**Comprendre le Sanskrit : de la Grammaire de Pāṇini aux Machines d'Eilenberg**

Gérard Huet

Académie des Sciences

Directeur de Recherches Emérite Inria

**Résumé**

Le sanskrit est la langue savante de la civilisation indienne. Il a été figé dès le 4ème siècle avant l'ère commune par une grammaire formelle exacte due au lettré Pāṇini, qui fut non seulement un linguiste de génie, mais un informaticien précurseur.

Le traitement informatique du Sanskrit a nécessité la mise au point d'algorithmes d'analyse morpho-phonétique sophistiqués, utilisant une méthodologie de programmation relationnelle inspirée des travaux du mathématicien Samuel Eilenberg, et pilotés par les règles d'euphonie de Pāṇini. L'exposé présentera brièvement cette méthodologie, illustrée par le traitement informatique d'exemples caractéristiques.

**Mystérieux sanskrit**

Comprendre le sanskrit, c'est un vaste programme. D'abord, qu'est-ce que le sanskrit ? Toute personne cultivée sait qu'il s'agit d'une langue de l'Inde ancienne. Mais, au contraire des langues classiques comme le grec ancien et le latin, le sanskrit a une aura de mystère, de bizarre, d'hermétisme, d'ésotérisme. Récemment, dans une série policière à la télévision, l'enquêteur, qui n'arrive pas à se faire comprendre de son subordonné, se fâche et lui dit : « Comment faut-il vous l'expliquer ? En sanskrit ? En largonji des louchébems ? » et cette juxtaposition du sanskrit et de l'argot des bouchers en dit long sur la place du sanskrit dans l'inconscient collectif populaire.

Même dans des cercles intellectuels sophistiqués, ce préjugé est tenace. Au collège Balliol, qui à l'Université d'Oxford a le privilège d'héberger la chaire de sanskrit, il est de tradition que le Master du collège, en recevant un candidat Fellow, l'interroge ainsi : « En tant que Fellow de Balliol College, vous serez amenés à dîner à la High Table. Il pourra se faire que vous y soyez assis à côté du Professeur de sanskrit. De quoi parlerez vous avec lui ? » Cette question sous-entend qu'un universitaire qui arrive à trouver un sujet de discussion intéressant avec le Professeur de sanskrit sera capable de se débrouiller socialement en toute circonstance : « The topic of Sanskrit is assumed to be dislocated from all customary intellectual pursuits, positioned, as it were, in the academic world's realm of imaginary numbers » [Minkowsky\_2006].

Mais enfin, pourquoi cette réputation mystérieuse ? Le sanskrit est reconnu comme l'une des grandes langues de civilisation, au même titre que le grec ancien ou l'hébreu, même s'il ne fut découvert en Occident que tardivement. La chaire de langue et littérature sanscrites au Collège de France fut inaugurée par Antoine-Léonard de Chézy le 16 janvier 1815, en même temps qu'Abel Rémusat inaugurait la chaire de langue et littérature chinoises [Levi\_1895]. La découverte du sanskrit en Occident, et de ses affinités avec de nombreuses langues européennes, fut le départ de la linguistique comparative, en mettant en évidence la branche des langues indo-européennes. Le linguiste suisse de Saussure étudia le sanscrit et la phonétique indo-européenne dès 1876, et publia sa thèse sur le génitif absolu en sanscrit en 1880. Sa connaissance de la grammaire de Pāṇini est à la base de son cours de Linguistique Générale, qui est généralement considéré comme l'origine de la linguistique occidentale moderne. Une introduction au sanskrit fait partie encore aujourd'hui de l'enseignement de base de la discipline.

Pourtant, la position du sanskrit dans la grande classification des langues de l'humanité reste controversée. Le védique et son frère jumeau l'avestique figurent parmi les plus anciennes langues attestées de la famille indo-européenne. Mais le védique, langue hiératique des poèmes du Véda, s'il est certes un ancêtre manifeste du sanskrit, est une langue morte qui n'est plus comprise que par une poignée de spécialistes, alors que le statut du sanskrit classique est tout autre, puisqu'il possède une littérature immense, composée dans un idiome pratiquement invariant depuis 25 siècles, et qui continue d'être compris dans l'Inde moderne par les personnes cultivées. Le Rāmāyaṇa, ou geste épique du héros déifié Rāma, est récité quotidiennement avec ferveur dans des milliers de temples hindous en Inde. Les indiens cultivés, qui ont souvent eu un apprentissage élémentaire du sanskrit à l'école primaire, s'y remettent volontiers à leur départ en retraite, et lisent la Bhagavādgīta dans le texte. Peut-on dire pour autant que le sanskrit est encore de nos jours une langue vivante ?

Le sanskrit fut certes inscrit dans la constitution indienne en 1949 parmi les 22 langues officielles du pays. Pourtant, son nombre de locuteurs est sujet à caution. Les différents recensements donnent des chiffres fluctuants et peu fiables, culminant à 20000 locuteurs, c'est-à-dire à un pourcentage infime de la population du pays. Le sanskrit n'est d'aucune utilité comme langue vernaculaire en Inde, nonobstant les légendes urbaines de villages parlant le sanskrit. Il existe bien des mouvements de renaissance de la langue, comme *Samskrita Bharati*, mais leur impact est très modeste. Pourtant, le sanskrit est encore aujourd'hui dans une certaine mesure une langue véhiculaire entre lettrés traditionnels. L'un d'eux me confiait avec une certaine coquetterie “ Je suis plus à l'aise en sanskrit qu'en hindi !”

Langue vivante ou langue morte, le statut du sanskrit est entre les deux. Des philologues prestigieux comme Sydney Pollock ont certes analysé que l'âge d'or du sanskrit en tant que littérature créative était révolu depuis le 17ème siècle et ainsi déclaré la mort du sanskrit [Pollock\_2001], alors que d'autres spécialistes ont une opinion plus nuancée [Hanneder\_2002]. Bien sûr cette question soulève des polémiques féroces avec les milieux nationalistes indiens. Mais le problème est plus compliqué qu'il n'y paraît, puisqu'après tout le sanskrit n'a jamais été à aucune époque la langue maternelle de quiconque. En effet, les femmes ne parlaient pas le sanskrit autrefois, à de rares exceptions près. Dans le théâtre classique, les rois et les prêtres s’expriment en sanskrit, mais les femmes et les serviteurs parlent le prakrit, la langue vernaculaire régionale. Donc de tous temps le sanskrit a été une langue apprise. Les jeunes brahmanes, à l'âge de 8 ans, prenaient pension chez leur maître d'école, qui leur inculquait le sanskrit par des récitations guidées, avant de leur réciter les textes en sanskrit de leur spécialité. Sur les photos suivantes on voit le maître inculquer l'accent védique par des mouvements de tête de l'élève, tandis que l'élève compte le mètre sur ses doigts (Plate 15 p. 180 de [Staal\_1983] Vol. 1. Photographies Adelaide deMenil).



Enseignement de l’accent védique

En fait, le prakrit (*prakṛta-bhāṣā*) signifie étymologiquement la langue naturelle. Les différents prakrits, ou dialectes moyen-indiens, étaient les langues vernaculaire du nord de l'Inde, descendantes du védique, comme la *magadhī*, que parlait le Bouddha. Ces langues ont plus tard évolué vers les langues modernes comme le hindi, le népali ou le bengali. Alors que le sanskrit (*saṃskṛta-bhāṣā*) signifie étymologiquement la langue raffinée ou apprêtée, c'est à dire une langue de registre élevé, obéissant strictement à une grammaire. Autrement dit, le sanskrit n'est pas à proprement parler une langue naturelle, c'est une langue artificielle dans une certaine mesure, un artefact de grammairien. Cette langue normalise la langue naturelle (*prakṛt*) utilisée à l'époque de Pāṇini (et des grammairiens qui l'ont précédé, et qu'il mentionne, mais dont il ne reste aucune autre trace).

On ne peut donc pas comparer le sanskrit et l'espéranto, qui est une création de toutes pièces comme patchwork de langues latines par un linguiste optimiste. Le langage décrit par la grammaire de Pāṇini était une langue familière aux locuteurs indigènes de l'époque, mais régularisée par des procédés morphologiques uniformes, et de plus avec une exigence de *faire sens*. Les formes complexes de la langue védique devenues caduques dans l'usage ne sont pas conservées, sauf à titre d'exception répertoriée comme telle (*chandasi*). Le gros de l'ouvrage définit une langue raffinée, châtiée, de registre élevé, et porteuse de sens précis. Une langue littéraire plus que véhiculaire.

La grammaire de Pāṇini ou Aṣṭādhyāyī (« En huit tomes ») explique tous les procédés productifs, à de minuscules exceptions près relevées par le grammairien Kātyāyana et analysées par le philosophe de la langue Patañjali dans son ouvrage magistral Mahābhāṣya (vers -150) qui fixa définitivement la norme du langage. La partie non védique de la grammaire devint donc le standard de fait du sanskrit classique. De descriptive, elle devint prescriptive, et le sanskrit cessa d'évoluer. Le sanskrit classique est donc une exception dans les langues de l'humanité : il est défini non par un corpus d'énonciations attestées, comme il est d'usage, mais par une grammaire formalisée. Même un ouvrage aussi ancien que le Mahābhārata, issu de la tradition orale de générations de bardes chantant des exploits héroïques, ne s'écarte de la norme qu'en une infime proportion de ses 90000 vers.

En fait, l'usage de la langue continua d'évoluer, mais strictement dans les limites de la grammaire. Ainsi, du temps de Pāṇini, les mots composés se limitaient à deux ou trois segments. Mais comme il avait décrit leur formation au moyen d’une règle récursive, la grammaire autorise des composés de longueur arbitraire, et les poètes se sont évertués à en construire des formes monstrueuses, mais justifiables grammaticalement.

**Pāṇini le grand raffineur**

Le grand linguiste américain Leonard Bloomfield, père de la linguistique structurelle, considérait Pāṇini comme son maître, et il disait de sa grammaire qu'elle était un modèle indispensable de la description des langues [Emeneau\_1988]. Il déclara même que l'Aṣṭādhyāyī était « one of the greatest monuments of human intelligence » . Elle a traversé les siècles à l'instar des *Éléments* d'Euclide, auxquels elle a d'ailleurs été comparée [Staal\_1965].

La grammaire de Pāṇini est un curieux document. Elle ne peut se comparer ni avec la grammaire de Port Royal, ni avec les grammaires formelles au sens de Chomsky. Elle se présente comme un formulaire de 4000 règles exécutables (*sūtra*). C'est en fait un programme (non déterministe) au sens informatique du terme. Mais ce n'est pas un programme qui est exécuté séquentiellement, de la règle {I,1,1} à la règle {VIII,4,68}, seulement la dernière partie de l'ouvrage a ce caractère séquentiel.

Le ‘programme’ est en fait une suite d'invocations des formules, sélectionnées par l'utilisateur pour exprimer la synthèse d'une énonciation sensée. Il s'agit donc d'une grammaire générative, où les énoncés du langage sont élaborés par un calcul.

La nature précise des calculs mis en jeu par l'utilisation de la grammaire de Pāṇini est complexe. Certains *sūtras* sont des définitions, d'autres sont des méta-règles régissant l'usage de la grammaire. Les plus nombreux sont des injonctions (*viddhi*), qui prescrivent un calcul. Un certain nombre de documents annexes sont nécessaires, comme une table des racines verbales données avec leurs paramètres combinatoires (*dhātupatha*). La grammaire manipule des formes, qui sont des suites de phonèmes accompagnées de marqueurs méta-linguistiques. Par exemple, à partir de la racine *kṛ* ‘faire’, on peut former un substantif d'agent de l'action par le sūtra {III,1,133}, qui prescrit l'utilisation du suffixe primaire ‘ṇvul’, constitué entièrement de marqueurs (*anubandha*). Le marqueur ‘ṇ’ spécifie une opération phonétique sur la racine, qui est portée au degré vocalique élevé (*vṛddhi*) par une autre règle qui produira le morphème *kār*. Le marqueur ‘vu’ est défini par le sūtra définitionnel {VII,1,1} comme le morphème *aka*, qui est affixé à la base *kār* pour produire *kāraka*. Le marqueur ‘l’ y spécifie la position de l'accent. La forme obtenue, signifiant « ce qui agit » peut maintenant être utilisée comme thème nominal pour se décliner par d'autres règles prescrivant les suffixes qui vont former ses formes fléchies, afin de produire par exemple l'instrumental *kārakeṇa*.

D'autres règles produisent les formes conjuguées des racines, leurs formes participiales, infinitives, absolutives aux différents aspects, temps, modes et voix, et leurs nuances sémantiques par affixation de préverbes. A partir de ces formes primaires, on forme des substantifs secondaires, à l'aide d'autres suffixes. Finalement, on peut fondre des formes nominales en une forme composée, en effaçant la déclinaison du segment initial, et suivant un certain nombre de schémas sémantiques. Au total, une morphologie complexe, mais expliquée très exactement par des procédés de réécriture de mots qui n'ont été dégagés en Occident qu'avec les systèmes de Post, un modèle universel de calcul basé sur la réécriture de chaînes de symboles [Post\_1936].

Les procédés morphologiques de formation des mots étant récursifs, notamment avec les composés nominaux, le vocabulaire est en fait infini. Les grammairiens ont inventé une anecdote mythologique où Indra, roi des dieux, décide d'être expert en sanskrit, et à cette fin demande à Bṛhaspati, le précepteur des dieux, de lui réciter la litanie des mots corrects. Mille années divines n'y suffirent pas, ils y sont encore. Les lettrés de Puṇe n’ont pas tenu compte de cet avertissement. Ils se sont lancés dans la construction d'un dictionnaire exhaustif, et 50 ans plus tard n'ont pas fini la lettre ‘*a*’.

La génération morphologique n'est qu'une partie de la grammaire. Un autre étage garantit la cohérence sémantique d'un ensemble de mots pour former une phrase. A cette fin, un réseau de dépendances sémantiques doit être établi. Les formes substantives sont pourvues d'un rôle sémantique (nommé *kāraka*, mot analysé plus haut) qui doit être cohérent avec le cas de leur déclinaison. Par exemple, une forme à l'instrumental peut recevoir le rôle d'agent pour satisfaire le ‘désir d'agent’ d'une forme verbale passive, ou bien le rôle d'instrument complément d’une forme verbale active. Plus généralement, la grammaire spécifie le désir (*ākāṅkṣā*) de chaque unité, et tous les désirs doivent être satisfaits pour que la phrase fasse sens dans une sémantique de situations. On retrouve là les notions de syntaxe structurale de Tesnière [Tesnière\_1959], qui ont conduit à la technologie des grammaires de dépendances en linguistique computationnelle contemporaine. Du point de vue calculatoire, la vérification de la cohérence des situations correspond à un calcul de satisfaction de contraintes combinatoires.

Le dernier étage du dispositif consiste à former à partir d'un ensemble sémantiquement cohérent de mots fléchis la liste de leurs formes dans un ordre d'énonciation spécifique - le *padapāṭha* ou suite des mots de la phrase. Une dernière passe fait un lissage phonétique facilitant la prononciation de l'énonciation correspondante, par application d'un procédé de fusion (*sandhi)* qui généralise la liaison du français. Ce lissage est effectué par des opérations répertoriées exhaustivement dans l'Aṣṭādhyāyī, et qui témoignent de l'expertise des grammairiens indiens antiques dans la science phonétique. Elles sont définies par des règles particulières à 4 composantes (x,u,v,w) spécifiant que dans un contexte gauche x, la fin de mot u s'agglutine au début du mot suivant v pour former la chaîne phonémique w. On reconnaît la forme des présentations de règles morpho-phonétiques de la morphologie computationnelle contemporaine [Kaplan\_1994, Mohri\_1997], développée en Occident seulement dans les années 1990.

Les différentes règles de la grammaire, utilisables aux différents niveaux structurels de la langue, sont agencées dans un ordre *a priori* déroutant, mais qui est nécessaire pour expliquer leurs incompatibilités mutuelles. Les règles sont condensées à l'extrême, pour permettre la mémorisation complète des 4000 règles du formulaire par un entraînement approprié. Qu'on en juge sur un exemple. La règle {III,1,133}, mentionnée ci-dessus comme engendrant des substantifs d'agents en -*aka*/-*ikā*, permet aussi d'engendrer une autre famille de substantifs d'agents en *-tṛ*/*-trī* avec le suffixe ‘tṛc’. Son énoncé est condensé dans la formule *‘*ṇvultṛcau’ où la finale ‘au’ évoque une forme au duel, maisdont le thème ‘ṇvultṛc’ est incompréhensible pour un non-spécialiste. Ces formules ne sont compréhensibles que par une étude approfondie, sous laconduite d'un maître, des commentaires de l'Aṣṭādhyāyī comme la glose de Bénarès (*kāśikāvṛtti*). Les *sūtras* ne sont pas plus du sanskrit que laformule mathématique ‘n!/k!(n-k)!’ n'est du français, ce qui n'empêche pas certains nationalistes indiens enthousiastes mais ignorantsd'affirmer que le sanskrit est le langage de programmation ultime,étudié en secret par la NASA, etc.

**Pāṇini linguiste et informaticien**

Les quelques remarques ci-dessus sur l'Aṣṭādhyāyī montrent que Pāṇini est assurément un linguiste de génie. La première évidence est l'utilisation d'un système de phonèmes de la langue, où les voyelles ont 3 degrés, tandis que les consonnes sont rangées dans cinq ‘couleurs’ (*varṇa*) représentant 5 points d'articulation, du larynx aux lèvres, et portent 4 traits phonétiques combinant les distinctions sourd/sonore et aspiré/non aspiré. On est stupéfié par cette rigueur toute algébrique de la discrétisation de l'espace phonétique de la parole. L'alphabet du français apparaît comme un galimatias absurde à cette lumière.

Rappelons que la linguistique moderne attribue l'invention de la notion de phonème au linguiste Troubetskoï du cercle de Prague (vers 1925), qui il est vrai connaissait l'œuvre de F. de Saussure, et donc le sanskrit, comme Bloomfield. Alors que les 50 phonèmes du sanskrit sont déjà en place dans la grammaire de Pāṇini. La recherche phonétique préalable avait déjà été accomplie par les écoles de récitation du Véda (*pratiśākhās*), et Pāṇini n'a fait qu'adopter la norme des phonéticiens qui l'ont précédé. Mais une innovation apparaît au tout début de la grammaire. Elle est en effet précédée d'un prélude de 14 formules très spéciales appelées *pratyāhārasūtrāṇi*, et qui définissent 14 listes de phonèmes, séparées par des marqueurs.

Par exemple, la 7ème liste énumère les 5 nasales. Cette curieuse structure de données sert en fait de définition très concise de tous les sous-ensembles de phonèmes utiles à conditionner les règles de la grammaire. On désigne un tel *pratyāhāra* ou énoncé condensé par la paire d'un phonème et d'un marqueur. Il dénote l'ensemble des phonèmes qui figure dans la table entre le phonème donné et le marqueur. Par exemple, le *pratyāhāra* ‘ac’ comprend toutes les voyelles, depuis {a, i, u} listés dans la première ligne, terminée par le marqueur ‘ṇ’, jusqu'aux diphtongues {ai, au} listés dans la quatrième ligne, terminée par le marqueur ‘c’ spécifié. Alors que ‘ic’ est ce même ensemble, auquel on a retiré a, ‘aṇ’ est juste {a, i, u}, etc. Ce codage permet une grande économie de notation, puisqu'une représentation de tous les sous-ensembles de l'ensemble de n phonèmes serait *a priori* de taille exponentielle en n, alors que l'agencement particulièrement astucieux des *pratyāhārasūtrāṇi* est linéaire : aucun phonème n'est répété, à une seule exception près (h).

La nature exceptionnelle de ce codage n'a été révélée que récemment par la mathématicienne allemande Wibke Petersen, qui a montré par une analyse fine de la combinatoire que le codage des *pratyāhārasūtrāṇi* est optimal [Petersen\_2004]. La méthode de synthèse de cette structure de données étonnante (attribuée par la tradition au dieu Śiva) est inconnue et très intrigante, au vu de la taille de l'espace de configurations possibles.

Nous avons déjà évoqué la description par Pāṇini des procédés morpho-phonétiques de description morphologique de la langue, ainsi que sa théorie des *kārakas* pour rendre compte des rôles sémantiques et de leur cohérence au sein d'une logique de situations. Elle est particulièrement élégante par son traitement du nominatif, qui ne désigne ni le sujet du verbe à l'actif, ni son objet au passif, mais sert seulement à les nommer, optionnellement, le rôle sémantique étant porté par la forme verbale. Ce système élégant et symétrique est supérieur aux approches transformationnelles de la syntaxe formelle [Kiparsky\_2009].

Au delà des nombreux concepts linguistiques de l'Aṣṭādhyāyī, montrant un état de connaissance des phénomènes de langue par des grammairiens indiens antiques qui ne sera redécouvert en Occident qu'au 20ème siècle, il convient de constater que Pāṇini a organisé leur utilisation effective au sein d'un modèle génératif de calcul qui le place définitivement parmi les pionniers de l'informatique. Les *sūtras* sont organisés de façon hiérarchique, pour permettre la factorisation des conditions contextuelles de leurs utilisations, dont la récurrence n'a pas à être répétée grâce au procédé de rémanence (*anuvritti*). Lorsque plusieurs règles sont applicables, elles sont listées par ordre décroissant de généralité, le cas général étant énoncé avant les cas particuliers, dont une métarègle de levée d'ambiguïté spécifie la priorité. Ce mécanisme évoque l'appel par filtrage dans les langages de programmation modernes. Un autre aspect de la solution apportée au problème du contrôle des calculs utilise une description hiérarchique des opérations qui évoque les classes des langages orientés-objet [Deo\_2003]. Enfin, partout dans la grammaire le souci de brièveté se traduit par des codages et méthodes de partage sophistiqués, qui montrent une prescience aiguë de la théorie de l'information. Pāṇini est donc non seulement un linguiste de génie, c'est un informaticien précurseur, 25 siècles avant l'apparition des ordinateurs et des langages de programmation.

De fait, la grammaire de Pāṇini est d'une telle complexité que son émulation exacte par un simulateur logiciel n'a pas été à ce jour complètement achevée, malgré des tentatives acharnées. En fait, il existe plusieurs difficultés conceptuelles. Par exemple, la grammaire ne se borne pas à manipuler les énonciations signifiantes, mais elle considère également leur signifié. Ainsi, certaines opérations ne sont permises sur une racine que si elle est utilisée dans un sens spécifique. Les constructions morphologiques sont exprimées comme des constructions sémantiques autant que combinatoires. Le suffixe primaire ‘tṛc’, qui construit des noms d'agent en -*tṛ*, est distingué de son homologue ‘tṛn’, qui construit les mêmes signifiants, mais avec un sens de procédé habituel, comme par exemple pour dénoter les professions ou les fonctions officielles. La langue peut ainsi être modélisée de manière très fine. Il se pose alors le problème de l'interférence de ces procédés sémantiques avec la combinatoire formelle des signifiants.

Cette difficulté se reflète dans différentes opinions sur la nature ontologique de cette grammaire. Une école de pensée considère la grammaire comme une espèce de vérificateur qui, étant donnée une énonciation correcte, en fournit le sens présumé. Une autre interprétation explique que la grammaire utilise le sens présumé comme guide pour fournir une énonciation grammaticalement correcte compatible avec ce sens. Aucun de ces deux points de vue n'est complètement satisfaisant, les conditions sémantiques et combinatoires étant mutuellement dépendantes. Il faut donc combiner ces deux points de vue, et considérer que les règles de Pāṇini manipulent en fait la paire (signifiant, signifié), à tous les niveaux de granularité, c'est à dire précisément la notion de signe linguistique inventée par de Saussure [Saussure\_1916]. On peut alors reconsidérer la grammaire comme une machine abstraite opérant sur les signes du sanskrit [Huet\_2016], une vision très sophistiquée du rôle d'une grammaire.

Alors, si l'on revisite la notion de programme sous-tendant les opérations grammaticales, ce programme ne doit être compris ni comme produisant le signifié à partir du signifiant, ni comme produisant le signifiant à partir du signifié, mais comme engendrant récursivement le signe combinant le signifiant et le signifié. Mais alors, quelle est la nature de la donnée d'entrée du programme ? Eh bien, c'est l'utilisateur qui sert d'oracle, en spécifiant progressivement les choix des *sūtras* pertinents. Cette donnée est à proprement parler l'intention communicative (*vivakṣā*) du locuteur. Il est significatif que le dialogue entre le maître et l'élève est de nature maïeutique : si l'élève produit une énonciation incorrecte, la charge de la preuve lui incombe : il doit donner la prakriyā, ou liste des *sūtras* qu'il pense pouvoir justifier la dérivation, jusqu'à ce que le maître en pointe l'application incorrecte, ou la production d'un sens non désiré.

Cette interprétation permet de rendre compte du fait que le contexte d'utilisation des règles doit être mémorisé, car certains marqueurs ont un rôle rémanent, même après leur élimination par substitution dans l'élaboration du signifiant. Il convient donc de garder l'historique des dérivations pertinentes à la synthèse des constituants du signe courant, c'est-à-dire précisément leur *prakriyā*. Cet élément contextuel de la sémantique n'est pas sans évoquer la notion de pleine abstraction (*full abstraction*) dégagée en sémantique dénotationnelle des langages de programmation.

Le sens de chaque construction est expliqué par une paraphrase canonique en sanskrit. Lorsque l'on compose ces paraphrases constituantes du signifié, le sanskrit utilisé est en fait un métalangage de description de la sémantique de la phrase. Les logiciens indiens ont intégré ces paraphrases à un langage général de description de concepts dans un mouvement intellectuel appelé nouvelle logique (*navya-nyāya*). C'est ainsi qu'une phrase telle que « le chat mange la souris » sera traduite en une formule telle que : « il existe un nombre non différent de un de chats mâles opérant comme agent d'une action de manger au présent de l'indicatif en voix active et dont l'objet est un nombre non différent de un de souris femelles ». La portée de cette méthodologie est importante : si la langue-objet des signifiants est bien une langue particulière, le sanskrit classique, la langue des signifiés est, elle, un sanskrit formalisé qui peut être considéré dans une certaine mesure comme une langue pivot universelle pour en décrire le sens [Bharati\_1995].

**Un corpus gigantesque**

Maintenant que nous avons fait ce tour rapide de la grammaire, il nous faut reconsidérer notre but initial : comprendre le sanskrit. Et, si possible, le comprendre suffisamment pour pouvoir le traiter informatiquement et en extraire le sens à l’aide de logiciels.

En effet, la littérature sanskrite est gigantesque. D'abord le Veda, préservé oralement d’une manière parfaite pendant plus de 3 millénaires à l'aide de procédés mnémotechniques redondants et ainsi correcteurs d'erreurs. Mais le traitement du védique est délicat, et sort de notre propos. Pour le sanskrit classique, par contre, nous avons un corpus relativement homogène énorme s'étendant sur 25 siècles.

Les manuscrits se comptent par millions, les œuvres imprimées en milliers, dans tous les domaines de la connaissance. On imagine à tort que le sanskrit s'occupe exclusivement de vérités profondes, de métaphysique, d'injonctions rituelles et de récits mythologiques. Mais la philosophie indienne, où le pouvoir de la parole tient un rôle prépondérant, est indissociable de la linguistique, à laquelle elle a donné naissance. Nous avons évoqué Pāṇini, Kātyāyana et Patañjali, mais c'est par milliers que se content les ouvrages traitant de l'analyse de la langue (*śābdabodha*). Les grammairiens y disputent avec les logiciens, les ritualistes y développent la pragmatique et la cohérence du discours, 20 siècles avant la théorie des *actes de langage* d’Austin, Searle et Grice.

Des traités astronomiques développent la trigonométrie sphérique, l'analyse, l'algèbre. Des traités de phonétique développent les suites arithmétiques et la combinatoire. Des considérations architecturales rituelles induisent la résolution d'équations diophantiennes. Un poème écrit sur un échiquier syllabe par syllabe démontre l’existence d’un chemin Hamiltonien dans le graphe du parcours d’un cavalier, 800 ans avant Euler. Les mathématiques sont partout, et très souvent en avance sur l'histoire officielle. D'énormes compilations de savoir médical traitent des pathologies et de leur traitement chez l'homme, le cheval, l'éléphant. Le théâtre a été théorisé, dans une théorie esthétique du goût (*rasa*) qui a mené à une tradition élaborée de critique littéraire, établissant des canons poétiques exploitant toutes les subtilités du sanskrit.

Dans tous ces domaines, le sanskrit est plus qu'une langue véhiculaire. Il modèle l'architecture même des raisonnements, et sert de support logique, épistémique et rhétorique à l'ensemble du savoir académique (*śāstra*). Ce savoir est organisé comme raffinement progressif à partir de l'enseignement de maîtres antiques. C'est donc un réseau de discussions, de thèses et d'antithèses débattues au fil de commentaires, de gloses, de réfutations. Il en résulte un corpus très structuré, dense d'intertextualité. Son analyse est un défi pour l'informatique appliquée aux humanités numériques.

**L'analyse du sanskrit, problème inverse ?**

Une langue formalisée par une grammaire, un corpus homogène fascinant. Le sanskrit est donc un terrain de choix pour la linguistique computationnelle, ou traitement informatique de la langue. Néanmoins, une approche trop naïve est vouée à l'échec. Initialement, les pandits indiens, spécialistes de grammaire, ne comprenaient pas qu'il y eût un problème. Si Pāṇini a expliqué de manière effectivement calculatoire comment engendrer exactement du sanskrit correct, il s'ensuit qu'il n'y a qu'à le suivre à l'envers pour analyser les énonciations de sanskrit de sa littérature. Certes, la méthode est théoriquement possible, sauf que le texte écrit qui a été transmis par les manuscrits n'est pas le signe engendré par le fonctionnement de la grammaire, mais au mieux sa partie signifiante. Ce n'est pas même ce signifiant modulé par l'accent, qui est systématiquement déterminé par la grammaire en fonction des opérations, et sert à les désambiguïser, mais a été omis en écriture, sauf pour les écrits védiques. Il ne reste donc comme matériau à traiter que la chaîne des phonèmes obtenue à la fin du processus, et qui peut être extraite exactement de l'écriture syllabique.

Rappelons que le processus grammatical se termine par le lissage phonétique appelé *sandhi*. Le texte écrit est un pur signal phonétique, juste discrétisé en phonèmes. L'avantage, c'est qu'il se prononce comme il s'écrit, il n'y a pas de mystère comme en français où « les poules du couvent couvent ». On sait prononcer le sanskrit exactement comme Pāṇini le parlait il y a 25 siècles, puisque les phonèmes du sanskrit sont décrits avec tous leurs traits phonétiques, ce qui suffit à les identifier dans l'alphabet phonétique international. En contrepartie, il faut faire précéder le traitement de la langue par le traitement de la parole, afin de retrouver les mots du discours dans le flux de l'énonciation continue.

Là aussi, on peut avoir l'illusion qu'il suffit d'inverser les règles de *sandhi* données par Pāṇini pour segmenter le texte en la liste de ses mots. Mais en l'absence d'aucune indication d'où peuvent se trouver les césures, on se trouve face à une combinatoire colossale. Autant le *sandhi* est relativement déterministe, autant son inversion est hautement non-déterministe, menant à une exploration combinatoire exponentielle. Un problème analogue est d’inverser toutes les multiplications effectuées pour produire un nombre entier, afin d'en tirer ses facteurs premiers. Autant la multiplication est une opération triviale pour un ordinateur, autant la décomposition d'un entier en facteurs premiers est difficile, au point de fournir la base des procédés cryptographiques contemporains.

Toutefois, le problème est soluble si l'on dispose d'une base de données des formes fléchies des mots, car on peut les chercher progressivement dans la chaîne d'entrée, en défaisant le *sandhi* au vol aux endroits pertinents. En effet, les règles de *sandhi* induisent une relation rationnelle, c'est à dire représentable par un automate fini. Le décryptage du *sandhi* peut alors s'exprimer comme composition de relations rationnelles sur les chaînes de phonèmes, et son inverse définit une transduction rationnelle opérant sur l'énonciation continue, et produisant le *padapāṭha* ou suite des mots de la phrase. Cette décomposition n'est pas unique, elle est encore hautement ambiguë, mais on peut espérer la focaliser en analysant la cohérence sémantique des segments, par résolution des contraintes de dépendance. En effet, les formes fléchies peuvent être systématiquement construites en utilisant la grammaire, et donc être munies des caractères morphologiques (comme le cas des déclinaisons) qui peuvent être reliés aux rôles sémantiques par la théorie des *kārakas*. Cette technologie a été mise en œuvre dans le logiciel Zen, et a permis le développement du « Reader assistant » et du « Corpus Manager » associés au site Web « Héritage du Sanskrit » [Huet\_2006, Goyal\_2013, Goyal\_2016, Huet\_2018].

**Un exemple commenté**

Examinons le traitement d'un exemple par ces outils.

Nous consultons le corpus numérisé, sur le chemin Heritage\_citations. En cliquant sur la citation 95, on obtient l'analyse représentée graphiquement sur la figure Fig. 1.



Fig. 1

Chaque segment porte un lien hypertexte local, menant à son analyse morphologique sous forme d'un lemme se référant au lexique. Par exemple, en cliquant sur le segment rouge *śobhate* on obtient le lemme : [ śubh1 ]{ pr. [1] md. sg. 3 } où la racine *śubh1* réfère au lexique, dans le sens de ‘briller’. Là, on peut vérifier que l'indicateur de classe (*gaṇa*) [1] pointe sur la table des conjugaisons de ce verbe, où on trouve bien la forme *śobhate* au présent en voix moyenne à la 3ème personne du singulier. L'examen des autres segments nous suggère le sens « L'imbécile ne brille que tant qu'il se tait ».

Mais un mystère demeure : comment le tronçon *tāvacchobhate* a-t-il été segmenté en *tāvat* et *śobhate* ? Examinons plus en détail ce processus en mode Tagging. La segmentation de *tāvacchobhate* produit en première solution ces deux segments, avec l'indication <t | ś ⇒ cch> qui donne la règle de *sandhi* à appliquer.

On peut vérifier que cette règle (en rouge sur la figure Fig. 2 extraite du programme source Ocaml de compilation du *sandhi*), est justifiée par une suite de *sūtras* de la grammaire de Pāṇini. Elle postule que la finale *t* se lie avec l'initiale *ś* (notée *z* dans le code) pour produire *cch*.



Fig. 2 Source du transducteur, renseigné par les *sūtras*

On remarquera à sa suite la règle optionnelle permettant le remplacement par *cś*. Elle permet par exemple de remplacer l'entrée par *tāvacśobhate*, et d'obtenir la même analyse. Ceci montre que la grammaire de Pāṇini, bien que normative, tolère des différences de prononciation régionales.

Le segmenteur est piloté par un transducteur, dont le graphe de transitions est construit à partir du lexique des formes fléchies, décoré par des transitions non-déterministes aux points de choix où le *sandhi* peut s'appliquer. Cette technologie, et sa complétude, sont expliquées dans [Huet\_2005].

**Eilenberg entre en scène**

Samuel Eilenberg (1913-1998) était un mathématicien américain d'origine polonaise, qui fonda la théorie des catégories avec Saunders Mac Lane. Il était membre du groupe Bourbaki, et avec Henri Cartan écrivit en 1956 l'ouvrage ‘Homological Algebra’, devenu un classique.

Dans les années 1970 il collabora avec Marcel-Paul Schützenberger dans le domaine de la théorie des langages et automates. On lui doit le traité magistral ‘Automata, Languages and Machines’, dont le Volume A parut en 1974 and le Volume B en 1976. Il décéda avant de compléter les volumes C and D, et les Volumes A and B ne furent jamais réimprimés, ce qui fait que ce travail est largement ignoré. Dans le volume A, il donna une définition de X-machine, dont l'étude était prévue dans les derniers volumes [Eilenberg\_1974]. Cette notion importante est maintenant appelée en son honneur machine d'Eilenberg. Ces machines sont des modèles mathématiques généraux de programmation relationnelle. Elles généralisent la notion d'automate fini non-déterministe, qui est utilisée pour rendre compte de leur contrôle, par un graphe fini de transitions, étiquetées avec l'alphabet des actions de la machine. Mais elles possèdent également une composante de données, où les actions sont maintenant interprétées comme des relations binaires opérant sur un ensemble paramétrique de valeurs X.



Samuel Eilenberg , Marcel-Paul Schützenberger et Seymour Ginsburg, ICALP 1972

Benoît Razet, dans son travail de thèse [Razet\_2009], sut adapter cette notion mathématique à un paradigme informatique de calcul relationnel, en remplaçant l'ensemble abstrait X par un domaine de calcul, les relations étant maintenant des relations calculables sur le domaine. On obtient ainsi une notion de machine d'Eilenberg effective, élégante combinaison d'un contrôle fini et d'un modèle de calcul relationnel complet au sens de Turing. Ce modèle de calcul est approprié à représenter l'exploration non-déterministe de l'espace des solutions de problèmes combinatoires, tels que notre problème de segmentation. De fait, le transducteur rationnel sous-tendant l'algorithme de segmentation peut être modélisé directement comme machine d'Eilenberg finie, l'espace des solutions étant fini [Huet-Razet\_2006, Huet-Razet\_2015].

Il est plaisant qu'un problème applicatif épineux, l'inversion des règles de *sandhi* du linguiste antique Pāṇini, trouve sa résolution expliquée dans un cadre mathématique élégant. Lorsque le modèle fut trouvé, le logiciel informatique put être recomposé en termes de machines d'Eilenberg abstraites grâce au mécanisme de modules paramétriques du langage fonctionnel Ocaml. Une dernière touche fut ajoutée par Pawan Goyal, qui exprima la décoration de l’automate de segmentation par les règles de *sandhi* en termes des conditions données par Pāṇini dans les *sūtras* pertinents, comme montré sur la figure Fig. 2.

**Extraction du sens**

Considérons un autre exemple, la citation 158. Voir la figure Fig 3.



Fig. 3 Phrase 2

La phrase est un dicton qui signifie « Un corbeau ne devient pas colombe en buvant du lait ». La forme verbale *bhavati* est le présent de la 3ème personne du singulier de la racine *bhū* , devenir. L’agent de l’action est le corbeau *kākaḥ*, qui est également agent de la clause absolutive *kṣīram pītvā* ‘ayant bu du lait’.

Sur cet exemple, il est possible de pousser l’analyse en obtenant la concordance des rôles sémantiques (*kāraka*) par un calcul de contraintes sur un graphe de dépendances. Un tel logiciel d’analyse a été développé par Amba Kulkarni [Kulkarni\_2013], et est accessible en prolongement du segmenteur-étiqueteur de ‘Sanskrit Heritage’. On en voit le résultat à l’analyse de cette phrase sur la figure Fig. 4. On remarque le partage du rôle agent (*kartā*) du corbeau (*kākaḥ*) par la clause principale et sa subordonnée absolutive *kṣīram pītvā*, où l’action de boire (*pītvā*) trouve son objet (*karma*) dans le lait (*kṣīram*), et est indiquée comme effectuée dans un temps précédent (*pūrvakālaḥ*).



Fig. 4 Graphe de dépendances de la phrase 2

**Conclusion**

Au terme de notre quête de la compréhension du sanskrit, nous voyons que les outils mathématico-informatiques, correctement renseignés par les méthodes d’analyse de la langue dues à Pāṇini et ses successeurs, peuvent nous aider effectivement à trouver le sens de ses énoncés. Cette méthodologie fait dialoguer à 25 siècles de distance un grammairien antique et un mathématicien contemporain.

Ces outils sont suffisamment avancés pour commencer à traiter le corpus considérable de cette langue, et à mettre au point des outils philologiques modernes, traitant l’information au niveau d’un modèle de situations exprimant le sens.

**Bibliographie**

[Bharati\_1995] Akshar Bharati, Vineet Chaitanya & Rajeev Sangal. Natural Language Processing. A Paninian Perspective. Prentice-Hall of India, New Delhi, 1995.

[Burrow\_1955] T. Burrow. The Sanskrit Language. Faber and Faber.

[Deo\_2003] Ashwini Deo. Derivational Morphology in Inheritance-based Lexica: Insights from Pāṇini. Lingua [117] (2003).

[Eilenberg\_1974} Automata, Languages, and Machines, volume A. Academic Press, 1974.

[Emeneau\_1988] M. B. Emeneau. Bloomfield and Pāṇini. Philosophy East and West [64,4] (1988) pp. 755-760. Society for Comparative Study of Society and History.

[Goyal\_2013] Pawan Goyal & Gérard Huet. Completeness Analysis of a Sanskrit Reader. Recent Researches in Sanskrit Computational Linguistics, eds. Malhar Kulkarni and Chaitali Dangarikar. D.K. Printworld (2013), pp. 130-171.

[Goyal\_2016] Pawan Goyal & Gérard Huet. Design and analysis of a lean interface for Sanskrit corpus annotation. Journal of Linguistic Modeling [4,2] (2016), pp. 117-126.

[Hanneder\_2002] J. Hanneder. On 'The Death of Sanskrit'. Indo-Iranian Journal [45,4] (2002), pp. 293-310. Brill.

[Hellwig\_2009] Oliver Hellwig. Extracting Dependency trees from Sanskrit texts. Sanskrit Computational Linguistics 3, eds. Amba Kulkarni & Gérard Huet, Springer-Verlag LNAI 5406, pp. 106-115.

[Huet\_2005] Gérard Huet. A Functional Toolkit for Morphological and Phonological Processing, Application to a Sanskrit Tagger. J. Functional Programming [15,4] (2005), pp. 573-614.

[Huet\_2006] Gérard Huet. Shallow syntax analysis in Sanskrit guided by semantic nets constraints. Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries, Kolkata. ACM http://doi.acm.org/10.1145/1364742.1364750.

[Huet-Razet\_2006] Gérard Huet & Benoît Razet. The Reactive Engine for Modular Transducers. Algebra, Meaning and Computation, Essays Dedicated to Joseph A. Goguen on the Occasion of His 65th Birthday, eds. Kokichi Futatsugi, Jean-Pierre Jouannaud & José Meseguer. Springer-Verlag LNCS vol. 4060, pp. 355-374.

[Huet\_2015] Gérard Huet & Benoît Razet. Computing with Relational Machines. Mathematical Structures in Computer Science [10] (2015), pp. 1-20. Cambridge University Press.

[Huet\_2016] Gérard Huet. Sanskrit signs and Pāṇinian scripts. Sanskrit and Computational Linguistics, ed. Amba Kulkarni. D.K. Printworld, New Delhi, 2016.

[Huet\_2018] Gérard Huet & Idir Lankri. Preliminary Design of a Sanskrit Corpus Manager. Computational Sanskrit & the Digital Humanities, eds. Gérard Huet & Amba Kulkarni. D.K. Printworld, New Delhi", 2018.

[Kaplan\_1994] Ronald M. Kaplan and Martin Kay. Regular Models of Phonological Rule Systems. Computational Linguistics [20,3] (1994) pp. 331-378.

[Kiparsky\_2009] Paul Kiparsky. On the Architecture of Pāṇini's Grammar. Sanskrit Computational Linguistics 1 & 2, eds. Gérard Huet, Amba Kulkarni & Peter Scharf, Springer-Verlag LNAI 5402 (2009).

[Kulkarni\_2013] Amba Kulkarn. A Deterministic Dependency Parser with Dynamic Programming for Sanskrit. Proceedings of the Second International Conference on Dependency Linguistics, pp. 157-166. Charles University Matfyzpress, Prague (2013).

[Levi\_1895] Sylvain Lévi. Génie de l'Inde. Leçon inaugurale de la chaire de langue et littérature sanscrites délivrée au Collège de France le 16 janvier 1895. Reprint Allia, Paris, 2008.

[Minkowsky\_2006] Christopher Minkowski. King David in Oudh: a Bible story in Sanskrit and the Just King at an Afghan court. Inaugural Lecture for the Boden Professorship, University of Oxford (2006).

[Mohri\_1997] Mehryar Mohri. Finite-State Transducers in Language and Speech Processing. Computational Linguistics [23,2] (1997), pp. 269-311.

[Petersen\_2004] Wibke Petersen. A Mathematical Analysis of Pāṇini's Śivasūtras. Journal of Logic, Language and Information [13,4] (2004), pp. 471-489.

[Pollock\_2001] Sheldon Pollock. The Death of Sanskrit. Comparative Studies in Society and History [43,2] (2001), Society for Comparative Study of Society and History, pp. 392-426.

[Post\_1936] Emil L. Post. Finite Combinatory Processes - Formulation 1. Journal of Symbolic Logic [1] (1936), pp. 103-105.

[Razet\_2009] Benoît Razet. Machines d'Eilenberg Effectives. Thèse de doctorat d'informatique, Université Denis Diderot, Paris (2009).

[Renou\_1966] Louis Renou. La Grammaire de Pāṇini. Ecole Française d'Extrême-Orient, Paris, 1966.

[Sampad\_2011] Sampad & Vijaya. The Wonder that is Sanskrit. Sri Aurobindo Society, Puducherry.

[Saussure\_1916] Ferdinand de Saussure. Cours de linguistique générale. Payot (1916).

[Sharf\_2009] Peter Scharf. Modeling Pāṇinian Grammar. Sanskrit Computational Linguistics 1 & 2, eds. Gérard Huet, Amba Kulkarni & Peter Scharf. Springer-Verlag LNAI 5402, pp. 95-126.

[Sharma] Rama Nath Sharma}. The Aṣṭādhyāyī of Pāṇini. Munshiram Manoharlal Publishers, 6 vols, 1987-2003.

[Staal\_1965] Euclid and Pāṇini. Language [15] (1965), pp. 99-116.

[Staal\_1983] Frits Staal. Agni - The Vedic Ritual of the Fire Altar. Asian Humanities Press, Berkeley, 1983.

[Tesnière\_1959] Lucien Tesnière. Éléments de Syntaxe Structurale. Klincksieck, Paris, 1959.