaire (tarskienne) de la vérité est si féconde, alors 1. nous devons prendre au sérieux son rôle explicatif, et 2. chercher à montrer que l'on peut rendre compte des phénomènes, dont cette théorie rend compte par des moyens plus faibles, serait une manœuvre scientifiquement douteuse.[64]

[63] Ketland (2005), p.87.

[64] C'est ce qui disqualifie aux yeux de Ketland la tentative de Tennant (2002) de soustraire le déflationniste à l'argument de la conservativité. Ketland ne met pas en doute que d'autres « explications » de la cohérence, n'employant pas le concept de vérité, soient possibles, mais le sens

Il n'est pas nécessaire d'être en désaccord avec Ketland sur le caractère *ad hoc* des reconstructions instrumentalistes des explications scientifiques, ni avec l'idée que la fécondité de la théorie des champs donne une raison (parmi d'autres) de considérer le concept de champ comme dénotant une propriété naturelle robuste. Mais notre point est précisément que la comparaison entre théorie de la vérité et équations de Maxwell est trompeuse. En effet, ce que montre l'argument précédent, c'est que, malgré la non-conservativité, la théorie de la vérité ne nous aide pas à faire quelque chose qui ait une valeur épistémologique comparable à celle d'une authentique « prédiction » dans les sciences empiriques, prédiction dont la réalisation viendrait en retour justifier la théorie de départ. La théorie de la vérité n'a pas de ces conséquences dont on pourrait affirmer qu'elles ne sont pas déjà des conséquences logiques (formelles) de l'expression complète de ce qui nous était connu dès avant sa formulation. Ou encore, pour le dire en un mot, la théorie de la vérité n'est pas une théorie arithmétiquement *risquée*, au sens où, selon Popper, une bonne théorie scientifique doit être empiriquement risquée, et où la théorie des champs l'est. *Et c'est bien ainsi* : les axiomes récursifs de la théorie tarskienne ne sont pas des lois abstraites formulées à titre hypothétique pour rendre compte (expliquer) de certains phénomènes arithmético-syntaxiques, et qui se trouveraient ensuite recevoir une justification ou une réfutation par ses conséquences ; ce sont sont au contraire des vérités connues *a priori* avec une certaine évidence. La situation diffère donc du tout au tout non seulement de celle qui prévaut dans la formulation de la théorie des électrons et des champs magnétiques, et plus généralement avec les hypothèses théoriques des sciences empiriques mais aussi, pour aller plus loin, de la situation qui prévaut, selon Gödel, dans les mathématiques elles-mêmes (et donc dans les domaines de la connaissance *a priori*), lorsqu'il s'agit de formuler certaines hypothèses hautement abstraites, en particulier des hypothèses ensemblistes, dont nous apprécions la plausibilité en fonction des conséquences qu'elles nous permettent de dériver.[65]

À nouveau, on peut rendre compte du fait que l'on peut prouver la vérité d'un énoncé qui n'était pas prouvable lorsque l'on étend PA par les axiomes tarskiens pour la vérité formulés *via* le codage de la syntaxe dans l'arithmétique et que l'on étend le schéma d'induction, de la façon suivante. On peut commencer par remar-

qu'il y a à ne vouloir considérer que celles-là (sur ce point, il se distingue de Shapiro. Voir le chapitre suivant.). *Par ailleurs* Ketland doute que Tennant ait réellement fourni une telle explication.

[65] Nous aurons l'occasion de revenir sur cette dernière comparaison au chapitre suivant.

quer tout d'abord que les axiomes tarskiens pour la vérité ne sont que des biconditionnels, qui n'affirment par eux-mêmes la vérité d'aucun énoncé. De plus, les nouveaux axiomes d'induction de l'arithmétique étendue n'affirment pas non plus par eux-mêmes la vérité d'aucun énoncé, tous les nouveaux axiomes ayant la forme de conditionnels (*si* $\phi(0)$ etc, *alors* $\forall x\phi(x)$). Pourquoi alors, avec ces nouveaux axiomes, on peut maintenant *prouver* que certains énoncés sont vrais (et donc ces énoncés eux-mêmes), qui n'étaient pas prouvables auparavant. Intuitivement, ce qui se passe d'un point de vue logique est la chose suivante. Les axiomes d'induction affirment chacun d'une propriété (en extension) que si elle est vraie de 0 et héréditaire alors elle est vraie de tous les entiers. Plus le langage ambiant permet d'exprimer de propriétés différentes, plus ces axiomes contraignent conjointement l'extension possible de la notion d'entier. A la limite, lorsque l'on peut parler de toutes les propriétés (en extension) dans le langage (comme en second ordre), alors on a identifié les entiers comme le plus petit ensemble contenant 0 et clos par succession. Dans le cas où il n'est pas possible de quantifier sur toutes les propriétés dans le langage, notre caractérisation n'est que partielle. En outre, elle se trouve renforcée chaque fois que nous introduisons dans le langage, d'une façon ou d'une autre, un nouveau moyen expressif. C'est ce qu'il se passe quand on introduit la notion de vérité par les axiomes tarskiens. Ces axiomes n'affirment la vérité d'aucun énoncé, mais le seul fait d'introduire ainsi dans le langage par une définition inductive axiomatique un prédicat qui n'est pas définissable explicitement, enrichit les moyens expressifs à notre disposition. Or notre compréhension des entiers est telle qu'en formulant le schéma d'induction nous entendions nous engager non pas seulement sur l'affirmation que l'ensemble des entiers est le plus petit ensemble clos par succession *parmi* les ensembles définissables arithmétiquement. Au contraire notre engagement était ouvert : c'est le plus petit, non seulement parmi ceux exprimables dans ce langage-ci, mais également parmi ceux qu'il sera peut-être possible d'exprimer à l'avenir. Vue sous cet angle, la non-conservativité des extensions aléthiques est compatible avec le thème déflationniste : le prédicat de vérité est un moyen d'expression qui a permis d'expliciter certains engagements que nous avions pris implicitement en formulant notre théorie de base (avec un schéma *ouvert*).[66]

[66] Un lecteur pourra s'étonner : après tout le prédicat introduit par les équivalences-T seulement, lui non plus n'est pas définissable dans le langage de l'arithmétique ; pourtant, les équivalences-T étendent conservativement PA. Le problème est que les équivalences-T, on l'a vu, ne sont pas suffisantes dans PA (et dans tout système sans règle "infinitaire") pour prouver la moindre généralisation contenant "vrai". Or pour que la nouvelle propriété puisse « faire du travail » dans le

3.3.4.1 Quelqu'un a-t-il changé de sujet ?

Pour résumer, il est possible d'admettre que savoir qu'une théorie interprétée A est vraie soit suffisant pour savoir que A est cohérente. Néanmoins, connaître la vérité de A ne donnera jamais aucune nouvelle connaissance des faits relevant du domaine d'investigation qui est celui de la théorie A, à moins que nous n'ayons su depuis le début que A était d'une manière ou d'une autre une formulation défectueuse de notre connaissance réelle de son domaine, et n'ayons eu de plus le moyen de reconnaître quelques extensions strictes de A comme étant correctes relativement à cette connaissance. Autrement, comment pourrions-nous être certains que les conclusions que nous tirons sont sûres ? Ou, pour faire écho à des réflexions que Shapiro a faites ailleurs[67] : si ces conditions n'étaient pas réunies, comment pourrions-nous être certains que nous n'avons pas changé de sujet ? que nous ne sommes pas subrepticement passés d'un discours sur les n*a*mbres à un discours sur les nombres ?

Le Problème de la frontière était celui de donner une explication de la distinction entre notion explicative et notion purement expressive. Il est naturel de ramener ce problème à un problème de disctinction entre des *énoncés* ayant une valeur explicative et les autres, la question étant alors de savoir de quel côté de la frontière tombent les énoncés qui permettent de rendre compte des usages jugés essentiels du concept de vérité. La Thèse de la Réflexivité impliquait que la théorie minimale de la vérité ne permet pas de rendre compte de ces usages, tandis que la théorie tarskienne le peut, et la question était alors de savoir si les axiomes de la théorie tarskienne sont ou non des principes explicatifs. Shapiro et Ketland prennent position sur le Problème de la frontière en déclarant explicatifs relativement à une

schéma d'induction, il faut pouvoir être en mesure de prouver des généralités concernant cette propriété, en particulier des énoncés de la forme "$\forall x(\phi(x) \rightarrow \phi(x'))$", où ϕ contient le nouveau prédicat. C'est pourquoi les équivalences-T, bien qu'elles introduisent en effet un prédicat inéliminable, ne renforcent pas les axiomes de PA.

[67] Shapiro (1998*a*). Pages 618 et suivantes, Shapiro reconnaît lui-même qu'il y a probablement quelque chose à dire en faveur de la thèse que le prédicat de vérité est logique. Il ajoute :

> Dans un autre article [Shapiro (1998*b*)], je soutiens que l'échec de la conservativité de théories telles que A" [il s'agit de A augmentée de principes aléthiques] parle contre ce qu'on appelle les théories déflationnistes de la vérité. Comment cette notion peut-elle être « fine » ou « sans substance » si l'on peut utiliser cette notion pour obtenir de nouvelles connaissances (p.ex. à propos des entiers naturels) ? Néanmoins, il y a une intuition naturelle selon laquelle en ajoutant un prédicat de vérité à A, et en adoptant A'', l'on a pas vraiment changé de sujet.(Shapiro (1998*a*), p.619)

théorie A les ensembles d'énoncé qui étendent non-conservativement A, et pensent ainsi pouvoir réfuter le déflationnisme. Mais d'une part, il n'est pas clair qu'il n'y ait pas une autre façon de tracer la frontière qui permettrait de résoudre positivement le Problème de la stabilité en respectant les intuitions déflationnistes. D'autre part, inversement, il n'est pas certain que la disctinction par la conservativité rende justice à ces intuitions. C'est ce que nous avons suggéré en soulignant l'existence de phénomènes de non-conservativité des lois logiques. Surtout, dans le cas des extensions aléthiques, l'observation d'un phénomène de non-conservativité sur la théorie de base dépend essentiellement du caractère schématique de cette théorie de base, et il est alors difficile de tirer des conclusions définitive, *quand bien même le critère de conservativité serait accepté* : car dans ces conditions, le gain en pouvoir explicatif résulte conjointement des axiomes aléthiques et des nouveaux axiomes de la théorie de base, et il semble impossible de trancher sur cette seule base du caractère explicatif ou non des axiomes pour la vérité.

3.4 Sans issue

Dans la section précédente, je suis resté assez vague sur la définition exacte du cadre formel de travail à partir duquel je tirais mes conclusions, j'ai seulement noté que, suivant Shapiro et Ketland, je considérais une théorie de base contenant sa propre syntaxe, éventuellement de façon codée. J'ai noté en passant que je m'autorisais ce flou, parce que la validité de mes remarques n'était pas sensible à la définition exacte de ce cadre de travail. Je veux maintenant étayer cette dernière affirmation en envisageant différentes possibilités. Je désignerai par PA l'arithmétique de Peano en premier en ordre et par Q, la théorie arithmétique plus faible obtenue à partir de PA en supprimant le schéma d'induction.[68] Je noterai $Sat_{\mathcal{L}}$ la théorie axiomatique tarskienne de la satisfaction (et de la vérité) pour le langage $\mathcal{L}$. Pour toute théorie T, je désignerai par Syn_T une théorie standard du premier ordre dans laquelle est formulée la syntaxe de T [69], avec son vocabulaire et ses axiomes spécifiques.

[68] Voir chapitre 3, section 2.2 pour une définition.
[69] Conçue selon le modèle présenté au chapitre 2.

3.4.1 Théorie de base : PA, extension aléthique : $PA + Sat_{\mathcal{L}_{\mathcal{PA}}}$

Pour commencer, prenons pour théorie de base PA et pour extension aléthique $PA + Sat_{\mathcal{L}_{\mathcal{PA}}}$, la syntaxe étant codée dans le vocabulaire de l'arithmétique. Si l'on accepte dans les formules d'induction le vocabulaire de Sat, alors nous devons admettre que notre connaissance de l'arithmétique n'était pas totalement explicitée dans la théorie de base.[70] Si, pour ce qui est de l'arithmétique, nous nous en tenons à PA dans la métathéorie, alors le vocabulaire aléthique n'est pas pas autorisé à figurer dans les formules d'induction de la métathéorie, la cohérence ne peut pas être prouvée, et en fait aucun phénomène de non-conservativité n'est observé.[71]

3.4.2 PA, $PA + Syn_{PA} + Sat_{\mathcal{L}_{\mathcal{PA}}}$

Supposons maintenant que, notre théorie de base étant toujours l'arithmétique en premier ordre PA, notre métathéorie est constituée de PA, la syntaxe de PA, Syn_{PA}, formulée en premier ordre et dans un vocabulaire spécifique, et des axiomes gouvernant la satisfaction ($Sat_{\mathcal{L}_{\mathcal{PA}}}$). Il y a une formule du *vocabulaire de* Syn_{PA} qui affirme la cohérence de PA, notons-la COH_{PA}[72], en majuscules. Un principe de preuve important de la théorie Syn_{PA} est le « schéma d'induction »[73], et si nous supposons que les instances de ce schéma d'induction peuvent contenir le vocabulaire de $Sat_{\mathcal{L}_{\mathcal{PA}}}$, alors cet énoncé, COH_{PA}, est prouvable dans l'extension aléthique. *Mais nous n'avons pas encore de cas de non-conservativité parce que l'énoncé de la cohérence en question n'est pas un énoncé du vocabulaire de* PA. Pour pouvoir dériver une contrepartie de cet énoncé dans le langage de PA il faut poser certains principes de correspondance (de « traduction ») entre la syntaxe Syn_{PA} et PA. Le problème est que les principes de correspondance nécessaires ne sont justifiés que si l'on accepte que les notions aléthiques peuvent apparaître également dans l'induction arithmétique. Nous n'avons aucune raison d'accepter ces principes de correspondance si tout ce que nous savons de l'arithmétique est couché dans PA.

[70] Et ce que montre la non-conservativité est que cette explicitation n'est pas sans effet.
[71] Je renvoie à nouveau à Halbach (1999*b*) pour une démonstration.
[72] Cette formule affirme qu'il n'y a aucune démonstration de $0 = 1$ dans PA.
[73] Voir chapitre 3, section 2.2.1.

3.4.3 Syn_{PA}, $Syn_{Syn} + Sat_{\mathcal{L}_{Syn}}$

Le problème ne peut pas être contourné en prenant pour théorie de base une théorie se présentant elle-même comme une théorie syntaxique, formulée dans un vocabulaire dédié. Pour le voir, on pourrait prendre pour théorie de base la syntaxe de PA elle-même.[74] L'extension aléthique serait constituée de la théorie de la syntaxe de la théorie de base et de notre théorie de la satisfaction Sat. Mais nous avons toujours le même problème : pour prouver l'énoncé de la cohérence exprimé dans le vocabulaire de Syn il nous faut faire appel, dans l'extension aléthique, à une théorie syntaxique strictement plus forte que celle de Syn_{PA}, et utiliser un principe d'induction où il est permis au vocabulaire de Sat de figurer !

Il semble donc que, du seul fait qu'en permettant au vocabulaire aléthique d'apparaître dans le principe d'induction il résulte en un gain de force logique, on ne peut pas conclure directement que les notions aléthiques sont « substantielles » en un sens pertinent. Il est naturel alors de se demander si la question du rôle de la vérité ne pourrait pas être tranchée en nous tournant vers un cadre logique dans lequel le *principe complet* d'induction arithmétique que nous semblons accepter quand nous acceptons PA pourrait être formulé ou, pour le dire autrement, un cadre dans lequel les principes spécifiques qui gouvernent notre compréhension de ce que sont les entiers naturels pourraient être totalement explicités. Si une telle théorie existe, l'arithmétique du second ordre paraît être un bon candidat.

3.4.4 Q^+, $Q^+ + Sat_{\mathcal{L}_{Q^+}}$

Il y a encore un choix de configuration des théories, dont j'ai déjà dit un mot en passant, mais sur lequel je voudrais revenir à présent en conclusion de cette section. Cette configuration m'intéresse parce que l'*impossibilité* d'y appliquer l'argument de la conservativité, du moins de la façon dont Shapiro et Ketland l'ont présenté, cette impossibilité me frappe comme une conséquence indésirable de leur argument. Considérons comme théorie de base une arithmétique strictement plus faible que PA, sans axiome d'induction, mais néanmoins encore suffisamment forte pour se qualifier comme « capable d'exprimer sa propre syntaxe ».[75] Prenant pour théorie-

[74] Ou envisager la possibilité d'une théorie qui ne soit autre chose que sa propre syntaxe. Je ne connais pas de tentative pour formuler une telle théorie, mais je ne vois rien qui en interdise en principe la possibilité.

[75] Toutes les fonctions récursives y sont fortement représentables, comme dans Q.

objet une théorie de ce type, V. Halbach a montré que les axiomes de Tarski pour la vérité étendaient conservativement cette théorie.[76] Mais puisque la cohérence de cette théorie est « exprimable » par un énoncé de son langage, et que cet énoncé est indépendant de la théorie, quelle conclusion devrions-nous tirer de ce fait dans la perspective de l'argument de la conservativité ? Il me semble que, puisque cette théorie est une arithmétique plus faible que PA, nous devrions pouvoir dire que son pouvoir explicatif est plus faible. Mais si nous interprétons la conservativité comme Shapiro et Ketland semblent nous le demander, nous sommes confrontés à cette conséquence que les axiomes tarskiens pour la vérité sont des prinicipes explicatifs arithmétiques lorsqu'ils étendent PA, mais pas lorsqu'ils étendent l'arithmétique *plus faible*, ce qui semble absurde. La réponse de repli naturelle est de dire que c'est l'ajout conjoint des axiomes tarskiens *et* de principes syntaxiques (éventuellement de façon codée) suffisamment forts qui engendre de nouveaux prinicipes explicatifs. En effet, en ajoutant par exemple à notre arithmétique faible les axiomes de PA et les axiomes traskiens, le résultat obtenu est bien non-conservatif (non seulement sur l'arithmétique faible, mais également sur PA). Mais devons alors à nouveau faire face à la même difficulté d'identifier les causes de cet effet : s'il n'est pas possible d'identifier le rôle spécifique des axiomes aléthiques dans ce processus, en tant que distinct du rôle des principes arithmétiques, la non-conservativité ne suffit pas pour soutenir la thèse d'un concept de vérité dont la théorie aurait un rôle explicatif substantiel.

On aura beau tourner les choses dans tous les sens, on se retrouvera toujours confronté à cette même observation : si pour observer un phénomène de non-conservativité sur A il faut reconnaître que certaines formulations de nos principes de preuves dans A étaient incomplètes et en accepter une reformulation plus forte, alors il n'y a aucune raison de conclure que les vérités aléthiques jouent le moindre rôle explicatif. Au contraire, ce résultat corrobore l'idée que le prédicat de vérité sert à *expliciter ce qui était implicite* dans notre acceptation de la théorie de base. Le simple fait de reconnaître la vérité des clauses tarskiennes pour la vérité ou la satisfaction ne permet nullement, à soi seul, d'expliquer de « nouvelles » vérités *du langage de la théorie de base*, à la façon dont dont les « lois théoriques » permettent d'expliquer des phénomènes observables, et les lois des éléments mathématiques idéaux d'expliquer des « phénomènes » mathématiques « réels ».

[76] Voir à nouveau Halbach (1999*b*), p.359, Th.3.1. L'absence schéma d'induction est cruciale ici.

3.5 Conclusion sur la thèse de la conservativité

En accordant provisoirement la Thèse de la Réflexion, j'ai accordé que la théorie minimale n'est pas une théorie adéquate* de la vérité. Et le problème soulevé était alors : à supposer que la théorie récursive de la vérité soit adéquate, peut-on en conclure que la notion de vérité a un pouvoir explicatif ? Shapiro et Ketland ont attiré notre attention sur le fait que la théorie tarskienne n'étend pas conservativement certaines théories de base. S'inscrivant dans une longue tradition philosophique d'interprétation des phénomènes de non-conservativité où ces derniers sont compris comme l'indice du caractère irréductible et indispensable des notions « théoriques », et reprenant à leur compte un modèle néo-positiviste de l'explication, ils concluent que les lois tarskiennes de la vérité permettent un emploi explicatif substantiel de la notion de vérité. Je crois que Shapiro et Ketland manquent une distinction essentielle entre, d'une part, le fait qu'une notion ou certains principes ont une utilité épistémologique, et qu'ils puissent même être indispensables à ce titre et, par suite, irréductibles à d'autres, et d'autre part le fait que cette notion ou ces principes aient un emploi proprement explicatif. Le déflationniste, jusqu'à preuve du contraire, n'a aucune difficulté avec l'idée vague du caractère épistémologiquement nécessaire de la notion de vérité, puisqu'il juge en effet que cette dernière est un indispensable moyen d'expression.

Dans ce chapitre, j'ai essayé de montrer qu'il était possible de concilier non-conservativité des principes aléthiques tarskiens sur certaines théories et la thèse déflationniste selon laquelle la vérité n'est qu'un outil d'expression, même en accordant momentanément la thèse de Réflexion. Ma conclusion est que la thèse de la conservativité, non seulement ne semble pas de prime abord faite pour rendre compte de la distinction recherchée par le déflationniste, mais encore échoue à le faire par ses propres lumières. Plus spécifiquement, j'ai suggéré (et non prouvé) que la factorisation suivante du processus conduisant à l'observation de phénomène de non-conservativité pouvait être éclairante :

1. Il est partie intégrante de notre compréhension de l'arithmétique que le schéma d'induction soit compris comme une règle. On ne peut pas comprendre complètement ce que sont les entiers naturels, sans être en position de reconnaître certaines théories de l'arithmétique plus forte que PA comme correctes.
2. Il y a de l'implicite et de l'explicite dans la représentation logique de nos connaissances lorsque nous formulons des théories où les schémas sont com-

pris comme règles. Les connaissances laissées implicites ont des conséquences logiques latentes, mais ces conséquences ne sont pas déployées dans le langage ambiant. Autrement dit, ce qui suit logiquement d'une théorie dont certains schémas sont compris comme règles n'épuise pas ce qui suit logiquement de ce qui était compris comme le contenu de la théorie.

3. Les principes de preuves ouverts deviennent naturellement plus forts quand leur environnement expressif s'enrichit. Dans ce processus, certains principes de preuves dictés par notre compréhension du sujet de la théorie de départ sont explicités.
4. C'est à ce point que la vérité entre en jeu : la vérité est un outil expressif, et elle permet d'expliciter certains engagements théoriques que nous avons relativement à la théorie de départ.
5. Donc la raison véritable de la non-conservativité des théories de la vérité sur certaines théories de bases, lorqu'elle est avérée, est à chercher dans le caractère *ouvert* de notre compréhension de la théorie de base (le caractère ouvert des principes de preuves arithmétiques) et le pouvoir expressif de la vérité.

Ce tableau de la situation semble raisonnable[77] et constitue une réponse attractive à l'argument de la conservativité.

Ce chapitre nous laisse donc avec des questions dans les deux directions, du problème de la frontière et du problème de l'usage. Il y a d'une part (1) la question de savoir si les usages du prédicat de vérité permis par la théorie tarskienne sont des usages explicatifs ou purement *expressifs*. Et cette question suppose que l'on a répondu à cette autre : (2) comment distinguer le caractère épistémologique spécifique des principes explicatifs ? Dans l'autre direction subsistent les questions relatives aux critères d'adéquation des théories de la vérité, au-delà du critère tarskien nous garantissant que le prédicat dont nous faisons la théorie est bien un prédicat de vérité. (3) De quels usages du prédicat de vérité une théorie de la vérité doit-elle rendre compte ? (4) La thèse de la Réflexion est-elle correcte ? Et à supposer donnée la réponse à ces questions, (5) quelle théorie de la vérité rend le mieux

[77] Field (1999) a lui aussi brièvement contesté l'argument de la conservativité en remarquant qu'on observait un phénomène de non-conservativité sur PA à la seule condition d'étendre les axiomes d'induction. Puisque l'induction, affirme Hartry Field, est un principe mathématique, et non aléthique, il conclut simplement que le phénomène de non-conservativité est une indication de la « substance » du principe d'induction, non de la vérité. Nous sommes plus prudents dans notre conclusion.

compte de *ces* usages ? Enfin, à l'intersection de ces deux familles de questions, le Problème de la stabilité demeure, de savoir si une théorie est possible qui rendent compte des usages essentiels de la notion de vérité et soit compatible avec l'idée que le prédicat de vérité n'a pas d'usage explicatif.

Au chapitre 5, je reviens aux théories de la vérité et au rôle du concept de vérité dans les explications. Dans ce chapitre historique, je donne quelques coups de sonde dans la littérature de la philosophie des mathématiques du vingtième siècle pour essayer d'éclairer la façon dont le rôle de la vérité dans les preuves a été compris et poser les bases d'une réflexion sur le statut épistémologique des preuves par la vérité. C'est seulement au chapitre 7, à partir des éléments de réflexion réunis jusque-là, que je présenterai ma tentative de réponse aux questions (1)-(6).

Chapitre 4

Preuves par la vérité et épistémologie des mathématiques

Dans ce chapitre, je prolonge la réflexion commencée au chapitre précédent sur la portée épistémologique des « preuves de cohérence par la vérité » en présentant, en opposant aux conclusions de Shapiro et Ketland d'autres éléments d'interprétation philosophique classiques des phénomènes d'incomplétude. L'intérêt de ce travail pour la suite est double. D'une part, il va me permettre de montrer que la thèse du caractère épistémologiquement modeste du rôle de la vérité procède d'une *intuition partagée*, d'autre part il fournit la clé d'une nouvelle distinction épistémologique qui permettra de donner à cette thèse une signification plus précise.

Je vais considérer trois réflexions philosophiques auxquelles ont donné lieu les théorèmes d'incomplétude et plus particulièrement le caractère épistémologiquement « spécial » que semblent revêtir ces énoncés indécidables simples dont Gödel a donné la méthode de construction systématique. Dans la première partie, je fais quelques remarques sur la position de Gödel lui-même. Ces remarques doivent permettre de prévenir certains malentendus et de présenter brièvement une première fois une distinction entre « extension intrinsèque » et « extension extrinsèque » d'une théorie, idée qui jouera un rôle important au chapitre 7. Dans la deuxième partie, je discute, sans volonté d'exhaustivité, les programmes de Feferman et Myhill pour élaborer une notion de conséquence « réflexive » d'une théorie qui permette de rendre

compte formellement de la relation de consécution qui semble exister entre la théorie mathématique A (supposée coder sa propre syntaxe), et ses énoncés gödéliens. Dans la troisième partie, je m'arrête à un programme dual des précédents. Il s'agit du programme d'Isaacson visant à montrer qu'en un certain sens *épistémologique* de la notion de complétude, l'arithmétique de Peano en premier ordre est complète. J'essaie de montrer que cette thèse, si elle est vraie, malgré des apparences défavorables, est néanmoins compatible avec l'idée que notre emploi des axiomes aléthiques tarskiens n'est pas un emploi explicatif. J'avertis que mon ambition n'est pas de présenter une discussion détaillée des thèses ou des programmes philosophiques portés par ces auteurs et, au contraire, je tenterai de ne présenter, à chaque fois, que ce qui a trait directement à notre discussion du rôle épistémologique du concept de vérité, en réduisant au minimum compatible avec l'intelligibilité du propos la présentation du contexte théorique dans lequel elles apparaissent.

4.1 Remarques sur Gödel et la vérité

Nous avons discuté du rôle du concept de vérité dans les preuves principalement en relation avec les preuves de cohérence des systèmes formalisés et avec à l'esprit les résultats d'incomplétude de Gödel. Il est donc intéressant de chercher dans les écrits de Gödel lui-même une réponse à la question qui nous intéresse. Mais nous risquons d'être déçus. La raison en est que les termes dans lesquels nous avons posé notre question -préciser la nature du rôle épistémologique du concept de vérité- est étrangère aux problématiques philosohiques de Gödel lui-même. Ces problématiques étaient certes porteuses de questions en apparence assez proches de celle qui nous intéresse, mais apparemment seulement - la cohérence est-elle un critère suffisant d'existence en mathématique ? Les mathématiques se réduisent-elles à un simple jeu de manipulation de symboles ? Plus généralement : quelle est la nature et comment connaissons-nous les vérités mathématiques ? Au contraire *notre* problème philosophique est essentiellement *post-tarskien*[1] : nous tenons pour acquise la légitimité du concept de vérité, le fait que les énoncés auxquels ils s'appliquent *ont* un contenu, ou encore la distinction entre vérité et prouvabilité, alors que les réflexions de Gödel tendent à les asseoir. Les risques d'anachronismes étant bien réels, il est parfois difficile de dire si, derrière des remarques de Gödel où apparaît le concept de vérité, il y a plus que l'apparence d'un intérêt pour la question qui nous occuppe. Sans prétendre

[1] Ou peut-être faudrait-il aussi bien dire « post-gödélien ».

offrir l'étude historique que la question mériterait, nous voudrions à présent donner quelques indications d'une réponse négative à la question.

La relation de Gödel à la notion de vérité en 1931 est marquée par la prudence, comme y a insisté Feferman (1989) : Gödel, dans son article de 1931, avait souhaité éviter de parler en terme de « vérité », quoiqu'il ait sans doute eu en sa possession la définition rigoureuse que Tarski en proposera par la suite.[2] Cette prudence était motivée par la crainte des résistances que pourrait susciter l'emploi du concept de vérité dans un contexte idéologique qui était largement défavorable à une conception « réaliste » du discours mathématique, et où même le caractère contentuel d'une large part du discours mathématique pouvait être mise en doute. Mais cette prudence, comme le rappelle Feferman[3], n'en allait pas moins de pair avec, d'une part, une conviction philosophique profonde que des programmes réductionnistes du type « formalisme » étaient fondamentalement erronés et, d'autre part qu'une conception philosophique correcte avait constitué une condition essentielle de la découverte de son propre résultat d'incomplétude, comme en témoigne un passage d'une lettre de Gödel cité dans Wang (1974) (p.9 et reproduit également dans Feferman (1989), p.106-107) :

> J'ajouterais que ma conception objectiviste des mathématiques et des métamathématiques en général, et du raisonnement transfini en particulier, a été également fondamental pour mon autre œuvre de logique. Comment en effet quelqu'un pourrait-il penser *exprimer* les métamathématiques *dans* les systèmes mathématiques eux-mêmes, si ces derniers sont considérés consister en des symboles dénués de signification qui acquièrent un substitut de signification seulement *à travers* les métamathématiques. ... il doit être remarqué que le principe heuristique de ma construction de propositions de la théorie des nombres indécidables dans les systèmes formels des mathématiques est le concept hautement transfini de « vérité mathématique objective » en tant qu'*opposé* à celui de « démontrabilité » [parenthèse omise] avec lequel il était généralement confondu avant mon travail et celui de Tarski.(Lettre de Gödel à Wang, cité dans Feferman (1989), p.106-107)

Si ce passage témoigne clairement d'une préoccupation du rôle de l'ontologie pour une épistémologie correcte plutôt que du rôle du concept de vérité lui-même, les

[2] Voir Feferman (1989), p.106
[3] Feferman, *op. cit.*

choses sont moins claires lorsque l'on se tourne vers d'autres écrits, publiés ou non, comme cet extrait d'une lettre de Gödel à A.W. Burks, citée dans Neuman (1966) (p.55-56) et reprise dans Feferman (1989), p.104-105 :

> Je crois que celui de mes théorèmes auquel réfère von Neumann n'est pas celui sur l'existence de propositions indécidables ou celui sur la longueur des preuves mais plutôt le fait qu'une description épistémologique complète d'un langage A ne peut être donnée dans le même langage A, parce que le concept de vérité d'énoncés de A ne peut être défini dans A. C'est ce théorème qui est la véritable raison de l'existence de propositions indécidables dans les systèmes formels contenant l'arithmétique. Je ne l'ai toutefois pas formulé explicitement dans mon papier de 1931 mais seulement dans mes Conférences de Princeton en 1934. Le même théorème fut prouvé par Tarski dans son article sur le concept de vérité publié en 1933 [...]

Est-ce dire que pour Gödel, la vérité était un concept « substantiel » en un sens qui serait déplaisant pour un déflationniste ? C'est à ce passage, et d'autres du même genre[4], que je pense lorsque je dis qu'ils ont une apparence trompeuse. Je voudrais maintenant mettre en regard d'autres considérations de Gödel pour étayer cette thèse.

Dans d'autres passages, il est clair que les considérations de Gödel visent avant tout les programmes « formalistes », l'idée que les contenus mathématiques n'iraient pas au-delà des contenus que Hilbert appellerait « réels ». Replacé dans ce contexte, le passage suivant qui ouvre l'article « On an extension of finitary mathematics which has not yet been used » [5] illustre plus clairement les enjeux philosophiques qui intéressaient Gödel au premier chef :

> P. Bernays a indiqué [note omise] en plusieurs occasions qu'étant donné le fait que la cohérence d'un système formel ne peut être prouvée par aucune procédure de déduction disponible dans le système lui-même, il

[4] Voir par exemple les remarques p.108-109 dans Feferman (1989). Relativement à la note 48^a de Gödel (1986), Feferman écrit :

> Since Gödel did not write part II to (1931) and never commented further on footnote 48^a, we do not know whether he saw it necessary to give a set theoretic analysis of the concept of truth in order to justify his claim. (The significance of footnote 48^a in this respect was brought to my attention by S. Kripke.) (Feferman (1989), p.109)

[5] Gödel (1990)

> est nécessaire d'aller au-delà du cadre de travail des mathématiques finitaires au sens de Hilbert pour prouver la cohérence des mathématiques classiques ou même de la théorie des nombres classique. Puisque les mathématiques finitaires sont définies [note omise] comme les mathématiques de l'*intuition concrète*, cela semble impliquer que des *concepts abstraits* sont nécessaires pour la preuve de la cohérence de la théorie des nombres. Une extension du finitisme par de tels concepts a été explicitement suggérée par Bernays (1935 : 69 p.271). Par concepts abstraits, dans ce contexte, il faut entendre des concepts qui sont essentiellement de second ordre ou d'ordre supérieur, i.e. qui n'ont pas pour contenu des propriétés ou des relations d'*objets concrets* (comme des combinaison de symboles), mais plutôt de *structures de pensée* ou de *contenus de pensée* (par exemple, des preuves, des propositions douées de sens, et ainsi de suite), où dans les preuves des propositions à propos de ces objets mentaux des insights sont nécessaires qui ne peuvent pas être dérivés d'une réflexion sur les propriétés (spatio-temporelles) des symboles les représentant, mais plutôt d'une réflexion sur les *significations* en jeu. (Gödel (1990), p.271)

Il est important de garder à l'esprit que les résultats qui suivent de l'application du concept de vérité au discours mathématique n'établissent pas le caractère contentuel du discours mathématique, mais le *suppose*.[6] Les conséquences philosophiques que Gödel tire des théorèmes d'incomplétude ont ainsi trait davantage au contenu du discours mathématique, ou à la nature des objets mathématiques, qu'au concept de vérité lui-même (sauf pour le fait qu'il faut distinguer vérité et prouvabilité).

Par ailleurs, comme nous l'avons déjà indiqué en passant, nous devons à Gödel l'idée de la distinction entre extension intrinsèque et extrinsèque d'un système d'axiomes, une distinction introduite dans le cadre de sa discussion sur les extensions possibles de la théorie des ensembles dans « What is Cantor's continuum problem » :

> D'autres axiomes d'infinité ont d'abord été formulés par P. Mahlo [note omise]. Ces axiomes montrent clairement que le système axiomatique de la théorie des ensembles tel qu'il est utilisé aujourd'hui est incomplet, mais également qu'il peut être complété [*supplemented*] sans arbitraire

[6] Tarski était clair sur ce point, voir chap. 3.

> par de nouveaux axiomes qui déploient [*unfold*] seulement le contenu du concept d'ensemble expliqué plus haut [N.d.T : le concept itératif]. (Gödel (1964), p.476-477)

et, un peu plus bas :

> Deuxièmement, néanmoins, même en laissant de côté la nécessité intrinsèque d'un nouvel axiome[7], et même dans le cas où il n'a aucune nécessité intrinsèque du tout, une décision probable à propos de sa vérité est possible aussi d'une autre façon, nommément, inductivement en étudiant son « succès ». (Gödel (1964), p.476-477)

A s'en tenir pour l'instant à la théorie des ensembles, on peut donc distinguer au moins deux types d'incomplétude, relatifs aux deux types d'énoncés indécidables dans ZFC (disons) :

1. Les énoncés indécidables dans ZFC mais immédiatement reconnus comme vrais en réfléchissant sur la preuve de leur indécidabilité : il s'agit en particulier des énoncés indécidables construits spécialement pour la preuve des théorèmes d'incomplétude.
2. Les énoncés indécidables dans ZFC mais décidables dans une extension intrinsèque (comme certains axiomes de grands cardinaux).
3. Les énoncés indécidables dans ZFC mais dont la vérité peut être tranchée par des considérations extrinsèques (on peut penser ici à des axiomes comme l'axiome de détermination)

Mais si cette distinction a un sens[8], elle justifie alors que nous distinguions entre plusieurs types de phénomène de non-conservativité ! Une extension non-conservative de ZFC dont la non-conservativité se borne, pour ainsi dire, à décider les énoncés de type (1) doit être distinguée d'une extension décidant seulement des énoncés de type (3).

Or quel est le statut des extensions aléthiques ? Considérons d'abord une réponse indirecte à cette question ouverte par les remarques de Gödel pour le cas d'une extension aléthique de ZFC. Gödel affirme que le concept itératif d'ensemble n'est pas

[7] N.d.T : entre autres, ceux auxquels il a été fait allusion dans la citation précédente

[8] Gödel aurait fait une distinction plus fine, distinguant par exemple les énoncés indécidables du type de ceux construits dans son article de 1931 des énoncés intrinsèquement justifiés. Les premiers sont décidables en réfléchissant simplement sur la preuve de leur indécidabilité, sans qu'il soit besoin de référer à nos concepts ensemblistes. Néanmoins nous laisserons ces distinctions de côté ici, n'en ayant pas besoin. VoirAtten et Kennedy (2009).

complètement déployé dans ZFC. Le même processus de construction d'ensembles à partir d'ensembles donnés peut être poursuivi plus loin, et en ajoutant à ZFC des axiomes garantissant l'existence de cardinaux inaccessibles, on n'a fait que déployer un peu plus de concept itératif d'ensembles. Cette nouvelle théorie étendue, ZFC^+, est formulée dans le même langage que ZFC. Or tous les énoncés du langage de la théorie des ensembles qui sont prouvables dans l'extension aléthique tarskienne de ZF le sont dans ZFC^+. On pourrait itérer la remarque : pour toute extension aléthique tarskienne $T(A)$ d'une théorie A du concept itératif d'ensemble, il existe une extension intrinsèque de A qui prouve tous les théorèmes ensemblistes de l'extension aléthique $T(A)$. Il est donc assez naturel de ne pas s'arrêter à la remarque de la non-conservativité de l'extension aléthique de ZF sur ZF, mais de chercher à *qualifier* ce phénomène non-conservativité. On observe alors que les extensions aléthiques ne nous disent pas davantage que les extensions intrinsèques.

Plus généralement, n'est-il pas naturel, comme nous l'avons fait nous-mêmes, d'interpréter les extensions aléthiques de théories contentuelles quelconques [9] comme des exemples d'extension *intrinsèques* de ces théories, c'est-à-dire dont la justification est fondée sur notre saisie du contenu conceptuel des notions apparaissant dans la théorie de base. Ainsi, des conclusions « déflationnistes » analogues à celles que nous venons de tirer relativement à la portée épistémologique des extensions aléthiques de ZFC vaudraient-elles aussi pour les extensions aléthiques de PA. Je ne sais pas si Gödel a jamais couché sur le papier une remarque de cet ordre, mais elle me semble pourtant en harmonie avec sa réflexion sur l'incomplétude. Peter Koellner par exemple, dans un article par ailleurs essentiellement consacré à la théorie des ensembles, illustre l'idée gödélienne d'axiomes intrinsèquement justifés ainsi :

> Voyons maintenant quelques exemples. Considérons d'abord la conception des entiers naturels qui sous-tend le système de l'arithmétique de Peano (PA). Cette conception des entiers naturels non seulement justifie l'induction mathématique pour le langage de PA mais pour toute extension du langage de PA qui a un sens. Par exemple si nous étendons le langage de PA en ajoutant le prédicat de vérité de Tarski et étendons les axiomes de PA en ajoutant les axiomes tarskiens de la vérité, alors, sur la base de notre conception des entiers naturels, nous sommes justifiés à accepter des instances de l'induction mathématique contenant le prédicat

[9] Supposées capables d'interpréter leur propre syntaxe

> de vérité. Dans le système qui en résulte on peut prouver $Con(PA)$. Ce processus peut alors être itéré. De plus, il y a d'autres exemples d'axiomes intrinsèquement justifiés sur la base de notre conception des entiers ; par exemple les principes de réflexion preuve-théoriques [note de l'A. : Pour davantage sur ce sujet voir FEFERMAN (1991) et les références qui s'y trouvent.]. Par contraste, la proposition Π^1_0, $Con(ZF + AD)$ [NdT : « AD » est l'acronyme de « Axiome de détermination »] n'est sans aucun doute pas intrinsèquement justifiée sur la base de notre conception des entiers naturels ; sa justification découle plutôt d'un réseau intriqué de théorèmes de théorie des ensembles contemporaine.(KOELLNER (2009), p.208)

Les extensions aléthiques d'une théorie en sont certainement des extensions intrinsèques au sens de Gödel. Néanmoins, nous ne voudrions pas laisser se consolider l'analogie entre l'extension par les axiomes de grands cardinaux et les extensions aléthiques. Il s'agissait pour nous d'introduire une idée centrale : que toutes les extensions d'une théorie ne sont pas sur un pied d'égalité épistémologique et que Gödel avait ouvert la voie à une réflexion qui chercherait à *qualifier* les phénomènes de conservativité. Il y a, à côté de la distinction *ex post* que nous pouvons faire entre différentes extensions en en comparant les conséquences, une différence plus fondamentale qui a trait à la nature des justifications que nous avons pour ces extensions. On peut voir la thèse selon laquelle certaines extensions d'une théorie sont justifiées par la saisie même du concept que cherche à articuler cette théorie comme une forme de « déflationnisme » relativement à ces extensions. Mais toutes les extensions intrinsèques d'une théorie n'ont pas le même statut épistémologique. Il y a certainement une distinction à faire entre les axiomes de grands cardinaux dont, selon Gödel, nous percevons directement la vérité en vertu de notre saisie du concept itératif d'ensemble et, d'un côté, les extensions aléthiques dont nous percevons la vérité *indépendamment même des concepts spécifiques que la théorie de départ vise à articuler* ou, d'un autre côté encore, les extensions d'une théorie que l'on obtient en y ajoutant simplement à titre de nouvel axiome un de ces énoncés Π_1 construits par Gödel pour ses preuves d'incomplétude (les énoncés G ou Coh dont il a déjà largement été question). Il faut donc naturellement distinguer *à l'intérieur* des extensions intrinsèques différents sous-genres, correspondant aux différents types d'incomplétude attachés à une théorie : aux extensions dont la vérité se comprend à la lecture de la preuve d'incomplétude gödélienne de ces théories correspondrait une

catégorie, aux extensions aléthiques une autre, à celles dont la vérité ne se perçoit plus spécifiquement qu'en vertu de notre saisie des concept de la théorie de base, une troisième, et pourquoi pas d'autres que révèlerait une analyse plus fine encore.

Nous ne chercherons pas ici à préciser plus en détail ce qui distingue les extensions aléthiques des autres type d'extensions intrinsèques[10], nous y reviendrons au chapitre 6. Il nous suffit ici de noter que les considérations précédentes nous paraissent de nature à étayer notre thèse sur la fonction (interne) du concept de vérité : l'utilisation du concept de vérité (*via* la théorie tarskienne) permet de déployer formellement une partie de ce que nous comprenons être le contenu de la théorie de base, indépendamment de notre saisie de la nature spécifique des concepts qui y figurent. Sous cette interprétation, la vérité n'a donc de rôle explicatif qu'en un sens vague, qui ne résiste pas à une distinction entre deux rôles épistémologiques distincts : celui de rôle expressif, permettant de développer un contenu laissé implicite à un stade antérieur du travail de formalisation d'un côté, et le rôle de principe explicatif authentiquement nouveau d'un autre côté. Une fois cette distinction faite, les remarques précédentes valent indication que le concept de vérité tombe dans la première catégorie et nous avons les moyens d'expliquer pourquoi l'extension aléthique de PA peut à bon droit être appelée une extension intrinsèque de PA, ou pourquoi l'on peut dire que cette extension est justifiée par notre saisie du contenu de PA. Etant dit que Gödel lui-même aurait très certainement accepté la thèse que les extensions aléthiques d'une théorie en constituent des extensions intrinsèques, ne faut-il donc pas conclure pour finir que, malgré des affirmations aux apparences rétrospectivement trompeuses, il n'y a pas de raison de penser que Gödel aurait eu à objecter à l'idée que la notion de vérité joue un rôle « modeste » dans les preuves de cohérence ?[11]

[10]Le lecteur trouvera dans Feferman (1991), §2, des informations précises relativement au pouvoir preuve-théorique exacte des extensions aléthiques tarskienne.

[11] Nous signalons en passant l'intérêt que pourrait avoir dans notre perspective une étude de l'importance que semble avoir prise pour Gödel dans les années soixante l'idée de « connaissance de la raison par elle-même » comme source du savoir mathématique. Comme le rapporte Mark van Atten, dans un brouillon de lettre au Prof. Tillich daté de Juin 1963 (Atten (2006)), Gödel écrit :

> J'ai dit que dans le raisonnement mathématique l'élément non calculatoire (i.e. intuitif) consiste en intuitions d'infinités de plus en plus élevées [higher and higher infinities]. C'est tout à fait vrai mais [note de l'éditeur rapportant l'existence d'un passage raturé ici] cette situation peut être analysée plus avant et il apparaît alors qu'ils résultent (comme il devient parfaitement clair lorsque les choses sont traitées dans le détail) d'une connaissance de la raison par elle-même de plus en plus profonde [pour être plus précis d'une connaissance rationnelle de plus en plus complète de l'*essence* de la raison (essence dont la faculté de connaissance de soi est elle-même

4.2 Ketland et le programme de Feferman

La remarque que nos engagements épistémologiques ne reçoivent pas leur expression complète à la fin du processus ordinaire de formalisation des théories, et en particulier si nos théories contiennent des schémas ouverts, cette remarque doit rappeler les réflexions fondationalistes de Feferman et d'autres (notamment Kreisel) nées des résultats d'incomplétude de Gödel. Notre distinction entre *implicite* et *explicite* elle-même est une réminiscence de l'usage qu'en a fait Feferman, s'inspirant lui-même de la distinction de Gödel. Il est par conséquent d'un intérêt particulier de noter que Ketland se réfère au travail de Feferman sur la réflexivité pour appuyer son argument. Dans cette section, je voudrais essayer de clarifier la situation.

Pour comprendre le travail de Feferman sur la vérité et la « réflexion » dans les années quatre vingt dix[12], il peut être utile de revenir brièvement sur ses travaux antérieurs. L'un des problèmes qui ont durablement inspiré son travail de logicien, pourrait être formulé grossièrement de la façon suivante : identifier ce qui est prédicativement acceptable étant donné la notion d'entier. Dans un premier temps, le problème est compris comme un problème de définissabilité : de quels *ensembles* d'entiers doit-on admettre l'existence étant donné que l'on accepte l'existence des entiers (la quantification sur l'ensemble des entiers) ? Ou : étant donné la notion d'entier, quels sont les ensembles d'entiers qui sont définissables sans présupposer de notions ou d'entités qui n'aient elles-mêmes été définies antérieurement. On peut donner assez simplement une idée informelle de la façon dont Feferman s'y prend. Donnons-nous un langage arithmétique du second-ordre. Au départ, puisque nous n'acceptons que les entiers, les variables du second ordre sont interpétées comme parcourant un domaine vide, le reste du langage recevant son interprétation attendue. Appelons le parcours initial des variables du second-ordre $\mathcal{R}_0$: $\mathcal{R}_0 = \emptyset$. Ceci étant dit, avec les formules de notre langage, interprété comme venons de l'indiquer, nous pouvons définir des ensembles d'entiers : ces ensembles qui sont l'extension de

> une part constituante)] [je crois que la raison calculatoire résulte également de la connaissance de soi de la raison mais non de la connaissance essentielle mais factuelle] Il me semble que c'est là une vérification (dans le champ des mathématiques) de certains tenants de la philosophie idéaliste.

Van Atten indique qu'une source probable d'inspiration de Gödel à cette époque se trouve dans les *Idées* de Husserl (en particulier sections 75 et 77-79). Pour une étude historique sur l'intérêt de Gödel pour les thèses du second Husserl dans les années soixante, voir Atten et Kennedy (2003).

[12]Le texte de référence ici est Feferman (1991).

formules du langage en question, par exemple, disons, l'ensemble des entiers pairs.[13] Appelons $\mathcal{R}_1$ l'ensemble de ces ensembles. Puisque ces ensembles sont définis sans quantification sur les ensembles d'entiers, leur existence n'est pas problématique, et nous pouvons les admettre au titre de valeurs possibles de variables du second-ordre. Nous réinterprétons donc maintenant nos variables du second-ordre comme parcourant, non pas $\mathcal{R}_0$, mais $\mathcal{R}_1$. Dans ce nouveau langage, on peut à nouveau considérer les ensembles d'entiers qui sont définissables par une formule de *ce* langage. On appelle $\mathcal{R}_2$ l'ensemble de ces ensembles, qui sont des valeurs admissibles de nos variables du second ordre, et l'on réinterprète nos variables du second-ordre comme parcourant $\mathcal{R}_2$. Et ainsi de suite. Cette construction donne naissance à une hiérarchie transfinie, que l'on peut définir de la façon suivante :

- $\mathcal{R}_0 = \emptyset$,
- $\mathcal{R}_{\alpha+1} = \{$ X : X est définissable par une formule du langage dont les variables du second ordre parcourent $\mathcal{R}_\alpha\}$,
- $\mathcal{R}_\lambda = \bigcup_{\beta<\lambda} \mathcal{R}_\beta$ pour λ limite.

Peut-on dire de tout ensemble d'entiers appartenant à $\mathcal{R}_\alpha$ pour un ordinal α, qu'il est définissable « prédicativement étant donné la notion d'entier » ? Non bien sûr, la définition de la hiérarchie des ensembles analytiques faisant un usage essentiel de la notion *d'ordinal*. La question est alors de savoir quels sont les ordinaux acceptables étant donné la notion d'entier, autrement dit les ordinaux « prédicatifs ». Ici plusieurs réponses ont été proposées. Une première proposition est d'appeler *prédicatifs* les ordinaux qui sont le type d'ordre de bons ordres *définissables* prédicativement. L'idée est alors qu'un ensemble d'entiers est prédicativement définissable s'il appartient à un $\mathcal{R}_\alpha$ pour un ordinal α *prédicatif*.[14] Cette caractérisation détermine l'ordinal ω_1^{ck}, le plus petit ordinal non récursif, comme le plus petit ordinal non prédicatif. Une condition plus stricte sera ensuite développée, fondée sur l'idée que les ordinaux prédicatifs ne sont pas tous les ordinaux définissables prédicativement, mais seulement ceux dont on *prouver* prédicativement qu'ils sont des ordinaux. Pour donner un sens à cette idée de « prédicativement prouvable », Feferman a développé une hiérarchie, non pas de langages, mais de systèmes de

[13] Définissable par la formule « $\exists y(x = 2.y)$ ».

[14] La notion de bon ordre est définissable par une formule du second ordre. C'est un ordre (linéaire, total) qui satisfait de plus le principe que tous ses « sous-ensembles » ont un plus petit élément :

$$A \text{ est bien ordonné par } < \text{ ssi } \forall X(\forall x(Xx \to Ax) \to \exists z \forall w(Xz \wedge Xw \wedge z \neq w \to z < w))$$

preuves fondés sur ces langages. Il appelles ces systèmes RA_α des systèmes d'*analyse ramifiée*. Les détails de ces travaux sont complexes et nous renvoyons le lecteur intéressé à l'article de présentation générale Feferman (2005) et aux références qui y sont citées. Disons simplement que le plus petit ordinal imprédicatif, sous cette définition, existe, qu'il a été découvert indépendamment par Feferman et Schütte au milieu des années soixante, et que c'est un ordinal récursif appelé Γ_0. [15]

A partir des années soixante-dix, le programme de Feferman connaît une inflexion : plutôt que sur la question de la définissabilité (quels sont les ensembles d'entiers prédicativement définissables ?), le programme est recentré sur la notion de prouvabilité et reformulé dans des termes plus généraux, la question de savoir quels systèmes arithmétiques sont acceptables étant donné notre acceptation d'une théorie arithmétique donnée n'étant plus qu'un cas particulier d'une question plus générale. Cette question plus générale, que Feferman fait sienne, est formulée par Kreisel de la façon suivante :

> Quels principes de preuves reconnaissons-nous comme valides une fois que nous avons compris (ou, comme on dit parfois, « accepté ») certains concepts donnés ? (Kreisel (1970), p.489, cité dans Feferman (2005).)

Les réponses de Feferman à cette question ont pris dans le cours ultérieur de sa réflexion plusieurs formes, sur lesquelles nous ne reviendrons pas, avec, dans le cas de l'arithmétique, des convergences répétées avec ce qu'il avait identifié dès les années soixante comme les limites des théories formalisées prédicativement justifiables. Ce qui importe ici surtout, c'est que c'est en réponse à *cette question* de Kreisel, question qui fait aussi écho à l'idée gödélienne de « nouveaux axiomes » d'une théorie qui peuvent en être une « continuation naturelle »[16], que Feferman développera (provisoirement) la notion de *clôture réflexive d'une théorie schématique* , telle qu'elle est expliquée dans Feferman (1991). Et c'est dans le contexte du développement de la notion de clôture réflexive d'une théorie schématique que Feferman emploie la notion

[15] Je ferai simplement deux remarques en passant, en lien direct avec la discussion du chapitre précédent. D'une part, on peut montrer que l'ensemble des (codes d') énoncés vrais du langage de PA est un ensemble prédicatif au sens de Feferman, autrement dit un ensemble qui peut être défini prédicativement à partir dans le processus décrit ci-dessus. D'autre part, que l'extension aléthique tarskienne de PA que nous avons appelée $T(PA)$, qui est non conservative sur PA, est un système prédicativement justifié au sens de Feferman, autrement un système justifiable sur la seule base de notre compréhension du concept d'entier. Dans tous les cas la notion de vérité, ou l'ensemble des énoncés vrais, n'apporte rien d'essentiellement, ou peut-être faudrait-il dire d'« irréductibement » nouveau, à la différence des ensembles d'entiers non-prédicativement réductibles à la notion d'entier qui seraient introduits par une quantification du second ordre sur tous les ensembles d'entiers.

[16] Voir Feferman (1999*a*).

vérité, emploi dont Jeffey Ketland[17] pense qu'il vaut réfutation du déflationnisme. Je voudrais maintenant montrer ce qui ne va pas, à mon sens, dans cette interprétation, en présentant en même temps les remarques de Ketland et des indications permettant de comprendre le sens de ce travail de Feferman sur la vérité.[18]

Défendant l'interprétation des travaux de Feferman que je conteste, J. Ketland, après avoir rappelé que les principes de réflexions et les énoncés de la cohérence sont des conséquences logiques des extensions aléthiques adéquates de PA, met en avant l'Obligation Epistémique Conditionnelle suivante, qu'il dit tirer de l'interprétation standard du théorème de Gödel, :

Obligation épistémique conditionnelle *Si* quelqu'un accepte une théorie mathématique de base S, alors il doit [*is committed* to] accepter un certain nombre *d'affirmations supplémentaires* dans le langage de la théorie de base (parmi lesquelles l'affirmation de l'énoncé de Gödel G). (Ketland (2005), p.79)

Il continue en citant le travail de Feferman à l'appui de sa condition conditionelle et de l'idée que la vérité joue un rôle dans sa justification :[19] :

> Les théorèmes de Gödel montrent l'inadéquation d'un système formel *seul* (...). Pourtant ils pointent en même temps vers la possiblité d'engendrer systématiquement des systèmes de plus en plus larges dont l'acceptation est implicite dans l'acceptation de la théorie de départ. Les moteurs pour y parvenir sont ce qu'on en est venu a appeler les *principes de réflexion.*(Feferman (1991), p.1)

Rapidement, Feferman explique avoir cherché pendant des années une façon plus naturelle d'engendrer ces extensions et annonce qu'il va présenter ses dernières découvertes dans cette direction :

> Ce que nous présentons ici est une notion nouvelle et simple de *clôture réflexive d'une théorie schématique* qui peut être appliquée de façon très générale. (Feferman (1991), p.1)

Qu'est-ce que la clôture réflexive d'une théorie schématique ?

> En fait, deux notions de clôture réflexive d'une [théorie] schématique $S(P)$ sont introduites dans cet article. La première est conçue pour

[17] Ketland (2005).

[18] Je me contenterai ici du sens très général de son travail. Pour plus d'information, je renvoie le lecteur à Feferman (1991), qui contient des précisions historiques et tous les détails techniques.

[19] Je cite un extrait de Feferman (1991) un tout petit peu plus long que celui cité par Ketland, mais sans différence essentielle.

> répondre à la question : *quels énoncés dans le langage de base L de S [...] doivent être acceptés si l'on a accepté les axiomes et règles de base de S ?* La réponse est donnée par une théorie ordinaire $Ref(S)$ formulée dans un langage $L(T,F) = L \cup \{T,F\}$ où T et F sont des prédicats partiels de vérité et fausseté applicables à eux-même au sens où ils s'appliquent aux (codes des) énoncés de $L(T,F)$. Les axiomes de $Ref(S)$ sont ceux de $S(P)$ appliqués au langage $L(T,F)$ avec des axiomes dénotés par $Self\text{-}TrAx_{L(T,F)}$, où ces derniers expriment les propriétés des prédicats partiels de vérité (et fausseté) du genre de ceux explicitement construits par S. Kripke (1975), entre autres. Ainsi, par exemple, nous pouvons raisonner dans $Ref(PA)$ par induction sur la vérité d'énoncés qui contiennent la notion de vérité, et ainsi parvenir à des énoncés de la forme :
>
> $$\forall x[Prov_{PA}(x) \rightarrow T(x)],$$
>
> et répétant ce genre d'arguments, dériver des principes de réflexions itérées pour l'arithmétique.
>
> La seconde notion générale de clôture réflexive pour [une théorie] schématique $S(P)$ introduite ici répond à la question : *quels schéma $A(P)$ dans le langage de $S(P)$ doivent être acceptés si l'on a accepté les axiomes (schématiques) et règles de base de $S(P)$?* [...] (Feferman (1991), p.2.

Aux oreilles de Ketland, tout ceci sonne comme si Feferman avait prouvé que la vérité était substantielle en un sens qui devrait déplaire au déflationniste. Juste après avoir cité Feferman, il commente ainsi :

> En bref nous pouvons expliquer l'Obligation épistémique conditionnelle en faisant appel à la notion de vérité. Mais la notion de vérité est-elle indispensable à l'explication de l'obligation épistémique conditionnelle ? Il semble que le déflationniste soit coincé [cornered]. Car si la notion de vérité est indispensable à cette explication de l'obligation épistémique conditionnelle, alors les axiomes pour la vérité (qui sont essentiels à la preuve des principes de réflexion et des autres énoncés) *doivent être non-conservatifs*. Et ceci viole la contrainte de conservativité. Si le déflationniste insiste sur la contrainte de conservativité, alors il ne peut pas *expliquer* pourquoi, étant donné que nous acceptons une théorie de base S, nous devons accepter l'énoncé réflexif plus fort : « Tous

> les théorèmes de S sont vrais ». Le déflationniste ne peut pas avoir les deux.(Ketland (2005), p.80)

Que penser de cet argument ? Tout d'abord, nous sommes d'accord que Ferferman soutient bien quelque chose comme la thèse de l'obligation épistémique conditionnelle, et avec cette thèse elle-même, *pourvu qu'elle soit formulée avec plus de précaution* :

Obligation épistémique conditionnelle prudente Si l'on accepte une théorie mathématique de base S *contenant certains schémas ouverts*, alors on doit accepter un certain nombre d'autres énoncés du langage de la théorie de base, du moins si cette théorie est assez forte pour parler de sa propre syntaxe (et l'un des d'eux est l'énoncé de Gödel G).

Quelle est la différence entre l'Obligation épistémique conditionnelle prudente et la condition originale de Ketland ? La différence importante est dans la restriction aux théories schématiques dans l'antécédent du conditionnel. Sans cette restriction, la condition originale de Ketland n'est pas correcte et, je crois, ne reflète pas en fait les vues de Feferman sur son propre travail. Car Feferman insiste, dans l'extrait cité plus haut, sur le fait que son intérêt concerne les théories schématiques, et l'implicite dont elles sont porteuses. Le meilleur moyen de clarifier la façon dont Feferman comprend son travail sur la vérité est peut-être de regarder, non pas son travail antérieur dans la même direction[20], mais son travail plus tardif. Feferman a en effet abandonné par la suite son approche en terme de vérité au profit d'une nouvelle formulation de son programme au centre de laquelle figure la notion de « déploiement » [*unfolding*] d'une théorie. Voici ce qu'écrit Feferman de sa propre entreprise dans le contexte d'un panorama de son œuvre sur la prédicativité :

> Il est de l'essence de la notion de déploiement [unfolding] que nous ayons à faire à des systèmes formels présentés schématiquement. [...]. La philosophie informelle derrière l'usage des schémas dans le concept de déploiement est leur ouverture [open-endedness]. C'est-à-dire qu'ils ne sont pas conçus comme s'appliquant à un langage spécifique dont le stock de symboles de base est fixé à l'avance, mais plutôt comme applicables à n'importe quel langage que nous pourrions en venir à reconnaître comme contenant des notions de base douées de sens. En

[20]Nous pensons bien entendu à son travail sur les progressions de théories. Voir Feferman (1962). Franzén (2004*b*) et Franzén (2004*a*) contiennent un résumé instructif de son travail.

> d'autres termes, il est implicite dans l'acceptation de schémas donnés que nous acceptons toutes ses instances de substitution douées de sens. Mais il n'est pas besoin que soit déterminées par avance quelles substitutions doivent être acceptées. Ainsi par exemple, si quelqu'un accepte les axiomes et les règles du calcul propositionnel classique donné sous forme schématique, il acceptera toutes les instances de substitution de ces schémas dans n'importe quel langage que l'on viendra à employer. La question à laquelle la notion de déploiement est supposée répondre est : étant donné un système schématique S, quelles opérations et quels prédicats - et quels principes les concernant - doivent être acceptés si l'on a accepté S ? (Feferman (2005), p.24).

Par conséquent la question à laquelle la notion de déploiement répond est exactement la même que celle à laquelle répond la notion de clôture réflexive d'une théorie. Dans ce dernier cas, la vérité n'avait qu'un rôle instrumental pour Feferman et, pour autant que je le comprenne, il n'a jamais tiré de conclusions « substantialistes » concernant le prédicat de vérité de ses résultats sur la clôture réflexive.

Revenant maintenant à la conclusion de Ketland, non seulement sa formulation du principe de l'Obligation épistémique conditionnelle est trompeuse (de notre point de vue), mais nous avons vu que le premier membre de l'alternative mise en avant dans son dilemme n'est pas correcte non plus : tout notre propos dans le chapitre 3 a justement été de montrer qu'en présence de schémas ouverts dans la théorie de base, la contrainte de conservativité n'est pas justifiée, et est en fait incorrecte (clairement violée dans certains cas). Il n'y a donc aucun dilemme.

Mais une fois abandonné le dogme de la contrainte de conservativité, la référence de Ketland au programme de Feferman paraît inappropriée. Souvenons-nous de la motivation de Feferman pour introduire sa notion de clôture réflexive d'une théorie schématique dans les passages cités précédemment. En voici un autre, tiré cette fois de la conclusion de Feferman (1991) :

> Les notions de clôture réfexives introduites ici sont relatives à une théorie au sens où elle nous disent ce qui *doit* être accepté *si* on a accepté les notions de base et les principes schématiques de cette théorie.(Feferman souligne. Feferman (1991), p.44)

Supposons maintenant que quelqu'un accepte PA et, au-delà de PA accepte

aussi une théorie des ensembles comme ZFC.[21] Accepter la notion d'ensemble théorisée dans ZFC engage certainement, *modulo* certaines lois de correspondance entre le langage ensembliste et le langage de l'arithmétique, l'acceptation de nouvelles propositions arithmétiques, au-delà de celles exprimées par les théorèmes de PA. Mais accepter PA, inversement, n'engage certainement pas l'acceptation de la notion d'ensemble telle que théorisée dans ZFC. Et en effet la notion d'ensemble théorisée dans ZFC est substantielle relativement aux principes explicatifs formulés dans PA. Maintenant, par contraste, il semble que le fait même qu'accepter PA engage épistémiquement à accepter l'extension aléthique de PA, doit être interprété comme une preuve que les principes aléthiques *ne* sont *pas* substantiels, y compris dans la perspective de Feferman. Une fois abandonnée la conservativité comme mesure de substance épistémologique, ce qui doit frapper dans le processus de réflexion de Feferman ce n'est pas que les extensions sémantiques de PA ne sont pas conservatives sur PA, mais plutôt que ces extensions sont comprises comme un moyen de rendre explicite quelque chose qui était déjà là dans nos engagements pris vis-à-vis de la théorie de base. Pour le dire encore une fois, ce qui est significatif du point de vue de Feferman, c'est précisément que les clôtures réflexives *n'ajoutent pas* de contenu mathématique substantiel ou encore, pour reprendre le terme de Gödel, extrinsèque, au contenu attendu de la théorie de base.[22]

4.3 Les réflexions de Myhill sur la notion de preuve

Le programme de Feferman, inspiré par Gödel et Kreisel,[23] n'est pas le seul se proposant d'élucider la notion de conséquence informelle appelée par les théorèmes d'incomplétude. Nous voudrions maintenant rapporter, pour consolider nos intuitions, quelques réflexions qu'inspirait à John Myhill la considération de notre condition épistémique face aux théorèmes d'incomplétude, dans Myhill (1960).

[21] Ou des théories plus fortes encore. On pourra se référer au programme de Harvey Friedman tel que présenté par exemple dans Feferman et al. (2000).

[22] Pour l'origine de la distinction entre extensions axiomatiques intrinsèques et extrinsèques voir à nouveau Gödel (1964). Pour un développement approfondi du point de vue de Feferman sur la distinction entre extensions intrinsèques et extrinsèques, spécialement dans le contexte de la théorie des ensembles, voir Feferman (1999*a*) ; Feferman et al. (2000).

[23] Voir p.ex. Feferman (2005), p.23, où Feferman rapporte les termes de la question posée par Kreisel qui ont motivés sa recherche :

> What principles of proof do we recognize as valid once we have understood (or, as one sometimes says, 'accepted') certain given concepts ?(Kreisel (1970), p.489)

Quoique la direction de recherche proposée par Myhill se révèle, pour finir, différente de celle de Feferman, *Myhill préférant* in fine *à cette dernière une approche axiomatique de la notion informelle de prouvabilité*, leur point de départ est semblable. Partant du premier théorème d'incomplétude, Myhill y insiste, le théorème ne montre pas seulement que pour toute théorie formalisée suffisamment riche il existe un énoncé (de son langage) indépendant d'elle, mais plutôt qu'il y a un énoncé indépendant d'elle qu'il est *rationnel* de croire, ou pour le dire en termes de conséquence, l'énoncé de Gödel d'un système formalisé S, non seulement est vrai et indépendant de S, mais *suit rationnellement* de S. Myhill souligne que notre compréhension de cette notion « absolue » de prouvabilité n'a rien à voir avec les notions sémantiques :

> J'affirme qu'il y a un sens absolu de « prouvable », ni syntaxique, ni sémantique ni psychologique, et qu'en ce sens de « prouvable », les énoncés indécidables de Gödel sont prouvables.(Myhill (1960), p.463)

un peu plus loin, Myhill précise :

> Nous avons des raisons [*grounds*] logiquement contraignantes de *croire* l'énoncé indécidable de Gödel p. Comme cet énoncé affirme seulement la cohérence de S, si nous ne sommes pas fondés à croire p nous ne sommes pas fondés à croire non plus rien de ce qui est établi par l'usage de S. Le fait que nous ayons besoin d'un système plus fort que S pour donner une preuve formelle de la cohérence de S, ne doit pas nous donner à croire, à tort, que la cohérence de S est moins 'certaine' que les théorèmes établis par l'usage de S. Il est possible de *prouver* p par des méthodes que devons admettre comme correctes si nous admettons les méthodes disponibles dans S comme correctes. (Myhill (1960), p.466)

Ces lignes, guidées par une intuition proche de celle de Feferman[24], semblent devoir exclure que le concept de vérité puisse avoir un rôle explicatif réel dans notre reconnaissance de la vérité de p. La raison en est simple : si en effet l'acceptation des méthodes pour prouver p est toute entière justifiée par notre acceptation de S, on voit mal alors le rôle que pourrait y jouer notre concept de vérité, lequel n'est en général pas théorisé dans la théorie de départ (S), sauf à voir en lui un simple moyen *d'exprimer* ou *d'expliciter* ces méthodes de preuves que, d'emblée, dès notre

[24]La dernière phrase en particulier.

acceptations de S (S dont nous supposons que le concept de vérité n'y figurait pas), nous avions accepté.

Mais c'est sans doute dans les premières pages de l'article que Myhill est le plus clair. Le passage suivant mérite d'être cité tout entier :

> Je vais maintenir qu'il y a un sens d' 'inférence correcte' ou 'preuve' qui d'un côté n'est pas une notion *syntaxique* relative à un système particulier, et d'un autre côté n'est pas réductible à la simple préservation de la vérité, ou à aucune autre notion *sémantique*. En fait il me semble que l'usage du terme 'preuve' dans les discussions mathématiques non-philosophiques ordinaires est assez clairement ni un terme syntaxique ni un terme sémantique.
>
> Il est aussi contradictoire d'employer des méthodes de preuves sans admettre leur correction, que de faire des affirmations sans admettre leur vérité. (Je n'utilise pas 'contradictoire' au sens de la logique formelle, mais à peu près comme un synonyme pour 'irrationnel'). Par conséquent si une personne qui a utilisé certaines méthodes pour prouver des théorèmes arithmétiques parvient à expliciter ces méthodes, il est *ipso facto* commis à la proposition parfaitement définie que l'emploi de ces méthodes ne peut conduire à une affirmation arithmétique fausse, par exemple l'affirmation que 0 est égal à 1. Par la technique d'arithmétisation de Gödel, qui traduit toute affirmation de dérivabilité formelle en un énoncé arithmétique [*arithmetic statement*], une telle personne est obligée d'admettre un nouvel énoncé arithmétique, nommément la version arithmétisée de l'affirmation que ses méthodes ne peuvent conduire à une preuve de l'affirmation que 0 est égal à 1. Par le théorème de Gödel, elle n'aurait pas pu établir cette proposition par ses méthodes antérieures. Donc, dès qu'une personne explicite les outils qu'elle a utilisés dans la construction des preuves arithmétiques, elle est *ipso facto* en position d'obtenir de nouvelles preuves arithmétiques qu'elle n'aurait pu obtenir en utilisant ces seuls outils. L'ensemble du processus est lié de près à ce que le philosophe logicien britannique W.E. Johnson appelait 'induction intuitive' ; nous constatons que nous faisons certaines inférences et réalisons là-dessus que le modèle [*pattern*] de ces inférences est de nature à conférer la validité à des arguments dans lesquels elles ap-

> paraissent. Cette prise de conscience est un pas rationnel et démonstratif en dehors de toute question de formalisation, quoique bien entendu les *résultats* de cette induction intuitive peuvent être formalisés après que l'induction ait eu lieu. (Souligné par l'auteur. Myhill (1960), 461-462)

Si l'on pense avec Myhill que l'énoncé de la cohérence d'une théorie A est quelque chose comme une conséquence rationnelle, sinon de la théorie A du moins de notre acceptation de A ; si l'on s'accorde avec lui que le pas décisif dans l'appréhension de cette conséquence est l'acte réflexif par lequel nous formons des croyances d'« ordre supérieur », c'est-à-dire ayant pour objet des propositions à propos de propositions et des règles que nous avons acceptées ; si l'on croit avec lui avec que la justification de l'inférence réflexive est autonome relativement au concept de vérité et de tout autre concept sémantique, et plus précisément qu'elle procède d'une réflexion sur notre attitude relativement à certains moyens de preuves et non sur la nature de la vérité ; et si l'on pense enfin que les *résultats* de cette induction intuitive peuvent être formalisés après coup alors il devient naturel d'envisager la possibilité que le rôle de la vérité dans les preuves de cohérence, loin d'introduire des principes d'explication et de justification nouveaux, est seulement de permettre de formuler les résultats de cette « induction intuitive ». À nouveau, toute interprétation des phénomènes d'incomplétude tendant à établir que les énoncés indécidables gödéliens d'une théorie en sont, d'une manière ou d'une autre, des conséquences implicites, que leur justification est « déjà là » dans nos justifications pour accepter la théorie en question, va dans le sens d'une révision à la baisse de la portée épistémologique qui confèreraient les axiomes tarskiens à la notion de vérité : ces axiomes, après tout, loin d'introduire un notion riche et radicalement nouvelle relativement à la théorie de départ, une notion qui fournirait de nouveaux supports conceptuels à nos justifications, ne sont peut-être que le moyen d'expliciter des justifications qui étaient « déjà là ». Et cette interprétation va dans un sens radicalement opposé à celui dans lequel allait l'interprétation de Shapiro et Ketland.[25]

[25] Il vaut de noter la façon dont Myhill aborde la notion de prouvabilité informelle, ou absolue, plus en détail, par contraste avec la méthode de Feferman. Puisque la notion de prouvabilité absolue n'est réductible ni à la syntaxe ni à la sémantique, l'approche axiomatique semble la plus « naturelle »[26]. Myhill propose les trois principes suivants pour un prédicat de prouvabilité absolue, noté B s'appliquant à des énoncés arithmétiques. :

1. $B^{\ulcorner}p^{\urcorner} \rightarrow p$
2. $B^{\ulcorner}p \rightarrow q^{\urcorner} \rightarrow (B^{\ulcorner}p^{\urcorner} \rightarrow B^{\ulcorner}q^{\urcorner})$
3. La règle de preuve suivante : Si p est un théorème, alors $B^{\ulcorner}p^{\urcorner}$ est un théorème.

4.4 Le programme d'Isaacson et la vérité arithmétique

Daniel Isaacson, dans deux articles importants[28], a défendu un programme de recherche qui, de prime abord, s'oppose radicalement à l'idée que certaines propositions, comme la cohérence de PA pourraient être des conséquences implicites de PA. Nous verrons que les thèses d'Isaacson, si elles sont correctes, posent une difficulté pour la position que nous défendons ici ; toutefois, et c'est ce que je tenterai de montrer, il n'est pas certain que ces difficultés soient insurmontables.

La thèse d'Isaacson, pour l'exprimer crûment, est que l'arithmétique de Peano en premier ordre, PA, est une théorie complète. Bien entendu cette thèse n'est pas une thèse *logique*, et ce n'est pas de la complétude au sens familier aux logiciens dont il s'agit ici. La thèse que défend Isaacson est de nature *épistémologique*. En deux temps, l'idée est la suivante : premièrement, PA occupe une place conceptuelle particulière et privilégiée parmi nos théories de l'arithmétique ou, pour le

Ces principes, précise Myhill, sont à adjoindre à une théorie de base ayant ses principes de preuves propres et suffisamment d'arithmétique pour permettre de parler des énoncés et prouver des faits élémentaires de leur syntaxe. Supposons pour simplifier un peu les choses que la théorie de base, A, soit finiment axiomatisée par un énoncé σ. Dans la métathéorie contenant A et même PA (pour avoir l'induction), et les axiomes et règles gouvernant le prédicat de prouvabilité « absolue » que nous venons de spécifier, à partir de σ et de la règle 3 nous pouvons inférer $B(\sigma)$, et de là $\forall x(Ax(x) \rightarrow B(x))$.

La généralisation universelle de l'axiome 2,

$$\forall x, y, z, [x = \ulcorner y \rightarrow z \urcorner \rightarrow (B(x) \rightarrow (B(y) \rightarrow B(z)))]$$

garantit que le *modus ponens* préserve la prouvabilité.[27] On ferait de même avec la règle de généralisation universelle, en ajoutant l'axiome affirmant que si une formule à une variable libre est prouvable alors sa clôture universelle l'est aussi. Nous pouvons donc prouver dans le système étendu, par induction que :

$$\forall x(Pr_A(x) \rightarrow B(x)).$$

Or

$$B(\ulcorner 0 = 1 \urcorner) \rightarrow 0 = 1$$

est un théorème de ce système (d'après l'axiome 1), ainsi que « $0 \neq 1$ » (théorème de A), donc « $\neg B(\ulcorner 0 = 1 \urcorner)$ » également. Donc

$$\neg Pr_A(\ulcorner 0 = 1 \urcorner)$$

est un théorème du système étendu. Nous avons là, pour reprendre, l'expression de Shapiro, une « preuve convaincante » de la cohérence de A, et cette preuve ne fait pas appel à la vérité mais au concept de prouvabiltié absolu. Et à leur tour les axiomes gouvernant ce prédicat de prouvabilité sont justifiés par une réflexion sur nos pratiques inférentielles, sans appel au prédicat de vérité.

[28] Isaacson (1996), parus pour la première fois en 1987, et Isaacson (1992).

dire dans termes qui ne sont pas ceux de l'auteur, l'ensemble des théorèmes de PA forme quelque chose comme une espèce épistémologique naturelle dans l'ensemble plus large des vérités arithmétiques. Alors que les théorèmes d'incomplétude ont montré le caractère incomplet et incomplétable de toute théorie formelle de l'arithmétique, Isaacson remarque que PA semble se distinguer comme une axiomatisation intrinsèquement importante. Il suggère que les raisons en sont conceptuelles, et non seulement historiques. La seconde partie de la thèse est une tentative d'élucidation de cette propriété épistémologique qui caractérise PA en propre : les théorèmes de PA sont plus précisément ces vérités arithmétiques qui peuvent être perçues comme telles sur la base d'une analyse conceptuelle de la notion d'entier uniquement. Isaacson :

> L'idée est que [PA] consiste en ces vérités qui peuvent être perçues comme vraies directement à partir du contenu purement arithmétique d'une analyse conceptuelle catégorique de la notion d'entier naturel. Les vérités exprimables dans le langage de l'arithmétique (du premier ordre) qui se tiennent au-delà de cette région sont telles qu'il n'y a aucun moyen par lequel leur vérité peut être perçue en termes purement arithmétiques. (Isaacson (1996), p.203)

Insistons qu'Isaacson ne conteste pas que notre compréhension complète de la notion d'entier engage des ressources conceptuelles qui ne sont pas disponibles dans PA, en particulier des notions d' « ordre supérieur » comme la quantification sur des ensembles d'entiers. Ces deux thèses sont clairement contrastées dans le passage suivant, entre autres :

> [...] je *n*'affirme *pas* que PA pourrait constituer en elle-même une base conceptuelle adéquate pour notre compréhension du concept d'entier naturel. Loin de là, je considère que nous ne pouvons arriver à un tel système que sur la base d'une compréhension d'ordre supérieur. Le système PA émerge [*arises*] comme constituant le contenu purement arithmétique de notre compréhension complète du concept d'entier naturel, où cette compréhension est implicitement et de façon inhérente d'ordre supérieur. (Isaacson (1996), p.209).

Notre propos dans les lignes qui suivent ne sera pas d'examiner la signification exacte de la thèse elle-même ni l'ensemble des arguments, souvent riches et difficiles, qu'Isaacson apporte en sa faveur. Notre nous concentrerons sur les rela-

tions qui existent entre la thèse d'Isaacson et notre propre interprétation du rôle épistémologique du concept de vérité, en particulier sur la question de leur compatibilité. Il semble en effet exister une tension entre trois idées, ou plutôt entre deux idées et un fait : l'idée que nous utilisons le concept de vérité seulement pour exprimer ce qui est 'déjà là' dans une théorie à l'état latent, en particulier dans PA, le fait que cette expression rend effectifs de nouveaux principes de preuves, enfin l'idée que PA est l'expression complète de notre compréhension des entiers s'obtenant par analyse conceptuelle de la notion d'entiers uniquement. En effet, si la thèse d'Isaacson est vraie alors, si l'extension aléthique de PA n'est pas conservative sur PA, le contenu de l'extension aléthique de PA va au-delà de ce qui peut s'obtenir par l'analyse conceptuelle de la notion d'entier dans le langage de PA. Notre objet principal sera de réfuter l'affirmation que, de cette implication, cette conséquence s'en suit : une authentique analyse conceptuelle de la notion de vérité est à l'œuvre dans la justification de l'extension aléthique de PA et de la vérité de l'énoncé de sa cohérence. La thèse d'Isaacson impliquerait donc pour finir que Shapiro et Ketland ont raison contre nous à propos du rôle épistémologique de la vérité. De façon intéressante, Leon Horsten a précisément soutenu une sorte d'extension de la thèse d'Isaacson, qu'il a défendue contre les objections gödéliennes en « chargeant » l'épistémologie de la vérité. Avant d'en revenir à la portée de la thèse d'Isaacson lui-même, nous allons donc présenter brièvement la thèse de L. Horsten, en insistant sur sa défense contre lesdites objections.

4.4.1 Horsten : vérité des énoncés mathématiques et vérité mathématique

> Nous allons soutenir que, contrairement aux premières apparences et avec quelques précisions, la thèse que la classe des vérités mathématiques coïncide avec la classe des théorèmes de ZFC est philosophiquement défendable. (Horsten (2001), p.173)

Et telle est, en substance, la thèse d'Horsten. À nouveau, nous n'entrerons pas dans l'ensemble des difficultés que soulève cette thèse et les précisions qu'elle requiert. Ce qui nous intéresse, c'est la réponse de Horsten à l'objection immédiate qu'elle soulève, analogue à celle soulevée par la thèse d'Isaacson :

> Une objection vient immédiatement à l'esprit. Si tous les théorèmes de ZFC sont des vérités mathématiques, alors certainement l'énoncé de

> Gödel pour ZFC (G_{ZFC}) et l'énoncé exprimant la cohérence de ZFC (Coh_{ZFC}) sont des vérités mathématiques ? (Horsten (2001), p.176)

Confronté à l'objection analogue, Isaacson répondait que ces énoncés gödéliens étaient des énoncés *métamathématiques*. Quoiqu'ils fussent vrais et formulés dans le langage de l'arithmétique, ces énoncés n'étaient pas arithmétiques dans son sens : pour percevoir leur vérité, il faut percevoir certains faits non perceptibles sur la seule base de notre analyse conceptuelle de la notion d'entier (comme par exemple la cohérence de PA, et la possibilité d'exprimer, *via* le mécanisme du codage, des faits non arithmétiques dans le langage de l'arithmétique). Il n'en restait pas moins que, pour Isaacson, la cohérence de PA était un fait, sinon arithmétique (en son sens), du moins un fait mathématique. La thèse d'Horsten ne concernant pas les vérités arithmétiques spécifiquement mais les vérités mathématiques en général, il doit pousser cette défense un peu plus loin.

Tout d'abord, Horsten concède que G_{ZFC} et Coh_{ZFC} sont vrais, et en veut pour preuve ce que nous avons déjà appelé « la preuve par la vérité » faisant appel aux axiomes tarskiens. Mais il ajoute ensuite qu'un défenseur de sa thèse qui accepte la correction de la preuve par la vérité

> insistera que *pour être une vérité mathématique, il ne suffit pas d'appartenir au langage des mathématiques et d'être vrai.*(Horsten (2001), p.177).

L'idée de Horsten est alors de distinguer parmi les énoncés mathématiques (ou métamathématiques) vrais ceux qui ont une preuve mathématique (les *vérités mathématiques*) de ceux qui ont une preuve philosophique (les *énoncés mathématiques vrais*).[29] Et, précisément, la preuve par la vérité de la cohérence de ZFC est, selon Horsten, un exemple de preuve du second genre :

> C'est simplement que de telles preuves gödéliennes ne sont pas des preuves *purement* mathématiques. Car elles contiennent essentiellement la notion de *vérité*, qui n'est pas elle-même une notion mathématique mais philosophique.(Horsten (2001), p.177)

et Horsten de préciser :

[29] Comme Isaacson a distingué un sens épistémologique de la notion de vérité arithmétique du simple concept de vérité exprimable dans le langage de l'arithmétique, Horsten distingue un sens épistémologique de la notion de vérité mathématique de la notion d'énoncé mathématique vrai. Les deux thèses sont très proches en esprit.

> Ce n'est pas nier que les mathématiques peuvent être appliquées pour produire des théories de la vérité intéressantes [note omise]. C'est seulement que les théories mathématiques de la vérité, de ce point de vue, appartiennent non pas aux mathématiques pures mais au mieux aux mathématiques *appliquées*, ou à la partie la plus mathématique de la philosophie.(Horsten (2001), p.177)[30]

Il y a une certaine interprétation de cette conclusion, nous semble-t-il, qui n'est pas compatible avec la transparence épistémique du concept de vérité. Il est crucial pour l'argument de Horsten que la vérité soit un concept contentuellement riche, c'est-à-dire dont l'analyse conceptuelle soit la source de justifications originales propres. C'est parce que les lois de la vérité sont distinguées par une analyse proprement *philosophique* que la preuve par la vérité des propositions gödéliennes est une preuve philosophique. Le caractère épistémologiquement efficace du concept philosophique de vérité en est une conséquence immédiate : il y a un contenu non mathématique de la notion de vérité, totalement hétérogène au contenu de ZFC, et c'est bien ce contenu qui est à l'œuvre dans l'explication des propositions gödéliennes. Cette analyse de la situation est donc en harmonie avec les arguments de Shapiro et Ketland sur la substance de la vérité, et incompatible avec notre thèse que le rôle du prédicat de vérité est seulement d'expliciter nos engagements épistémologiques. [31,32]

4.4.2 Isaacson, la complétude épistémologique de l'arithmétique et le concept de vérité.

La proposition d'Horsten, si elle était correcte[33], s'appliquerait également au cas des preuves par la vérité des énoncés de Gödel pour PA, si bien que la thèse

[30] Citer en note les éléments allant dans son sens que Leon croit trouver chez Gödel, p.179

[31] Remarquons tout de même que malgré leur oppositions sur la nature du concept de vérité, ces deux thèses s"accordent sur le fait que la nature des preuves par la vérité est différente des preuves mathématiques ordinaires en ce sens qu'elle fait appel à des modes du connaître ou des sources de justifications qui ne se réduisent pas aux modes ordinaires ou canoniques de la connaissance mathématique. Sur ce point fondamental, je m'accorde avec Horsten.

[32] Dans un article à paraître (Horsten ([forthcoming])), Leon Horsten défend par ailleurs une thèse d'inspiration déflationniste selon laquelle le concept de vérité est un outil « inférentiel ». Nous nous sentons en accord avec les idées directrices de l'article qui, par leur esprit, se rapprochent des nôtres, quoique les arguments et les détails de la thèse diffèrent largement. Reste que cette thèse nous paraît *prime abord* en tension avec celle présentée ici.

[33] Et si nous l'avons interprétée correctement.

d'Isaacson (et celle de Horsten avec elle), pourrait être vue comme la prémisse d'un argument possible contre la transparence de la vérité. Mais ce n'est pas la voie choisie par Isaacson pour soutenir sa propre thèse. En fait Isaacson, pour autant que nous sachions, ne s'est jamais attardé à considérer en propre la signification qu'a pour sa thèse le phénomène de la non-conservativité des extensions aléthiques de PA sur PA elle-même. Mais il en dit beaucoup, néanmoins, dans le cours de considérations connexes, et nous trouverons un appui textuel solide dans Isaacson (1992). Pour le dire d'emblée, je ne crois pas au bout du compte que la thèse d'Isaacson soit incompatible avec notre thèse sur le rôle de la vérité.

Nous nous attelons donc maintenant à cette tâche : défendre la compatibilité de la thèse d'Isaacson avec l'idée que le rôle épistémologique du concept de vérité n'est pas celui d'un concept explicatif. Pour commencer, il nous faut souligner qu'Isaacson lui-même défend incidemment l'idée que la vérité *a* un pouvoir d'explication, mais d'une façon qui ne met pas en danger sa propre thèse, comme nous allons le voir. Dans ce qui suit, nous suivrons la terminologie d'Isaacson et nous dirons qu'une proposition ou un énoncé est *arithmétique* (en un sens épistémologique, donc), *non seulement si elle formulable par un énoncé du langage de PA en premier ordre* (quantification uniquement sur les entiers, sans passage par des notions d'ordre supérieur), mais encore si *sa vérité peut être perçue directement à partir de notre analyse conceptuelle de la notion d'entier*. Si, comme le veut Isaacson, PA est complète relativement à cette notion d'énoncé arithmétique, tous les théorèmes de PA doivent être arithmétiques ; par conséquent, non seulement les axiomes de PA doivent être arithmétiques, mais la conséquence logique doit encore préserver la propriété d'être arithmétique. Ce point donne lieu a une objection potentielle, qu'Isaacson formule ainsi :

> On pourrait arguer, contre cette vue [*que tous les théorèmes de PA sont arithmétiques*], que la dépendance aux dérivations logiques rend une proposition non-arithmétique [...] puisque la justification d'une déduction logique est en termes du concept général de vérité, plutôt qu'à partir de notre conception fondamentale des entiers naturels.(Isaacson (1992), p.92)

La formulation de cette objection témoigne d'une différence de point de vue entre Isaacson et nous quant à la nature du rôle de la vérité : Isaacson tient que le concept de vérité permet de justifier les lois de la logique, alors que nous pensons

qu'il n'en est rien.[34] Dont acte. Notons cependant déjà ici que, si ce désaccord est réel, ses conséquences restent circonscrites. Revenons donc à la réponse d'Isaacson à l'objection qu'il a lui-même formulée. C'est la suivante :

> Le point est qu'une explication [*account*] de la vérité arithmétique peut, et en fait doit, procéder sur la base du fait que nous comprenons le langage de l'arithmétique. Ce langage contient les notions spécifiquement arithmétiques « zéro », « successeur », « plus », « multiplié » et peut-être d'autres, que nous comprenons *via* notre saisie de la structure des entiers naturels. Mais il doit aussi inclure des notions telles que « et », « ou », « non », « si...alors », « tous », « quelque ». Notre saisie de ces dernières notions valide les principes de la logique du premier ordre. Pour autant que nous comprenons le langage de l'arithmétique du premier ordre, nous sommes en possession d'une base permettant de percevoir la correction des principes de la logique du premier ordre, ainsi que la vérité de chaque axiome de l'arithmétique de Peano appliquée à la structure des entiers naturels.(Isaacson (1992), p.96-97)

Nous avons donc besoin du concept de vérité pour justifier les lois qui gouvernent le langage de la logique du premier ordre, mais il n'en reste pas moins qu'une fois acceptées les lois de la logique du premier ordre, les théorèmes de PA sont perçus comme vrais dans le langage de l'arithmétique directement en vertu de notre analyse des nombres entiers. L'objection étant désamorcée, nous en venons à la question qui intéresse au premier chef Isaacson, et nous avec lui, celle de savoir, non pas si tous les théorèmes de PA sont arithmétiques, mais si PA est *complète* pour cette notion d'énoncé arithmétique.

Pour défendre sa réponse positive à cette question, Isaacson doit bien sûr soutenir que l'énoncé gödélien de la cohérence de PA n'est pas un énoncé arithmétique. Son idée est que la vérité des énoncés gödéliens n'est pas perceptible directement à partir de notre perception de la structure des entiers : c'est la vérité d'un fait non-arithmétique qui est perçue dans le cours d'une méta-réflexion sur notre théorie et non sur les entiers, et le fait que cette vérité peut être codée dans le langage de l'arithmétique. Pour percevoir la vérité de l'énoncé de la cohérence, il faut en effet percevoir le fait de la cohérence de PA, et cette perception ne peut provenir d'une réflexion sur le concept d'entier uniquement mais doit reposer également sur une

[34] Voir ci-dessus les sections 4.2 et 4.3, et les chapitres 6 et 7.

réflexion « d'ordre supérieur », à savoir une réflexion sur notre propre réflexion sur les entiers telle qu'elle a trouvé sa forme dans PA. À ce point, Isaacson ne dit pas si pour opérer cette réflexion d'ordre supérieur un concept de vérité est requis. Mais réfuter le caractère arithmétique (en son sens) des énoncés gödéliens n'est pas suffisant pour soutenir la thèse d'Isaacson. Pour lui donner sa crédibilité, il faut encore passer en revue différentes extensions strictes de PA et montrer qu'elle ne sont pas justifiables sur la base de notre perception de la structure des entiers. Nous allons nous arrêter à l'une de ses extensions parce que l'analyse qu'en donne Isaacson est directement pertinente pour notre recherche sur la vérité.

Isaacson (1992) est consacré à l'étude d'une famille d'extensions de PA obtenues par adjonction d'une version ou d'une autre de l'ω-règle. Dans la mesure où l'adjonction d'une version de l' ω-règle à PA produit en général une extension non-conservative de PA, Isaacson doit donc montrer que la correction de ces différentes versions de l'ω-règle, celles du moins dont l'adjonction à PA produit des extensions non conservatives, n'est pas une conséquence immédiate de notre conception fondamentale des entiers. La version qui nous intéresse est la suivante, ϕ étant une formule du langage de l'arithmétique :

$$\omega\text{-règle}\ \frac{\text{Pour tout entier naturel n, } PA \vdash \phi(\underline{n})}{\forall x \phi(x)}$$

Isaacson remarque que si nous supposons que nous savons :

> (1) Tout énoncé prouvable dans le système PA est vrai dans la structure des entiers naturels

alors nous pouvons établir l'ω-règle comme une extension de PA de la façon suivante. Nous citons Isaacson :

> Supposons la prémisse de l'ω-règle, i.e.
>
> (2) Pour tout entier naturel n, $PA \vdash \phi(\underline{n})$
>
> A partir de (2), par (1), nous avons
>
> (3) Pour tout entier naturel n, « $\phi(\underline{n})$ » est vrai
>
> Ce qui suit est la clause pour le quantificateur universel dans la définition inductive de la vérité, dans le cas où chaque élément du domaine est dénoté par un terme canonique du langage (si bien que l'induction peut être faite directement sur le prédicat de vérité, plutôt que via la relation de satisfaction) :

> (4) Pour tout élément d du domaine, « $\phi(\underline{d})$ » est vrai (dans le domaine donné) si et seulement si « $\forall x \phi(x)$ » est vrai (dans ce domaine),
> où « $\underline{\text{d}}$ » signifie le terme canonique du langage qui dénote l'élément d du domaine.
> La condition (3) et la moitié du biconditionel (4), spécialisé au cas du domaine des nombres entiers, donne
> (5) « $\forall x \phi(x)$ » est vrai dans le domaine des entiers naturels
> La condition d'adéquation fondamentale [*basic*] pour tout prédicat de vérité, « ... est vrai », est le schéma (la « convention-T » de Tarski)
> (6) « p » est vrai si et seulement si p
> La ligne (5) et l'instance appropriée de (6) donne
> (7) $\forall x \phi(x)$
> interprété dans le domaine des entiers naturels. La dérivation de (7) à partir de (2) est l'ω-règle.(Isaacson (1992), p.107)

La question, pour Isaacson, est donc de savoir si cette dérivation est arithmétique. Comme chaque pas de l'inférence est valide en logique du premier ordre, et que le langage de l'arithmétique est clos pour la conséquence en premier ordre (voir argument précédent), ce sont les prémisses de l'argument qu'il faut scruter. Autrement dit il faut se demander si (1), (4) et (6) sont arithmétiquement acceptables. Mais (4) et (6) ne sont pas problématiques en vertu du fait qu'ils découlent de notre acceptation de la conséquence logique pour le langage du premier ordre :

> La clôture de la vérité arithmétique pour la conséquence logique du premier ordre porte en elle un engagement à accepter la théorie générale de la vérité, puisque la propriété fondamentale de la conséquence logique est la préservation de la vérité [*note omise*], et, en conséquence, cette notion de vérité arithmétique contiendra (4) (par exemple en validant le principe de généralisation universelle) et (6) [*note omise*].

Reste donc la prémisse (1). Or la prémisse (1) n'est rien d'autre que le principe de Réflexion pour PA. La question d'Isaacson est donc maintenant exactement celle que nous explorons : quelle est la base de la justification du principe de Réflexion ? Plus précisément, dans les termes de l'auteur : peut-on percevoir la vérité du principe de Réflexion sur la base de notre seule perception de la structure des entiers et des principes qui justifient notre acceptation de la notion de conséquence logique en premier ordre, au rang desquels, d'après le passage cité à l'instant, il faut compter

« la théorie générale de la vérité » (c'est-à-dire les axiomes récursifs à la Tarski) ? Voici la réponse d'Isaacson :

> Je ne nie pas, et en fait je considère certainement, que nous pouvons voir que *tous* les axiomes et les théorèmes de l'arithmétique de Peano sont vrais dans la structure des entiers naturels, mais le point est que le faire est un pas au-delà de notre saisie de la structure fondamentale des entiers naturels, et donc viole une contrainte de minimalité. Ce qui est arithmétique doit être exprimable dans le langage de l'arithmétique du premier ordre [...] et perceptible comme vrai *seulement* à partir de notre compréhension du concept fondamental d'entier naturel. Dans la seconde étape du processus permettant d'établir le fait que tous les axiomes et les théorèmes de l'arithmétique sont vrais dans la structure des entiers naturels, nous allons au-delà du corps de vérités initialement perceptibles comme vraies à partir du concept fondamental d'entier naturel par le processus *supplémentaire* de réflexion sur le fait que les axiomes auxquels nous sommes arrivés par ce premier processus sont, en vertu de ce processus, vus comme vrais dans la structure des entiers naturels. Dans ce second processus de réflexion, ce sur quoi nous réfléchissons n'est pas le concept fondamental des entiers naturels, mais notre processus initial de réflexion sur ce concept.(Isaacson (1992), p.109)

En d'autres termes, puisque, selon Isaacson, notre acceptation des lois de la vérité est implicite dans notre acceptation de PA comme théorie de l'arithmétique, et que le principe de réflexion n'est pas arithmétique, il faut donc que notre acceptation des lois de la vérité ne soient *pas* suffisantes, en conjonction avec nos justifications purement arithmétiques d'accepter PA, pour justifier le principe de réflexion.

Il vaut la peine de rapprocher les remarques d'Isaacson de choses que nous avons déjà vues en passant : pour Isaacson, des axiomes et théorèmes de PA au principe de réflexion il y a quelque chose comme un changement de sujet : nous passons d'une réflexion sur les entiers à une réflexion sur cette réflexion (dernière ligne de la citation précédente). C'est aussi ce processus de réflexion que Myhill plaçait au cœur de son explication de notre capacité à prouver les énoncés de Gödel, même si pour sa part il n'y voyait pas un changement de sujet mais, indépendamment de tout sujet, quelque chose comme un prolongement rationnel de nos moyens de preuves. L'insistance

de Myhill à distinguer dans ce processus l'indication de l'existence d'un concept de prouvabilité absolu indépendant de toute notion sémantique, cette insistance montre que l'on peut interpréter l'usage du prédicat de vérité dans la formulation du principe de réflexion comme épistémologiquement neutre, au sens où sa présence ne nous dit rien du rôle de la saisie du concept de vérité dans la reconnaissance de la correction de ce principe. De même la citation précédente me semble indiquer que, pour Isaacson, c'est la réflexion sur nos moyens de preuves, *et non sur la notion de vérité*, qui permet de justifier, sur la base de PA, le principe de réflexion pour PA. Les « concepts d'ordre supérieur » qui, selon Isaacson, sont crucialement à l'œuvre dans nos justifications réflexives des énoncés de Gödel, ce ne sont pas les concepts sémantiques, mais les concepts *d'axiomes de* PA, de *preuve dans* PA, *de cohérence* d'un système formel. Le cœur rationnel de ces différentes formulations me semble en harmonie avec la thèse que je tente de mettre en avant, à savoir que c'est dans une réflexion, non sur la nature de la vérité, mais sur nos propres supports de justification et états épistémiques que nous trouvons des ressources pour étendre nos moyens de justification.

4.5 Conclusion

Il n'est pas sans doute pas nécessaire de revenir en conclusion sur le détail de ce qui se présente déjà soi-même comme un résumé analytique d'un certain nombre de positions philosophiques. J'insisterai donc simplement sur les trois point importants que je retiens de ce bref examen dans la perspective générale de cette thèse.

Le premier point, gödélien, nous rappelle à la distinction qu'il faut savoir maintenir entre logique et épistémologie : c'est plus précisément l'idée que tous les phénomènes de non-conservativité n'ont pas la même signification épistémologique.

La second est intimement lié au premier : c'est l'idée que nous avons dans certaines circonstances une voie privilégiée de reconnaissance de la vérité de certains énoncés, en particulier des énoncés gödéliens d'une théorie. Les énoncés gödéliens d'une théorie A seraient, en un certain sens, des conséquences *informelles* de A ; leur assertabilité serait *implicite* dans celle de A, et il reviendrait à quelque chose comme un processus de *réflexion* sur une théorie de permettre d'expliciter la vérité de G et de l'énoncé de la cohérence.

Enfin, si nous avons besoin du concept de vérité dans ce processus, il semble que ce soit pour formuler le produit de cette réflexion, et non que nous fassions

appel à un concept et à des lois dont la connaissance est radicalement étrangère à notre compréhénsion des principes sur lesquels nous réfléchissons. L'utilisation de vérités aléthiques ici, n'a pas le même sens que le recours à des nouveaux principes d'explication, mais n'est que le moyen d'expliciter ce qui est déjà là. À ce titre, l'extension d'une théorie par les principes aléthiques tarskiens semble très différente de son extension par des principes mathématiques « substantiels ».

Dans la troisième partie de ce travail, je proposerai une explication du rôle du concept de vérité dans notre appréhension des théories qui permette de rendre compte de ces intuitions.

Conclusion de la seconde partie

Au chapitre 4, nous avons montré que le critère de non-conservativité comme réponse au Problème de la frontière ne justifiait pas une interprétation du rôle de la vérité dans les preuves de cohérence qui serait incompatible avec les thèses déflationnistes. Au chapitre 5, les réflexions de Gödel et Ferferman sont venues encourager et donner un contenu plus précis à cette idée que les phénomènes de non-conservativité associés aux preuves de cohérence par la vérité ont un caractère tout à fait spécial. En particulier, pour Gödel, pour Feferman, comme pour Myhill, une des leçons des théorèmes d'incomplétude est qu'il faut tracer une frontière épistémologique entre les extensions qui ne font qu'expliciter, formuler ou, pour Myhill, suivre « rationnellement », des contenus déjà acceptés. Et, pour Feferman du moins, l'usage des principes aléthiques est précisément celui d'un moyen d'explicitation. Je propose au chapitre 6 de faire de cette idée, en la précisant et en essayant de lui ôter tout air de mystère, le socle d'une réponse au Problème de la frontière et d'en développer les conséquences.

Troisième partie

Réflexion et vérité

Introduction

Dans la première partie de ce travail, j'ai présenté les données du problème : les thèses déflationnistes relativement à l'usage de la notion de vérité et à sa portée épistémologique d'un côté et, d'un autre côté, les ressources théoriques nécessaires à l'explication de certains usages de la notion de vérité. Dans la seconde partie j'ai examiné de façon critique une réponse possible aux problèmes de la stabilité, de la frontière et de l'usage[35] soulevé par les thèses déflationnistes. Dans cette partie, je voudrais essayer de présenter ma propre réponse à ces problèmes.

Cette troisième partie est néanmoins composée de deux chapitres bien distincts dont le point commun est seulement, en un sens vague, de chercher à comprendre les liens existant entre certains principes de réflexion et la nature de notre rationalité. Le premier chapitre constitue une recherche pour ainsi négative, et un *excursus* dans le programme de cette thèse ; il peut être passé en première lecture. Nous avons vu dans les chapitres précédents que la cohérence d'une théorie A pouvait être *déduite* de A et de principes généraux *a priori* concernant la notion de vérité (la théorie tarskienne) et de syntaxe. L'existence de cette déduction montre qu'un sujet qui accepte une théorie A et ces principes aléthiques peut construire une justification de la cohérence de A fondée sur ces prémisses, et que par conséquent un sujet acceptant une théorie donnée doit, en un sens, en accepter la cohérence. Mais n'y a-t-il pas d'autre moyen de rendre raison de l'idée qu'un sujet acceptant une théorie *doit*, rationnellement, en accepter la cohérence ? Un sujet acceptant une théorie mais refusant la notion de vérité ne peut-il pas justifier lui aussi son acceptation de la cohérence de la théorie ? Dans ce chapitre, j'explore la possibilité d'une justification pragmatique fondée sur l'analyse des engagements rationnels qui sont contractés par un sujet acceptant une théorie donnée.

C'est seulement au chapitre suivant que je renoue avec le fil principal de ce travail

[35] Voir l'introduction au chap. 3.

et la question du rôle de la notion de vérité dans les preuves. J'essaie d'y répondre aux questions laissées ouvertes à la fin de la partie précédente en développant une notion épistémique de *conséquence réflexive*. Cette notion me permet de proposer une réponse cohérente au problème de la stabilité et de donner une interprétation déflationniste du rôle théorique des axiomes récursifs pour la vérité.

Chapitre 5

Réflexion et rationalité

Pourquoi j'ai décidé d'être cohérent, à la réflexion

> [...] je me résolus de feindre que toutes les choses qui m'étoient jamais entrées en l'esprit n'étoient non plus vraies que les illusions de mes songes. Mais aussitôt après je pris garde que, pendant que je voulois ainsi penser que tout étoit faux, il falloit nécessairement que moi qui le pensois fusse quelque chose ; et remarquant que cette vérité, je pense, donc je suis, étoit si ferme et si assurée ...
>
> Descartes, *Discours de la méthode*, Quatrième partie.

L'argument de Shapiro et Ketland pour le pouvoir explicatif de la notion de vérité présenté au chapitre 4 faisait un usage essentiel de la notion de conservativité, et c'est sur ce point que j'ai cherché à prendre l'argument en défaut. J'ai montré d'une part que, comme critère de démarcation des emplois explicatifs, appliqué aux extensions aléthiques de théories *schématiques*, la non-conservativité n'étayait pas les conclusions inflationnistes de Shapiro et Ketland ; et que, d'autre part, le critère ne rendait pas justice à l'intuition déflationniste du caractère à la fois épistémiquement indispensable (pour l'expression de certaines généralisations) et non-explicatif des emplois du prédicat de vérité.

Dans le présent chapitre je laisse momentanément de côté la question de savoir quelle est l'interprétation correcte des preuves de cohérence par la vérité. Si pour un sujet acceptant une théorie donnée la preuve de la cohérence par la vérité articule des raisons d'accepter que cette théorie est cohérente fondées en partie sur son acceptation de certaines lois aléthiques et en partie sur le contenu de la théorie

elle-même, je voudrais m'interroger sur un autre genre de raisons que peut avoir un tel sujet d'accepter une telle conclusion. Il y a un sens intuitif dans lequel, me semble-t-il, si un sujet accepte un théorie donnée A, alors il *doit* accepter que A est cohérente, et cela pour des raisons relevant de sa compréhension sa propre activité théorique, indépendamment du contenu de la théorie qu'il accepte et de toute réflexion sur la notion de vérité. Je vais essayer de donner un sens précis à l'idée qu'il existe une justification rationnelle *par défaut* d'accepter que ce que l'on accepte est cohérent, et que cette justification n'a rien à voir avec le contenu de la théorie elle-même, pas plus qu'avec la question de sa vérité. Ma conclusion sera qu'un sujet qui accepte une théorie donnée peut bien en effet justifier, en un certain sens, son acceptation de la cohérence de cette théorie, en employant des ressources conceptuelles non aléthiques, mais qu'une telle justification est d'une nature et d'une portée épistémologique différentes de celle qui se fonde sur la preuve de la cohérence par la vérité. Cette dernière, mais non la première, constitue une *explication* de la cohérence et peut fonder la *croyance* en la cohérence de la théorie.

Le chapitre est divisé en quatre parties. Dans la première je reviens brièvement, dans une perspective épistémologique, sur la preuve canonique de la cohérence d'une théorie mobilisant les principes aléthiques tarskiens. Je précise les conditions sous lesquelles une telle dérivation vaut justification, pour un sujet, à accepter que cette théorie est cohérente et j'introduis le fil des idées qui seront développées dans la suite. Dans une seconde partie, j'introduis en la motivant la notion d' *acceptation* d'une théorie, en insistant sur certaines de ses propriétés pertinentes pour les fins que je poursuis. Dans la troisième parti j'essaie de voir si les outils classiques, bayésiens, d'étude de la logique de la *décision* permettent de rendre compte de l'idée que, quoique la cohérence A ne soit pas une conséquence déductible de A, la *décision d'accepter* la cohérence s'impose à celui qui a *décidé d'accepter* la théorie, indépendamment de son contenu. Néanmoins il apparaîtra que, sur un point, cet argument fondé sur la méthode des *Dutch Book*, ne donne pas autant que l'on en attend : car pour *montrer* que la première décision contraint rationnellement la seconde, il n'est pas clair que nous puissions nous dispenser de la notion de vérité. Dans la quatrième et dernière partie j'aborde donc le problème sous un autre angle en discutant de la notion de rationalité elle-même. Je considère les normes auxquelles est soumise l'action d'accepter et distingue parmi elles la norme de cohérence. Puis je me fais l'avocat de l'idée qu'un principe général de *Responsabilité en première personne* est un principe constitutif de notre rationnalité : Si un agent rationnel

accepte une théorie X (et le temps qu'il accepte X), il doit accepter « X est accept*able* ».

5.1 L'argument de Shapiro pour le pouvoir explicatif de la vérité

Nous avons vu au chapitre 3 par quel argument informel Shapiro a soutenu que le concept de vérité avait un pouvoir explicatif.[1] Nous avons vu alors que le passage d'une preuve de la cohérence de A dans la métathéorie à la preuve de l'énoncé du langage-objet G (ou coh_A, pour ce qui importe) n'était pas trivial. En substance, ce passage requiert que l'on enrichisse les principes de preuves que nous acceptons dans le langage de A. Mais laissons ce point de côté et considérons l'argument de Shapiro jusqu'à la conclusion, *dans le métalangage*, que A est cohérente. N'avons-nous pas là un exemple simple *d'explication* ? [2] Dans toute la suite de ce chapitre, quand je parlerai de la cohérence d'une théorie, et de la possiblité de prouver cette cohérence, ce ne sera plus spécifiquement de l'énoncé standard qui « encode » cette proposition dans le langage-objet (quand il existe) que je parlerai. D'une manière générale, toute la discussion qui suit doit être comprise comme indépendante de la nature de la théorie-objet. Celle-ci, je l'appellerai génériquement A, pourrait être une théorie physique ou biologique formalisée, ou une théorie mathématique interprétée mais trop peu expressive pour « encoder sa propre syntaxe », ou encore, bien sûr, l'arithmétique, la théorie des ensembles etc. Enfin, pour faciliter les

[1] Pour la commodité du lecteur, je rappelle la citation ici :

> Retournons à notre théorie arithmétique A et à son énoncé de Gödel G (ou Coh). Supposons qu'un professeur de logique asserte que G est vrai, et qu'un étudiant étonné [*puzzled*] demande une explication. L'étudiant croit l'affirmation du professeur que G est vrai, mais il veut qu'on lui montre pourquoi il est vrai. L'étudiant veut quelque chose comme une preuve convaincante ou une preuve explicative. La réponse naturelle est de faire remarquer que tous les axiomes de A sont vrais et que les règles d'inférence préservent la vérité. Il suit que « 0=1 » n'est pas un théorème et donc A est cohérente. L'énoncé de Gödel est équivalent à la cohérence de A. Il me semble que cette version informelle de la dérivation de Coh et G est une bonne *explication* s'il en est. L'argument montre pourquoi G est vrai. Faire remarquer que Coh et G sont (après tout) des conséquences logiques ou sémantiques de la théorie originale A ne sera d'aucune aide au déflationniste parce que c'est le fait qu'il s'agissait d'expliquer. Notre étudiant veut savoir pourquoi G est vrai -ou pourquoi G est une conséquence- et le passage par la notion de vérité fournit cette explication.(Shapiro (1998*b*), p.505).

[2] Sans non-conservativité relativement à la théorie de départ, donc.

références ultérieures, j'appellerai l'explication de la cohérence de A proposée par Shapiro la *preuve par la vérité* de la cohérence de A.

Bien entendu, cette preuve de la cohérence de A ne nous justifie à accepter la cohérence de A que pour autant que nous sommes justifiés à accepter les axiomes de A eux-mêmes et les axiomes pour la vérité et la syntaxe. Par conséquent cette preuve ne peut convaincre de la cohérence de la théorie qu'un sujet convaincu de la vérité de la théorie elle-même et donc d'un certain nombre de faits qui sont décrits ou prédits par la théorie elle-même. Si la théorie A est la théorie de la gravitation de Newton, la preuve de sa cohérence sera fondée en partie sur des hypothèses physiques, par exemple sur la valeur de la constante universelle de gravitation. Pour la même raison, si le sujet pense que la théorie en question est fausse - non pas absurde, mais simplement fausse -, ou s'il cherche, à la façon de Hilbert, à justifier son usage tout en adoptant une attitude instrumentaliste ou anti-réaliste relativement à cette théorie, la « preuve par la vérité » ne lui donnera pas ce qu'il en attend.[3] Puisque figurent parmi ses prémisses les axiomes de A elle-même, la « preuve par la vérité » de la cohérence de A n'est pas de nature à augmenter ma confiance en la théorie A.

5.1.1 Une variation sur l'argument de Shapiro

Pour clarifier ce que j'ai en vue dans ce chapitre, je vais commencer par présenter une variation l'argument de Shapiro. Imaginons que nous sachions qu'agent rationnel de notre connaissance, appelons-le Pierre, croit tous les théorèmes d'une théorie A. Supposons de plus que nous sachions que Pierre est parfaitement rationnel en un sens qui implique que ses croyances sont cohérentes[4] et qu'il accepte toutes les conséquences logiques de ses croyances. Considérons alors la « preuve » suivante de la cohérence de A :

> Pierre croit tous les théorèmes de A. « $0 \neq 1$ » est un théorème de A.
> Donc Pierre croit « $0 \neq 1$ ». Puisque Pierre est parfaitement rationnel

[3] Puisque figurent parmi ses prémisses les axiomes de A elle-même, la « preuve par la vérité » de la cohérence de A n'est pas de nature à augmenter ma confiance en la théorie A. Il ne s'agit donc par d'une preuve de cohérence du genre de celles que Hilbert cherchait et qui auraient eu, si elles avaient été possibles, une valeur fondationnelle. Il s'agissait en effet pour Hilbert de donner dans une théorie, sur laquelle nous avions peu de raisons de nourrir des doutes, une preuve de la cohérence de théories *plus fortes* sur la cohérence desquelles nous avions des raisons de nourrir des doutes.

[4] La cohérence des croyances peut sembler être une condition de rationalité extrêmement forte. Néanmoins, c'est une condition de rationalité en général considérée comme minimale dans les études sur la rationalité idéalisée, par exemple dans le cadre bayésien.

> (cohérent), il ne croit pas « 0 = 1 ». Donc « 0=1 » n'est pas un théorème de A, et donc A est cohérente. L'énoncé de Gödel est équivalent à la cohérence de A etc.

Nous pouvons parvenir à la même conclusion à partir de prémisses légèrement plus faibles, avec un peu de raisonnement arithmétique, comme Shapiro :

> Pierre croit tous les axiomes de A et croit toutes les conséquences de ses croyances. Donc il croit tous les théorèmes de A.

La fin de la preuve est comme précedemment, en bref : puisque Pierre est parfaitement rationnel, A est cohérente.

Ces « preuves » de cohérence ne font nullement appel concept de vérité. Et Pierre peut avoir des raisons extravagantes de croire les axiomes de A. Pierre en est peut-être venu à croire les axiomes de A pour des raisons esthétiques, ou par simple mimétisme. Ou peut-être Pierre est-il conventionaliste, et pense-t-il que A n'est qu'un « choix de langage ».[5] Quelles ques soient ses raisons de croire, d'accepter ou d'asserter les axiomes de A, tant que nous savons que Pierre est parfaitement rationnel dans ses croyances (assertions, etc.), nous pouvons inférer de son acceptation ou de son assertion des axiomes de A que A est cohérente. Et le concept de vérité n'a aucune part dans cette explication.

La preuve précédente de la cohérence de A reposait sur certaines informations concernant la structure des croyances de Pierre et son acceptation des axiomes de A (pour faire bref, appelons cette preuve la preuve-par-Pierre de la cohérence) et non sur le fait que les croyances de Pierre étaient vraies, ni qu'elles étaient dans une certaine relation de correspondance avec des faits arithmétiques, ni encore sur le fait que Pierre lui-même pouvait avoir un concept de vérité. Néanmoins, bien entendu, la conclusion de la preuve-par-Pierre de la cohérence de A n'est convaincante que dans un scénario dans lequel ses prémisses sont elles-mêmes supposées être convaincantes. Mais nous ne connaissons pas Pierre, et plus généralement nous n'avons aucune justification *a priori* pour penser qu'un certain individu existe, croyant tous les axiomes d'une certaine théorie A et parfaitement rationnel ; et par conséquent les hypothèses de la preuve-par-Pierre ne font pas de cette dérivation une explication convaincante, pour nous ou pour l'étudiant de Shapiro, de la cohérence de A. Si la preuve de la cohérence par la vérité en constitue, elle, une « explication

[5] Auquel cas il vaudrait mieux dire que Pierre accepte l'objet formel qu'est la théorie A comme outil d'inférence, plutôt qu'il ne la *croit*. Je reviendrai sur ce point.

convaincante », c'est que ses prémisses *ont*, pour nous et pour l'étudiant de Shapiro, une *évidence* qui manquait à celles figurant dans notre parodie d'explication : dans le scénario envisagé par Shapiro où la théorie A a été raisonnablement choisie, l'étudiant et nous *croyons* les axiomes de A, et croyons les lois *a priori* de la vérité pour le langage de A. Et de ces prémisses il est possible de *déduire* la cohérence de A.

Nous avons vu que le cœur logique de la preuve de la cohérence par la vérité était le suivant : à partir de A et des lois tarskiennes de la vérité, on peut prouver que les théorèmes de A appartiennent à un ensemble cohérent d'énoncés (ici l'ensemble des énoncés vrais). Quelle autre justification l'affirmation que tous les théorèmes de A appartiennent à un ensemble cohérent d'énoncé peut-elle avoir ? Comme nous l'avons dit, il serait suffisant pour la justifier de savoir qu'ils sont acceptés par un agent rationnel, au sens par exemple, de la théorie classique, bayésienne, du choix rationnel. Nous avons cherché autour de nous un tel sujet, et nous ne l'avons pas trouvé. Mais peut-être n'avons-nous pas cherché au bon endroit, et peut-être avons-nous trop demandé. : peut-être n'est-il pas nécessaire de *savoir* qu'un tel sujet est rationnel pour justifier l'acceptation de la cohérence d'une théorie, s'il s'agit d'une théorie que *nous* acceptons et si le sujet n'est autre que *nous-même*. La question qui m'intéresse ici est la suivante : notre acceptation de A justifie-t-elle l'acceptation *par nous* du fait qu'elle est acceptée par un agent rationnel ? Que nous croyions ou pas la théorie A, que nous ayons adopté vis-à-vis d'elle et de son langage une attitude réaliste, descriptiviste, ou au contraire anti-réaliste, intrumentaliste, ne devons-nous pas, *si* nous l'acceptons pour les fins d'une activité rationnelle, accepter du même coup, de ce seul fait, que la théorie en question est cohérente ? La cohérence de nos *décisions*, à défaut du contenu des théories que nous acceptons, ne nous contraint-elle pas à accepter que ce que nous acceptons est cohérent, quand bien même nous n'en avons pas de *preuve*, quand bien même nous n'aurions aucun *indice* de cette cohérence ? C'est donc cette idée que, dans ce chapitre, je voudrais essayer d'explorer : l'idée qu'un sujet possède une justification rationnelle, mais pragmatique et défaisable, qui n'est pas fondée sur le contenu de la théorie elle-même et ne requiert pas le concept de vérité, pour accepter la cohérence d'une théorie qu'il accepte.

5.2 Pourquoi accepter ?

J'ai parlé d'*accepter* une théorie et de la possibilité de justifier l'*acceptation* de la cohérence, en réfléchissant sur la logique de nos décisions. Il y a au moins trois raisons raisons pour lesquelles j'emploie le terme d'*acceptation* et non simplement celui de « croyance« , de « croyance justifiée » ou même de « connaissance ». La première est que la contrainte rationnelle dont je cherche à rendre compte lie aussi bien un sujet qui adopte (pour le dire vaguement) une théorie parce qu'il la croit vraie qu'un sujet ayant une attitude instrumentaliste vis-à-vis de la théorie qu'il adopte et qui, par conséquent, ne *croit* pas la théorie en question. La seconde est que ce qui m'intéresse, le moyen par lequel il me paraît plausible de soutenir la thèse précédente, est l'étude de la rationnalité des *décisions* ; la croyance ou la connaissance sont des attitudes involontaires qui ne procèdent d'aucune décision de croire ou de connaître.[6] La troisième est lié au type d'attitude vis-à-vis de la proposition de la cohérence pour laquelle on peut envisager qu'elle puisse être justifiée en l'absence d'indice de la vérité de la proposition en question. Et c'est aussi, entre autres, pour ces mêmes raisons pour lesquelles elle m'intéresse ici que la notion d'acceptation a trouvé une place en épistémologie et en philosophie des sciences. Elle a servi à qualifier la relation de l'anti-réaliste aux théories qu'il emploie, elle est utile pour rendre compte de l'idée de *décisions* scientifiques, pour rendre compte de l'idée qu'il y a des devoirs épistémiques, elle a servi enfin, de façon plus problématique sans doute, pour rendre compte de l'idée de *justification par défaut* qui semble nécessaire pour rendre compte de la possibilité d'un fondement de nos justifications dans la tradition fondationaliste d'analyse de la connaissance comme croyance vraie justifiée. Je me propose dans cette section de revenir sur ces trois points en précisant la notion d'acceptation que j'ai en vue.

Le terme d'acceptation est polysémique, même si d'une façon générale l'importance d'une distinction entre l'attitude de croyance et l'attitude d'acceptation (ou une variété de telles attitudes) a été soulignée à plusieurs reprises dans la littérature. Elle a été thématisée par exemple dans Cohen (1989) et un cas particulier d'une telle distinction a été défendu dans Fraassen (1980). La notion d'acceptation spécifique développée par van Fraassen[7], par exemple, est telle qu'accepter une théorie engage, mais engage seulement, à tenir pour vraies ses conséquences empi-

[6] Il faudrait nuancer : pensons par exemple au pari de Pascal.

[7] Dans Fraassen (1980), p.11-12

riques. Une théorie scientifique, pour van Fraassen, est une théorie semi-interprétée pour la quelle la question de la vérité ne se pose pas à proprement parler. Les termes théoriques n'y ont qu'une signification partielle, celle qui leur est conférée indirectement par leur conséquences observables, et le scientifique anti-réaliste l'utilise simplement comme un outil d'inférence. Van Fraassen décrit cette attitude comme une attitude d'acceptation avec les engagements vis-à-vis des conséquences de la théorie que cette attitude impose. Mettre l'acceptation, en ce sens, plutôt que la croyance, au centre de l'activité scientifique, est donc un moyen pour van Fraassen de développer sa position antiréaliste.[8] Mais c'est plus généralement toute une famille d'attitudes distinctes que révèle l'observation de nos états mentaux : l'acceptation au sens de tenir-pour-vrai en est une[9], celle de van Fraassen un autre, et l'on en trouverait autant que de variations possibles associées à des *buts* et des *normes* différents.

Malgré cette variété, on peut distinguer la croyance de l'acceptation par certains traits généraux. Nous suivons ici la comparaison utile proposée par Pascal Engel (1998)[10], en la simplifiant sur quelques points secondaires. P. Engel distingue quatre caractères principaux de la croyance :

1. Les croyances sont involontaires et ne sont pas normalement sujettes à un contrôle volontaire direct,
2. Les croyances visent le vrai,
3. Les croyances sont informées par ce qui compte comme « indice »ou éléments probants [*evidence*] pour ce qui est cru,
4. Les croyances sont sujettes à un idéal d'intégration ou d'agglomération,

et y oppose point par point la notion d'acceptation :

1. L'acceptation est volontaire ou intentionnelle, contrairement à la croyance,
2. L'acceptation ne vise pas la vérité mais l'utilité ou le succès (on peut accepter ce que l'on croit faux),

[8] Le même terme s'impose lorsque l'on veut décrire l'attitude d'un conventionnaliste en mathématique ou en logique. Quelle est l'attitude d'un Carnap, disons, relativement au langage et aux énoncés de la logique ? Ces énoncés n'ont pas de contenu propre, et s'agit pas de les croire mais seulement de choisir des les accepter ou non.

[9] Voir Engel (1998) sur la distinction entre croire et tenir pour vrai.

[10] Engel (1998), p.143-144

3. L'acceptation n'est pas nécessairement informée par les preuves ou les indices,
4. L'acceptation n'est pas régulée par un idéal d'intégration rationnelle au sens où les croyances le sont.[11]

Avec ces caractéristiques, la notion d'acceptation reste largement ouverte. Même dans le cadre restreint de l'enquête scientifique, plusieurs types d'attitudes satisfaisant aux critères qui viennent d'être donnés sont possibles.[12] Les points 2, 3 et 4 ne caractérisent l'acceptation que négativement, au sens où ils signalent que certaines règles gouvernant la croyance ne s'appliquent pas nécessairement à la notion d'acceptation. Cependant, il n'est pas interdit de penser que certaines propriétés présentées comme caractéristiques des croyances s'appliquent en fait, *mutatis mutandis*, à une notion d'acceptation donnée. Il se peut, en particulier, que l'unique but poursuivi par les agents soit la recherche de la vérité ou, en des termes moins chargés, de ce qui est le plus avantageux *en termes purement épistémiques* d'accepter.[13] Par conséquent (point 2), l'utilité ou le succès, dans ce contexte, se mesurent aussi, en autres, à l'aune de la vérité. De même, à l'intérieur de ce contexte restreint, on pourrait soutenir que l'acceptation *est* contrainte par les raisons *probantes* que sont les indices ou les preuves, même si, en leur absence l'acceptation est toujours possible en vertu d'autres types de justifications. En ce sens, il est possible que, dans un « code » qui régirait les normes de l'acceptation figure une règle comme : « ne pas accepter p en présence d'indices que non-p ». On pourrait également soutenir que l'élaboration du corps de la science, compris comme un ensemble accepté de vérités, doit également viser un certain idéal d'intégration. La notion d'acceptabilité

[11] Un point que je n'ai pas mentionné est que les croyances sont susceptibles de degrés, tandis que l'acceptation, à strictement parler, ne l'est pas. Toutefois, comme le remarque P. Engel, on peut quantifier indirectement l'acceptation, en associant à l'action d'accepter l'espérance d'utilité qui lui est associée, ce qu'en anglais on appelerait « acceptance worthiness ». Dans le cas très particulier dans lequel il est question de l'acceptation dans le cadre de l'enquête scientifique dont le but est la connaissance, on pourrait voir le degré auquel on accepte une théorie ou une hypothèse simplement comme le degré auquel nous nous croyons justifié à le faire étant donné que notre but est la connaissance. Nous supposerons ici que l'on peut accepter à des degrés divers, exactement comme l'on peut croire à des degrés divers, et que les méthodes bayésiennes d'étude de la cohérence des croyances s'applique identiquement à la cohérence de ce qui est accepté. Sur la question de la place de la notion d'acceptation dans l'épistémologie bayésienne, on pourra consulter la discussion critique de Maher (1993) et les références qui s'y trouvent.

[12] Pour une discussion plus approfondie des différentes notions d'acceptation, nous renvoyons à nouveau à Engel (1998), Cohen (1989), et aux essais dans Engel (2000).

[13] La question de savoir ce qu'il est avantageux épistémiquement d'accepter étant elle-même une question ouverte. La simplicité d'une théorie est-elle un avantage épistémique ? Sa fécondité ? etc.

par un agent que nous avons en vue est principalement celle qui est à l'œuvre dans ce genre de contexte scientifique idéalisé.

Si la notion d'acceptation ouvre la voie à la possibilité d'autres types de justifications que celles qui comptent ordinairement comme la justification d'une croyance, il est important également pour notre projet d'insister sur le fait que l'acceptation, au contraire de la croyance, est volontaire. Dans le sens du terme que j'ai en vue, l'acceptation est spécifiquement le produit d'un examen critique, d'une délibération (parfois conduite *in petto*, parfois au sein d'une communauté) : c'est un acte *réfléchi*. À l'issue de ce processus de délibération rationnelle, il y a une *décision d'accepter* ou non une proposition et, c'est le point sur lequel je veux maintenant insister, précisément parce qu'elle procède d'une décision l'acceptation engage notre *responsabilité*. Tant que nous acceptons une proposition, nous devons répondre de la conformité de notre décision à une certaine norme, en général publique : je ne me regarde comme ayant le droit d'accepter p que pour autant que j'accepte qu'une certaine norme d'acceptabilité a été satisfaite par moi.

Maintenant que j'ai précisé un peu la notion d'acceptation que j'ai en vue, et avant d'aller plus loin, je veux faire deux remarques supplémentaires pour motiver son introduction. La première me permettra de préciser une certaine conception de la notion de « justification », et la seconde d'illustrer la possibilité de « justification sans preuve ».

La notion de justification peut être comprise en plusieurs sens, mais c'est une notion classique, *interne*, introspectible de justification qui m'intéresse dans le contexte présent, correspondant à l'idée selon laquelle la croyance (traditionnellement) est justifiée lorsqu'elle est fondée en raison.[14] En ce sens, une idée naturelle est que nous sommes justifiés croire ce que nous croyons s'il nous est *permis* de le faire une fois qu'ont été satisfaites un certain nombre *d'obligations*. Autrement dit, quelque chose comme une « éthique de la croyance », selon l'expression de William Clifford, doit gouverner nos pratiques spécifiquement épistémiques, à la façon dont l'éthique ordinaire doit gouverner l'action de façon plus générale. Cette idée d'un code de conduite épistémique est au cœur de toutes les tentatives d'édicter des *règles pour la direction de l'esprit* en vue de fonder l'entreprise de la connaissance, et de toute réflexion sur la méthode scientifique. Or justement, l'idée même que

[14]Par opposition à une notion « externaliste » de justification selon laquelle, par exemple, un agent est justifié à croire que p si les mécanismes qui sous-tendent la formation de sa croyance que p sont « fiables », en un sens à préciser.

nos croyances pourraient être contraintes par certaines obligations dictées par une réflexion méthodologique semble supposer que nous pouvons exercer un contrôle sur nos croyances. Si, conformément à la maxime kantienne, *à l'impossible nul n'est tenu*, la possiblité d'une éthique de la croyance, de règles relatives à ce qu'il est obligatoire ou permis de croire dans telles ou telles circonstances, suppose que la croyance est soumise à notre volonté. Un problème est que, selon toute vraissemblance, elle ne l'est pas, du moins *pas directement*. Je ne peux pas croire, par la seule volonté, que la lune est verte, quand bien même on m'offrirait une somme avantageuse pour le faire, ou aussi grande ma motivation soit-elle, ni, inversement, *cesser* de croire que la terre est plus ou moins sphérique. Cet argument, connu sous le nom d' « argument du volontarisme »[15], semble obliger à reformuler toute conception déontique de la justification. Tout au plus avons-nous un contrôle, dans une certaine mesure, sur certaines *conditions de formation de nos croyances*. L'acceptation, au contraire, nous l'avons dit, procède d'un acte volontaire et la conception déontique de la justification de l'acceptation est immunisée contre l'argument présenté à l'instant si l'attitude qui est l'objet de nos justifications est l'acceptation. L'analyse internaliste et déontique classique de la justification[16] paraît s'appliquer plus directement à une forme d'acceptation réfléchie, qu'à la croyance ordinaire, et l'idée même que « l'homme de science » a le devoir régler son esprit selon les exigences de la méthode suggère que ce n'est pas la justification de *croyances* qui est primitivement en jeu dans l'enquête scientifique : celui qui s'est engagé sur cette voie doit parfois accepter des choses qu'il ne croit pas.[17]

J'en viens à mon second point. Un exemple fameux de règle d'« éthique de

[15]Développé principalement par Alston (1985) et Alston (1988)

[16]Telle qu'elle est par exemple présentée dans Chisholm (1977). Voir aussi Engel et Dutant (2005) pour une présentation des conceptions internalistes de la justification.

[17] Mentionnons néanmoins qu'il existe une littérature importante sur la question de savoir si la notion de justification est une notion déontique. D'une part cette question est liée au débat entre internalisme et externalisme : Plantinga (1993) chap.1 et Goldman (1999) ont soutenu l'idée que le principal argument en faveur des conceptions internalistes de la justification est lié au caractère *prima facie* déontique de la notion de justification épistémique des croyances. L'argument du volontarisme contre la conception déontique de la justification n'a pas convaincu tous les philosophes. Contre les attaques d'Alston, Steup (1988) rappelle que les croyances sont souvent *indirectement* volontaires. Dans Steup (2000) il va jusqu'à discuter en profondeur la question de savoir ce qu'il faut entendre par *contrôle volontaire* des croyances et *action libre*. La question épistémologique apparaît alors lié au problème classique de la possiblité de la liberté de l'action. Parmi d'autres options possibles pour désamorcer l' « argument du volontarisme », mentionnons aussi que Feldman (2000) nie que la question du caractère volontaire ou non de nos croyances ait une réelle pertinence épistémologique.

la croyance », que l'on pourrait reformuler en termes d'éthique de l'acceptation épistémique, est la règle que Clifford (Clifford (1879)) reprend à Locke, en substance :

Proportionnez vos croyances aux éléments probants [*evidence*] disponibles

Dans une entreprise dont le but est la connaissance, est-il nécessaire de s'astreindre à une telle règle ? Il existe différents arguments pour penser que non, mais je voudrais esquisser un argument pour l'affirmation plus forte qu'il n'est *pas possible*, en ce qui concerne l'acceptation, que ne soit rationnellement justifiée que l'acceptation des propositions pour lesquelles nous possédons des preuves ou des indices de leur vérité. Cet argument est implicite dans la réponse que donne Crispin Wright[18] aux défis sceptiques. Le point de départ est la considération des défis sceptiques radicaux lancés contre, sinon la possibilité de toute connaissance, du moins contre la possibilité de *justifier* la moindre connaissance. En un mot, la stratégie du scepticisme radical, celle mis en œuvre par Descartes au moment du doute hyperbolique, est de montrer qu'une certaine régression infinie est inévitable : pour être justifié à accepter que p, il faudrait que je sois justifié à accepter que p est justifié, et la recherche de justifications (preuves, indices etc.) qui permettrait de fonder ne serait-ce que l'acceptation d'*une* proposition n'aurait pas de fin. Par conséquent, il semble que l'acceptation d'aucune proposition ne soit justifiée. Mais c'est absurde. Par conséquent puisque, si nous devions tout justifier, nous entrerions dans une régression sans fin, et que nous ne pourrions par conséquent pas être justifiés à accepter quoi que ce soit, il faut admettre que l'acceptation de certaines propositions est justifiée *par défaut, en l'absence de preuves ou d'indices de leur vérité.*[19] Il est possible que nous ne soyons pas justifiés à *croire* de telles propositions, mais leur acceptation rend la croyance justifiée possible : si je suis justifié à accepter que

(*) je ne suis pas en ce moment même le jouet d'un malin génie

alors ma perception de cette table devant moi justifie ma croyance qu'il y a là une table, quoique je ne dispose présentement d'aucune preuve de (*).

Bien entendu, il y a la difficulté de principe de délimiter le champ de ces propositions (et de ces situations) pour l'acceptation desquelles il rationnellement permis

[18] Wright (2004*b*)

[19] J'utilise les termes d'« éléments probants », d'« indice » et de « preuve » en combinaison pour rendre un équivalent du terme anglais *evidence*.

de se passer d'indice de leur vérité. Ces propositions dont l'acceptation est justifiée par défaut en l'absence d'indice sont d'un type particulier ; ce sont celles, par exemple, sans l'acceptation desquelles aucun entreprise de connaissance ne serait même possible, ce que Crispin Wright appelle « les pierres de touche » d'un projet épistémique donné. Cela pourrait être la croyance que je suis bien en train d'observer ce que je suis en train d'observer tandis que je suis en train d'effectuer une vérification visuelle, ou la croyance que les inférences logiques, sans lesquelles aucune pensée articulée n'est possible, préservent les justifications. Nous n'avons pas de preuve spécifique disponible pour ces croyances, et pourtant elles semblent rationnelles. C'est le genre d'idée développée par Crispin Wright[20], qui préfère distinguer deux espèces de justifications [*warrant*], réservant le terme de « justification » [*justifications*] aux justifications « évidentielles » à proprement parler, et parlant de *permission* [*entitlement*], à propos de ces justifications *par défaut*, *a priori*, et *défaisables*.[21] Mais l'idée que nous sommes justifiés à accepter, ou à croire, certaines propositions sans preuve, n'est pas une thèse nouvelle. C'était par exemple déjà la substance de la célèbre réponse que W. James avait faite à W. Clifford[22] dans laquelle il soutient qu'il peut être bénéfique de croire au-delà de ce qui est permis par la seule évidence disponible, distinguant utilement deux buts, dans une certaine mesure concurrents, à l'œuvre dans l'entreprise de connaissance : le désir de croire le maximum de vérités et le désir d'éviter le maximum d'erreurs. La règle édictée par Clifford, selon James, ne procéderait que d'une précellence arbitraire accordée à la seconde de ces maximes sur la première.

Maintenant que j'ai un peu précisé la notion d'acceptation que j'ai en vue, que j'ai illustré qu'il était plausible qu'un sujet soit rationnellement justifié, par défaut, à accepter une proposition sans posséder de preuve spécifique de cette proposition, et que j'ai rappelé qu'en ce sens d'acceptation nous étions responsables de ce que nous acceptons, j'en viens à ma première tentative pour montrer qu'est justifiée rationnellement l'acceptation de la cohérence d'une théorie par un sujet acceptant la théorie en question.

[20] En particulier dans Wright (2004*a*)et Wright (2004*b*)

[21] La notion d'*entitlement* en jeu ici est proche de celle présentée par Burge (1993). On trouve un emploi un peu différent dans Dretske (2000). Pour un panorama des recherches récentes sur la notion de justification, on pourra consulter Pryor (2001).

[22] Dans (James (1956).

5.3 De l'irrationalité de parier contre sa propre cohérence

5.3.1 Dutch Book et rationnalité

Une réponse bayésienne classique à la question de savoir à quelles conditions l'ensemble des croyances d'un agent est rationnel est la suivante : les croyances (avec leur degré associé) sont cohérentes si elles sont conformes au calcul des probabilités. Mais il existe également un autre notion bayésienne, ou quasi-bayésienne, de rationnalité, voisine mais distincte de la précédente, celle pour laquelle Ramsey avait imaginé le test suivant : un agent entretient un ensemble de croyances incohérent (i.e. est irrationnel) si ces croyances le rendent vulnérable à un *dutch book*.[23] Un agent est vulnérable à un dutch book si ses croyances sont telles, qu'il est juste pour lui d'accepter un ensemble de paris dont l'issue certaine, connaissable *a priori* et sans information privée, sera un perte nette pour lui. On considère parfois que l'intérêt de prouver la vulnérabilité d'un agent à un *dutch book* réside dans le fait que c'est un moyen commode de montrer que les degrés de croyances de l'agent ne sont pas conformes au calcul des probabilités. Dans cette perspective, c'est la conformité des degrés de croyances au calcul des probabilités qui est la norme de la rationalité[24], et l'usage de la notion de (in)vulnérabilité à *dutch book* , adéquatement précisée, est seulement dérivée, et justifiée dans les cas où peut prouver certains résultats de correspondance entre les deux notions. Notre approche ici est différente. Nous regardons la vulnérabilité à un *dutch book* comme primitive, et faisons l'hypothèse qu'elle renferme une authentique explication du caractère défectueux de certains ensembles d'états doxastiques.[25] Il nous semble intuitivement plausible que : « Il

[23] Ramsey (1931), p.183 :

> Having any definite degree of belief implies a certain measure of consistency, namely willingness to bet on a given proposition at the same odds for any stake, the stakes being measured in terms of ultimate values. Having degrees of belief obeying the laws of probability implies a further measure of consistency, namely such a consistency between the odds acceptable on different propositions as shall prevent a book being made against you.

Pour un panorama critique des arguments par dutch book, voir p. ex. Hajèk (2008).

[24] C'est l'approche bayésienne connue sous le nom de « probabilisme ».

[25] Nous ne sommes pas seuls à endosser cette approche, et la force intuitive de formes variées de *dutch book* est communément admise. En un sens, l'approche de Fraassen (1984) en témoigne, qui commence par présenter le problème qui l'intéresse sous ce jour, même s'il affirme pour finir que la vulnérabilité à un *dutch book. Nous allons y revenir.* ne saurait constituer à ses yeux un

est incorrect de conformer ses degrés d'acceptabilité envers des propositions à la distribution P *parce que* le faire nous expose à un *dutch book* » est une explication plausible du caractère rationnellement défectueux d'un état doxastique. Avec cette notion bayésienne, au sens large, de rationalité, je propose maintenant de regarder brièvement un emploi qu'en a fait van Fraassen[26] et qui s'apparente, par certains aspects, à celui que nous avons en vue.

5.3.2 Une justification non-évidentielle d'un principe de réflexion : van Fraassen

L'idée que, le temps que nous acceptons un certain bagage doxastique donné, une certaine logique de la *décision* contraint notre acceptation de principes supplémentaires pour la vérité desquels nous n'avons pas d'indice disponibles, en particulier certains principes réflexifs, n'est pas nouvelle. Cette idée de contrainte réflexive rationnelle s'apparente à une thèse mise en avant par Bas van Fraassen que je voudrais maintenant présenter brièvement.

Contre l'« évidentialisme » d'un Clifford, Bas van Fraassen prend le parti de James : l'enquête scientifique, et la construction de nos meilleures théories, pourrait exiger que nous adoptions parfois des propositions pour la vérité desquelles nous n'avons pas d'indice spécifique. Mais l'espace de ces propositions qu'il serait possible d'adopter, remarque van Fraassen, semble devoir obéir à certaines contraintes. L'exemple particulier qui retient l'attention de van Fraassen est un principe bien spécial, un principe relatif à notre rationalité diachronique : quelque absence d'indice que je puisse avoir *aujourd'hui* sur ce que sera *demain* mon degré de croyance dans une proposition H (supposée donnée), je ne suis pas libre aujourd'hui, étant donné mon degré de croyance actuel en H, d'adopter n'importe quel degré de croyance relativement à ce que sera mon degré de croyance en H demain. Plus formellement, le Principe de Réflexion dont van Fraassen se fait l'avocat est donc le suivant :

$$p_t(A|p_{t+x}(A) = r) = r$$

c'est-à-dire : le degré de croyance d'un agent au moment t en la proposition A sous l'hypothèse que son degré de croyance en A à $t + x$ sera r, doit être égal à r.

argument suffisant pour répudier les croyances d'un agent comme irrationnelles.

[26] Fraassen (1984).

Pour justifier informellement son principe, van Fraassen propose de le scénario suivant.[27]

> Soit H l'hypothèse en jeu - la théorie de l'évolution, disons - et soit E la proposition que Bas van Fraassen croira pleinement que H (dans un an, disons). Pour fixer les idées, supposons que $P(E)$ - mon degré de croyance que E sera le cas - est de 0,4 et $P(\neg H \wedge E)$ - mon degré de croyance que j'en viendrai à croire pleinement, à tort, H - est égal à 0,3.

Je suis donc dans un situation où mes états doxastiques sont en violation du principe de Réflexion. Supposons maintenant qu'un bookmaker me propose l'ensemble de paris suivants :

> le premier paie 1 si j'en viens à croire H et H est en fait fausse - il demande 0,2 pour celui-ci. Le second paiera 0,5 si je n'en viens pas à croire H, et il m'en demande 0,3. Le troisième paie 0,5 si j'en viens vraiment à croire H ; celui-là coûte 0,2.

Tous ces paris sont acceptables étant donné l'état doxastique de l'agent décrit précédemment, et ils ont un coût total de 0,7. Mais maintenant de deux choses l'une :

> Dans un scénario, je n'en viens pas à croire H ; je gagne le second pari et perd les deux autres. Dans l'autre scénario, j'en viens à embrasser l'hypothèse ; maintenant je perds le second pari, dis moi-même [au bookmaker] que H est vraie, donc je n'obtiens rien pour le premier pari (quoique je reçoive une maigre pitance quand je lui revends un pari sur $\neg H$ pour presque rien), et je gagne le troisième.

Dans tous les cas, le calcul est simple, je subis une perte sèche. Si l'on accepte que la vulnérabilité à ce *dutch book* est un indicateur du caractère irrationnel des croyances de l'agent, alors il faut conclure que le principe de Réflexion (celui de van Fraassen mentionné ci-dessus) est un principe de rationnalité.

Je ne veux pas aller plus loin dans la présentation des arguments apportés par van Fraassen pour justifier son principe[28], et j'en viens directement aux leçons que

[27] Voir Fraassen (1984), p.237-238.

[28] En particulier sur l'analyse « volontariste » et non simplement « descriptiviste » qu'il doit donner du statut des jugements épistémiques du type « Il est probable qu'il pleuvra demain », en mettant en relief l'existence d'*engagements* contractés par celui qui énonce ce genre de proposition. Voir Fraassen (1984), p.250-255. J'aurai à venir à ce genre de considérations par la suite, mais dans le contexte plus simple, synchronique, qui est le nôtre.

j'en tire.

Le Principe de Réflexion de van Fraassen n'est pas identique au principe de réflexion épistémique que nous souhaitons défendre, mais l'esprit, sinon la lettre, de la remarque de van Fraassen, est applicable plus largement : une fois qu'un ensemble de propositions a été accepté par un agent rationnel alors, s'il réfléchit sur ce qu'il accepte il doit en tirer les conséquences normatives, à savoir non seulement que ces propositions ont été acceptées, mais encore qu'elles sont acceptables. Voyons comment ces considération pourraient s'appliquer au cas qui nous intéresse.

5.3.3 Qu'il faut parier sur sa propre cohérence

L'idée générale motivant la transposition de l'argument de van Fraassen au cas qui nous occupe est simple. Supposons par exemple que nous soyons conventionnalistes en matière d'arithmétique. L'arithmétique de Peano n'est à propos de rien, il s'agit simplement d'un choix de langage, et notre acceptation de PA est simplement une décision guidée par des considérations pragmatiques. Nous souhaitons néanmoins que nos décisions soient rationnelles : nous acceptons donc toutes les conséquences logiques de nos conventions et nous regardons nos choix conventionnels comme défaisables dans le cas où ils s'avèreraient incohérents. Clairement, le fait que nous ayons accepté PA n'implique pas que PA est cohérente. Nos décisions pourraient bien se révéler incohérentes pour finir. Pourquoi un conventionnaliste devrait-il accepter que PA est cohérente ? La réponse, analogue à celle de van Fraassen, est que celui qui décide d'adopter une théorie A prend une forme d'engagement, et que tout engagement d' « ordre supérieur » qui serait en contradiction avec le *sens* de cet engagement initial doit être regardé comme pragmatiquement incohérent : celui qui a accepté les théorèmes de A doit accepter qu'ils sont accept*ables*, et en particulier cohérents. C'est une affaire de *stabilité des décisions*, selon l'expression de Dubucs (2003).

Dans la situation de jeu que nous allons présenter, un joueur prend des paris portant sur la réalisation des certaines propositions contenues dans une théorie A et d'une proposition métathéorique concernant la théorie A elle-même, à savoir la cohérence A. La nature des propositions sur lesquelles portent les paris appellent une remarque. Tout d'abord, pour ce qui nous occupe, la nature de la théorie A et le langage dans lequel elle est formulée a peu d'importance : il peut s'agir aussi bien d'une théorie arithmétique que d'une théorie empirique, ou ce que l'on voudra.

En second lieu la proposition de la cohérence d'une théorie donnée est elle-même une proposition mathématique, peut-être arithmétique, ou syntaxique[29], en tout cas telle que, vraie ou fausse, elle l'est *nécessairement*. Mais ceci peut sembler poser un problème pour notre projet. En effet, dans le cadre de travail qui est habituellement celui de l'étude de la rationnalité idéalisée, *les vérités mathématiques sont supposées réalisées dans tous les états possibles de la nature*, si bien que le cadre de travail force à qualifier d'irrationnel un degré de croyance non nul en un énoncé mathématique faux, ou un degré de croyance strictement inférieur à 1 en une proposition arithmétique vraie. Mais si croire un énoncé mathématique faux, quel qu'il soit, est irrationnel, alors montrer qu'un agent qui accepte une théorie A, tout en acceptant pas qu'elle est cohérente, n'est pas rationnel, manque tout à fait son but : car la cohérence étant un fait mathématique, si A est cohérente alors il est de toute façon irrationnel d'en être moins que sûr.[30] Cette objection nous semble davantage participer d'une critique des limites du cadre de travail idéalisé ordinaire sur la rationnalité que d'une objection sérieuse à notre projet, fort modeste. C'est une question de philosophie des mathématiques qu'il ne nous est pas possible de traiter en profondeur ici que de déterminer si les croyances mathématiques fausses sont irrationnelles. Disons simplement que sur ce ce point, nous pensons avec d'autres qu'il n'en est rien et c'est ce que nous supposerons.[31] Dans ce qui suit nous considérerons que les théories mathématiques ne sont pas essentiellement différentes sur ce point des théories empiriques ordinaires et que rien ne s'oppose par principe à ce qu'un agent rationnel entretiennent des croyances mathématiques fausses.[32] Ces points étant précisés, voyons à présent pourquoi un agent acceptant une théorie A n'est pas libre relativement aux propositions métadoxastiques qu'il peut rationnellement

[29] Au sens de la syntaxe au chapitre 2.

[30] Je remercie Leon Horsten et Mikael Cozic pour avoir attiré mon attention sur cette objection.

[31] Voir par exemple Detlefsen :

> There is nothing inherently irrational (rational) *per se* about mathematical falsehood (truth) in my view (Detlefsen (1979), f.n. 1)

[32] Du point de vue d'un traitement formel, des aménagements au cadre de travail habituel seraient donc nécessaires, par exemple l'introduction d'états de la nature dans lesquelles les vérités mathématiques sont violées, et donc par exemple de certains types de « mondes possibles impossibles ». Du fait des complications posées par ce genre de formalisation, nous resterons ici à un niveau informel. Notons tout de même en passant que nous ne renonçons pas à l'idéalisation de l'omniscience logique des agents : nous maintenons qu'il est irrationnel d'avoir des degrés de croyances non nuls envers des contradictions logiques et des degrés de croyances inférieurs à 1 envers des vérités logiques. Pour fixer les idées, on supposera que la « logique » s'arrête à la logique du premier ordre.

accepter.

Un Dutch Book gödélien Imaginons que nous acceptions chaque axiome d'une théorie A au degré 1, mais que nous n'acceptions la cohérence de A qu'à un degré plus petit que 1, disons 0,5. Un bookmaker hollandais, appelons-le Kurt, nous propose d'acheter 0,4 un pari qui paye 1 si A n'est pas cohérente, rien sinon. Etant donné nos croyances, ce pari est acceptable, et même avantageux. Maintenant supposons que A est cohérente : alors le pari est perdu, nous perdons 0,4 et Kurt gagne 0,4. Si maintenant A *n*'est *pas* cohérente : alors nous gagnons 0,6 sur ce pari, mais puisque A n'est pas cohérente et que nous acceptons au degré 1 tous les théorèmes de PA, nous sommes vulnérable à un dutch book dont l'issue est une perte certaine de 1 pour nous. Au bout du compte, nous perdrions donc 0,4 et Kurt gagnerait à nouveau 0,4. Par conséquent, si notre degré d'acceptation des axiomes de A est 1, nous ne devrions pas accepter de payer pour un pari sur l'incohérence de A.

Il semble de prime abord que nous soyons parvenu à nos fins, puisque l'argument est supposé montrer qu'il est irrationnel d'accepter la cohérence de A à un degré moindre de celui auquel nous acceptons A. La rationalité, en tout cas de prime abord, n'a rien à voir avec la vérité, au sens où un agent rationnel peut n'avoir aucune croyance vraie d'une part et où, d'autre part, il n'est pas nécessaire de posséder le concept de vérité pour être rationnel : les contraintes de rationalité dont est porteur le critère de vulnérabilité à un dutch book sont purement structurelles. Avons-nous donc ce que nous cherchions ? Non, un problème subsiste : il se peut que le joueur qui parie à la fois sur A et sur l'incohérence de A soit irrationnel, quoique la conjonction de A et de la proposition que A est incohérente ne soit pas logiquement contradictoire, mais comment le joueur peut-il reconnaître cela ? Comment peut-il savoir qu'il est irrationnel de parier à la fois sur A, et sur l'incohérence de A ? Certes, il lui suffirait pour s'en aviser de refaire le raisonnement que nous avons nous-même produit pour reconnaître le caractère défectueux de son attitude et justifier son acceptation de la cohérence. Le problème est que ce raisonnement met en jeu le concept de vérité, de façon cachée : car comment savons-*nous* que, si A n'est pas cohérente, alors Karl perdra de l'argent, sinon parce que nous avons dérivé logiquement *a priori* de la vérité de A (« Tous les théorèmes de A sont vrais ») la cohérence de A ? Si nous devions écrire dans le détail la façon dont nous avons calculé les gains et les pertes associés au contrat précédent, nous nous apercevrions

que nous avons précisément fait le genre de raisonnement présenté par Shapiro en début de chapitre, *via* un détour par la notion de vérité pour rendre compte du fait que la cohérence de A est une conséquence *a priori* de A.

Nous avons beau avoir montré qu'il est irrationnel d'accepter à la fois A et l'incohérence de A, ce détour nous ramène simplement à cette vérité que les seules lois de la vérité mises à contribution dans les explications de la cohérence sont des lois *a priori* ; et nous n'avons donc pas expliqué comment un agent acceptant une théorie A peut expliquer qu'il est justifié à accepter la cohérence de A, sans en passer par la preuve-par-la vérité. Il semble donc qu'il faille nous y prendre autrement que par la méthode des *dutch book*. Dans la section suivante, plutôt que d'essayer de chercher à appliquer un autre critère de rationalité, je propose de procéder directement en réfléchissant sur certaines contraintes qui doivent peser sur l'action rationnelle.

5.4 Cohérence en première personne : Réflexion épistémique et Responsabilité

Je voudrais montrer qu'il est possible de dériver le caractère rationnel de l'acceptation de la cohérence d'une théorie que l'on accepte à partir d'une réflexion *a priori* sur nos normes d'acceptation, d'une part, et sur certaines obligations auxquelles est soumis un sujet rationnel. Je propose donc de procéder en deux temps : je commencerai par m'interroger sur certaines des lois *a priori* qui doivent gouverner une bonne notion d'acceptabilité, puis je discuterai du caractère rationnel du principe de Responsabilité présenté en introduction.

5.4.1 Acceptabilité

La preuve de la cohérence par la vérité faisait un détour par le Principe de réflexion aléthique sur A :

Principe de Réflexion aléthique :

A est vraie

Une fois ce principe justifié, il suffisait de prouver dans un second temps que l'ensemble des énoncés vrais est cohérent, ce dont une analyse conceptuelle de la notion de vérité nous assure, pour inférer la cohérence de A. Mais la cohérence n'est

pas l'apanage de l'ensemble des énoncés vrais et je voudrais à présent soutenir, pour commencer, que la cohérence de A doit suivre également du principe de « réflexion épistémique sur A » suivant :

Principe de Réflexion épistémique : A est acceptable

En effet, que signifie l'affirmation que je suis justifié à accepter la proposition ou la théorie A, ou comme je dirai aussi de façon synonyme, que A est acceptable (par moi, maintenant) ? Cela signifie que mon acceptation de A satisfait à une certaine norme, que j'affirme qu'il m'est permis, au regard d'une certain code tacite d'« éthique épistémique », d'accepter A.

Par « acceptable » j'entends ce qui satisfait aux normes qui gouvernent l'action d'accepter au sens que nous avons donné à ce terme. On peut accepter un énoncé, une théorie, à tort ou à raison, et c'est à la notion d'acceptabilité de rendre compte de ce qui doit compter comme accepter *à tort* ou *à raison*. Mais par ailleurs l'acceptabilité n'a rien à voir avec la vérité, ce sont des normes différentes. Supposons qu'il soit vrai que la première planète habitée hors du système solaire soit à moins de deux miliards d'années-lumières de la terre. La proposition que la première planète habitée est à moins de deux milliards d'années-lumières de la terre est-elle pour autant acceptable (par moi, aujourd'hui) ? Non, évidemment : je n'ai aucune *raison* de l'accepter. Inversement, il est plausible que certaines hypothèses des sciences physiques contemporaines soient acceptables (par nous aujourd'hui) bien qu'elles soient fausses. Par conséquent, il n'est pas besoin d'y insister trop longtemps, la vérité et l'acceptabilité sont deux propriétés bien distinctes.

Mais alors quelles sont les règles qui gouvernent *a priori* la notion d'acceptabilité ? Revenons à la question de savoir ce que prescrit le « code » auquel nous devons soumettre nos décisions d'accepter, c'est-à-dire les conditions sous lesquelles nous sommes justifiés à accepter ce que nous acceptons. La question de savoir ce que doit contenir un tel code éthique est aussi difficile que la question de la nature de la justification elle-même, et nous n'envisageons pas d'essayer de dire quelque chose de neuf à ce sujet. Faut-il y inscrire l'injonction de Descartes de n'accepter que ces énoncés dont la vérité est claire est distincte ? D'autres règles opératoires sont sans nul doute moins problématiques. On peut penser qu'il n'est pas permis d'accepter une hypothèse en présence seulement d'indices de sa fausseté ; ou qu'il n'est permis d'accepter un rapport d'observation que si nous avons vérifié que les conditions de cette observation remplissaient un certain nombre de conditions variées (conditions

d'éclairage, reproductibilité etc.). C'est sans doute seulement lorsque de telles conditions sont réunies que je suis justifié à accepter une hypothèse, et seulement lorsque je me suis assuré de la conformité de mes actions à cette « éthique épistémologique » que je peux me reconnaître justifié à accepter ce que j'accepte. Bien entendu il ne s'agit pas chercher à énumérer ici les articles d'une méthode universelle : la seule chose qui m'intéresse est de montrer que ces lois doivent avoir pour conséquence qu'un ensemble d'énoncés qui est *acceptable* (pour un sujet donnée, à un moment donné) est *cohérent*. Pourquoi, parmi les conditions structurelles les plus générales sous lesquelles un sujet est justifié à accepter l'ensemble des propositions qu'il accepte, doit figurer la condition que cet ensemble doit au minimum être cohérent ? Tout simplement parce que c'est ce principe qui permet de rendre compte du fait que ne sommes pas prêts à accepter une théorie que nous tenons pour incohérente. Que les principes de justification soient conçus comme devant nous assurer de la vérité probable de tous les énoncés que nous acceptons, ou qu'il s'agisse seulement d'organiser ou de systématiser de façon rationnelle un ensemble de données observationnelles, il semble que l'ensemble de ce qui est accepté *doit* être cohérent. Pourquoi ? Le problème n'est pas seulement qu'une théorie incohérente doive être fausse (car après tout, à nouveau, cette idée n'a qu'une application limitée pour un instrumentaliste). Le problème est qu'une théorie incohérente est inutile. L'usage essentiel des théories consiste dans le fait qu'elles rendent un certain nombre d'inférences possibles, qu'on les conçoive comme des outils d'explication, de prédiction ou de systématisation. Puisqu'il faut accepter les conséquences logiques de ce que l'on accepte, accepter une théorie incohérente reviendrait à tout accepter, c'est-à-dire tout et son contraire. Or, quelle que soit la façon précise dont on en conçoit le but, la racine du projet d'élaboration d'une théorie est de discriminer certains énoncés, ceux que l'on accepte, de ceux que l'on accepte pas. Si tout est acceptable, alors c'est la raison même de l'activité théorique qui disparaît. Par conséquent, il est hautement plausible que, de même qu'une analyse conceptuelle de la notion de vérité révèle que l'ensemble des énoncés vrais est cohérent, de même une analyse conceptuelle de la notion de justification ou d'acceptabilité doit révéler que si un agent est justifié à accepter un ensemble d'énoncés, ou si un ensemble d'énoncés est acceptable (par un sujet à un moment donné), alors cet ensemble d'énoncés est également cohérent. Par conséquent il semble que les principes suivants doivent être vrais *a priori* :

Principes minimaux d'acceptabilité :

1. Acceptable (X)→ Acceptable(Cn(X))
   où X est un ensemble d'énoncés et $Cn(X)$ la clôture de X par les règles d'inférences.
   Autrement dit les conséquences d'un ensemble acceptable d'énoncés sont acceptables.
2. ¬Acceptable($\perp$)
   C'est-à-dire : les contradictions ne sont pas acceptables.
   Par conséquent, de façon dérivée :
3. Acceptable (X)→ Cohérent(X)
   Si X est un ensemble acceptable d'énoncés alors X est cohérent.

Mais si ce que nous venons de dire est correct, la dérivation de la cohérence de A *à partir* du principe de Réflexion épistémique sur A est quasiment immédiate :

1. A est acceptable (Principe de réflexion épistémique)
2. Si A est acceptable, alors A est cohérente (réflexion sur les normes d'acceptabilité)
3. Donc A est cohérente. (par 1, 2)

Nous avons une dérivation de la cohérence de A à partir du principe de réflexion épistémique qui est tout à fait analogue à la dérivation de la cohérence de A *à partir* du principe de réflexion aléthique. Au lieu de faire appel à des lois *a priori* de la vérité, néanmoins, cette dérivation ne fait appel qu'à une analyse élémentaire des buts qui gouvernent l'action d'accepter telle nous l'avons présentée, à la signification que nous donnons au terme « acceptable ».

Mais, bien entendu, cette dérivation est encore loin de constituer en elle-même une explication de notre thèse de départ, à savoir qu'un agent rationnel acceptant une théorie A doit rationnellement accepter la cohérence de A. Il y a un fossé conceptuel apparemment infranchissable entre l'acceptation (de A) par un agent et l'acceptabilité (de A dans les conditions qui sont celles de l'agent), entre le fait qu'un agent accepte une théorie et le fait que cet agent soit justifié à accepter cette théorie.

5.4.2 Le principe de responsabilité en première personne

Le principe qui permet de faire le pont entre la petite dérivation de la section précédente et cette idée qu'un sujet acceptant une théorie donnée doit accepter que cette théorie est cohérente est le suivant, que j'appellerai le Principe de Responsabilité :

Principe de Responsabilité :

> Si un agent rationnel S accepte un ensemble d'énoncés X, S doit accepter « X est acceptable ».

Si ce principe est correct, en effet, nous avons l'explication cherchée :

1. S accepte A (notre hypothèse de départ)
2. Donc S doit accepter « A est acceptable »(par 1 et Responsabilité)
3. Or si X est acceptable, alors X est cohérent (réflexion sur les normes d'acceptabilité/justification)
4. Donc S doit accepter « A est cohérent » (2 et 3)

La question est donc savoir si le principe de Responsabilité est correct. Bien entendu, le principe suivant est faux :

$$\text{J'accepte } X \ \rightarrow \ X \text{ est acceptable}$$

Il peut être *vrai* qu'un agent rationnel accepte de fait la théorie A, sans que pour autant A satisfasse aux critères d'acceptabilité. C'est une situation banale dans laquelle l'agent s'est simplement trompé et une illustration parmi d'autres du fossé entre ce qui est et ce qui doit être. Comment alors le principe de responsabilité peut-il être une contrainte *rationnelle* ?

Une façon voir ce qui est en jeu est de commencer par considérer le cas d'un agent rationnel qui accepterait :

(*) La terre tourne autour du soleil, mais je ne suis pas justifié à accepter que la terre tourne autour du soleil.

Les *conditions de vérité* de cet énoncé ne sont pas problématiques, pas plus que, pour prendre un exemple célèbre entre tous, la négation du cogito cartésien (« Je ne suis pas ») n'est une contradiction logique. Le caractère paradoxal de ces affirmations n'est pas à chercher dans le contenu sémantique de ce qui est affirmé. Le paradoxe

est pragmatique : ces énoncés ne sont pas paradoxaux, c'est leur affirmation, ou leur acceptation qui l'est.

Moore[33] avait déjà fait remarquer qu'il y a quelque chose de paradoxal à affirmer des énoncés de la forme suivante :[34]

- Il y a un lingot d'or dans le jardin mais je ne crois pas qu'il y ait un lingot d'or dans le jardin.
- Il y a un lingot d'or dans le jardin mais je crois qu'il n'y a pas de lingot d'or dans le jardin
- Il y a un lingot d'or dans le jardin mais je ne sais pas s'il y a un lingot d'or dans le jardin
- Il y a un lingot d'or dans le jardin mais je ne pense pas qu'il y ait un lingot d'or dans le jardin
- etc.

En quoi consiste le caractère paradoxal de ces affirmations ? À nouveau, un énoncé de Moore n'est pas logiquement contradictoire, ses deux membres peuvent être vrais conjointement. D'autre part, le caractère paradoxal de l'affirmation disparaît dès que certaines modifications sont opérées :

- Passer du présent au passé fait disparaître le paradoxe. Affirmer qu'il y avait un lingot dans le jardin mais que je ne croyais pas qu'il y avait un lingot dans le jardin ne pose pas de problème.
- L'emboîtement dans un contexte logique plus large fait disparaître le paradoxe. Il n'y a rien de paradoxal dans mon affirmation de « Il est possible qu' il y ait un lingot d'or dans le jardin mais que je ne le croie pas », ou « Si il y a un lingot d'or dans le jardin et que je ne le croie pas, alors il y a peu de chance que je le trouve jamais ».[35]
- De même passer de la première personne à la seconde ou la troisième fait disparaître le paradoxe. Rien de paradoxal dans l'affirmation qu'il y a un lingot d'or dans le jardin et que Pierre ne croit pas qu'il y ait un lingot d'or dans le jardin.

[33] Moore (1942), p.512

[34] Ce sont sans doute les remarques qu'y consacre Wittgenstein dans Wittgenstein (2004) (IIx) qui ont suscité pour le paradoxe de Moore un intérêt durable. Wittgenstein envisage la possibilité que les attributions de croyance en première personne ne puisseent pas être conçues comme de simples *descriptions* de nos propres états mentaux.

[35] Remarquons que le caractère paradoxal semble en revanche survivre au passage au futur.

Nous avons suivi Moore et posé le problème en terme d'affirmation (ou d'assertion), mais le paradoxe ne se réduit pas à la violation d'une maxime conversationnelle du type « asserter p seulement si l'on sait/croit/pense que p »[36] La difficulté est la même si l'on considère un agent rationnel *croyant*, *acceptant*, pour soi-même une proposition mooréenne :[37] quelle est la règle qui est violée lorsqu'un agent *croit* un énoncé de Moore ?

Il me semble que, dans le cas de l'assertion des énoncés pradoxaux de Moore comme dans le cas de l'agent qui accepte l'énoncé $(*)$ ci-dessus c'est plus profondément un principe du même type qui est violé. Ce qui ne vas pas dans ces différentes actions c'est qu'elles violent une *principe de responsabilité constitutif de notre rationalité* : les croyances constituées dans l'examen réflexif de nos propres croyances et les croyances examinées doivent d'une certaine façon être *coordonnées* pour former un ensemble rationnel. Ma croyance que je crois à tort (maintenant, à cet instant précis) qu'il n'y a pas de lingot d'or dans le jardin est irrationnelle parce qu'elle une telle croyance à propos de croyances miennes *doit* être concomitante d'une révision de mes croyances « de premier ordre ».

Nous avons dit que l'acceptation au sens où nous employons ce terme est une action délibérée, réflexive, critique, telle qu'elle est à l'œuvre dans un contexte scientifique idéalisé. Dans ces conditions l'acceptation d'une hypothèse ou d'une théorie est *lumineuse*, pour reprendre le terme de Williamson[38], au sens où si nous les acceptons nous *savons* que nous les acceptons. Dès lors, la justification du principe de responsabilité est à chercher dans la relation que doit entretenir un agent rationnel au contenu de ce qu'il accepte, et qu'il n'ignore pas. Mais ce vers quoi pointent les remarques du paragraphe précédent est que cette relation ne peut pas être simplement conçue sur le modèle observationnel, celui d'un agent *constatant* simplement qu'il accepte une hypothèse ou une théorie donnée : s'il n'y avait que cela, alors observer que l'on accepte une théorie A ne contraindrait nullement l'acceptation de

[36] Il est clair néanmoins que ces assertions semblent en effet une *maxime* gricéenne de *qualité* : *Ne dites pas ce que vous croyez être faux.* (Grice (1991) p.27)

[37] Ce point est noté également dans Williams (2004) pour la croyance. La stratégie de Williams pour rendre compte du caractère défectueux d'une croyance en première personne d'un énoncé de Moore est d'offrir une justification au principe de Evans :

> Ce qui justifie ma croyance que p justifie également ma croyance que je crois que p. (Williams (2004), p.349).

Nous n'adhérons pas à cette maxime, faute de la comprendre complètement. Les conséquences qu'en tire Williams sont néanmoins proches des nôtres en esprit (voir en particulier *op. cit.* p. 352-353).

[38] Williamson (2000).

la cohérence. Cette observation serait essentiellement analogue à l'observation du fait qu'*autrui* accepte A[39]. En en effet *cette observation* ne me donne *en soi* aucune raison d'accepter que A est justifiée ou satisfait aux normes minimales d'acceptabilité. Au contraire, la *rationalité* d'un sujet commande que la nature de l'articulation de ses jugements à propos de ses propres jugements, avec ces derniers jugements eux-mêmes, incorpore essentiellement le fait que les uns comme les autres sont *ses* pensées. On a pas la même relation épistémologique, en termes de droits comme en termes de devoirs, avec le contenu de ses propres pensées et avec le contenu des pensées d'autrui, quand bien mêmes ces contenus seraient identiques d'un point de vue sémantique. L'idée qu'il existe un lien essentiel entre la rationalité et la nature de notre rapport à nos propres pensées, bien sûr, n'est pas une idée nouvelle. C'est au contraire un thème classique des études philosophiques sur la connaissance de soi. Tyler Burge (Burge 1996) écrit par exemple :

> Trouver de façon justifiée ses propres raisons invalides ou ses pensées injustifiées, est normalement en soi une raison paradigmatique, du point de vue des pensées examinées (ainsi que dans la perspective de l'examen), de les altérer [...]. L'examen des raisons qui est partie intégrante du raisonnement critique inclut l'examen et les attitudes examinées en un unique point de vue. Le modèle observationnel simple traite l'examen et le système examiné comme dissociés d'une façon qui est incompatible avec les normes de l'examen critique. Il fait du système examiné un objet d'investigation, mais non une partie du point de vue de l'investigation. [...] Nous sommes épistémiquement responsables seulement parce que nous sommes capables d'examiner nos pensées et nos raisonnements.[...] Notre responsabilité lorsque nous réfléchissons sur nos pensées s'étend immédiatement à l'ensemble du point de vue. (Burge (1996), p.110-111)

Si ces remarques sont correctes, alors le principe de Responsabilité :

si S accepte X, S *doit* accepter « X est acceptable ».

est bien un principe de *rationalité en première personne*. Et puisque en outre, il semble que pour le justifier il ne soit nullement besoin de faire appel à la notion de vérité, mais seulement à des considérations générales sur la nature de la rationalité, la justification de son acceptation de la cohérence d'une théorie par un sujet accep-

[39] Ou, par exemple, que nous l'avons accepté dans le *passé*, peut-être un passé immédiat.

tant la théorie en question peut être donnée en première personne sans recours à la notion de vérité. Elle a la forme suivante :

1. J'accepte A
2. A est acceptable (par 1 et Responsabilité)
3. Si un ensemble d'énoncés est acceptable il est cohérent (théorie *a priori* de l'acceptabilité)
4. Donc A est cohérente (par 2 et 3).

Bien entendu, cette suite de jugements ne constitue nullement une *preuve* de la cohérence de A sous l'hypothèse que j'accepte A. Ce n'est pas une preuve parce que les étapes de cette justification ne préservent pas la vérité. Plus précisément c'est la justification du passage de (1) à (2) qui n'est pas une justification au sens ordinaire, mais une étape défaisable du raisonnement : sa validité est défaite par n'importe quel indice de ce que A n'était pas, en fait, acceptable. Je vais revenir sur ce point en conclusion. En attendant je remarque simplement, au-delà de la différence fondamentale qui sépare cette justification de la cohérence de l'autre, celle qui est développée dans la « *preuve*-par-la vérité », l'analogie qui demeure dans la seconde partie du raisonnement, celle qui conduit, dans un cas, de la vérité de A à la cohérence de A, et dans l'autre de l'acceptabilité de A à la cohérence de A, à partir de principes *a priori* gouvernant la vérité et l'acceptabilité respectivement. Il se peut très bien que la théorie A se révèle être incohérente à l'usage. Qu'en conclurions-nous ? *Via* la théorie de l'acceptabilité, simplement que la théorie A n'est pas acceptable, de même que, dans la théorie de la vérité[40] on peut prouver, sous l'hypothèse que A n'est pas cohérente, que A n'est pas vraie.

5.5 Conclusion

Les deux propriétés suivantes, qu'une théorie est susceptible de posséder, n'ont à peu près aucun rapport : (1) la propriété d'être vraie (2) la propriété d'être crue vraie, ou assertée, acceptée, éventuellement de façon justifiée, par un agent. De toute évidence, on peut le regretter, aucune des deux n'implique l'autre : une théorie peut être vraie sans être crue telle, et elle peut être crue vraie sans l'être. Et la même

[40] C'est-à-dire ici la théorie constitutée des principes aléthiques tarskiens et d'une théorie de la syntaxe, à l'exclusion des axiomes de la théorie de base, ce que j'ai noté MT au chapitre 2. Voir chapitre 2, section 5 pour une définition.

chose est vraie si l'on remplace la simple croyance par la croyance (l'acceptation) justifiée. Revenons en conclusion sur les deux justifications de la cohérence d'une théorie que nous avons examinées, celle qui est fournie à un sujet par la preuve par la vérité, et celle qui est fournie par réflexion sur les normes qui gouvernent ses décisions épistémiques. Le raisonnement de Shapiro est une *preuve* de sa conclusion sous les hypothèses spécifiées :

1. A
2. A est vraie [par 1 et les lois tarskiennes de la vérité]
3. A est cohérente [par 2, et principes *a priori* de la vérité (les équivalences-T)]

Si les prémisses sont vraies, la conclusion l'est aussi. Donc tout indice que les prémisses sont vraies vaut indice que la conclusion l'est aussi. Par conséquent, si l'on on est justifié à croire les prémisses, on est justifié à croire la conclusion.

Mais si le raisonnement de Shapiro doit compter *pour moi* comme une preuve convaincante de la cohérence de A, alors je dois en fait accepter ses prémisses. Donc en fait en il doit être vrai que : j'accepte A. Mais s'il est vrai que j'accepte A, il est vrai que je *dois*, en un autre sens, accepter la cohérence de A, c'est ce que montre le second raisonnement.

1. J'accepte A
2. A est acceptable [par 1 et responsabilité en première personne]
3. A est cohérente [par 2, et principes *a priori* gouvernant la norme d'« acceptabilité »]

Mais crucialement, à la différence du raisonnement précédent, il ne s'agit pas ici d'une déduction logique, le principe de responsabilité est faillible : il se peut que la prémisse 1 soit vraie mais que 2 ne soit pas vrai. Donc tout indice de la vérité de 1 (que l'on peut obtenir par introspection) n'est pas un indice de la vérité de 3. C'est ce qui fait de ces deux justifications des justifications radicalement différentes dans leur portée épistémologique. Oui, celui qui se croit justifié à croire A et les lois récursives de la vérité peut se croire justifié à croire en la cohérence de A ; mais celui qui se croit justifié à croire A, n'a pas de ressources conceptuelles aléthiques, et raisonne seulement sur ses devoirs épistémiques, peut seulement voir qu'il est justifié à *accepter* la cohérence de A, non à *croire* en la cohérence de A.

Cette justification de l'acceptation de la cohérence d'une théorie que nous acceptons (aussi longtemps que nous l'acceptons) n'est pas une *preuve*, c'est donc une

forme de justification *par défaut* dont les étapes sont défaisables. Ce qui compte comme une justification de la cohérence cessera de compter pour tel en présence d'indice d'incohérence (en même temps que je devrai modifier mon attitude relativement à la théorie en question). L'idée qu'il est rationnel de tenir certaines propositions pour justifiées par défaut, nous l'avons dit, a été défendue récemment dans la littérature par Crispin Wright (Wright 2004*a*) dans un effort pour tirer des leçons épistémologiques du scepticisme radical. Les propositions que Wright vise à justifier de la sorte sont ce qu'il appelle « les pierres de touche de toute entreprise cognitive », ces hypothèses sans lesquelles nous ne pourrions regarder aucune de nos méthodes de justifications pour correctes dans une entreprise de connaissance de donnée (les lois de la logique, le fait que nous ne sommes pas le jouet d'un malin génie, le fait que nous sommes en ce moment même trompés par nos sens etc.). Il est possible que la cohérence de ce que nous acceptons fasse partie de ces pierres de touche. La présente approche fait le lien entre la possibilité d'une telle justification et les exigences spéciales de la rationalité en première personne. C'est cette perspective qui donne du sens à l'idée de responsabilité : parce que *je* décide d'accepter une théorie donnée, certaines décisions supplémentaires s'imposent à *moi*.

Reprenant pour terminer le fil général de ce travail, je conclurai provisoirement sur ce constat d'échec : les ressources conceptuelles aléthiques semblent bien indispensables pour dériver, prouver, la cohérence, et donc rendre compte du fait qu'un sujet est justifié à *croire* une théorie donnée est justifié à *croire* qu'elle est cohérente.[41] Mais ce n'est pas le dernier mot de ma réponse à l'argument de Shapiro. Je donnerai au chapitre 6 une interprétation épistémologique des preuves de cohérence par la vérité qui rende plausible la thèse que le concept de vérité n'y joue pas, malgré tout, de rôle explicatif réel.

Un dernier avertissement pour conclure : dans la suite de ce travail, j'utiliserai à nouveau le terme « accepter » dans un sens plus naïf, indépendamment de ce que j'ai ai dit ici, et la même chose sera vraie du terme de « justification ».

[41] On peut bien sûr prouver la cohérence de A dans une métathéorie où le concept de vérité n'apparaît pas explicitement, par exemple dans une métathéorie « essentiellement plus riche », au sens de Tarski, que la théorie de départ. Mais dans ce cas on peut en fait définir la vérité dans la métathéorie, et par conséquent la notion de vérité est bien « implicitement » présente dans la métathéorie.

Chapitre 6

Conséquence réflexive

Les théories de la vérité, pour quoi faire ?

À la fin du chapitre 3, nous avons laissé un certain nombre de questions en suspens. Il y avait d'une part la question de savoir quels sont les usages du prédicat de vérité dont une théorie de la vérité doit rendre compte. C'était le Problème de l'usage. Nous avons vu que du point de vue déflationniste l'utilité du prédicat de vérité résidait dans son rôle pour l'expression de certaines généralisations. Mais que veut dire cette thèse exactement ? De quelles assertions une théorie de la vérité est-elle comptable ? Faut-il par exemple admettre la Thèse de la Réflexion[1], et plus généralement le fait qu'une théorie de la vérité doit permettre de *prouver* certaines généralisations ? Le problème connexe de déterminer, une fois donnée une réponse au Problème de l'usage, quelle théorie de la vérité permet d'en rendre compte est donc lui aussi resté sans réponse. D'autre part, il y avait la question de savoir à quelles conditions on peut reconnaître dans l'usage d'un énoncé ou d'un ensemble d'énoncés un rôle explicatif, et non seulement expressif. Après avoir écarté le critère de non-conservativité au chapitre 3, le Problème de la frontière était resté ouvert. L'objet de ce chapitre est de formuler une réponse cohérente à ces questions et de donner ainsi des éléments de réponse positifs au Problème de la stabilité des thèses déflationnistes auquel Shapiro et Ketland ont donné son acuité.

Avant d'entrer dans le vif de mon argument, je voudrais commencer par mettre le lecteur sur la voie que je vais suivre. Nous avons vu au chapitre 3 qu'un sujet acceptant une théorie A, et capable de mobiliser la notion de vérité de façon appropriée, était en position de construire une preuve de la cohérence de A. On

[1] Présentée au chapitre 3 également.

peut penser, comme Shapiro et Ketland, que ce fait d'usage étaye la Thèse de la Réflexion, à savoir qu'il doit être possible de *dériver* la cohérence de A à partir de A elle-même et d'une théorie de la vérité.[2] Mais il y a, à l'œuvre dans le passage de l'observation de départ à la Thèse de la Réflexion, une hypothèse injustifiée. La remarque fondamentale que je voudrais mettre en avant relève de l'examen critique des conditions sous lesquelles il est possible de dire qu'un sujet croit véritablement une *théorie*.[3] Cet examen révèle qu'il n'est en général pas possible à un sujet d'exercer pleinement sa faculté de juger sur une théorie A[4] sans l'aide médiate du concept de vérité, et cela pour des raisons qui se ramènent simplement, d'une part, à celles déjà évoquées au chapitre 1, à savoir la nécessité d'avoir recours à la montée sémantique pour formuler certaines généralisations et, d'autre part, au caractère infinitaire de l'objet qu'est une théorie. Ma thèse est qu'une fois les conditions de cet exercice clairement dégagées, il apparaît alors cette conséquence remarquable : si c'est un fait de logique que la cohérence de A n'est pas déductible de A, il apparaît comme un fait épistémologique que la cohérence de A est déductible de ce qui doit en effet être le contenu des jugements *effectivement* portés par un sujet acceptant la théorie A. Ce genre de considérations conduit alors à une réinterprétation radicale de l'usage de la notion de vérité qu'illustre l'observation des phénomènes d' « explications sémantiques » de la cohérence.

Mon plan sera le suivant. Dans une première partie, j'introduis la notion de conséquence réflexive d'une théorie, qui sera le fondement de ma démarcation entre usages explicatifs et usages expressifs d'une notion. Je commencerai par introduire une notion intermédiaire de conséquence α-réflexive, dont le seul objet est de faire le lien entre la notion de conséquence réflexive qui sera présentée ensuite et certaines idées gödéliennes présentées au chapitre 4. Dans la seconde partie, j'entreprends l'examen épistémologique annoncé, à savoir celui des conditions de possibilité de la croyance en une théorie par un sujet fini. Je montre alors que le principe de réflexion aléthique sur A et la cohérence de A sont des « conséquences réflexives » de A. Dans la troisième partie je reviens sur les théories de la vérité : à la lumière de ce qui précède, je défends alors la théorie minimale de la vérité, tout en présentant une interprétation de la théorie tarskienne qui fait droit à son rôle d'objet théorique

[2] Ici, comme précédemment, je suppose que A contient sa propre syntaxe.

[3] Je dirai aussi aussi « accepte » une théorie, mais c'est toujours des théories contentuelles que j'aurai en vue, et dans ce contexte j'utilisai « accepter » en un sens relâché, plus ou moins comme un synonyme de « croire ».

[4] Je dirai plus précisément ce qu'il faut entendre par là dans la suite.

central.

6.1 Conséquence réflexive

Au chapitre 3, j'ai présenté la réponse de Stewart Shapiro et Jeffrey Ketland au Problème de la frontière : une théorie A est « substantielle » relativement à une théorie B si A n'est pas conservative sur B.[5] Dans cette section, je propose une autre explication (au sens de Carnap) de la frontière que le déflationniste entend tracer : A n'est pas substantielle relativement à B si A n'est pas conséquence réflexive de B. Comme annoncé en introduction, je commence par introduire une notion intermédiaire de « conséquence α-réflexive » pour faire le lien entre les considérations du chapitre 4 et ce qui suit.

6.1.1 Préliminaire : Gödel

À côté de la notion paradoxale d'ensemble comme extension d'une formule quelconque, il existe une autre notion naturelle d'ensemble. C'est la notion cumulative ou itérative d'ensemble : un ensemble est n'importe quel objet que l'on peut former à partir d'un univers donné d'individus en itérant l'opération consistant à former l'ensemble des objets satisfaisant telle ou telle conditions parmi ceux déjà donnés. Mais si Gödel pense que cette notion d'ensemble est, en un sens, parfaitement définie,[6] il ne pense pas pour autant qu'il y en ait une théorie accessible déterminée et achevée. Le processus cumulatif fondé sur l'opération « ensemble de » joue plutôt le rôle d'une idée directrice que la raison semble pouvoir déployer à l'infini. Pour faciliter la discussion qui suit, je forcerai un peu la terminologie : je parlerai de *l'intuition* de la notion itérative d'ensemble, que l'on peut se figurer comme la source à laquelle s'alimente le processus de formation des idées qui donne lieu à la formulation des axiomes, comme d' un contenu de nature propositionnelle, et qui comme tel peut avoir des conséquences logiques, ou être lui-même conséquence logique d'autres contenus propositionnels. Mon propos n'étant pas exégétique, ces approximations sont acceptables.

[5] Comme au chapitre 3, je dirai tantôt que ce qu'il faut expliquer est la démarcation entre rôle explicatif et rôle expressif, tantôt, quand le contexte est suffisamment clair, que ce qu'il faut expliquer est le terme « substantiel ».

[6] Au sens suivant : étant donné cette notion d'ensemble, tous les énoncés ensemblistes doivent avoir une valeur de vérité déterminée (même s'il n'est pas forcément possible de décider laquelle.).

Ceci étant posé, et comme nous l'avons vu au chapitre 4, Gödel distingue alors entre différents types d'extensions de ZF. Certains axiomes, quoique *logiquement indépendants* de ZF, forment des extensions *intrinsèquement justifiées* de ZF. Gödel entend par là qu'il s'agit d'énoncés dont nous pouvons reconnaître la vérité directement en vertu de notre compréhension de la notion itérative d'ensemble. Il en donne pour exemple les axiomes de grands cardinaux. Pour d'autres énoncés ensemblistes, affirme Gödel, il en va tout autrement : notre compréhension de la notion d'ensemble ne nous permet pas de décider de leur valeur de vérité. L'hypothèse du continu est de ce second genre, et Georges Boolos a soutenu que l'axiome du choix en était un autre exemple.[7] On peut représenter la situation par une série de schémas sur lesquels nous avons, d'une part, les axiomes de ZF, qui sont indépendants les uns des autres,

$$ZF_{Ax1} \qquad ZF_{Ax2} \qquad \cdots$$

et d'autre part ce que nous avons appelé l' « intuition » du concept itératif, un contenu propositionnel que je note X sur le schéma ci-dessous, que nous acceptons ou percevons, et qui *implique* les axiomes de ZF :

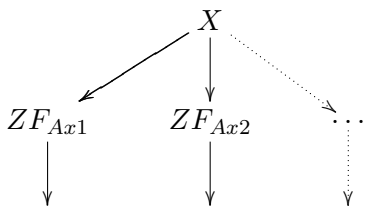


FIG. 6.1: Les axiomes de ZF ne sont pas conditionnellement indépendants.

Or X implique aussi, outre les axiomes de ZF et ses conséquences, un certain nombre d'autres énoncés, logiquement indépendants de ZF, comme par exemple certains axiomes affirmant l'existence de grands cardinaux (voir figure 6.2).

[7] Voir BOOLOS (1971).

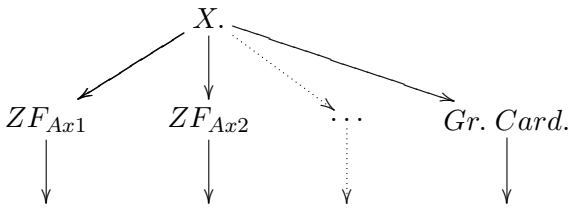


FIG. 6.2: Les axiomes de grands cardinaux ne sont pas conditionnellement indépendants des axiomes de ZF.

La valeur de vérité de l'hypothèse du continu, en revanche, n'est pas reconnaissable par réflexion sur notre idée directrice d'ensemble. Autrement dit, tandis que, dans le système de nos croyances justifiées, les axiomes de ZF et de grands cardinaux ont un parent commun, ce n'est pas le cas de l'ensemble des axiomes étendus à l'hypothèse du continu ou à l'axiome du choix.

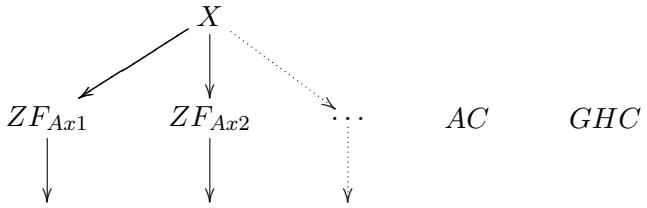


FIG. 6.3: L'hypothèse (généralisée) du continu et l'axiome du choix sont conditionnellement indépendants des axiomes de ZF.

Ajouter un axiome ensembliste indépendant de ZF à ZF, c'est toujours et invariablement former une extension non-conservative de ZF. Cela, c'est la logique qui le dit. La remarque de Gödel n'est pas une remarque de logique, c'est une remarque d'épistémologie. En substance, il y a une signification épistémologique des phénomènes d'extensions non-conservatives qui ne peut apparaître qu'à la lumière d'une analyse de notre *accès* aux différents axiomes : certains sont des manifestations de l'introduction de principes introduisant des aperçus radicalement nouveaux

sur le sujet étudié, hétérogènes à ceux déjà admis, d'autres ne font que développer, exprimer, formaliser, des principes explicatifs déjà admis. C'est à préciser cette idée que je vais consacrer la section suivante.

6.1.2 Principes explicatifs et conséquence α-réflexive

La conservativité est une relation à deux arguments : on parle de la conservativité d'une théorie sur une autre théorie.[8] Sur la voie de la recherche d'un nouveau critère de démarcation qui ne soit pas seulement logique mais épistémologique, d'un critère qui fasse droit à la relation qu'un sujet entretient avec les propositions qu'il accepte, on s'attend naturellement à ce que le nouveau critère dépende d'un troisième paramètre représentant ce sujet de la connaissance.

6.1.2.1 Conséquence α-réflexive

Pour saisir l'idée gödélienne d'une proposition qui n'introduit « rien de nouveau » dans le schéma général d'explication d'un sujet donné, je propose la notion (informelle) de conséquence α-réflexive d'une théorie :

Définition 12. *Soit α un agent. ϕ est une conséquence α-réflexive d'un ensemble d'énoncés Γ si et seulement si les justifications de α pour Γ permettent à α de justifier ϕ.*

Plutôt qu'un objet de nature propositionnel, je considérerai une justification comme une *suite d'actions*, actions de deux catégories distinctes : des jugements et des inférences. Dans cette perspective, une justification d'un énoncé ϕ est une action composée, constituée d'un ensemble de jugements de départ dont le contenu sont des propositions tenues pour vraies par le sujet (les prémisses), et d'une suite d'actions inférentielles et de jugements intermédiaires dont le dernier terme est le jugement que ϕ. Dans le cas où une justification ne comporte aucune inférence et se réduit à un simple jugement, je parlerai de justification *directe*. Typiquement, pour tenir pour justifiés les *axiomes* d'une théorie, en tant qu'ils doivent être tenu pour *évidents*, nous n'avons pas besoin de faire de détour par une justification inférentielle. Par ailleurs, de façon générale, nous n'avons pas besoin de supposer que les inférences par lesquelles le sujet construit une justification de la conclusion sont des inférences

[8] Voir, à nouveau, chapitre , section 1.

logiques : il peut s'agir d'inférences inductives, d'inférences analytiques[9], etc. Un dernier point enfin, sur lequel nous aurons l'occasion de revenir : s'il est vrai que, lorsqu'un sujet juge que ϕ, il sait qu'il juge que ϕ, un sujet peut mener à bien une suite d'inférences qui justifie ϕ, établissant ainsi une preuve de ϕ, sans être en position d'affirmer qu'il a une preuve ou une justification de ϕ. Il pourrait par exemple avoir réflexivement un doute sur la validité d'une des inférences en jeu, et ne pas être capable d'établir cette validité à sa propre satisfaction, quoique cette inférence soit en fait valide. Dans ce cas, le sujet possède seulement ce que j'appellerai une justification *aveugle* pour ϕ.[10] Pour *voir* qu'il possède une justification pour ϕ, il faut de plus que le sujet sache que les inférences en jeu dans sa justification sont elles-mêmes, d'une manière ou d'une autre, justifiées.[11]

Ces quelques précisions visent simplement à spécifier une certaine compréhension ordinaire de la notion de justification et devraient être suffisantes pour les besoins très spécifiques et limités qui sont les nôtres ici. Un cas particulier, mais typique, de ce que j'ai en vue quand je parlerai de conséquence α-réflexive est une situation dans laquelle à partir des *preuves*, que possède un sujet α pour les propositions de Γ, où une preuve est n'importe quelle démonstration (*éventuellement en une ligne*) dont toutes les hypothèses sont tenues pour vraies, α peut construire une preuve de ϕ.

Comme attendu, je propose, avec la terminologie que nous avons fixée, de reformuler l'idée gödélienne présentée dans la section précédente (et en admettant les idéalisations de sa position que nous avons faites) comme la thèse que les axiomes de grand cardinaux sont des conséquences α-réflexives de ZF pour certaines valeurs de α. Pourquoi pour « certaines valeurs de α » seulement ? N'est-ce pas vrai pour toute valeur du paramètre α ? Pas nécessairement. Supposons en effet que la justification que possède un sujet a pour croire les axiomes de ZF soit *déférentielle*, au sens où

[9] Par exemple : $\dfrac{\text{Pierre est célibataire}}{\text{Pierre n'est pas marié}}$

[10] « Aveugle » parce que le sujet suit les règles « à l'aveugle », au sens où Kripkenstein (Wittgenstein tel que reconstruit par S. Kripke (1984), selon l'expression consacrée) note que nous suivons les règles « blindly ».

[11] Ceci en vertu du principe de justification inférentielle tel qu'il apparaît dans la tradition classique, fondationnaliste, de la justification, dans laquelle je m'inscris tacitement :

Justification inférentielle :

Pour être justifié à croire P sur la base de E, il faut :

1. être justifié à croire E
2. et être justifié à croire que E supporte P.

Je reviendrai sur ce point.

a croit les axiomes de ZF parce qu'il les a entendus affirmés par un sujet b en qui il a toute confiance[12], et pour cette raison seulement. Autrement la justification de a chaque axiome $A_{ZF,i}$ est de la forme :

1. Tout ce qu'affirme b est vrai,
2. b affirme $A_{ZF,i}$, donc
3. $ax_{ZF,i}$ est vrai, et donc
4. $A_{ZF,i}$.

Dans ce cas il semble clair que les axiomes de grands cardinaux ne sont pas des conséquences a-réflexives de ZF, a n'ayant aucun moyen de reconnaître la vérité des axiomes de grands cardinaux en réfléchissant sur ses justifications pour accepter les axiomes de ZF. Rien de ce qu'il regarde comme une preuve justifiant sa croyance dans les axiomes de ZF ne peut lui fournir de preuve justifiant la croyance dans les axiomes de grands cardinaux.

J'aurai l'occasion, dans la suite, de revenir sur d'autres exemples de conséquences α-réflexives.[13] Pour conclure, j'introduis une notion duale de celle de conséquence α-réflexive, pour faciliter les discussions ultérieures. La définition est celle attendue et se passe de commentaires.

Définition 13. *Soit S un énoncé, B un corps d'énoncés, α un agent.*

S est α-substantiel relativement à B si et seulement si S n'est pas justifié par les justifications que possède α pour B.

[12] On suppose que a a de bonnes raisons de mettre sa confiance en b, une confiance fondée par exemple sur son expérience passée avec lui.

[13] Avant de conclure, je fais simplement en passant une remarque sur le lien entre la notion de conséquence α-réflexive esquissée ici, et ce que pourrait être une étude bayésienne de la structure des croyances de α. Il semble en effet que dire d'une proposition p qu'elle est une conséquence α-réflexive d'un ensemble de proposition Γ doive essentiellement impliquer, par exemple, que la distribution probabilité P_α de α est telle que $P_\alpha(\Gamma \wedge p) \neq P_\alpha(\Gamma) \cdot P_\alpha(p)$, autrement dit que les probabilités (subjectives) de α pour Γ et p ne sont pas indépendantes. Ce genre de considérations pourrait motiver l'entreprise d'une étude des phénomènes que nous avons en vue en termes purement bayésiens. Néanmoins, comme je l'ai déjà noté au chapitre 5, une véritable étude bayésienne exigerait une adaptation majeure des concepts classiques pour s'appliquer aux croyances mathématiques. Je mentionne également que les notions développées dans ce chapitre ne sont pas sans rapport avec la notion de « compacité épistémologique » présentée par Detlefsen (1979). Detlefsen écrit :

> We shall say that T is « epistemically compact » when the dubitability of the whole theory is, in a sense, « reflected » in a finite portion of the theory. More precisely, we shall say that T is epistemically compact when there is finite subset T_f of the axiomes of T that is as dubitable as the entire axiom-set of T.(Detlefsen (1979), §4)

Néanmoins, une comparaison détaillée entre les différentes notions nous mènerait trop loin et j'en remets l'étude à plus tard.

6.1.2.2 Comparaison avec la non-conservativité

Si grossière cette première notion de conséquence α-réflexive soit-elle, on peut néanmoins utilement la comparer à la notion de conservativité. J'ai déjà remarqué, et il est important de noter, que la notion de pouvoir explicatif d'un énoncé (relativement à un ensemble d'énoncés) appartient à l'épistémologie, non à la logique. Par conséquent, le pouvoir explicatif d'une proposition relativement à un corps de propositions pour un agent ne peut pas être expliqué par des moyens logiques uniquement, contrairement à ce qui est supposé dans le critère de non-conservativité. C'est donc un point à mettre au crédit de l'approche présentée ici que le caractère *α-substantiel* d'un énoncé dépende explicitement d'un paramètre subjectif.

Les deux critères sont-ils *indépendants* ? D'une part, c'est ce que nous venons de voir, il est clair que des axiomes peuvent étendre non-conservativement A sans être α-substantiels relativement à A. Nous avons déjà vu l'exemple des axiomes de grands cardinaux relativement à ZF, mais de nombreux autres exemples sont possibles. Pour en rester à l'arithmétique, nous pouvons reprendre un exemple que nous avions déjà utilisé au chapitre 3 : PA avec induction pour toutes les formules arithmétiques est une extension non-conservative de l'arithmétique PA_{At} dont le schéma d'induction est limité aux formules atomiques uniquement, mais l'on peut penser que nos justifications pour la première sont aussi de fait nos justifications pour la seconde.

Il est plus difficile de se prononcer sur la réciproque. Il y a bien les cas triviaux dans lesquels on considère deux théories sans concepts communs, et sans qu'aucune loi de correspondance ne soit posée entre les deux. Ainsi, en l'absence de lois de correspondance entre les entiers et certains ensembles finis par exemple, $ZF+PA$ étend conservativement PA ; pourtant ZF est bien intuitivement α-substantiel relativement à PA. Mais ces cas de conservativité ne sont pas intéressants, parce que nous savons en fait comment interpréter les entiers dans les ensembles et qu'une fois posées les lois de correspondance idoines l'extension est non-conservative.[14] Plus intéressant est le cas suivant : en ajoutant à l'arithmétique de Peano les axiomes de la théorie des ensembles finis (et en nous donnant des lois de correspondance

[14] En fait de lois de correspondance, on peut penser à ce que Quine appelle une « fonction de délégation », permettant d'opérer une réduction de l'univers des entiers à l'univers des ensembles, à la façon par exemple de Frege ou de von Neumann. Voir Quine (1976*b*), chapitre 20, « Ontological reduction and the World of Numbers ».

entre les deux), nous obtenons une extension conservative de PA.[15] Pourtant, il est plausible que nos justifications pour les axiomes de ZF privés de l'axiome de l'infini sont tout à fait extrinsèques à nos justifications pour les axiomes de PA. En fait on est dans un cas où deux champs conceptuels étudiés au départ de façon séparée sont *in fine* identifiés l'un à l'autre. On sait que ce genre d'unification de différents domaines a souvent un pouvoir éclairant, les manières de penser typiquement associées à l'un pouvant alors être mis à profit dans l'autre. La théorie des ensembles finis, alors, est-elle α-substantielle relativement PA ? On a envie de dire « oui » si la première introduit réellement une « nouvelle manière de voir » les entiers et, en effet, l'idée que nos justifications pour accepter la théories des ensembles finis sont hétérogènes à nos justifications pour accepter PA semble correspondre à cela. Néanmoins, il convient d'être prudent ici. Sans précision sur la nature exacte de la notion de justification que nous utilisons, et sans clarification des processus mobilisés dans la justification des principes arithmétiques et ensemblistes, les notions d'α-substance et α-conséquence que nous avons introduites sont trop grossières pour permettre une analyse de l'exemple précédent qui soit suffisamment fine pour être réellement intéressante. Je me bornerai donc ici à ces indications vagues, et à noter qu'une telle analyse irait bien au-delà des ambitions de ce travail.[16]

6.1.3 Conséquence réflexive

J'en viens maintenant à la notion centrale de ce chapitre, celle de *conséquence réflexive*. J'ai noté tout à l'heure qu'un sujet pouvait avoir pour accepter des axiomes de ZF des justifications très indirectes, ayant par exemple déféré l'analyse de la notion d'ensemble à une communauté d'experts en laquelle il a des raisons de placer

[15] Voir Isaacson (1987).

[16] Je note également en passant que si l'on adopte la façon de voir sous-jacente dans le dernier exemple, on devra sans doute dire également que les principes de la syntaxe d'un langage sont α-substantiels relativement à l'arithmétique de Peano. En un sens, il me semble qu'il y a là un point correct en effet : les principes décrivant la syntaxe d'un langage (d'un langage déjà en usage) ne sont pas reconnus vrais en vertu de notre perception des entiers naturels ; la similarité de structure entre les deux théories est plutôt quelque chose qui se révèle à nous après coup, après que nous ayons réfléchi sur l'un et l'autre objet pour son propre compte et selon des voies spécifiques. On pourra donc vouloir relâcher un peu la notion d'α-substance en identifiant parfois des théories de vocabulaires et concepts différents, en particulier l'arithmétique de Peano et la syntaxe théorique qu'on lui fait ordinairement correspondre ou, dans un autre registre, PA et la théorie des ensembles finis, quitte peut-être à abandonner l'idée qu'il serait possible d'avoir des cas non triviaux de principes qui seraient conservatifs sur une théorie donnée tout en étant α-substantiels sur cette théorie. Quoi qu'il en soit, les choix que l'on peut faire ici n'auront pas d'incidence pour la suite.

sa confiance. Plus généralement, il semble que la classe de justifications qu'un sujet peut avoir pour sa croyance en un ensemble de propositions soit très variée. La notion de conséquence réflexive peut être vue comme une façon d'extraire ce qu'il y a de commun à toutes les relations concevables de conséquence α-réflexive, en faisant abstraction des modalités particulières par lesquelles un sujet particulier justifie ses propres croyances, l'effet recherché de cette généralité étant de ne retenir que certaines conditions très générales de l'exercice du jugement.

Définition 14. *ϕ est une conséquence réflexive de l'ensemble d'affirmations Γ si et seulement si tout ensemble de justifications qu'un sujet peut avoir pour Γ justifie ϕ.*

Même si le paramètre correspondant au sujet de la pensée a disparu, la notion de conséquence réflexive est bien une notion épistémique, définie par référence à toute justification possible. Ma seconde remarque concerne ce qui doit compter comme la construction d'une justification dans ce contexte général où il n'est plus fait référence à aucun sujet particulier. Je supposerai que c'est plus spécifiquement la possibilité de construire une démonstration *logique* de ϕ à partir de n'importe quel ensemble de propositions pouvant servir de base à une justification de Γ qui est en jeu ici. Cette condition remplace l'exigence antérieure, plus lâche, où cette restriction concernant les règles inférentielles à l'œuvre dans le processus de justification n'était pas en vigueur. Si maintenant l'on pense à la vieille conception médiévale de conséquence logique d'après laquelle les conclusions suivent logiquement des prémisses si elles sont déjà *contenues* en elles[17], alors dire que ϕ est conséquence réflexive de Γ c'est dire, non que ϕ est « contenue » dans Γ, mais qu'elle est « contenue » dans des propositions que nous tenons pour vraies et grâce auxquelles nous justifions Γ. Ma troisième et dernière remarque est que si Γ est un ensemble de propositions (ou d'énoncés), l'ensemble des conséquences réflexives de Γ n'est pas vide : il contient au moins toutes les conséquences logiques déductibles de Γ. Bien entendu, la notion de conséquence réflexive n'est intéressante que si elle ne se confond pas avec celle de conséquence déductive. C'est une question que j'aborderai dans la section suivante.

[17] Voir par exemple Sundholm (1998) pour des références.

6.2 Le rôle de la notion de vérité dans l'exercice du jugement

Au chapitre 3, nous l'avons vu, la réponse de Shapiro et Ketland au Problème de la frontière était supposé impliquer que les axiomes tarskiens jouaient le rôle de principes explicatifs relativement à une théorie arithmétique. Shapiro et Ketland parvenaient à cette conclusion par la preuve de la cohérence par la vérité : l'extension tarskienne de l'arithmétique A permet de prouver la cohérence de A, que A ne peut prouver.

À identifier le caractère « substantiel » d'une proposition ou d'un ensemble (relativement à une théorie donnée) à la non-conservativité (de l'extension de cette théorie par cette proposition), on voit bien également que l'énoncé de la cohérence de A lui-même est « substantiel » relativement à A. Le même raisonnement montre que le Principe de réflexion sur A (i.e. « Tous les théorèmes de A sont vrais ».) est « substantiel » relativement à l'extension aléthique *minimale* de A. Inversement, il est clair que si l'on parvenait à montrer qu'en un sens ou un autre l'énoncé de la cohérence de A n'est pas une vérité épistémiquement « substantielle » relativement A, le fait que l'on puisse dériver cet énoncé dans l'extension aléthique tarskienne de A ne constituerait plus un indice du caractère épistémiquement « substantiel » des principes aléthiques tarskiens eux-mêmes. Dans cette section, je ne vais pas chercher à analyser la théorie tarskienne directement. Ma stratégie va consister en une attaque directe du caractère épistémiquement « substantiel » de l'énoncé de la cohérence de A et du principe de réflexion sur A, relativement à A. Je reviendrai sur le statut de la théorie tarskienne dans la section suivante.

6.2.1 Introduction

Nous avons défini la notion de conséquence réflexive. D'après la définition, il est clair que si un énoncé ϕ est logiquement dérivable d'un ensemble d'énoncés Γ, alors ϕ est une conséquence réflexive de Γ : on peut montrer que nos justifications pour Γ justifient ϕ, la démonstration consistant dans la preuve de ϕ sous les prémisses Γ. En introduction, je voudrais maintenant attirer l'attention sur le fait suivant :

Fait 1. *Il y a des ensembles d'énoncés tels que l'ensemble de leurs conséquences logiques est un sous-ensemble strict de l'ensemble de leurs conséquences réflexives.*

Considérons en effet la théorie T axiomatisée par une infinité de formules atomiques attribuant chacune la propriété P à un objet d'un univers du discours supposé infini, les entiers par exemple, et de sorte que P est attribuée à tous les objets de l'univers : T :

1. P1
2. P2
3. ...
4. Pn
5. ...

Pour présenter cette théorie, nous avons recours à une astuce d'écriture, les points de suspension, moyennant laquelle il n'y a pas de difficulté à comprendre de quelle théorie il s'agit. De plus, il est facile de voir que la théorie T n'a pas pour conséquence déductive l'énoncé général affirmant que tous les objets sont P : un énoncé universel n'est pas déductible de ses seules instances. Maintenant considérons cette théorie, non pas simplement comme un ensemble d'énoncés, ou de propositions, mais dans sa relation à l'exercice du jugement d'un sujet, comme objet possible de cet exercice. Supposons à nouveau qu'un axiome correspond à l'expression d'un jugement réellement effectué par un sujet, action qui scelle le fait qu'il se tient pour justifié à accepter cet axiome, à partir duquel il peut ensuite inférer d'autres jugements. Un sujet ne peut être dans ce genre de relation l'ensemble des axiomes d'une théorie que s'il a effectué un certain nombre de jugements, qui sont autant d'actions mentales. Mais un sujet fini ne peut jamais accomplir qu'un nombre fini d'actions. Un sujet fini ne peut donc juger ou accepter chaque axiome de T individuellement, ces axiomes étant en nombre infini. Il est donc clair que s'il accepte (ou croit) les axiomes de T, ce ne peut être que d'une façon indirecte, et en vertu du fait qu'il accepte (ou croit) d'abord un énoncé du type « Tous les objets (entiers) sont P », peut-être quelque autre généralisation plus faible, mais portant toujours sur une infinité d'objets. Or aucune de ces généralisations infinies qu'il est requis qu'un agent accepte pour pouvoir être dit accepter (indirectement) toutes les instances de T n'est conséquence logique des axiomes de T. Par conséquent, il y a quelque chose qu'un agent doit accepter (croire) s'il accepte les axiomes de T et qui n'est pas conséquence logique de ces axiomes. Le point important est que l'on ne peut croire les axiomes de T que si l'on croit en fait des axiomes qui justifient une théorie

logiquement plus forte que T, et cela quand bien même chaque axiome de T serait par ailleurs évident par lui-même.

6.2.2 Théorie minimale et finitude

Avant d'en venir à la défense de la thèse principale de cette section, une étape est encore nécessaire. Son rôle est double. D'une part, parce que dans la suite j'aurai besoin de tenir le compte rigoureux des ressources théoriques requises pour accomplir différents types de tâches, il me faut revenir, très brièvement, à l'étude des propriétés logiques de la théorie minimale de la vérité pour ajouter quelques précisions à ce que nous en avons dit au chapitre 2. D'autre part, sans préjuger encore de ce que sera ma réponse au Problème de l'usage, je voudrais fixer une terminologie en arrêtant stipulativement une certaine signification de la phrase « le prédicat de vérité permet d'exprimer des généralisations ».

Nous avons déjà vu aux chapitres 2 et 3 que $M(A)$ l'extension aléthique minimale de A, était une extension conservative de A.[18] Nous avons vu également que $T(A)$, l'extension aléthique tarskienne de A[19], est à plusieurs égards une théorie plus forte que $M(PA)$.[20] Les points importants sont résumés dans le tableau suivant :

[18] À nouveau, A contient sa propre syntaxe.

[19] On suppose toujours que A contient sa propre syntaxe. Pour simplifier la présentation, je supposerai également que A est une théorie arithmétique contenant un nom pour chaque entier. Cette hypothèse permet de se passer du détour par la notion de satisfaction dans la formulation de la théorie tarskienne. Voir note suivante.

[20] Pour rappel :
J'appelle théorie minimale de la vérité pour un langage L l'ensemble d'énoncés suivants :

$$Vr(\ulcorner\phi\urcorner) \leftrightarrow \phi \ (\phi \in L)$$

J'appelle *extension aléthique minimale* de la théorie A, et noterai $M(A)$ la théorie constituée des axiomes suivants :

- Les axiomes de A
- toutes les instances du schéma
$$Vr(\ulcorner\phi\urcorner) \leftrightarrow \phi \quad (\text{ for } \phi \in \mathcal{L}_A)$$
- toutes les nouvelles instances des schémas de la théorie A que l'on peut former dans le vocabulaire étendu (celui de A et des termes aléthiques).

Par constraste, et pour rappel, la théorie tarskienne de la vérité, T, est composée des axiomes suivants :

1. Un nombre *fini* d'instance du schéma-T :
$$Vr(\ulcorner\phi\urcorner) \leftrightarrow \phi, \text{ pour } \phi \text{ atomique seulement}$$
2. des clauses inductives quantifiant sur les énoncés :

	$T(PA)$	$M(PA)$
conservativité sur PA	Non	Oui
$?\vdash \forall x(Vr(x) \vee Vr(\neg x))$	Oui	Non
$?\vdash$ Tous les axiomes de A sont vrais	Oui	Non
$?\vdash$ Tous les théorèmes de A sont vrais	Oui	Non
$?\vdash Coh_{PA}$	Oui	Non

Nous l'avons dit à plusieurs reprises, l'extension aléthique minimale d'une théorie ne permet pas de *prouver* des *généralisations* portant sur la vérité, c'est la source de la faiblesse de la théorie.[21]

Toutefois, et c'est mon premier point, il faut noter que cette faiblesse ne concerne que les généralisations ayant une *infinité* d'instances. Si nous voulons généraliser sur un nombre *fini* d'instances, la théorie minimale donne ce que l'on en attend. Supposons par exemple que A est *finiment* axiomatisée. Alors il est possible dans ce cas de prouver dans $M(A)$ que tous les axiomes de A sont vrais :
À partir de la définition dans la syntaxe de A de l'ensemble des axiomes de A, c'est-à-dire de l'énoncé

$$\forall x(Ax_A(x) \leftrightarrow x = \ulcorner \phi \urcorner)$$

où ϕ est la conjonction des axiomes de A, et de

$$\phi \leftrightarrow Vr(\ulcorner \phi \urcorner) \text{ (axiome de la théorie minimale),}$$

on infère

$$\forall x(Ax_A(x) \rightarrow Vr(x))$$

– $\forall \phi \quad Vr(\ulcorner \neg \phi \urcorner) \leftrightarrow \neg Vr(\ulcorner \phi \urcorner)$
– $\forall \phi \quad Vr(\ulcorner \phi \wedge \psi \urcorner) \leftrightarrow Vr(\ulcorner \phi \urcorner) \wedge Vr(\ulcorner \psi \urcorner)$
– $\forall \phi \quad Vr(\ulcorner \forall x \phi \urcorner) \leftrightarrow \forall \text{terme clos } c, Vr(\ulcorner \phi(c) \urcorner)$

L'extension aléthique tarskienne de A, notée $T(A)$ est composée des axiomes suivants :

– Les axiomes de A
– Les axiomes de T (dans le vocabulaire approprié)
– + toutes les instances des schémas de A que l'on peut formuler dans le vocabulaire étendu.

Ces formulations n'ont de sens qui si nous supposons dans chaque cas que la syntaxe de A est codée dans la théorie A, et que le langage de A contient des termes clos pour désigner chaque individu de l'univers du discours de A. Dans le cas contraire, il faut reformuler la théorie en terme de satisfaction. Il n'y a rien là d'essentiel pour notre propos. Pour fixer les idées on peut supposer que A n'est autre que PA, la théorie de l'arithmétique de Peano en premier ordre.

[21] Voir les chapitres 2 et 3.

Plus généralement on a :

Fait 2 (Théorie minimale et généralisations finies)**.** *Soit A une théorie capable de représenter sa propre syntaxe, et E un ensemble fini de théorèmes de A. Alors dans $M(A)$ on peut représenter E par une formule ϕ et prouver : $\forall x(\phi(x) \rightarrow Vr(x))$.*

Démonstration. On pose $E = \{\sigma_0, \cdots, \sigma_n\}$, et $\phi(x) \leftrightarrow_{def} x = \sigma_1 \vee \cdots \vee x = \sigma_n$. Puisque les énoncés de E sont des théorèmes de A, on a : $M(A) \vdash Vr(\sigma_i)$, pour $0 \leq i \leq n$. Donc $M(A) \vdash \forall x((x = \sigma_1 \vee \cdots \vee x = \sigma_n) \rightarrow Vr(x))$ et donc $M(A) \vdash \forall x(\phi(x) \rightarrow Vr(x))$. □

J'en viens à mon second point, concernant l'expression des généralisations. Il y a certainement un sens clair dans lequel le déflationniste doit exiger que le prédicat de vérité permette d'*exprimer* des généralisations : au sens où un énoncé général (universel disons) liant une variable qui est un argument du prédicat de vérité doit *impliquer* non seulement les instances de substitutions ordinaires de cet énoncé général, mais également les énoncés obtenus à partir du premier en remplaçant $Vr(x)$ par un énoncé quelconque. Il faut que d'une affirmation de l'énoncé « Tous les axiomes de l'arithmétique sont vrais » il soit possible de dériver tout énoncé dont on serait en mesure de prouver qu'il est un axiome de l'arithmétique. Je dirai que d'une théorie de la vérité qui permet de rendre compte de la dérivabilité des instances (au sens où j'emploie ce mot ici) d'un énoncé général, que son prédicat de vérité permet d'*exprimer des généralisations*, ceci, bien entendu, sans préjuger de la réponse à donner au Problème de l'usage.

Ceci posé, il est crucial de noter que la théorie minimale de la vérité est également suffisante pour *exprimer* des généralisations infinies, au sens que nous venons de préciser. Volker Halbach a montré la chose suivante [22] :

Fait 3. *Soit $\Phi(x)$ une formule de L_{PA} dont x est la seule variable libre, Ref_Φ l'énoncé « $\forall x(\Phi(x) \rightarrow Vr(x))$ » et soit PA' la théorie PA étendue par tous les énoncés de la forme*

$$\Phi(\ulcorner\psi\urcorner) \rightarrow \psi$$

où ψ est un énoncé de L_{PA}. Alors PA' et $M(PA) + Ref_\Phi$ prouvent exactement les mêmes formules de L_{PA}.

[22] Halbach (1999*b*), p.13.

Par conséquent si, dans PA, on peut prouver $\Phi(\ulcorner s \urcorner)$ pour tout énoncé s satisfaisant le prédicat Φ[23], alors l'énoncé Ref_Φ implique dans $M(PA)$ tous les énoncés s satisfaisant Φ. La théorie minimale de la vérité permet de rendre compte du fait qu'en affirmant « tous les énoncés qui sont Φ sont vrais » je suis en mesure de prouver tous les énoncés dont je peux prouver qu'ils sont Φ.

L'analogie avec la notion ordinaire d'expression de la généralité est la suivante : de même que l'on ne peut inférer de ce que *10 est pair*, *100 est pair*, etc., que *tous les multiples de 10 sont pairs*, tandis que l'on peut inférer de ce que *tous les multiples de 10 sont pairs*, et de ce que *10 est un multiple de 10*, *100 est un multiple de 10*, etc., que *10 est pair*, *100 est pair* etc., de même la théorie minimale ne rend pas compte de l'inférence (incorrecte) de ce que *0 n'est pas un successeur*, que *tous les entiers ont un successeur* etc., à la conclusion que *tous les axiomes de Peano sont vrais*, mais rend compte du fait que l'on peut inférer de ce que *tous les théorèmes de PA sont vrais*, et de ce que *« 0 n'est pas un successeur » est un axiome de PA*, *« Tous les entiers ont un successeur » est un axiome de PA*, etc., que *0 n'est pas un successeur*, etc. En ce sens, si la théorie minimale ne permet pas de prouver des généralités, elle permet d'*exprimer* des généralités, ou plus précisément d'expliquer l'*usage du prédicat de vérité pour l'expression de généralisations* sur des positions d'énoncés.[24]

En résumé : 1. $M(PA)$ est conservative sur PA, 2. $M(PA)$ ne permet de prouver aucune généralisation infinie sur la vérité, mais 3. $M(PA)$ permet de prouver des généralisations finies et 4. d'*exprimer* des généralisations infinies. Nous reviendrons sur la question du caractère adéquat* ou non de la théorie minimale de la vérité à la fin de ce chapitre. Entre temps, nous adopterons la théorie minimale et exploiterons librement les propriétés que nous venons de rappeler.

6.2.3 Les axiomes comme expression du jugement

Un axiome d'une théorie contentuelle, au sens classique du terme « axiome », n'est pas seulement un énoncé interprété ou même une proposition, au sens moderne de ce terme, c'est également l'expression d'un *jugement*, au sens où Kant et Frege

[23] Ce qui sera le cas pour les ensembles d'énoncés récursivement énumérables, en particulier les prédicats syntaxiques comme « être un axiome » d'une théorie récursivement axiomatiée, ou « être un théorème » d'une théorie récursivement axiomatisée.

[24] Voir chapitre 1.

employaient ce mot,[25] et d'un jugement dont la correction est tenue pour non-problématique, et mois problématique en général que les jugements qui peuvent ensuite être inférés d'eux. Les axiomes doivent être *évidents*, c'est-à-dire que nous les tenons pour justifiés sans que la médiation d'autres jugements ne soit nécessaire à la reconnaissance de leur vérité.

Reprenons maintenant le raisonnement déjà esquissé plus haut à propos d'une théorie infiniment axiomatisée. Dans le cas où une théorie est *finiment* axiomatisée, il se peut que chaque axiome soit l'expression d'un jugement qui a effectivement été porté par un sujet qui en aurait reconnu l'évidence. Je peux juger que 0 n'est pas un successeur ou que tout nombre entier a un successeur, il n'y a là aucune difficulté. Mais dans ce cas, c'est ce que nous avons vu au paragraphe précédent, il est possible de prouver dans l'extension aléthique minimale de cette théorie[26] que tous les axiomes de la théorie en question sont vrais. Plus précisément, si A est finiment axiomatisée, on peut prouver dans l'extension aléthique minimale de A l'énoncé « Pour tout x, si x est un axiome de A, alors x est vrai ».

Mais en fait la plupart des théories qui nous intéressent ont un ensemble infini d'axiomes. C'est vrai de l'arithmétique de Peano en premier ordre, c'est vrai de la théorie des ensembles ZF (en premier ordre), et c'est vrai en général de toutes les théories pour l'axiomatisation desquelles nous avons recours aux *schémas d'axiomes*. Dans ce cas, la relation d'un sujet aux axiomes de la théorie doit être quelque peu différente de celle que nous venons de décrire. Comme nous l'avons vu tout à l'heure, nous ne pouvons jamais avoir porté, à un moment donné, qu'un nombre fini de jugements. Par conséquent l'ensemble des jugements que nous avons portés et qui nous qualifie comme acceptant un ensemble infini d'axiomes ne peut lui-même pas être infini. Par conséquent, si nous pouvons être dits accepter ces axiomes, il ne peut s'agir que d'une acceptation indirecte, conséquence de quelque autre jugement que nous avons porté.

Quel jugement ? Il y a une analogie structurelle entre cette limitation de notre faculté de juger (l'impossibilité de porter une infinité de jugements), et l'impossibilité de faire une infinité d'affirmations ou d'assertions, le jugement pouvant être vu comme la contrepartie mentale de l'acte linguistique d'assertion.[27] Il est donc assez

[25] Sur le sens anciennement reçu du terme de *proposition*, et le rapport avec l'emploi actuel du terme « jugement », nous renvoyons aux remarques de Martin-Löf (1996).

[26] L'extension aléthique contenant également la syntaxe de la théorie de base.

[27] L'idée que le jugement et l'assertion sont les contreparties mentale et linguistique l'une de l'autre se trouve chez Frege (voir chapitre 1). On pourrait nuancer les choses, et associer plutôt

naturel, pour comprendre quels jugements nous portons effectivement (en vertu desquels nous pouvons être dits accepter les axiomes de la théorie), de tourner notre regard vers ce que nous affirmons, quand nous affirmons les axiomes d'une telle théorie. Ce que nous faisons, faute de capacités infinies d'expression, c'est formuler les axiomes de ces théories infiniment axiomatisées en employant des *schémas* (ou des points de suspension). Mais un schéma, à strictement parler, n'exprime aucune proposition, c'est un simple artefact de présentation. D'où la question : quelle est la proposition que nous exprimons réellement en utilisant ces schémas, et quel énoncé standard, doué de sens, l'exprime ? Nous sommes dans un cas typique où le besoin de généralité nous force à opérer une montée sémantique parce que nous avons besoin de généraliser en position d'énoncés, comme nous l'avons vu au chapitre 1. L'explication naturelle et plausible de ce que nous affirmons lorsque, écrivant les axiomes de Peano, nous écrivons schématiquement

$$\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x')) \rightarrow \forall x \phi(x)$$

est que nous affirmons la proposition suivante

(R_1) Tous les énoncés de la forme $\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x')) \rightarrow \forall x \phi(x)$ sont vrais.

Ne peut-on *affirmer* ces axiomes sans recourir à la notion de vérité ? Si l'on admet, et je l'admets, la thèse quinienne selon laquelle les autres voies connues que sont les quantification substitutionnelles, propositionnelles ou du second-ordre sont problématiques ou supposent en fait pour être comprises de recourir à la notion de vérité[28], il semble bien que le passage par la vérité soit indispensable. Certes, on pourrait peut-être décrire ce que nous faisons, ou ce que nous acceptons, simplement en disant :

(R_2) J'accepte/j'asserte tous les énoncés de la forme

assertion et ce que Kant désignait plus spécifiquement comme l'une des modalités du jugement, le jugement assertorique. Mais en fait, et je fais cette remarque simplement en passant, il ne semble pas que le problème que nous pointons ait fondamentalement à voir avec le type d'acte linguistique ou mental que nous considérons : que l'on parle d'assertion, d'interrogation, ou d'autre chose encore, le problème est le même. Ce qui est en jeu est acte plus primitif, un acte qui semble être constitutif de tous ces actes et qui a à voir avec la prédication elle-même, en tout cas avec la simple *expression* d'un contenu, que ce contenu soit ensuite affirmé, mis en doute, interrogé, etc. On ne peut pas affirmer une infinité de phrases, mais l'on ne peut pas davantage poser une infinité de questions. Quoi qu'il en soit de ce point, je m'en tiendrai au registre du jugement ici, en notant tout de même que le problème que je soulève ne dépend pas essentiellement de la question de savoir si le jugement est, dans une terminologie kantienne, « assertorique » ou »problématique ».

[28] Voir chapitre 1, section 4.

$$\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x')) \rightarrow \forall x \phi(x)$$

R_2 pourrait être une description correcte de ce que nous faisons, mais justement ce que nous acceptons lorsque nous acceptons le schéma d'induction ne porte pas sur ce que nous faisons ou acceptons, mais seulement sur les entiers. Le contenu de R_2 est largement différent du contenu de ce que nous acceptons, et R_2, par lui-même, n'implique pas les instances du schéma d'induction. R_2 impliquerait bien les axiomes d'induction moyennant un principe auxiliaire du type :

$$\text{J'asserte } s \rightarrow p$$

où il faut remplacer « p » par un énoncé et « s » par un nom de cet énoncé , mais il est clair qu'un tel principe est tout simplement faux. R_1, en revanche, semble bien accomplir exactement le travail demandé : en disant que toutes les instances du schéma d'induction sont vraies, nous ne faisons ni plus ni moins, en substance, qu'affirmer ces instances en référant à leur squelette syntaxique.[29] Il y a un petit surplus de contenu hétérogène aux axiomes eux-mêmes, à cause de la référence à la syntaxe, mais c'est la meilleure approximation que nous ayons, et en cela le rapport de R_1 aux axiomes de départ est en tout point comparable à celui qu'entretient l'ensemble des énoncés singuliers à l'énoncé général obtenu *via* un prédicat auxiliaire.[30] Nous sommes exactement dans une de ces situations où, comme le notait Quine, pour les besoins de l'expression d'une certaine généralité, nous devons avoir recours à la montée sémantique, puis au prédicat de vérité pour « annuler » la référence linguistique.

Le point général est que, puisque nous ne pouvons porter qu'un nombre fini de jugements, il faut, en présence d'une infinité d'axiomes, que notre pensée se rapporte à eux par une *description*, une description sous laquelle leur vérité s'impose à nous. Mais une description n'est qu'un nom, ce n'est pas l'expression d'un jugement, ni une affirmation, et nous avons alors *besoin* du prédicat de vérité pour dénominaliser ces axiomes, pour faire de cet ensemble d'axiomes, sous la description qui en a été donnée, le contenu d'un jugement possible.

[29] On pourrait à partir de là construire un argument en faveur de l'idée que nous avons de toute façon besoin d' un prédicat de vérité déflationniste transparent pour accomplir ce genre de tâche expressive. Mais ce que nous voulons montrer à terme est plus fort, à savoir que nous n'avons besoin que de ce prédicat de vérité, au moins pour rendre compte des usages de la vérité dans les preuves de cohérence.

[30] Sur le modèle : Tom est mortel, Jack est mortel etc. ⇒ Tous *les hommes* sont mortels, où « homme » est le prédicat auxiliaire qui fixe ce que j'appellerai *le champ* de la généralité.

Ce que nous avons dit dans le cas des axiomes d'inductions est que, si nous les acceptons, le jugement que nous portons, et en vertu duquel nous pouvons être dits accepter ces axiomes, celui qui scelle l'engagement de notre pensée avec leur contenu, ne peut prendre d'autre forme que celle de l'attribution d'un prédicat aux axiomes présentés sous une certaine description, une des (éventuellement) multiples descriptions sous lesquelles nous connaissons les axiomes en question. Dans le cas des axiomes de Peano, la façon dont nous les présentons indique que ce jugement n'est autre que R_1 : R_1 n'ajoute rien au contenu des énoncés qui sont jugés vrais, en dehors de l'appareillage conceptuel impliqué dans la description elle-même, et par ailleurs ces probablement sous cette description. Dans ce cas cette description est une description syntaxique, et ce sont surtout ces cas qui nous intéressent. Mais dans le cas général la description peut être ensembliste, par exemple « Tous les énoncés vrais dans le modèle standard de l'arithmétique », ou tout à fait extra-mathématique « Ce que Pierre m'a dit hier », « La première phrase écrite par Platon » etc. De fait, nous pouvons effectivement déterminer le ou les référents de la description dans certains cas seulement. Dans le cas du prédicat $Th_A(x)$=« x est un théorème de la théorie A », où A est une théorie récursivement axiomatisée, si s est dans l'extension du prédicat Th_A alors on peut prouver $Th_A(s)$ dans PA, mais si s n'est pas un théorème, on ne peut pas prouver $\neg Th_A(s)$ dans PA.

Si ce que nous avons dit jusqu'ici est correct, alors une conséquence s'impose : si nous avons une justification pour les axiomes d'induction, ce ne peut être qu'une justification sous une certaine description de ce genre, et en dernière instance, il s'agira d'une justification de l'énoncé R_1 lui-même. Ainsi, pour continuer avec l'exemple des axiomes de Peano en premier ordre, nous voyons que s'ils sont justifiés, ils ont le même ensemble de justifications que la théorie finiment axiomatisée suivante, que j'appelerai PA^+ :

1. $\forall x \neg(0 = Sx)$
2. $\forall x \forall y(Sx = Sy \rightarrow x = y)$
3. $\forall x(\neg(x = 0) \rightarrow \exists y(x = Sy))$
4. $\forall x(x + 0 = x)$
5. $\forall x \forall y(x + Sy = S(x + y))$
6. $\forall x(x.0 = 0)$
7. $\forall x \forall y(x.S(y) = (x.y) + x)$

8. Tous les énoncés de la forme $\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x')) \rightarrow \forall x\phi(x)$ sont vrais.

A partir là, et en vertu de ce que nous avons noté au paragraphe précédent sur les propriétés de la théorie minimale, il est facile de poursuivre et de prouver que tous les axiomes de la théorie sont vrais, dans une extension conservative de la théorie de la syntaxe de PA. Pour peu que l'on identifie PA et la théorie de sa syntaxe, alors nous aurons la preuve que tous les axiomes de PA sont vrais dans PA^+ augmenté des seules équivalences-T.[31]

L'idée que nous cherchons à défendre n'est pas que, pour un sujet donné, il pourrait se trouver que son acceptation des axiomes soit justifiée par le jugement que « tous les axiomes sont vrais » tandis que, pour d'autres sujets, cette justification viendrait d'ailleurs. Il ne s'agit pas de dire ici que le principe de réflexion *sur les axiomes* est une conséquence α-réflexive des axiomes, pour un certain agent α. Ce que nous voulons dire est que, pour tout sujet fini, *toute justification de l'ensemble des axiomes permettra d'inférer que tous les axiomes sont vrais*, puisque au fond

[31] A partir de PA^+ et des équivalences-T on dérive :

1. « $\forall x\neg(0 = Sx)$ » est vrai
2. « $\forall x\forall y(Sx = Sy \rightarrow x = y)$ » est vrai
3. « $\forall x(\neg(x = 0) \rightarrow \exists y(x = Sy))$ » est vrai
4. « $\forall x(x + 0 = x)$ » est vrai
5. « $\forall x\forall y(x + Sy = S(x + y))$ » est vrai
6. « $\forall x(x.0 = 0)$ » est vrai
7. « $\forall x\forall y(x.S(y) = (x.y) + x)$ » est vrai
8. Tous les énoncés de la forme $\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x')) \rightarrow \forall x\phi(x)$ sont vrais.

Dans PA on peut définir les axiomes de PA, comme nous l'avons déjà vu :

x est un axiome de PA ssi

1. x= « $\forall x\neg(0 = Sx)$ »
2. ou x= « $\forall x\forall y(Sx = Sy \rightarrow x = y)$ »
3. ou x=« $\forall x(\neg(x = 0) \rightarrow \exists y(x = Sy))$ »
4. ou x=« $\forall x(x + 0 = x)$ »
5. ou x=« $\forall x\forall y(x + Sy = S(x + y))$ »
6. ou x=« $\forall x(x.0 = 0)$ »
7. ou x=« $\forall x\forall y(x.S(y) = (x.y) + x)$ »
8. ou x est de la forme $\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x')) \rightarrow \forall x\phi(x)$.

A partir de la définition des axiomes de PA et des énoncés précédents, on dérive sans difficulté :

Tous les axiomes de PA sont vrais

donner une justification d'un *ensemble* d'axiomes donné ne signifie rien d'autre que donner une justification de ce que tous les axiomes de cet ensembles sont vrais. Reprenons par exemple le cas où un sujet possède une justification déférentielle pour les axiomes d'un théorie, PA par exemple, pour contraster la différence de statut entre le principe de réflexion sur les axiomes d'une théorie et le statut des axiomes de grands cardinaux relativement à la théorie des ensemble. Une telle justification déférentielle pour les axiomes de PA, puisque toute justification ne peut se fonder que sur un ensemble fini de jugements, ne pourrait elle-même contenir qu'un nombre fini d'instances de principes de déférence du type

Si Kurt affirme que p alors p,

ce qui constituera un support insuffisant pour la justfication de *tous* les axiomes de PA. Par conséquent cette justification déférentielle doit en fait contenir comme prémisse la généralisation elle-même, par exemple ici

Tout ce qu'affirme Kurt est vrai.

Or, à partir de « Tout ce qu'affirme Kurt est vrai » et de

Kurt affirme tous les axiomes de Peano,

si le sujet peut certes déduire les axiomes de Peano eux-mêmes[32], il lui est également facile de déduire, grâce à la théorie minimale uniquement

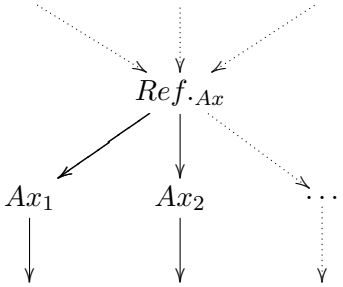
Tous les axiomes de Peano sont vrais.

Par conséquent, un sujet qui justifie les axiomes de Peano par un argument d'autorité est aussi en position de justifier l'affirmation que tous les axiomes de Peano sont vrais.

Je conclus sur la question de la justification des axiomes en résumant ce que nous avons vu. En un mot, si nous ne pouvons penser le contenu de l'ensemble des axiomes que *via* une description de ces axiomes, parce que cet ensemble est infini, alors le jugement le plus modeste à même de réaliser la relation épistémique dans laquelle se trouve le sujet du jugement qui endosse l'ensemble des axiomes d'une théorie est le jugement que tous les axiomes de cette théorie sont vrais. Si ce point

[32] À condition toutefois qu'il soit capable d'identifier les énoncés qui se cachent derrière la phrase nominale « les axiomes de Peano ».

est correct, alors *a fortiori*, toute justification qu'un agent peut avoir pour croire les axiomes de la théorie doit en fait être une justification du jugement que tous les axiomes de la théorie sont vrais, où « vrai » peut être simplement compris comme le prédicat minimal de vérité. Or puisque dans le cas où les axiomes sont en nombre fini, il est possible également de construire une justification de l'affirmation que tous les axiomes sont vrais à partir de n'importe quelle justification des axiomes et à l'aide de la théorie minimale de la vérité[33], le principe de réflexion sur les axiomes d'une théorie A est dans tous les cas, que les axiomes soient en nombre fini ou infinis, une conséquence réflexive des axiomes de A ou de l'extension aléthique minimale de A.[34]



Je conclus en reprécisant trois points. Nous avons vu si A contient sa propre syntaxe (ce que nous n'avons cessé de supposer), le principe de réflexion aléthique sur les axiomes de A implique les axiomes de A (modulo la théorie minimale de la vérité) et n'est pas impliqué par eux (et la théorie minimale de la vérité) : je n'affirme pas maintenant que le principe de réflexion est conséquence logique du contenu de A ni de l'extension aléthique minimale de A. Le second point est que je ne suis pas en train de proposer une notion « irréductiblement informelle » de conséquence logique telle que, en ce sens de la notion, le principe de réflexion sur les axiomes de A serait conséquence du contenu sémantique des axiomes A ou de l'extension aléthique minimale de A. Il n'y a rien d'irréductiblement informel dans ce qui est en jeu ici, et ce que j'avance n'a aucun rapport, de près ou de loin, avec une critique de la notion de conséquence logique en premier ordre. Enfin, je n'affirme pas non plus qu'accepter le principe de réflexion sur les axiomes de A serait *simplement*

[33] On suppose toujours que la syntaxe est disponible dans A

[34] Si l'on compte les règles minimales d'introduction et d'élimination du prédicat de vérité comme des règles logiques qui sont toujours à la disposition du sujet, on peut formuler la conclusion de façon plus nette en éliminant le deuxième membre de la conjonction.

une obligation rationnelle à laquelle est soumis tout sujet acceptant les axiomes A, obligation à laquelle par ailleurs il pourrait ou non se conformer. Ce que j'affirme est que le principe de réflexion sur les axiomes de A est une *conséquence déductive* de tout contenu qui est effectivement le contenu du jugement d'un sujet qui aurait fait siens le axiomes de A[35], jugement en vertu duquel le sujet peut être dit accepter les axiomes de A.[36] Si nous pouvions porter une infinité de jugements indépendants, les choses seraient radicalement différentes. Peut-être que dans ce cas, ayant fait nôtre une infinité d'axiomes par l'exercice direct de notre jugement, serions-nous tout de même dans l'« obligation rationnelle » de juger également que tous ces axiomes sont vrais ?[37] En tout cas le point qui était le mien n'aurait plus de raison d'être. Mais puisque nous sommes des êtres finis, nous ne pouvons porter ces jugements que médiatement, *via* un effort de conceptualisation permettant de décrire le champ de la généralité visée, puis en dénominalisant ces énoncés décrits, grâce au prédicat de vérité.

6.2.4 Jugements et théories

Au paragraphe précédent j'ai parlé de la relation d'un sujet acceptant une théorie aux axiomes de cette théorie. Pour aller plus loin, il faut maintenant se demander ce que cela signifie que d'accepter une *théorie*. Quels types de jugements un sujet doit-il avoir effectué pour pouvoir être dit accepter une théorie donnée ?

Il y a pour chaque théorie une infinité de façons de la décrire. Je peux parler de PA comme de la théorie la plus souvent mentionnée dans ce travail, ou comme de la théorie arithmétique la plus populaire chez les philosophes, et de mille autres manières encore. Chaque théorie possède cependant ce que l'on pourrait appeler un ensemble de *descriptions canoniques*, la décrivant comme un ensemble d'énoncés engendré par un certain nombre d'axiomes et de règles d'inférence. Par exemple, une description canonique de PA est donnée par la spécification de l'ensemble des

[35] Pour être tout à fait précis, je devrais dire : conséquence déductive d'une *extension conservative* du contenu du jugement porté par le sujet. Ceci pour les cas où A est finiment axiomatisée, et où la réflexion sur les axiomes s'obtient dans l'extension aléthique minimale de A, qui est conservative sur A. Je suppose toujours tacitement que A contient sa syntaxe.

[36] Comme me l'a fait remarquer Alejandro Perez Carballo, si ce qui précède est correct, alors un sujet qui affirmerait que ses propres croyances arithmétiques lui permettent de justifier tous les axiomes de PA en premier ordre, *et rien d'autre*, ment ou se trompe. C'est une conséquence que j'accepte sans difficulté.

[37] Dans ce cas, de toute façon, la question perdrait tout intérêt, et il est inutile de chercher à y répondre.

axiomes écrits au chapitre 2 de cette thèse et des règles classiques d'inférences logiques. D'une façon générale, une description canonique d'une théorie est un couple (Ax, R) formé d'un ensemble Ax d'énoncés et d'un ensemble R de règles, la théorie étant l'ensemble des énoncés engendré par R à partir de Ax. Mais bien que la donnée des règles fasse ordinairement partie de la façon dont nous décrivons ces ensemble d'énoncés que sont les théories, il faut garder à l'esprit que ces règles d'inférence ne font pas partie de qui est *affirmé* par la théorie, mais seulement ce moyen pour nous de décrire ce qui est affirmé par la théorie, c'est-à-dire par l'ensemble infini de ses énoncés (ses « théorèmes » justement).

La question qui m'intéresse est celle de savoir ce en quoi cela consiste que d'accepter ou de croire une théorie, ou pour poser la question autrement : quels jugements un sujet doit-il avoir effectué pour compter comme acceptant une *théorie* donnée ? Il est clair qu'accepter une théorie ce n'est pas seulement accepter des axiomes. Un logicien intuitionniste et un logicien classique pourraient se trouver d'accord sur leurs jugements relativement aux axiomes d'une théorie donnée, ils n'en accepteraient pas pour autant les mêmes théories, ne serait-ce que d'un point de vue extensionnel : le premier accepte l'ensemble des énoncés qui suivent des axiomes par les règles intuitionnistes, le second l'ensemble des énoncés qui suivent des axiomes par les règles classiques, et ces deux ensembles d'énoncés sont distincts. Mais on peut maintenant distinguer deux analyses de ce que cela signifie pour un sujet que d'accepter une théorie, un sens fort et un sens faible qui me semblent reconnaissables dans l'usage courant. Au sens fort, accepter une théorie c'est accepter une théorie sous une description, ce qui est une manière indirecte d'accepter explicitement tous ses énoncés. Au sens faible, accepter une théorie c'est l'accepter en un sens partiellement « dispositionnel » : un sujet juge que A, suit les règles R, donc il accepte virtuellement tous les énoncés dérivable de A par la règle R, alors même qu'il pourrait ne pas avoir identifié exactement ce que sont ces règles, et donc en un sens ne pas savoir ce qu'il accepte. Je veux montrer dans cette section que le principe de réflexion sur la *théorie* A, c'est-à-dire l'affirmation « tous les théorèmes de A sont vrais » est une conséquence réflexive, non de l'acceptation des axiomes de A, mais de l'acceptation de la théorie A. Puisqu'il semble y avoir deux sens dans lesquels on peut dire qu'un sujet accepte une théorie, je propose de procéder en deux temps. Je commence par le sens fort du terme accepter, parce que c'est pour elle que l'analyse est la plus rapide étant analogue à ce que nous avons vu pour les axiomes, puis je proposerai mon analyse de la relation « faible » et des conclusions

qu'il faut en tirer.

Au sens fort

Une théorie, comme ensemble d'énoncés clos pour un certain ensemble de règles R est un ensemble infini d'énoncés, la pensée ne peut se rapporter à eux que par description, et accepter une théorie au sens fort c'est accepter l'ensemble de ses théorèmes sous quelque description, par exemple *tous les théorèmes de A* ou, de façon plus précise, *l'ensemble des énoncés engendrés par tels et tels axiomes et telle et telle règle d'inférence*, *la théorie la plus souvent mentionnée dans ce travail*. En vertu de l'argument développé dans la section précédente le sujet qui accepte la théorie en ce sens fort juge en fait que

tous les théorèmes de A sont vrais

ou que

l'ensemble des énoncés engendré à partir des axiomes
par telle et telle règles d'inférence sont vrais

pour une certaine description sous laquelle il est en position de décrire l'ensemble des énoncés qu'il accepte. Dans tous les cas, le jugement porté par le sujet va lui permettre de justifier en fait le principe de réflexion sur la théorie A, à savoir l'énoncé « tous les théorèmes de A sont vrais », sous l'hypothèse modeste que le sujet sait que la théorie est un ensemble de théorèmes et que « A » est un nom de la théorie dont il parle.

Plus précisément, ce que le sujet est en mesure de justifier dépend de la description sous laquelle il identifie la théorie. Si, par exemple, un sujet accepte l'arithmétique de Peano sous la description

« La théorie la plus souvent mentionnée dans ce travail »

alors, sachant qu'une théorie est simplement l'ensemble de ses théorèmes, il peut bien tirer de ce qui justifie son jugement que

tous les théorèmes de la théorie la plus souvent
mentionnée dans ce travail sont vrais,

mais il ne pourra pas justifier l'affirmation que

« 0 n'est pas un successeur » est vrai,

ou que tous les énoncés dérivables par telles et telles règles à partir des axiomes ayant telle et telle formes syntaxique sont vrais. Enfin, il ne pourra pas même dériver les axiomes et les théorèmes de la théorie elle-même... Pour pouvoir construire une justification pour ces énoncés à partir du jugement mentionné plus haut, il faut que le sujet puisse « identifier la théorie sous la description », que la théorie lui soit « présentée », ce qui sera le cas lorsqu'il aura reconnu sous la description ce que j'ai appelé une description canonique de la théorie. Je laisserai de côté les cas où cette condition n'est pas remplie et je m'en tiendrai aux cas dans lesquels le sujet *sait de quoi il parle*.

Dans la situation qui nous intéresse, celle d'un sujet qui *connaît* une théorie, qui en connaît une description canonique, non seulement les jugements généraux du type « Tous les axiomes (ou théorèmes) de PA sont vrais », mais également les attributions de vérité aux descriptions syntaxiques précises de ces axiomes et théorèmes[38], sont des conséquences réflexives de la théorie. Par conséquent, si la relation d'un sujet à une théorie est celle de l'acceptation au sens fort, alors toute justification qu'un sujet peut avoir pour accepter une théorie A justifie les formulations arithmético-syntaxiques précises de « Tous les théorèmes de A sont vrais ». Le principe de réflexion sur les théorèmes de A est donc une conséquence réflexive de A.

Au sens faible

Que penser à présent de la relation à une théorie A $(= (Ax, R))$ d'un sujet tel que les énoncés de Ax expriment le contenu de jugements qu'il a effectué et qui se trouve *suivre* les règles de R à l'aveugle, c'est-à-dire sans savoir qu'il les suit et sans les tenir pour justifiées. Dans ce cas il me semble que le sujet n'accepte pas véritablement la théorie en question, mais qu'il l'accepte seulement potentiellement ou implicitement. Je dirai qu'il accepte$_i$ les R-conséquences de A, ou qu'il accepte seulement implicitement la théorie en question.[39]

[38] Puisque l'on suppose toujours que le sujet possède au moins les moyens de faire la syntaxe élémentaire de tout langage.

[39] Il est donc naturel de s'attendre à ce que dans un processus d'actualisation de cette forme d'acceptation potentielle qu'est l'acceptation$_i$ une forme de généralité apparaisse dans le contenu de ce qui est accepté, correspondant à la reconnaissance de la généralité dont les règles sont implicitement porteuses.

Ce que nous voulons montrer est que toute justification que peut se donner un sujet pour accepter une théorie, en un sens très faible de justification, doit en fait justifier que tous ses théorèmes sont vrais. Dans le cas d'un sujet qui accepte$_i$ une théorie, précisément en tant qu'il n'accepte pas de façon pleine et complète la théorie, n'a pas encore formulé le contenu latent dans ce qu'il accepte implicitement, ce qu'il devrait tenir pour justifié de croire s'il se tenait justifié à croire ou à accepter au plein sens du terme la théorie en question. Or quel est ce contenu, quelle justification un sujet peut-il avoir pour accepter$_i$ une théorie ? Cette justification ne pourra être seulement une justification des *axiomes* de la théorie, puisque accepter une théorie ce n'est pas seulement accepter les axiomes de la théorie. Il faut encore que le sujet justifie sa pratique inférentielle. Voyons à quelles conditions un sujet peut-il tenir sa propre pratique inférentielle pour justifiée. Il faut d'abord que le sujet explicite cette pratique pour en juger. L'effort réflexif par lequel le sujet parvient à dégager, en les formulant explicitement, les règles qui guident sa pratique inférentielle, est une première étape de ce processus, étape dans laquelle la notion de vérité n'a aucune part.[40] C'est une réflexion dont le résultat est couché dans la syntaxe sous la forme de propositions *générales* qui décrivent, à l'aide d'un prédicat binaire spécial, que je noterai « $\vdash$ », et de variables syntaxiques, les règles de la théorie en question.[41] Par exemple, si le sujet suit la règle du *modus ponens*, le fruit

[40]Il y a ici un problème bien connu lié à cette façon de décrire le processus : comment pouvons-nous jamais reconnaître que toutes les inférences que nous avons faites sont des instances d'une certaine règle donnée, puisque en principe nous n'avons jamais effectué qu'un nombre fini d'inférence, et qu'il existe une infinité de règles différentes dont ces inférences particulières sont des instances ? Plus profondément, en vertu de quel genre de *fait* pouvons-nous être dits suivre une certaine règle et non une autre ? Faut-il donc admettre la factualité de dispositions psychologiques ? Des explications non-psychologiques sont-elles possibles ? Une vaste littérature philosophique a été consacrée à ce problème, d'abord présenté dans Wittgenstein (2004) §185-243, dont les considérations sceptiques sont (largement) développées, précisées et discutées dans S. Kripke (1984). L'état de l'art dressé en 1989 par Paul Boghossian est toujours utile aujourd'hui (*cf.* Boghossian (1989)). En tout état de cause, ce n'est pas une question que nous avons l'intention de discuter ici. Disons que, si nécessaire, nous *supposons* que, d'une façon ou d'une autre, nous suivons des règles, que ces règles sont, pour certaines du moins, identifiables. De façon plus plausible, les règles que nous explicitons ne sont pas des formulations de règles que nous suivrions à notre insu, mais des extrapolations à partir de notre pratique, d'une portée beaucoup plus générale, et dont la correction s'impose à nous une fois que nous les avons formulées. Si l'on adopte une position sceptique en contestant l'idée que pourrait exister *le fait* que nous « suivons des règles », alors l'idée même que nous pourrions accepter$_i$ une théorie n'a plus de sens. Donc si un sujet accepte une théorie, il l'accepte au sens fort ou il ne l'accepte pas du tout.

[41]Voir le chapitre 2, pour une esquisse de définition des notions de la syntaxe d'une théorie. Attention, l'usage du symbole « $\vdash$ » ici est simplement compris comme un usage *descriptif* d'une certaine pratique ; on ne suppose pas *a priori* que la relation $\vdash$ satisfait de « bonnes » propriétés supposées saisir une certaine explication de la notion de validité, que cette explication soit sémantique

de cette réflexion sera couché dans l'énoncé

$$\forall x, y, ((x \text{ et } y \text{ sont des énoncés de mon langage}) \rightarrow \{x, x * \Rightarrow * y\} \vdash y$$

Il faut noter que cette étape du raisonnement dans laquelle le sujet doit formuler des propositions générales décrivant sa pratique inférentielle met crucialement en jeu des aperçus syntaxiques. C'est parce qu'il est capable d'identifier certains modèles fondamentaux et descriptibles syntaxiquement auxquels se conforme sa pratique inférentielle, qu'il peut décrire celle-ci en toute généralité. Mais bien entendu affirmer que pour tout énoncé x et tout énoncé y,

$$\{x, x * \Rightarrow * y\} \vdash y$$

ce n'est jamais que *décrire* une certaine pratique, ce n'est pas l'endosser ni poser une affirmation qui, si elle était justifiée, justifierait notre pratique. On aura beau avoir justifié le jugement que

$$\forall x, y(\{x, x * \Rightarrow * y\} \vdash y),$$

on n'aura nullement justifié la pratique du *modus ponens*, car le contenu de *ce* jugement est une simple description de cette pratique dans un métalangage d'où nous parlons de notre langage de départ. La question demeure donc : que doit justifier un sujet, quel jugement doit-il porter, pour compter comme regardant sa pratique inférentielle comme justifiée ?

Une règle n'est pas un contenu propositionnel, ce n'est pas l'objet d'une attitude propositionnelle. Mais ce pour quoi l'on peut construire des justifications sont des contenus propositionnels, des conclusions possibles de raisonnements (éventuellement de raisonnements triviaux en une seule étape). Il doit donc exister un énoncé, exprimant un contenu propositionnel, qui puisse être l'objet d'un jugement J tel que, parce qu'il a effectué J, un sujet puisse être dit tenir pour justifiées les règles en question, un jugement J qui doive être justifié par tout ce qui peut compter comme une « justification des règles ». Chercher un tel jugement J, ce n'est pas nier que la validité des règles puisse être *évidente*, ou que les règles elles-mêmes puissent être « auto-justifiantes », selon l'expression de Dummett ; c'est simplement dire qu'il y a là un jugement, le jugement de la validité des règles justement, tel que, si nous ne

ou inférentielle. Je reviendrai sur ce point.

l'avions pas porté, nous ne regarderions pas comme justifiées les conclusions inférées de nos jugements selon ces règles.[42]

Pour comprendre ce qu'une justification de notre pratique inférentielle doit justifier, considérons une autre représentation standard des règles, par exemple de la règle d'introduction du connecteur $\wedge$, telle qu'elle est donnée en déduction naturelle :

$$\frac{A \qquad B}{A \wedge B}$$

Ce format de représentation présente l'avantage de nous rappeler ce en quoi cela consiste que de suivre la règle d'introduction de la conjonction par un artifice de présentation dans lequel « A », « B » et « $A \wedge B$ » semblent maintenant occuper des *positions d'énoncés*.[43] Suivre une règle d'inférence c'est en effet passer d'un jugement, ou de plusieurs jugements, à un autre jugement (éventuellement dans des contextes hypothétiques), d'une affirmation à une autre. Il semble alors que pour justifier la pratique inférentielle dont le modèle a été explicité dans la présentation de la règle ci-dessus,[44] il soit suffisant, et nécessaire, en vertu de la signification de « si...alors », de justifier le schéma

$$\text{Si } A \text{ et si } B \text{ alors } A \wedge B$$

où A et B sont des pseudo-variables en positions d'énoncés et où $A \wedge B$ est également en position d'énoncé, *schéma dont chaque instance est dérivable par un sujet suivant la règle en question et la règle d'introduction de « si...alors... »*.[45] Et à nouveau, ce

[42] Nous sommes certainement justifiés à accepter les conséquences logiques d'une proposition justifiée que nous acceptons même si nous ne reconnaissons pas que nous le sommes. Mais la notion de justification à l'œuvre dans ces deux dernières remarques est une notion « externaliste » de justification. La notion que nous avons en vue, nous l'avons déjà souligné, est la notion *interne* de justification. C'est cette notion qui doit permettre de rendre compte du savoir comme *itinerarium mentis* du sujet de la connaissance, et de l'accès du sujet à ses justifications.

[43] Je dis « semble », puisque en fait, ce que nous faisons véritablement est décrire dans un métalangage une certaine relation binaire entre ensembles d'énoncé et énoncés, notons-la R, et affirmer simplement

$$\forall x, y(\{x, y\} R x * \overline{\wedge} * y)$$

Mais la représentation utilisée plus haut, en paraissant omettre le symbole prédicatif, qu'on le note R ou autrement (mais en fait, bien entendu, il est là, c'est simplement le trait horizontal) donne à voir la pratique elle-même dont elle fait la théorie.

[44] À nouveau dans le processus d'explicitation de ce modèle, une rôle essentiel est joué par l'identification de la *forme logique*, ou syntaxique, de certains énoncés, et cette explicitation mobilise donc des ressources conceptuelles qui relèvent de la syntaxe. Et à nouveau, la notion de vérité n'a pas de part à cette partie du processus.

[45] Si je suis la règle

que nous voulons dire quand nous disons que nous acceptons le schéma en question, ce n'est rien d'autre que le fait que nous acceptons la proposition générale

Pour tout énoncé x et pour tout énoncé y, si x est vrai et si y est vrai, alors $\ulcorner x \wedge y \urcorner$ est vrai,

où maintenant x, y sont des variables objectuelles standards, où « $\wedge$ » est mentionné et non plus utilisé[46], et où le prédicat de vérité est utilisé dans son rôle d'auxiliaire pour « l'expression des généralisations », au sens de la section 2.2. Autrement dit, à *suivre la règle* $\wedge$-Introduction, on s'engage à *justifier la proposition* générale que, pour tous énoncés x et y, si x est vrai et y est vrai, alors $\ulcorner x \wedge y \urcorner$ est vrai. Pour justifier notre pratique consistant à suivre la règle $\wedge$-Introduction, il faut au moins justifier cette proposition générale.[47]

Ce que nous venons de voir pour la règle d'introduction de la conjonction peut être généralisé à toutes les règles. On conclurait alors que pour justifier nos règles il faut justifier, par exemple, que si x est vrai et $\ulcorner x \rightarrow y \urcorner$ est vrai alors y est vrai ; et que si, si x est vrai alors y est vrai, alors $\ulcorner x \rightarrow y \urcorner$ est vrai ; etc. Autrement dit on montrerait que ce qu'il faut justifier est que chaque règle d'inférence « préserve la

$$\mathrm{R}\ \frac{A}{B}$$

pour un certain type d'énoncés A et B, alors, si je comprends le sens de « $\rightarrow$ », je peux justifier « $A \rightarrow B$ » trivialement de la façon suivante :

$$\frac{\overset{R}{\dfrac{[A]^1}{B}}}{\overset{(1)}{A \rightarrow B}}$$

[46] Conformément à l'idée déjà notée ci-dessus, selon laquelle la formulation de la règle dans sa généralité met essentiellement en jeu un aperçu syntaxique, aperçu dans lequel la possession du concept de vérité n'a aucune part.

[47] J'insiste sur le fait qu'il y a deux aspects bien distincts dans le processus de « propositionalisation » de la règle : l'introduction du conditionnel et l'introduction du prédicat de vérité. Le prédicat de vérité n'est utile que parce que les règles qui nous intéressent sont *générales* et il sert l'expression de la généralisation sur des énoncés en positions phrastiques. Si ce qui nous intéresse est une règle sans généralité, par exemple la règle d'inférence du jugement que Pierre n'est pas marié au jugement que Pierre est célibataire, la propositionalisation se fait simplement par l'introduction du conditionnel : pour justifier la pratique de l'inférence en question, je dois justifier le jugement conditionnel que si Pierre n'est pas marié, alors il est célibataire, et je n'ai nullement besoin du prédicat de vérité pour exprimer ce qu'il faut justifier.

Par ailleurs j'ai pris pour exemple une règle logique, mais c'est sans importance ici. Le caractère *logique* de la validité est un tout autre problème, et en effet pour justifier qu'une inférence est *logiquement* valide il faut montrer davantage que la seule préservation de la vérité. Il faut en outre qu'elle la préserve « nécessairement », le problème étant de dire précisément ce que signifie la nécessité en pareil cas. En tout état de cause, ce point est sans incidence ici.

vérité ». De plus, puisque les règles que nous suivons sont en nombre fini, il serait alors facile de montrer qu'avec une théorie syntaxique adéquate, on peut conclure de ce que chaque règle préserve la vérité que *toutes* les règles préservent la vérité, par un raisonnement analogue à celui qui nous a permis de montrer plus haut que tous les axiomes d'une théorie étaient vrais à partir d'une description des axiomes de la théorie et de la vérité de ces axiomes.[48]

Deux remarques pour conclure. Tout d'abord, à nouveau, le prédicat de vérité dont nous avons besoin est simplement le prédicat de vérité *minimal*, celui dont la théorie étend *conservativement* la théorie de la syntaxe : il est suffisant pour permettre l'expression de généralisations au sens précisé à la section 2.2. En second lieu, il est essentiel de garder à l'esprit, pour comprendre ce que nous sommes en train de faire, que nous *n'affirmons pas* au terme du raisonnement précédent que la préservation de la vérité est une *explication* ou une justification intéressante de la validité des règles, que la vérité joue un rôle dans l'élucidation de cette notion de validité des règles logiques. Nous disons *au contraire* qu'il s'agit là seulement de l'*expression* de la validité des règles, et que par conséquent les jugements de « préservation de la vérité » par une règle sont des jugements que *toute* justification des règles doit justifier.

6.2.5 Le principe de Réflexion aléthique et l'énoncé de la cohérence

Si ce que nous avons dit est correct, accepter au sens fort l'ensemble infini d'énoncés qui constitue une théorie donnée A, c'est posséder une description de cet ensemble d'énoncés et juger que tous les énoncés tombant sous cette description sont vrais. Si le sujet est en mesure d'identifier cet ensemble d'énoncés comme une certaine théorie A, ce que l'on peut supposer ici, alors il est en mesure de prouver l'énoncé « tous les théorèmes de A sont vrais », et donc toute justification qu'il a pour A peut être étendue en une justification de ce que tous les théorèmes de A sont vrais.

D'un autre côté, nous avons vu que si un sujet accepte$_i$, c'est-à-dire implicitement, une théorie A, tout ensemble de justifications qu'il peut avoir pour son acceptation implicite de A permet de justifier que :

1. Tous les axiomes de A sont vrais

[48] Voir la section 2.2. de ce chapitre.

2. Les règles d'inférences préservent la vérité

Or dans l'extension *minimale* de la théorie de la syntaxe ($M(PA)$), le Principe de Réflexion sur A :

Tous les théorèmes de A sont vrais

est une conséquence déductible des deux énoncés précédents.[49] Par conséquent toute justification que pourrait posséder un sujet pour accepter la théorie A lui permet de justifier le principe de réflexion sur A.

À présent, toujours dans l'extension minimale de la syntaxe, on peut conduire le raisonnement suivant :

1. $\vdash \forall x(Th_A(x) \rightarrow Vr(x))$
   (« Tous les théorèmes de A sont vrais »)
2. $\vdash 0 \neq 1$
   (ou tout autre théorème de A)
3. $\vdash Th_A(\ulcorner 0 \neq 1 \urcorner)$
   (syntaxe de A)
4. $\vdash Vr(\ulcorner 0 \neq 1 \urcorner)$ (1. et 3.)
5. $\vdash \neg Vr(\ulcorner 0 = 1 \urcorner)$ par 4. et les Vr-règles, par la dérivation suivante :
$$\dfrac{\dfrac{Vr(\ulcorner \neg(0=1) \urcorner)}{\neg(0=1)} \quad \dfrac{[Vr(\ulcorner 0=1 \urcorner)]^1}{0=1}}{\dfrac{\bot}{\neg Vr(\ulcorner 0=1 \urcorner)}\,1}$$
6. $\vdash \neg Th_A(\ulcorner 0 = 1 \urcorner)$. (1. et 5.)

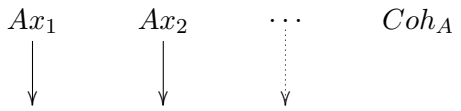
Autrement dit, à partir du principe de réflexion sur A, et de l'extension aléthique minimale de A, on peut construire une preuve de la cohérence de A.

Ces considérations appuie donc la thèse suivante, qui était attendue depuis le début du chapitre :

Proposition 1 (Thèse de la conséquence réflexive)**.** *Etant donnée une théorie A, le Principe de Réflexion sur A et l'énoncé de la cohérence de A sont des conséquences réflexives de A.*

[49] On suppose toujours bien entendu que le schéma d'induction dans la syntaxe peut être instancié par des formules contenant le prédicat de vérité.

L'analogie avec la situation que nous avons décrite en première partie de ce chapitre est la suivante. L'énoncé (standard) de la cohérence de la théorie A (supposée contenir sa propre syntaxe) est logiquement indépendant de la théorie A.



Mais toute justification que nous pourrions avoir pour accepter la théorie A[50] doit permettre de justifier le jugement que tous les axiomes de A sont vrais et que les règles d'inférences préservent la vérité, et par conséquent que tous les théorèmes de A sont vrais. Il y a néanmoins deux différences fondamentales entre la situa-

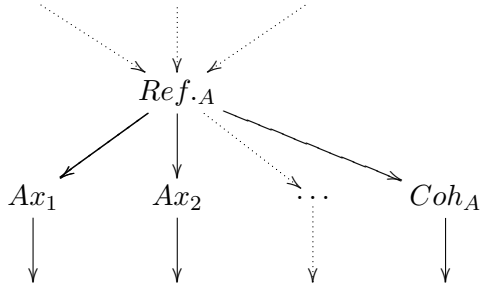


FIG. 6.4: L'énoncé de la cohérence de A est une conséquence réflexive de A.

tion épistémologique décrite en début de chapitre et celle-ci. La première est que la justification des axiomes de ZF par ce que nous avons appelé, faute de mieux, l'intuition de la hiérarchie cumulative, n'est pas *obligatoire*. Nous pourrions imaginer, et nous avons vu, qu'il est possible de justifier ces axiomes autrement (par déférence). En particulier il y a comme un excès de « contenu » de notre intuition de l'univers cumulatif relativement aux axiomes de ZF[51]. La seconde différence est que le principe de réflexion sur la théorie A justifie non seulement les axiomes de A, mais également les règles de A, au lieu que le contenu de l'intuition rationnelle de la

[50] Pas seulement les axiomes de A.

[51] Sinon, toute justification pour les axiomes de ZF justifierait potentiellement le contenu de l'intuition cumulative. Je n'exclus pas principe cette possibilité, mais je ne suis pas sûr que cette idée soit fidèle à l'esprit de Gödel.

hiérarchie cumulative ne justifie pas à soi seul que nous soyons justifiés à accepter les conséquences des axiomes de ZF.

Nous avons vu que notre acceptation du principe de réflexion sur A pouvait être vue, dans le cas où A était acceptée au sens faible, comme fondée sur l'explicitation des engagements que nous prenions dans notre *pratique* inférentielle consistant à suivre les règles de A. Il est possible que nous suivions à l'aveugle certaines règles, et qu'ainsi nous acceptions implicitement une théorie A (« accepter au sens faible »). Mais alors, d'une part, nous n'avons en général que des justifications aveugles pour les théorèmes autres que les axiomes et, d'autre part, nous contractons par notre pratique une dette théorique dont le montant est explicité, au terme d'un processus de réflexion, dans le principe de réflexion sur A. Accepter la théorie A, non pas implicitement, mais en toute connaissance de cause, c'est accepter ce principe. Il est tentant alors de formuler une définition rendant compte de ce qu'affirmer l'incohérence d'une théorie que l'on accepte est une contradiction pragmatique. C'était déjà en substance l'idée de Myhill[52], qui remarquait que :

> Il est aussi contradictoire d'employer des méthodes de preuves sans admettre leur correction, que de faire des affirmations sans admettre leur vérité. (Je n'utilise pas 'contradictoire' au sens de la logique formelle, mais à peu près comme un synonyme pour 'irrationnel').

On pourrait définir une notion (très circonstanciée) de contradiction pragmatique, c'est-à-dire d'actions contradictoires (et non de contenus propositionnels contradictoires) en posant qu'il y a contradiction pragmatique lorsque le sujet contracte par sa pratique un ensemble d'engagements (ici des engagements purement épistémiques formulés en termes de justification) qui ne peuvent pas être simultanément satisfaits ; en particulier lorsqu'il ou accepte$_i$ une théorie A et un énoncé ϕ tel que ϕ n'est pas consistant avec l'ensemble des conséquences réflexives A.[53]

[52] Voir chapitre 4

[53] À nouveau, l'explicitation de ces engagements passe par la découverte de l'ensemble R des règles suivies par l'agent puis l'expression des engagements dont elles sont porteuses exprimé en toute généralité. Si un sujet α suit les fameuses règles pour « tonk » données par Prior (1960),

$$\frac{A}{A \text{ tonk } B} \qquad \frac{A \text{ tonk } B}{B}$$

(et les règles usuelles pour la flèche), il peut montrer, pour tous les énoncés particuliers A et B, l'énoncé correspondant de la forme

$$A \rightarrow A\text{tonk}B$$

6.3 Théorie minimale et théorie tarskienne

Adéquation* de la théorie minimale

Si ce que nous avons dit est correct, nous avons également à présent une réponse au Problème de l'usage. Le déflationniste affirmait que le prédicat de vérité devait permettre *d'exprimer* des généralisations. Nous avons vu que le prédicat minimal de vérité permet de s'acquitter de cette tâche si l'on admet qu'« exprimer des généralisations » doit s'entendre comme l'idée que d'une attribution de vérité il doit être possible de dériver les énoncés auxquels la vérité est attribuée.[54]

Il y a des usages putatifs que la théorie minimale de la vérité n'explique pas, nous l'avons vu aux chapitres 2 et 3 : c'est la possibilité d'inférer d'une théorie A et d'une théorie de la vérité ce que nous avons appelé au chapitre 2 les généralisations logiques et théoriques, en particulier le principe de réflexion sur A. Shapiro avait suggéré qu'un tel usage était ancré dans notre emploi discursif du concept de vérité.

simplement en utilisant la règle pour « tonk » et l'introduction de la flèche. La généralisation affirme, elle, que

$$\text{pour tous énoncés } x \text{ et } y, \text{ si } x \text{ est vrai alors } x * \overline{tonk} * y \text{ est vrai}$$

(où « vrai » est le prédicat minimal de vérité).

Cette remarque m'amène à préciser un point. Toute notre approche suppose la donnée d'un sujet du jugement, disons α, et tel que α soit porteur d'un certaine pratique inférentielle. Cette pratique inférentielle détermine une certaine notion « interne » de justification, que l'on pourrait appeler α-justification : un jugement j est α-justifié par un ensemble de jugements J s'il existe une suite d'inférences pratiquées par α permettant d'inférer le jugement j à partir des jugements dans J. Pour simplifier les choses, j'ai supposé antérieurement par défaut que la pratique de α était *correcte*, au sens où une α-justification d'un jugement j justifie bien j. Mais en droit notre point ne dépend pas de cette hypothèse. Le sujet qui nous intéresse « pourrait » suivre les règles pour « tonk ». Réfléchissant sur sa pratique (toujours avec des ressources syntaxiques pour pouvoir en décrire les détails) il constaterait qu'il suit une règle R telle que :

$$\forall x, y \quad x \vdash_R x * \overline{tonk} * y$$

et, « réalisant » que $\vdash_R$ décrit *sa* pratique inférentielle, il devrait ensuite embrasser la généralisation mentionnée plus haut, qui est l'expression standard de la validité de la règle en termes de préservation de la vérité. Mais bien entendu, en affirmant cet énoncé, le sujet n'a pas justifié sa pratique inférentielle elle-même ! Pour cela il faudrait qu'il *explique* la notion de *validité*, ce en quoi cela consiste pour une règle que d'être valide, et qu'il montre ensuite que les règles qu'il suit sont en effet valides. Sur ce point, et les différentes façon de s'y prendre (voir chapitre suivant). Ce que je voulais noter ici est que la conséquence réflexive n'a rien à voir, en droit, avec ce processus complexe de justification ; ce qui est en jeu dans ce chapitre est seulement la reconnaissance des engagements implicites dans notre pratique inférentielle telle qu'elle est.

[54] À condition d'être capable d'identifier syntaxiquement les énoncés qui tombent sous la description.

Il en voulait pour preuve la capacité d'un sujet acceptant une théorie A (et la syntaxe de A) à en inférer le principe de réflexion sur A. Mais ce que nous avons vu est précisément qu'il n'y a pas de tel usage du prédicat de vérité. Personne n'a jamais *inféré* d'une « théorie » A le principe de réflexion ; c'est dans l'autre sens que les choses se passent : celui qui accepte tous les théorèmes d'une théorie A ne les accepte que médiatement, *via* l'acceptation d'un contenu dont il faut au bout du compte que le principe de réflexion soit déductible. De même, nous n'avons pas plus de raison d'attendre d'une théorie de la vérité et de la quantification qu'elles nous permettent de déduire de l'ensemble des instances de « la loi tiers exclus », c'est-à-dire des instances du schéma $p \vee \neg p$, la généralisation $\forall x(Vr(x) \vee Vr(\neg x))$, que nous n'avons de raison d'attendre d'une théorie de la quantification qu'elle nous permette de déduire de $Po_1, \cdots, Po_n, \cdots$ la généralisation « Tous les o_n sont P ». Accepter la loi du tiers-exclu, *c'est* accepter la généralisation. Il semble donc qu'on ait surestimé d'un côté l'extension des usages du prédicat de vérité dont une théorie de la vérité doit rendre compte (Problème de l'usage), tandis que l'on a sous-estimé de l'autre les usages dont la théorie minimale de la vérité peut rendre compte. Lorsque ces erreurs ont été rectifiées, la thèse suivante apparaît alors plausible :

Proposition 2 (Adéquation* de la théorie minimale de la vérité)**.** *La théorie minimale de la vérité permet de rendre compte des usages de la notion de vérité jugés essentiels par le déflationniste.*

La théorie tarskienne

Pourtant, je veux maintenant montrer pourquoi la théorie tarskienne est néanmoins un objet théorique privilégié, et pourquoi, en quel sens plutôt, elle est bien « la bonne » théorie de la vérité. Nous avons essayé de montrer que le Principe de Réflexion sur A et l'énoncé de la cohérence de A sont des conséquence réflexives de A. Par conséquent, le fait que ces énoncés soient prouvables dans l'extension tarskienne de A ne montre pas que les principes aléthiques tarskiens soient utilisés comme de nouveaux principes explicatifs, ou qu'ils recèleraient quelques aperçus fondamentaux sur la nature de la vérité. À cette conclusion négative, je voudrais ajouter des considérations positives concernant le statut de l'extension tarskienne elle-même. Plus précisément, je voudrais formuler une thèse *a priori* plus forte que celle défendue dans la section précedente, à savoir :

Proposition 3. *$T(PA)$ est elle-même une conséquence réflexive de PA.*[55,56]

Rappelons la composition de la théorie tarskienne de la vérité dans le cas où le langage contient un terme pour tous les individus de l'univers du discours, comme c'est le cas avec PA :

1. Les instances du schéma-T pour les formules atomiques :

$$Vr(\ulcorner\phi\urcorner) \leftrightarrow \phi, \text{ pour } \phi \text{ atomique}$$

2. des clauses inductives quantifiant sur les énoncés :
   - $\forall\phi \quad Vr(\ulcorner\neg\phi\urcorner) \leftrightarrow \neg Vr(\ulcorner\phi\urcorner)$
   - $\forall\phi \quad Vr(\ulcorner\phi \wedge \psi\urcorner) \leftrightarrow Vr(\ulcorner\phi\urcorner) \wedge Vr(\ulcorner\psi\urcorner)$
   - $\forall\phi \quad Vr(\ulcorner\forall x\phi\urcorner) \leftrightarrow$ Pour tout terme clos $c, Vr(\ulcorner\phi(c)\urcorner)$

Nous voulons montrer que les clauses de « compositionalité »[57] apparaissent naturellement comme des *conséquences* de ce que nous acceptons quand nous endossons nos règles d'inférences, après réflexion sur leur forme générale, à la façon de ce qui a été indiqué dans la section précédente.[58] Considérons par exemple les règles usuelles d'introduction et d'élimination de la conjonction :

$$\frac{A \qquad B}{A \wedge B}\ (\wedge\text{-Int}) \qquad \frac{A \wedge B}{A}\ (\wedge\text{-Elim}) \qquad \frac{A \wedge B}{\text{B}}\ (\wedge\text{-Elim})$$

Endosser ces règles après les avoir formulées en toute généralité, c'est affirmer :

1. $\forall x, y\ ((Vr(x) \wedge Vr(y)) \rightarrow Vr(\ulcorner x \wedge y\urcorner))$
2. $\forall x, y\ (Vr(\ulcorner x \wedge y\urcorner) \rightarrow Vr(x))$
3. $\forall x, y\ (Vr(\ulcorner x \wedge y\urcorner) \rightarrow Vr(y))$

À partir de (2) et (3), on peut inférer :

[55]Cette thèse est plus forte que la précédente, parce que dans $M(PA)+$ « Tous les théorèmes de PA sont vrais » on ne peut pas dériver les axiomes de $T(PA)$, autrement dit $T(PA)$ est logiquement plus forte que $M(PA)+Ref_{PA}$. L'idée est la suivante : une preuve dans $M(PA)+Ref$ mobilise au plus un nombre fini d'équivalences-T. Or pour un ensemble d'énoncés dont la complexité est bornée, le prédicat de vérité est définissable par une formule de PA. Donc une preuve dans $M(PA)+Ref$ peut être transformée en une preuve dans $PA + \{Th_{PA}\ulcorner\phi(x)\urcorner \rightarrow \phi(x)\}$. Mais ce dernier système, appelé $PA+$ Réflexion Uniforme, est connu pour être moins moins fort que $T(PA)$ (Je remercie Volker Halbach de cette remarque.)

[56] Je me limite pour la formulation de cette thèse au cas où la théorie de base est PA. Le cas général demanderait une étude plus approfondie, voir les quelques remarques ci-dessous.

[57]C'est-à-dire toutes les clauses tarskiennes à l'exception de la première.

[58]Par conséquent ce que nous voulons montrer est que les équivalences-T pour les seuls énoncés atomiques sont suffisantes, en présence des ces énoncés explicitant ce que nous acceptons en acceptant nos règles d'inférences, pour dériver toutes les clauses de la définition.

$$\forall x, y\ (Vr(\ulcorner x \wedge y \urcorner) \rightarrow (Vr(x) \wedge Vr(y)),$$

et de là, avec (1),

$$\forall x, y\ ((Vr(x) \wedge Vr(y)) \leftrightarrow Vr(\ulcorner x \wedge y \urcorner)),$$

c'est-à-dire la clause de compositionnalité du prédicat de vérité et de la conjonction des axiomes récursifs tarskiens. Pour faciliter la lecture on peut représenter le raisonnement que l'on vient de faire en déduction naturelle.[59] La première étape consiste à reformuler ce que l'on endosse en acceptant les règles d'inférence de la conjonction en utilisant un prédicat de vérité. Les règles pour la conjonction deviennent les « règles explicites » suivantes :

$$\frac{Vr(A) \qquad Vr(B)}{Vr(A \wedge B)}\ (\wedge\text{-Int})$$

$$\frac{Vr(A \wedge B)}{Vr(A)}\ (\wedge\text{-Elim}) \qquad \frac{Vr(A \wedge B)}{\text{Vr(B)}}\ (\wedge\text{-Elim})$$

où, les variables « A » et « B » sont libres. À partir de là nous avons dérivé la clause tarskienne usuelle pour la conjonction de la façon suivante :

$$\dfrac{\dfrac{\dfrac{[Vr(A \wedge B)]^1}{Vr(A)} \qquad \dfrac{[Vr(A \wedge B)]^1}{Vr(B)}}{Vr(A) \wedge Vr(B)}}{Vr(A \wedge B) \rightarrow Vr(A) \wedge Vr(B)}\ (1)$$

et

$$\dfrac{\dfrac{\dfrac{[Vr(A) \wedge Vr(B)]^1}{Vr(A)} \qquad \dfrac{[Vr(A) \wedge Vr(B)]^1}{Vr(B)}}{Vr(A \wedge B)}}{(Vr(A) \wedge Vr(B)) \rightarrow Vr(A \wedge B)}\ (1)$$

Nous traitons maintenant le cas de l'implication matérielle dont nous reformulons les règles ordinaires d'introduction et d'élimination avec un prédicat de vérité de la façon attendue :

[59] Pour faciliter la lecture, j'omets systématiquement les guillemets « ⌜ » et « ⌝ ». J'écris par exemple $Vr(A \wedge B)$ au lieu de $Vr(\ulcorner A \wedge B \urcorner)$. Les symboles de connecteurs apparaissant dans la portée d'un prédicat de vérité doivent donc aussi être compris comme désignant les fonctions syntaxiques correspondantes.

$$\dfrac{\dfrac{}{Vr(A)}\,(1)}{\dfrac{\vdots}{\dfrac{Vr(B)}{Vr(A \to B)}\,(1)}} \qquad \dfrac{Vr(A \to B) \qquad Vr(A)}{Vr(B)}$$

A partir de ces règles, à nouveau, nous pouvons dériver les clauses de compositionnalité correspondantes :

$$\dfrac{\dfrac{Vr(A \to B) \qquad [Vr(A)]^1}{Vr(B)}}{Vr(A) \to Vr(B)}\,(1)$$

et

$$\dfrac{\dfrac{Vr(A) \to Vr(B) \qquad [Vr(A)]^1}{Vr(B)}}{Vr(A \to B)}\,(1)$$

Pour montrer que la loi composition du prédicat de vérité avec la négation est implicite dans l'acceptation des lois de la logique, je supposerai ici que les deux lois suivantes sont des explicitations de lois de la logique[60] :

$$Vr(A) \vee Vr(\neg A)$$

Autrement dit, « pour tout énoncé x, x est vrai ou $\neg * x$ est vrai ». Et

$$\neg(Vr(A) \wedge Vr(\neg A)),$$

autrement dit, « pour tout énoncé x, ce n'est pas le cas que x et $\neg * x$ sont tous les deux vrais », que l'on peut reformuler de façon classiquement équivalente de la façon suivante :

$$\neg Vr(A) \vee \neg Vr(\neg A).$$

Moyennant quoi les deux dérivations suivantes donnent ce qu'on en attend :

$$\dfrac{\dfrac{Vr(A) \vee Vr(\neg A) \qquad [\neg Vr(A)]^1}{Vr(\neg A)}}{\neg Vr(A) \to Vr(\neg A)}\,(1) \qquad \dfrac{\dfrac{\neg Vr(A) \vee \neg Vr(\neg A) \qquad [Vr(\neg A)]^1}{\neg Vr(A)}}{Vr(\neg A) \to \neg Vr(A)}\,(1)$$

[60] Ceci simplement pour simplifier l'exposé. Voir la note suivante.

où dans chacune des deux dérivations, la première inférence est une instance de raisonnement par cas.[61] La question du quantificateur universel est plus délicate. Pour la règle d'élimination du quantificateur universel[62], les choses se passent de la façon attendue. La règle est la suivante :

$$\frac{\forall x A}{A[x/t]}$$

[61]Cette présentation des choses suppose que les règles logiques sont les règles classiques, sans quoi les prémisses ne peuvent pas être vues comme l'« explicitation de lois logiques », et ne peuvent être dérivées de nos lois « explicites ». Néanmoins, il n'est pas nécessaire d'admettre les lois de la logique classique pour prouver la « compositionnalité de Vr avec la négation ». Pour être plus général, considérons « $\neg\phi$ » comme une abréviation pour « $\phi \to \bot$ » ; ce qu'il faut alors montrer est la chose suivante :

$$Vr(\ulcorner \phi \to \bot \urcorner) \leftrightarrow (Vr(\ulcorner \phi \urcorner) \to \bot)$$

Les deux dérivations suivantes, utilisant les règles explicites pour « $\to$ » mentionnées plus haut, et les règles standard, montrent comment faire :

$$\dfrac{\dfrac{[Vr(\ulcorner \phi \to \bot \urcorner)]^2 \qquad [Vr(\ulcorner \phi \urcorner)]^1}{\dfrac{\dfrac{Vr(\ulcorner \bot \urcorner)}{\bot}}{Vr(\ulcorner \phi \urcorner) \to \bot}\,(1)}}{Vr(\ulcorner \phi \to \bot \urcorner) \to (Vr(\ulcorner \phi \urcorner) \to \bot))}\,(2) \qquad \dfrac{\dfrac{[Vr(\ulcorner \phi \urcorner) \to \bot]^2 \qquad [Vr(\ulcorner \phi \urcorner)]^1}{\dfrac{\dfrac{\bot}{Vr(\ulcorner \bot \urcorner)}}{Vr(\ulcorner \phi \to \bot \urcorner)}\,(1)}}{(Vr(\ulcorner \phi \urcorner) \to \bot) \to Vr(\ulcorner \phi \to \bot \urcorner)}\,(2)$$

En se plaçant dans un calcul dans un calcul multi-conclusions, mieux adapté au traitement inférentiel de la logique classique, on obtient une dérivation plus directe. Dans le calcul des séquents multi-conclusions de Gentzen, les règles pour la négation sont les suivantes :

$$\vdash \neg \frac{\Gamma, A \vdash \Delta}{\Gamma \vdash \neg A, \Delta} \qquad \neg \vdash \frac{\Gamma \vdash A, \Delta}{\Gamma, \neg A \vdash \Delta}$$

Leur explicitation (partielle) avec prédicat de vérité est donc :

$$\vdash \neg \frac{\Gamma, \mathrm{Vr(A)} \vdash \Delta}{\Gamma \vdash \mathrm{Vr}(\neg\ \mathrm{A}), \Delta} \qquad \neg \vdash \frac{\Gamma \vdash \mathrm{Vr(A)}, \Delta}{\Gamma, \mathrm{Vr}(\neg\ \mathrm{A}) \vdash \Delta}$$

Il est alors facile de dériver les règles de « compositionalité » suivantes :

$$\frac{\Gamma \vdash \mathrm{Vr}(\neg \mathrm{A}), \Delta}{\Gamma \vdash \neg \mathrm{Vr(A)}, \Delta} \qquad \frac{\Gamma \vdash \neg \mathrm{Vr(A)}, \Delta}{\Gamma \vdash \mathrm{Vr}(\neg \mathrm{A}), \Delta}$$

Nous le faisons pour la première :

$$\dfrac{\dfrac{\dfrac{Vr(A) \ \vdash \ Vr(A)}{Vr(A), Vr(\neg A) \ \vdash}}{Vr(\neg A) \ \vdash \ \neg Vr(A)} \qquad \Gamma \vdash Vr(\neg A), \Delta}{\Gamma \vdash \neg Vr(A), \Delta}$$

[62]Ou, de façon analogue pour la règle d'introduction du quantificateur existentiel.

où « A[x/t] » est la formule A dans laquelle on a substitué à une ou plusieurs occurrences libres de x le terme clos t. Reformulée de façon explicite avec le prédicat de vérité on obtient :

$$\frac{Vr(\ulcorner \forall x A \urcorner)}{\text{pour tout terme clos } t \quad Vr(\ulcorner A[x/t] \urcorner)}$$

où « $A[x/t]$ » est l'énoncé obtenu en substituant t à x, et donc :

$$(1)\ \frac{\dfrac{[Vr(\ulcorner \forall x A \urcorner)]^1}{\text{pour tout terme clos } t \quad Vr(\ulcorner A[x/t] \urcorner)}}{Vr(\ulcorner \forall x A \urcorner) \rightarrow \text{pour tout terme clos } t\ Vr(\ulcorner A[x/t] \urcorner)}$$

soit la dernière clause de la définition axiomatique tarskienne, lue de droite à gauche. Mais les choses sont moins claires lorsque l'on considère la règle d'introduction du quantificateur universel, à savoir :

$$\frac{A(t)}{\forall x A(x)}$$

où le terme t est « libre » au sens où rien n'est supposé à son propos dans les hypothèses de la dérivation, qu'il est donc arbitraire. Le point délicat est de savoir comment rendre compte de ce caractère « arbitraire » du choix de t. Il est naturel d'expliciter ce caractère arbitraire du terme t sous la forme d'une généralité, et reformulant la règle comme précédemment on obtient alors :

$$\frac{\text{pour tout terme clos } t \quad Vr(\ulcorner A[x/t] \urcorner)}{Vr(\forall x A(x))}$$

et donc au bout du compte exactement l'autre direction de la clause pour le quantificateur universel dans la définition axiomatique tarskienne de la vérité (dans le cas où le langage possède un terme pour chaque individu) :

$$(1)\ \frac{\dfrac{[\text{pour tout terme clos } t \quad Vr(\ulcorner A[x/t] \urcorner)]^1}{Vr(\ulcorner \forall x A(x) \urcorner)}}{\text{pour tout terme clos } t \quad Vr(\ulcorner A[x/t] \urcorner)] \rightarrow Vr(\forall x A(x))}$$

On constate que la clause tarskienne suit de la réflexion et de l'« affirmation de la règle » en question. Mais la règle en question est-elle correcte ? Seulement sous l'hypothèse tacite que pour tout individu x, il existe un terme clos y, tel que y dénote x, hypothèse qui était la nôtre pour nous autoriser cette définition axiomatique de la vérité sans en passer par la notion de satisfaction. Pour aller plus loin, il faudrait donc à présent reprendre toute notre démonstration dans un

cadre plus général. Considérer les clauses compositonnelles gouvernant la notion de satisfaction d'une formule et de dénotation d'un terme, se donner d'autre part des analogues de la théorie minimale de la vérité pour les relations de satisfaction et de dénotation, et enfin essayer de voir si l'on peut dériver les axiomes de la théorie compositionnelle de la satisfaction à partir d'une explicitation des règles logiques à l'aide de prédicats « déflationnistes » de satisfaction et de dénotation. Cette entreprise exigerait donc une discussion de la conception déflationniste de la dénotation, par exemple, et d'autres sujets que nous avons préféré laisser de côté ici pour contenir notre discussion dans des limites de complexité et d'étendue raisonnables. Je m'en tiendrai donc ici à ces résultats partiels concernant le prédicat de vérité et le cas du langage de l'arithmétique.

Dans ce cas au moins (et l'on peut que faire des conjectures à propos du cas général), donc, nous avons établi la thèse 2 : nous sommes parvenu à la conclusion que non seulement le principe de réflexion sur PA est une conséquence réflexive de PA, mais les clauses tarskiennes elles-mêmes n'introduisent « rien de neuf », en ce sens qu'elles doivent être justifiées par toute justification de nos pratiques inférentielles. L'extension aléthique tarskienne d'une théorie A apparaît elle-même ne faire qu'expliciter certains de nos engagements implicites dans notre acceptation de la théorie A, à savoir notre acceptation des règles d'inférence de A.

Mais inversement, si l'on peut dériver les clauses tarskiennes à partir des lois qui explicitent les engagements que nous prenons en acceptant les règles d'inférence, la réciproque est également vraie : à partir des clauses tarskiennes on peut dériver la validité des règles d'inférence, c'est ce qu'avait montré Tarski. Par conséquent, on peut voir les clauses tarskiennes elles-même comme le moyen exact d'expliciter ce qui est implicitement accepté par un sujet qui suit les règles d'inférences logiques (gouvernant les constantes logiques habituelles). Autrement dit :

Proposition 4. *Les lois récursives de la vérité sont équivalentes aux lois énonçant la validité des règles logiques.*

On pourrait noter que cette équivalence est « robuste » au sens où elle reste vraie que la logique sous-jacente soit la logique classique ou intuitionniste.[63]

[63]En particulier, on ne peut pas dériver des clauses tarskiennes l'énoncé général du tiers-exclu en utilisant seulement des règles d'inférence intuitionnistes. Autrement dit les clauses tarskiennes par elles-mêmes sont neutres relativement à une supposée « nature » de la vérité. En revanche les équivalences-T restent dérivables des clauses récursives en utilisant seulement les règles intuitionnistes. Ces deux points ont déjà été remarqués dans la littérature, voir par exmeple Putnam

Les remarques précédentes tendent à montrer que $T(A)$, l'extension aléthique tarskienne de A, est une conséquence réflexive de A. Mais l'on pourrait aller encore un peu plus loin et se demander si $T(A)$ épuise l'ensemble des conséquences réflexives de A : y a-t-il d'autres conséquences réflexives de A que les conséquences déductives de $T(A)$? Si par « accepter une théorie », on n'entend rien de plus (mais rien de moins non plus) qu'accepter ses axiomes et accepter implicitement ses règles, il semble que non.[64] Ceci suggère donc la conjecture suivante, où l'ensemble des conséquences réflexives de la théorie engendrée par les axiomes A $(= (Ax_A))$ et des règles de la logique classique FOL est noté $Cn_{Ref}((Ax_A, FOL))$:

Proposition 5. $Cn_{Ref}((Ax_A, FOL)) = T(A)$
Autrement dit : l'ensemble des conséquences réflexives d'une théorie classique A identifiée par l'ensemble de ses axiomes et des règles d'inférences classiques est identique à l'ensemble des conséquences déductives de $T(A)$, son extension aléthique tarskienne.[65]

Je reviens rapidement en conclusion sur le tableau général esquissé dans cette section. Au chapitre 3, nous avions présenté la Thèse de la Réflexion, une thèse invoquée par Tarski, Shapiro et d'autres, pour justifier la préférence donnée à la théorie tarskienne contre la théorie minimale de la vérité. En substance la thèse était :

Thèse de la Réflexion Une *bonne* theorie de la vérité *doit* permettre de prouver le principe de réflexion

Nous sommes maintenant en situation de comprendre ce qui pose problème dans cette thèse : elle est fondamentalement équivoque. « Bonne », la théorie tarkienne l'est certainement, mais quelle fonction remplit-elle ? Plutôt que de reprendre la thèse de la réflexion à notre compte, nous avons cherché à préciser un peu les choses : si une théorie de la vérité doit simplement articuler notre compréhension de la

(1978*b*) et Sundholm (2004). Voir aussi Tennant (1987) pour un exposé plus détaillé.

[64] Quand les axiomes sont en nombre infini, nous avons vu qu'ils ne peuvent eux-mêmes être acceptés qu'indirectement. Tout ce qui est en fait accepté explicitement par le sujet, ajouté à l'explicitation de ce qui est implicitement accepté par lui, permet de dériver le principe de réflexion sur les axiomes, lequel est donc une conséquence réflexive de A (puisque dans le cas fini, il est aussi dérivable dans une extension conservative des axiomes de A). Mais ce principe de réflexion sur les axiomes est bien une conséquence de l'extension aléthique tarskienne de la théorie.

[65] Avec la terminologie qui nous avons introduite dans la section précédente, nous avons donc : Soit α un agent acceptant la théorie A et suivant les règles de la logique classique. *Affirmer* ϕ est une contradiction pragmatique si et seulement si $T(A) \vdash \neg\phi$.

notion de vérité comme correspondance et rendre compte de nos usages inférentiels du terme « vrai », il semble que la théorie minimale dise tout ce qu'il est nécessaire de dire, puisqu'elle permet au prédicat de vérité de jouer son rôle expressif dans les contextes de montée sémantique. Que cette théorie, qui rend compte de la fonction du prédicat de vérité, ne permette pas de prouver quoi que ce soit de nouveau, est à mettre à son crédit. Mais si par « bonne théorie de la vérité » nous entendons une théorie qui explicite ce qui est implicitement accepté par un agent acceptant une théorie donnée, alors en effet la théorie minimale n'est pas une « bonne » théorie de la vérité, et la théorie tarskienne lui est supérieure sous ce rapport. Si inversement, par la recherche d'une « bonne » théorie de la vérité, nous voulons dire la recherche d'axiomes contenant le concept de vérité et tels que tous les théorèmes qui sont le produit de leur adjonction à une théorie de base doivent être admis comme justifiés sur la base du fait que la théorie de base est justifiée - c'est-à-dire si nous cherchons le meilleur moyen de déployer les justifications nécessaires pour notre théorie de base - alors certainement la simple théorie minimale n'est pas une bonne théorie de la vérité ! Chercher une telle théorie est un but parfaitement légitime. Mais encore une fois, la force logique des théories que nous trouverons sur ce chemin ne montre nullement que nous utilisons la vérité comme un concept explicatif « substantiel ».

6.4 Prolongements : justifications itérées et théories informelles

Nous avons soutenu que pour justifier notre acceptation de la *théorie* PA, il nous faut justifier le principe de réflexion sur PA, c'est-à-dire donner des raisons d'accepter les axiomes de la théorie $PA + Ref_{PA}$.[66] La théorie $PA + Ref_A$ déploie explicitement dans ses conséquences des engagements qui étaient implicites dans l'acceptation de PA. Mais le raisonnement peut être itéré : nous ne pouvons avoir de justification directe de toutes les conséquences de $PA + Ref_A$, et toute justification de notre acceptation de la *théorie* $PA + Ref_{PA}$ doit être une justification de $PA +$

[66] Si ce que nous avons dit dans la section précédente est correct, ce que nous disons est vrai également de $T(PA)$. Les considérations de cette section sur les extensions de PA par des principes de réflexion valent, *mutatis mutandis*, pour les extensions aléthiques tarskiennes de PA. Le choix de l'un des types d'extensions plutôt que de l'autre est sans importance pour ce qui nous occupe ici, en particulier rien de ce que nous disons ne dépend de la question de savoir si $T(PA)$ déploie toutes les conséquences réflexives de PA.

$Ref_{PA+Ref_{PA}}$, où $Ref_{PA+Ref_{PA}}$ est l'énoncé affirmant que tous les théorèmes de $PA + Ref_{PA}$ sont vrais. Et ainsi de suite.

Cette progression appelle plusieurs remarques. La première est une remarque de prudence, motivée par le fait que nous sommes en train de transgresser un interdit que nous avions formulé en introduction à ce travail relativement aux contextes d'emploi du terme « vrai ». La dernière occurrence du terme « vrai », dans le paragraphe précédent, prend en effet pour arguments des énoncés contenant eux-mêmes le terme « vrai ». Or, pour que le prédicat de vérité fasse son office d'outil expressif, nous avons vu qu'il doit être gouverné par toutes les instances du schéma-T obtenues par substitution aux lettres schématiques d'énoncés du langage duquel nous voulons affirmer une infinité d'énoncés. Quand ce langage contient le prédicat de vérité lui-même, nous sommes face à une difficulté bien connue : l'ensemble des instances du schéma-T dans un langage contenant son prédicat de vérité et les règles d'inférences logique classiques est contradictoire.[67]

Pour contourner le problème, de nombreuses solutions ont été proposées dans la littérature. La recherche contemporaine sur les paradoxes tend à s'orienter vers des logiques non-classiques, tout en maintenant la demande d'intersubstituabilité d'un énoncé p et de « $Vr(p)$ » en tout contexte extensionnel pour conserver le pouvoir d'expression de la généralité à l'aide du prédicat de vérité.[68] Discuter ces solutions, et leur compatibilité avec notre thèse sur le rôle du prédicat de vérité nous mènerait au-delà des limites que nous nous sommes fixés dans ce travail. Nous nous contenterons ici d'adopter la solution radicale de Tarski lui-même en supposant une hiérarchie de prédicat de vérité, chacun ne s'appliquant qu'à des énoncés ne contenant pas de prédicat de vérité ou des prédicats de vérité de rang strictement inférieur à lui-même. Plutôt qu'une unique théorie minimale de la vérité, nous faisons donc provision d'une infinité de telles théories, une pour chaque prédicat de la hiérarchie, que nous pouvons choisir d'utiliser quand nous en avons besoin pour exprimer sans risque une infinité d'énoncés. Ceci étant précisé, il nous faut reformuler en conséquence notre affirmation de début de paragraphe : toute justification de $PA + Ref^1_{PA}$, où Ref^1_{PA} est l'énoncé « *tous les théorèmes de PA sont vrais*$_1$ » doit être une justification de $PA + Ref^2_{PA+Ref^1_{PA}}$, où $Ref^2_{PA+Ref^1_{PA}}$ est l'énoncé « *tous*

[67] Voir chapitre 2.

[68] Voir par exemple en particulier les recherches de Hartry Field sur les paradoxes, explicitement motivées par la volonté de préserver le pouvoir expressif de la vérité. Field (2002), Field (2003), et plus récemment Field (2008).

les théorèmes de $PA + Ref_{PA}$ *sont vrai*$_2$ ». Cette théorie est à son tour logiquement plus forte que $PA + Ref^1_A$. Et ainsi de suite.

Il semble que le déploiement des jugements qu'il doit être permis de tenir pour corrects si nous sommes justifiés à accepter une théorie, ne soit pas moins infini que le déploiement de ses conséquences à partir de ses axiomes. Les extensions successives de la théorie de base par les principes réflexifs donnent lieu à une hiérarchie de théories de plus en plus fortes. Mais d'un point de vue logique, la situation, pour être complexe, n'en est pas moins connue. Des suites de théories d'un type proche de celles que nous venons d'esquisser ont été étudiées sous le nom de *progressions récursives de théories* par Feferman en 1962, à la suite des travaux de Turing sur les extensions par les principes de cohérence, un programme de recherche antérieur, mais lié, à celui que nous avons présenté au chapitre 4. Une présentation détaillée des travaux logiques de Ferferman sur les progressions de théories dépasse de beaucoup le cadre de ce travail. On peut considérer notre travail comme une contribution modeste au programme de Feferman. Tandis que le travail de Feferman *commence* avec le constat que, si nous sommes justifiés à accepter une théorie nous le sommes à en accepter diverses extensions, en particulier son extension par des affirmations de cohérence ou des principes de réflexions, notre travail est un travail philosophique qui cherche à *expliquer* le constat qui est au commencement de ce programme : comment est-il possible qu'étant justifiés à accepter une théorie donnée, nous puissions reconnaître que nous le sommes *ipso facto* à accepter une théorie logiquement plus forte ? L'explication que nous en proposons a le mérite de la simplicité : nous la fondons seulement sur une analyse déflationniste du rôle du concept de vérité, une réflexion sur nos limitations épistémiques (en particulier notre capacité à juger) et des hypothèses minimales sur la notion de justification. Nous n'affirmons ni que l'énoncé de la cohérence de A est conséquence du contenu de la théorie A (même de façon cachée ou informelle etc.), ni que son acceptation serait simplement une conséquence rationnelle de l'acceptation de A. Nous affirmons que l'énoncé de la cohérence de A est une conséquence *déductive* du contenu que nous acceptons *en fait* si nous acceptons A explicitement. Nous avons également proposé une analyse du mouvement réflexif par lequel un sujet est amené à accepter ce dont il a reconnu qu'il accepte implicitement par dans sa pratique déductive, en montrant comment ce processus faisait émerger des contenus propositionnels porteurs d'une généralité resté à l'état latent » dans les dispositions du sujet à suivre certaines règles *ad libitum* mais absente des contenus acceptés au départ.

Avant de conclure sur ce point, je voudrais souligner deux choses. D'une part, si les extensions réflexives d'une théorie sont plus fortes que cette théorie elle-même, l'itération indéfinie de l'opération de réflexion, ou d'explicitation, pour peu que l'on mette en œuvre de façon raisonnable le détail de ce processus d'itération, *ne permet pas de justifier n'importe quel énoncé vrai du langage de la théorie de départ*. En particulier, il semble que, dans le cas de la théorie des ensembles, aucune hypothèse ensembliste intéressante ne soit décidée dans les extensions aléthiques tarskiennes itérées de ZF.[69] Dans le cas de l'arithmétique, la question est plus compliquée et les détails de la définition du processus d'itération prennent de l'importance : parle-t-on d'itération finie uniquement ? Ou bien voulons-nous considérer un processus d'itération transfini ? Dans ce dernier cas, sur quels ordinaux le processus est-il défini ? Les ordinaux dénombrables, les ordinaux qui sont les types d'ordre de bons ordres définissables dans une extension du second-ordre conservative sur la théorie sur laquelle opère la réflexion ? Ou même qui seraient prouvablement tels ? Et comment définit-on la théorie obtenue dans le processus d'itération aux ordinaux limites ? Ces questions sont difficiles, tant du point de vue logique, que du point de vue de l'évaluation de la correction méthodologique des réponses que l'on a pu y apporter.[70] Malgré ces difficultés, le point que nous faisions reste correct : en

[69] Aatu Koskensilta, tandis que je demandais s'il existait des hypothèses ensemblistes intéressantes décidées dans des extensions aléthiques de ZF ou ZFC, m'a fait par courrier les remarques suivantes, qui permettent de clarifier la situation d'un point de vue logique :

> Nous pouvons noter que toute extension de ZFC avec un prédicat de vérité itéré (et nous pouvons être généreux, autorisant l'itération sur des bons-ordres qui sont des classes propres et deviennent définissables dans le cours du processus d'itération du prédicat de vérité) est naturellement interprétable dans la théorie des ensembles de Morse-Kelley. Je ne connais aucune « hypothèse ensembliste intéressante » qui est indécidable dans ZFC mais ne l'est pas dans MK.

On trouvera l'échange original complet, en date des 12 et 15 juin 2007, dans les archives du forum Foundation of Mathematics, à l'adresse suivante : *http ://cs.nyu.edu/pipermail/fom/*

[70] Pour une discussion méthodologique du fameux théorème de complétude de l'arithmétique obtenu par Feferman par les méthodes de réflexion, nous renvoyons à nouveau le lecteur à l'article Franzén (2004*b*). Feferman (1962) utilise des résultats de Feferman (1960) où il est montré que, si les axiomes d'une théorie donnée sont décrits par un certain type de formules « bien » choisies, il est possible de donner dans la théorie en question une preuve de l'énoncé affirmant qu'aucune contradiction ne peut être dérivée de la théorie dont les axiomes sont ceux énumérés par la formule en question. Autrement dit, il est possible de prouver dans une théorie comme PA certains énoncés dont on peut dire, en un sens, qu'ils « expriment » la cohérence de PA. Bien entendu, on ne peut pas prouver dans PA que cet énoncé est équivalent à l'énoncé standard de la cohérence de PA, ce qui contredirait le second théorème d'incomplétude. Les conditions sous lesquelles on est prêt à reconnaître qu'une formule énumérant les axiomes d'une théorie est « naturelle » est donc de la première importance, et se pose de façon d'autant plus aiguë lorsque nous n'avons pas à notre

général, l'itération du processus de réflexion sur une théorie ne permet de décider qu'un certain genre d'énoncés du langage de la théorie de départ. La progression de la force logique des théories dans la hiérarchie des justifications est indéfinie[71], mais elle doit néanmoins être *bornée*.

D'autre part, c'est ma seconde remarque, il faut être attentif au fait que le caractère strictement croissant de la force logique des extensions réflexives successives est relatif à une certaine mesure de la « force logique » d'une théorie. Nous avons dit : parce que nos capacités d'expression et de pensée sont finies, nos justifications doivent se révéler être strictement plus fortes, logiquement, que ce que nous voulons justifier. Or, que signifie « logiquement strictement plus fort » ici ? La théorie à justifier, A, et la théorie J que doit justifier toute théorie qui justifie la théorie A, ne sont en général pas formulées dans le même langage. Comme nous l'avons vu, nous avons recours dans J au prédicat de vérité, et à un appareil descriptif permettant de parler des énoncés et des règles de A dont nous cherchons à justifier le contenu, appareil dont nous n'avons pas en général de raison de penser qu'il est disponible dans la théorie à justifier elle-même. Pour pouvoir comparer la force *logique* de A à celle de J, il est donc naturel de se tourner[72] vers la notion de *conservativité* : J est logiquement plus forte que A ssi J n'est pas conservative sur A. Au chapitre 3, nous avons défini la conservativité d'une théorie A sur une théorie B de la façon suivante : A est conservative sur B si et seulement si tout énoncé du langage de B prouvable dans A est prouvable dans B. Mais il existe une définition plus souple de la notion de conservativité, qui permet de moduler la notion de « plus grande force logique ».

Définition 15. *Soient A et B deux théories. A est conservative* sur B si et seulement si, tout énoncé du langage de B qui est* conséquence logique *de A est* conséquence logique *de B.*

Dans cette définition, « conséquence logique » réfère à ce que l'on appelle parfois la notion *sémantique* de conséquence ($\models$), expliquée en termes modèle-théoriques. Les deux notions de conservativité et de conservativité* coïncident lorsque c'est à la logique du premier ordre que nous pensons. Mais pour ces logiques dont la relation

disposition de présentation axiomatique priviligiée d'une théorie donnée, comme c'est le cas pour la ω-ième itération de l'extension aléthique d'une théorie. Pour une discussion de la signification épistémologique des preuves de cohérence à la Feferman, on pourra consulter Resnik (1989).

[71] Ne serait-ce qu'à s'en tenir aux itérations finies.

[72] À nouveau, mais cette fois de façon appropriée.

de conséquence n'est pas récursivement énumérable, ou qui ne sont pas complètes, les deux notions ont des extensions différentes. Si maintenant nous comparons la force logique de différentes théories à l'aide de la notion de conservativité*, il n'est plus vrai que la théorie réflexive $PA + Ref_A$ d'une théorie A est nécessairement plus forte logiquement que A en ce sens. Pour s'en convaincre, il nous suffit de prendre pour exemple PA^2, l'arithmétique de Peano en second ordre, et L^- un calcul (forcément incomplet) pour la logique du second ordre. Le principe de réflexion sur PA^2 (« Toutes les L^--conséquences déductives de PA^2 sont vraies ») étend non-conservativement PA^2, mais est une extension *conservative** de PA^2.[73] La situation est donc la suivante. Parce que nos capacités d'expression et de pensée sont limitées, nous ne pouvons juger de ce qu'une théorie n'est justifiée que par la médiation d'un jugement dont le contenu est strictement plus fort, au sens de la logique du premier ordre, que cette théorie. Si la théorie A est formulée dans une logique dont le « contenu calculatoire » et le « contenu descriptif » sont distincts (i.e. quand la relation de conséquence au sens sémantique n'est pas récursivement énumérable), néanmoins, la théorie $PA + Ref_A$, quoique possédant un contenu calculatoire strictement plus fort que celui de A (au sens où elle n'est pas conservative sur A), peut posséder exactement le même contenu descriptif (au sens où elle serait conservative au sens sémantique sur A).

Pour conclure, je voudrais revenir sur la notion de *théorie*, et plus particulièrement sur les conditions d'individuation des théories. Comme nous avons déjà eu l'occasion de le noter, au sens qui est celui ordinairement adopté en logique une théorie est un ensemble déductivement clos d'énoncés, où « déductivement clos » doit s'entendre comme « clos pour les règles de déduction formelles ». On peut aussi vouloir distinguer deux théories extensionnellement identiques mais dont « les axiomes » ou « les règles d'inférences », dont le rôle est si important dans la façon dont les théories au sens précédent nous sont données, sont distincts. Laissant de côté les règles d'inférences, que je suppose être les règles classiques, une théorie peut alors être vue, par exemple, comme un couple (A, Ax), constitué d'un ensemble d'énoncés déductivement clos A et d'un ensemble d'énoncés Ax, tel que la clôture déductive de Ax est identique à A. Sous cette nouvelle condition d'individuation, deux axiomatisations distinctes Ax_1 et Ax_2 d'un même ensemble d'énoncés déductivement clos A engendrent deux théories (A, Ax_1) et (A, Ax_2) équivalentes mais distinctes.

[73] En effet, l'arithmétique de Peano en second ordre est catégorique, donc sémantiquement complète.

Les deux types de conditions d'individuation peuvent avoir leur utilité, c'est une question de buts poursuivis.

Mais il existe une autre distinction, plus difficile à cerner, entre les théories formelles, ou plutôt les théories *formalisées* (et donc contentuelles), et ce que l'on appelle parfois les théories *informelles*. Depuis les théorèmes d'incomplétude de Gödel, l'idée a souvent été exprimée qu'il existerait un reste irréductible dans l'opération de formalisation par laquelle nous passons des théories informelles, qui constitueraient effectivement le contenu de la pensée et de ce qui est accessible à partir de ce contenu par les opérations de la raison, aux théories formalisées comprises comme ensemble de d'énoncés clos par les règles de raisonnement codifiées par la logique du premier ordre, comme si le processus de formalisation entraînait une forme de dégradation au terme de laquelle la raison ne s'y retrouvait plus tout à fait. La tentation de pensée qui apparaît alors est qu'il existe une notion irréductiblement informelle de *conséquence*, pour laquelle l'énoncé de la cohérence d'une théorie, ou l'énoncé de Gödel de cette théorie, est une conséquence informelle *de la théorie*. Une théorie informelle est alors l'ensemble des conséquences informelles, en ce sens vague, d'un ensemble d'énoncés donnés. Cette façon d'envisager les choses est insatisfaisante pour plusieurs raisons. D'abord, elle nous laisse avec un mystère, celui d'expliquer la notion de conséquence informelle. Comment alors pourrait-on ne serait-ce que poser la question de savoir si les règles de consécution informelle sont justifiées, ou seulement nous demander réflexivement si nous les acceptons, affirmer éventuellement que nous les acceptons, si nous ne savons pas même ce qu'elles sont ? En outre, l'idée que l'énoncé de la cohérence d'une théorie serait une conséquence informelle *de cette théorie* et de rien d'autre, n'a de sens que si l'on suppose que le contenu sémantique (vériconditionnel) dont on « infère informellement », l'énoncé de la cohérence d'une théorie, est autre que le contenu sémantique de l'ensemble des énoncés de la théorie pris ensemble, qui n'est autre que celui de ses axiomes. Or, il semble plus raisonnable, d'après le tableau dressé dans ce chapitre, d'affirmer seulement que toute *justification* que nous pouvons avoir pour accepter la théorie permet de construire une justification des énoncés gödéliens, tout en reconnaissant que, du point de vue du contenu propositionnel, les jugements qui fondent une telle justification possèdent un contenu plus riche que celui de la théorie de départ.[74] Or

[74] Puisque dans la théorie de départ (A) on ne peut pas prouver l'ensemble des instances de « $Th_A(\ulcorner\phi\urcorner) \rightarrow \phi$ », et moins encore la généralisation formulée à l'aide du prédicat de vérité (ce que j'ai appelé ici le principe de réflexion sur A, et que l'on trouve dans la littérature sous le nom de

une relation de *conséquence logique*, formelle ou informelle, doit permettre, à partir d'un ensemble d'énoncés, de déployer des engagements liés au *contenu sémantique* de ces énoncés. Je propose donc d'envisager la notion de théorie informelle un peu différemment, en utilisant la notion de conséquence réflexive développée dans ce chapitre, dont l'objet n'est pas de rendre compte des engagements véhiculés par le contenu sémantique des théories, mais des engagements épistémiques contractés par un sujet lorsqu'il accepte une théorie.

Le point de départ est de considérer qu'accepter une certaine théorie, d'un certain point de vue, c'est accepter un ensemble d'*engagements* épistémiques. Pour le dire de façon imagée, la façon ordinaire d'identifier les théories ne prête attention qu'aux engagements « descendants », c'est-à-dire aux propositions qu'un sujet s'engage à accepter en acceptant les axiomes parce que ces propositions sont des conséquences de ces axiomes. Un sujet qui juge justifiés les axiomes n'en a pas pour autant construit effectivement une justification de toutes les conséquences de ces axiomes, mais il a contracté par là un engagement à tenir pour justifiée toute conséquence dont on lui présenterait une preuve à partir des axiomes. Or dans l'autre sens, on peut faire valoir qu'existent également des engagements « ascendants » associés à l'acceptation d'une théorie au sens précédent : en acceptant une proposition et en contractant les engagements descendants envers ses conséquences, on s'engage aussi à tenir pour justifiées les propositions qui explicitent ces engagements. Or nous avons vu que les propositions qui renferment l'expression de ces engagements ne sont pas des conséquences logiques de la théorie de départ. Supposons maintenant que, conformément à ce que nous avons dit plus haut, en acceptant la théorie A et en l'identifiant comme la clôture déductive de certains axiomes par certaines règles, je sois *ipso facto* engagé à accepter le principe de réflexion sur A. En vertu de mes premiers engagements à accepter les conséquences logiques de ce que j'accepte, accepter ce principe m'engage à nouveau, en aval, à en accepter toutes les conséquences ; mais alors ensuite également, en amont, à accepter une nouvelle explicitation des engagements qui en résultent, telle que formulée dans une nouvelle itération du principe de réflexion. Autrement dit, mon engagement à accepter une théorie A m'engage à accepter la théorie $Cn_{Ref}(A)$, et celle-ci à accepter $Cn_{Ref}(Cn_{Ref}(A))$. Notons $Cn_{ref,n}(A)$ la nè itération de l'opération consistant à prendre l'ensemble des conséquences réflexives de la théorie A, en partant de A.

« Principe général de réflexion ». En outre on aura par exemple introduit de nouveaux concepts, les concepts syntaxiques, le concept de vérité, qui étaient absents de la théorie de départ.

Puisque, du point de vue des engagements épistémiques contractés par un sujet acceptant une théorie, l'individuation des théories par leurs conséquences logiques est insatisfaisante, l'option qui paraît à présent naturelle, est d'individuer les théories en tenant compte également des engagements « ascendants ». Soient A et B deux ensembles d'énoncés, la théorie informelle engendrée par A et la théorie informelle engendrée par B (et les règles classiques d'inférence) sont identiques si et seulement s'il existe deux entiers n et n' tels que $Cn_{Ref,n}(A) = Cn_{Ref,n'}(B)$.[75] Sous ce critère, et si ce que nous avons dit précédemment est correct, alors PA et $T(PA)$ sont une seule et même théorie.[76]

Notre projet n'est pas de développer ici une théorie de la justification ou de la notion de théorie informelle, ni de dégager de façon définitive l'extension de ce que nous avons identifié comme l'implicite de l'acceptation d'une théorie. Notre projet était plus modeste, c'était simplement celui d'examiner le rôle du prédicat de vérité dans ce type de processus, et nous avons donné des éléments pour étayer la thèse selon laquelle le prédicat de vérité joue essentiellement un rôle « expressif ». Revenons en effet sur la comparaison entre les deux interprétations de la portée philosophique de la force logique qui est celle de l'extension aléthique tarskienne d'une théorie, disons $T(PA)$. Pour les « substantialistes », si $T(PA)$ permet de démontrer l'énoncé de la cohérence de PA, c'est parce que nous y avons introduit un nouveau concept explicatif, avec les lois qui le gouvernent. $T(PA)$, selon cette lecture, est une extension de PA qui s'apparente à l'extension de notre théorie physique par de nouvelles lois, celles par exemple de la théorie des champs.[77] Les raisons que nous avons d'adopter PA et les raisons spécifiques que nous avons d'adopter $T(PA)$ ne sont pas pensées dans leur unité profonde. L'idée que, d'une part, $T(PA)$ ne fait qu'expliciter les engagements implicites dans l'acception de PA et que, d'autre part, le rôle du prédicat de vérité dans ce processus est seulement celui de permettre l'expression de la généralité, paraît au contraire en accord avec les thèses déflationnistes sur la vérité.[78]

[75] Je laisse de côté les discussions relatives à la question de savoir si n ne peut pas être pris dans les ordinaux, plutôt que dans les seuls entiers. Voir nos remarques sur le travail de Feferman dans la section.

[76] Puisqu'alors $Cn_{Ref,0}(T(PA)) = Cn_{Ref,1}(PA)$, même si $Cn(Ax_{PA}) \neq Cn(Ax_{T(PA)})$. En fait, il est naturel de penser que les résultats de Feferman sur les limites de la prédicativité « étant donné les entiers » sont pertinents ici, et l'on pourrait être tenté, dans le cas particulier de l'arithmétique, d'identifier toutes les théories extensions aléthiques de PA jusqu'à l'ordinal Γ_0, par exemple, comme formant le contenu d'une unique « théorie informelle ».

[77] Voir chapitre 3.

[78] Dans la continuité de notre analyse, il faudrait présenter une interprétation des théories non-

6.5 Conclusion

Notre réponse aux problèmes développés depuis le chapitre 3 procède d'un examen critique des relations épistémiques qu'un sujet peut entretenir avec une théorie. La remarque de départ, simple et contraignante, est que l'impossibilité d'exercer son jugement *ad infinitum* impose certaines contraintes sur les contenus qui peuvent être effectivement les objets de jugements d'un sujet qui accepte une théorie donnée. Une théorie étant toujours constituée d'une infinité d'énoncés, et ces énoncés ne pouvant avoir été chacun l'objet d'un jugement distinct, il faut donc que ce soit en vertu du fait qu'il a effectué un jugement dont le contenu est strictement plus fort que celui des énoncés de la théorie réunis, que le sujet accepte en effet la théorie en question, quelque chose comme un énoncé général dont les « instances » sont les théorèmes de la théorie en question. Ainsi, il est possible de reconnaître qu'un sujet acceptant une théorie A soit en effet en situation d'affirmer que tous les théorèmes de cette théorie sont vrais et que la théorie est cohérente, comme dans le scénario imaginé par Shapiro et présenté au chapitre 3. Mais ceci ne montre pas qu'il y a un usage du prédicat de vérité par lequel un sujet *infère* d'une théorie A la proposition que cette théorie est vraie. C'est pour ainsi dire dans l'autre sens que le sujet fait ses inférences déductives : un sujet juge d'abord qu'une théorie est vraie, faute de pouvoir exercer son jugement sur chacun des énoncés de la théorie.[79] Comme dans le cas des généralisations classiques sur les positions objectuelles, la généralisation sur des expressions en position phrastique est donnée dans la pensée sans que toutes les instances ne le soient jamais : s'il se trouve que quelques instances peuvent être la source psychologique de la formulation de l'énoncé général, c'est l'énoncé général qui est premier dans la l'ordre théorique, sans qu'il y ait de solution de continuité logique entre les premières et les dernières.[80]

Dans le présent chapitre, en nous appuyant sur les remarques déflationnistes

typées de la vérité à la lumière de la méthode utilisée ici, à savoir se demander de quel implicite ces théories de la vérité non-typées sont des expressions, de quel type d'action constituent-elles la justification ou explicitent-elles les engagements théoriques. C'est un travail que je remets à des recherches ultérieures.

[79] Et en général, pour le mathématicien ou le logicien, sa justification est plus détaillée : il a examiné des axiomes et des règles, et a jugé que les premiers étaient vrais et que les secondes préservaient la vérité.

[80] Un lecteur pourrait remarquer ceci : les énoncés logiques généraux vrais (valides), eux, sont prouvables à partir des règles gouvernant les expressions logiques. Par exemple, on peut prouver dans la logique du premier ordre

$$\forall x(Px \vee \neg Px).$$

à propos de l'usage du prédicat de vérité, nous avons soutenu que, pour compter comme acceptant la théorie A, nous devions endosser réflexivement au moins le contenu du principe de réflexion aléthique sur A : « Tous les théorèmes de A sont vrais », où « vrai » n'est autre que le prédicat de vérité minimal.[81] Nous avons ensuite soutenu qu'il n'y avait pas, d'un côté, le prédicat de vérité minimal et, de l'autre, le prédicat de vérité tarskien, dont les théories saisiraient des concepts de vérité distincts : les axiomes de la théorie tarskienne elle-même se comprennent comme des formulations, à l'aide du prédicat de vérité, de ce que nous acceptons lorsque nous acceptons de suivre les règles logiques. Celui qui accepte les axiomes d'une théorie A et, explicitement, toutes les règles de déductions, accepte en fait l'extension aléthique tarskienne de A, $T(A)$, et non seulement $(A + Ref_A)$. D'un côté, la théorie minimale suffit à rendre compte de tous les usages du prédicat de vérité mis en avant par le déflationniste, d'un autre côté, l'extension tarskienne d'une théorie A est un objet théorique privilégié en tant qu'explicitation de ce qu'un sujet accepte lorsqu'il accepte la théorie A.

Au chapitre 3, j'ai repoussé l'argument de la conservativité en notant que les extensions aléthiques tarskiennes n'étendaient non-conservativement que les théories contenant des schémas *ouverts*. J'ai alors soutenu que les phénomènes de non-conservativité observés lorsque les schémas de la théorie de base sont étendus au vocabulaire aléthique étaient *a priori* compatibles avec une interprétation « déflationniste » dans laquelle ces phénomènes illustreraient le « pouvoir *expressif* » du prédicat de vérité. Dans le cas de l'arithmétique de Peano en premier ordre, le principe d'induction est renforcé, avons-nous vu, lorsque de nouveaux moyens expressifs sont à disposition dans le langage. À propos de ce phénomène par lequel une théorie

Si le prédicat de vérité est un prédicat logique, comme j'ai annoncé que j'allai le soutenir, et si les règles qui gouvernent le prédicat de vérité se réduisent les équivalences-T, pourquoi ne peut-on pas prouver, par exemple, la loi générale du tiers exclu, « Pour tout énoncé, soit cet énoncé est vrai soit sa négation est vraie », à partir des seules règles logiques et des équivalences-T ? La réponse est simplement qu'un tel énoncé n'est pas purement logique et que, comme nous l'avons dit, la formulation de ces énoncés généraux exige essentiellement une réflexion *syntaxique* préalable sur la forme logique des énoncés (réflexion où la vérité n'a aucune part). Or les lois de la syntaxe ne sont pas à proprement parler des lois logiques, et l'on a pas de théorie syntaxique complète effective. Si l'on se donne une théorie syntaxique complète, par exemple en ajoutant à notre syntaxe ordinaire en premier ordre une règle infinitaire de type ω-règle, alors en effet la loi générale du tiers-exclu est conséquence de cette théorie et des équivalences-T. Voir chapitre 2.

[81] Cette justification du principe de réflexion aléthique n'a pas d'analogue pour les principes de réflexions doxastiques ou épistémiques, ce qui le distingue fondamentalement de ces derniers. Notons que dans aucun des deux cas nous n'affirmons que la cohérence de A est une conséquence logique ou analytique de A.

schématique se trouve comme automatiquement renforcée quand le langage environnant est lui-même enrichi, j'ai proposé de parler d'un mouvement d'explicitation de certains engagements implicites dans notre acceptation de la théorie de base. En l'espèce, l' « implicite » que la présence de la théorie tarskienne était vue comme permettant d'expliciter, était localisé assez précisément comme un *reste* du processus de formalisation d'un schéma *ouvert* dans un langage fixé. Dans le présent chapitre, j'ai également affirmé que le prédicat de vérité permettait d'expliciter des engagements implicites dans notre acceptation d'une théorie donnée, et là encore j'ai assez précisément localisé cet implicite, mais je lui ai assigné un autre lieu : le prédicat de vérité permet maintenant d'expliciter le contenu que nous sommes justifiés à accepter si nous sommes justifiés à accepter une théorie donnée (ses axiomes, ses règles). Je n'ai pas changé d'avis entre le chapitre 3 et celui-ci. Il se trouve qu'en explicitant, *via* la théorie tarskienne de la vérité, ce qui est implicite dans l'acceptation des règles d'inférences, il devient possible d'expliciter des engagements arithmétiques ou syntaxiques qui n'avaient pas reçu de formulation dans la théorie de départ en libéralisant notre schéma d'induction. À aucun moment nous n'avions affirmé que la théorie tarskienne articulait la signification du prédicat de vérité, et cette signification seulement. Nous avions indiqué que, même si c'était le cas, si les axiomes tarskiens étaient les principes primitifs de notre compréhension de la notion de vérité, et non des lois dérivées, il ne suivrait pas des conditions sous lesquelles la non-conservativité est observée que le prédicat de vérité est un concept « explicatif ». Nous sommes parvenus à présent à une affirmation plus forte : les axiomes tarskiens ont eux-mêmes un statut dérivé, et ce sont les équivalences-T qui sont conceptuellement primitives. Ce point ne contredit pas le premier, il ne fait que le renforcer et l'éclairer sous un autre jour.

Ces conclusions marquent une étape de ce travail. D'une part nous en avons terminé avec notre tentative pour montrer que le prédicat de vérité ne jouait pas de rôle explicatif dans les preuves de cohérence, et d'autre part nous avons soutenu que la théorie minimale était suffisante pour rendre compte des usages du prédicat de vérité pour l'« expression » des généralisations. Le terrain est ainsi préparé pour la dernière partie, dans laquelle je cherche à défendre plus spécifiquement la thèse du caractère logique du prédicat de vérité.

Conclusion de la troisième partie

En conclusion de cette partie, je voudrais mettre en regard les chapitres 5 et 6. En un sens ces deux chapitres diffèrent du tout au tout, tant par leurs méthodes que par leurs conclusions. Au chapitre 5, j'ai essayé de donner un argument pragmatique pour montrer qu'un sujet acceptant un théorie donnée devait accepter que cette théorie est cohérente, et qu'il pouvait justifier cette acceptation sans ressources aléthiques. L'idée était la suivante. Un sujet acceptant et sachant qu'il accepte une théorie (A, disons) est en mesure d'affirmer

(1) « J'accepte A. »

En vertu du principe de responsabilité, avons-nous soutenu, il *doit* alors accepter

(2) « A est acceptable »

quand bien même (1) pourrait être vrai et (2) faux. (C'est cette étape qui fait que la justification pragmatique de la cohérence n'en est pas une *preuve*.) Puis, si le sujet possède une théorie adéquate de l'acceptabilité, théorie qui doit affirmer, entre autres, que la conséquence logique préserve l'acceptabilité et que les contradictions ne sont pas acceptables, il doit alors pouvoir déduire de (2) et de cette théorie que

(3) A est cohérente

À nouveau, la prémisse principale de l'argument, à savoir (1), pourrait bien être vraie sans que (3) le soit, mais tout le sens de notre développement était de montrer que, malgré tout, la rationnalité de l'action engageait le sujet acceptant une théorie à accepter sa cohérence. La clé de l'explication ici est donc la capacité du sujet à réfléchir sur certaines normes (l'acceptabilité) et la façon dont cette réflexion met

à jour l'obligation dans laquelle est le sujet de lier synchroniquement le *contenu* de ce qu'il accepte et l'*action* d'accepter.

Au chapitre 6, j'ai exploité la tension existant entre le caractère infini de l'objet qu'est une théorie comprise comme ensemble d'énoncés et la finitude des capacités de jugement d'un sujet, pour montrer que ce qu'il est possible de dériver logiquement des contenus des jugements portés effectivement par un sujet qui accepte une théorie excède ce qu'il est possible de dériver simplement des axiomes de la théorie elle-même ; dans cet excès on peut dériver la cohérence elle-même. Une des clés, cette fois, était de reconnaître que le processus par lequel le sujet actualise par un jugement ce qui n'était qu'une acceptation potentielle implique l'explicitation d'un contenu duquel l'énoncé standard de la cohérence de la théorie peut être dérivé.

Nous avons deux descriptions de prime abord radicalement distinctes des processus par lesquels un sujet acceptant une théorie donnée se trouve en position de justifier l'acceptation de la cohérence de la théorie qu'il accepte. Mais ces deux explications ont une description abstraite, et équivoque, commune : il s'agit dans les deux cas de montrer que la réflexion sur les engagements épistémiques qu'il a contracté en acceptant certaines normes de justification permet au sujet de justifier (son acceptation de) la cohérence de la théorie. Il est possible néanmoins que cette description commune, avec ces ambiguïtés calculées sur les termes de « justification », « d'acceptation » et de « réflexion »[82] recèle davantage qu'une simple équivocité. Je laisse pour une prochaine occasion le travail, nécessaire mais trop éloigné de notre sujet pour être poursuivi ici, de clarifier les relations qui existent entre ces deux explications.

[82] « Justifier », recouvre dans un cas l'action de construire une preuve déductive, dans l'autre des processus plus larges (inférences pragmatiques ?, mais peut-on encore parler d'inférence ici dans le passage de (1) à (2) ?), « accepter » signifie tantôt « juger » tantôt recouvre une attitude plus large n'impliquant pas la croyance, et la « réflexion » est tantôt une réflexion permettant de mettre au jour les contenus réels ou latents de nos attitudes tantôt une réflexion sur nos attitudes elles-mêmes.

Quatrième partie

Logicité et vérité

Chapitre 7

Logicité et vérité

7.1 Introduction

Si ce que nous avons dit dans la partie précédente est correct, la théorie minimale de la vérité est suffisante pour rendre compte des usages du prédicat de vérité jugés essentiels par le déflationniste. Elle suffit à articuler notre compréhension du prédicat de vérité comme « correspondance » et rend compte de son emploi comme outil pour l'expression de certaines généralisations. Les deux thèses déflationnistes selon lesquelles, d'une part, la notion de vérité ne joue pas de rôle explicatif substantiel et, d'autre part, est un outil de généralisation, s'avèrent donc être cohérentes entre elles, mais également en cohérence avec la thèse selon laquelle les équivalences-T « disent tout ce qu'il y a à dire » sur la notion de vérité. Dans cette dernière partie, je voudrais considérer une quatrième thèse, qui prolonge naturellement les thèses précédentes : la thèse du caractère logique de la notion de vérité.

Dire que la vérité est une notion logique, ou que les emplois déflationnistes de la notion de vérité sont des emplois logiques, soulève immédiatement une difficulté : qu'est-ce qu'une notion logique ? Si nous n'avons pas de critère de démarcation entre les notions logiques et les autres, ou entre les vérités logiques et les autres, comment soutenir, et quel sens peut avoir, la thèse du caractère logique de la notion de vérité ? Or le problème de la démarcation de la logique est un problème philosophique largement ouvert. Il y a bien un certain nombre d'intuitions partagées concernant le propre de la logique, mais aucune explication précise de ces intuitions ne fait l'objet d'un consensus philosophique suffisamment large pour que nous puissions fonder nos remarques sur elle seule. D'un autre côté, il ne peut pas être question

dans le cadre d'un travail sur la vérité et les thèses déflationnistes de reprendre la question du propre de la logique à nouveaux frais. Par conséquent, il est clair d'emblée qu'on ne peut espérer donner ici un argument absolument conclusif en faveur de la thèse du caractère logique de la notion de vérité. Mais cela ne signifie pas qu'il n'y ait rien de philosphiquement significatif à dire sur la question. En effet on peut néanmoins présenter un faisceau d'indices en faveur de la thèse de la logicité de « vrai », un ensemble d'éléments de réflexion qui, en éclairant au cas par cas la façon dont notre compréhension de la notion de vérité répond à nos différentes intuitions et tentatives d'explication de la notion de logicité, et si ces éléments convergent, confèrent ensemble à la thèse une certaine plausibilité. Cette méthode d'argumentation donnera donc par force au développement un caractère haché, dans la mesure où la complémentarité des approches n'est pas totalement comprise. Et puisque, en outre chacune de ces approches souffre la critique et revêt un caractère inabouti, il faut comprendre ce travail moins comme une défense de la thèse de la logicité de la notion vérité que comme une exploration philosophique de liens existant entre les thèses déflationnistes et l'idée de logicité de la vérité.

Il existe deux traditions dominantes dans la recherche d'un critère de démarcation entre les notions logiques et les notions non logiques. Une tradition que j'appellerai *inférentialiste*, et une tradition que j'appellerai *sémantique*. Chacune de ces traditions présente un certain nombre d'intuitions préthéoriques proches relativement à la nature de la logicité et développe une méthodologie d'analyse raisonnée qui lui est propre. Mon plan sera donc de procéder en deux étapes. Dans la première, je reviens sur les grandes lignes du programme inférentialiste. Je commence par considérer la notion « inférentialiste » de validité d'un argument, et je montre que les règles minimales pour la vérité sont valides en ce sens. Puis je considère plus spécialement la question de la validité *logique* et discute du caractère logique du prédicat de vérité dans cette perspective. Dans la seconde partie, je propose un cadre d'analyse sémantique pour rendre compte des « emplois déflationnistes » du prédicat de vérité et en défendre le caractère logique.

7.2 Sémantique inférentielle et vérité

7.2.1 La validité sans la vérité

Au chapitre 6 j'ai insisté sur le fait que, s'il accepte certaines règles d'inférence, un sujet accepte également la préservation de la vérité par les règles en question. Mais *affirmer*, simplement, que les règles sont valides, ce n'est pas *expliquer* en quoi consiste la validité d'une inférence, ni expliquer pourquoi précisément les règles que l'on accepte *sont* valides. Qu'une inférence soit valide cela doit dépendre de la signification des énoncés en jeu uniquement, et l'on attend d'une *explication* de la validité qu'elle rende compte de ce fait par une explication de ce en quoi cela consiste, pour une expression, d'avoir la signification qu'elle a. On peut voir l'explication standard de la conséquence logique,[1] que l'on pense plus particulièrement à une explication modèle-théorique ou à une explication en termes de vérité *simpliciter*, comme une explication de ce type, dans laquelle la signification d'un énoncé est « expliquée » en termes de conditions de vérité, et la signification des constantes logiques en termes de leur contribution aux conditions de vérité des énoncés.[2] Dans cette perspective, les lois de compositionnalité du prédicat de vérité et des constantes logiques, c'est-à-dire les énoncés comme

$$\forall x, y, \text{ si } x \text{ et } y \text{ sont des énoncés alors}$$
$$(Vr(\ulcorner x \vee y \urcorner) \leftrightarrow Vr(x) \vee Vr(y)),$$

qui précisent exactement ce qu'est la contribution du mot « $\vee$ » aux conditions de vérité d'un énoncé, sont vues comme des *explications* de la signification du mot

[1] Un mot, ici, de la relation entre « inférence » et « conséquence ». Une inférence est une action par laquelle un sujet passe d'un jugement, ou d'un ensemble de jugements, à un jugement, et l'on peut voir une règle d'inférence comme un opération définie sur des ensembles de jugements à valeur dans un ensemble de jugements ; tandis que la notion de conséquence logique est une relation entre un ensemble de propositions (ou d'énoncés) et une proposition (un énoncé). Ce sont donc des objets de catégories distinctes. Néanmoins, je supposerai dans ce qui suit qu'il y a un lien intangible entre conséquence (logique) et inférence (logiquement) valide : une inférence d'un ensemble de jugements J à un jugement j est (logiquement) valide si et seulement si l'ensemble des propositions (énoncés) Γ qui sont des contenus de jugements dans J ont pour conséquence (logique) la proposition (énoncé) qui est le contenu du jugement j. Ce principe nous autorise à passer des discussions sur les inférences à des discussions sur la conséquence et réciproquement. Je note en passant que ce principe pourrait être critiqué en ce qu'il force la compacité de la relation de conséquence logique. Discuter de la portée de ce point nous amènerait trop loin, je le laisse de côté ici.

[2] *Via* la définition de la vérité dans une structure, ou *via* les clauses récursives de la théorie compositionnelle de la vérité *tout court*.

« $\vee$ ».

En tant que modèle général d'explication de la notion de conséquence *logique*, cette façon de procéder appelle deux remarques ici. Tout d'abord, telle quelle, cette explication *suppose*, plutôt qu'elle n'explique, la distinction entre expressions logiques et expressions non-logiques, en ne disant rien de ce qui distingue la signification des constantes logiques.[3] D'autre part, cette explication n'est pas, de prime abord, en accord avec ce que nous avons dit au chapitre 6 concernant l'usage de la notion de vérité. En effet, nous avons soutenu que les clauses compositionnelles se comprenaient d'abord comme une façon d'expliciter ce que nous acceptons en suivant les règles d'inférence que nous suivons. De ce point de vue, en affirmant l'énoncé de la composition de « $\vee$ » et « Vr » ci-dessus, je ne suis pas en train de donner une explication profonde de ce en quoi consiste ma compréhension de la signification de « $\vee$ », mais seulement en train de formuler en toute généralité une affirmation à laquelle m'engage le lien qui existe entre mes affirmations d'énoncés de la forme $\ulcorner x \vee y \urcorner$ et mes affirmations x et y, en vertu de mon allégeance à une certaine pratique inférentielle (les règles d'inférence logique). Il est clair, néanmoins, qu'un locuteur qui est en position de faire ce genre d'affirmation (l'affirmation de la compositionnalité) sait quelque chose de la signification de « $\vee$ », puisqu'il faut bien en dernière instance que ce soit en vertu de la signification de « $\vee$ », quelle que soit la manière dont on veut analyser celle-ci, que ma pratique des règles gouvernant « $\vee$ » soit justifiée ; mais la dérivation des clauses de compositionnalité par réflexion sur ma pratique inférentielle ne peut compter inversement comme une explication de ce pourquoi ma pratique inférentielle est justifiée, et donc comme *explication* de la validité de ces règles. De la manière dont j'ai rendu compte de l'usage de « vrai », un sujet qui suit une règle R exprime ce qu'il accepte par là en affirmant que la règle préserve la vérité : ceci n'explique pas en quoi consiste la validité d'une règle, mais montre seulement que le sujet doit tenir pour valide une règle qu'il suit.[4] Par conséquent, d'après notre tableau de l'usage du prédicat de vérité, il semble plus naturel de chercher à expliquer la validité d'une règle autrement que par la

[3] Tarski, toutefois, en dit davantage dans Tarski (1966/1986). Voir l'appendice A1.

[4] Rappelons, qu'inversement, il est possible de dériver la validité de chaque règle logique à partir des clauses de compositionnalité. Par conséquent, si nous parvenions à justifier les clauses de compositionnalité sans faire explicitement l'hypothèse de la correction de notre pratique inférentielle, nous serions en possession d'une justification de notre pratique inférentielle. Bien entendu, s'il fallait prendre comme un fait brut les clauses de compositionnalité, cela ferait du prédicat de vérité une notion explicative. Mais c'est précisément à montrer qu'il ne s'agit pas d'une loi brute de la vérité que nous avons consacré une partie du chapitre 6.

signification véricontionnelle des expressions qui y figurent.[5]

Michael Dummett, Dag Prawitz, et d'autres, ont pris au sérieux l'idée que ce n'est pas par la détermination des conditions de *vérité* d'un énoncé que se constitue la compréhension de sa signification par sujet. Le point de départ de leur approche est que ce qui compte comme la signification d'une expression doit être déterminable par la pratique manifeste des locuteurs, et que par conséquent cette signification doit être fixée par l'usage, ou certains usages, qui en sont faits dans les inférences. La signification des énoncés est alors identifiée à leur *rôle inférentiel*, et la signification de leurs expressions constituantes doit être expliquée par la façon dont elles contribuent systématiquement au rôle inférentiel des énoncés, autrement dit leur contribution à ce qui doit compter comme justifiant un énoncé d'une part et, d'autre part, à ce qui peut être justifié par cet énoncé. Cette contribution spécifique des constituants subphrastiques à la signification inférentielle des phrases est explicitée dans deux types de règles d'inférences : celles par lesquelles une expression peut être introduite en position dominante dans un énoncé (règles d'introduction) et celles par lesquelles une expression en position dominante dans cet énoncé peut en être éliminée (règles d'élimination). À chaque expression est ainsi associée un ensemble restreint de règles qui *constituent*, ou *fixent*, la signification inférentielle de cette expression. La validité des inférences en général, et des inférences logiques en particulier, est alors expliquée par la signification inférentielle des expressions qui figurent dans leurs prémisses et leur conclusions, et donc en dernière instance par la pratique inférentielle qui détermine la signification des expressions.[6]

Une approche « inférentialiste » de la sémantique présente des traits intéressants pour une étude comme la nôtre. D'une part, nous l'avons dite, l'indépendance de l'analyse relativement au concept de vérité fait de l'inférentialisme une approche

[5]Pour comparaison :

Affirmation/Interprétation	Approche « sémanticiste »	Approche « déflationniste »
$Vr(x * \overline{\vee} * y)$ $\leftrightarrow$ $Vr(x) \vee Vr(y)$	Explique la signification de « $\vee$ »	Explicite ce qui est implicite dans une pratique de justification dont fait partie l'emploi des règles pour « $\vee$ »
Clauses récursives (+ Syntaxe) $\Rightarrow$ « Les règles logiques préservent la vérité »	Explication de la validité des règles logiques obtenue par analyse de la signification des constantes logiques	*Si* l'on avait une justification de chacune des règles on pourrait justifier l'affirmation globale de la validité des règles

[6]Le schéma général est donc analogue à celui de l'explication standard de la validité en termes de la signification des termes qui figurent dans les énoncés en jeu, sauf que c'est désormais la « signification inférentielle » qui remplace l'explication vérionditionnelle de la signification.

privilégiée pour le déflationniste : si ce dernier a répudié le caractère explicatif de la vérité dans les théories de la signification, il est tentant pour lui d'embrasser une approche appartenant au paradigme de la signification-comme-usage. L'analyse inférentialiste n'est qu'une espèce de ce genre, se concentrant sur les inférences déductives. C'est pourquoi l'analyse inférentielle est de prime abord davantage en harmonie avec la méthodologie déflationniste générale qu'une explication de la validité qui se fonderait d'emblée sur une analyse vériconditionnelle de la signification et qui serait formulée en termes de préservation de la vérité. D'autre part, et plus simplement, l'analyse inférentielle ne présuppose pas une compréhension antérieure de la notion de vérité. Puisque le concept de vérité n'a pas d'occurrence essentielle dans la boîte à outil conceptuelle de l'analyse inférentielle de la signification et de la validité, on peut penser que cette analyse est neutre quant aux conclusions que l'on pourra en tirer sur la notion de vérité, au sens où l'on ne risque pas en quelque façon, en adoptant un point de vue inférentialiste sur la « sémantique » de la notion vérité, d'avoir d'emblée présumé de sa nature.

Dans ce qui suit, l'esquisse de ce que pourrait être une explication non-véri-conditionnelle de la validité logique sert donc un double but : elle permet de conforter dans un domaine particulier l'intuition déflationniste que la vérité n'a pas de rôle explicatif substantiel à jouer dans les explications[7], et d'autre part, surtout, d'appliquer cette analyse inférentielle au prédicat de vérité lui-même, en espérant de cette analyse « sémantique » pouvoir tirer un éclairage sur sa nature.

Harmonie Revenons maintenant à l'explication de la validité des inférences proposée par les inférentialistes. Elle est simple dans son principe. Supposons que certaines règles d'inférence déterminent, ou même constituent, la signification de certaines expressions, quoi que cela veuille dire exactement. Ces règles sont donc valides par définition, en vertu de la signification de ces expressions. Appelons ces règles des règles d'inférences *canoniques*. Supposons qu'un énoncé ϕ soit prouvable

	Analyse vériconditionnelle	Analyse en termes d'usage
Signification d'un énoncé =	Conditions de Vérité	Rôle Inférentiel
Signification d'un constituant = exemple connecteur δ	contribution aux CV $Vr(x * \overline{\delta} * y)$ ssi $Vr(x)...Vr(y)$	contribution au RI règles δ-introduction et δ-élimination
Validité logique d'une inférence/d'un énoncé	déterminée par la signification des mots logiques	*idem*

[7]Ici, l'explication de la signification des expressions logique et l'explication de la correction, ou de la validité, des inférences.

en ne faisant appel qu'à des inférences canoniques : alors ϕ est certainement valide, étant la conclusion d'une suite d'inférences valides. À partir de là, comment définir la validité d'une inférence quelconque ? Il est naturel de proposer qu'une inférence d'un ensemble d'énoncés Γ à un énoncé ψ est valide, ou que ψ est conséquence de Γ, si l'on peut dériver ϕ de Γ en ne faisant appel qu'aux règles d'inférences canoniques.[8]

Sur le fond de cette esquisse, deux questions importantes émergent :

1. Tout ensemble de règles d'introduction et d'élimination d'une expression vaut-il définition d'une expression, c'est-à-dire confère-t-il une signification inférentielle

[8]Le détail de la définition de la validité d'une inférence change substantiellement d'un auteur à un autre, mais discuter ces détails me mènerait trop loin. Dans ce qui suit, je m'en tiens autant que possible aux considérations générales qui gouvernent l'ensemble des propositions. Voir par exemple Prawitz (1971), Dummett (1991) pour une présentation des définitions respectives de ce deux auteurs. Prawitz (2006) contient une comparaison des deux approches.Voir aussi Hacking (1979) pour une approche encore différente.

Pour fixer les idées, je signale ici une spécificité de l'approche de Dummett et Prawitz. Ces deux auteurs s'inspirent plus directement d'une remarque de Gentzen devenue célèbre :

> Les règles d'introduction représentent, pour ainsi dire, les « définitions » des symboles concernés, et les règles d'élimination ne sont pas davantage, en dernière analyse, que les conséquences de ces définitions. Ce fait pourrait être exprimé de la façon suivante : en éliminant un symbole, on ne peut utiliser la formule dont nous élimnons le symbole dominant que « dans le sens permis par la règle d'introduction de ce symbole » (Gentzen (1969), p.80)

Dans cette perspective, pour Dummett comme pour Prawitz, ce sont les *règles d'introduction* d'une expression, et elles seules, qui sont comprises comme « définissant » l'expression. Ces règles explicitent plus spécialement la contribution de l'expression aux conditions de prouvabilité ou de « vérification » de l'énoncé, et en ce sens ce choix peut être compris comme un parti pris « vérificationniste » sur la signification. Dans cette perspective, la validité des règles d'élimination *n*'est *pas* définitionnelle, mais doit être justifiée comme celle de n'importe quelle autre inférence, par la signification conférée à l'expression par sa, ou ses, règles d'introduction. Anticipant quelque peu la discussion qui va suivre, je précise que lorsque Dummett parle d'*harmonie* entre une règle d'élimination d'une expression et sa règle d'introduction il entend spécialement l'idée que l'on peut justifier la validité de la première par la seconde.

Mais l'idée que, dans la représentation des inférences en déduction naturelle, seules les règles d'introduction d'un symbole constituent sa signification inférentielle, tandis que les règles d'élimination sont secondes, cette idée n'est pas contraignante (et ne va pas sans poser des difficultés, voir la discussion lucide de Dummett sur son « hypothèse fondamentale » dans Dummett (1991) et n'est pas une hypothèse généralement acceptée. Les auteurs qui rejettent l'idée d'une priorité d'un type de règle sur l'autre préfèrent en général raisonner sur une représentation des arguments en calcul des séquents, qui ne contient formellement que des règles d'introduction des expressions (en plus des règles structurales), tantôt dans les *hypothèses*, tantôt dans les *conclusions* d'une dérivation. La question de l' « harmonie » entre les deux types de règles d'introduction se pose toujours, mais la « symétrie » sous laquelle elles se présentent ôte tout sens immédiat à l'idée que l'une des deux seulement fixerait la signification de l'expression devrait permettre de justifier l'autre. Ce type d'approche est illustré dans Hacking (1979) ou Kremer (1988). Quelle que soit l'approche adoptée, le point essentiel reste le même : il s'agit de justifier la validité d'une inférence en général par la validité définitionnelle des règles qui fixent la signification des expressions.

à cette expression ?

2. A quelles conditions une expression définie par un ensemble de règles d'introduction et d'élimination est-elle une expression logique ?

La réponse à la première question doit être Non, puisque l'analyse de ce qui constitue la signification des expressions doit avoir une portée normative et permettre d'expliquer la notion de validité. Ce sont les règles proposées par Prior (1960) pour l'expression dissonante « tonk » qui illustrent sans doute le mieux la nécessité d'un critère normatif, que Dummett appelle « harmonie » entre règles d'introduction et règles d'élimination. La signification inférentielle de « tonk » est supposée être fixée par les règles d'introduction de la conjonction et d'élimination de la disjonction, de la façon suivante : [9]

$$\text{Tonk-int}\ \frac{\Gamma \vdash A}{\Gamma \vdash A \text{ tonk } B} \qquad \frac{\Gamma \vdash A \text{ tonk } B}{\Gamma \vdash B}\ \text{Tonk-elim}$$

Il est facile de voir qu'ajouter les règles pour « tonk »' à un système déductif produit un désastre logique. Voici par exemple une preuve de l'existence de Dieu :

$$\text{Tonk-elim}\ \frac{\text{Tonk-int}\ \dfrac{\vdash\rightarrow\ \dfrac{A \vdash A}{\vdash (A \rightarrow A)}}{\vdash (A \rightarrow A) \text{ tonk Dieu existe}}}{\vdash \text{ Dieu existe}}$$

Il y a donc certainement un problème avec « tonk », et nous avons là une illustration de la nécessité de soumettre l'ensemble des règles gouvernant la signification d'un symbole à un critère normatif si l'on veut pouvoir rendre compte de l'idée de validité. En l'espèce, si nous sommes capables de trouver un critère d'harmonie satisfaisant, ce critère doit permettre de montrer que les règles pour *tonk ne* sont *pas* en harmonie (et accessoirement que l'argument présenté à l'instant n'est pas valide).

Il y a deux façons de répondre au problème. La première est d'en juger par les effets, et d'inscrire dans la notion d'harmonie l'obligation pour les règles de *n*'avoir *pas* certaines conséquences délétères sur leur environnement déductif. C'est la voie

[9] Une remarque. Il ne faut pas confondre le calcul des séquents avec une simple présentation dans le style des séquents d'un calcul de déduction naturelle. Le calcul en déduction naturelle possède des règles d'introduction et d'élimination des constantes logiques. Le calcul des séquents n'a que des règles d'introduction, à gauche et à droite. La déduction naturelle au format séquents n'est qu'une variante notationnelle de déduction naturelle où l'ensemble des hypothèses dont dépend une preuve à une étape donnée est marqué à cette étape à gauche du symbole « $\vdash$ ». Nous adoptons cette notation plus compacte ici.

proposée par Belnap (1961) et reprise, avec une variante, par Dummett (1991) sous le nom d'harmonie *globale*.

Critère de conservativité La proposition de Dummett est la suivante : ajouter une constante logique gouvernée par ses règles d'introduction et d'élimination à un langage contenant déjà éventuellement un certain nombre de règles pour d'autres constantes logiques doit produire une extension conservative de ce langage.

Bien entendu, les règles pour « tonk » ne satisfont pas ce critère, du moins si elle sont ajoutées à un système de règles qui n'est pas déjà contradictoire. Mais le critère de conservativité, tel que nous l'avons interprété[10], n'invalide pas seulement les règles pour « tonk ». Ainsi, lorsqu'on ajoute les règles classiques pour la négation au fragment implicatif de la logique classique[11], on obtient une extension non-conservative : on sait par exemple que la loi de Pierce[12], qui n'est pas dérivable dans le fragment implicatif de la logique classique présentée en déduction naturelle (mono-conclusion), y devient dérivable en présence des règles classiques pour la négation. C'est donc un critère extrêmement contraignant, trop peut-être. Et il y a des complications :

[10]Je note simplement ici la variante plus faible de la contrainte de conservativité proposée par Belnap. L'ajout des règles produit une extension non-conservative, en son sens, si ajouter ces règles à un calcul déductif ne valide pas de nouvelles *lois générales de la déduction*, c'est-à-dire de nouvelles lois purement structurales. La théorie de la déduction de base est constituée seulement des axiomes suivants

1. (Identité) $A \vdash A$
2. (Affaiblissement) $\dfrac{A_1, \ldots, A_n \vdash C}{A_1, \ldots, A_n, B \vdash C}$
3. (Permutation) $\dfrac{A_1, \ldots, A_i, A_{i+1}, \ldots, A_n \vdash C}{A_1, \ldots, A_{i+1}, A_i, \ldots, A_n \vdash C}$
4. (Contraction) $\dfrac{A_1, \ldots, A_n, A_n \vdash C}{A_1, \ldots, A_n \vdash C}$
5. (Transitivité) $\dfrac{A_1, \ldots, A_m \vdash B \qquad C_1, \ldots, C_n, B \vdash D}{C_1, \ldots, C_n, A_1, \ldots, A_m \vdash D}$

On pourrait montrer que l'ajout des règles pour « tonk » donnée par Prior à un système déductif satisfaisant la loi de la transitivité ci-dessus, valide la nouvelle loi structurale suivante : $A \vdash B$. Par conséquent, les règles pour « tonk » ne sont pas harmonieuses relativement à un système déductif de ce type. Il est facile de voir que la conservativité au sens de Dummett est une condition plus forte que la conservativité au sens de Belnap : ajouter les règles classiques aux règles intuitionistes en déduction naturelle produt une extension non-conservative au sens de Dummett, mais pas au sens de Belnap.

[11] En déduction naturelle mono-conclusion.

[12]C'est-à-dire :

$$(((P \to Q) \to P) \to P).$$

car si ajouter les règles classiques pour la négation résulte en une extension non-conservative du fragment implicatif de la logique classique, ajouter ces règles à un langage ne contenant aucune constantes logiques (dans un calcul monoconclusion), par exemple, *est* conservatif. Ainsi, l'introduction d'une même constante logique à un système de règles pourra ou non résulter en une extension conservative du système de départ selon les cas, et un même ensemble de règle pour une expression apparaîtra comme correct ou non selon le « moment » auquel il a été introduit dans le système. De plus, la conservativité est aussi sensible aussi au cadre de représentation dans lequel sont données les règles.[13] Dans un calcul où les déductions sont représentées avec des conclusions multiples, on peut formuler les règles pour « non » de façon à ce qu'elles étendent conservativement le fragment implicatif de la logique classique ! Ces problèmes montrent qu'en tant que condition *nécessaire* d'harmonie, la conservativité n'est pas satisfaisante et demanderait à être raffinée. Mais le critère est aussi insatisfaisant en raison de son caractère « extrinsèque » : on peut admettre qu'il faut en effet que des règles définitionnelles, c'est-à-dire fixant la signification d'une expression, soient non-créatives, mais le critère de conservativité ne dit pas *pourquoi* tel ensemble de règles est harmonieux et non tel autre, il nous permet seulement de le constater. Où est passée l'idée de Gentzen selon laquelle les règles d'élimination sont des « conséquences » des règles d'introduction ? Pour rendre compte de cette idée, il faut un second critère, un critère d'harmonie « locale ».

En quoi consiste l'harmonie locale des règles gouvernant l'usage d'un symbole ? L'idée générale est assez claire, même si, dans le détail, les propositions sont nombreuses : c'est celle d'un équilibre qui doit prévaloir entre les conditions qui justifient l'affirmation d'un énoncé, et les conséquences qu'il est permis de tirer de l'affirmation de cet énoncé. Il s'agit de s'assurer que les conséquences les plus fortes que l'on peut inférer directement d'un énoncé $\sigma(\delta)$, dont « δ » est l'expression dominante, par la règle d'élimination de « δ », n'excèdent pas ce qu'il est justifié d'accepter sur la base des conditions les plus faibles sous lesquelles il est possible d'inférer directement $\sigma(\delta)$ *via* sa règle d'introduction. Dans le cadre d'une représentation des arguments dans le format de la déduction naturelle, avec ses règles d'introduction et d'élimination des symboles, ce principe est saisi de façon plus précise par ce que Dag Prawitz a appelé un *principe d'inversion* :

[13] Ce qui explique l'obligation des réserves portant sur le caractère « mono-conclusion » du calcul dans les lignes précédentes.

> Soit α une application d'une règle d'élimination ayant B pour conséquence. Alors, les déductions qui satisfont la condition suffisante [...] pour dériver la prémisse majeure de α,[14] lorsqu'elles sont combinées aux déductions de la prémisse mineure de α (s'il y en a), « contiennent » déjà une déduction de B ; la déduction de B peut donc s'obtenir directement à partir des déductions données sans l'ajout de α. (Prawitz (1965), p. 33)

Voyons comment le principe fonctionne sur un exemple. La règle d'introduction usuelle de l'implication est la suivante :

$$\rightarrow\text{-intro}\ \frac{\begin{array}{c}[A]\\ \Pi\\ B\end{array}}{A \rightarrow B}$$

où Π est une dérivation de B sous l'hypothèse A, et où l'hypothèse A est déchargée au moment de l'introduction de l'implication. La règle d'élimination usuelle de l'implication est la suivante (*Modus Ponens*) :

$$\rightarrow\text{-elim}\ \frac{\begin{array}{cc}\Pi_1 & \Pi_2\\ A & A \rightarrow B\end{array}}{B}$$

où Π_1 et Π_2 sont respectivement des dérivations de A et de $A \rightarrow B$. Pour montrer que les règles usuelles d'introduction et d'élimination de l'implication sont en harmonie au sens de Prawitz, il suffit de remarquer qu'une dérivation dans laquelle la règle d'introduction et d'élimination de l'implication de la flèche se succèdent peut être *réduite* :

$$\rightarrow\text{-elim}\ \frac{\begin{array}{cc}\Pi_1 & \begin{array}{c}[A]\\ \Pi\\ B\end{array}\\ A & \rightarrow\text{-intro}\ \dfrac{}{A \rightarrow B}\end{array}}{B} \quad \Rightarrow \quad \begin{array}{c}\Pi_1\\ [A]\\ \Pi\\ B\end{array}$$

On peut transformer la première preuve de B, où la prémisse majeure de la règle $\rightarrow$-élim a été obtenue par la règle d'introduction définitionnelle de « $\rightarrow$ », en une preuve de B sous les mêmes hypothèses mais dont l'occurrence de la règle d'introduction et de la règle d'élimination a disparu. L'existence de ce type de procédure

[14] N.d.t : La prémisse majeure, c'est-à-dire l'occurrence de l'énoncé dont l'expression dominante a été éliminé par l'application α de la règle.

de *réduction* montre que la succession d'une inférence introduisant l'implication et d'un *modus ponens* n'est qu'un simple détour dans l'argument : ce qui est dérivable par élimination de l'expression principale d'une formule est « déjà là » dans toute preuve de cette formule dont la dernière règle est une règle d'introduction. L'idée que la règle d'élimination de l'expression pourrait être *justifiée* par la règle d'introduction trouve ici un sens précis.[15] Nous avons un second critère d'harmonie, local celui-là :[16]

Critère de réduction locale Les règles d'introduction et d'élimination d'une expression sont en harmonie si ces règles satisfont le principe d'inversion.

Il est facile de se convaincre que les règles d'introduction et d'élimination de « tonk » ne sont pas en harmonie au sens où la règle d'élimination de « tonk » n'est pas justifiée par sa règle d'introduction. Dans l'argument donné plus haut pour l'existence de Dieu, la règle d'introduction de « tonk » était suivie immédiatement de sa règle d'élimination, mais ce détour était essentiel, indispensable à la preuve de la conclusion. Un argument valide pour les prémisses de la règle d'introduction ne peut être transformé en un argument valide pour la conclusion de la règle d'élimination.

[15] Lorsque, dans un calcul, avec les règles d'introduction et d'élimination pour tous les symboles du langage considéré, tous les détours dans tous les arguments sont éliminables, on dit que le calcul est normalisable. On peut voir la procédure de normalisation d'une preuve comme une procédure de justification de sa validité.

Quand on abandonne l'idée du caractère prioritaire des règles d'introduction sur les règles d'élimination dans la constitution de la signification inférentielle, et que l'on représente les arguments dans le calcul des séquents, il faut formuler les choses un peu différemment. L'analogue du principe d'inversion est l'*élimination des coupures* sur les occurrences de formules dont le connecteur principal est celui dont on cherche à justifier localement les règles. En prenant l'exemple des règles pour la conjonction, on peut voir qu'une dérivation avec coupure sur une formule conjonctive qui a été introduite à gauche et à droite par ses règles définitionnelles peut être transformée en une dérivation d'où la coupure sur cette occurrence de la formule a disparu. En partant de

$$\dfrac{\dfrac{\dfrac{\Gamma_1 \vdash \phi, \Delta_1 \qquad \Gamma_2 \vdash \psi, \Delta_2}{\Gamma_1, \Gamma_2 \vdash \phi \wedge \psi, \Delta_1, \Delta_2}\,(\vdash \wedge) \qquad \dfrac{\Gamma_3, \phi \vdash \Delta_3}{\Gamma_3, \phi \wedge \psi \vdash \Delta_3}\,(\wedge \vdash)}{\Gamma_1, \Gamma_2, \Gamma_3 \vdash \Delta_1, \Delta_2, \Delta_3}\,(\text{cut})$$

on obtient

$$\dfrac{\dfrac{\Gamma_1 \vdash \phi, \Delta_1 \qquad \Gamma_3, \phi \vdash \Delta_3}{\Gamma_1\Gamma_3 \vdash \Delta_1, \Delta_3}\,(\text{cut})}{\Gamma_1, \Gamma_2, \Gamma_3 \vdash \Delta_1, \Delta_2, \Delta_3}\,(\text{Affaiblissement})$$

L'analogue de la normalisation est le théorème d'élimination des coupures de tous les arguments. Voir par exemple, à nouveau, Hacking (1979), Kremer (1988).

[16] Il se peut que ce critère soit satisfait sans que premier le soit. J'en dis plus sur la relation entre ces deux critères dans l'appendice à ce chapitre.

Nous avons donc deux critères normatifs permettant de statuer sur la question de savoir si un ensemble de règles permet de fixer la signification d'une expression, la conservativité et la réduction locale. Pour arriver à nos fins, il faut encore dire quelque chose de la logicité.

Logicité L'analyse précédente, en toute généralité, ne présumait pas du type d'expression dont des règles d'introduction et d'élimination sont supposées fixer la signification : il pouvait s'agir d'expressions logiques ou non, le problème n'était pas là. Mais pour savoir ce qu'il y a de spécial avec les expressions logiques, ce qui distingue leur signification, il suffit maintenant s'intéresser à ce qui distingue, ou doit distinguer, les règles qui fixent leur signification.

1. Considérons d'abord quelques règles putatives fixant la signification de termes clairement non-logiques.

   a) Par exemple, des règles d'inférence gouvernant le prédicat « être un carré », noté « $Carre(x)$ » :

   $$\text{Carré-Intro}\ \frac{\Gamma \vdash Parallelogramme(x) \qquad \Gamma \vdash Aunangledroit(x)}{\Gamma \vdash Carre(x)}$$

   $$\text{Carré-elim}\ \frac{\Gamma \vdash Carre(x)}{\Gamma \vdash Parallelogramme(x)}$$

   $$\text{Carré-elim}\ \frac{\Gamma \vdash Carre(x)}{\Gamma \vdash Aunangledroit(x)}$$

   b) Ou encore les fameuses règles données par Dummett pour l'expression péjorative « Boche » [17] :

   $$\text{Boche-int}\ \frac{\Gamma \vdash Allemand(x)}{\Gamma \vdash Boche(x)} \qquad \text{Boche-elim}\ \frac{\Gamma \vdash Boche(x)}{\Gamma \vdash Cruel(x)}$$

   c) Enfin les règles pour le terme théorique « neutrino », élaborées dans l'esprit de la méthode de Carnap-Ramsey[18], où $\exists X T(X)$ est l'énoncé de Ramsey de la théorie supposée introduire le terme « neutrino » :

   $$\text{neutrino-int}\ \frac{\Gamma \vdash \exists X T(X)}{\Gamma \vdash T(neutrino)} \qquad \text{neutrino-elim}\ \frac{\Gamma \vdash T(neutrino)}{\Gamma \vdash \exists X T(X)}$$

2. Et considérons d'un autre côté les règles d'introduction et d'élimination du connecteur habituel de conjonction, « $\wedge$ », en déduction naturelle :

[17] Par exemple dans Dummett (1973), p.454. Les règles ne sont pas en harmonie.
[18] Voir Carnap (1966), chap. 26-28 et Lewis (1970).

$$\wedge\text{-int}\ \frac{\Gamma \vdash A \qquad \Gamma \vdash B}{\Gamma \vdash A \wedge B} \qquad \wedge\text{-elim}_1\ \frac{\Gamma \vdash A \wedge B}{\Gamma \vdash A} \qquad \wedge\text{-elim}_2\ \frac{\Gamma \vdash A \wedge B}{\Gamma \vdash B}$$

La différence frappante entre ces groupes de règles réside en ce qu'elles diffèrent fondamentalement quant à l'appareil conceptuel d'arrière-plan qu'elles mobilisent dans le processus de « constitution de la signification ». Ce qui compte comme justification pour introduire « neutrino » est hautement chargé théoriquement, et nécessite la compréhension spécifique de certains termes autres que le terme défini par la règle. Il en est de même pour les deux autres règles : dans les règles pour « Boche » comme dans celles pour « parallélogramme », les fondements et les conséquences des assertions contenant le terme défini comme expression principale ont besoin pour être décrits de prédicats spécifiques, « Allemand », « cruel », « paraléllogramme ». À l'inverse, ce qui est distinctif des dernières règles, c'est que les fondements pour l'assertion d'un énoncé $\sigma(\wedge)$, et les conséquences directes que l'on peut en tirer, sont formulées de façon purement *structurales*, c'est-à-dire simplement à l'aide d'un signe primitif d'inférence et de variables pour les énoncés.[19]

C'est ce qui fait d'elles des règles à caractère formel : aucune connaissance du « contenu » des prémisses n'est nécessaire pour l'usage correct de l'expression, ni donc pour la détermination de sa signification. Dans la mesure où ce qui détermine la signification inférentielle de l'expression peut être décrit syntaxiquement uniquement à l'aide de variables[20], du symbole de relation de déductibilité « $\vdash$ » et d'un nom de l'expression dont les règles sont supposées déterminer la signification, la signification de l'expression introduite peut être dite ne refléter rien d'autre que les aspects les plus généraux de notre pratique déductive elle-même. Si l'on appelle *formelle* une règle dont la description satisfait ces propriétés, nous obtenons donc

[19] On pourrait comparer les ressources mises en jeu dans la formulation des règles d'introduction pour « $\wedge$ » et pour « boche » (par exemple) dans un langage plus net, en reprenant les notations de la syntaxe présentée au chapitre 2 et en se donnant un symbole pour la déductibilité (« $\vdash$ »). La règle d'introduction de « Boche » devient :

Pour tout ensemble fini σ d'énoncés du langage langage-objet, pour tout entier naturel k,
si $\sigma \vdash \overline{Allemand} * \overline{(} * v_k * \overline{)}$ alors $\sigma \vdash \overline{Boche} * \overline{(} * v_k * \overline{)}$

La règle d'introduction pour la conjonction devient :

Pour tout ensemble fini σ d'énoncés (du langage-objet), pour tous énoncés x et y,
si $\sigma \vdash x$ et $\sigma \vdash y$ alors $\sigma \vdash \overline{(} * x * \overline{\wedge} * y * \overline{)}$

[20] Les variables de notre syntaxe, prenant pour valeurs les énoncés du langage-objet, celui pour lequel nous décrivons les règles.

le critère de logicité suivant :[21]

Critère de logicité Une expression est logique si et seulement si les règles qui fixent sa signification inférentielle sont formelles.

7.2.2 Approche inférentielle du prédicat de vérité

Il nous reste à présent à appliquer ces considérations au cas du prédicat de vérité, en commençant par présenter les règles « définitionnelles » qui nous intéressent.

La théorie minimale Si ce que nous avons dit au chapitre 6 est correct, tout ce dont nous avons besoin pour expliquer nos usages inférentiels de la notion de vérité est la théorie minimale de la vérité. La théorie minimale suffit à rendre compte de l'intuition de la correspondance (elle est adéquate au sens de Tarski) et elle suffit également pour rendre compte de l'usage du prédicat de vérité pour l'« expression de généralisations ».

Nous savons également que ces règles ne suffisent pas, en conjonction avec les règles logiques habituelles, pour prouver des généralisations contenant « vrai », comme par exemple ce que Tarski appelait la « loi du tiers-exclu » :

$$\text{(TE)}\ \forall x,\ \text{Enoncé}(x) \rightarrow (Vr(x) \vee Vr(\ulcorner \neg x \urcorner))$$

Cette remarque a plus d'une fois été faite pour servir une critique sur l'insuffisance explicative de la théorie minimale.[22] Nous avons déjà répondu à cette critique en remarquant au chapitre 6 que la formulation et la justification de ces généralisations exigeait tout d'abord une réflexion syntaxique dans laquelle la notion de vérité n'a aucune part.[23] Mais nous avons également noté qu'un énoncé comme (TE) ne relevait pas de la simple description de la syntaxe d'un langage. La syntaxe n'y

[21] J'adopterai ce critère sur cette base, et je renvoie le lecteur à Dummett (1991), p.250 *sqq.*, ainsi qu'à Sambin, Battilotti et Faggian (2000), pour un approfondissement.

[22] Voir par exemple Gupta (1993).

[23] Même du point de vue de l'analyse tarskienne de la validité logique *étant donné un choix de constantes logiques*, quand bien même nous déclarerions que le prédicat de vérité est une constante logique dont l'interprétation ne doit pas être considérée comme variable, l'énoncé (TE) n'en serait pas pour autant déclaré logiquement valide, et l'on ne peut donc demander qu'il soit dérivable de règles logiques uniquement. Tarski parlait de « la loi du tiers-exclu », mais (TE) est une vérité de la syntaxe logique, non le schéma que l'on désigne parfois sous ce nom et dont toutes les instances sont des validités logiques. Si l'on s'en tient à la terminologie de Tarski, les « lois de la logique » ne sont pas les énoncés valides, même s'il s'agit deux choses pour lesquelles on peut parler de « vérités logiques ».

joue que le rôle d'un moyen pour décrire une classe d'énoncés que nous endossons à telles et telles conditions. D'après notre analyse de la situation au chapitre 6, notre justification pour (TE) ne relève pas seulement de la syntaxe de notre langage, ni de cette syntaxe *et* de ce que nous comprenons par « vrai » : la vérité et la syntaxe ne sont que le moyen d'exprimer quelque chose de neuf, ce principe (TE) qui explicite ce que doit justifier toute justification de notre pratique déductive. De ce point vue, l'impossibilité de *prouver* (TE) d'après les principes gouvernant notre compréhension du prédicat de vérité et la syntaxe descriptive d'un langage est un résultat attendu. Conversement, l'impossibilité de *prouver* (TE) dans la théorie minimale et la syntaxe descriptive d'un langage ne montre pas que la théorie minimale de la vérité ne dirait pas tout ce en quoi consiste notre compréhension du prédicat de vérité.[24]

Nous avons déjà vu au chapitre 2 que les équivalences-T ne définissaient pas le concept de vérité au sens où l'on parle ordinairement de « définition » d'un prédicat. Nous ne sommes nullement, avec une telle « définition », en situation de donner une formule ϕ telle que

$$\forall x(Vr(x) \leftrightarrow \phi(x))$$

c'est-à-dire permettant d'éliminer le prédicat de vérité en tout contexte extensionnel. Mais les questions qui nous intéressent maintenant sont : fixent-elles correctement la *signification inférentielle* du prédicat de vérité, et cette signification est-elle celle d'une expression *logique* ? Et cette question est bien ouverte.

Jusqu'ici nous avons présenté la théorie minimale comme un ensemble de biconditionnels, mais l'on peut la formuler sous forme d'une paire de règles d'introduction et d'élimination du prédicat de vérité, une représentation qui est plus conforme au cadre d'analyse que nous adoptons ici. L'hypothèse que nous faisons, soutenue par les considérations précédentes, est donc qu'un sujet possède une compréhension complète du concept de vérité si, et seulement si, il accepte inconditionnellement les deux règles d'inférence suivantes :

$$\text{Vr-intro}\ \frac{A}{Vr(a)} \quad \text{and} \quad \text{Vr-elim}\ \frac{Vr(a)}{A} \quad ,$$

[24] J'ajoute en passant que le fait qu'un énoncé comme (TE), qui exprime en un sens la loi du tiers-exclu, ne soit pas lui-même une vérité logique au sens d'énoncé logiquement valide (en particulier n'est pas lui-même une instance du schéma du tiers-exclu), ne signifie pas pour autant que le prédicat de vérité, quant à lui, comme constituant de cet énoncé, n'est pas une expression logique, même en acceptant la façon dont le choix des constantes logiques détermine l'ensemble des vérités logiques dans l'analyse classique de Tarski (Tarski (1936/2009)).

où la lettre « A » doit être remplacée par un énoncé du fragment du langage ne contenant pas « Vr », et où « a » un nom de cet énoncé.[25]

D'un point de vue logique, que l'on présente la théorie minimale sous forme de biconditionnels ou sous forme des règles revient au même, pour autant que l'on comprenne la signification du conditionnel de façon standard.[26]

En outre, nous n'attachons pas de priorité conceptuelle au formalisme de la déduction naturelle, et l'on pourrait adopter aussi bien le formalisme du calcul des séquents :[27]

$$\mathrm{Vr} \vdash \frac{\Gamma, A \vdash \Delta}{\Gamma, Vr(a) \vdash \Delta} \qquad \vdash \mathrm{Vr} \ \frac{\Gamma \vdash A, \Delta}{\Gamma \vdash Vr(a), \Delta}$$

Une dernière remarque préliminaire, enfin, sur ce que *n*'est *pas* le sens philoso-

[25] En outre, pour être clair sur ce point, on suppose que la règle est applicable dans les contextes hypothétiques : la règle d'introduction doit se lire comme affirmant que si l'on peut inférer A sous les hypothèses Γ, alors on peut inférer $Vr(a)$ sous les hypothèses Γ, et de façon analogue pour la règle d'élimination.

[26] Plus précisément, ces règles permettent de dériver les équivalences-T et, inversement, en présence des équivalences-T les règles sont admissibles, dès que la logique contient un conditionnel gouverné par des règles standard. Dans un sens : supposons que L contienne une « implication », c'est-à-dire un mot logique, notons-le $\rightarrow$, tel qu'un « théorème de la déduction » vaille :

$$\frac{A_1, \ldots, A_n \vdash B}{A_1, \ldots, A_{n-1} \vdash A_n \rightarrow B}$$

Alors les équivalences-T sont des conséquences des règles que nous nous sommes données :

$$\dfrac{\dfrac{\dfrac{A \vdash A}{A \vdash Vr(\ulcorner A \urcorner)}}{\vdash A \rightarrow Vr(\ulcorner A \urcorner)} \qquad \dfrac{\dfrac{A \vdash A}{Vr(\ulcorner A \urcorner) \vdash A}}{\vdash Vr(\ulcorner A \urcorner) \rightarrow A}}{\vdash A \leftrightarrow Vr(\ulcorner A \urcorner)}$$

Lorsque l'on travaille dans certaines logiques non-classiques, il devient parfois important de distinguer entre une formulation en termes de règles d'inférences et une formulation en termes de biconditionnels. Dans la « logique forte » de Kleene par exemple, qui est une logique trivalente, il n'y a pas de conditionnel satisfaisant le théorème de déduction.

[27] On a parfois objecté aux analyses inférentialistes conduites dans des calculs multi-conclusions de ne pas représenter des opérations de la pensée, de n'être qu'un formalisme *ad hoc*, ou de supposer implicitement la compréhension du « ou » classique. De nombreuses discussions critiques de ces objections existent dans la littérature, motivées en partie par le fait que la possibilité d'une justification inférentielle de la logique classique (et non seulement intuitionniste) dépend du caractère admissible du calcul multi-conclusion dans l'entreprise d'analyse inférentielle de la signification des constantes logiques. De nombreuses voies ont été explorées pour donner une interprétation de ce calcul qui ne tombe pas sous le coup des objections mentionnées à l'instant. Voir en particulier Smiley (1996), Restall (2005), Rumfitt (2000) et les références citées dans ces articles. Ces questions n'ayant pas d'incidence directe ici, je les laisse de côté.

phique de ce que nous sommes en train de faire. En affirmant que la signification inférentielle du prédicat de vérité est donnée par les règles précédentes, nous *ne* sommes *pas* en train d'adopter un thèse antiréaliste sur la vérité, nous *n*'identifions *pas* non plus la notion de vérité à la notion d'assertibilité ou de prouvabilité, ne *ne* soutenons *pas* davantage que toute vérité est connaissable etc. Il est facile de voir qu'avec les règles pour la vérité que nous nous sommes données et les règles de la logique classique il est possible de dériver toutes les instances du schéma

$$Vr(\ulcorner A \urcorner) \vee Vr(\ulcorner \neg A \urcorner),$$

ce qui est ordinairement compris comme une marque d'engagement envers une forme de réalisme métaphysique. Nous n'identifions pas non plus vérité et assertabilité : il n'y a aucune raison de penser, sur la seule base de la signification du mot « assertable », que le principe analogue du tiers-exclu pourrait valoir pour l'assertabilité. Pour autant que l'on puisse dire, certains des énoncés exprimant, pour un énoncé A, qu'il est assertable ou que sa négation est assertable (i.e. « Assertable(« A ») ou Assertable(« non A ») »), non seulement ne sont pas analytiques, c'est-à-dire dérivables des règles qui fixeraient la signification du mot « assertable »,[28] mais sont matériellement faux. D'un autre côté, la possibilité de dériver, grâce aux règles pour « Vr » que nous nous sommes données, les instances du schéma $Vr(A) \vee Vr(\neg A)$, dépend crucialement du fait que nous puissions utiliser aussi les règles d'inférence classiques, en particulier celles concernant « $\vee$ » et « $\neg$ », et en dernier lieu du fait que les instances du schéma $p \vee \neg p$ soient logiquement valides. En fait, affirmer simplement que la signification inférentielle du prédicat de vérité est donnée par les règles précédentes est totalement neutre relativement à la question métaphysique du réalisme.[29]

[28] On peut imaginer que les règles pour l'assertabilité sont typiquement les règles « de preuves » correspondant à l'introduction et l'élimination de l'opérateur de nécessité en logique modale :

$$\frac{A}{\text{Assertable(« A »)}} \qquad \frac{\text{Assertable(« A »)}}{A)}$$

où, contrairement aux règles gouvernant la notion de vérité, l'introduction n'est possible que si l'on dispose d'une preuve de la prémisse et non seulement d'une dérivation sous hypothèses. Le fait que l'on ne puisse pas introduire le prédicat d'assertabilité dans les contextes hypothétiques a pour conséquence que l'on ne peut pas dériver des règles les instances de

$$\text{Assertable(A)} \vee \text{ Assertable(non A)}$$

sauf si l'on a par ailleurs une preuve de A ou une preuve de la négation de A, bien entendu.

[29] Ce qui importe pour soutenir une forme de réalisme contre lequel Dummett et quelques

Analyse inférentielle de la vérité Ce que nous voulons montrer, c'est que les règles-Vr satisfont les conditions d'harmonie qui caractérisent les définitions inférentielles correctes, et que ces règles définissent une expression logique, au sens que l'analyse preuve-théorique donne à ce terme.

La question de l'harmonie des règles va sans difficulté majeure. Il n'est pas difficile en effet de voir que ces règles satisfont les deux critères d'harmonie globale et locale que nous avons présentés dans la section précédente. Commençons par le critère de conservativité. Le critère de conservativité (en dépit des doutes que nous avons émis à son égard) est satisfait, comme le montre le résultat de conservativité des extensions minimales des théories par les *équivalences-T* que nous avons déjà mentionné[30], et en vertu de la correspondance entre règles-Vr et équivalences-T expliquée plus haut. [31] Toutefois, ce que montre ce résultat est la conservativité sur le fond d'une théorie syntaxique donnée en avance, qui permette d'affirmer l'existence de certaines expressions, et peut-être les lois fondamentales de la formation de certaines classes d'expressions particulières. Ces règles ne sont pas conservatives sur la logique pure, puisque que l'on peut en inférer l'existence de plusieurs objets.[32] Le critère de conservativité est un critère relatif, et l'harmonie globale de règles données est relative aussi à un langage dans lequel les règles sont introduites. Ce que l'on peut dire est que les règles pour la vérité satisfont le critère d'harmonie globale dans les contextes contenant un minimum de ressources syntaxiques.[33]

inférentialistes se sont élevés, est donc de savoir 1. s'il existe une autre manière plausible d'expliquer la signification des expressions que par leur contribution à la signification inférentielle des énoncés et 2. si non, de savoir s'il est possible de donner une explication de la signification inférentielle de « ¬ » et « ∨ » qui permette de rendre compte de la validité du tiers-exclu. Sur ce dernier point, voir la note précédente. Par ailleurs, nous avons déjà noté au chapitre précédents que les clauses récursives tarskiennes pour la vérité sont également neutres relativement à ces questions.

[30] Voir chapitre 2.

[31] Je mentionne en passant que les règles-T satisfont également un critère d'unicité pour les définitions inférentielles présenté par Belnap dans l'article déjà cité. Si nous supposons que les règles pour la notion de déductibilité sont les règles structurales usuelles, on peut montrer que les règles pour Vr définissent une unique constante en introduisant un nouveau symbole Tr, supposé gouverné par les mêmes règles que celles que nous avons données pour Vr, et en raisonnant de la façon suivante :

$$\dfrac{\dfrac{A \vdash A}{A \vdash Vr(a)} \qquad \dfrac{A \vdash A}{Tr(a) \vdash A}}{Tr(a) \vdash Vr(a)}\ \text{Cut}$$

(On obtient l'autre sens en intervertissant toutes les occurrences de chacun des prédicats.)

[32] Nous avons déjà noté pourquoi au début du chapitre 3, section 1.2, note 27.

[33] En revanche les règles satisfont sans restriction le critère de conservativité tel que formulé par

Pour ce qui est maintenant du critère d'harmonie locale, considérons par exemple les règles-Vr en déduction naturelle et le critère d'inversion de Prawitz. Clairement ici le principe d'inversion est valable : étant donné une dérivation dans laquelle une application de la règle d'introduction de Vr est immédiatement suivie d'une application de sa règle d'élimination, nous pouvons trouver une dérivation qui ne fait pas ce détour. Le pas de réduction se fait simplement de la façon suivante :

$$\begin{array}{c} \Gamma \\ \Pi_1 \\ A \\ \text{Vr-intro}\ \dfrac{}{\mathrm{Vr(a)}} \\ \text{Vr-elim}\ \dfrac{}{A} \end{array} \quad \Rightarrow \quad \begin{array}{c} \Gamma \\ \Pi_1 \\ A \end{array}$$

Par conséquent les règles sont en harmonie et fixent correctement la signification de l'expression.[34] [35]

Reste donc la question de la logicité. Il y a un sens dans lequel les règles pour la vérité sont quasi-schématiques :

$$\text{Vr-intro}\ \frac{A}{Vr(a)} \quad \text{and} \quad \text{Vr-elim}\ \frac{Vr(a)}{A}$$

La prémisse de la règle d'introduction (et de même pour la conclusion de la règle d'élimination) est schématique, la spécification des conditions d'introduction du prédicat de vérité ne réfère à aucun contenu particulier. Ce point accrédite la thèse de la logicité du prédicat de vérité. Le problème, néanmoins, est que la règle n'introduit pas seulement le prédicat de vérité, mais également une *autre* expression, à savoir un nom de l'énoncé qui sert de prémisse à la règle. Par conséquent, d'une part, la règle ne donne pas la signification inférentielle du prédicat de vérité seulement (sa contribution propre au condition de prouvabilité de l'énoncé « $Vr(a)$ ») mais la signification inférentielle d'un énoncé complet, à savoir « $Vr(a)$ » lui-même, contenant non seulement le prédicat de vérité mais aussi un terme désignant un énoncé. Simultanément, la règle perd une propriété fondamentale attendue de toute

Belnap.

[34] De façon équivalente, en calcul des séquents, on voit que les coupures sur une occurrence de formule de la forme « Vr'A' » sont réductibles à des coupure sur « A » :

$$\dfrac{\dfrac{\Gamma', A \vdash \Delta'}{\Gamma', \mathrm{Vr\langle\langle\ A\ \rangle\rangle} \vdash \Delta'} \quad \dfrac{\Gamma \vdash A, \Delta}{\Gamma \vdash \mathrm{Vr\langle\langle\ A\ \rangle\rangle}, \Delta}}{\Gamma, \Gamma' \vdash \Delta, \Delta'} \quad \Rightarrow \quad \dfrac{\Gamma', A \vdash \Delta' \qquad \Gamma \vdash A, \Delta}{\Gamma, \Gamma' \vdash \Delta, \Delta'}$$

[35] Sur ce point, voir aussi l'appendice à ce chapitre.

règle logique, et que la schématicité doit permettre de garantir dans une situation non pathologique, à savoir la préservation de la validité de la règle par substitution uniforme d'énoncés à d'autres : si, dans une instance de la règle, je remplace uniformément un énoncé par un autre, je dois encore avoir une instance de la règle, c'est ce qui fait de la règle une règle purement formelle. Les règles pour « Vr » ne sont stables que pour une *double* substitution simultanée, d'un énoncé à un autre *et* d'un nom de cet énoncé au nom de l'autre. On ne peut pas contourner le fait que l'emploi du prédicat de vérité est conditionné à la possession de certaines ressources permettant de parler du langage, et que l'inférence de la vérité d'un énoncé désigné sous un certain nom à cet énoncé, et réciproquement, n'est possible que si l'on peut reconnaître l'énoncé sous le nom. On ne peut *inférer* d'un énoncé x le fait que « Vrai » s'applique à y que si l'on sait que y est un nom de l'énoncé x, et inversement je ne peux inférer de « La deuxième phrase des Poésies de Lautréamont est vraie » l'énoncé « les premiers principes doivent être hors de discussion » que si je sais que la deuxième phrase des Poésies de Lautréamont est l'énoncé en question. D'une manière générale, les ressources nécessaires pour pouvoir inférer selon les règles qui sont supposées fixer la signification de « vrai », loin d'être purement logiques, peuvent alors être considérables. La question est de savoir si ce fait est de nature à réfuter la thèse d'après laquelle la notion de vérité *elle-même* est logique. La réponse, en retour, va dépendre de notre capacité à *isoler* ces ressources conceptuelles auxiliaires de ce qui est proprement en jeu dans l'emploi du mot « vrai ».

Une première approche consisterait à se rabattre sur une affirmation beaucoup plus modeste. C'est la stratégie de Hodes (2004) (parmi les rares auteurs à avoir proposé une analyse d'inspiration inférentialiste du prédicat de vérité). Hodes maintient que la signification de « Vr » est fixée par les « règles générales » que nous avons données pour commencer, dont les instances sont du type

$$\frac{\text{La neige est blanche}}{\text{a est vrai}} \qquad \text{où « a » est un nom } \textit{quelconque}$$

$$\frac{\text{a est vrai}}{\text{La neige est blanche}} \qquad \text{de l'énoncé « La neige est blanche ».}$$

Il remarque alors qu'il est nécessaire d'avoir une certaine connaissance « sémantique », celle de la dénotation de « a », pour pouvoir éliminer « vrai ».[36] Il en conclut

[36] Et de même pour la règle d'introduction. Le cas de la règle d'élimination est simplement plus parlant, reflétant une difficulté typique à laquelle peut se trouver confronté un locuteur apprenant qu'un certain énoncé est vrai.

que le sens de « Vr » n'est pas constitué par des règles purement « syntaxiques »[37], et cette étrangeté l'amène à qualifier le prédicat de vérité de constante « semi-logique » unique en son genre...

Mais il doit y avoir une erreur ici. En fait, il est évident que l'on peut parfaitement comprendre la signification du prédicat de vérité sans être capable d'éliminer le prédicat de vérité en tout contexte d'usage. Que les règles générales pour le prédicat de vérité que nous avons données soient valides au sens où, si la prémisse est vrai, alors la conclusion l'est aussi[38], ne signifie pas que ce sont ces règles qui *déterminent* la signification inférentielle de « Vr », qu'il faille *reconnaître* la validité de chaque instance pour comprendre complètement la signification inférentielle du prédicat de vérité. En renonçant à isoler la contribution propre de notre compréhension du prédicat de vérité dans notre reconnaissance, quand elle est possible, de la validité des inférences du type général

$$\frac{\text{A}}{\text{a est vrai}} \qquad \frac{\text{a est vrai}}{\text{A}}\ ,$$

où « A » doit être remplacé par un énoncé et « a » par un nom *quelconque* de cet énoncé

on s'interdit une analyse de la nature de la notion de vérité. La catégorie des expressions « semi-logiques » ne fait alors que masquer cette absence.

Mais l'on peut faire mieux, en révisant à la baisse les règles qui sont supposées fixer la signification inférentielle du prédicat de vérité et en cherchant à « factoriser » la validité des règles « générales » en deux processus bien distincts : la validité de règles plus modestes fixant la signification inférentielle de « Vr » et la validité d'autres règles où la notion de vérité n'apparaît pas. En effet, nous avons déjà vu qu'il existe pour tout langage une classe de noms élémentaires de ses énoncés : ceux que l'on obtient par une description structurelle[39] et dont la théorie est couchée dans la « syntaxe » dont les grandes lignes ont été présentées au chapitre 2. À l'aide de quelques noms primitifs d'expressions du langage, d'un opérateur de concaténation des expressions[40], et de quelques lois gouvernant leur emploi, il est possible de donner une description structurelle de tous les énoncés. On peut soutenir que pour cette classe de noms, l'inférence de l'énoncé « A » à l'énoncé « Vr($\ulcorner A \urcorner$) », où « $\ulcorner A \urcorner$ » est un nom de ce genre, est, sinon une inférence logique, du moins une inférence

[37] J'adopte ici la façon de parler de Hodes.

[38] Les règles de Hodes ne sont essentiellement qu'une formulation, sous forme de règles, du *schéma-T* tel que le formulait Tarski dans le cas où langage-objet est inclu dans le métalangage.

[39] Voir la présentation de la syntaxe au chapitre 2

[40] Que j'ai noté « * » au chapitre 2 et au début de ce chapitre.

elle-même élémentaire, disons logico-syntaxique. En partant de ce point solide, on pose que ce sont ces règles particulières d'introduction et d'élimination du prédicat de vérité, employant ces noms élémentaires d'énoncés, qui fixent la signification inférentielle de « Vr » dans un contexte contenant la théorie élémentaire de ces noms, et donc une méthode simple permettant de construire le nom de tout énoncé à partir d'un nombre fini de symboles. Les nouvelles règles supposées fixer la signification inférentielle de « Vrai » peuvent donc maintenant être écrites de cette façon :

$$\frac{\mathrm{A}}{\mathrm{Vr}(\ulcorner A \urcorner)} \qquad \frac{\mathrm{Vr}(\ulcorner A \urcorner)}{\mathrm{A}}$$

Cette révision[41] à la baisse de nos règles n'est pas un problème, elle oblige simplement à considérer les autres cas d'introduction et d'élimination sur des noms ou des descriptions plus complexes pour ce qu'intuitivement ils sont, à savoir des emplois dérivés qui ne sont pas valides *seulement* en vertu de la signification inférentielle du prédicat de vérité. En traitant dans cet esprit l'exemple des Poésies de Lautréamont, l'inférence éliminative est possible, et n'est possible que, lorsque je suis en mesure d'identifier l'objet de la description sous un de ses noms élémentaires. En traitant la description « La deuxième phrase des Poésies de Lautréamont » comme un nom, disons « p2 » :

$$\text{Vr-élim}\ \frac{\text{Subst. Id.}\ \dfrac{Vr(p2) \qquad p2 = \ulcorner \text{Les premiers principes doivent être hors de discussion} \urcorner}{Vr(\ulcorner\ \text{Les premiers principes doivent être hors de discussion} \urcorner)}}{\text{Les premiers principes doivent être hors de discussion}}$$

Cette décomposition de ce qui est en jeu dans l'inférence de « $Vr(p2)$ » à « Les premiers principes doivent être hors de discussion » est naturelle et explique rétrospectivement le caractère contre-intiutif de la thèse selon laquelle ce serait les « règles générales »d'introduction et d'élimination qui expliqueraient la signification inférentielle de prédicat de vérité : des règles bien plus faibles y suffisent. Ces

[41] Réécrite plus complètement pour tenir clairement le compte des ressources en jeu (en reprenant les notations de base de la syntaxe donnée au chapitre 2), notre description de la règle d'introduction (par exemple) dit :

$$\text{Pour tout énoncé } x \text{ du langage [auquel s'applique le prédicat de vérité, disons } L\text{]},\quad x \vdash \overline{Vr} * \overline{(} * cod(x) * \overline{)}$$

où « $\vdash$ » est un prédicat binaire que nous utilisons pour décrire les règles d'inférence et où « cod » est une fonction qui à tout énoncé x du langage L associe le nom qui est canoniquement associé à x par la méthode de description des énoncés qui est supposée faire partie des ressources du langage que l'on décrit (celui, global, contenant également le prédicat de vérité, disons ML, et pour lequel la règle en question est une règle d'inférence.

considérations permettent de corriger nettement le tableau pessimiste dressé par Hodes : on peut conclure maintenant que le prédicat de vérité, s'il n'est pas logique, est du moins logico-syntaxique, au sens où son emploi suppose tout de même quelques ressources conceptuelles couchées dans une théorie capable de décrire la syntaxe élémentaire du langage.[42]

Mais ce processus peut encore être mené un pas plus loin si, pour expliquer notre usage inférentiel du prédicat de vérité, il n'est pas besoin de supposer que nous disposions véritablement d'une théorie syntaxique, avec ses notions d'expressions et ses lois de concaténation. Il y a, me semble-t-il, un sens dans lequel chaque énoncé du langage possède un nom canonique véritablement transparent : quelque chose comme ce que l'on obtient par la mise entre guillemets de cet énoncé. Bien entendu, une analyse possible de ce qu'est, en fait, la mise entre guillemets, nous ramènerait tout droit aux descriptions structurelles : l'énoncé entre guillemets est simplement un nom de l'énoncé obtenu en utilisant les symboles comme noms d'eux-mêmes et la juxtaposition comme opérateur de concaténation. Mais cette reconstruction rationnelle ne rend pas compte du caractère d'*immédiateté* du processus par lequel je suis capable de former des noms d'énoncés de mon langage par la mise entre guillemets, et de reconnaître un énoncé sous un nom de ce genre. Si ce point est correct, n'est-il pas naturel de vouloir considérer ces noms d'énoncés comme donnés d'emblée avec le langage lui-même, comme des primitifs que je suis capable d'invoquer à volonté, à charge peut-être pour une théorie auxiliaire de rendre compte des relations entre les nom de ce type et les descriptions structurelles ? Peut-être, peut-être pas. Dire oui, c'est dire que la théorie qui permet de rationnaliser la constitution d'un objet doit être comptée parmi les ressources permettant de rendre compte de la possibilité de l'usage de cet objet. Dire non, c'est semble-t-il faire droit à l'idée que l'inférence de la vérité d'un énoncé que je me *présente* comme objet de pensée, à l'affirmation de cet énoncé lui-même, ne mobilise aucune véritable syntaxe, c'est faire comme si l'intimité de notre commerce avec nos énoncés était telle que nous aurions un *nom propre* de chacun de ceux que nous utilisons.[43]

[42]Nous avons vu au chapitre 2 que la vérité pour un langage du premier ordre peut être définie explicitement dans la syntaxe du deuxième ordre (dans un langage contenant aussi le vocabulaire de la théorie-objet. On aurait pu en conclure que la notion de vérité pour un langage du premier ordre est une notion logico-syntaxique. Mais c'est alors en un sens très étendu, dans lequel on admet la logique du second-ordre comme faisant partie de la logique. Le sens en jeu ici est bien plus modeste.

[43] Les énoncés sont en nombres infinis, bien entendu, et c'est pourquoi il faut une théorie « récursive » pour les engendrer, et donc quelque chose comme de l'arithmétique. D'un certain

Si, donc, l'on se donne un nom simple de chaque énoncé, représenté par exemple par l'énoncé mis entre guillemets, alors il est naturel de réduire encore nos emplois définitionnels des règles d'introduction et d'élimination du prédicat de vérité et de poser que sa signification inférentielle est donnée simplement par ces règles « minimales » :

Définition 16 (Définition inférentielle minimale du prédicat de vérité)**.**

$$\frac{A}{\mathit{Vrai}(\text{« } A \text{ »})} \qquad \frac{\mathit{Vrai}(\text{« } A \text{ »})}{A}$$

Autrement dit, plus explicitement, dans le langage de la syntaxe dans lequel *nous* décrivons ces règles :

Pour tout ensemble fini σ d'énoncés du langage,
pour tout énoncé x de ce langage ne contenant pas le prédicat de vérité,
si $\sigma \vdash x$ alors $\sigma \vdash \overline{Vr} * \overline{(} * \overline{\text{«}} * x * \overline{\text{»}} * \overline{)}$

où « $\overline{\text{«}}$ » et « $\overline{\text{»}}$ » sont des noms des guillemets ouvrants et fermants respectivement (on formulerait de façon analogue la règle d'élimination). Avec ces nouvelles règles, la « preuve » précédente de ce que les premiers principes doivent être hors de discussion doit être revue, puisque la règle d'élimination du prédicat de vérité opérait sur un nom complexe d'énoncé et non sur un de ces noms simples distingués que nous nous sommes donnés. Pour pouvoir montrer que l'inférence des prémisses à la conclusion conclusion est valide, il faut maintenant en outre être capable d'identifier ce nom composé à l'un de ces noms élémentaires qui nous sont par hypothèse donnés d'emblée, i.e. il faut dans la preuve précédente ajouter une étape avant l'application de la règle Vr-élim, à l'effet que

$\ulcorner$ Les premiers principes doivent être hors de discussion$\urcorner$ =« Les premiers principes doivent être hors de discussion »

Autrement dit, de manière générale, la démonstration de la validité de l'inférence de la vérité d'un énoncé à cet énoncé lui-même est conditionnée maintenant à notre capacité à identifier le nom qui est l'argument du prédicat de vérité à l'un de ces noms canoniques que nous nous sommes donnés au départ et qui présentent de façon transparente l'énoncé qui va être utilisé.

point de vue ce n'est pas grand chose, mais ce n'est pas rien. D'un autre côté, nous n'utilisons jamais qu'un nombre fini d'énoncés.

Si cette idée est tenable, les seules ressources dont nous avons besoin de supposer qu'elles sont nécessaires la compréhension du prédicat de vérité sont les axiomes habituels de l'égalité, du type $\vdash t = t$ et principe de substitution des identiques, et les axiomes particulièrement élémentaires suivants :

Si A, B sont des énoncés de L et que $A \neq B$:
$\vdash$« A »$\neq$« B »

Cette théorie n'est pas à strictement parler une théorie purement logique. Dans cette théorie on peut prouver, pour tout entier n, qu'il existe au moins n objets. On admet en général que la logique pure doit être ontologiquement neutre et, si c'est vrai, l'existence de n objets n'est pas une vérité logique. Si l'idéologie de la règle est faible, elle vient tout de même avec certains engagements ontologiques. Mais la théorie des objets nécessaire pour rende compte des inférences aléthiques est *particulièrement* faible.

Il semble difficile de réduire davantage les ressources théoriques auxiliaires nécessaires pour rendre compte de la signification inférentielle du prédicat de vérité. Ma conclusion provisoire est donc la suivante : la signification inférentielle du prédicat de vérité est expliquée par les règles minimales d'introduction et d'élimination sur des noms canoniques élémentaires, ou sur des descriptions structurelles élémentaires. La théorie dont nous avons minimalement besoin pour rendre compte des noms d'énoncés qui sont en jeu dans les règles d'introduction et d'élimination du prédicat de vérité, quelle qu'elle soit, est une théorie faible.

Mais maintenant il semble que nous ayons atteint une limite fondamentale de l'approche inférentialiste dans la perspective qui est la nôtre. Car d'un côté il ne semble pas *a priori* absurde de tenir que la vérité soit une notion logique sans pour autant tenir que les instances de

$$Vr(\ulcorner A \urcorner) \rightarrow A$$

ou même de

$$Vr(\text{« } A \text{ »}) \rightarrow A$$

seraient des vérités logiques à strictement parler. Mais d'un autre côté, l'approche inférentielle, par nature, est une analyse contextuelle, où le sens d'une expression est analysé dans ses contextes d'usages comme sa contribution au rôle inférentiel des énoncés dans lesquels elle figure. Or, tous les emplois du prédicat de vérité que

l'on pourra analyser dans ce cadre seront des cas d'application à des termes, qui ne sont pas, eux, des expressions logiques. Le problème est, en partie, celui de la catégorie grammaticale du prédicat de vérité,[44] en partie celui de la nature des termes auxquels il s'applique. Par conséquent, au vu de l'analyse précédente, *le prédicat de vérité est aussi logique qu'un prédicat s'appliquant à des énoncés peut l'être d'après les lumières de l'analyse inférentielle de sa signification*, c'est-à-dire de sa contribution au rôle inférentiel des énoncés dans lesquels il figure.

7.2.3 Conclusion sur l'approche inférentielle

Résumons rapidement ce que nous venons de voir. Tout d'abord j'ai commencé par rappeler les fondements de l'analyse inférentialiste, ou preuve-théorique, de la signification et de la validité. Ce premier mouvement m'a permis d'esquisser ce qu'une analyse non véricondition nelle de la signification pourrait être, et la façon dont elle pourrait être mise en œuvre pour la définition d'un critère de validité. Puis nous avons appliqué les outils présentés à une analyse inférentielle du prédicat de vérité lui-même. Nous avons montré que les règles pour la vérité étaient en harmonie, mais également que des règles très faibles devaient suffire à comprendre sa signification inférentielle. Ces règles possèdent un caractère quasi-schématique, qui détermine son emploi comme un emploi logico-syntaxique. Nous avons remarqué que c'était le mieux que l'on pouvait espérer d'une analyse sémantique contextuelle en termes d'usages du prédicat de vérité, et que cette analyse ne pouvait donc pas totalement rendre compte de l'idée que la vérité pourrait être une notion logique quand bien même l'énoncé « si A alors Vr(« A ») » ne serait pas lui-même logique. Dans la section suivante, je reviens sur cette l'idée de la logicité du prédicat de vérité par une autre voie, celle de l'approche sémantique, en essayant de voir si cette façon d'aborder la question ne permet pas de surmonter ce problème.

[44] L'analyse inférentielle de l'identité pose également des diffcultés, bien que l'identité soit communément comprise comme une notion logique. Pour un argument en faveur de la logicité de « = », voir par exemple Tarski (1966/1986).

7.3 Une approche sémantique

7.3.1 La logique : un sujet sans objet ?

L'approche sémantique du problème de la nature de la logique cherche à caractériser le *domaine* de la logique, ce dont la logique parle, ce à propos de quoi sont notions logiques. Il y a historiquement deux intuitions concurrentes majeures ici. Selon la première, ce qui distingue la logique est sa *généralité*, le fait que ce qu'elle nous dit s'applique à tous les domaines de la réalité. Selon la seconde, ce qui distingue la logique est son *absence d'objet*, ou sa formalité, le fait que, en un certain sens, elle ne nous dise rien du monde. Ces deux intuitions ne sont pas, d'ailleurs, inconciliables. Elles cohabitent chez Kant par exemple.[45] Il y a dans la *Critique de la raison pure*, en particulier dans l'Introduction à la logique transcendantale, des formulations explicites de ces deux traits de la logique :

> Une logique *générale* mais *pure* ne s'occupe donc que des principes *a priori* ; elle est un *canon de l'entendement* et de la raison, mais seulement par rapport à ce qu'il y a de formel dans leur usage, quel qu'en soit d'ailleurs le contenu (qu'il soit emirique ou transcendental).(Kant Critique de la Raison Pure 76/49. Traduction française *in* Kant (2006))

Dans cette science « courte et aride », le logicien « doit toujours avoir en vue » la règle suivante :

> Comme logique générale, elle fait abstraction de tout le contenu de la connaissance de l'entendement, et de la diversité de ses objets, et elle n'a à s'occuper que de la simple forme de la pensée.(*idem*)

On mentionnera aussi ce passage de la *Logique*, où Kant écrit :

> Mais si nous mettons de côté toute connaissance que nous devons emprunter aux seuls *objets*, et si nous réfléchissons seulement à l'usage de l'entendement en général, nous découvrons ces règles qui sont absolument nécessaires à tous égards et sans considération des objets particuliers de la pensée, puisque sans elle nous ne pourrions pas penser du

[45] John MacFarlane a montré de façon convaincante que c'est Kant qui a, le premier, *caractérisé* la logique par sa généralité et sa formalité. Cette thèse est défendue dans MacFarlane (2000), en particulier chapitre 4. Nous renvoyons à cet ouvrage pour une défense détaillée de ce point, laquelle demanderait un examen de la façon dont la spécificité de la logique a été conçue d'Aristote à Leibniz en passant par Port-Royal.

> tout.[...] Et de là vient que les règles universelles et nécessaires de la pensée en général ne peuvent concerner que sa seule *forme* et aucunement sa *matière*. Par conséquent la science qui contient ces règles universelles et nécessaires est simplement une science de la forme de notre connaissance intellectuelle ou de la pensée. (*Logique*, 12-13. Traduction française par M. Guillermit *in* KANT (2000).)

Les deux aspects, généralité et formalité, sont liés dans la philosophie kantienne. Dans l'ordre de l'analyse kantienne, la formalité de la logique suit de la généralité, *via* de la notion kantienne de contenu : parce que les jugements de la logique pure et générale ne sont pas informés par l'intuition ou les formes *a priori* de la sensibilité, ils n'ont pas de contenu. Mais c'est un trait particulier de l'épistémologie kantienne que la notion de généralité y implique que la logique soit également formelle. Les deux notions sont, en droit, indépendantes. Frege, par exemple, qui rejette l'idée kantienne selon laquelle les contenus de la pensée ne peuvent être donnés que par l'intuition sensible, tient à son tour que la logique est générale mais rejette, en 1884, l'idée qu'elle serait formelle *stricto sensu* : oui, la logique étudie les lois de la pensée comme telle[46] mais, contre la formalité, la raison a ses objets propres, les entiers.[47]

Inversement, pour Carnap et les positivistes, la logique est formelle mais n'est pas une science générale. Carnap maintient une distinction stricte entre sciences formelles et sciences du réel, et la logique, comme les mathématiques, appartiennent aux premières, celles qui n'ont pas de contenu « réel ». Etant le produit de décisions sans autre justification que pragmatiques, leurs lois n'ont rien de nécessaires et ne valent pas norme pour toute pensée possible : en un sens on peut voir le principe de

[46] Au tout début de l'*Idéographie*, Frege écrit ainsi :

> La démonstration la plus solide est manifestement celle qui est purement logique, qui, abstraction faite de la caractérisation particulière des choses, se fonde seulement sur les lois sur lesquelles toute connaissance repose. (FREGE (2000), préface, p.5)

Sur ce point Frege n'a pas varié, et l'on trouverait des passages plus tardifs d'une teneur semblable.

[47] Ce qui ne contredit pas le passage cité dans la note précédente : c'est au terme d'un processus de réflexion faisant abstraction des lois particulières de l'esprit en tant qu'il s'applique à des contenus particuliers, et en ne se reposant que sur le raisonnement le plus « solide », le raisonnement logique, que Frege découvre les entiers comme objets de la raison pure. Ainsi dans la conclusion des *Fondements de l'arithmétique* :

> On pourrait, en modifiant une proposition connue, dire que l'objet propre de la raison est la raison. L'arithmétique traite d'objets dont nous ne prenons pas connaissance comme d'un élément étranger, apporté de l'extérieur par la médiation des sens ; ces objets sont donnés immédiatement par la raison, et elle peut les pénétrer totalement, comme ce qui lui est le plus propre. (FREGE (1884/1969), tr. fr. p.225)

tolérance comme la négation de l'idée de généralité absolue des lois de la logique.

Certaines formulations de Russell laissent penser que, lui aussi, accepte le caractère général de la logique tout en refusant la formalité. On lit naturellement dans ce sens la fameuse remarque :

> [...] la logique est concernée par le monde réel, quoique sous ses traits les plus abstraits et les plus généraux, tout autant que la zoologie (Russell (1919), tr. fr. p.202)

Par d'autres côtés, pourtant, la formalité de la logique semble concommittante de la distinction opérée par l'atomisme logique entre forme et contenu des propositions. Ainsi chez le Russell de *Theory of Knowledge*, les deux idées de généralité et formalité coexistent-elles clairement :

> Toute notion logique, en un sens très important, est ou implique un *summum genus*, et résulte d'un processus de généralisation qui a été conduit à son extrême limite. C'est une particularité de la logique, et une pierre de touche par laquelle les propositions logiques peuvent être distinguées de toutes les autres. Une proposition qui mentionne une entité définie quelconque, qu'elle soit particulière ou universelle, n'est pas logique : aucune entité définie, d'aucune sorte, n'est jamais un constituant d'aucune proposition vraiment logique. Les « constantes logiques », qui peuvent sembler être des entités ayant une occurrence dans les propositions logiques, ne sont vraiment concernées que par la pure *forme*, et ne sont pas en fait des constituants des propositions dans l'expression verbale desquelles leur nom apparaît. (Russell (1913/1992) chap.9, p. 97. Notre traduction.)

Des passages comme le suivant, tiré des *Principles of Mathematics*, vont également dans le sens de la formalité de la logique :

> Admettons - je pense que nous pouvons le faire - que les formes des propositions puissent être représentées par la forme des énoncés dans lesquels elles sont exprimées, sans mot spécial destiné à représenter la forme : nous parviendrons alors à un langage où tout le formel appartiendrait à la syntaxe, et non au vocabulaire. Dans un tel langage, nous pourrions exprimer toutes les propositions des mathématiques même sans connaître un seul mot du langage. Le langage de la logique mathématique,

> poussé à son point de perfection, serait un langage de ce genre (Russell (1903), trad. fr. p.200-2001)

On se souviendra également de la formulation lapidaire de Wittgenstein :

> 5.4611 Die logischen Operationszeichen sind Interpunktionen.[48] (Wittgenstein (1922/1993))

Il y a donc au moins deux intuitions sémantiques distinctes et historiquement attestées du propre de la logique qui sont en droit indépendantes. Ces intuitions participent peut-être d'une matrice générale commune, mais alors il est difficile de la dégager tant l'expression fondamentale de cette intuition semble devoir mettre en jeu des thèses métaphysiques, sémantiques et même épistémologiques relatives à ce qui doit compter comme le « contenu » du discours. Confronté à cette variété, il faut faire un choix, et, parce qu'il me paraît bien adapté pour aborder la question de la logicité de la notion de vérité, je ferai celui d'adopter la façon de voir qui sous-tend les thèses de l'atomisme logique et en vertu de laquelle, en un certain sens du mot « contenu », les expressions logiques sont les expressions qui sont dénuées de contenu propre. Je ne me prononce pas, avec Carnap[49], pour un relativisme et contre une conception « absolutiste » des lois de la logiques, et je ne dis pas non plus qu'une analyse en termes de généralité ne pourrait pas être féconde,[50] mais j'adopterai seulement cette idée que les mots logiques, en un sens *ne parlent de rien.*

Dire que les expressions logiques[51] ne sont à propos de rien, toutefois, ce n'est pas dire que les expressions logiques ne sont pas essentielles comme moyens d'expression à tout discours sur le monde. Il y a des choses que je peux dire du monde grâce aux expressions logiques que je ne peux pas dire sans elles. Si je dis

> « Fido est un chien farouche *ou* Madeleine ne sait pas s'y prendre avec les animaux »

je dis bien quelque chose du monde qu'il ne me serait pas possible d'exprimer sans aucune expression logique, uniquement à l'aide d'énoncés élémentaires (« atomiques ») formulés dans le vocabulaire des concepts et des noms qui figurent dans

[48] *Les signes d'opérations logiques sont des marques de ponctuations*

[49] Carnap dont je parlerai peu, mais dont l'influence est bien là dans la suite.

[50] C'est par exemple l'approche de Tarski (1966/1986), et c'est en fait l'approche que j'utilise dans l'appendice 1.

[51] Et je parlerai aussi indifféremment de « notions » logiques, puisque c'est d'expression inteprétées qu'il s'agit ici.

l'énoncé ci-dessus. Pour en rester à des termes très généraux pour le moment, les expressions logiques sont des expressions qui n'ont pas de contenu propre mais qui permettent d'*articuler* des contenus.

7.3.2 Le contenu de « vrai »

Il y a maintenant un rapport assez clair entre cette façon de comprendre le propre des notions logiques et ce que les déflationnistes regardent comme l'emploi essentiel du mot « vrai ». Revenons en effet à l'intuition de la transparence de la vérité[52] selon laquelle, lorsque nous attribuons la vérité à un énoncé interprété, ce n'est pas vraiment de cet énoncé que nous voulons parler, mais *de ce dont parle cet énoncé*. Souvenons-nous par exemple de la remarque de Quine selon laquelle en attribuant la vérité à l'énoncé « la neige est blanche », nous ne faisons en substance qu'attribuer la blancheur à la neige :

> Le prédicat de vérité a son utilité là où, quoique nous soyons toujours concernés par la réalité, nous devons du fait d'une certaine complication technique mentionner des énoncés. Ici le prédicat de vérité sert, pour ainsi dire, à pointer vers la réalité à travers l'énoncé ; il sert de rappel au fait que quoique les énoncés soient mentionnés, la réalité est toujours tout le point. (Quine (1970), p.11)[53]

L'intuition qui guide ces remarques sur le sujet réel des affirmations contenant le prédicat de vérité se vérifie dans ses usages ordinaires d'où il ressort que contrairement aux termes « électrons » ou « taux d'inflation », disons, « vrai » ne pointe vers aucun sujet propre, n'est pas mis à contribution pour parler d'une propriété qui épouserait la nature « selon ses articulations naturelles ».[54] Comparons les deux affirmations suivantes, par exemple, supposées faites dans le registre de la communication scientifique ordinaire :

1. Si certains électrons entrent dans l'état énergétique α ce mois-ci alors nous entrerons dans une période de récession l'an prochain.
2. Si ce que Schumpeter a dit des cycles économiques est vrai alors nous entrerons dans une période de récession l'an prochain.

[52] Voir chapitre 1
[53] Sur le sens cette « complication technique », voir le chapitre 1
[54] Platon, *Phèdre*, 265d.

D'un côté, l'affirmation (1) se lit naturellement comme indiquant que l'état énergétique de certains électrons est causalement ou nomologiquement pertinent pour la situation économique de l'année prochaine. Mais, de l'autre côté, nous ne lisons certainement pas (2) comme impliquant que la vérité a un rôle causal pertinent en économie. L'exposé des causes putatives du conséquent de (2) est présumablement à chercher dans la théorie de Schumpeter elle-même. « Vrai » est neutre parce que la possibilité de la montée sémantique est ouverte dans tous les domaine du discours, y compris dans le registre du discours scientifique, et sans que cette montée sémantique implique aucunement un *changement de sujet* : du discours nous ne changeons que le *mode*. Mais s'il y a un sens raisonnable dans lequel, dans le passage du mode matériel au mode formel du discours, nous ne changeons pas de sujet, alors il doit y avoir un sens raisonnable dans lequel « vrai » n'*est à propos de rien*.

Une autre façon de faire le même point est de considérer la différence des types d'*engagements* épistémiques que prend un locuteur affirmant :

1. La théorie économique de Schumpeter est vraie.
2. La théorie économique de Schumpeter est belle.

Ce que nous appelons les engagements épistémiques, ce sont simplement les engagements à produire une justification de ce qui est affirmé que contracte le locuteur affirmant une proposition. Lorsque nous disons que la théorie économique de Schumpeter est belle, nous portons un jugement esthétique sur une théorie. À supposé qu'il ne s'agisse pas simplement d'une « façon de parler »[55], ce sont donc des arguments esthétiques qu'il nous faudra donner pour défendre notre affirmation, des arguments dont l'évaluation appartient à la science du Beau. Mais lorsque nous disons que la théorie de Schumpeter est vraie, il en va tout différemment. Il n'y a pas de science du Vrai, il n'y a que la science tout court : en dernière analyse, pour défendre notre affirmation, ce sont des arguments *économiques*, qu'il nous faudra trouver.[56]

[55] MacAllister (1996), par exemple, soutient que des critères esthétiques ont un rôle décisif dans les choix théoriques opérés en période de science révolutionnaire.

[56] Au niveau individuel de l'enquête scientifique, des arguments que l'on pourrait qualifier de *déférentiels* sont possibles et courants : on peut être justifié à croire qu'une théorie ou une hypothèse est vraie sur la foi du témoignage d'une personne que l'on estime compétente. Mais dans la mise en commun des efforts collectifs, et dans le corps scientifique total qui en résulte, toutes les hypothèses « déférentielles » doivent pouvoir être déchargées, pour ne laisser place qu'à des hypothèses scientifiques ordinaires. Je note en passant que, vu l'importance du prédicat de vérité dans les justifications déférentielles (voir les quelques remarques à ce sujet au chapitre 6), le contraire mettrait le déflationniste dans une situation fâcheuse ; mais avec lui également le statut des explications scientifiques.

Les considérations de ce type nous semblent renforcer la thèse que le concept de vérité, malgré sa grammaire de surface, est en général utilisé de façon « neutre », « universelle » ou « atopique ».

Bien entendu, il y a bien un autre sens dans lequel, lorsque nous affirmons « tous les axiomes de la géométrie d'Euclide sont vrais », nous ne parlons pas des propriétés du plan mais disons aussi quelque chose à propos des axiomes d'Euclide, peut-être des objets bien particuliers et localisés comme ces phrases qu'Euclide a qualifié d'axiomes dans son traité. Selon que l'on comprend cette phrase dans l'un ou l'autre sens, c'est tantôt une phrase d'abord à propos d'autres phrases, tantôt une phrase essentiellement à propos des propriétés de l'espace. Afin de faciliter la discussion, introduisons, dans une terminologie aussi neutre que possible, une distinction entre « contenu$_A$ » et « contenu$_B$ »[57] d'une attribution de vérité : le contenu$_A$ est celui qui correspond à la grammaire de surface (c'est l'attribution d'une certaine propriété, la vérité, à un certain contenu propositionnel, disons, celui dont la vérité est prédiqué), tandis que le contenu$_B$ est l'état du monde indirectement décrit par l'attribution de vérité.[58] Ainsi le contenu$_B$ de la proposition qu'il est vrai que la neige est blanche est simplement que la neige est blanche.

Maintenant, à nouveau, ce que nous voulons montrer est que l'emploi du prédicat de vérité *tel qu'il est compris par le déflationniste* est un emploi logique. Mais si une attribution de vérité à un énoncé ou à un contenu propositionnel est essentiellement un moyen *indirect* d'affirmer quelque chose *de ce à propos de quoi* est ce contenu propositionnel ou cet énoncé, alors c'est seulement en tant que moyen d'exprimer indirectement ces états du monde qu'il nous faut l'analyser, non comme moyen de parler des énoncés. Autrement dit, ce qu'il nous faut examiner, c'est ce que les contenu$_B$ des énoncés attribuant la vérité à d'autres énoncés nous permettent de dire du monde. Si nous parvenons à faire cela, nous pourrons alors examiner ce qu'ils nous permettent de dire et qu'il ne serait pas possible de dire dans le langage « objet », sans prédicat de vérité. Si le déflationniste est cohérent, le contenu$_B$ d'énoncés contenant des attributions de vérité à d'autres énoncés doit permettre de dire des choses qu'il n'est pas possible dire uniquement avec des énoncés du langage-objet. Simultanément, si l'emploi cet emploi du prédicat de vérité l'apparente à une expression logique, ce surplus d'expressivité ne doit pas signifier que le prédicat de vérité posséderait un « contenu » propre, qu'il permettrait de décrire une propriété

[57] Parler de « contenu matériel » et de « contenu formel » me paraît pouvoir prêter à confusion.
[58] C'est-à-dire : qui doit être realisé si l'attribution est correcte.

« réelle » du monde qu'il ne serait pas possible de décrire dans lui, mais qu'il sert bien uniquement à *articuler des contenus*. On voit que, vu sous cet angle, la thèse de la logicité de la vérité (telle que cette notion est théorisé par le déflationniste) devient plus plausible qu'il ne pouvait paraître de prime abord. Je voudrais maintenant essayer de lui donner un contenu précis. Je commence, dans la section qui suit, par proposer une sémantique permettant de rendre compte de l'idée de contenu$_B$ d'une attribution de vérité avant, dans la suivante, de préciser mon analyse de la logicité et de conclure.

7.3.3 Un cadre déflationniste pour la vérité

Nous avons besoin d'un cadre dans lequel analyser sémantiquement ce qui est exprimé par l'emploi déflationniste d'un prédicat de vérité. La difficulté est que, d'un côté, l'emploi du prédicat de vérité dans un énoncé nécessite l'usage d'un appareil conceptuel auxiliaire permettant de parler du langage, et dont le contenu ne fait pas partie du contenu$_B$ exprimé par l'énoncé en question. Nous avons donc besoin d'outils pour parler d'énoncés, ou d'ensembles d'énoncés, auquel le prédicat de vérité puisse s'appliquer afin de servir son rôle expressif, tout en essayant de bien séparer ce qui, dans une attribution de vérité, relève de ce qui est exprimé au sens $_A$ et au sens $_B$. Pour satisfaire ce cahier des charges, nous allons introduire une sémantique spéciale « à étage »[59], et considérer une forme de *langage sorté* contenant deux sortes d'expressions, en connexion avec ces structures d'interprétation sortées.

La distinction entre les deux sortes d'expressions fait écho à la distinction entre langage-objet et métalangage : les expressions de la première sorte sont les expressions du « langage-objet », le langage pour lequel le prédicat de vérité est un prédicat de vérité, tandis que les expressions de la seconde sorte sont le prédicat de vérité lui-même ainsi que l'ensemble des expressions du « méta-langage » dans lequel nous faisons nos attributions de vérité et qui permettent de décrire le langage-objet.[60] En raison de cette analogie, nous appellerons « expressions du métalangage » les expressions de la seconde sorte (le prédicat de vérité et les noms d'objets auquel il

[59] Nous devons cette idée à Denis Bonnay qui l'a introduite lors d'un travail commun.

[60] À la différence de ce l'on appelle en général « métalangage », ce que nous appellerons ici « métalangage » désigne spécifiquement un ensemble d'expressions dédiées au discours sur le langage-objet. En outre on ne supposera pas que ces expressions permettent parler du langage uniquement d'un point de vue syntaxique, il ne se limite donc pas *a priori* au langage de la syntaxe.

s'applique etc.) par opposition à celui constitué des expressions de la première sorte (le langage *dont* on parle dans le métalangage). Autrement dit, nous considérons un langage (au sens large) qui est constitué de deux niveaux : un langage-objet d'une part et un métalangage d'autre part. Ces deux parties du langage sont distinguées syntaxiquement et tenues à part, les expressions de chacune étant de « sortes » différentes. Du point de vue sémantique, les structures d'interprétation de ce langage seront à leur tour bipartites, pour permettre de rendre compte de l'idée que nous n'utilisons les expressions du « métalangage » que pour parler « indirectement » du « monde », c'est-à-dire de ce dont parle le langage-objet lui-même : une partie de la structure interprétera les expressions du langage objet, tandis qu'une autre partie interprètera les expressions du métalangage. Par commodité nous appellerons ces deux parties l'infrastructure et la métastructure de la structure totale.[61]

Le pouvoir d'expression conjoint, en termes de contenu$_B$, du prédicat de vérité, des outils associés de description des énoncés du langage-objet (prédicat de vérité et outils qui sont interprétés dans la partie « méta-structure » des structures), et du langage de départ sera mesuré par les classes de structures d'interprétation *du langage-objet* qu'il permet indirectement de définir. Avant de voir comment les choses se passent dans le détail, notons tout de suite que le pouvoir expressif ainsi mesuré va dépendre de l'ensemble de ces outils que nous nous donnons dans le métalangage pour parler des énoncés interprétés, et non seulement du prédicat de vérité lui-même. Si la seule façon dont je peux me référer aux énoncés du langage (-objet) est par des noms de ces énoncés entre guillemets, pour prendre un exemple informel, et que je n'ai aucune autre façon d'en parler, l'utilisation du prédicat de vérité ne m'apportera pas grand chose : je pourrai seulement dire des choses comme

« la neige est blanche » est vrai,

dont j'aurai aussi bien pu exprimer le contenu$_B$ en disant simplement

La neige est blanche.

Plus les moyens qui sont à ma disposition pour décrire des énoncés ou des ensembles d'énoncés de mon langage seront puissants, plus le pouvoir expressif associé à l'utilisation du prédicat de vérité est important. D'un autre côté, c'est le pouvoir

[61] « Superstructure » est un terme français, mais « métastructure » renvoie plus clairement à l'idée de « métalangage », et cela peut être utile pour comprendre ce que nous sommes en train de faire.

expressif du prédicat de vérité qui nous intéresse, et non une combinaison spécifique du prédicat de vérité et d'un ensemble de prédicats, fonctions, ou noms d'énoncés, c'est-à-dire un « métalangage » spécifique donné. Plutôt que d'arrêter une liste plus ou moins arbitraire des outils descriptifs que nous sommes prêts à admettre à titre de prédicats, symboles de fonctions et noms dans le « métalangage », nous raisonnerons donc sur des choix arbitraires de « métalangage », c'est-à-dire à ressources descriptives métalinguistiques donnée quelconque. L'idée générale de ce qui suit étant un peu précisée, voyons comment les chose se passent dans le détail.

Soient L_1 un langage de signature σ fondé sur un ensemble de constantes logiques K et défini de façon usuel. K peut être l'ensemble usuel des constantes logiques du premier ordre, mais nous n'avons pas besoin de cette restriction et l'on peut imaginer qu'il contienne également des quantificateurs du second ordre, ou bien des conjonctions infinies, ou encore, pour rester plus proche du langage naturel, des quantificateurs ordinaires comme « la plupart », « au moins un » etc. Soit L_2 un langage fondé sur le même ensemble de symboles logiques K, de signature τ, contenant de surcroît un prédicat unaire distingué « Vr », et dont les variables et symboles non-logiques sont distingués de ceux de L_1. Nous noterons en caractère gras les variables de L_2, $\mathbf{x}, \mathbf{y}, ...$, ainsi que les symboles appartenant à la signature $\tau : \mathbf{P_1}, \mathbf{P_2}, ...$, et réserverons les caractères ordinaires au langage L_1. L'ensemble des expressions, termes, et formules de L_1 et L_2 étant définis de la façon usuelle, on définit le *langage aléthique* $L_{\sigma,(\tau,Vr)}$ de signature (σ, τ) comme simplement la réunion $L_{\sigma,(\tau,Vr)} = L_1 \cup L_2$. Par là, il faut seulement entendre que toutes les notions standards de la syntaxe de $L_{\sigma,(\tau,Vr)}$ sont définies comme la réunion des notions correspondantes de L_1 et L_2. Ainsi l'ensemble des variables de L est la réunion de l'ensemble des variables de L_1 et de l'ensemble des variables de L_2, l'ensemble des termes de L est la réunion de l'ensemble des termes de L_1 et de l'ensemble des termes de L_2, l'ensemble des énoncés de L est la réunion de l'ensemble des énoncés de L_1 et de l'ensemble des énoncés de L_2 etc.[62] Pour faciliter les références, appelons

[62]Nous ne le ferons pas, mais l'on pourrait en fait se donner un langage plus « intégré ». Une première étape serait d'étendre l'ensemble des énoncés de L aux combinaisons booléennes d'énoncés de L_1 et L_2. Nous pourrions le faire, cela ne changerait rien à ce qui suit. Une seconde étape serait de permettre à l'intérieur des énoncés de L l'alternance de quantificateurs liant des variables de L_1 et quantificateurs liants des variables de L_2, *mais en supposant toujours que les quantificateurs polyadiques et les symboles de relations sont homogènes, au sens où ils ne prennent qu'un seul type de variable*. On peut alors définir L de façon plus naturelle et directe, à la manière d'un langage sorté. C'est ce que nous faisons avec Denis Bonnay et Julien Boyer dans un article en préparation « Logicality and truth ». La preuve du fait 2 s'en trouve considérablement compliquée, et les

τ la *signature aléthique* du langage total, $L_{\sigma,(\tau,Vr)}$, et σ l'*empreinte* de $L_{\sigma,(\tau,Vr)}$. La signature aléthique de $L_{\sigma,(\tau,Vr)}$ contient donc tout le vocabulaire du *métalangage* (excepté le prédicat de vérité), c'est-à-dire ces expressions qui « parlent » du langage-objet et seulement celles-là, tandis que l'empreinte de L contient le vocabulaire du langage-objet, toute les notions qui servent dans $L_{\sigma,(\tau,Vr)}$ à parler directement « du monde ». Nous n'interdisons pas que le langage-objet lui-même permette de parler d'un langage, et pourquoi pas de sa propre syntaxe puisque c'est un sujet comme un autre. Mais l'idée qui présidera à la sémantique du langage est que seule l'empreinte du langage parle « réellement » du monde et que, lorsque nous parlons du langage-objet *dans* le métalangage avec le vocabulaire de la signature aléthique, nous le faisons d'une façon spéciale, uniquement *en vue* de parler « du monde ». La distinction entre les deux parties du langage ne correspond pas à une différence de *nature* dans les expressions qui y figurent, mais sert à saisir à la distinction entre deux *modes* distincts du discours.

Exemple 1 (Syntaxe des langages aléthiques)**.** *Soit* $K = \{\forall, \exists, Q_{laplupart}\}$*,* $\sigma = \{R^2, P^1\}$ *et* $\tau = \{\mathbf{P}, \mathbf{a}\}$*. Notons* L *le langage aléthique* $L_{K,Vr}$ *de signature* (σ, τ)*.*

- *Les énoncés suivants sont des énoncés* L *:*
  - $\forall x \exists y(Rxy)$
  - $\forall \mathbf{x}\mathbf{P}\mathbf{x}$
  - $Vr(\mathbf{a}) \vee (\forall \mathbf{x}(Vr(\mathbf{x}) \rightarrow \exists \mathbf{y} \neg P\mathbf{y}))$
  - $QxyRxy$
- *Les suites de symboles suivantes ne sont pas des expressions de* L *:*
  - $\exists x \mathbf{P}x$ *(le prédicat et la variable ne sont pas de la même sorte)*
  - $Vr(\mathbf{a}) \wedge \exists x Px$ *(le premier membre de la conjonction est un énoncé du métalangage, pas le second).*
  - $Qx\mathbf{y}(Px \wedge \mathbf{P}\mathbf{y})$ *(Le quantificateur lie deux variables de sortes différentes)*

Nous donnons maintenant la sémantique des langages aléthiques. Soit $L = L_1 \cup L_2$ un langage aléthique[63], comme précédemment.

Définition 17. *Une structure aléthique d'interprétation du langage* L *est un couple de la forme* $(\mathcal{M}, \langle \mathcal{A}, \overline{Vr} \rangle)$*, que nous noterons* $\mathcal{M} \otimes \mathcal{A}$ *et où :*

propriétés spécifiques des constantes logiques choisies devant être prises en compte. C'est pourquoi nous avons préféré nous en tenir à cette présentation simplifiée. La perte du côté de l'élégance mathématique nous semble largement compensée par le gain en accessibilité pour un résultat qui, du point de vue conceptuel, reste essentiellement le même.

[63] J'abrège $L_{\sigma,(\tau,Vr)}$ par L pour alléger la notation

– $\mathcal{M}$ *une* σ*-structure,*
– $\mathcal{A}$ *une* τ*-structure* dont le domaine A est l'ensemble des énoncés de L_1,
– *et où* $\overline{Vr}$ *est l'ensemble d'éléments de* A *défini par :* $\phi \in \overline{Vr} \leftrightarrow \mathcal{M} \vDash \phi$.

Je noterai $\mathcal{A}_\mathcal{M}$ la structure $\langle \mathcal{A}, \overline{Vr} \rangle$. Autrement dit, avec $\tau = \{\mathbf{P}_1, \cdots, \mathbf{P}_\alpha\}$ et $\mathcal{A} = \langle A, \mathbf{P}_1^\mathcal{A}, ..., \mathbf{P}_\alpha^\mathcal{A} \rangle$, on a $\mathcal{A}_\mathcal{M} = \langle A, \mathbf{P}_1^\mathcal{A}, ..., \mathbf{P}_\alpha^\mathcal{A}, \overline{Vr} \rangle$, où l'interprétation $\overline{Vr}$ du prédicat « Vr » du métalangage est l'ensemble des énoncés du langage-objet qui sont vrais dans l'*infrastructure* de la structure aléthique. (Pour préciser ce que nous avons déjà dit : nous appelons $\mathcal{M}$ l'infrastructure de $\mathcal{M} \otimes \mathcal{A}$, et nous appellerons $\mathcal{A}_\mathcal{M} = \langle \mathcal{A}, Vr \rangle$ la métastructure de $\mathcal{M} \otimes \mathcal{A}$.)

Une *assignation complète* dans une L-structure aléthique $\mathcal{M} \otimes \mathcal{A}$ est une fonction qui à toute variable $x \in var$ (où var est l'ensemble des variable de L_1) associe un élément du domaine M de l'infrastructure de $\mathcal{M} \otimes \mathcal{A}$ et à toute variable $\mathbf{x} \in \mathbf{var}$ (où **var** est l'ensemble des variables de L_2) associe un élément du domaine A de la métastructure de $\mathcal{M} \otimes \mathcal{A}$. Si σ est une assignation complète sur $\mathcal{M} \otimes \mathcal{A}$, on note σ_{inf} (σ_{sup}) la fonction réduite de σ sur var (**var**, respectivement). La notion de satisfaction d'une formule ϕ dans une structure aléthique est définie de la façon suivante :

Définition 18 (Satisfaction dans une structure aléthique)**.** *Pour toute formule* $\phi \in L$*, toute structure aléthique* $\mathcal{M} \otimes \mathcal{A}$*, toute assignation complète* σ *sur* $\mathcal{M} \otimes \mathcal{A}$*,*

$$\mathcal{M} \otimes \mathcal{A} \vDash \phi[\sigma]$$

ssi

– $\phi \in L_1$ *et* $\mathcal{M} \vDash \phi[\sigma_{inf}]$,
– *ou* $\phi \in L_2$ *et* $\mathcal{A}_\mathcal{M} \vDash \phi[\sigma_{sup}]$

Nous avons défini les notions de langage aléthique, structure aléthique et de satisfaction dans une structure aléthique. Il nous faut maintenant rendre compte de l'idée que ce nous avons ici appelé le « méta-langage » du langage aléthique global n'est que le moyen, par un détour sémantique, de parler du monde. Autrement dit, ce que « dit » un énoncé du métalangage doit être identifié à ce qu'il nous dit de l'infrastructure, c'est-à-dire du domaine où est interprété l'*empreinte* du langage global. Ce sont les objets de ce domaine que nous regardons comme les objets réels de notre discours, ce dont on continue à parler, mais sur le mode formel, dans le métalangage, lorsque l'on dit que tels et tels énoncés sont vrais. Pour rendre compte de cette idée, je vais préciser la notion de contenu$_B$ d'un énoncé.

Considérons un langage formalisé standard de signature σ, muni d'une sémantique standard, et un énoncé de ce langage, disons ϕ. Une façon standard de représenter le « contenu » (extensionnel) de ϕ est de l'identifier à la classe $Mod(\phi)$ des σ-structures qui satisfont ϕ, c'est-à-dire à l'ensemble des structures qui sont compatibles avec la vérité de ϕ. La question qui nous intéresse est celle de savoir quels sont les « contenus » exprimés par les énoncés du métalangage, en termes des classes de structures d'interprétation *du langage-objet* qu'ils permettent de définir, et il nous faut donc raisonner à interprétation du métalangage *fixée*. Si l'on ne veut pas, et l'on ne veut pas, identifier le contenu d'un énoncé du métalangage à la classe de toutes les structures *aléthiques* dans lequel il est satisfait, mais seulement à la classe des structures d'interprétation du langage-objet avec lequel sa vérité est compatible, il faut *se donner* une interprétation fixée de la signature aléthique du langage global. On peut alors définir le contenu$_B$ d'un énoncé ϕ de façon analogue à la façon dont on a défini le contenu d'un énoncé standard :

Définition 19 (Contenu$_B$ d'un énoncé)**.** *Soit $L_{\sigma,(\tau,Vr)}$ un langage aléthique, $\mathcal{A}$ la structure qui interprète le métalangage et ϕ un énoncé du langage $L_{\sigma,(\tau,Vr)}$. Le contenu$_B$ de l'énoncé ϕ , noté $Mod_{\mathcal{A}}(\phi)$ est défini par*

$$Mod_{\mathcal{A}}(\phi) = \{\mathcal{M} : \mathcal{M} \textit{ est une } \sigma\textit{-structure et } \mathcal{M} \otimes \mathcal{A} \vDash \phi\}.$$

et le pouvoir expressif du langage total, pour un choix de signature aléthique et d'une interprétation du métalangage, est identifié à la classe de ses classes élémentaires en ce sens, i.e. $El(L_{\sigma,(\tau,Vr)}, \mathcal{A}) = \{C : \exists \phi \in L_{\sigma,(\tau,Vr)} \textit{ tel que } C = Mod_{\mathcal{A}}(\phi)\}$.

La notion de contenu$_B$ permet de rendre compte des deux intuitions déflationnistes sur l'emploi du prédicat de vérité, la transparence et l'expressivité.

Transparence. D'abord le prédicat de vérité est *transparent* dans ces applications à un nom d'énoncé, au sens où l'on vérifie facilement que le contenu$_B$ de l'énoncé

« La neige est blanche » est vrai[64]

est *identique* au contenu de l'énoncé

La neige est blanche.

[64]La remarque vaut quel que soit le terme dénotant l'énoncé « la neige est blanche » que l'on voudrait mettre à la place du nom de l'énoncé entre guillemets, pourvu qu'il dénote bien l'énoncé en question.

La définition permet plus généralement de rendre compte de l'idée qu'en attribuant la vérité à un ensemble d'énoncés on ne fait qu'affirmer ces énoncés eux-mêmes : l'énoncé du méta-langage

$$\forall \mathbf{x}\ (\mathbf{Pape}(\mathbf{x}) \rightarrow Vr(\mathbf{x}))$$

où $\mathbf{Pape}(\mathbf{x})$ est interprété par l'ensemble des énoncés du « langage-objet » L_σ qui ont été prononcés par le Pape, définit la classe *des σ-structures* dans lesquelles tous les énoncés de L_σ prononcés par le Pape sont vrais. De même, si le métalangage contient un prédicat $\mathbf{Ax_{PA}}$ interprété par l'ensemble des énoncés du langage de l'arithmétique L_{PA} qui sont des axiomes de PA, alors le contenu$_B$ de

$$\forall \mathbf{x}\ (\mathbf{Ax_{PA}} \rightarrow Vr(\mathbf{x}))$$

définit une certaine classe de L_{PA}- structures, celles précisément où les axiomes de PA sont vrais (mais elle est définie en une phrase cette fois, au lieu d'une infinité d'axiomes).

Dans chaque cas, le contenu$_B$ de l'énoncé aléthique est identique au contenu exprimé par la classe des énoncés auxquels la vérité est attribuée. On saisit ainsi précisément l'emploi du prédicat de vérité que le déflationniste avait en vue : un énoncé qui attribue la vérité à un ensemble d'autres énoncés ne dit en réalité que ce que disent ces énoncés eux-mêmes.

Expressivité. Mais ce cadre sémantique permet de rendre compte également de l'idée que les emplois du prédicats de vérité (combinés qu'ils sont toujours avec l'emploi de moyens de descriptions des énoncés, prédicat, constantes, quantificateurs, fonctions, etc.), permettent d'augmenter les *capacités expressives du langage*. Nous venons d'en voir un exemple avec les axiomes de Peano en premier ordre : un unique énoncé du méta-langage permet de définir la classe des modèles des axiomes de Peano, alors que l'on sait qu'il n'existe aucune axiomatisation finie de PA en premier ordre (dans le langage de PA).[65]

On peut aller plus loin dans cette direction, par exemple en remarquant la chose suivante. Supposons à nouveau que la logique sous-jacente soit la logique du premier

[65] C'est le théorème de Ryll-Nardzewski. J'en profite pour noter que, d'un point de vue logique plus usuel, la possibilité d'axiomatiser finiment « la plupart » des théories en utilisant un prédicat de vérité et des outils syntaxiques a été démontrée dans Kleene (1952). Bien entendu pour que l'axiomatisation soit finie, il faut que la syntaxe soit finiment axiomatisée tout en étant suffisamment forte (ce qui est possible) et il faut que les axiomes de la vérité eux-mêmes soient en nombre finis (Kleene utilise l'axiomatisation tarskienne).

ordre et considérons le langage-objet de signature σ vide. Considérons ensuite les énoncés ϕ_n de la forme

$$\exists x_1, ..., x_n \forall y \ (y = x_1 \vee ... \vee y = x_n)$$

, qui affirment chacun l'existence d'exactement n objets, et choisissons un métalangage possédant un prédicat $\mathbf{F}$ interprété exactement par cet ensemble d'énoncés (i.e. tel que $\mathbf{F}^{\mathcal{A}} = \{a \in A \ / \ a = \phi_n \text{ pour un } n\}$.) Alors

$$\exists \mathbf{x} \ (\mathbf{F}(\mathbf{x}) \wedge Vr(\mathbf{x}))$$

est un énoncé de $L_{\sigma,(\tau,Vr)}$ qui définit la classe des σ-structures finies. Par compacité, il n'existe pas même un ensemble infini d'énoncé de L_σ qui définisse cette classe.

Nous venons de voir que la sémantique proposée permet de rendre compte des emplois de la notion de vérité distingués par le déflationniste. Mais l'on peut aussi se demander, à l'inverse, ce que cette sémantique nous dit de la vérité. En faisant maintenant varier l'interprétation langage-objet, à interprétation du métalangage fixée, quels sont les énoncés du langage aléthique global qui sont vrais quelle que soit la manière dont on interprète les constantes non-logiques du langage-objet ? Définissons la notion de *validité aléthique* d'une formule ϕ du langage althique relativement à une interprétation du métalangage de la façon suivante :

Définition 20 (Validité aléthique)**.** *Soit $L_{\sigma,(\tau,Vr)}$ un langage aléthique, $\mathcal{A}$ la structure qui interprète le métalangage et ϕ un énoncé du langage $L_{\sigma,(\tau,Vr)}$. ϕ est une $\mathcal{A}$-validité aléthique ($\vDash_{\mathcal{A}} \phi$) si et seulement si pour toute σ-structure $\mathcal{M}$,*

$$\mathcal{M} \otimes \mathcal{A} \vDash \phi$$

On vérifie alors facilement que, pour tout langage aléthique, on a :

$$\vDash_{\mathcal{A}} \phi \leftrightarrow Vr(\mathbf{x}) \ [\sigma : \mathbf{x} \rightarrow \phi],$$

et que si de plus $\mathcal{A}$ contient un nom $\ulcorner \phi \urcorner$ pour chaque énoncé ϕ de L_σ, alors les « équivalences-T » deviennent valides :

$$\vDash_{\mathcal{A}} \phi \leftrightarrow Vr(\ulcorner \phi \urcorner).$$

Et plus l'outillage descriptif contenu dans le métalangage est important, plus nous enrichissons la classe des validités aléthiques. Supposons par exemple que $L_{\sigma,(\tau,Vr)}$ possède un prédicat $\mathbf{non}(\mathbf{x}, \mathbf{y})$ interprété de la façon suivante : pour tout $a, b \in A$, $(a, b) \in \overline{non}^{\mathcal{A}}$ si et seulement si l'énoncé b est la négation de l'énoncé a. Alors :

$$\vDash_{\mathcal{A}} \forall \mathbf{x}, \mathbf{y}\ (\mathbf{non}(\mathbf{x}, \mathbf{y}) \rightarrow (Vr(\mathbf{x}) \leftrightarrow \neg Vr(\mathbf{y})))$$

De même, à supposer que la métastructure contienne une relation ternaire $\mathbf{et}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ telle que pour tout $a, b, c \in A$, $(a, b, c) \in \overline{et}^{\mathcal{A}}$ si et seulement si l'énoncé a est la conjonction des deux énoncés b et c (dans cet ordre), nous aurons

$$\vDash_{\mathcal{A}} \forall \mathbf{x}, \mathbf{yz}\ [\mathbf{et}(\mathbf{x}, \mathbf{y}, \mathbf{z}) \rightarrow (Vr(\mathbf{x}) \leftrightarrow Vr(\mathbf{y}) \wedge Vr(\mathbf{z}))]$$

Et ainsi de suite.

Maintenant que nous avons un cadre pour discuter du contenu$_B$ d'un énoncé, il nous faut poursuivre.

7.3.4 La vérité comme notion logique

Si notre artifice de construction permet de modéliser l'emploi du prédicat de vérité que le déflationniste a en vue, en tenant séparés les comptes de ce qui relève du sujet apparent du discours métalinguistique où sont faites les attributions de vérités et de ce qui relève de ce qui est resté tout le « point » du discours, il reste encore à montrer que le prédicat de vérité, entendu comme le moyen de cet usage là, est logique.[66]

Dans ce qui suit, je ne vais pas chercher à défendre un candidat contre un autre au titre de ce qui doit compter comme la « véritable logique ». C'est une question

[66]Il existe dans la littérature contemporaine une tradition d'études de la logicité remontant Tarski (1966/1986) (au moins) et que j'ai déjà eu l'occasion de mentionner dans ce travail. Cette approche est fondée sur l'étude des propriétés d'invariance de l'extension des notions logiques (dont on aura donné une sémantique appropriée), dans un esprit prolongeant le programme d'Erlangen de classification des géométries de Klein. Cette tradition d'étude la logicité inaugurée par Tarski a ensuite été poursuivie par des philosophes et logiciens tels que Gila Sher, Solomon Feferman, Van McGee, et tout récemment par Denis Bonnay. L'idée fondamentale est donc d'étudier l'extension des constantes logiques (les quantificateurs essentiellement) et de chercher à déterminer quelles sont leurs propriétés d'invariance caractéristiques. Si par exemple les quantificateurs sont vus comme des classes de structures, selon la conception moderne des quantificateurs généralisés, on pourra se demander pour quelle relation R entre structures les quantificateurs logiques (puisque au sens généralisé de « quantificateur » certains quantificateurs peuvent ne pas être logiques) sont-ils clos ou, pour le dire autrement, pour quelles genre de relation R un quantificateur, s'il est logique, ne doit pas permettre de distinguer entre deux structures qui sont dans la relation R ? Ces propriétés doivent permettre de rendre compte de l'intuition de la généralité et de la formalité de la logique. Malheureusement, étudier les propriétés d'invariance du prédicat de vérité n'est pas un point de départ facile pour notre projet, toujours pour cette même raison que c'est le contenu$_B$ et non le contenu$_A$ des attributions de vérité qui nous intéresse ici. Il faut donc trouver une voie de contournement. Il se trouve cependant que la proposition que je ferai ici pour étudier le prédicat de vérité peut être également formulée de façon naturelle dans le cadre de travail qui est celui de cette tradition tarskienne. Je renvoie le lecteur à l'appendice A1 pour les détails.

complexe, multiforme, et je n'en ai pas besoin. La notion dont je vais avoir besoin est celle, *a priori* beaucoup plus simple, de « notion logique *étant donné* le choix d'un ensemble L de constantes logiques ». De quoi s'agit-il ? Supposons que, par un moyen ou un autre, nous soyons parvenus à nous mettre d'accord sur le caractère logique d'un certain nombre d'expressions (en vertu de leur signification). Appelons L l'ensemble de ces constantes logiques. Il est alors possible que ce choix détermine le caractère logique d'autres expressions qui ne figurent pas dans L : pour prendre un exemple simple, si nous avons déclaré logiques les expressions « $\wedge$ » et « $\neg$ », alors ce choix détermine le caractère logique de « $\vee$ », par exemple, tout simplement parce que nous pouvons regarder la disjonction comme une abbréviation de certaines combinaisons de négation et de conjonction. L'idée d'un critère de « logicité d'une notion étant donné le choix d'une logique L » obéit à ce genre de considérations. Une fois que nous aurons présenté ce critère, nous montrerons alors que, quel que soit le choix de constantes logiques arrêté, les emplois déflationnistes de la notion de vérité apparaissent comme des emplois logiques relativement à ce choix de départ.

Notions logiques

On se souvient de l'intuition fondamentale qui gouvernait l'idée de la logique pour Wittgenstein et pour Russell : les constantes logiques ne sont pas des constituants de la proposition mais articulent la forme logique de cette proposition. Plus précisément, on peut distinguer entre, d'un côté, les propriétés (je parlerai pour ma part de prédicats), les individus (les noms), et d'un autre côté l'articulation de ces propriétés et de ces individus en une proposition qu'il revient aux mots logiques de décrire ou d'exprimer. Je ne chercherai pas à discuter de façon critique cette idée ici, ni à en donner un compte-rendu historiquement fidèle. Au contraire, je *partirai* simplement de cette intuition informelle pour justifier les termes dans lesquels je vais poser le problème du propre de la logique.

On peut voir l'enjeu qui sous-tend le choix d'une logique comme celui de la façon dont doit être résolu la tension existant entre d'une part, l'idée que les expressions logiques ne doivent pas avoir de contenu -disons de contenu « descriptif » - et, d'autre part, l'idée qu'elles doivent permettre de décrire des états du monde qui ne sont pas descriptibles sans elles. Pour donner un peu plus de sens à cette idée, supposons que nous représentions un état du monde relativement à un vocabulaire « descriptif » de départ $\mathcal{V}$ sur un domaine donné, par une $\mathcal{V}$-structure d'interprétation sur le domaine

en question. *Nous* faisons donc cette description, non dans langage de vocabulaire $\mathcal{V}$ lui-même, mais dans un métalangage ensembliste. Nous voyons que deux « états du monde » relativement à $\mathcal{V}$ peuvent être distingués de plusieurs manières. Tout d'abord, certains peuvent l'être par des énoncés atomiques *du langage de vocabulaire* $\mathcal{V}$, comme par exemple les deux états suivants relativement à un vocabulaire $\mathcal{V} = \{P^1, c_1, c_2\}$ sur le domaine $\{1, 2, 3\}$:

$$\mathcal{M}_1 = \langle \{1, 2, 3\}, \overline{P} = \{1\}, \overline{c_1} = \{1\}, \overline{c_2} = \{2\} \rangle$$

et

$$\mathcal{M}_2 = \langle \{1, 2, 3\}, \overline{P} = \{1, 2\}, \overline{c_1} = \{1\}, \overline{c_2} = \{2\} \rangle$$

Dans le second état, l'énoncé « Pc_2 » est vrai, tandis qu'il est faux dans la première. Dans d'autres cas, il n'est pas possible de distinguer entre deux de nos descriptions d'états du monde par un simple énoncé atomique du langage $\mathcal{V}$, mais il doit rester possible de décrire leur différences comme une différence d'*arrangement* de ces propriétés et de ces individus dans le domaine : en vertu de l'idée que les expressions logiques sont celles qui n'ont pas de contenu mais permettent d'articuler des contenus, ce sont les différences dont il doit être possible de rendre compte par un énoncé logiquement complexe construit sur le vocabulaire $\mathcal{V}$. Par exemple, il n'est pas possible de distinguer entre

$$\mathcal{M}_3 = \langle \{1, 2, 3\}, \overline{P} = \{1\}, \overline{c_1} = \{1\}, \rangle$$

et

$$\mathcal{M}_4 = \langle \{1, 2, 3\}, \overline{P} = \{1, 2, 3\}, \overline{c_1} = \{1\} \rangle$$

par un énoncé atomique, mais l'énoncé complexe « $\forall x P(x)$ » permet de distinguer entre l'état de choses décrit par $\mathcal{M}_4$ et celui décrit par $\mathcal{M}_3$.

Enfin, parfois, pour le choix d'une logique L donnée, il n'est pas possible de distinguer entre deux de nos descriptions d'états par un énoncé, quel qu'il soit, du vocabulaire considéré. Dans ce dernier cas, ce qui doit compter comme « la logique » ayant été fixé, on doit conclure simplement que la différence que fait apparaître notre représentation ensembliste entre les deux états n'est pas une différence qui peut être décrite comme simplement une différence d'arrangement des propriétés et des individus qui sont le sujet de $\mathcal{V}$ sur le domaine en question, au sens d'« arrangement » qui est implicite dans le choix de L. Ce n'est pas dire que ces deux états du monde

décrits dans notre métalangage ensembliste sont indiscriminables en général, mais seulement que la différence que révèle notre description doit être vue comme devant faire appel à de nouveaux prédicats, de nouvelles constantes, éventuellement dans le contexte d'un domaine plus grand que celui de l'état décrit : il s'agit d'une différence qui met en jeu essentiellement autre chose qu'une simple différence d'*articulation* des notions et des noms de $\mathcal{V}$. En effet, il est assez clair que se prononcer d'emblée sur le fait que toute différence entre deux de nos représentations ensemblistes de deux états du monde relativement à un vocabulaire $\mathcal{V}$ sur un domaine donné est simplement une différence qu'il devrait être possible de faire dans le vocabulaire $\mathcal{V}$ avec une logique satisfaisante, c'est *présumer* de ce qui doit compter comme logique en permettant à des différences qui sont descriptibles dans le langage de départ *augmenté* du langage de la théorie des ensembles, de compter comme des différences qui pourraient en droit être décrites comme de simples différences d'*articulation* entre nos notions de départ.

Le problème de savoir quelle logique est la « vraie » logique, dans ce cadre, pourrait alors se poser de la façon suivante : étant donné un stock $\mathcal{V}$ de noms et de prédicats, quelles descriptions du monde *relativement à* $\mathcal{V}$ doit-il être possible de distinguer par ce qui doit compter comme simplement une *articulation logique* de ces noms et de ces concepts ? Où faut-il mettre la frontière entre ces différences qui demandent à être décrites par l'introduction de propriétés nouvelles, parce ce que ce sont en somme des différences « substantielles », et celles qui peuvent être décrites simplement comme des différences d'« arrangements » des propriétés et les individus donnés au départ dans le domaine du discours ?

Pour illustrer cette façon de poser la question, je propose d'examiner quelques exemples. Considérons le langage ordinaire de l'arithmétique, avec un symbole de relation d'ordre, le symbole de la fonction successeur, de l'addition, de la multiplication et une constante d'individu « 0 ». Si « la » logique est la logique du premier ordre, on sait par compacité qu'il existe des modèles non-standards de l'arithmétique, c'est-à-dire des structures élémentairement équivalentes au modèle standard de l'arithmétique $\mathcal{N}$, mais non isomorphes à lui.[67] Si $\mathcal{M}$ est un modèle non-standard, non seulement il n'existe aucun énoncé atomique qui nous permettrait de distinguer entre $\mathcal{N}$ et $\mathcal{M}$, mais l'on ne peut saisir dans le vocabulaire $\mathcal{V}$ aucun trait qui distinguerait $\mathcal{N}$ de $\mathcal{M}$ même par un énoncé logiquement complexe. Dans

[67] Par le théorème de Löwenheim-Skolem, on sait qu'il y a en un de même cardinalité que le modèle de départ.

les termes que nous avons adoptés pour commencer, il y a donc deux descriptions d'états du monde relativement au vocabulaire de l'arithmétique sur un domaine, qui peuvent être distinguées dans le langage dans lequel sont faites ces descriptions (celui de la théorie des ensembles), mais qui ne peuvent pas être faites dans le langage de l'arithmétique du premier ordre lui-même. Pour le dire métaphoriquement, la différence entre $\mathcal{N}$ et $\mathcal{M}$ ne peut être saisie simplement comme une différence d'*arrangement* des propriétés et des objets pour lesquels nous avons des concepts et des noms dans $\mathcal{V}$, du moins si notre notion de ce qui compte comme simple articulation de contenus n'excède pas ce qu'il est permis d'entendre par là si la logique du premier ordre est « la » logique. À nouveau, cela ne veut pas dire qu'on ne peut pas distinguer les deux états du monde en logique du premier ordre (c'est ce que nous faisons dans notre description ensembliste), par exemple en introduisant de nouveaux concepts dans notre langage, des concepts ensemblistes, ou/et en plongeant cette description dans une description d'état de domaine plus large, mais simplement que cette distinction n'est pas une distinction que la logique peut faire à l'aide des seules ressources descriptives données au départ. D'un autre côté, si l'on admet que la logique du second-ordre est véritablement de la logique, le tableau change du tout au tout (dans l'exemple précédent), parce qu'il est maintenant possible, avec les seules ressources conceptuelles-descriptives de départ (le vocabulaire non-logique), de former un énoncé, c'est-à-dire un arrangement logique des concepts de départ, qui permet de distinguer $\mathcal{N}$ et $\mathcal{M}$, à savoir par exemple l'énoncé exprimant le fait que la relation d'ordre est un bon ordre, puisque les modèles non-standards de l'arithmétique complète du premier ordre ne sont pas bien-ordonnés et que la propriété de bon ordre est exprimable en second ordre.[68,69] Si la logique du second-ordre est de la logique alors, d'après la conception de la logique esquissée en début de section, ce qui distingue les états $\mathcal{N}$ et $\mathcal{M}$ peut être vue comme une simple différence d'articulation des concepts et individus sur le domaine, ne mettant pas essentiellement en jeu de nouvelles propriétés ou nouveaux concepts hors de ceux qui sont consignés dans notre vocabulaire de base.

[68] Par exemple : pour tout X, il existe x, tel que pour tout y, $(Xx \wedge Xy \wedge x \leq y)$.

[69] Dans le vocabulaire que l'on s'est donné, on peut axiomatiser catégoriquement l'arithmétique, et donc *a fortiori* si $\mathcal{M}$ est une structure non isomorphe à $\mathcal{N}$,

$$\mathcal{N} \not\equiv_{SOL} \mathcal{M}$$

où « $\equiv_{SOL}$ » est la relation qui existe entre deux structures si et seulement si elles satisfont exactement les mêmes énoncés du second-ordre.

Ce qui m'importe dans l'exemple précédent, ce n'est pas tant qu'il plaiderait en faveur, ou contre, l'idée que la logique du second-ordre est de la logique, mais qu'il peut être compris comme une façon d'illustrer le fait qu'en passant de la logique du premier ordre à la logique du second ordre on change véritablement de notion de logicité. Si la logique du premier ordre est *la* logique, alors ce que l'on appelle « logique du second ordre » n'est pas de la logique, mais autre chose. On peut également trouver des exemples parfaitement clairs de notions qui ne sont pas logiques. S'il l'on se donnait par exemple un quantificateur monadique Q_R défini de la façon suivante :

$$\mathcal{M} \vDash Q_R(x)\phi(x)[\sigma]$$
si et seulement si,
il existe un objet rouge qui satisfait ϕ[70],

on pourrait alors distinguer entre deux descriptions d'états du monde sur la base de la couleur de certains individus du domaine, quel que soit le vocabulaire relativement auquel l'état de chose est décrit. L'idée que l'emploi d'une telle expression ne participerait que de l'articulation de nos concepts descriptifs de départ est contre-intuitive, parce qu'alors qu'est-ce qui ne compterait pas comme logique ? Pour toute décision *raisonnable* de ce qui doit compter comme logique, il est clair que le quantificateur Q_R ne doit pas compter comme logique. Inversement, si nous acceptons l'idée que la conjonction et la négation sont des constantes logiques, alors la description que nous avons faite de ce qui est en jeu dans le choix d'une logique conduit naturellement à affirmer que la disjonction est également une notion logique : en ajoutant la disjonction à la conjonction et à la négation, il n'est pas possible *dans les langage de vocabulaire* $\mathcal{V}$ de faire entre deux de nos descriptions-d'états-relativement-à-$\mathcal{V}$ des distinctions qu'il ne serait pas possible de faire simplement avec la conjonction et la négation.

Pour résumer ce que nous venons d'esquisser, si l'enjeu qui sous-tend le choix d'une logique peut être vu comme celui de la façon dont doit être résolu la tension existant entre d'une part, l'idée que les expressions logiques ne doivent pas avoir de contenu « descriptif » et, d'autre part, l'idée qu'elles doivent permettre l'expression de contenus qui ne sont pas exprimables sans elles, alors on peut représenter le choix d'une logique L, c'est-à-dire le choix d'un ensemble de constantes logiques, comme

[70] Plus précisément : si et seulement s'il existe un objet rouge a tel que toute assignation σ' tel que $\sigma'(x) = a$ et identique à σ sinon, satisfait ϕ.

le choix de ce qui doit compter, entre deux $\mathcal{V}$-structures, *pour un vocabulaire* $\mathcal{V}$ *quelconque*, comme simplement des différences d'articulation des notions de $\mathcal{V}$ sur un domaine donné. Maintenant, à supposer donné le choix d'une logique L, je dirai que deux descriptions d'états du monde relativement au vocabulaire $\mathcal{V}$ sont $L_{\mathcal{V}}$-indiscriminables s'il n'existe pas d'énoncé formé à partir du stock de concepts et de noms de départ (le vocabulaire « descriptif » du langage) dans les formes qui sont permises par le choix de la logique L, et qui soit vrai dans l'une mais fausse dans l'autre. Autrement dit :

Définition 21 ($L_{\mathcal{V}}$-indiscriminabilité)**.** *Soit* $\mathcal{V}$ *un vocabulaire non logique,* $\mathcal{M}$ *et* $\mathcal{M}'$ *deux* $\mathcal{V}$*-structures, et* L *une logique.*

$$\mathcal{M} \textit{ et } \mathcal{M}' \textit{ sont } L_{\mathcal{V}}\textit{-indiscriminables si et seulement si } \mathcal{M} \equiv_L \mathcal{M}'$$

où $\equiv_L$ *est la notion d'équivalence élémentaire associée à la logique* L*.*[71]

J'en viens maintenant à la notion qui m'intéresse plus directement ici, celle de « logicité d'une notion étant donné le choix d'une logique L ». Supposons à nouveau que nous ayons une idée claire de qui ce doit compter comme logique, c'est-à-dire des types de différences entre structures d'interprétation qui peuvent être descriptivement prise en charge par la forme logique, et de celles qui doivent être faites par l'introduction de nouveaux concepts, et supposons que nous ayons la logique correspondante L, c'est-à-dire un stock de constantes logiques *complet* au sens où il nous permet de faire toutes (et seulement) les distinctions visées entre structures qui satisfont les mêmes énoncés atomiques mais qui ne sont pas $L_{\mathcal{V}}$-indiscriminables. Si ce que nous avons dit est correct, il y a un sens clair dans lequel on peut discuter de la question de savoir si une notion qui n'est pas lexicalisée parmi les constantes de L est, ou n'est pas, une notion logique : il suffit d'observer son comportement relativement à la notion d'équivalence L-élémentaire. Si ajouter comme constante interprétée aux côtés de nos L-constantes logiques l'expression « δ » ne permet pas de discriminer des structures qui sont $L_{\mathcal{V}}$-indiscriminables, et ce quel que soit le vocabulaire $\mathcal{V}$ considéré, alors « δ » est logique étant donné la notion de logicité véhiculée par le choix de L ; si non, cela signifie que sa contribution à l'expressivité du langage n'est pas une contribution purement logique : la notion δ doit être vue comme véhiculant un authentique « contenu » et non seulement comme

[71] Deux $\mathcal{V}$-structures sont L élémentairement équivalentes si et seulement si elles satisfont exactement les mêmes énoncés du langage engendré par L sur le vocabulaire $\mathcal{V}$.

un moyen d'articuler des contenus. Ces considérations motivent l'introduction du critère suivant de logicité d'une notion étant donné un choix de constantes logiques L :

Proposition 6 (Critère de logicité d'une notion (étant donné un choix de constantes logiques))**.** *La notion « δ » est logique selon L si et seulement si*

$$\equiv_L = \equiv_{L+\delta}$$

où L désigne un ensemble complet de constantes logiques, « $\equiv_L$ » la notion d'équivalence élémentaire associée à L, et « $\equiv_{L+\delta}$ » la notion d'équivalence élémentaire associée à $L \cup \{\delta\}$.

Le caractère logique des emplois du prédicat de vérité

Le critère que nous avons proposé présente l'intérêt d'être applicable dans le cadre de travail que nous avons mis en place pour séparer le pouvoir expressif du prédicat de vérité des ressources auxiliaires dans lequel il doit être plongé pour être utilisé : pour savoir si l'emploi du prédicat de vérité tel qu'il est compris par le déflationniste est un emploi logique, il suffit de s'enquérir des discriminations entre structures d'interpétation (du langage-objet) qu'il permet. Et le résultat attendu est quasiment immédiat :

Fait 4. *Etant donné un métalangage interprété dans une métastructure $\mathcal{A}$, pour toute σ-structure $\mathcal{M}$,*

$$\textit{si } \mathcal{M} \equiv_{L_\sigma} \mathcal{M}', \textit{ alors } \mathcal{M} \otimes \mathcal{A} \equiv_{L_{\sigma,(\tau,Vr)}} \mathcal{M}' \otimes \mathcal{A}$$

Démonstration. Si $\mathcal{M} \equiv_{L_\sigma} \mathcal{M}'$ alors, pour tout énoncé de L_σ, $\mathcal{M} \vDash \phi$ ssi $\mathcal{M}' \vDash \phi$. Par définition des structures aléthiques, il s'en suit que $\mathcal{A}_\mathcal{M} = \mathcal{A}'_{\mathcal{M}}$. La définition 9 permet de conclure. □

L'ajout d'un prédicat de vérité interprété à une logique selon les modalités que nous avons détaillées dans la section précédente, et quel que soit l'appareil conceptuel auxiliaire mis en œuvre, ne change pas la notion d'équivalence élémentaire : si deux structures (dans un vocabulaire donné) ne peuvent pas être distinguées par un énoncé (du vocabulaire de la structure) couché dans la logique de départ, alors elles ne peuvent pas être distinguées par l'emploi déflationniste du prédicat de vérité. Mais si le contenu$_B$ des énoncés contenant le prédicat de vérité ne permet pas de

faire des distinctions que la logique (de départ) ne fait pas, alors le prédicat de vérité, dans son emploi déflationniste, fonctionne lui aussi comme un prédicat logique au sens en jeu dans la logique de départ : en ajoutant un prédicat de vérité à notre langage, on n'a pas introduit de nouveau « contenu ». En vertu du critère de logicité d'une notion relativement à une logique L que nous avons proposé, nous concluons donc sur l'affirmation suivante :

Proposition 7. *Pour tout logique L, les emplois déflationnistes de la notion de vérité sont logiques au sens de L.*

Avant de conclure brièvement, je ferai deux remarques. Tout d'abord ce que nous avons remarqué dans la section précédente à propos du pouvoir expressif aditionnel donné par le métalangage est compatible avec le Fait 4 : certes, une structure finie et une structure infinie ne sont pas élémentairement équivalentes dans le langage aléthique, mais elles ne sont pas élémentairement équivalentes en logique du premier ordre non plus. Ma seconde remarque est que ce résultat ne montre pas que les *autres* notions du métalangage seraient logiques également, ce qui serait contrintuitif. En effet, il faut se souvenir que le contenu de ces autres notions a été mis en prarenthèse, isolé : le contenu$_B$ d'un énoncé du métalangage ne contenant pas de prédicat de vérité est nul dans la sémantique que nous avons proposée tout à l'heure, et c'était précisément à cet effet que nous avions introduit la notion de contenu$_B$. Mais dire que ces expressions n'ont pas de contenu$_B$, ou remarquer que leur contribution *au contenu$_B$* d'un énoncé du métalangage est tel ou tel, ce n'est pas dire que ces expressions n'ont pas de contenu, de contenu au sens ordinaire. Si l'on veut étudier le contenu de *ces* expressions, alors mettons-les en position de langage-objet dans notre construction. Et à leur tour, il sera possible de parler de ce dont elles parlent, soit en les utilisant, sur le mode matériel du discours, soit en parlant de ces expressions elles-mêmes et en utilisant la notion de vérité, sur le mode formel du discours, avec le secours d'expressions du métalangage dont la contribution au contenu$_B$ des énoncés métalinguistiques aura été isolée.

Nous avons proposé une analyse sémantique des emplois déflationnistes de la notion de vérité, puis une représentation possible des enjeux sémantiques qui sont ceux sous-jacent au choix d'une logique, et nous avons proposé sur cette base un argument dont la conclusion est que les emplois déflationnistes de la notion de vérité sont des emplois logiques. Pour autant que nos explications des emplois déflationnistes de la notion de vérité et de la notion de logicité sont correctes, cette conclusion montre

que la thèse selon laquelle le prédicat de vérité est une notion logique est une thèse déflationniste cohérente.

7.3.5 Conclusion sur l'approche sémantique

Je reviens brièvement sur ce que j'ai essayé de montrer dans cette « approche sémantique » de la logicité de la notion de vérité. Le problème de départ était double. Il fallait d'une part proposer un critère permettant de décider si une notion est logique ou pas, et d'autre part trouver un moyen d'identifier ce qui est exprimable par l'emploi du prédicat de vérité au sens déflationniste, c'est-à-dire indépendamment de ce qui est exprimé par les moyens auxiliaires qui rendent possible cette attribution. L'introduction des « structures aléthiques » était le moyen de résoudre le second problème : elles nous permettent, derrière la montée sémantique et ses vicissitudes, de continuer à regarder « vers le monde » et d'analyser au clair ce que nous en disons. La résolution du second problème s'est révélée facilitée par une découverte inattendue : pour montrer que les emplois « déflationnistes » du prédicat de vérité sont des emplois logiques, il n'est pas nécessaire d'avoir résolu le problème général de la démarcation des notions logiques. La raison en est que, quelle que soit la notion de logicité qu'on aura pu trouver, le critère de « logicité d'une notion relativement à un choix de constantes logiques donné » que nous avons formulé nous assure que l'emploi déflationniste du prédicat de vérité est un emploi logique. Quant à ce critère lui-même, il était motivé d'abord par l'esprit d'une certaine tradition sémantique, selon laquelle les notions logiques sont comprises comme ne possédant pas de contenu, mais servant uniquement à articuler des contenus. Dire cela, néanmoins, ce n'est pas dire que les contenus articulés par des formes logiques complexes ne permettent pas d'exprimer davantage que les contenus simples. Ce surplus expressif que permet l'articulation complexe des contenus relativement à un langage qui serait dépourvu de mots logiques est en quelque sorte la contribution propre de la logique au pouvoir expressif du langage. L'idée est alors que, une tierce notion est logique si sa contribution au pouvoir expressif du langage n'excède pas cette marge qu'il revient de droit à la logique d'occuper.[72] Dans les appendices A1

[72]Dans le cadre de travail que j'ai proposé, étant donné un langage fondé sur un ensemble de constantes logiques et un vocabulaire non-logique, si l'ajout d'une notion prétendue logique à ce langage peut bien permettre de définir des classes de structures qu'il n'était pas possible de définir sans lui (c'est l'utilité des notions logiques), cet ajout ne doit pas modifier la relation d'équivalence élémentaire associée au langage de départ.

et A2, après une présentation historique du travail de Tarski sur les notions logiques *via* la notion d'invariance, je présente les choses sous un jour un peu différent, et j'explique le lien entre mon approche et les travaux classiques fondés sur la notion d'invariance.[73]

Pour conclure, comme je l'ai déjà dit, je ne regarde pas ces conclusions comme définitives. Les « structures aléthiques » sortées, de même que le critère de « logicité d'une notion étant donné un stock de constantes logiques », sont des constructions qui demanderaient à être étudiées de façon plus approfondie et systématique pour être mieux comprises. Si les structures aléthiques permettent de saisir *un* aspect de la sémantique du discours sur la vérité, et il n'a jamais été question de dire plus que cela, quelles autres emplois d'autres notions ce type de construction ne donnerait-t-il pas pour logique ? Au contraire, la notion de vérité est-elle dans une position singulière ici ? C'est à une étude de ce genre, que nous remettons à des recherches ultérieures, qu'il reviendrait maintenant d'étayer et de préciser les conclusions que nous avons formulées ici.

7.4 Conclusion

La notion de vérité telle qu'elle est analysée par le déflationniste présente plusieurs propriétés qui l'apparentent à une notion logique. C'est d'abord une notion universelle, au sens où elle s'applique à tous les domaines de discours indifféremment, pour peu que l'on admette avec le déflationniste que la montée sémantique, ou le passage du *mode* matériel au *mode* formel du discours ne constitue pas un changement de *sujet*. C'est cette même intuition de la transparence des attributions de vérité qui était derrière l'idée que le prédicat de vérité ne joue pas de rôle « substantiel » dans les explications. Et cette idée est à son tour en harmonie avec cette autre que le prédicat de vérité n'a pas de contenu, mais qu'il sert plutôt à *articuler* des contenus.

Nous avons vu que cette analogie entre l'analyse déflationniste des usages du prédicat de vérité et l'usage des expressions logiques pouvait en fait être menée plus loin sur le terrain des analyses formelles de la logicité. En adoptant le cadre conceptuel qui est celui de la théorie de la preuve philosophique, j'ai montré que les règles minimales d'introduction et d'élimination du prédicat de vérité, si elles

[73] Ces travaux permettent également de faire le lien avec l'intuition de la logique en termes de généralité.

ne permettent pas de définir explicitement le prédicat de vérité (ou de fixer son extension dans un modèle), peuvent néanmoins être vues comme définissant, en un autre sens, la « signification inférentielle » de l'expression, à la façon dont les règles d'inférence usuelles pour les constantes logiques peuvent être dites en fixer la signification. Ce point prolonge nos conclusions du chapitre précédent sur le caractère suffisant de la théorie minimale pour rendre compte de nos usages effectifs du prédicat de vérité pour l'« expression des généralisations ». Qui plus est, ces règles apparaissent comme constituant la signification d'une expression *presque logique*, et en fait aussi logique qu'une analyse de la contribution du prédicat de vérité à la *signification inférentielle* des énoncés puisse le révéler. Avec ces conclusions, il semblait en effet que nous touchions aux limites de l'analyse inférentielle pour notre projet : une analyse de l'*usage* (inférentiel) du prédicat de vérité ne permet pas d'isoler sa signification propre de la contribution spécifique des termes auxquels il s'applique à la signification des énoncés. L'intuition que le prédicat de vérité pourrait être logique, bien qu'un énoncé du type

si « la neige est blanche » est vrai alors la neige est blanche

ne le soit pas, suggérait que la conclusion atteinte à ce stade pouvait n'être pas définitive.

Cette intuition qu'une bonne analyse du prédicat de vérité devrait permettre de faire *la part des choses*, fait écho à cette autre intuition, sémantique, qui est derrière l'analyse déflationniste de la signification des attributions de vérité comme ne parlant pas, en fait, des énoncés auxquels elle est attribuée, mais de ce dont parlent ces énoncés eux-mêmes. Cette intuition a motivé l'introduction d'un cadre de travail particulier devant permettre une analyse sémantique de *ces* usages du prédicat de vérité, en isolant ce qui exprimé par une attribution de vérité au sens voulu par le déflationniste. Ce cadre de travail sémantique était fourni par les « structures aléthiques ». Sur cette base, j'ai proposé d'examiner ce que l'introduction d'un prédicat de vérité dans un langage ajoutait sur le plan expressif à ce langage. Après avoir montré que la sémantique des attributions de vérité proposée rendait compte des intuitions déflationnistes fondamentales, j'ai essayé de montrer que la contribution du prédicat de vérité à l'expressivité du langage, telle qu'analysée ici, était en harmonie avec une certaine idée de la logicité. Plus précisément, j'ai posé la question suivante : l'introduction du prédicat de vérité (en sens en jeu ici) permet-elle de faire des distinctions entre des « descriptions d'états de choses relativement à un

vocabulaire donné » que la logique de départ ne permet pas de faire ? Et nous avons vu que la réponse à cette question était Non. J'ai suggéré que l'on pouvait faire valoir ce résultat pour étayer la thèse selon laquelle les usages du prédicat de vérité tels qu'ils sont compris par le déflationnistes sont des usages purement logiques. Nous avons vu également que ce résultat ne contredisait pas l'idée que le prédicat de vérité pouvait avoir une valeur expressive, au sens où, même si l'introduction du prédicat de vérité ne permet pas de faire entre des structures des distinctions qui n'auraient pu être faites sans lui, elle permet en revanche en général de *définir* des classes de structures qui ne sont pas définissables sans lui. Ce résultat illustre d'une autre manière l'idée d'une expression qui n'aurait pas de contenu « substantiel », mais aurait un rôle expressif important, qui était l'intuition principale de la logicité dont nous étions parti.

Comme je l'avais annoncé en introduction à ce chapitre, aucune de ces analyses n'est complètement conclusive. Néanmoins, elle permettent ensemble de mieux cerner les relations qu'entretiennent l'idée de logicité et la conception que les déflationnistes se font de la notion de vérité. Jusqu'à un certain point, elles me paraissent montrer que l'idée de logicité du prédicat de vérité est bien cohérente avec les thèses déflationnistes.

7.5 Annexe au chapitre 7

Harmonie locale et paradoxes

L'analyse inférentielle du prédicat de vérité peut-elle nous apprendre quelque chose sur les paradoxes de la vérité ? Les quelques remarques qui suivent n'ont pas l'ambition de répondre à cette question, mais seulement celle, infiniment plus modeste, d'indiquer pourquoi elle se pose. Elles permettent également de donner de donner quelques éléments de réflexion sur la portée normative des critères d'harmonie que nous avons donnés.

Nous avons remarqué que le critère d'harmonie globale fournie par la conservativité n'épousait qu'assez mal les considérations générales qui motivent l'analyse inférentielle de la signification : si pour justifier des règles il faut le garde-fou de la conservativité alors, si la validité des règles ne dépend que de la signification des termes qui y figurent, il n'est pas clair que l'on a pas abandonné en cours de route l'idée que toute la signification des termes est donnée par les règles. Stephen Read

Read (2000) a plaidé pour une distinction stricte entre le problème de l'harmonie des règles introduisant et éliminant une expression et les propriétés globales du système de logique qui en résulte, qu'il s'agisse de la conservativité, de la cohérence, ou de la normalisation globale des preuves. Read veut s'en tenir à l'idée que (1) les règles d'introduction d'une expressions fixent complètement la signification d'une expression, (2) que ces règles justifient, en vertu de la signification qu'elles confèrent à l'expression, certaines règles d'élimination et (3) que toute dérivation qui n'utilise que ces règles est justifiée *en vertu la signification* des expressions dont les règles sont utilisées dans la dérivation.

Pour déterminer quelles règles d'élimination sont justifiées, Read fournit une procédure générale systématique permettant d'associer à toute règle d'introduction d'une expression une règle générale d'élimination correspondante (GE), telle que la règle d'élimination en question puisse être déclarée valide en vertu de la signification inférentielle conférée à l'expression par sa règle d'introduction. L'idée fondamentale derrière la notion de justification est toujours celle de Gentzen-Prawitz-Dummett, à savoir qu'une règle d'élimination est valide si les conclusions qu'elle autorise sur la base d'un énoncé où l'expression « δ » est dominante (et d'hypothèses auxiliaires Γ) sont justifiées par (Γ et) ce qui compte comme une justification pour l'introduction de l'énoncé en question (tel que spécifié par sa règle d'introduction). Guidé par cette idée, la forme de la procédure générale de « dérivation » de la règle d'élimination canoniquement associée à une règle d'introduction donnée développée par Read est la suivante. Si δ est une expression, dont les règles d'introduction sont les suivantes (où les Π_i sont des dérivations)

$$\frac{\Pi_1}{A}\,\delta I \qquad \frac{\Pi_2}{A}\,\delta I \quad \cdots$$

La règle d'élimination générale associée (GE) sera la suivante (où les Π_i sont déchargées par l'élimination de δ) :

$$\frac{\begin{array}{cccc} & (\Pi_1) & (\Pi_2) & \cdots \\ A & C & C & \cdots \end{array}}{C}\,\delta GE$$

Pour voir que la règle d'élimination est justifiée par la règle d'introduction, il suffit de considérer le cas où la règle d'élimination succède immédiatement à une règle d'introduction :

$$\frac{\begin{array}{cccc} \dfrac{\Pi_i}{A}\,\delta I & (\Pi_1) & (\Pi_2) & \cdots \\ & C & C & \cdots \end{array}}{C}\,\delta GE$$

Mais dans ce cas l'occurrence visible de la formule A[74] dans le schéma de preuve ci-dessus peut être éliminée puisque en fait nous avons déjà une preuve de C par Π_i. Cette façon d'engendrer la règle d'élimination canonique associée à un ensemble de règles d'introduction étant donnée, une règle d'introduction et une règle d'élimination donnée sont en harmonie, au sens de Read, si et seulement si la règle d'élimination peut être dérivée[75] de la règle générale d'élimination canoniquement associée à la règle d'introduction.

L'absence de conditions d'harmonie globale portant sur le système qui résulte de l'introduction des règles pour une expression et le caractère strictement local de la propriété d'harmonie n'est pas si faible qu'il serait au bout du compte dépourvu de toute valeur normative et il est facile de voir que le critère est suffisant pour disqualifier la validité des règles d'élimination pour « tonk » : la règle d'élimination de « tonk » donnée par Prior n'est pas dérivable de la règle d'élimination générale canoniquement associée à sa règle d'introduction par la procédure de Read.[76] La force normative du critère est donc bien réelle. Ce qui est intéressant, pourtant, est l'autre question : le critère local de Read a-t-il pour conséquence que l'introduction de règles harmonieuses à un système de logique est une extension conservative, en préserve la cohérence, ou garantit la normalisation des preuves ? Et à nouveau la réponse est non. Précisément, les règles d'introduction et d'élimination *naïves* du prédicat de vérité (au sens où l'on parle de théorie naïve des ensembles), c'est-à-dire sans restriction sur le langage de la classe des énoncés qui peuvent être des arguments du prédicat Vr, illustrent ce phénomène.

En effet, il est facile de voir en général que les règles naïves d'introduction et d'élimination du prédicat de vérité satisfont les critères d'harmonie strictement locaux. Si l'on veut adopter la forme que donne Read à ce critère, on peut vérifier qu'étant donné la règle d'introduction de Vr, sa procédure donne la règle généralisée d'élimination (GE) suivante :

$$\frac{\Gamma \vdash \text{Vr« A »} \qquad \Gamma', A \vdash B}{\Gamma, \Gamma' \vdash B}$$

Or il est facile de voir que cette règle d'élimination est équivalente à notre règle (ordinaire) d'élimination. Pour voir que (GE) implique (Vr-elim), il suffit de prendre

[74] Cette occurrence qui est en position de conclusion d'une règle d'introduction d'une expression et en position de prémisse principale d'une règle d'élimination de cette expression.

[75] Au sens strict cette fois où l'on pourrait montrer qu'elle est, selon le terme technique consacré, « admissible » dans le calcul.

[76] Voir à nouveau Read (2000) pour les détails.

$\Gamma = \emptyset$ et $B = A$ dans (GE). Ceci donne (Vr-elim). Pour la réciproque, il suffit d'utiliser l'admissibilité des coupures en déduction naturelle :

$$\dfrac{\dfrac{\Gamma \vdash \mathrm{Vr}\text{« A »}}{\Gamma \vdash A}\ \mathrm{Cut} \qquad \Gamma', A \vdash B}{\Gamma, \Gamma' \vdash B}$$

Par conséquent, selon l'analyse de Read, non seulement les règles « tarskiennes », restreintes, pour la vérité, mais également les règles naïves pour le prédicat de vérité sont en harmonie et, comme il est facile de s'en convaincre, la même conclusion s'impose pour tous les critères d'harmonie strictement *locaux* de ce type, en en particulier pour les deux variations sur ce critère que nous avons mentionnés dans le corps du chapitre, à savoir l'obéissance au principe d'inversion (en déduction naturelle) ou son *alter ego* dans le calcul des séquents[77] Mais puisque qu'il ne fait aucun doute que le critère d'inversion *a* une portée normative, il faut en conclure que les règles naïves pour Vr, bien qu'engendrant des contradictions, sont d'une nature différente de celle des règles pour *tonk*, et que les premières sont en effet *correctes* en un sens où les secondes ne le sont pas.

D'un point de vue technique, on sait que la propriété d'harmonie locale des règles, dans le cas des constantes logiques ordinaires, est à la base des théorèmes de normalisation (déduction naturelle) et d'élimination des coupures (calcul des séquents) de Prawitz et Gentzen pour la logique du premier ordre, et que ces résultats permettent de prouver la cohérence des règles de la logique du premier ordre.[78] Mais le passage de l'harmonie des règles à la normalisation du calcul et à la cohérence n'est possible que si les règles satisfont une autre propriété, la propriété de la sous-formule, qui consiste en ceci que les prémisses d'une règle d'introduc-

[77] C'est-à-dire l'élimination locale d'une coupure sur deux occurrences d'une formule obtenues par les règles canoniques d'introduction (à droite et à gauche) de son expression dominante.

[78] La normalisation d'un système de règles nous assure que tout énoncé qui peut être dérivé de certaines hypothèses par application des règles d'introduction et d'élimination peut être dérivé sans faire de "détours", c'est-à-dire en commençant par une succession de règles d'élimination des symboles figurant dans les hypothèses, suivi d'une succession de règles d'introduction des symboles figurant dans la conclusion (en déduction naturelle). La normalisation nous assure donc qu'en faisant des "détours", en appliquant par exemple une règle d'introduction d'un symbole puis plus tard une règle d'élimination, on ajoute rien. Le théorème d'élimination des coupures est un résultat analogue en calcul des séquents, montrant que tout ce qui est dérivable dans le calcul peut être dérivé de façon « directe », uniquement par l'emploi des règles d'introduction canonique. Pour les définitions précises, je renvoie par exemple à Prawitz (1965) et aux références qu'ils contient. La possibilité de normaliser un calcul est parfois vue comme un autre critère d'harmonie globale du système. Je renvoie le lecteur à Dummett 1991.

tion sont toujours des sous-formules de leur conclusion.[79] Mais si l'on ajoute les règles naïves pour la vérité à un système, on perd cette propriété de la sous-formule qui permet de réduire pas à pas la complexité des formules maximales dans la démonstration de la normalisation. Le problème est que la règle d'introduction de Vr fait en général *décroître* la complexité des formules : $Vr(\ulcorner \exists x Vr(x) \urcorner)$ est atomique alors que $\exists x Vr(x)$ ne l'est pas. Vouloir adapter la notion de complexité d'une formule pour faire en sorte que les règles d'introduction de l'ensemble du système accroissent toujours la complexité des formules en ce nouveau sens est sans espoir quand le prédicat de vérité est « auto-applicatif » : il y a un conflit irréductible entre le fait que $Vr(\ulcorner \exists x Vr(x) \urcorner$ s'obtient de $\exists x Vr(x)$ par Vr-intro, et le fait que la seconde s'obtient de la première par $\exists$-intro.[80] Dans le cas où l'on restreint l'application du prédicat de vérité à des énoncés ne contenant pas de prédicat de vérité le problème disparaît, puisqu'alors on ne peut obtenir le premier énoncé par la règle d'introduction à partir du second.

De façon concomitante, les dérivations de contradictions dans un système de règles contenant les règles naïves pour le prédicat de vérité ne sont pas des preuves normalisables (ou sans coupures). Pour voir comment les choses se passent, considérons dérivation suivante en calcul des séquents, sous l'hypothèse que « λ » est un terme tel que $\lambda =$« $\neg Vr(\lambda)$ »[81] :

[79] Où la notion ordinaire syntaxique de sous-formule est étendue en sorte que les instances d'une formule universelle compte comme des sous-formules de la formule universelle elle-même, de sorte que les règles d'introduction accroissent toujours la complexité des formules

[80] Voir sur ce point les remarques de Kremer (1988), p. 260

[81] À strictement parler, cette dérivation sera simplement réfutation de l'hypothèse de départ, à savoir que $\lambda =$« $\neg Vr(\lambda)$ ». Mais d'une part l'idée que l'on pourrait réfuter *a priori* le fait que « λ » ne dénote pas l'énoncé « $\neg Vr(\lambda)$ » est bien absurde (pourquoi n'en aurais-je pas décidé ainsi ?) est donc authentiquement paradoxale même si elle n'est pas contradiction logique, et d'autre part on réfuterait l'arithmétique par un raisonnement analogue (avec des nom complexes des énoncés, cette fois, pour pouvoir prouver dans la syntaxe l'existence d'un énoncé équivalent à la négation de sa vérité.) La preuve en déduction naturelle montre un phénomène de « jeu » analogue avec la négation.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{Vr(\lambda) \vdash Vr(\lambda)}{\vdash \neg Vr(\lambda), Vr(\lambda)}\,{\scriptstyle \vdash \neg}}{\neg Vr(\lambda) \vdash \neg Vr(\lambda)}\,{\scriptstyle \neg \vdash}}{\neg Vr(\lambda) \vdash Vr(\langle\!\langle\, \neg Vr(\lambda)\, \rangle\!\rangle)}\,{\scriptstyle \vdash Vr}}{\neg Vr(\lambda) \vdash Vr(\lambda)}\,{\scriptstyle \text{subst. des identiques}}}{\neg Vr(\lambda), \neg Vr(\lambda) \vdash}\,{\scriptstyle \neg \vdash}}{\neg Vr(\lambda) \vdash}\,{\scriptstyle \text{contraction}} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{Vr(\lambda) \vdash Vr(\lambda)}{\neg Vr(\lambda), Vr(\lambda) \vdash}}{\neg Vr(\lambda) \vdash \neg Vr(\lambda)}}{Vr(\langle\!\langle\, \neg Vr(\lambda)\, \rangle\!\rangle) \vdash \neg Vr(\lambda)}\,{\scriptstyle \text{subst. des identiques}}}{Vr(\lambda) \vdash \neg Vr(\lambda)}}{\vdash \neg Vr(\lambda), \neg Vr(\lambda)}\,{\scriptstyle \text{contraction}}}{\vdash \neg Vr(\lambda)}}{\vdash}\,{\scriptstyle \text{cut}}$$

En suivant dans la dérivation le devenir, dans le sous-argument de gauche, de la première occurrence de la formule « $\neg Vr(\lambda)$ » (ligne 2), on s'aperçoit qu'elle est « devenue » la formule « $Vr(\lambda)$ » à la ligne 5 : uniquement avec des règles canoniques d'*introduction* d'expressions (et la substitution des identiques) nous sommes parvenu à faire « disparaître la négation », à transformer une formule « plus complexe » en formule « moins complexe ».

Une question naturelle, si l'on adopte comme Read une perspective strictement locale sur l'harmonie, est donc de se demander comment interpréter, d'un point de vue philosophique et non seulement technique, le double fait d'une part de l'harmonie locale des règles naïves pour la vérité et d'autre part de l'impossibilité de normaliser les preuves de contradictions utilisant des règles naïves pour la vérité. La réponse de Read serait de dire que les règles naïves pour la vérité fixent bien la signification d'une expression, contrairement à celle de « tonk », mais qu'il s'agit de la signification d'une expression contradictoire.[82]Le problème de cette réponse, c'est qu'elle rompt le lien qui avait été tissé depuis le début entre signification et validité : les inférences valides étaient telles en vertu de la signification des expressions qui y figurent, et l'on se proposait de rendre compte de cette signification par le rôle des expressions dans les inférences ; désormais il faut séparer les expressions qui ont une signification « acceptable » des autres. Et où trouvera-t-on le critère de démarcation recherché, sur quoi le fondera-t-on ?

Que l'on suive ou non Read sur cette voie, la question se pose de savoir si une approche inférentialiste de la signification, plus généralement, ne permet pas de mettre au jour quelque chose d'intéressant ou d'important sur les paradoxes de la vérité, quelque chose qu'une approche sémantique vériconditionnelle, avec son redoublement de la notion de vérité dans le métalangage de l'explication, rendrait

[82]La preuve de contradiction présentée à l'instant serait donc bien justifiée en vertu de la signification (contradictoire) de « vrai ».

moins facilement décelable, voire invisible à force d'obstination. Ces questions nous mèneraient trop loin, et j'en remets l'examen éventuel à un travail ultérieur.[83]

[83]Pour une approche différente de celle de Read de ce qu'essentiellement ce type de remarque sur l'impossibilité de normaliser les dérivations du type « Menteur » nous apprend sur les paradoxes de la vérité, on pourra également consulter Tennant (1982). C'est, à ma connaissance, le seule tentative du genre.

Conclusion

Au commencement de ce travail j'ai annoncé un projet : contribuer à une analyse épistémologique du rôle de la notion de vérité. Au premier chapitre j'ai montré quelles relations existaient entre un tel projet et le déflationnisme contemporain. Ce déflationnisme en matière de vérité apparaît à la jonction, d'une part, d'une réflexion sur la notion de vérité, réflexion dans laquelle l'intuition de la transparence de certains emplois du prédicat de vérité (« Il est vrai que la neige est blanche » vs. « La neige est blanche ») et de son caractère indispensable pour l'expression de certaines généralisations (« Toutes les conséquences logiques d'une proposition vraie sont vraies ») ont une part centrale, et d'autre part d'une réflexion sur la méthodologie des sciences inscrite dans la postérité de l'empirisme logique. Le déflationnisme apparaît alors comme une thèse sur le statut de la notion de vérité (et des notions sémantiques en général) dans les explications scientifiques, sur fond de naturalisme méthodologique et métaphysique. L'originalité de la thèse déflationniste est de tenter de concilier deux idées jusque-là opposées : la vérité serait irréductible à des propriétés naturelles et, néanmoins, la notion classique de vérité aurait sa place dans le langage de l'explication scientifique. C'est que, selon le déflationniste, une fois que son emploi a été correctement compris, il apparaît que la notion de vérité ne joue pas de rôle « substantiel » dans l'explication : elle n'y figure qu'à la façon d'un outil auxiliaire permettant l'*expression* de certaines vérités. Lorsqu'une propriété figurant dans une explication putative n'est pas de prime abord une propriété naturelle (qu'il s'agisse d'une propriété mathématique, sémantique, ou de n'importe quelle propriété de « haut niveau ») , le naturaliste métaphysique avait traditionnellement deux options : montrer que la propriété en question peut être réduite à des propriétés naturelles, ou montrer que l'on peut s'en dispenser pour expliquer tout ce qu'il y a à expliquer. Le déflationniste ouvre une troisième voie au naturaliste : montrer que les emplois du prédicat correspondant

ne sont pas des emplois dans lesquels la propriété correspondante joue un rôle explicatif ou, de façon plus brève et plus vague, que la propriété en question n'est pas « substantielle ».

Généralisée hors des sciences empiriques, la thèse que le prédicat de vérité ne joue pas de rôle substantiel dans les explications fut le point de départ de ce travail, la tâche que je m'assignais étant alors d'analyser plus spécialement le rôle des lois *a priori* de la vérité dans les explications. Au fil des chapitres, j'ai discuté et proposé des réponses à quelques-unes des questions qu'elle soulève. Comment distinguer les emplois explicatifs des emplois expressifs de la notion de vérité (Problème de la frontière) ? De quels usages une théorie ou une définition de la notion de vérité doit-elle rendre compte si le prédicat de vérité doit pouvoir jouer le rôle que le déflationniste lui assigne (Problème de l'usage) ? Une telle théorie est-elle possible, et si elle l'est, ne légitime-t-elle pas aussi des emplois explicatifs de la notion de vérité (Problème de la stabilité) ?

L'exploration de ce terrain ne pouvait que commencer par une présentation et une discussion du travail Tarski, ce que j'ai fait au chapitre 2. Tarski a en effet montré comment *définir* la notion de vérité pour un langage, quand cette définition est possible, et donné des indications cruciales sur ce que sont les principes qui doivent gouverner toute théorie axiomatique de la notion de vérité pour un langage donné lorsque une définition explicite pour ce langage n'est pas possible. J'ai clarifié la distribution des ressources théoriques mises en œuvre par Tarski en fonction des différents usages du prédicat de vérité dont elles permettent de rendre compte. Puis, au chapitre 3, j'ai discuté une réponse globale aux questions posées par la thèse déflationniste et mentionnées au paragraphe précédent. Cette réponse hostile au déflationnisme est présentée sous la forme d'un argument dont la substance, en termes vagues mais corrects, était la suivante : c'est parce que nous avons connaissance de certaines « lois de la vérité » que nous sommes en mesure de mener à bien certaines explications, en particulier que nous pouvons expliquer que certaines théories que nous acceptons sont cohérentes. J'ai repoussé cet argument en critiquant la réponse qu'il suppose au Problème de la frontière en termes de non-conservativité, et j'ai donné une première formulation du type spécial de processus à l'œuvre dans une preuve de cohérence par la vérité en terme de mise au jour d'un contenu implicitement accepté. Au chapitre 4, cette intuition était alors étayée et raffinée par l'examen d'un certain nombre de réflexions épistémologiques auxquelles ont donné lieu les phénomènes d'incomplétude chez Kurt Gödel , Solomon Feferman,

et Daniel Isaacson. Non seulement cette idée que la preuve de la cohérence d'une théorie par la vérité pouvait être vue un processus d'explicitation d'un contenu implicite se trouvait-elle confortée par cet examen, mais les réflexions de Gödel sur la distinction entre extensions intrinsèque et extension extrinsèque d'une théorie en termes des modalités de la reconnaissance de leur vérité fournissaient également les clés permettant de rendre cette idée plus précise au chapitre 6.

Une preuve de la cohérence d'une théorie où figurent parmi les hypothèses tous les axiomes de la théorie elle-même n'est pas de nature à renforcer notre confiance en la cohérence de la théorie au-delà de la confiance que nous avons en la théorie elle-même. Les preuves de cohérence d'une théorie, empirique ou non, que l'on obtient en mobilisant des lois *a priori* de la vérité sont de ce genre. Mais inversement, ces preuves montrent aussi qu'un sujet acceptant une théorie, s'il est justifié à le faire, est *a priori* justifié à accepter la cohérence de la théorie. Au chapitre 5, j'ai tenté d'explorer cette relation singulière entre acceptation d'une théorie et acceptation de sa cohérence sous un autre angle en montrant qu'un sujet acceptant une théorie possédait une justification pragmatique de la cohérence de cette théorie ne mobilisant aucune ressource conceptuelle liée à la notion de vérité ni les hypothèses de la théorie elle-même. Pour y parvenir, j'ai présenté une notion déontique et épistémique d'acceptation réfléchie, critique, délibérée, assumée. J'ai soutenu la thèse selon laquelle cette notion devait obéir à un principe de responsabilité : Si S accepte p, alors S accepte que p est acceptable. Sur cette base, et en réfléchissant sur les normes minimales d'acceptabilité, il semble que l'on puisse justifier notre acceptation de la cohérence d'une théorie que nous acceptons. Mais tandis que la preuve de cohérence par la vérité est une justification infaillible de la cohérence sous les hypothèses de la théorie, la justification pragmatique est une justification défaisable dont la prémisse principale est l'observation que j'accepte les axiomes de la théorie et dont le chaînon défaisable est l'étape justifiée par le principe de responsabilité. Au bout du compte, ce que montre cet *excursus*, c'est que si un sujet peut bien en un sens justifier son *acceptation* de la cohérence d'une théorie qu'il accepte, il n'en a pas pour autant produit quelque chose qui pourrait compter, d'après ses propres lumières, comme une *explication* de la cohérence de la théorie, ce que faisait bien la preuve par la vérité.

Au chapitre 6, j'ai présenté mes propres réponses aux questions laissées ouvertes par notre première formulation de la thèse déflationniste. Pour ce faire, j'ai d'abord cherché à montrer que la simple *formulation* de ce qui était accepté par un sujet

acceptant une théorie donnée, si cette formulation doit être accessible à un sujet fini, doit en passer par l'usage du concept de vérité. Les contraintes que font peser la finitude du sujet sur la formulation explicite d'une théorie A qu'il reconnaît accepter et dont il doit pouvoir juger, sont telles que cette formulation est logiquement plus forte que le seul ensemble des énoncés qui constituent A. Il apparaît alors que l'énoncé de la cohérence de A, s'il n'est pas conséquence de A, est une conséquence de cette formulation. Par conséquent, un sujet qui accepte une théorie A accepte *en fait* une théorie qui implique la cohérence de A. C'est ce que j'ai appelé un phénomène de conséquence réflexive : l'énoncé de la cohérence de A n'est pas conséquence logique de A, mais c'est une conséquence épistémologique, pour ainsi dire, des modalités de notre accès à A. Le développement de cette thèse m'a également permis de défendre la théorie minimale de la vérité, en tant que théorie dont la maîtrise est suffisante pour la compréhension et l'usage du concept de vérité, parce que le concept minimal est suffisant pour rendre de compte de cet usage du prédicat de vérité dans la formulation de généralisations. Quel statut donner alors aux lois tarskiennes de la vérité ? Quelle est la nature des justifications que nous avons pour ces lois ? Ma réponse a été de montrer que ces lois pouvaient être vues comme des formulations explicites, à l'aide du prédicat de vérité minimal, de ce qui est implicitement accepté par celui qui accepte les règles d'inférences logiques qui sous-tendent le langage de toute théorie. Autrement dit, c'est essentiellement parce que nous acceptons les lois de la logique (et que nous avons un prédicat de vérité minimal), que nous sommes justifiés à accepter les lois tarskiennes. Tirant les conséquences de ce point, j'ai suggéré que l'extension aléthique tarskienne d'une théorie donnée doit être comprise comme une explicitation des engagements que nous avons pris en acceptant la théorie de départ. Cette explicitation est le fruit d'un double mouvement. Un mouvement réflexif par lequel un agent constitue des objets de pensée en prenant du recul par rapport à sa propre activité discursive : ici la description des modèles généraux d'inférence que nous suivons exige l'emploi de ressources variées, en particulier de ressources syntaxiques dans le cas des inférences logiques, mais la notion de vérité n'y joue aucun rôle. Puis un second mouvement dans lequel le sujet réaffirme ce que la description de son modèle d'activité le montrait disposé à affirmer : mais si ici le prédicat de vérité est indispensable, c'est dans son rôle de moyen auxiliaire (avec d'autres, comme les quantificateurs) pour l'expression des généralisations.

Dans le dernier chapitre, la poursuite de l'idée que le prédicat de vérité est un

concept purement expressif donne finalement lieu l'exploration de la thèse selon laquelle la notion de vérité est une notion *logique*. Le projet est justifié par l'idée que les concepts logiques sont, dans une certaine tradition, conçus comme des concepts dépourvus de contenu. Après quelques considérations générales permettant de donner une assise intuitive à la discussion, cette thèse est discutée dans une perspective inférentialiste, en prenant pour acquis, sur la base de ce qui avait été présenté au chapitre précédent, que notre compréhension du concept de vérité est articulée par les équivalences-T, ou plus précisément par les simples règles de citation et de décitation. Après un rappel de tenants et aboutissants de l'analyse inférentialiste de la logicité, j'ai indiqué que le prédicat de vérité satisfait aux contraintes d'harmonie et de schématicité qui fondent le caractère logique d'une expression dans cette tradition. Puis, la même entreprise est reconduite avec les outils de la sémantique extensionnelle et d'analyse de la logicité développés originellement par Tarski. À nous en tenir à un certain usage du prédicat de vérité compris comme moyen de parler du monde « à travers » les énoncés, et après avoir dégagé un critère de logicité fondée en dernière instance sur la notion d'invariance, j'ai montré que l'on pouvait étayer la thèse de la logicité des emplois déflationnistes de la notion de vérité.

Les résultats de ce travail dans leur ensemble permettent de renforcer la cohérence de la position déflationniste. L'idée que le prédicat de vérité est logique a parfois été suggérée dans la littérature déflationniste sans jamais avoir été approfondie. En ce sens, ce travail peut être pris comme une explication possible de cette thèse, et une tentative d'en établir le bien-fondé. En second lieu, de nombreux déflationnistes ont suggéré que notre *compréhension* du prédicat de vérité était déterminée par notre acceptation des équivalences-T. Notre travail montre non seulement qu'une telle thèse est plausible, mais rend compte du sa relation à la thèse du caractère logique du prédicat de vérité. Enfin, notre travail contribue à l'effort déflationniste d'analyse du rôle de la vérité dans les explications. Nous nous sommes concentrés dans ce travail sur les lois *a priori* de la vérité, leur rôle dans les recherches logiques et mathématiques, et plus particulièrement sur le cas paradigmatique de l'emploi du prédicat de vérité dans les preuves de cohérence. On pourrait donc voir nos conclusions, si elles sont correctes, comme confortant, dans le domaine de la logique et des mathématiques, les thèses déflationnistes sur l'emploi de la notion de vérité dans les sciences naturelles. Vues sous cet angle, les deux thèses sont complémentaires mais indépendantes.

Annexe A

Logicité et Invariance

Note : Dans la dernière partie du chapitre 7, j'ai montré qu'ajouter un prédicat de vérité interprété à une logique donnée ne changeait pas la notion d'équivalence élémentaire associée à cette logique. Ce résultat permettait de donner une borne à l'expressivité du prédicat de vérité dans l'usage ciblé par le déflationniste. J'ai soutenu que ce résultat pouvait être vu comme une indication du caractère logique des emplois déflationnistes de la notion de vérité. Néanmoins cette approche sémantique de la logicité n'est pas usuelle. L'approche standard dans ce domaine est celle initiée par Tarski (1966/1986) en termes d'invariance. Dans cet appendice, je fais une présentation historique de l'approche de Tarski et de ses motivations, et j'explique le lien entre cette tradition et la proposition qui était la mienne au chapitre précédent.

Dans la première partie je présenterai le travail de Tarski sur les notions logiques, et en particulier la notion d'invariance, en insistant sur la façon dont elle permet de capturer les spécificités sémantiques des notions logiques. Dans une seconde partie, je présente certains prolongements contemporains du travail de Tarski dûs à Gila Sher, Solomon Feferman, Van McGee et Denis Bonnay. J'introduis une notion nouvelle mais implicite dans littérature, celle de complétude fonctionnelle d'une logique du premier ordre. Dans troisième partie, le prédicat de vérité entre à nouveau en scène, et je présente un résultat analogue, dans ce cadre, à celui obtenu dans le chapitre précédent.[1]

A.1 Tarski et le problème de la démarcation des constantes logiques

Dans son texte séminal sur le *Concept de conséquence logique*[2], Tarski avait une première fois été confronté au *problème de la démarcation* des constantes logiques, que nous formulerons simplement de la façon suivante, pour les références ultérieures :

Problème de la démarcation Caractériser les concepts logiques.

C'est qu'une solution à ce problème semblait requise pour que l'élucidation de la notion de conséquence *logique* fût complète. La définition de la conséquence logique proposée par Tarski, en effet, est la suivante :

> Nous disons que l'énoncé X *suit logiquement* des énoncés de la classe K si et seulement si tout modèle de la classe K est aussi un modèle de l'énoncé X (Tarski (1936/2009), p.93)

Il y a au cœur de cette définition la notion tarskienne de *modèle*[3], et la définition de cette notion *suppose* établie la distinction entre constantes logiques et constantes non-logiques :

> Examinons une classe quelconque d'énoncés L et *remplaçons toutes les constantes extra-logiques* [je souligne] des énoncés de la classe L par

[1] Ce travail est issu d'une recherche en cours conduite en collaboration avec Denis Bonnay et Julien Boyer. Qu'ils soient ici remerciés.

[2] Le texte a paru pour la première fois en polonais en 1936 et a été repris en anglais dans le volume des écrits de Tarski édité par Woodger Tarski (1956) Voir Tarski (1936/2009) pour une traduction française.

[3] Il ne faut pas confondre cette notion avec la notion de modèle aujourd'hui en usage en logique, qui en diffère sur des points importants.

> des variables correspondantes (les mêmes constantes étant remplacées par les mêmes variables, les constantes distinctes étant remplacée par des variables distinctes) ; nous obtenons alors une classe de fonctions propositionnelles L'. Toute suite quelconque d'objets qui satisfait toutes les fonctions propositionnelles de la classe L' sera appelée *modèle de la classe d'énoncés* L.(Tarski (1936/2009), p.93)

Pour illustrer la dépendance du contenu des concepts de modèle et de conséquence logique définis par Tarski au problème de la démarcation, considérons par exemple l'argument suivant :

$$\frac{\begin{array}{c}\text{Barbey est au moins aussi grand que Villiers}\\ \text{Villiers est au moins aussi grand que Huysmans}\end{array}}{\text{Barbey est au moins aussi grand que Huysmans}}$$

Si la relation « être au moins aussi grand que » est distinguée comme une relation logique alors, si l'explication de Tarski est correcte, l'argument précédent doit compter comme un cas de raisonnement *logiquement* valide. Sinon, l'argument est seulement matériellement correct, ou analytiquement valide.[4]

Tarski reconnaît l'importance du problème et admet l'existence d'une certaine distinction imprécise entre termes logiques et non logiques fondée dans l'intuition ; il reste cependant sceptique quant à la possibilité de présenter un critère de démarcation objectif :

> Sous-jacente à toute notre construction est la division de tous les termes des langages déjà étudiés, en logiques et extra logiques. Cette division n'est certainement pas entièrement arbitraire : si, par exemple, nous voulions ne pas compter parmi les signes logiques les signes d'implication et les quantificateurs, la définition donnée de la conséquence pourrait

[4]Nous avons à dessein choisi un exemple qui peut s'avérer troublant. On peut peut en effet avoir le sentiment que l'argument précédent possède une propriété très proche de celle de la validité logique, la construction « être au moins aussi X que » semblant forcer la transitivité. Et en effet le rendu suivant de l'argument en fait un argument logiquement valide dans sa forme propositionnelle :

$$\frac{\begin{array}{c}\text{Si Villiers est grand alors Barbey est grand}\\ \text{Si Huysmans est grand alors Villiers est grand}\end{array}}{\text{Si Huysmans est grand alors Barbey est grand}}$$

La notion de conséquence logique développée par Tarski est une notion de conséquence formelle. Supposé résolu le problème de la démarcation, un argument peut être logiquement valide au sens de Tarski, et un argument obtenu en substituant synonymes pour synonyme de façon non uniforme dans le précédent n'être pas logiquement valide.

> conduire à des conséquences contredisant clairement les intuitions courantes. D'un autre côté je ne connais aucune base objective qui nous permette de dresser une frontière nette entre ces deux catégories de termes. (Tarski (1936/2009), p.95.)
>
> D'autres recherches sans doute éclaireront le problème qui nous intéresse ; peut-être serait-il possible à l'aide d'arguments importants et objectifs, de justifier la limite imposée par la tradition entre termes logiques et extra-logiques. Mais je ne serais toutefois guère étonné si le résultat de ces recherches était franchement négatif et si nous étions alors forcé de considérer des concepts comme ceux de « conséquence logique », d'« énoncé analytique » et de « tautologie » comme des concepts relatifs qui doivent être rapportés à une division bien définie, quoique plus ou moins arbitraire, des termes logiques et extra-logiques. Ce caractère arbitraire de la division reflèterait naturellement, dans une certaine mesure, la fluctuation que l'on peut observer dans l'usage du concept de conséquence dans les discours de tous les jours.(Tarski (1936/2009), p.97)

Ces lignes témoignent de la double position deTarski sur problème de la démarcation à la fin du *Concept de conséquence logique*. D'une part un pluralisme assumé : il n'y a rien de tel qu'un unique usage du terme « expression logique » (ou de la notion de « conséquence logique »), et différentes démarcations peuvent être élaborées pour rendre compte de ces différents sens. D'autre part un scepticisme plus profond concernant la possibilité, pour une démarcation donnée, d'en fournir une explication plus intéressante que la simple énumération bipartite correspondante des termes du langage.[5] C'est sur ce second point que Tarski reviendra trente ans plus tard.

En mai 1966, Tarski donna à Londres une conférence intitulée « What are logical notions ? ». Sans abandonner l'idée qu'il n'y a pas *un* problème de la démarcation mais autant que de sens que le terme « logique » peut en recevoir dans l'usage, Tarski fait un pas décisif en avant en affirmant désormais que, du moins pour un certain sens de l'expression « notion logique », une solution conceptuellement fondée au problème de la démarcation est possible :

> Laissez-moi vous dire par avance qu'en répondant à la question « Que

[5] C'est de cette façon que nous comprenons la dernière partie de la dernière phrase citée dans le passage ci-dessus.

> sont les notions logiques ? » ce que je ferai est une suggestion ou une proposition à propos d'un usage possible du terme « notion logique ». Cette suggestion me semble en accord, sinon avec tout usage prévalent du terme « notion logique », du moins avec un usage qui est rencontré en pratique. Je pense que le terme est utilisé dans de nombreux sens différents et que ma suggestion rend compte de l'un d'eux.(Tarski (1966/1986), p.145)

Avant de présenter le travail de Tarski il nous faut affiner un peu la définition de ce que nous avons appelé le *problème de la démarcation* et distinguer deux problèmes voisins. Notre formulation du problème était faite en termes de la distinction entre deux catégories de *concepts*. Mais les conditions d'identité des concepts étant notoirement difficiles à élaborer, cette formulation du problème rend difficile sa résolution en termes rigoureux.[6] On pourrait donc vouloir distinguer deux problèmes. Le premier serait formulé en termes linguistiques :

Problème de la démarcation (version linguistique) :

Caractériser les *expressions* logiques.

Si, sous cette formulation, on peut envisager que des recherches syntaxiques puissent permettre de répondre au problème posé, il est clair néanmoins que, dans la perspective sémantique qui est la nôtre ici, le problème linguistique n'est qu'une simple reformulation du problème de départ : sous l'hypothèse que le caractère logique ou non d'une expression dépend de son *sens* uniquement alors ce problème se réduit au précédent. Mais l'on peut penser qu'indépendamment d'une caractérisation de la logicité des expressions, il existe une propriété de logicité qui se manifeste au niveau de la *dénotation* ou de l'*extension* des expressions. Ces extensions peuvent être des individus, des classes d'individus, des relations entre individus, des classes

[6] Quoique qu'il n'y ait pas de terminologie universellement acceptée, nous tenons ici pour acquis qu' concept est désigné ou associé à chaque expression, et que ses conditions d'individuation sont identiques à celles du sens de cette expression. Ainsi les concepts d'homme et de bipède sans plume sont distincts, mais également ceux de multiplication par deux et de multiplication par le quotient de quatre par deux. Les concepts, comme les propositions, sont donc non seulement des objets intensionnels, mais encore hyperintensionnels. Il existe un cadre logique approprié pour traiter des concepts individués de façon *simplement intensionnelle*, c'est celui des mondes possibles. L'intension d'un concept, dans cette acception du terme, est la fonction qui associe à chaque monde possible l'extension du concept dans ce monde. Ce cadre permettrait de distinguer *homme* de *bipède sans plume*. Néanmoins il ne permettrait pas la distinction entre *multiplication par deux* et *multiplication par le quotient de quatre par deux*. Dans ce qui suit nous nous bornerons à distinguer le problème de la démarcation des extensions logiques, qui sera l'objet de notre étude, du problème de la distinction des concepts logiques.

de classes d'individus etc. Ce sont ces objets extensionnels que Tarski appelle « notions logiques », et c'est le problème de la démarcation entre notions logiques et notions non-logique *en ce sens* auquel il propose une réponse. Si l'on identifie tout objet de la hiérarchie des types sur un domaine donné à sa fonction caractéristique (un opérateur), nous pouvons donc formuler le problème de Tarski ainsi :

Problème de la démarcation (version sémantique et extensionnelle) :

Caractériser les opérateurs logiques.

Cette distinction étant faite, comment les deux problèmes sont-ils liés ? Si une expression est logique, alors son extension doit l'être. Cette direction ne semble pas poser de problème particulier dès lors que l'on accepte la thèse, que Tarski va éloborer et défendre, selon laquelle il y a un sens à distinguer entre extensions logiques et extension non-logiques. Inversement, nous devons nous attendre à ce que des expressions possédant une extension logique ne soient intuitivement pas logiques[7]. Considérons par exemple le quantificateur $\forall^{\bullet}$, défini comme un prédicat de second ordre de la façon suivante :

$\forall^{\bullet}$ est vrai d'un prédicat ϕ si et seulement si
tout individu a la propriété ϕ et la neige est blanche.

$\forall^{\bullet}$ a la même extension que $\forall$ (si, comme c'est le cas, la neige est blanche dans l'univers du discours), et pourtant l'un est intuitivement logique et l'autre pas. Une solution à la version extensionnelle du problème de la démarcation donnera donc seulement une condition nécessaire de logicité des expressions. Mais existe-t-il une condition nécessaire et suffisante liant la solution du problème extensionnel de la démarcation au problème intensionnel ? Nous n'avons pas de réponse précise à cette question et notons simplement que la thèse suivante, formulée par McGee, est, de prime abord plausible :

Thèse de McGee :

Une expression est une expression logique si et seulement si il suit de la signification de l'expression que son extension est logique.

Nous adopterons cette thèse sans justification, et dans ce qui suit nous concentrerons notre attention sur le problème extensionnel de la démarcation.

Ceci étant précisé, nous présentons maintenant la solution de Tarski au problème (extensionnel) de la démarcation. La proposition de Tarski est d'étendre la méthode

[7] Voir également McGee (1996), p.578

développée dans le *programme d'Erlangen* de Klein pour la classification des géométries autour de la notion d'invariance et de l'appliquer à la logique elle-même. De quoi s'agit-il ? La multiplication des géométries au dix-neuvième siècle avait fait naître un besoin de clarification de leurs relations. L'idée de Klein fut de proposer une classification des géométries fondée sur une caractérisation leurs objets d'études respectives. Cette caractérisation, à son tour, était donnée en termes d'invariance : les notions d'une géométrie donnée sont caractérisées par le groupe des transformations de l'espace laissant invariantes les notions en question. Les notions caractéristiques de la géométrie euclidienne, « parallèle », « congruents », « homothétiques », etc. sont telles que les jugements associés de parallèlisme, congruences etc., conservent leur valeur de vérité sous certaines transformation de l'espace. Ainsi si deux droites sont parallèles, leur images seront parallèles après une translation, une rotation, ou une transformation homothétique de l'espace. Plus généralement, les notions de la géométrie euclidienne sont invariantes par la classes de similitudes.[8] Réciproquement, suivant la caractérisation de Kelin, toute notion invariante par similitude est en droit une notion qui appartient à la géométrie Euclidienne. La géométrie affine, et la topologie peuvent être caractérisée de façon analogue. La géométrie affine étudie les notions invariantes par la classes plus large des transformations affines, et la topologie les notions invariantes par la classe plus large encore des transformations continues de l'espace. Non seulement la notion d'invariance par un groupe de transformation permet-elle de caractériser ce dont une géométrie donnée est l'étude, mais elle permet également de classer ces géométries selon un ordre reflétant leur *généralité*. La groupe des isométries de l'espace étant un sous-groupe du groupe des transformations continues, la topologie apparaît comme une géométrie plus générale que la géométrie euclidienne. Plus la classe de transformations considérée est grande, plus la classe des notions invariantes par ces transformations est petite, et plus la géométrie dont les notions sont caractérisées par l'invariance par cette classe de transformation est générale. C'est à ce point que Tarski introduit sa proposition de caractérisation des notions logiques :

> Supposons maintenant que nous prolongions cette idée [de Klein], et considérions des classes de transformation encore plus grandes. Dans le cas extrême, nous considérerions la classe de *toutes* les permutations [*one-one transformation*] de l'espace, ou de l'univers du discours, ou du

[8] Les similitude sont les transformations qui préservent les rapports de distances.

> « monde », sur lui-même. Quelle sera la science qui s'occupe des notions invariantes par par cette plus grande classe de transformation ? Ici nous n'aurons que très peu de notions, toutes d'un caractère très général. Je suggère que ce sont les notions logiques, que nous appelions « logique » une notion si elle est invariante par toutes les permutations [*one-one transformation*] possibles du monde sur lui-même.(Tarski (1966/1986), p.149)

C'est ce que nous appellerons la thèse de Tarski :

Thèse de Tarski :

> Un opérateur est logique si et seulement si il est invariant par toute permutation des individus du domaine du discours.

L'idée que les notions logiques sont invariantes par permutations n'est pas entièrement nouvelle en 1966. Tarski et Lindenbaum avaient déjà montré en 1934 que toutes les notions des *Principia Mathematica* étaient invariantes par bijection[9]. Et dans Mostowski (1957)indexMostowski, Andrzej, où est discutée la notion de quantificateur généralisé, la demande que les notions logiques soient invariantes par bijection est explicite. Par ailleurs l'idée de *caractériser* les notions logiques par l'invariance par permutation n'est pas entièrement nouvelle non plus. Selon Steve Awodey, on trouve dans les notes manuscrites de Carnap des indications que ces idées étaient apparues lors de conversations entre Carnap, Tarski et Quine à Harvard dans les années quarante.[10] Surtout, une proposition de caractérisation de la logique très proche de celle de Tarski[11], fondée également sur l'extension du programme d'Erlangen, se trouvait déjà dans Mautner (1946).[12]

En quel sens le critère d'invariance de Tarski permet-il de capturer la notion de logicité ? Nous avons avons vu au chapitre précédent que la logique se distinguait d'un point de vue sémantique par trois propriétés : la généralité, l'insensibilité à l'identité des individus, et la formalité. Du point de vue de Tarski, le critère d'invariance doit permettre de capturer la *généralité* de la logique, comme l'indiquent clai-

[9] Tarski (1936/2009), Th. ?

[10] Nous nous appuyons ici sur une communication orale de Steve Awodey, qui a évoqué ces sources lors d'une conférence intitulée « Explicating 'Analytic' » donnée dans le cadre du colloque *Carnap's Ideal of Explication : Logic, Metalogic, and Wissenschaftslogik* organisé par Pierre Wagner à l'IHPST en Mai 2009.

[11] Mais dont Tarski ne semble pas avoir eu connaissance

[12] Bonnay (2006) chap. 3 contient un historique plus détaillé, précisant la nature des travaux de Mautner, ainsi que ceux du logicien franco-russe Krassner qui ont été redécouverts à cette occasion.

rement les passages cités plus haut et l'inspiration revendiquée du programme d'Erlangen. L'idée est que la logique est non seulement générale, mais la plus générale de toutes les sciences, et que cette généralité maximale peut être capturée par l'invariance des notions logiques par la plus grande classe de transformation, celle de toutes les permutations. Mais en même temps, quoique Tarski n'en dise rien, il est clair que le requisit d'invariance capture directement ce que nous avons appelé, au chapitre 7, l'insensibilité des notions logiques à l'identité des individus. Si une notion, vue comme une fonction d'un domaine d'objet dans l'ensemble {Vrai, Faux}, prend la même valeur pour tout argument, alors ce que capture cette notion n'est pas sensible à l'identité des objets en question. Enfin, notons que l'invariance par permutation semble également pouvoir prétendre capturer un aspect de ce que nous avons appelé la formalité de la logique. En effet, demander que des notions soient invariantes pour une certaine classe de transformation, c'est demander que ces notions ne soient pas sensibles aux changements d'identité des individus à l'intérieur de certaines classes, et c'est donc bien d'une certaine manière faire abstraction de certains contenus *spécifiques*. Mais il n'y a rien d'absurde, dès lors, à vouloir aller plus loin et soutenir dans le même mouvement une certaine idée de ce c'est, pour une notion, de faire abstraction de tout contenu, de toute relation aux objets : faire abstraction de tout contenu apparaît naturellement dans ce cadre comme un cas limite de l'abstraction de tout contenu spécifique. Demander que les notions logiques soient invariantes par permutation, c'est-à-dire par la plus grande classe de transformation, c'est alors demander que les notions logiques fassent abstraction complète de tout contenu, et qu'elles ne soient, en somme *à propos* de rien. Que cette idée soit acceptable, cela dépend de ce que nous entendons mettre sous la notion de « contenu ». Mais il n'y a rien d'absurde dans l'idée de considérer dialectiquement que l'invariance par permutation puisse nous permettre d'expliquer et de préciser l'idée qu'une notion puisse être dépourvue de tout contenu. C'est là cependant un pas que Tarski ne franchit pas : nulle part Tarski ne s'est engagé sur le caractère formel de la logique, en ce sens très fort de formalité.[13]

Après cette mise en perspective, voyons à présent un peu plus en détail comment fonctionne l'invariance par permutation.

[13]Il vaut de noter que nous parlons ici de *notion logique* et non d'*expression*. Il est évident que les expressions logiques ont un contenu, pour Tarski, au sens trivial qu'elles ont une signification, laquelle est représentée par des classes ou des fonctions appropriées qu'il appelle, justement, « les notions logiques ».

A.2 Invariance

L'invariance par permutation

Dans cette section nous allons nous donner des définitions exactes des outils dont nous aurons besoin par la suite, ainsi que quelques éléments d'illustration du critère d'invariance de Tarski. Nous suivons pour les définitions la présentation de Benthem (1989). Nous commençons par définir un ensemble de types.

Définition 22 (Types)**.** *L'ensemble des types est défini par induction :*

1. *e est un type*
2. *t est un type*
3. *si τ et τ' sont des types, alors (τ, τ') est un type.*

Selon l'interprétation que nous avons en tête, sur un domaine de base donné, chaque objet, classe d'objets, fonction etc. appartient à un domaine d'un certain type, e étant le type des individus, t le type des valeurs de vérité, (e, t) le type de l'interprétaion des prédicats unaires, $(e, (e, t)$ le types de l'interprétation des prédicats binaires, $((e, t), t)$ le type de l'interprétation des prédicats de second ordre etc. On peut définir la hiérarchie des domaines par induction :

Définition 23. *La hiérarchie des domaines typés sur un domaine de base D est définie par induction à partir de deux domaines de base :*

1. *$D_e = D$ de type e est l'ensemble.*
2. *$D_t = \{V, F\}$ de type t est l'ensemble $\{V, F\}$ des booléens*
3. *Si D_τ et $D_{\tau'}$ sont deux domaines de types τ et τ', $D_{(\tau,\tau')} = D_{\tau'}^{D_\tau}$, l'ensemble des fonctions de D_τ dans $D_{\tau'}$ est un domaine de type (τ, τ')*

Soit maintenant un domaine d'individu D par exemple $D = \{A, B, C\}$, et π une permutation des individus de D définie, disons, de la façon suivante : $\pi(A) = B$, $\pi(B) = C$, et $\pi(C) = A$. À chaque chaque objet de type τ dans la hiérarchie des domaines des types est naturellement associée son image par la *transformation induite* par π sur le domaine D_τ. Dans notre petit exemple, considérons un objet de type (e, t), c'est-à-dire une fonction unaire prenant ses arguments dans le domaine des individus et à valeur dans le domaine des booléens, et voyons quelle est son image par la transformation associée à π. Soit $f_{\{A,B\}}$ la fonction qui aux individus

A, B associe la valeur V, et qui associe la valeur F aux autres individus. L'image de $f_{\{A,B\}}$ par la transformation induite par π sur $D_{(e,t)}$ est simplement la fonction $f_{\{\pi(A),\pi(B)\}}$, c'est-à-dire $f_{\{B,C\}}$. On notera $\pi_{(e,t)}(f_{\{A,B\}}) = f_{\{B,C\}}$. On pourrait procéder de façon analogue pour les autres types, avec cette remarque que la permutation des individus n'a pas d'incidence sur le domaine booléen. Pour référence ultérieure, nous définissons en toute généralité les permutations associées à π sur les domaines typés de la façon suivante :

Définition 24 (Permutation sur les domaines typés)**.** *Soit π une permutation sur D, on définit les permutations π_τ pour chaque type τ par induction :*

1. $\pi_e = \pi$
2. $\pi_t = Id_t$
3. *Si τ et τ' sont deux types et π_τ et $\pi_{\tau'}$ les permutations associées, on définit $\pi_{(\tau,\tau')}$, pour tout objet O de type (τ, τ') en posant $\pi_{(\tau,\tau')}(O) = \{(\pi_\tau(x), \pi_{\tau'}(y) : (x, y) \in O\}$*

Nous sommes maintenant en mesure de définir précisément l'invariance par permutation d'un opérateur du D, au sens standard qui est aussi celui de Tarski.

Définition 25 (Invariance par permutation)**.** *Soit D un domaine, τ un type et $O \in D_\tau$, O est invariant par permutation si et seulement si, pour toute permutation π sur D, $\pi_\tau(O) = O$.*

Equipés de cette définition, regardons à présent comment fonctionne le critère de Tarski sur un exemple simple. Considérons à nouveau le domaine D à trois éléments $D = \{A, B, C\}$. On peut facilement donner sous forme de tableau le graphe de toutes les permutations possible du domaine.

D	π_1	π_2	π_3	π_4	π_5	π_6
A	A	A	B	B	C	C
B	B	C	A	C	B	A
C	C	B	C	A	A	B

Supposons maintenant de surcroît que A et B soient des chevaux et que C ne soit pas un cheval. L'interprétation de « être un cheval » sur le domaine D est donc une fonction qui aux individus A et B associe la valeur V (Vrai) et qui associe la valeur F (Faux) à l'individu C. Autrement dit $Cheval(x)$ est une fonction dont le graphe

est $\{(A,V),(B,V),(C,F)\}$. Cet opérateur est-il invariant par permutation ? Non. Pour le voir, demandons-nous quelle est l'image de $Cheval(x)$ par la permutation associée à π_2. Par définition

$$\pi_{2,(e,t)}(Cheval) = \{(\pi_2(A), Id(V)), (\pi_2(B), Id(V)), (\pi_2(C), Id(F))\}$$

Donc $\pi_{2,(e,t)}(Cheval) = \{(A,V),(C,V),(B,F)\} \neq Cheval$. *Cheval* n'étant pas pas invariant par permutation, il n'est pas déclaré logique par le critère de Tarski. Une autre façon de dire les choses est que *Cheval* distingue entre les individus, ou est sensible à l'identité des individus, ou encore n'est pas formel.

Voyons maintenant le cas d'un quantificateur familier, disons le quantificateur universel, $\forall$. Suivant l'analyse Frégéene, qui permet d'associer une dénotation (fonctionnelle) aux quantificateurs, $\forall$ est un prédicat de prédicat, ou de façon équivalente, une fonction à valeur dans les booléens qui prend pour argument des fonctions à valeur dans les booléens prenant elles-même leurs arguments dans le domaine de base. Autrement dit (l'interprétation de) $\forall$ est un objet de type $((e,t),t)$. En identifiant chaque fonction de type (e,t) à l'ensemble d'individus dont elle est la fonction caractéristique on peut alléger un peu les notations et définir le quantificateur universel sur D comme la fonction de l'ensemble des parties de D à valeur dans les booléens telle que :

$$\forall(A) = 1 \text{ si et seulement si } A = D$$

Cet opérateur est-il invariant par permutation ? Pour le voir il suffit de noter que, étant donné une permutation π sur le domaine des individus, l'image de D lui-même par la permutation associée sur le domaine (e,t) n'est autre que D lui-même. Donc finalement pour $\pi_{((e,t),t)}(\forall) = \forall$. Si une partie du domaine D contient tous les éléments de D alors alors son image après une permutation arbitraire des individus est toujours le domaine D tout entier. On pourrait de même montrer que le quantificateur $\exists$ sur D est invariant par permutation en remarquant que si une partie de D est non-vide, alors son image après une permutation des individus est toujours un ensemble non-vide.

Le fait que le critère d'invariance par permutation déclare que *Cheval* n'est pas logique mais que $\exists$ et $\forall$ sont logiques n'est pas un artefact lié à la différence d'ordre entre les (types des) opérateurs.[14] On peut montrer par exemple que l'égalité, qui

[14] Voir la définition d'ordre d'un opérateur ci-dessous.

est une relation du premier ordre, est logique. L'identité sur D est (identifiée à) la relation : $Id = \{(A, A), (B, B), (C, C)\}$. Mais pour toute permutation π sur les individus, on se convainc facilement que $\{(\pi(A), \pi(A)), (\pi(B), \pi(B)), (\pi(C), \pi(C))\} = Id$. Et inversemement, on peut trouver des propriétés de second ordre (de même type par exemple que les quantificateurs) et qui ne sont pas invariantes par permutation. Dans notre exemple précédent, supposons à cet effet que les objets A et C sont bleus et que B est rouge. Et considerons la propriété de second ordre « être un couleur », que nous noterons $Couleur$ et qui est vraie (de l'interprétation) d'un prédicat de premier ordre si et seulement si ce dernier est (l'interprétation d') un prédicat de couleur. Dans notre exemple il est facile de voir que, pour toute partie P de D,

$$Couleur(P) = 1 \text{ si et seulement si } P = \{A, C\} \text{ ou } P = \{B\}$$

Mais couleur n'est pas invariant par permutation. En effet, considérons la permutation associée à π_2 sur le type des prédicats de second ordre :

$$\pi_{2,((e,t),t)}(Couleur)(P) = 1$$
$$\text{si et seulement si}$$
$$P = \{\pi_2(A), \pi_2(C)\} \text{ ou si } P = \{\pi_2(B)\},$$

autrement dit, si et seulement si $P = \{A, B\}$ ou $P = \{C\}$. Donc $Couleur \neq \pi_{2,((e,t),t)}(Couleur)$, et « être une couleur » n'est pas une notion logique.

Remarquons enfin que dans le cadre typé que nous avons choisi, toutes les fonctions de vérité ($\vee$, $\wedge$, $\neg$, $\rightarrow$) sont invariantes par permutations des individus du domaine et donc logiques.[15] Nous voyons sur ces quelques exemples comment fonctionne le critère d'invariance et constatons qu'il donne, sur les quelques cas simples traités, des réponses en accord avec nos intuitions.

Nous concluons maintenant cette cette sous-section avec la définition de la notion d'ordre d'un type, qui nous servira à définir plus précisément les quantificateurs par la suite :

Définition 26 (Ordre d'un type)**.** *Par induction :*

1. e et t sont d'ordre 0

[15] La question de savoir si l'invariance par permutation telle que nous l'avons définie est un bon critère de logicité pour les connecteurs propositionnels, ou s'il ne faut pas introduire une certaine notion d'invariance par permutation des objets de type t, est discutée dans MacFarlane (2000), chap. 6, pp. 207 et suivantes, et dans Bonnay (2006), chap 3. pp. 122 et suivantes.

2. *Si τ est de la forme $(\tau_1, (\tau_2, (\cdots, (\tau_{n-1}, \tau_n) \cdots)))$, alors l'ordre de τ est égal à $max(\tau_1, \cdots, \tau_n) + 1$.*

La définition fait ce que l'on en attend intuitivement, comme l'on peut le voir les exemples suivants :

Exemple 2. *Quelques exemples :*

- *(e, t), le type des prédicats unaires, est d'ordre 1*
- *$(e, (e, t))$, le type des prédicats binaires, est d'ordre 1.*
- *$((e, t), t)$, le type des prédicats unaires de prédicats unaires est d'ordre 2.*
- *$((e, t), ((e, t), t)))$, le type des prédicats binaires de prédicats unaires, est d'ordre 2.*

Par extension on se permettra de parler de l'ordre d'un opérateur plutôt que de l'ordre de son type. Notons que, du fait des conventions prises, le quantificateur universel dit ordinairement « du premier ordre » $\forall$ est ici un opérateur d'ordre 2, et le quantificateur universel dit du second ordre est ici d'ordre 3.

Quantificateurs généralisés

Le critère de Tarski ne permet de tester la logicité que des notions définies de façon très restreinte, à savoir les opérateurs définis sur un seul domaine. Autrement dit, la notion de « notion logique » utilisée par Tarski est celle de « notion logique relativement à un domaine D ». Mais ce traitement est en porte-à-faux avec le traitement sémantique (modèle-théorique) moderne des constantes logiques. Pour pouvoir définir de façon uniforme la notion générale de satisfaction dans une structure, les constantes logiques doivent être définies sur *tous* les domaines. D'autre part nous avons conceptuellement à gagner à adopter le traitement moderne.

Pour voir comment le traitement moderne des quantificateurs diffère du traitement de Tarski, il suffit de considérer la définition standard de la vérité dans un modèle. Cette définition comporte un clause pour chaque quantificateur, par exemple :

$$\mathcal{M} \vDash \forall x \phi[\sigma] \Leftrightarrow \text{Pour toute x-variante } \sigma' \text{ de l'assignation } \sigma : \mathcal{M} \vDash \phi(x)[\sigma']$$

Pour que la définition de satisfaction d'une formule dans n'importe quelle structure d'interprétation ait un sens, il faut que le quantificateur $\forall$ soit défini non pas sur un seul domaine D, mais sur toutes les structures. L'opérateur que Tarski associe

au quantificateur universel n'est en fait, du point de vue modèle-théorétique moderne, que la dénotation $\forall_D$ du quantificateur universel sur le domaine particulier D. Par conséquent, en abandonnant cette restriction, nous suivrons la conception moderne (modèle théorique) des quantificateurs et nous nous éloignons de la doctrine tarkienne. Mais nous avons également annoncé tirer un bénéfice conceptuel de l'adoption du point le vue modèle-théorique moderne selon lequel les quantificateurs sont définis sur toutes les structures, un bénéfice qui n'est pas sans rappeler celui qu'on obtient en modifiant la notion tarskienne de conséquence pour permettre la variation du domaine. En effet rappelons que l'idée qui motive l'emploi du concept d'invariance est que ce dernier doit permettre de capturer des notions maximalement générale, c'est-à-dire aussi suffisamment abstraites pour n'être pas sensible à l'identité des individus, ou à un contenu spécifique. Mais si par accident tous les individus d'un domaine donné sont des chats, alors l'interprétation *Chat* de l'expression « Chat » dans le domaine en question se trouvera être invariante par permutation, et donc logique selon le critère de Tarski. Considérer des opérateurs définis sur toutes les structures permet de renforcer la puissance de notre outil : une généralisation immédiate du critère de Tarski est maintenant de requérir d'une notion logique définie à travers tous les domaines qu'elle soit invariante par *bijection* à travers les domaines. Comme il y a des structures dans lesquelles tous les individus ne sont pas des chats, le critère d'invariance par bijection nous assure que la fonction qui caractérise l'ensemble des chats dans chaque domaine n'est pas logique selon le nouveau critère. L"invariance par bijection, qui prolonge naturellement le critère de Tarski dans le cadre d'une sémantique modèle-théorique des quantificateurs, a été mis en avant par Sher (1991) comme critère nécessaire de logicité :

Thèse de Tarski-Sher Les notions logiques sont les notions invariantes par bijection.[16]

Pour préciser un peu les choses nous donnons maintenant la définition exacte que nous suivrons pour la notion de quantificateur. Cette définition nous permettra en même temps d'introduire la notion de quantificateur généralisé. Nous commençons par la définition introduite par Mostowski (1957), pour laquelle nous avons d'abord besoin de la notion d'extension d'une formule dans une structure. L'extension d'une formule à une variable libre dans une structure est simplement

[16]Nous simplifions un peu les choses ici. Sher ajoute en effet d'autres contraintes, en particulier une contrainte d'ordre : les notions logiques sont au plus d'ordre 2.

l'ensemble des éléments du domaine de la structure qui satisfont la formule dans $\mathcal{M}$:

Définition 27 (Extension d'une formule dans une structure)**.** *Pour une formule $\psi = \psi(x, y_1, \cdots, y_n) = \psi(x = \overline{y})$ et une suite $\overline{a}$ de n objets dans la structure $\mathcal{M}$ on définit : $\psi(x, \overline{a})^{\mathcal{M},x} = \{b \in M : \mathcal{M} \vDash \psi(b, \overline{a})\}$*

On pourrait généraliser cette définition pour qu'elle s'applique aux formules à plusieurs variables libres et aux énoncés, mais c'est une peine que nous ne nous donnerons pas ici.[17] Muni cette définition nous pouvons alors définir les quantificateurs de Mostowki :

Définition 28. *Le symbole $\overline{Q}$ est un quantificateur si et seulement si son interprétation Q_M sur chaque domaine M est une fonction de type $((e,t),t)$ ou, de façon équivalente, un ensemble de sous-ensembles de M. Sa signification est donnée par :*

$$\mathcal{M} \vDash \overline{Q}x\phi(x, \overline{a}) \Leftrightarrow \psi(x, \overline{a})^{\mathcal{M},x} \in Q_M$$[18]

ou de façon équivalente,

$$\mathcal{M} \vDash \overline{Q}x\phi(x, \overline{a}) \Leftrightarrow \langle M, \psi(x, \overline{a})^{\mathcal{M},x} \rangle \in Q$$

Comme nous avons déjà eu l'occasion de le mentionner, Mostowski demandait de surcroît, pour qu'un quantificateur soit *logique*, qu'il soit clos par isomorphisme, au sens suivant, équivalent à la demande d'invariance par bijection :

Définition 29. *Un quantificateur est clos par isomorphisme si et seulement si pour tout univers M, M', et tout $A \subseteq M$, $A' \subseteq M'$: si $|A| = |A'|$ et $|M - A| = |M' - A'|$, alors $A \in Q_M \Leftrightarrow A' \in Q_{M'}$*

On peut se convaincre que $\exists$, interprété dans tout domaine D par $\exists_D : \{A \subseteq D : A \neq \emptyset\}$ est un quantificateur logique au sens de Tarski-Sher et au sens de Mostowski. En revanche le quantificateur $Q^\bullet$ défini sur chaque domaine par : $Q^\bullet_M = \{A \subseteq M : \text{Le Pape Benoît XVI } \in A\}$ n'est pas logique.

[17] Voir par exemple Peters et Westerståhl (2006), p.56.

[18] A nouveau, nous avons identifié les fonctions aux ensembles dont elles sont les fonctions caracteristiques pour facilité la lisibilité.

Une dernière généralisation de la notion de quantificateur est proposée par Lindström (1966), qui donne lieu à ce que l'on appelle aujourd'hui les *quantificateurs généralisés*, qui subsument par exemple la notion de déterminants (objets de type $((e,t),((e,t),t))$). La généralisation, par rapport aux quantificateurs de Mostowski, consiste maintenant à regarder tout opérateur d'ordre 2, quel que soit son type, comme un quantificateur. En plus des quantificateurs unaires familiers de la logique « du premier ordre » ($\forall$, $\exists$), et des quantificateurs généralisés unaires introduits par Mostowski, sont maintenant introduits les quantificateurs *polyadiques* dont les expressions des langues naturelle comme « la plupart », « plus de la moitié de » sont des exemples simples bien connus des linguistes.

Notons enfin qu'un quantificateur (au sens généralisé) peut simplement être identifié à une classe de structures, comme il apparaît déjà dans la dernière partie de notre définition 7, ce qui permet de simplifier un peu les notations. Par exemple l'opération associée à « $\exists$ » est simplement la classe $Q_{\exists}$ de structures $\langle M, P\rangle$ telles que P n'est pas vide. La relation avec la clause de satisfaction habituelle pour $\exists x$ est la suivante (avec $\mathcal{M}$ une structure quelconque et τ une assignation abitraire sur $\mathcal{M}$) :

$\mathcal{M} \vDash \exists x\ \phi(x)\ \tau$
si et seulement si
il existe $a \in M$ tel que $\mathcal{M} \vDash\ \phi(x)\ \tau[x := a]$
si et seulement si
$\langle M, \psi(x)^{\mathcal{M},x}\rangle \in Q_{\exists}$

Un quantificateur de type $((e,t),t)$ sera identifié à une classe de structures de la forme $\langle M, A\rangle$, où A est un sous-ensemble de M, tandis qu'un quantificateur binaire de type $((e,t),((e,t),t))$ sera identifié à une classe de structures de la forme $\langle M, R\rangle$ où R est une relation sur M, et ainsi de suite. En fait, étant donné la généralité de la notion de quantificateurs introduite par Lindström, *toute classe* de structures d'un signature donnée peut être vue comme un quantificateur.

Avec ces nouvelles notations, nous pouvons reformuler les critères d'invariance par permutation et d'invariance par bijection de la façon suivante[19] :

Invariance par permutation Q est invariant par permutation ssi pour tout domaine M et toute bijection f de M dans lui-même, et tout sous-ensemble A de M :

[19]Nous nous limitons au cas de quantificateurs monadiques pour ne pas trop alourdir la notation

$$\langle M, A\rangle \in Q \leftrightarrow \langle M, f[A]\rangle \in Q$$

où $f[A]$ est l'image de A par la fonction f.

Invariance par bijection Q est invariant par bijection ssi pour tous domaines M et M' et toute bijection f de M dans M', pour tout sous-ensemble A de M :

$$\langle M, A\rangle \in Q \leftrightarrow \langle M', f[A]\rangle \in Q$$

Notons ici que l'invariance est aussi appelée *invariance par Isomorphisme*.[20]

Pour terminer cette section, nous introduisons finalement la notion de *définissabilité* d'un quantificateur généralisé dans un langage. Considérons un langage logique contenant pour seuls symbole de quantificateurs le symbole du quantificateur universel $\overline{\forall}$ et un symbole pour quantificateur généralisé unaire $\overline{Q}$, supposé dénoter les quantificateurs $\forall$ et Q respectivement . Soit σ une signature, $\sigma = \{R\}$ où R est un symbole de relation binaire. L'énoncé de L_σ

$$Qx\forall yRxy$$

définit une classe de structures, nommément la classe des σ-structures $\mathcal{M}$ telles que

$$\mathcal{M} \vDash Qx\forall yRxy$$

Mais par définition cette classe de structures définit un quantificateur *binaire*

$$Q' = \{(M, R) : \langle M, R\rangle \vDash Qx\forall yRxy\}$$

Si nous notons $\overline{Q'}$ le symbole dénotant Q', et que nous l'ajoutons à notre stock de constantes logiques, nous obtenons un langage logique enrichit. La clause de satisfaction associée au nouveau symbole de quantification, pour toute structure $\mathcal{M}$ (de n'importe quelle signature cette fois), est alors la suivante :

$$\mathcal{M} \vDash \overline{Q'}xy\phi(x, y) \Leftrightarrow \langle M, \phi(x, y)^{\mathcal{M},x,y}\rangle \in Q'$$

Remarquons que Q' a été défini à partir des seuls opérateurs logiques de L, Q et $\forall$, le symbole R apparaissant de façon non interprété dans la formule définissant Q'. Si $\forall$ et Q méritent le noms de constantes logiques, alors nous dirons dans ce cas que Q' a été défini de façon purement logique. La thèse que, dans ces conditions Q' doit également être logique est la thèse de *clôture par définissabilité*, à laquelle nous allons venir.

[20]Raison : dire que $\langle M, A\rangle$ est isomorphe à $\langle M', A'\rangle$ c'est exactement dire qu'il existe une bijection f de M dans M' telle que $A' = f[A]$.

Se ressembler, d'un point de vue logique

La question se pose de savoir si l'intuition de la généralité et l'intuition de la formalité sont correctement capturées par le critère d'invariance. Nous sommes passé de l'invariance à l'invariance pour capturer une plus grande généralité, mais ne devrait-on pas aller encore plus loin ? C'est ce que nous voulons voir maintenant. Notre intérêt pour cette question est double : d'abord, la caractérisation de la logique donnée par le critère de Tarski-Sher est sujette à controverse. En effet, toutes les notions mathématiques apparaissent comme logiques sous ce critère. Malgré le théorème de McGee (1996), à l'effet que

Théorème 4. *Q est logique au sens de Tarski-Sher si et seulement si pour chaque domaine M, Q_M est définissable dans $L_{\infty,\infty}$*

le sentiment que le critère surgénère est fort.[21] Il est donc important que le résultat que nous allons présenter ne lui soit pas lié de façon essentielle. Si le prédicat de vérité apparaissait comme logique seulement pour *cette* délimitation du domaine de la logique, on serait fondé à contester la portée conceptuelle de notre résultat. Notre deuxième motivation tient au fait qu'en discutant des moyens de généraliser un peu la notion d'invariance nous allons nous donner quelques outils conceptuels dont nous allons avoir besoin par la suite pour élaborer notre argument pour la logicité du concept de vérité.

L'idée qui présidait au choix de l'invariance par la « plus grande classe de transformations » était d'introduire la plus grande variété d'états possibles de l'univers du discours pour n'en retenir que les notions les plus générales. Certes, l'insensibilité des notions logiques à l'identité des individus rend nécessaire leur invariance par permutation et même par bijection. Mais l'invariance par bijection est-elle suffisante pour qu'une notion soit logique ? Notons que l'insensibilité des notions logiques à l'identité des individus n'a pas été présentée comme une condition suffisante de logicité par Tarski, lequel motivait l'invariance par permutation par la thèse de la généralité de la logique. Nous avons vu que l'invariance par permutation devait être renforcée en invariance par bijection pour faire droit à l'insensibilité des

[21] On peut en effet douter que la définissabilité d'une notion dans la logique $L_{\infty,\infty}$ soit une indication suffisante de son caractère logique, indication qui viendrait en somme confirmer le bien-fondé du critère de Tarski-Sher. Au contraire, les logiques fortement infinitaires comme $L_{\infty,\infty}$, dont les formules peuvent contenir une infinité arbitrairement grande de variables, de quantificateurs, ou de connecteurs, sont généralement considérées comme porteuses d'un contenu mathématique ensembliste fort, qui excède la capacité d'expression qui est celui de la logique en propre.

notions logiques à l'identité des individus, mais il est clair que l'invariance par bijection préserve encore des propriétés structurelles non-triviales, comme par exemple les propriétés de cardinalité du domaine. En effet, une condition nécessaire pour qu'existe une bijection entre deux structures est qu'elles aient même cardinalité. Par conséquent, tous les quantificateurs de cardinalité sont invariants par bijection, y compris des quantificateurs comme « Il existe $\aleph_3$ » qui de prime abord ne sont pas logiques mais relèvent plutôt d'un outillage *ensembliste* sophistiqué pour parler du monde. Avons-nous des raisons de principes de considérer ces propriétés structurelles fortes comme des propriétés logiques ? La seule raison que nous avons de le faire, à ce stade, c'est qu'elles sont les propriétés invariantes par isomorphisme, et que l'invariance par isomorphisme est la notion supposée capturer la généralité de la logique. Mais sommes-nous réellement tenus par l'analyse conceptuelle de la généralité en terme d'invariance par isomorphisme ?

Ce sont ces questions qui motivent une dernière généralisation du cadre de travail sur l'invariance dont nous allons avoir besoin par la suite. Le premier pas de cette réflexion consiste à s'interroger sur la primauté de la notion d'isomorphisme dans notre projet. Après tout, la notion d'isomorphisme n'est qu'*une* certaine notion de *ressemblance* entre structures, parmi d'autres possibles. En logique et en mathématiques bien d'autres notions de ressemblances entre structures sont souvent à l'œuvre et d'une importance conceptuelle cruciale. En algèbre, les différentes notions de morphisme (morphisme de groupe, d'anneaux etc) sont des objets centraux permettant de capturer l'idée d'une certaine ressemblance structurelle. En logique mathématique, la notion d'équivalence élémentaire entre structures associée à chaque logique L est également une notion conceptuellement privilégiée de « ressemblance » entre structures. Clairement, ces différentes notions de ressemblances entre structures préservent *moins* de propriétés structurelles que la notion de ressemblance associée aux isomorphismes. En particulier, les homomorphismes ou l'équivalence élémentaire en logique du premier ordre ne préservent pas certaines propriétés de cardinalité du domaine. Du point de vue que nous venons d'adopter, l'invariance par isomorphisme n'est qu'un cas particulier d'invariance relativement à une certaine notion de ressemblance entre structures. Mais une fois dégagée la notion de ressemblance entre structures derrière les contraintes spéciales que faisait peser sur elle la notion d'isomorphisme, cette dernière semble perdre son statut privilégié. En particulier, la relation d'isomorphisme n'est certainement pas la relation de ressemblance la plus générale, le mérite en revenant indiscutablement désormais,

en l'absence de toute contrainte supplémentaire sur la relation de ressemblance, à la relation universelle !

La notion technique de *relation de similarité* entre structures a été définie par Feferman[22] pour capturer précisément une certaine idée de ressemblance entre structures. En adaptant les notations, et en généralisant un peu[23] nous pouvons définir simplement une *relation de similarité*, S, comme une relation entre structures respectant les signatures et nous notons $\mathcal{M}S\mathcal{M}'$ si $\mathcal{M}$ et $\mathcal{M}'$ sont similaires. L'invariance d'une notion relativement à une relation de similarité S quelconque peut maintenant être définie en adaptant la notion d'invariance par isomorphisme :

Définition 30. *Une opération Q est S-invariante si et seulement si pour toutes structures $\mathcal{M}$, $\mathcal{M}'$: si $\mathcal{M}S\mathcal{M}'$ alors $\mathcal{M} \in Q$ si et seulement si $\mathcal{M}' \in Q$.*

Nous notons $Inv(S)$ la classe des opérations S-invariantes.

Une fois ce cadre défini une voie de recherche de style fondationnel s'ouvre naturellement, cherchant à répondre à la question « Qu'est-ce que la (vraie) logique ? », autrement dit : à quelles conditions sur les relations de similarité la classe des notions invariantes par similarité sont-elles les notions logiques ? C'est la question qui a intéressé les philosophes jusqu'ici : Gila Sher [24] soutient que c'est la notion d'isomorphisme, Solomon Feferman [25] plaide pour la relation d'homomorphisme fort [26], Enrique Casanovas[27] pour la relation d'homomorphisme, Denis Bonnay [28] pour la relation d'isomorphisme potentiel etc.[29] Ce n'est pas la voie que nous suivrons. Ce que nous voulons faire, c'est revenir à l'esprit du programme d'Erlangen, un esprit de classification. Plutôt que de demander quelle est la « vraie » logique, je demanderai comment distinguer deux logiques. Et c'est la possibilité qui est ouverte par ces travaux de donner *des conditions d'individuation sémantique* d'une logique donnée qui va nous intéresser.

[22] Voir Feferman (1999*b*), p.39.
[23] Voir Bonnay (2006).
[24] Sher (1991).
[25] Feferman (1999*b*).
[26] Avec quelques restrictions supplémentaires
[27] Casanovas (2007).
[28] Bonnay (2006).
[29] Voir Bonnay (2006), chap 4., pour une présentation synthétique.

Critère d'identité d'une logique

Il y a davantage de logiques aujourd'hui que de géométries au début du XXè siècle. Et quoi de commun entre ce que nous appelons la logique du premier ordre, les logiques infinitaires ou la logique du second ordre ? Lorsque nous enrichissons la logique du premier ordre d'une nouvelle constante logique, changeons-nous de sujet ou ne faisons-nous que nous donner de nouveaux moyens de la parler de la même chose ? Cela dépend probablement de la nature de la constante logique ajoutée, et c'est ce à quoi nous voulons donner un sens précis dans cette section.

A ce stade, l'idée que je veux mettre en œuvre va de soi, dans ses grandes lignes : il s'agit d'individualiser une logique L par la relation de similarité associée, c'est-à-dire la plus relation de similarité S la plus fine telle que toutes les notions de L sont S-invariantes. Pour faciliter la discussion, je dirai volontiers que S est le « sujet » de L[30] Considérons la logique du premier ordre ordinaire, dont les seules constantes logiques, outre les connecteurs propositionnels, sont les quantificateurs existentiels ($\exists$) et universels ($\forall$). En vertu de ce que nous avons dit plus haut, leurs dénotations sont des classes de structures de la forme $\langle M, A\rangle$, $A \subseteq M$. Pour comprendre quel est le « sujet » propre de la logique du premier ordre, il nous faut donc considérer la plus grande relation de similarité telle que les quantificateurs existentiels et universels sont invariants sous cette relation de similarité (cette proposition se généralise naturellement à toute logique $\mathcal{L}$). Notre proposition, provisoire, est de déclarer que si S cette relation de similarité, alors S caractérise le sujet de la logique du premier ordre au sens précédent.

On pourrait présenter les choses un peu différemment en introduisant la notion de *complétude fonctionnelle* d'une logique, issue la métalogique du calcul propositionnel. Lorsque nous étendons le fragment de la logique propositionnelle ordinaire fondé sur l'ensemble de connecteurs $\{\vee\}$ à la logique fondée sur l'ensemble $\{\neg, \vee\}$, avons-nous changé de logique ? En un sens important, il semble que non. Pour comprendre en quel sens il s'agit toujours de la même logique nous nous plaçons à un point de vue sémantique et nous disons qu'il s'agit de la logique des fonctions de vérité à un nombre fini d'arguments. Savoir que c'est *cela* le sujet de la logique propositionnelle nous permet d'affirmer qu'en passant de $\mathcal{L}_{\{\vee\}}$ à $\mathcal{L}_{\{\neg,\vee\}}$ nous n'avons pas changé de sujet. Inversement, l'extension d'un système de logique propositionnelle par des opérateurs non vérifonctionnels est comprise comme un authentique

[30]Il ne faut pas prendre trop au sérieux ce mot de « sujet » ici.

changement de logique. Ce que suggèrent ces observations à propos de la logique propositionnelle, c'est que la notion de complétude fonctionnelle est centrale pour notre classification des logiques. Il est naturel alors de vouloir identifier des logiques de la façon suivante : L_1 et L_2 sont des systèmes de la même logique si et seulement si leurs plus petites extensions fonctionnellement complètes sont identiques, ce qui ne serait qu'une autre façon de dire que ces logiques ont le même objet. Et une fois cet objet caractérisé, comme plus haut, nous sommes en mesure de donner un analogue à la notion de complétude fonctionnelle recherchée : la logique $\mathcal{L}$ est fonctionnellement complète si et seulement si tous les quantificateurs S-invariants sont définissables $\mathcal{L}$, ou autrement dit si, et seulement si toutes les classes de structures S-invariantes sont définissables dans $\mathcal{L}$.

Mais il y a ici une difficulté qui doit retenir attention et nous oblige à raffiner un peu notre proposition. En effet, suivant ici la littérature,[31], il est admis qu'une notion définissable à partir de notions logiques doit être elle-même logique. Si, par exemple, le quantificateur $\exists$ est logique, que l'égalité est une relation logique, ainsi que les fonctions de vérité, alors nous devons admettre que le quantificateur $\exists_{>2}$ (il existe plus de trois) est lui-même logique. En effet, il est définissable simplement par la formule suivante :

$$\exists_{\geq 2}(x)\phi(x) \text{ ssi } \exists x \exists y(x \neq y) \wedge (\phi(x) \wedge \phi(y))^{32}$$

Il y a donc une contrainte qui doit peser sur notre critère d'identité pour une logique, à l'effet que l'on ne sort pas du domaine d'une logique donnée en y ajoutant un opérateur définissable. De ce point de vue, le choix d'identifier une logique, considérée comme une classe d'opérateurs K, simplement à la plus grande relation de similarité laissant les opérateurs de K invariants, n'est pas satisfasant. Etant donnée une relation de similarité S quelconque et une classe K d'opérateurs S-invariants, il est tout à fait possible qu'existe un opérateur Q qui soit *définissable*

[31] *cf.* à nouveau Feferman (1999*b*)

[32] On pourrait raisonner aussi en termes de symboles logiques. Pour toute L-structure $\mathcal{M}$, pour toute formule ϕ :

$$\mathcal{M} \vDash \overline{Q}_{\geq 2}\phi(x) \text{ ssi } \mathcal{M} \vDash \exists x \exists y((\phi(x) \wedge \phi(y)) \wedge (x \neq y)).$$

Si la classe de structures $\langle M, P\rangle$ telles que P contient au moins deux éléments est définissable par une formule purement logique (i.e. où les seuls symboles interprétés sont des symboles logiques) alors le quantificateur $\overline{Q}_{\geq 2}$, qui dénote cette classe, est lui-même une constante logique. Plus généralement un symbole de qui dénote une classe de structures définissables en termes purement logiques doit être un symbole logique.

dans un langage contenant pour seules constantes logiques des symboles dénotants des opérateurs appartenant à K, et tel que Q lui-même ne soit pas S-invariant. C'est qu'il se passe par exemple si l'on choisit pour relation de similarité la relation d'isomorphisme partiel Iso_ω : le quantificateur Q_{Inf} (« Il existe une infinité de ») est Iso_ω invariant, mais l'on peut définir dans un langage relationnel contenant le quantificateur Q_{Inf} des classes de structures qui ne sont pas S-invariantes ! [33]

Ce que montrent ces considérations, c'est que toute relation de similarité n'est pas un bon candidat au titre de « sujet » possible d'une logique. Les seules relations de similarité qui doivent nous intéresser sont les relations de similarité closes par définissabilité :

Définition 31. *S est close par définissabilité ssi pour toute classe K d'opérateurs S-invariants, pour toute signature σ : toute classe de structures définissable par une formule de signature σ et ne contenant que des opérateurs de K au titre de l'interprétation de ses symbole logique, est S-invariante.*

Si nos notons L une logique identifiée à la classe des interprétations de ses symboles logiques primitifs, nous pouvons noter $S(L)$ la plus grande relation de similarité close par définissabilité et telle que tous les opérateurs de L sont S-invariants. Je dirai alors que deux logiques L_1 et L_2 ont même sujet si $S(L_1) = S(L_2)$.[34]

En résumé, en utilisant les outils développés dans la littérature pour comprendre la notion de constante logique, et en nous inspirant à notre tour du programme de classification des géométries de Klein, nous avons donné un critère raisonné de ce qui doit compter comme *une* logique. Ce point est important, c'est de son sens que dépend la portée de notre affirmation que la notion de vérité est-elle même une notion logique.

[33] Voir Bonnay (2008). Un exemple présenté par Bonnay est repris dans Galinon (2009), joint en annexe à ce travail, p. ? ?

[34] Cette définition permet également de définir une notion de complétude fonctionnelle étendue au-delà du calcule propositionnel en posant : L est fonctionnellement complète si et seulement si tous les opérateurs $S(L)$-invariants sont définissables dans L. [35] Finalement, étant donné une logique $\mathcal{L}$, nous appellerons *clôture fonctionnelle* de $\mathcal{L}$ toute logique fonctionnellement complète ayant le même objet que $\mathcal{L}$.

Les logiques fonctionnellement complètes sont vrai-complètes

Nous voulons prouver que, pour une logique quelconque, si cette logique est « fonctionnellement complète », alors l'enrichir d'un prédicat de vérité interprété n'accroît pas son pouvoir expressif. En reprenant ici le cadre formel des extensions aléthiques d'une logique développé au chapitre 7 pour rendre compte de l'expressivité du prédicat de vérité dans ses emplois « transparent », nous dirons qu'une logique est vraie-complète si ses extensions aléthiques n'augmentent pas son pouvoir expressif. Autrement dit,

Définition 32. *Une Logique L est vraie-complète si et seulement si pour toute signature σ, pour toute expansion aléthique $L_{\mathcal{A},Vr}$ de L, pour tout énoncé ϕ de $L_{\mathcal{A},Vr,\sigma}$, il existe un énoncé ϕ^* de L_σ tel que pour toute σ-structure $\mathcal{M}$:*

$$\mathcal{M} \otimes \mathcal{A} \vDash \phi \text{ si et seulement si } \mathcal{M} \vDash \phi^*$$

Autrement dit L est vraie-complète si et seulement si, si une classe de structures aléthiques est une classe élémentaire dans une certaine extension aléthique de L, alors la classe des structures réduites correspondante est une classe élémentaire de L. En quantifiant universellement sur les expansions aléthiques, nous demandons que l'énoncé ϕ^* existe quelque soit la richesse des extansions aléthiques que nous considérions, c'est-à-dire quelle que soit la richesse des moyens que nous nous donnons pour former des ensembles d'énoncés auxquels appliquer les prédicat de vérité.

Nous sommes alors en mesure de formuler le résultat suivant :

Théorème 5. *Si L est fonctionnellement complète alors L est vraie-complète.*

Démonstration. La preuve a été donnée par Denis Bonnay dans un travail commun en cours :

On note $Inv(S)$ la classe des quantificateurs invariants par S. Une logique L est fonctionnellement complète si et seulement si $El_L = Inv(S)$.

Soit maintenant L une logique et S une relation de similarité telle que $El_L = Inv(S)$. Soit σ une signature, $L_{\mathcal{A},Vr,\sigma}$ une expansion aléthique quelconque de L_σ, et ϕ un énoncé de $L_{\mathcal{A},Vr,\sigma}$. Il suffit de montrer que $Q_\phi = \{\mathcal{M}/\mathcal{M} \otimes \mathcal{A} \vDash \phi\}$ est clos pour S, car alors $Q_\phi \in El_L$ puisque $Inv(S) \subseteq El_L$. Supposons que $\mathcal{M} \in Q_\phi$ et que $\mathcal{M}S\mathcal{M}'$. Nous voulons montrer que $\mathcal{M}' \in Q_\phi$.

Par hypothèse $El_L \subseteq Inv(S)$, donc $\mathcal{M}S\mathcal{M}'$ implique que $\mathcal{M} \equiv_{L_\sigma} \mathcal{M}'$. Sinon il existerait un énoncé ψ de L_σ vrai dans $\mathcal{M}$ mais non dans $\mathcal{M}'$, ce qui implique que la classe des modèle de ψ n'est pas close par S, contredisant $El_L \subseteq Inv(S)$. Par le Fait 4, $\mathcal{M} \equiv_{L_\sigma} \mathcal{M}'$ implique que $\mathcal{M} \otimes \mathcal{A} \equiv \mathcal{M}' \otimes \mathcal{A}$. Puisque $\mathcal{M} \otimes \mathcal{A} \vDash \phi$, on a que $\mathcal{M}' \otimes \mathcal{A} \vDash \phi$, et donc $\mathcal{M}' \in Q_\phi$.

□

Autrement dit, étant donné une logique L et une extension aléthique L_{Vr} de cette logique, L et L_{Vr} sont identifiées comme la même logique dans la classification que nous avons donnée dans la section précédente. D'un point de vue logique, ajouter un prédicat de vérité ne fait pas « changer de sujet ».

Interprétation du résultat

Le résultat précédent peut être compris comme un résultat de définissabilité implicite : si une logique est fonctionnellement complète, alors le prédicat de vérité est « implicitement définissable » dans cette logique. Que faut-il entendre par là ? Il est bien connu que la vérité pour un langage donné (classique) n'est pas définissable explicitement dans ce langage. Notre usage de « définissabilité implicite » s'écarte également de celui qui en est fait par exemple par Hilbert ou de celui en jeu dans les théorèmes classiques de définissabilité de Beth. Dans notre perspective, les aspects syntaxiques de la logique sont en retrait et les aspects purement sémantiques qui sont mis en avant. Souvenons-nous que nos avons identifié un langage à la classe des classes de structures qu'il permet d'exprimer, autrement dit la classe de leurs classes élémentaires. Dans le même ordre d'idée, identifions le prédicat de vérité à la classe des classes de structures qu'il permet d'exprimer *via* toutes les formes d'énoncés disponibles dans le langage L où seuls les constantes logiques et le prédicat de vérité sont intérprétés, et appelons cette classe V_L. Alors si L est une logique vraie-complète la vérité est implicitement définissable dans L en ce sens que toutes les classes de structures qui sont dans V_L sont définissables par des énoncés ne contenant pas de prédicat de vérité. Ceci ne signifie nullement qu'il existe, pour les extensions aléthiques de logiques vraie-complètes, une procédure de traduction du langage total dans le fragment non-aléthique du langage. Il est possible qu'aucune définition uniforme de la vérité n'existe, mais seulement des traductions contextuelles, au sens où pour chaque énoncés contenant le prédicat de vérité on peut trouver un énoncé

équivalent, au sens précisé à l'instant, ne le contenant pas.

Le résultat précédent nous assure qu'il est suffisant que L soit fonctionnellement complète pour que la vérité soit implicitement définissable dans L. Quelles conclusions tirons-nous de ce résultat ? Tout d'abord le résultat est compatible avec le fait que l'ajout d'un prédicat de vérité interprété à un langage, en général, en augmente le pouvoir expressif. Mais souvenons-nous de la section 2. Nous avons remarqué qu'à chaque logique L est naturellement associée une relation d'indiscriminabilité, ou de similarité entre structure, et que L pouvait ou non capturer complètement les invariants de cette relation de similarité. Lorsque L ne capture pas tous les invariants de sa relation de similarité associée[36], nous avons dit que L n'était pas fonctionnellement complète, en analogie avec la notion bien connue qui cours dans la métathéorie du calcul propositionnel. D'un autre côté, pour qu'une notion soit logique au sens de L, il suffit que les distinctions que cette notion permet de faire ne soit pas plus fines que celles permises par la relation de similarité associée à L. Mais précisément, le résultat précédent montre qu'en ajoutant un prédicat de vérité à une logique L, on ne peut discriminer plus de structures qu'il n'est permis par la relation de similarité associée à L. En résumé, lorsque le prédicat de vérité augmente le pouvoir expressif d'une logique L, il le fait en respectant la notion de logicité sous-jacente à L, puisqu'il n'ajoute aucun pouvoir expressif à toute logique fonctionnellement complète.

Par analogie avec ce que nous avons montré aux chapitres précédent, nous pourrions présenter les choses de la façon suivante. Lorsqu'on ajoute un prédicat interprété à un langage, il se peut que l'on augmente le pouvoir expressif de ce langage. Mais de même qu'il est souhaitable de chercher à qualifier les phénomènes de non-conservativité lorsqu'ils surviennent, parce que tous n'ont pas la même signification, il est souhaitable de qualifier les phénomènes d'enrichissements expressifs. Et c'est ce que nous venons de faire : l'enrichissement expressif auquel donne lieu l'adjonction d'un prédicat de vérité à une logique est circonscrite par ce qui apparaît être le domaine de la logique de départ, tel qu'il se dégage de l'analyse l'analyse par invariance de la logique de départ.

On peut vouloir approfondir l'analogie entre non-conservativité et non-conservativité « expressive » par une comparaison de notre argument avec une version stylisée de l'argument de la conservativité de Shapiro et Ketland. L'approche de la

[36]Pour mémoire : la plus grande relation de similarité close par définissabilité et telle que tous les constantes logiques de L soit S-invariantes.

notion de vérité par la conservativité se concentre sur la force *déductive* différentielle de certaines théories relativement à leur extension par une *théorie de la vérité*. Ces théories de base étendues par des principes faisant appel au prédicat de vérité, ne sont pas présumées contraindre complètement l'interprétation du prédicat, mais néanmoins être être suffisamment fortes pour que certaines vérités mettant en jeu la notion de vérité en apparaissent comme des théorèmes. Le critère proposé qui doit permettre de juger du caractère substantiel ou non de la vérité est alors celui de la conservativité des extensions sur les théories de bases. Par contraste nous supposons donnée l'interprétation du prédicat de vérité, et cherchons à déterminer si le prédicat est logique. Pour déterminer si le prédicat de vérité est logique nous nous concentrons sur la force *expressive* différentielle de certains *langages* relativement à leur version augmentée par le prédicat de vérité interprété. Rappelons que le caractère logique d'une expression, si l'approche par l'invariance est correcte, a à voir avec le genre de distinction qu'une expression (interprétée) peut faire entre différents états possibles du monde (représentés dans notre modèle par des différentes structures). Il y a ces expressions, telles que « Rouge », qui distinguent entre différents états sur la base de certains de leur traits empiriques, d'autres, comme certaines expression mathématiques, qui distinguent différents états en vertu de différences structurelles complexes, tandis que les expressions logiques ne distinguent différents états que selon des traits qu'on se sera accordés à considérer comme « logiques », en un sens qui, présumablement, incorporera nos intuitions fondamentales sur la nature de la logique, à savoir sa formalité, sa généralité etc. Par conséquent, le critère proposé qui doit permettre de juger du caractère logique ou non de la vérité, un critère de « conservativité expressive » : ajouter un prédicat de vérité à une logique L (supposée être La Logique) ne doit pas enrichir son pouvoir expressif (ie son pouvoir de discrimination).

Vues sous cet angle, les deux approches partagent ce que l'on pourrait appeler un problème d'*instabilité*, assez similaire. En effet, quand vient le moment de mettre en œuvre la machinerie conceptuelle pour trancher la question qui l'a motivée, un constat s'impose : étendre une théorie de base par des principes aléthiques tarskiens, tantôt produit une extension conservative de la théorie de base, tantôt non ; d'un autre côté, ajouter un prédicat de vérité à une logique tantôt enrichit son pouvoir expressif, tantôt non. À ce stade, il manque donc un élément conceptuel crucial pour féconder ces approches et leur permettre de justifier les conclusions que leurs défenseurs veulent en tirer : il faut pouvoir expliquer pourquoi cette versatilité ne

met pas en danger les conclusions avancées, pourquoi les cas « défavorables » aux conclusions attendues ne le sont qu'en apparence.

Dans le cas des approches par la conservativité, il n'y a aucune réponse convaincante disponible à cette question. Comme nous l'avons mentionné au chapitre précédent, l'extension tarskienne de l'arithmétique de Peano en premier ordre en est une extension non conservative, tandis que l'extension tarskienne de l'arithmétique de Robinson est une extension conservative de cette dernière. À l'autre bout du spectre des théories arithmétiques, pourquoi ne pas mentionner aussi le fait bien connu des logiciens que les axiomes de Peano en premier ordre augmentés de l'ω-règle est une théorie complète, ce qui implique que son extension tarskienne en est une extension conservative. Mais alors pourquoi la non-conservativité observée dans un cas serait-elle plus significative que la conservativité observée dans les autre cas ? D'ailleurs, ne serait-ce pas plutôt le contraire ? Aucune réponse à ces questions n'est donnée par les défenseurs de l'approche par la conservativité.[37]

En revanche, notre approche nous permet de faire beaucoup mieux. Certes, ajouter un prédicat de vérité interprété à une logique tantôt augmente son pouvoir expressif, tantôt ne l'augmente pas : tout dépend de la logique dont on part. Mais nous avons découvert une condition nécessaire conceptuellement significative à l'observation de cas d'augmentation du pouvoir expressif de la logique : ce sont les cas de logiques fonctionnellement incomplètes. Si une logique est fonctionnellement complète, alors lui ajouter le prédicat de vérité n'augmente pas son pouvoir expressif.[38] Mieux, le pouvoir expressif que gagne une logique lorsque qu'on y ajoute un prédicat de vérité ne va pas au-delà du pouvoir expressif de se son extension fonctionnellement complète. Pourquoi voyons-là une indication du caractère logique du prédicat de vérité ? Souvenons-nous que le pouvoir expressif d'une logique peut être augmenté de façons qualitativement différentes. Par exemple, ajouter le prédicat interprété « Rouge » à une logique L augmentera certainement le pouvoir expres-

[37] Une réponse immédiate est bien entendu : il suffit qu'il existe *une* théorie non-conservativement étendue par son extension aléthique pour montrer que la vérité est substantielle. On peut considérer que la première partie de notre chapitre précédent, traitant des théories arithmétiques et du rôle des schémas, est une réponse à cette réponse.

[38] En fait, si nous élargissons notre définition des structures aléthiques de sorte que le domaine de la superstructure puisse être n'importe quel surensemble de l'ensemble des énoncés du langage de base, alors on peut montrer que la condition est également suffisante : L est vraie-complète si et seulement si elle est fonctionnellement complète. Voir l'article « Logicality and truth » en collaboration avec Denis Bonnay et Julien Boyer (en préparation). Certaines difficultés d'interprétation rendent néanmoins prématuré l'exposé du résultat.

sif de L, si L mérite le nom de « logique ». D'un autre côté, ajouter une nouvelle « constante logique » à L peut augmenter le pouvoir expressif de L *même lorsque la nouvelle constante est logique au sens de « logique » implicite dans* L *elle-même.*[39] Et ce que nous avons montré, c'est que le pouvoir expressif additionnel que donne la vérité est du second genre, non du premier. L'ajouter à une logique L respecte la logique de L, et le pouvoir additionnel qu'elle donne n'est pas « substantiel » en un sens qui irait au-delà de ce qui est permis au pouvoir expressif des expressions logiques. C'est en ce sens qu'on peut dire que le prédicat de vérité est logique.

Approche par la conservativité	Notre approche
Etudie les propriétés de conservativité d'une théorie du concept X	Etudie les propriétés expressives de X muni de sa signification attendue
Les propriétés non-substantielles ont des théories conservatives	Les expressions logiques n'enrichissent pas le pouvoir expressif d'un langage
Problème : La vérité apparaît comme substantielle ou non selon selon la théorie de base	Problème : La vérité apparaît comme logique ou non selon la logique de base
Réponse : ?	Réponse : Le prédicat de vérité n'enrichit que les logiques fonctionnellement incomplètes et pas au-delà de leur extension complète

[39] Implicite, si l'on accepte l'approche en terme d'invariance.

A.3 Conclusion

Revenons en conclusion sur le chemin parcouru. Qu'avons-nous montré au juste ? Nous sommes partis d'une analyse conceptuelle et informelle des traits distinctifs de la sémantique des constantes logiques. Les notions logiques, avons-nous dit, sont caractérisées par une sorte d'absence de contenu, absence que l'idée de *formalité* des notions logiques semble faite pour capturer. Cette idée de formalité recouvre en fait elle-même différentes notions, passibles d'analyses différentes, en particulier en termes de généralité ou en termes d'absence d'objet. Plutôt que d'insister sur ces différences, j'ai rappelé, suivant les pas de Tarski, comment la notion d'invariance pouvait être mise à profit pour rendre compte de la généralité de la logique et j'ai fait l'hypothèse que l'invariance par certaines classes de transformation permettait tout aussi bien de capturer l'idée informelle d'absence de contenu des notions logiques.

La seconde étape importante a été d'abandonner l'idée d'une caractérisation unique de la logicité, et de profiter de la versatilité de la notion d'invariance pour dégager l'idée de « sujet » d'une logique donnée. Il y a une gradation possible de l'idée de généralité, comme de l'idée d'absence de contenu, et la notion d'invariance permet de capturer ces degrés en termes de « taille » des classes d'invariance associées à une classe d'expressions données. Le « sujet » d'une logique n'est rien d'autre que la relation de « ressemblance logique », ou de « similarité » entre structures , sémantiquement associée à la classe de ses constantes logiques. À partir de l'idée du « sujet » d'une logique donnée il a été possible de donner un sens précis à l'affrimation que certaines logiques sont expressivement incomplètes, en un sens de « complétude » analogue, ai-je soutenu, à celui de « complétude fonctionnelle », la notion familière de la logique propositionnelle.

Parallèlement, nous avons repris la suggestion faite au chapitre 7 d'après laquelle l'usage ordinaire du prédicat de vérité indiquait qu'il partage, au moins de prime abord, certains caractéristiques importants des constantes logiques (neutralité, atopicité), malgré une grammaire de surface en vertu de laquelle la vérité se dit de certains types d'objets bien spécifiques, les énoncés. Cette hypothèse justifiait l'emploi des *structures aléthiques*, définies au chapitre 7, et permettant d'étudier le pouvoir expressif de la vérité en mettant entre parenthèse, pour ainsi dire, les énoncés qui lui servent de véhicules. À nouveau, on peut voir le choix de ce cadre formel comme une façon d'incorporer la thèse déflationniste en vertu de laquelle lorsque l'on parle en surface de la vérité des énoncés, on ne parle véritablement au

fond *que* du monde (i.e. de ce dont parlent ces énoncés).

Muni de la notion de structure aléthique, nous avons découvert que tout énoncé d'un langage fondé sur une logique fonctionnellement complète et muni d'un prédicat de vérité pouvait être « traduit » par un énoncé de ce langage ne contenant pas de prédicat de vérité. Autrement dit les nouvelles distinctions que permet de faire un prédicat de vérité lorsqu'il est introduit dans un langage fondé sur une logique L sont des distinctions purement logiques, au sens de « logique » implicite dans l'acceptation du caractère logique de L, et où l'élaboration de cette notion d' « implicite » est faite en terme d'invariance *via* la notion de complétude fonctionnelle. La conclusion est alors que si l'approche de la logicité en terme d'invariance est correcte, et pour autant que le contenu des attributions de vérité soit réductible au contenu des énoncés auxquels la vérité est attribuée, alors le prédicat vérité est logique.

Annexe B

A note on generalized functional completeness in the realm of elementary logic

Note : l'article qui suit a été publié en 2009 dans le *Bulletin of the section of logic* de l'Académie des Sciences de Pologne. Son objet est de montrer qu'une logique L est fonctionnellement complète, au sens donné à ce terme dans l'appendice A, si et seulement si toutes les classes de structures closes par L-équivalence élémentaire sont élémentaires dans L.

In « Logicality and Invariance » (2008), Denis Bonnay introduced a generalized notion of functional completeness. In this note we call attention to an alternative characterization that is both natural and elementary.

B.1 Maximizing the expressive power of a logic

Let L be a logic whose logical vocabulary contains truth-functional connectives (possibly infinitary ones), the first-order quantifiers and possibly some generalized quantifiers $\overline{Q_1}, ..., \overline{Q_n}$. Semantically, quantifiers are identified with classes of structures in a standard manner.[1] Satisfaction in a structure is defined in conformity with the meaning of the logical constants chosen, e.g., $\mathcal{M} \vDash \overline{Q}x\phi(x)$ iff $\langle M, \{a : \mathcal{M} \vDash \phi(a)\}\rangle \in Q$. Given those constraints, a logic L can be identified with a set of logical constants.[2] Naturally associated with L, we have

1. an elementary equivalence relation between structures ($\equiv_L$), and
2. the class of elementary classes of L (El_L).[3]

The two concepts are related to the intuitive notion of the expressive power of a logic : on the one hand, the more expressive a logic, the more fine-grained the partition of the universe of structures induced by the associated elementary equivalence relation. (Think of two non-isomorphic first-order equivalent models of PA. They are obviously not second-order equivalent.) On the other hand, the larger the collection of elementary classes associated with a logic, the more expressive the logic. (Adding second-order logical constants to FOL makes new classes of structures definable by a sentence ; for instance the isomorphism class of $\mathbb{N}$ is second-order elementary but not first-order elementary).[4]

Given a logic L, all elementary classes are of course closed under elementary equivalence ; but what about the converse ? In first-order logic, not all classes of

[1] See e.g. Westerståhl (2007), p.235. For instance $\forall$ is the class of structures $\langle M, A\rangle$ such that $M = A$.

[2] Note that this definition of a logic is not purely semantic, and thus is less general that the one usually found in abstract model theory. See for instance Ebbinghaus (1985), Definition 1.1.1. for the general definition.

[3] A class of structures is *L-elementary* in our terminology iff it is definable by a single sentence of L.

[4] See for instance Ebbinghaus (1985), Def. 3.1.4 and 3.1.5 for a similar distinction between the two notions of expressive strength of a logic.

structures closed under elementary equivalence are elementary.[5] However, we have no reason to rule out the possiblity that for some logic L all classes of structures closed under elementary equivalence for L are elementary in L, and such a case seems to be of conceptual interest. Thinking intuitively of the notion of L-elementary equivalence between structures as a criteria of (relative) identity between structures from the point of view of that logic, it seems natural to think of logics such that all classes closed under elementary equivalence are elementary as meeting a requirement of "expressive completeness » or « internal completeness ». We shall call such logics $\Delta\Sigma$-closed.

Given the fact that $\Delta\Sigma$-closed logics seem to form a natural object of attention, it may be worth remarking upon the fact that they also arise from other directions. The rest of this note will be devoted to the task of making precise the following remark that this simple condition of $\Delta\Sigma$-closure of a logic is equivalent to one of expressive completeness introduced recently by Denis Bonnay (in Bonnay (2008), p.35).[6] This remark may enforce the idea of « generalized functional completeness » as both natural and elementary.[7]

We first recast our foregoing discussion with precise definitions. The following is standard (see Bell et Slomson (1969), p. 141, for the first-order version) :

Définition 33. *Given a logic L and a fixed first-order signature :*

- *An elementary class of structures (El_L) is one axiomatizable by a sentence of L.*
- *A Δ-elementary class (El_L^{Δ}) is the intersection of a set of elementary classes, i.e. is axiomatisable by a theory of L.*
- *A Σ-elementary class (El_L^{Σ}) is the union of a set of elementary classes.*
- *A $\Delta\Sigma$-elementary class ($El_L^{\Delta\Sigma}$) is class which is the union of a collection of Δ-elementary class.*

Given the previous definition, the following fact is trivial :

[5] Think of the fact that, in first-order logic, when a theory has models of arbitrary finite cardinality it necessarily has a model of infinite cardinality. But a finite and an infinite structure cannot be elementary equivalent in FOL.

[6] From a historical point of view, the notion is rooted in the work of Tarski on logical constants Tarski (1966/1986).

[7] The phrase « elementary logic » in the title of this paper refers both to the informal sense of « simple » and to first order logic.

Proposition 8. *Let $\mathcal{K}$ be a class of structures. $\mathcal{K} \in El_L^{\Delta\Sigma}$ iff $\mathcal{K}$ is closed under elementary equivalence.*

Démonstration. The proof is similar to the first-order case (see Bell et Slomson (1969), Lemma 1.11 p.143). □

Proposition 1 justifies our talking of $\Delta\Sigma$-closed logic in the opening paragraph. Remark that, like Δ-interpolation [8], $\Delta\Sigma$-closure also seems to record a kind of balance between the syntax and the semantics of the logic, albeit in a quite different sense than Δ-interpolation is usually understood to do : here the balance is between the power to define (classes of structures) and the power to discriminate (between structures).

B.2 Generalized functional completeness

I now briefly recall the motivation and definition of generalized functional completeness given by Bonnay in Bonnay (2008). The problem is best understood by analogy with Propositional logic. Propositional logic is the logic of truth-functions (of finitely many arguments). In view of this general characterization of propositional logic, two things could naturally be required of any particular propositional logic : first, every logical form in the propositional logic should determine a truth-function (expressive adequacy), and second every truth-function should be expressible in the propositional logic (expressive completeness). Usual propositional logic based on the set of logical constants $\{\neg, \vee\}$ is expressively adequate and complete in this sense, while $\{\wedge, \vee\}$ and $\{\forall, \vee\}$ are not (not complete and not adequate respectively).

The study of functional completeness is of conceptual interest. First, it is *prima facie* a fundamental property that a logic may lack or enjoy, and we would like to understand how a property which is so congenial in the propositional case is supposed to be applied in non-propositional cases. Second, and more importantly, it is fundamental to understand how the many superficially different logics we know of connect to each other and to identify, beyond their variety, some illuminating classification. As we contend that $\{\vee\}$ and $\{\neg; \wedge\}$ are not really different logics,

[8] A logic has the Δ-interpolation property iff every class $\mathcal{K}$ of structures such that both $\mathcal{K}$ and its complement (in the given signature) are projectible class in L is also elementary in L. See for instance Ebbinghaus (1985), §7.2 for details.

one may well ask, for example, whether $L_{\infty,\omega}$ is best understood as an intrinsically justified extension of FOL or rather as an entirely different logic. In connection to this, understanding functional completeness may bear on the philosophical issue of determining which properties are logical properties, as oppsosed to, say, distinctively mathematical properties. For instance, starting from the assumption that properties definable in FOL are purely logical properties, it would not seem reasonable to deem every non-first-order definable property as non-logical in the eventuality that FOL were not functionally complete. Rather, it would seem that accepting FOL as logic commits one to accepting every property definable in the functional completion of FOL as being also a logical property.

Now, there is no widely accepted extension of the notion of functional completeness beyond the propositional case. This state of affait is explained by Bonnay thus, focusing on case of FOL :

> There is no such standard result for FOL, essentially because there is no standard answer to what FOL is about, which would be similar to the claim that PC is about truth-functions. (Bonnay (2008), p.35)

It is one of Tarski's merits to have set up a framework for studying precisely this question in rigourous terms. Tarski did it in the spirit of Klein's *Erlanger program* : because logic is the most general science, what a logic is about are the properties invariant under the largest class of transformations.[9]

In order to give a precise definition of the notion of expressive completeness as worked out by Bonnay, we first recall briefly his generalized framework for studying invariance properties [10]. À *similarity* relation S between structures is a binary relation over structures of same signature, and a class K of structures is said to be invariant under S if it is closed under S.[11] Moreover, given a relation of similarity S, we can define the operation $Inv(S)$ taking S to be the collection of classes of structures invariant under S, that is the collection of quantifiers invariant under S. Associated with $Inv(S)$ is $\overline{Inv(S)}$, the class of logical expressions whose elements denote the members of $Inv(S)$.

[9]Tarski was interested in the demarcation of the set of logical constants or, to put it otherwise, he wanted to determine which logic is the true logic, if any. G. Sher Sher (1991), S. Feferman Feferman (1999*b*) and D. Bonnay Bonnay (2008) have followed his path, improving on the framework and conceptual motivation.

[10]For reasons lying outside the scope of this paper, it has become clear that it is not satisfactory to restrict the notion of invariance to one of invariance under a class of transformation over a structure, as done by Tarski. See Bonnay (2008) for details.

[11]That is K is S-invariant iff, for every $\mathcal{M}, \mathcal{M}'$, if $\mathcal{M}S\mathcal{M}'$ then $\mathcal{M} \in K$ iff $\mathcal{M}' \in K$.

We are now in a position to state the generalization of the foregoing notions of adequacy and expressive completeness of a logic in complete analogy with the propositional case. À logic is to be understood as the logic of S-invariant notions, for a suitable S. Moreover two things are required of a logic L, understood as a set of logical constants, to ensure its adequacy and completeness relative to the target notion of logicality : every logical form in L should express an S-invariant notion (adequacy) and every S-invariant notions should be definable in L (completeness). Remark that on this view, for some choice of a similarity relation S, there may not exist a complete and adequate logic of S-invariant notions.[12] For an adequate S-logic to exist, S must be closed under definability, which means the following : the classes of structures definable by a sentence of $L_{\overline{Inv(S)}}$ should themselves be S-invariant (i.e. logical forms in $L_{\overline{Inv(S)}}$ should express S-invariants).[13] Finally, define a logic L to be *functionally complete* if, and only if, it is adequate and complete relative to S-invariants, for some S. It is easily seen that this definition can be rephrased thus :

Définition 34. *L is functionally complete iff for some S, $El_L = Inv(S)$*[14]

B.3 A logic is ΔΣ-closed iff functionally complete

The following is almost immediate given the definitions :

Proposition 9. *The following are equivalent :*

1. *L is $\Delta\Sigma$-closed*
2. *L is functionally complete*

Démonstration. It is obvious that $\Delta\Sigma$-closure implies functional completeness : If L is $\Delta\Sigma$-closed then $El_L = Inv(\equiv_L)$, because it is the same to speak of the classes of structures invariant under elementary equivalence and to speak of the class of structures closed by the same elementary equivalence relation, so that $Inv(\equiv_L)$ and $El_L^{\Delta\Sigma}$ are just notational variants. Hence there is an S such that $El_L = Inv(S)$.

[12] The case of $S = \equiv_{FOL}$ is a case in point. See below.

[13] See Bonnay (2008), p. 50 for the definition of closure under definability of a similarity relation. In contradistinction to $\equiv_{FOL}$, the relations of isomorphism and potential isomorphism between structures are closed under definability. See Bonnay (2008) p.51.

[14] L is functionally complete iff adequate and complete relative to S-invariants of some S. Moreover L is adequate to S-invariance iff $El_L \subseteq Inv(S)$. By definition L is complete relative to S-invariance iff all S-invariants are definable in L, that is $El_L \supseteq Inv(S)$.

To prove the converse implication, remark first that

Proposition 10. *If S is a similarity relation and $L_{\overline{Inv(S)}}$ the associated logic, then S is closed under definability iff $S = \equiv_{L_{\overline{Inv(S)}}}$.*

Démonstration. Only if : Assume that S is closed under definability, and that it is *not* the case that $\mathcal{M} \equiv_{L_{\overline{Inv(S)}}} \mathcal{M}'$. Then there is ϕ s.t. $\mathcal{M} \in Mod(\phi)$ and $\mathcal{M}' \notin Mod(\phi)$. Since S is closed under definability, $Mod(\phi)$ is S-invariant. Hence it is not the case that $\mathcal{M}S\mathcal{M}'$. The converse inclusion is obvious since if it is not the case that $\mathcal{M}S\mathcal{M}'$, there is a quantifier $\overline{Q}$ in $L_{\overline{Inv(S)}}$ such that $\mathcal{M} \in Q$ and $\mathcal{M}' \notin Q$, and a formula ϕ in $L_{\overline{Inv(S)}}$ s.t. $\mathcal{M} \models \phi$ and $\mathcal{M} \not\models \phi$.

For the *if* part : assume that S is not closed under definability. There is a ϕ in $L_{\overline{Inv(S)}}$ which distinguishes two similar structures. Hence they are not elementary equivalent in the sense of $L_{\overline{Inv(S)}}$. Hence $S \neq \equiv_{L_{\overline{Inv(S)}}}$. □

Now remark that if L is functionally complete, there is an S such that $El_L = Inv(S)$ (by definition). But $S = \equiv_L$, so that $El_L = Inv(\equiv_L)$ (i.e. L is ΔΣ-closed). □

So the notion of functional completeness, which grew out from fine-grained considerations pertaining to the denotation of logical constants and the nature of logic, has a straightforward counterpart in model-theoretical terms, namely in terms of the relation between elementary classes and the relation of elementary equivalence associated with a logic.

We conclude with some remarks and questions. We have already remarked that FOL is not ΔΣ-closed. On the other hand it can be shown [15] that $El_{\mathcal{L}_{\infty,G}} \supset El_{\mathcal{L}_{\infty,\omega}}$ while on the other hand it follows from a result of Barwise that $\mathcal{L}_{\infty,G}$ has the Karp property[16] and then, by a well-known theorem of Karp, that $\equiv_{\mathcal{L}_{\infty,\omega}} = \equiv_{\mathcal{L}_{\infty,G}}$.

[15] See Kolaitis (1985), Remark 1.1.4 p. 370

[16] See Kolaitis (1985), p.398-399, Th. 3.15 and 3.2.1. Incidentally, this fact implies that the game quantifier G is logical in the sense of Bonnay. In this connection it is interesting to note that in $L_{\infty,G}$ one can write a statement asserting that $<$ *is a well-ordering of type* $\gamma+\gamma$ *for some ordinal* γ, a statement not expressible in $L_{\infty,\infty}$ (by a result of Malitz, 1966). Thus « game quantifiers give rise to infinitary logics which are different from the usual infinitary logics $L_{\lambda,\kappa}$ » (Kolaitis (1985), p.370).

In general, however, it is not possible to improve freely on the elementary classes of L while keeping $\equiv_L$ fixed, for the process of adding more logical constants to L modifies $\equiv_L$ as it modifies El_L. This is true even if the added quantifiers are invariant under the relation of elementary equivalence in the base logic : the quantifier *There exists infinitely many* ($Q_{\geq\aleph_0}$) is $\equiv_{FOL}$-invariant, but adding it to FOL modifies the elementary equivalence relation.[17] This fact suggests the following question : is there a natural property of generalized quantifiers (Q) which would garantee that elementary equivalence for the extended logic (L+Q) is the same as the elementary equivalence for the base logic (L) ? [18]

Moreover, it follows from the fact that $\equiv_{FOL}$ is not closed under definability that there is no ΔΣ-extension L of FOL such that $\equiv_{\mathcal{L}}=\equiv_{FOL}$. Given the centrality of FOL to our ordinary logical theorizing, the question arises whether there are natural ΔΣ-extension of FOL at all. It is shown in Bonnay (2008)[19] that the least fine-grained similarity relation whose invariants give rise to a functionally complete logic is the relation of potential isomorphism between structures. Hence the logic of Iso_p invariance would be a possible candidate. Is there any known logic which is functionally complete relative to Iso_p-invariants ? It is proved in Barwise (1973) that an operator Q is Iso_p-invariant iff for any set M, Q_M is definable in $L_{\infty,\omega}$. À result involving *uniform definability* of Iso_p invariants and $L_{\infty,\omega}$ would have proved the ΔΣ-closure of $L_{\infty,\omega}$. But the already mentionned fact that $El_{\mathcal{L}_{\infty,G}} \supset El_{\mathcal{L}_{\infty,\omega}}$, while $\equiv_{\mathcal{L}_{\infty,\omega}}=\equiv_{\mathcal{L}_{\infty,G}}$, shows that this is not possible. Remark also that since it is possible to define game theoric logics stronger than $\mathcal{L}_{\infty,G}$ and enjoying the Karp property[20], we know that $\mathcal{L}_{\infty,G}$ is not itself ΔΣ-closed. However, we do not know whether there are ΔΣ-closed logics with the Karp property, nor do we know, more generally, whether there are any (natural ?) ΔΣ-closed extension of FOL[21].

[17] Let $\mathcal{M} = \langle M, R\rangle$ and $\mathcal{M}' = \langle M, R'\rangle$ be such that R and R' are equivalence relations with R having an infinite number of equivalence class of arbitrarily big finite cardinality, and $\mathcal{M}'$ being just like $\mathcal{M}$ except for the fact that it contains also an infinitary equivalence class. Let ϕ be the sentence « $\overline{R}$ is an equivalence relation and $\exists x Q_{\geq\aleph_0} y\ x\overline{R}y$ » Then $\mathcal{M} \nvDash \phi$ and $\mathcal{M}' \vDash \phi$.See Bonnay (2008), p.51.

[18] I thank a referee for pointing out this question.

[19] p.51, Theorem 3.10

[20] See again Kolaitis (1985), p.396-399 for an example.

[21] I wish to thank Denis Bonnay and an anonymous referee for helpful suggestions. I'm also grateful to Allen Mann and Giulia Terzian for proof-reading my english.

Bibliographie

Liste des ouvrages cités dans le corps du texte.

Alston, William P. (1985). "Concepts of Epistemic Justification". Dans : *The Monist* 68.

— (1988). "The deontological conception of justification". Dans : *Philosophical Perspectives* 2.

— (2001). "A Realist Conception of Truth". Dans : *The Nature of Truth*. Éd. par Michael Lynch. MIT Press.

Atten, Mark van (2006). "Two draft letters from Gödel on Self-Knowledge of Reason". Dans : *Philosophia Mathematica* 14.2, p. 255–261.

Atten, Mark van et Juliette Kennedy (2003). "On the philosophical development of Kurt Gödel". Dans : *Bulletin of Symbolic Logic* 9.4.

— (2009). ""Gödel's modernism : on set-theoretic incompleteness," Revisited". Dans : *Logicism, Intuitionism, and Formalism : What has become of them ?* Éd. par Sten Lindström, Krister Palmgren et Viggo Stoltenberg-Hansen. Synthese Library. Springer, p. 303–356.

Barwise, Jon (1973). "Back and Forth through infinitary logic". Dans : *Model Theory*. Éd. par M. D. Morley. T. 8. Studies in Mathematics. Mathematical Association of America, p. 5–34.

Bell, J. L. et A. B. Slomson (1969). *Models and Ultraproducts*. North-Holland Publishing Company.

Belnap, Nuel D. (1961). "Tonk, Plonk and Plink". Dans : *Analysis* 22, p. 130–134.

Benthem, Johan van (1989). "Logical constants across varying types". Dans : *Notre Dame Journal of Formal Logic* 30.3, p. 315–341.

Betti, Arianna (2008). "Polish Axiomatics and its Truth : On Tarski's Lesniewskian Background and the Ajdukiewicz Connection". Dans : *New essays on Tarski and philosophy*. Éd. par Douglas Patterson. Oxford University Press.

Black, Max (1949). *Language and Philosophy*. Ithaca : Cornell University Press.

Blanché, Robert et Jacques Dubucs (1997). *La logique et son histoire*. Armand Colin.

Boghossian, Paul A. (1989). "The rule-following considerations". Dans : *Mind* 98.392, p. 507–549.

— (1996). "Analyticity reconsidered". Dans : *Noûs* 30.3, p. 360–391.

— (2000). "Knowledge of Logic". Dans : *New Essays on the A Priori*. Éd. par Paul Artin Boghossian et Christopher Peacocke. Oxford ; New York : Oxford University Press.

Bonnay, Denis (2006). "Qu'est-ce qu'une constante logique ?" Thèse de doct. Université Paris-I.

— (2008). "Logicality and invariance". Dans : *Bulletin of Symbolic Logic* 14.1, p. 29–68.

Boolos, George (1971). "The iterative conception of set". Dans : *Journal of Philosophy* 68, p. 215–232.

— (1984). "To Be Is to Be a Value of a Variable (or to Be Some Values of Some Variables)". Dans : *Logic, Logic and Logic*. First published in *Journal of Philosophy* 81 : 430-448. Cambridge, MA : Harvard University Press, p. 54–72.

Brandom, Robert B. (1994). *Making it explicit : reasoning, representing, and discursive commitment*. Harvard University Press.

Bucholz, W. et al. (1981). *Iterative induction definitions and subsystems of analysis : recent proof-theoretical studies*. Lecture notes in mathematics 897. Springer Verlag.

Burge, Tyler (1993). "Content preservation". Dans : *The Philosophical Review* 102.4, p. 457–488.

— (1996). "Our entitlement to self-knowledge". Dans : *Proceedings of the Aristotelian Society* 96.

Burgess, John et Gideon Rosen (1997). *A Subject with No Object : Strategies for Nominalistic Interpretation of Mathematics*. Oxford University Press.

Carnap, Rudolf (1963). "Intellectual autobiography". Dans : *The philosophy of Rudolph Carnap*. Éd. par P. A. Schilpp. T. 11. Librairy of Living Philosophers. La Halle : Open Court.

— (1966). *Philosophical foundations of physics*. Basic Books. Traduction fr. *Les fondements philosophiques de la physique*, Colin, 1973.

Casanovas, Enrique (2007). “Logical operations and Invariance”. Dans : *Journal of Philsophical Logic* 36.1.

Chisholm, Roderick (1977). *Theory of knowledge*. New Jersey : Prentice Hall.

Churchland, Paul et Patricia Churchland (1998). *On the contrary : critical essays 1987-1997*. MIT Press.

Churchland, Paul M. (1984). *Matter and Consciousness*. revised. Cambridge, MA. : MIT Press.

Cieslinsky, Cezary (2007). “Deflationism, conservativeness and maximality”. Dans : *Journal of Philosophical Logic*.

Clifford, William K. (1879). “The ethics of belief”. Dans : *Lectures and essays*. T. II. London : Macmillan.

Coffa, Alberto (1991). *The Semantic Tradition from Kant to Carnap : To the Vienna station*. Edited by Linda Wessels. Cambridge : Cambridge University Press.

Cohen, Jonathan (1989). “Belief and Acceptance”. Dans : *Mind* 98.391, p. 367–389.

Damnjanovic, Nic et Stewart Candlish (2007). “A brief history of Truth”. Dans : *Handbook of the Philosophy of Sciences*. Éd. par Dale Jacquette. T. 5 : Philosophy of logic. Elsevier.

Davidson, Donald (1984). *Inquiries into Truth and Interpretation*. Oxford : Oxford University Press.

— (1996). “The Folly of Trying to Define Truth”. Dans : *Journal of Philosophy* 93.

De Vidi, D. et G. Salomon (1999). “Tarski on ‘essentially richer’ metalanguages”. Dans : *Journal of Philosophical Logic* 28, p. 1–28.

Detlefsen, Michael (1979). “On interpreting Gödel’s second theorem”. Dans : *Journal of Philosophical Logic* 8.3, p. 297–313.

— (1986). *Hilbert’s program*. Éd. par J. Hintikka. Kluwer academic publisher.

— (1990). “On an alleged refutation of Hilbert’s program using Gödel’s first incompleteness theorem”’. Dans : *The Journal of Philosophical Logic* 18. Reprinted in Detlefsen (1992) pp.199–235.

Douven, Igor et Frank Hindriks (2005). “Deflating the correspondence intuition”. Dans : *Dialectica* 59.3.

Dretske, Fred (1981). *Knowledge and the flow of information*. MIT Press.

— (2000). “Entitlement : epistemic rights without duties ?” Dans : *Philosophy and Phenomenological Research* 60.3.

DUBUCS, Jacques (2003). "Carnap, Gödel et la nécessité mathématique". Dans : *Carnap Aujourd'hui*. Éd. par François LEPAGE et François RIVENC. Paris : Vrin-Bellarmin.

DUMMETT, Michael (1973). *Frege : Philosophy of Language*. Harper et Row.

— (1991). *The Logical Basis of Metaphysics*. The William James lectures ; 1976. Cambridge, Mass. : Harvard University Press.

EBBINGHAUS, H.-D. (1985). "Extended logics : The general framework". Dans : *Model-theoretic logics*. Éd. par Jon BARWISE et Solomon FEFERMAN. North-Holland Publishing Company. Chap. II.

ENGEL, Pascal (1998). "Believing, holding true, and accepting". Dans : *Philosophical explorations* 1.2, p. 140–151.

— éd. (2000). *Believing and Accepting*. Springer.

— (2001). "Is Truth a Norm ?" Dans : *Interpreting Davidson*. Éd. par Petr KOTATKO, Peter PAGIN et Gabriel SEGAL. CSLI, p. 37–51.

ENGEL, Pascal et Julien DUTANT, éds. (2005). *Philosophie de la connaissance*. Vrin.

FEFERMAN, Solomon (1960). "Arithmetization of meta-mathematics in a general setting". Dans : *Fundamenta Mathematicae* LXIX.

— (1962). "Transfinite recursive progressions of axiomatic theories". Dans : *Journal of Symbolic Logic* 27, p. 259–316.

— (1989). "Kurt Gödel : Conviction and Caution". Dans : *Gödel's theorems in Focus*. Éd. par S.G. SHANKER. Routledge, p. 96–115.

— (1991). "Reflecting on Incompleteness". Dans : *Journal of Symbolic Logic* 51, p. 1–48.

— (1999*a*). "Does mathematics need new axioms ?" Dans : *The American mathematical monthly* 106.2, p. 99–111.

— (1999*b*). "Logic, Logics, and Logicism". Dans : *Notre Dame Journal of Formal Logic* 40.1, p. 31–54.

— (2005). "Predicativity". Dans : *The Oxford Handbook of Philosophy of Mathematics and Logic*. Éd. par Stewart SHAPIRO. Oxford University Press, p. 590–624.

FEFERMAN, Solomon et al. (2000). "Does mathematics need new axioms ?" Dans : *Bulletin of Symbolic Logic* 6.4, p. 401–446.

FELDMAN, Richard (2000). "The ethics of belief". Dans : *Philosophy and Phenomenological Research* LX.3.

FIELD, Hartry (1972). "Tarski's Theory of Truth". Dans : *Journal of Philosophy* 69, p. 347–375.

FIELD, Hartry (1977). "Logic, meaning and conceptual role". Dans : *Journal of Philosophy* 74.7, p. 379–409.
— (1978). "Mental Representation". Dans : *Erkenntnis* 13, p. 9–61.
— (1980). *Science Without Numbers : A Defence of Nominalism*. Library of philosophy and logic. Oxford : Blackwell.
— (1986). "Stalnaker On Intentionality". Dans : *Pacific Philosophical Quarterly* 67, p. 98–112.
— (1992). "Truth". Dans : *Philosophy of Science* 59.2, p. 321–330.
— (1999). "Deflating the Conservativness Argument". Dans : *Journal of Philosophy* 96, p. 533–540.
— (2001). *Truth and the Absence of Fact*. Oxford University Press.
— (2002). "Saving the Truth Schema from Paradox". Dans : *Journal of Philosophical Logic* 31, p. 1–27.
— (2003). "The semantic paradoxes and the paradoxes of vagueness". Dans : *Liars and Heaps*. Éd. par J. C. BEALL et Michael GLANZBERG. Oxford University Press.
— (2005). "Reply to Barry Loewer". Dans : *Philosohical Studies* 124, p. 110–118.
— (2008). *Saving truth from paradox*. Oxford University Press.
FODOR, Jerry (1974). "Special sciences (or the disunity of science as a working hypothesis)". Dans : *Synthese* 28, p. 97–115.
— (1975). *The language of thought*. Harvard University Press.
— (1987). *Psychosemantics : the problem of meaning in the philosophy of mind*. MIT Press.
FRAASSEN, Bas van (1980). *The Scientific Image*. Oxford University Press.
— (1984). "Belief and the Will". Dans : *The Journal of Philosophy* 81.5, p. 235–256.
FRANZÉN, Torkel (2004*a*). *Inexhaustability : A Non-Exhaustive Treatment*. Peters, A.K.
— (2004*b*). "Transfinite progressions : a second look at completeness". Dans : *Bulletin of Symbolic Logic* 10, p. 367–389.
FREGE, Gottlob (1884/1969). *Les Fondements de l'Arithmétique*. Seuil.
— (1971*a*). *Ecrits logiques et philosophiques*. Éd. et trad. par C. IMBERT. Paris : Seuil.
— (1971*b*). *On the foundations of geometry and formal theories of arithmetic*. Translated and with an introduction by Eike-Henner W. Kluge. New Haven ; London : Yale University Press.

FREGE, Gottlob (1999). *Ecrits posthumes*. Éd. par Philippe (de) ROUILHAN et Claudine THIERCELIN. Jacqueline Chambon.

— (2000). *Idéographie*. Vrin.

FRIEDMAN, Harvey et Michael SHEARD (1987). "An axiomatic approach to self-referential truth". Dans : *Annals of Pure and Applied Logic* 33, p. 1–21.

GALINON, Henri (2009). "A note on generalized functional completeness in the realm of elementary logic". Dans : *Bulletin of the Section of Logic* 38.1, p. 1–9.

GEACH, Peter (1965). "Assertion". Dans : *Philosophical Review* 74.4, p. 449–465.

GENTZEN, Gerhard (1969). "Investigations into logical deduction". Dans : *The collected papers of Gerhard Gentzen*. Éd. par M.E. SZABO. Amsterdam : North-Holland.

GOCHET, Paul (1992). *Quine en perspective*. Paris : Flammarion.

GÖDEL, Kurt (1964). "What is Cantor's continuum problem ? [1964]". Dans : *Philosophy of Mathematics : Selected Reading (2nd ed.)* Éd. par Paul BENACERRAF et Hilary PUTNAM. Cambridge University Press, p. 470–485.

— (1986). "Uber Formal Unentscheidbare Sätze der *Principia Mathematica* und Verwandter Systeme I [1931]". Dans : *Gödel Collected Works*. Éd. par Solomon FEFERMAN. T. 1. English translation 'On formally undecidable propositions of Principia Mathematica and related systems I' ; trans. Jean van Heijenoort. Oxford : Oxford University Press, p. 335–398.

— (1990). "On an extension of finitary mathmatics which has not yet been used [1972/1958]". Dans : *Gödel Collected Works*. Éd. par Solomon FEFERMAN. T. 2. Oxford University Press.

GOLDMAN, Alvin (1999). "Internalism exposed". Dans : *The Journal of Philosophy* 96.

GREENBERG, Mark et Gilbert HARMAN (2006). "Conceptual role semantics". Dans : *Oxford Handbook of philosophy of language*. Éd. par SMITH, BARRY et Ernest LEPORE. Oxford U.P.

GRICE, H. P. (1991). "Logic and conversation". Dans : *Studies in the Way of Words*. Publié pour la première fois dans *Syntax and Semantics*, volume 3. Peter Cole and Jerry L. Morgan, eds. (New York : Academic Press, 1975), pp.41-58. Harvard University Press.

GROVER, Dorothy (1992). *A Prosentential Theory of Truth*. Princeton University Press.

GROVER, Dorothy, Joseph CAMP et Nuel BELNAP (1975). "A Prosentential Theory of Truth". Dans : *Philosophical Studies* 27, p. 73–125.

GUPTA, Anil (1993). "Minimalism". Dans : *Philosophical Perspectives* 7.

HACKING, Ian (1979). "What is logic ?" Dans : *Journal of Philosophy* 76, p. 285–319.

HAJÈK, Alan (2008). "Dutch Book Arguments". Dans : *Oxford Handbook of Rational and Social Choice*. Éd. par Paul ANAND, Prasanta PATTANAIK et Clemens PUPPE. Oxford University Press.

HÀJEK, Petr et Pavel PUDLÀK (1993). *Metamathematics of first-Order Arithmetic*. Springer Verlag.

HALBACH, Volker (1999*a*). "Conservative theories of classical truth". Dans : *Studia Logica* 62, p. 353–370.

— (1999*b*). "Disquotationalism and Infinite Conjunctions". Dans : *Mind* 108, p. 1–22.

— (2001). "How Innocent is Deflationism ?" Dans : *Synthese* 126.1-2.

HARMAN, Gilbert (1982). "Conceptual role semantics". Dans : *Notre Dame Journal of Formal Logic* 23.2.

HECK Jr, Richard (1997). "Tarski, Truth and Semantics". Dans : *Philosophical Review* 107, p. 533–554.

HINTIKKA, Jaakko (1996). *Lingua Universalis vs. Calculus Ratiocinator : an ultimate presupposition of Twentieth-century logic*. Dordrecht : Kluwer academic publisher.

HODES, Harold (2004). "On the sense and reference of a logical constant". Dans : *Philosophical Quarterly*.

HORSTEN, Leon (2001). "Platonistic Formalism". Dans : *Erkenntnis* 54, p. 173–194.

— ([forthcoming]). "Levity". Dans : *Mind*.

HORWICH, Paul (1998*a*). *Meaning*. Oxford University Press.

— (1998*b*). *Truth*. 2nd. Oxford University Press.

— (2006). "The Value of Truth". Dans : *Noûs* 40.2, p. 347–360.

ISAACSON, Daniel (1987). "Arithmetical truth and hidden higher-order concepts". Dans : *Logic Colloqium'85*. Éd. par Paris Logic GROUP. Amsterdam : North-Holland, p. 147–169.

— (1992). "Some considerations on arithmetical truth and the omega-rule". Dans : *Proof, Logic and Formalization*. Éd. par Michael DETLEFSEN. Routledge, p. 94–138.

Isaacson, Daniel (1996). “Arithmetical truth and hidden higher-order concepts”. Dans : *The Philosophy of Mathematics*. Éd. par W.D. Hart. Oxford Readings in Philosophy. Oxford University Press. Chap. X, p. 203–224.

James, William (1956). “The Will to Believe [1896]”. Dans : *The will to believe and other essays in popular philosophy*. New York : Dover Publications.

Kant, Immanuel (2000). *Logique*. Éd. par Jaeger. Paris : Vrin.

— (2006). *Critique de la Raison Pure (1787)*. Éd. par A. Renaut. Paris : GF-Flammarion.

Kaplan, David (1989). “Demonstratives”. Dans : *Themes from Kaplan*. Éd. par J. Almog, J. Perry et H. Wettstein. New York ; Oxford : Oxford University Press. Chap. 17, p. 481–563.

Kaye, Richard (1991). *Models of Peano Arithmetic*. Oxford University Press.

Ketland, Jeffrey (1999). “Deflationism and Tarski’s Paradise”. Dans : *Mind* 108, p. 69–94.

— (2000). “A proof of the (strengthened) Liar formula in a semantical extension of Peano Arithmetic”. Dans : *Analysis* 60.1.

— (2005). “Deflationism and the Gödel-Phenomena : Reply to Tennant”. Dans : *Mind* 114.1, p. 75–88.

Kim, Jaegwon (1989). “The myth of non-reductive materialism”. Dans : *APA Proceedings*. T. 63. 3, p. 31–47.

— (1996). *Philosophy of Mind*. Westview Press.

Kleene, Stephen C. (1952). “Finite axiomatizability of theories in the predicate calculus using additional predicate symbols”. Dans : *Memoirs of the Amercian Mathematical society* 10, p. 27–68.

Koellner, Peter (2009). “On reflection principles”. Dans : *Annals of Pure and Applied Logic* 157.2-3, p. 206–219.

Kolaitis, Ph. G. (1985). “Game Quantification”. Dans : *Model-theoretic logics*. Éd. par Jon Barwise et Solomon Feferman. North-Holland Publishing Company. Chap. X.

Kotlarski, H. (1991). “Full Satisfaction Class : A Survey”. Dans : *Notre Dame Journal of Formal Logic* 32.4, p. 573–579.

Kreisel, Georg (1970). “Principles of proof and ordinals implicit in given concepts”. Dans : *Intuitionism and Proof Theory*. Éd. par A Kino, John Myhill et R. E. Vesley. North-Holland, p. 489–516.

KREMER, Michael (1988). "Kripke and the logic of truth". Dans : *Journal of Philsophical Logic* 17, p. 225–278.

KRIPKE, S. A. (1976). "Is there a problem about substitutional quantification ?" Dans : *Truth and Meaning : essays in semantics*. Éd. par Gareth EVANS et John MCDOWELL, p. 325–419.

KRIPKE, Saul (1975). "Outline of a theory of truth". Dans : *The Journal of Philosophy* 72.2, p. 690–716.

— (1984). *Wittgenstein on Rules and private language*. Blackwell.

LAUGIER, Sandra (2002). *L'anthropologie logique de Quine*. Paris : Vrin.

LEEDS, Stephen (1978). "Theories of Reference and Truth". Dans : *Erkenntnis* 13.1, p. 111–129.

LEITGEB, Hannes (2005). "What truth depends on". Dans : *Journal of Philosophical Logic* 34, p. 155–192.

— (2007). "What theories of truth should be like (but cannot be)". Dans : *Philosophy Compass* 2.2.

LESNIEWSKI, S. (1929). "Grundzüge eines neuen Systems der Grundlagen der Mathematik". Dans : *Fundamenta Mathematicae* 14, p. 1–81.

LEWIS, David (1970). "How To Define Theoritical Terms". Dans : *The Journal of Philosophy* 67.13, p. 427–446.

— (1975). "Language and languages". Dans : *Philosophical Papers*. T. I. First published in *Minessota Studies in the Philosophy of Science* Vol. VII (University of Minnesota Press, 1975) : 3–35. Oxford, New York : Oxford University Press, p. 163–188.

— (2001). "Forget about the "correspondence theory of truth"". Dans : *Analysis* 61.272, p. 275–280.

LINDSTRÖM, Per (1966). "First Order Predicate logic with generalized quantifiers". Dans : *Theoria* 32, p. 186–195.

LOAR, Brian (1980). "Ramsey's theory of belief and truth". Dans : *Prospects for pragmatism*. Éd. par D.H. MELLOR. Cambridge University Press.

— (1982). "Conceptual Role and Truth-Conditions". Dans : *Notre Dame Journal of Formal Logic* 23.2, p. 272–283.

LOEWER, Barry (1997). "A guide to naturalizing semantics". Dans : *The Blackwell companion to the philosophy of language*. Éd. par B. HALE et C. WRIGHT. Blackwell.

Loewer, Barry (2005). "On Field's *Truth and the absence of Facts* - comment". Dans : *Philosohical Studies* 124, p. 59–70.

MacAllister, James (1996). *Beauty and revolution in Science*. Ithaca, NY : Cornell University Press.

MacFarlane, John (2000). "What does it mean to say that logic is formal ?" Thèse de doct. University of Pittsburgh.

Maher, Patrick (1993). *Betting on theories*. Cambridge Studies in Probabilityn Induction and Decision Theory. Cambridge University Press.

Mancosu, Paolo (2008*a*). "Quine and Tarski on Nominalism". Dans : *Oxford studies in Metaphysics*. Éd. par Dean Zimmerman. T. 4.

— (2008*b*). "Tarski, Neurath, and Kokoszynska on the Semantic Conception of Truth". Dans : *New essays on Tarski and philosophy*. Éd. par Douglas Patterson. Oxford University Press.

— (2009). "Tarski's engagment with Philosophy". Dans : *The Golden Age of Polish Philosophy*. Éd. par Sandra Lapointe et al. T. 16. Logic, Epistemology, and the Unity of Science. Springer.

Martin-Löf, Per (1996). "On the meanings of logical constant and the justifications of the logical laws". Dans : *The Nordic Journal of Philosophy* 1.1, p. 11–60.

Mautner, F.I. (1946). "An extension of Klein Erlanger program : logic as invariant-theory". Dans : *American Journal of Mathematics* 68, p. 345–384.

McGee, Vann (1991). *Truth, Vagueness and Paradox*. Hackett Pub.

— (1992). "Maximal Consistent sets of instances of Tarski's Schema (T)". Dans : *Journal of Philosophical Logic* 21.

— (1996). "Logical operations". Dans : *Journal of Philosophical Logic* 25, p. 567–580.

— (2006). "In praise of free lunch". Dans : *Self-reference*. Éd. par V. Hendricks, T. Bolander et A. Pedersen. CSLI Publications.

McGrath, Matthew (2003). "Deflationism and the normativity of truth". Dans : *Philosophical Studies* 112, p. 47–67.

Montague, Richard (1970). "English as a Formal Language". Dans : *Linguaggi nella Società e nella Tecnica*. Éd. par Bruno Visentini et al. Milan : Edizioni di Comunità.

Moore, G.E. (1942). "A reply to my critics". Dans : *The philosophy of G.E. Moore*. Éd. par P.A. E. Schlipp. Evanston, IL : Northwestern University.

Moschovakis, Yiannis (2008). *Elementary induction on abstract structures*. Dover Publications.

Mostowski, A. (1957). "On a generalization of quantifiers". Dans : *Fundamenta Mathematicae* 44, p. 12–37.

Myhill, John (1960). "Some remarks on the notion of proof". Dans : *Journal of Philosophy* 57.14, p. 461–471.

Neuman, J. von (1966). "Theory of Self-reproducing automata". Dans : éd. par A.W Burks. Urbana : University of Illinois.

Neurath, O. (1936). "Erster Internationaler Kongress für Einheit der Wissenschaft in Paris 1935". Dans : *Erkenntnis* 1, p. 377–430.

Parsons, Charles (1983). *Mathematics in philosophy*. Cornell University Press.

Peters, S. et D Westerståhl (2006). *Quantifiers in Language and Logic*. Oxford : Oxford University Press.

Pietroski, Paul (2005). "Meaning before truth". Dans : *Contextualism in philosophy*. Éd. par G. Preyer et G. Peters. Oxford U.P.

Plantinga, Alvin (1993). *Warrant : the current debate*. Oxford University Press.

Prawitz, Dag (1965). *Natural Deduction*. Uppsala : Almquist et Wicksell.

— (1971). "Ideas and Results in proof-theory". Dans : *Proceedings of the 2. Scandinavian Logic Symposium*. Éd. par J. Fenstad. Noth-Holland, p. 237–309.

— (2006). "Meaning approached via proofs". Dans : *Synthese* 148, p. 507–524.

Price, Huw (1998). "Three norms of assertibility". Dans : *Philosophical Perspectives* 12, p. 240–254.

Prior, Arthur N. (1960). "The runabout inference-ticket". Dans : *Analysis* 21, p. 38–39.

Pryor, Jim (2001). "Highligts of recent epistemology". Dans : *British Journal of Philosophy of Science* 52, p. 95–124.

Putnam, Hilary (1971). *Philosophy of Logic*. New York : Harper et Row.

— (1978*a*). "Meaning and knowledge". Dans : *Meaning and the moral sciences*. London : Routledge.

— (1978*b*). *Meaning and the moral sciences*. London : Routledge et Kegan Paul.

Quine, W. V. O. (1950). *Methods of Logic*. New York : Henry Holt & Co.

— (1953*a*). *From a Logical Point of View*. Oxford : Oxford University Press.

— (1953*b*). "Notes on the theory of reference". Dans : *From a Logical Point of View*. Harvard University Press.

QUINE, W. V. O. (1960*a*). "Variables explained away". Dans : *Proceedings of the american philosophical society* 104.3, p. 343–347.

— (1960*b*). *Word and Object*. Cambridge, Mass. : MIT Press.

— (1966). "The Ways of Paradox". Dans : *The Ways of Paradox and Other Essays*. Harvard University Press. Chap. 1.

— (1969). "Epistemology naturalized". Dans : *Ontological Relativity and Other Essays*. Columbia University Press.

— (1970). *Philosophy of logic*. Cambridge, Mass : Harvard University Press.

— (1976*a*). "Carnap on logical truth". Dans : *The Ways of Paradox and Other Essays*. Cambride, MA : Harvard University Press.

— (1976*b*). *The Ways of Paradox and Other Essays*. 2e éd. First edition published 1966. Cambridge, MA et London : Harvard University Press.

— (1990). *Pursuit of Truth*. Harvard University Press.

— (1992). *Quiddités*. Paris : Seuil.

— (2008). *Philosophie de la logique*. Éd. par Tr. fr. par J. LARGEAULT. Aubier Flammarion.

RAMSEY, Frank P. (1927). "Facts and Propositions". Dans : *Proceedings of the Aristotelian Society* 7, p. 153–170.

— (1931). "Truth and Probability". Dans : *The Foundations of Mathematics and other Logical Essays*. Éd. par R. B. BRAITHWAITE. London : Routledge et Kegan Paul, p. 156–198.

— (1991). "On Truth". Dans : *Episteme* 16, p. 1–16.

RAY, Greg (2005). "On the matter of essential richness". Dans : *Journal of Philosophical Logic* 34, p. 433–457.

RAYO, Agustin et Gabriel UZQUIANO, éds. (2006). *Absolute generality*. Oxford University Press.

READ, S. (2000). "Harmony and autonomy in classical logic". Dans : *Journal of Philosophical Logic* 29, p. 123–154.

RESNIK, Michael D. (1989). "On the philosophical significance of consistency proofs". Dans : éd. par S.G. SHANKER. Routledge. Chap. VI, p. 115–130.

— (1990). "Immanent Truth". Dans : *Mind* 99.395, p. 405–424.

RESTALL, Greg (2005). "Multiple Conclusions". Dans : *Logic, Methodology and Philosophy of Science : Proceedings of the Twelfth International Congress*. Éd. par P. HAJEK, L. VALDES-VILLANUEVA et D. WESTERSTAHL. Kings' College Publications, 2005, 189–205., p. 189–205.

Rivenc, François (1993). *Recherches sur l'universalisme logique*. Paris : Payot.

— (1998*a*). "Ce que Ramsey a vraiment dit, ou la théorie prophrastique de la vérité". Dans : *Philosophie* 57, p. 16–50.

— (1998*b*). *Semantique et Vérité : de Tarski à Davidson*. Paris : Presses Universitaires de France.

— (2001). "Définition et critère de vérité". Dans : *Philosophie* 65.

— (2008). *Lecture de Quine*. Cahiers de Logique et d'Epistémologie. College Publications.

Rojszczak, Artur (2002). "Philosophical background and philosophical content of the semantic definition of truth". Dans : *Erkenntnis* 56, p. 29–62.

— (2005). *From the act of judging to the sentence : the problem of truth bearers from Bolzano to Tarski*. Springer Verlag.

Rouilhan, Philippe (de) (1984). "Le Menteur. Sur la théorie de la vérité de Tarski". Dans : *Le temps de la réflexion* 5, p. 271–291.

— (1998). "Tarski et l'universalité de la logique". Dans : *Le formalisme en Question*. Éd. par Frédérique Nef et Denis Vernant. Paris : Vrin.

— (2002). "On What There Are". Dans : *Proceedings of the Aristotelian Society* 102.2.

Rouilhan, Philippe (de) et Serge Bozon (2006). "La vérité de IF : Hintikka a-t-il vraiment exorcisé la malédiction de Tarski ?" Dans : *The Philosophy of Jaako Hintikka*. Éd. par R.E. Auxier et L. E. Hahn. Open Court.

Rumfitt, Ian (2000). ""Yes" and "No"". Dans : *Mind* 109.436, p. 781–823.

Russell, Bertrand (1903). *The Principles of Mathematics*. Cambridge : Cambridge University Press.

— (1913/1992). *Theory of knowledge*. Éd. par E. Ramsden et K. Blackwell. Routledge.

— (1919). *Introduction to mathematical philosophy*. London : Allen et Unwin.

Sambin, G., G. Battilotti et C. Faggian (2000). "Basic Logic : reflection, symmetry, visibility". Dans : *Journal of Symbolic Logic* 65, p. 979–1013.

Shapiro, Stewart (1983). "Conservativeness and Incompleteness". Dans : *Journal of Philosophy* 80.9, p. 521–531.

— (1998*a*). "Induction and indefinite extensibility : The Gödel sentence is true, but did someone change the subject ?" Dans : *Mind* 107.427, p. 597–624.

— (1998*b*). "Proof and Truth : Through Thick and Thin". Dans : *Journal of Philosophy* 95.10, p. 493–521.

Sher, Gila (1991). *The Bounds of Logic : A Generalized Viewpoint*. MIT Press.

Smiley, Timothy (1996). "Rejection". Dans : *Analysis* 56.1.

Smith, Nicholas J.J. (2000). "Frege's judgment stroke". Dans : *Australasian Journal of Philosophy* 78.2, p. 153–175.

Smorynski, C. (1977). "The Incompleteness Theorems". Dans : *Handbook of Mathematical Logic*. Éd. par J. Barwise. Amsterdam : North-Holland, p. 821–865.

Soulez, Antonia, éd. (1985). *Manifeste du Cercle de Vienne et autres écrits*. Presses Universitaires de France.

Stainton, Robert J. (2006). "Meaning and Reference - Some Chomskian Themes". Dans : *Handbook of the philosophy of language*. Éd. par E Lepore et B. Smith. Oxford University Press.

Steup, Matthias (1988). "The Deontic conception of justification". Dans : *Philosophical Studies* 53.

— (2000). "Doxastic volontarism and epistemic deontology". Dans : *Acta Analytica* 15.

Strawson, P. F. (1949). "Truth". Dans : *Analysis* 9, p. 83–97.

Sundholm, Goran (1998). "Inference vs. Consequence". Dans : *The Logica Yearbook 1997*. Éd. par Tim Childers. Filosofia, p. 26–35.

— (2004). "The proof-explanation of logical constants is logically neutral". Dans : *Revue Internationale de Philosophie* 58.4, p. 401–410.

Tarski, Alfred (1935). "The Concept of Truth in Formalized Languages". Dans : *Logic, Semantics and Metamathematics*. 2nd. JH Woodger (trans.) ; First published as 'Der Wahrheitsbegriff in Den Formaliserten Sprachen', *Studia Philosophica* I (1935). Indianapolis (1983) : Hackett Pub.

— (1936/2009). "Du concept de conséquence logique". Dans : *Philosophie de la logique*. Éd. par D. Bonnay et M. Cozic. Vrin.

— (1939). "On undecidable statements in enlarged systems of logic and the concept of truth". Dans : *Journal of Symbolic Logic* 4.3, p. 105–112.

— (1944). "The Semantic Conception of Truth and the Foundations of Semantics". Dans : *Philosophy and Phenomenological Research* 4.3, p. 341–376.

— (1956). *Logic, Semantics, Metamathematics : Papers from 1923 to 1938*. Oxford University Press.

— (1966/1986). "What are Logical Notions ?" Dans : *History and Philosophy of Logic* 7, p. 143–154.

— (1969). "Truth and proof". Dans : *Scientific American* June, p. 63–70.

Tarski, Alfred (1983). *Logic, Semantics, Metamathematics*. Éd. par John Corcoran. Hackett pub.

Tarski, Alfred et R.L. Vaught (1957). "Arithmetical extensions of relational systems". Dans : *Compositio mathematica* 13, p. 81–102.

Tennant, Neil (1982). "Proof and Paradox". Dans : *Dialectica* 36, p. 265–296.

— (1987). *Antirealism and logic. Truth as eternal*. Oxford et New York : Oxford University Press.

— (2002). "Deflationism and the Gödel-Phenomena". Dans : *Mind* 111.

Wang, Hao (1974). *From mathematics to philosophy*. Routledge et Kegan Paul.

Westerståhl, Dag (2007). "Quantifiers in formal and natural languages". Dans : *Handbook of Philosophical Logic, 2nd edition*. Éd. par Dov M. Gabbay et F. Guenthner. T. 14. Springer Netherlands, p. 223–338.

Williams, John N. (2004). "Moore's paradox, Evan's principle and self-knowledge". Dans : *Analysis* 64.4, p. 348–53.

Williamson, Timothy (2000). *Knowledge and its Limits*. Oxford : Oxford University Press.

Wittgenstein, Ludwig (1922/1993). *Tractatus logico-philosophicus*. Gallimard.

— (2004). *Recherches Philosophiques*. Gallimard.

Wolenski, Jan (2009). "The Rise and Development of logical semantics in Poland". Dans : *The Golden Age of Polish Philosophy*. Éd. par Sandra Lapointe et al. T. 16. Logic, Epistemology, and the Unity of Science. Springer.

Wolenski, Jan et Roman Murawski (2008). "Tarski his Polish predecessors on Truth". Dans : *New essays on Tarski and philosophy*. Éd. par Douglas Patterson. Oxford University Press.

Wright, Crispin (1999). "Truth : A Traditional debate reviewed". Dans : *Canadian Journal of Philosophy* 24.

— (2001). "On basic logical knowledge". Dans : *Philosophical Studies* 106.1-2, p. 41–85.

— (2004*a*). "Intuition, entitlement and the epistemology of logical laws". Dans : *Dialectica* 58.1.

— (2004*b*). "Warrant for nothing (and foundations for free) ?" Dans : *Proceedings of the Aristotelian Society* 78.1, p. 167–212.

Index des noms propres

Index des titres des articles et ouvrages cités

Index des notions

Table des figures

Liste des tableaux

Recherches sur la vérité - définition, élimination, déflation

Résumé : La vérité, dit-on, est un des buts de la science. Mais quelle est la place de la notion de vérité elle-même dans le langage de la science ? La notion de vérité peut-elle être suffisamment clarifiée ? Et si oui, quelle peut être sa contribution au discours scientifique, pour quels usages la notion de vérité peut-elle être mobilisée ? Ce travail cherche à répondre à ces questions. Sa thèse principale est que la notion de vérité s'apparente à une notion logique. Cette idée s'inscrit dans un courant de réflexion contemporain sur la vérité appelé « déflationnisme ». Mais la formulation elle-même est nouvelle, comme sont nouveaux les arguments mis en œuvre pour l'étayer. Négativement, une critique détaillée de certains arguments *a priori* développés contre le déflationnisme est présentée. Positivement, nous caractérisons, d'une part, une classe critique d'affirmations mettant en jeu la notion de vérité comme ensemble de moyens d'expliciter des contenus déjà implicitement acceptés, et nous introduisons d'autre part des considérations et des outils permettant de donner un sens précis à la thèse de la logicité de la notion de vérité.

Inquiries into the concept of truth - definition, elimination, deflation

Summary : Truth, it is said, is one of the aims of scientific inquiry. But what is the place of the notion of truth itself within scientific inquiry ? Can the notion be explained in a satisfactory way ? If so, how, in turn, could it contribute to scientific discourse, in what ways ? In this work we aim to answer those questions. Our main thesis is that truth is akin to a logical notion. The inspiration behind this claim is not new to this work, which actually can be seen as sequel to recent work on so-called « deflationary conceptions of truth ». The claim itself, however, is new, as are the arguments and ideas brought out to sustain it. On a defensive side, prominent *a priori* counterarguments to deflationism are discussed in details and carefully refuted on shared grounds. On the constructive side, we introduce and discuss a new criterion to distinguish between « substantial » and « non-substantial » uses of the notion of truth and new considerations to assess its logical character.

Discipline : Philosophie

Mots-clés : Philosophie de la logique ; Logique ; Vérité ; Épistémologie ; Philosophie des mathématiques ; Rationalité ; Constante logique ; Invariance

Équipe d'accueil : Institut d'Histoire et de Philsophie des Sciences et des Techniques (UMR 8590)
13, rue du Four
75006, Paris.

École doctorale : École doctorale de Philosophie de l'Université Paris 1 (ED 280)
1, rue d'Ulm
75005, Paris.