

Sciences, langages et langues

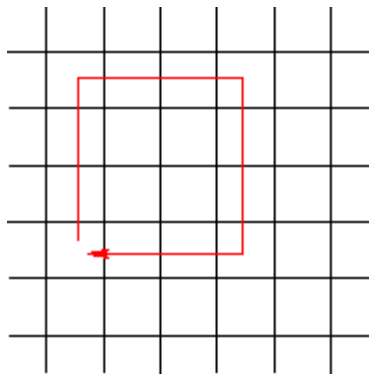
Gilles Dowek

C'est que l'écriture, Phèdre, a un grave inconvénient, tout comme la peinture. Les produits de la peinture sont comme s'ils étaient vivants ; mais pose leur une question, ils gardent gravement le silence. Il en est de même des discours écrits. On pourrait croire qu'ils parlent en personne intelligente, mais demande leur de t'expliquer ce qu'ils disent, ils ne te répondront qu'une chose, toujours la même.

Platon.

I. Les langages informatiques

Une manière de commander un robot qui se déplace sur une grille carrée est de lui donner une suite d'instructions, qui commandent chacune un mouvement dans une certaine direction. Ainsi la suite « nord nord nord est est est sud sud sud ouest ouest ouest » commande au robot de se déplacer de trois unités vers le nord, puis de trois unités vers l'est, de trois unités vers le sud, et de trois unités vers l'ouest enfin, et donc de parcourir les quatre cotés d'un carré.



Les expressions ainsi formées d'une succession de mots « nord », « sud », « est » et « ouest », forment un langage.

Mais, ce même mouvement peut aussi se décrire dans un autre langage, dans lequel une instruction commande au robot de se déplacer droit devant lui, alors que deux autres lui commandent de tourner sur lui-même d'un quart de tour vers la droite ou vers la gauche. Dans cet autre langage, ce même mouvement s'exprime de manière tout à fait différente : « avancer avancer avancer tourner-à-droite avancer avancer avancer tourner-à-droite avancer avancer avancer tourner-à-droite avancer

avancer avancer tourner-à-droite ». Ces deux langages permettent d'exprimer les mêmes mouvements et il n'est pas difficile de transformer une commande exprimée dans l'un d'eux, en une commande exprimée dans l'autre.

Avec ces exemples, nous venons d'ajouter deux langages aux milliers de langages que les informaticiens ont inventés pour commander des robots, programmer des ordinateurs, décrire des textes et des images, interroger des bases de données, etc. Et dès que nous voyons un tel langage, il nous vient des idées pour l'améliorer : par exemple ajouter une instruction qui permettrait de commander la répétition d'une autre instruction, ce qui permettrait d'écrire : « trois fois nord », plutôt que « nord nord nord ».

Ces langages, et plus généralement, tous les langages informatiques, sont soumis à une double contrainte : ils doivent être utilisables par les êtres humains qui commandent les robots, mais aussi par les robots qui exécutent ces commandes. Cette seconde contrainte explique, en partie, la différence entre ces langages et les langues naturelles que nous employons tous les jours.

Aux débuts de l'informatique, cette contrainte était prépondérante et seuls existaient des *langages machines*, relativement faciles à utiliser pour les machines, mais pas pour les êtres humains. L'histoire des langages de programmation est, en grande partie, celle de l'effort de concevoir des langages plus faciles à utiliser par des êtres humains, tout en gardant la possibilité de les traduire vers des langages utilisables par des machines. Autrement dit, une tendance dans l'évolution des langages de programmation est de se rapprocher des langues. Toutefois, nous serons amenés à relativiser cette idée : les langues ne sont pas nécessairement les meilleurs outils pour exprimer des algorithmes par exemple et la nécessité de rendre les programmes exécutables par des machines n'est donc peut-être pas l'unique raison pour laquelle les langages de programmation se sont éloignés des langues.

Avec la notion de langage s'introduit naturellement celle de traduction d'un langage à un autre. Mais, avant cette notion de traduction, s'introduit une notion de correction : l'expression « nord nord nord tourner-à-droite », par exemple, n'est correcte dans aucun des deux langages que nous avons définis, car elle mélange des mots d'un langage et de l'autre. Chaque langage est donc défini par une grammaire, qui est un algorithme qui indique si une phrase est correcte ou non dans ce langage.

La grammaire d'un langage peut elle-même s'exprimer dans un langage, que l'on doit à John Backus et qui est un cas particulier d'un langage plus général dû à Noam Chomsky. Par exemple, la grammaire du premier des langages que nous avons introduit s'exprime ainsi

$$I = \text{nord} \mid \text{sud} \mid \text{est} \mid \text{ouest}$$
$$E = \varepsilon \mid I E$$

Ce langage d'expression des grammaires contient des symboles pour des langages, ici I et E , les symboles = (égal), | (ou) et ε , qui désigne l'expression vide, et des symboles pour les symboles du langage décrit, ici « nord », « sud », « est » et « ouest ».

Cette définition se lit en deux temps. La première ligne définit un langage I qui ne contient que les quatre expressions « nord », « sud », « est » et « ouest ». La seconde définit un langage E qui contient d'une part l'expression vide et d'autres par toutes les expressions formées d'une expression du langage I suivie d'une autre expression du langage E , ce qui revient à dire que les expressions du

langage E sont des suites formées de zéro, un ou plusieurs mots « nord », « sud », « est » et « ouest ».

Cette grammaire définit uniquement si une expression est correcte ou non. Ce que le robot doit faire quand on lui donne une commande exprimée dans ce langage est défini, non par la grammaire de ce langage, mais par sa sémantique.

Il n'est pas difficile de définir, de même, la grammaire du second langage que nous avons introduit ou la grammaire du langage des grammaires.

II. Qu'est-ce qu'un langage ?

Les deux langages que nous avons présentés sont assez différents des langues, telles le français ou le chinois.

La spécialisation. Tout d'abord, ces langages permettent d'exprimer une commande donnée à un robot qui se déplace sur une grille carrée, mais pas un contrat ou un roman. De même, le langage des grammaires permet d'exprimer uniquement des grammaires, les langages de programmation permettent d'exprimer uniquement des algorithmes, etc. Ces langages sont donc spécialisés.

Une langue, en revanche, peut être utilisée non seulement pour exprimer une commande donnée à un robot, une grammaire, un algorithme, etc., même si c'est souvent de manière imprécise, mais aussi un contrat ou un roman. Elles sont donc universelles.

Le vocabulaire. Le vocabulaire des deux langages ci-avant est très limité : « nord », « sud », etc. pour le premier, « avancer », « tourner-à-droite », etc. pour le second. De même, le langage des grammaires ne contient que quelques « mots » : « = », « | », « ε », etc. et le vocabulaire d'un langage de programmation se limite souvent à quelques dizaines de mots. Le vocabulaire d'une langue naturelle, en revanche, en contient plusieurs dizaines de milliers.

En revanche, le vocabulaire d'un langage est souvent extensible. Ce n'est pas le cas des langages simples que nous avons présentés ci-avant, mais les langages de programmation, par exemple, comportent souvent un mécanisme de définitions qui permet d'associer un nom, par exemple « parcourir-un-carré », à une expression complexe, par exemple « nord nord nord est est sud sud sud ouest ouest ouest ». En utilisant un tel mécanisme, on ajoute au langage un nouveau mot « parcourir-un-carré ». Dans une langue, en revanche, il n'est pas possible de dire « On appellera "épagneul breton" un steak saignant. », puis « Je voudrais manger un épagneul breton. ».

Un mot ainsi introduit a souvent une portée limitée, c'est-à-dire qu'il peut être utilisé dans une certaine partie du texte mais non dans d'autres. Cette notion de portée, corollaire de cette possibilité d'étendre le langage par de nouveaux mots, semble elle aussi absente des langues.

La grammaire. La grammaire des deux langages que nous avons introduits est plus simple, que celle d'une langue. La grammaire du premier de ces langages, par exemple, se décrit en deux lignes

$I = \text{nord} \mid \text{sud} \mid \text{est} \mid \text{ouest}$

$E = \varepsilon \mid I E$

voire en une seule

$E = \varepsilon \mid \text{nord } E \mid \text{sud } E \mid \text{est } E \mid \text{ouest } E$

et la grammaire d'un langage de programmation se définit ainsi en quelques pages. Nous sommes loin des centaines de pages d'une grammaire du français, avec sa classification des verbes en plusieurs groupes, ses règles de concordance des temps, d'accord du participe passé, etc.

En revanche, la grammaire d'un langage est plus précise : un signe de ponctuation oublié dans un programme et celui-ci n'est plus correct, alors qu'il est difficile de trouver une phrase en français qui serait incorrecte, à cause d'un signe de ponctuation oublié.

De plus, la frontière entre les phrases correctes et les phrases incorrectes dans une langue n'est pas si bien définie. Par exemple, la phrase « Nous, on y est pas allé, mon frère et moi, au cinéma. » manifeste une moins grande maîtrise de la langue que la phrase « Ni mon frère ni moi ne sommes allés au cinéma. », mais est-elle incorrecte pour autant ? En revanche, il n'y a aucune discussion possible sur le fait qu'une expression soit grammaticalement correcte ou non dans un langage, car cette correction grammaticale est définie par un algorithme.

III. Une histoire de la notion de langage

L'apparition de la notion de langage dans notre culture, et sa distinction de celle de langue, semble donc un effet du développement de l'informatique et, plus précisément, de cette nécessité de trouver des moyens d'expression à la fois utilisables par des êtres humains et par des machines. Toutefois, comme toujours quand on identifie un nouveau concept, il n'est pas inutile de regarder vers le passé, et de se demander si ce concept n'existait pas déjà, de manière plus ou moins silencieuse. Et il semble qu'il faille remonter assez loin, pour retrouver les premières traces de la notion de langage : aux premières écritures des nombres, c'est-à-dire aux origines mêmes de l'écriture.

Les nombres. Quand nous insérons un nombre écrit en chiffres dans un texte écrit en français, par exemple dans la phrase « Il y a 237 habitants dans ce village. », nous sentons une rupture dans la langue utilisée. Il n'est par exemple pas possible d'insérer le mot « village », ou même le mot « et », entre le « 3 » et le « 7 ». L'expression « 237 » semble être une expression d'une autre langue insérée dans un texte en français, l'alphabet même dans lequel elle est exprimée est différent. Toutefois, ce n'est pas à une autre langue, mais à un langage, que l'expression « 237 » appartient. Nous retrouvons en effet, dans le langage des nombres, les trois caractéristiques que nous avons évoquées ci-avant : la spécialisation, le vocabulaire restreint et la grammaire simple. La seule nouveauté dans la grammaire des nombres

$C = 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$

$$E = \varepsilon \mid 0 E \mid C E$$

$$N = 0 \mid C E$$

par rapport à la grammaire du langage de commande d'un robot est qu'elle interdit que l'écriture d'un nombre commence par le chiffre « 0 », sauf si cette écriture est elle-même « 0 ».

La notation musicale. Comme l'écriture des nombres, l'écriture de la musique a demandé d'inventer un langage particulier. On aurait pu écrire par exemple « jouer un si, puis jouer un autre si, puis jouer un sol précédé d'un la bref, jouer ensuite un ré, deux fois plus long que le sol, puis un mi aussi long que le si, puis un ré, aussi long que le mi, une alternance rapide de ré et de do, suivie d'un si et d'un do brefs, pendant un temps deux fois plus long que le ré et une alternance rapide de do et de si pendant un temps trois fois plus long que le ré, tout cela gracieusement ». Mais il semble que depuis le XIII^e siècle, les instrumentistes préfèrent lire cela sous la forme



À nouveau ce langage est spécialisé : il ne permet d'écrire que la musique, son vocabulaire est restreint : do, noire, *forte*, trille, etc. quoi qu'il le soit un peu moins que le vocabulaire des nombres et sa grammaire est relativement simple, quoi qu'elle le soit un peu moins que la grammaire des nombres, mais précise : elle autorise de mettre six croches dans une mesure binaire à trois temps, mais non quatre noires.

La notation algébrique. La notation algébrique, qui date du XVI^e siècle, permet d'écrire

$$x^3 + 3x^2 = 20$$

l'équation qui se disait jadis : « un cube et trois carrés font vingt ». À nouveau, on retrouve là spécialisation, vocabulaire restreint et grammaire simple.

Les langages de la logique. Cette notation algébrique est l'une des inspirations de l'idéographie de Gottlob Frege qui est à l'origine du renouveau de la logique à la fin du XIX^e siècle et qui est à la fois l'ancêtre de la théorie des ensembles et de la logique des prédicats. Le langage de la logique des prédicats, qui permet par exemple d'écrire

$$\forall x \exists y (x = 2 \times y \vee x = 2 \times y + 1)$$

ce que l'on dit dans en français : « pour tout x , il existe un y tel que x soit égal à deux fois y ou x soit égal à deux fois y plus un », a aussi un vocabulaire restreint et une grammaire simple. Ce langage est également spécialisé : il ne permet d'exprimer que des propositions déclaratives, c'est-à-dire susceptibles d'être vraies ou fausses, et non, par exemple, un ordre, une question ou une injure. Mais il a l'ambition de permettre l'expression de toutes les propositions déclaratives et est donc beaucoup moins spécialisé que le langage des nombres ou celui des commandes d'un robot.

Le langage des expressions algébriques et celui de la logique se distinguent également des langues en ce qu'ils comprennent une notion de variable « x », qui ne semble pas avoir d'équivalent dans les langues.

La nomenclature chimique. La nomenclature des composés chimiques, qui date également du XIX^e siècle, est elle aussi un langage spécialisé, au vocabulaire restreint et à la grammaire simple mais précise. Il y a par exemple un chlorure d'aminométhylpyrimidinylhydroxyéthylméthythiazolium, mais pas de chlorure d'épagneul breton : seuls un nombre restreint de mots peuvent être employés dans ce langage. Outre la nomenclature, la chimie est une grande créatrice de langages : langage graphique des formules développées des molécules, langage des équations chimiques pour décrire les réactions, etc.

Des situations intermédiaires. Sans cesser d'être une langue, la langue scientifique a quelques traits qui rappellent les langages. Par exemple, si, comme nous l'avons vu, il n'est pas possible dans la langue courante de dire « On appellera "épagneul breton" un steak saignant. », puis « Je voudrais manger un épagneul breton. », on peut, dans la langue scientifique, dire « On appellera "travail" d'une force, son produit scalaire avec le déplacement de son point d'application. » puis utiliser le mot défini : « Le travail d'une force orthogonale au déplacement de son point d'application est nul. » Une fois défini, le mot « travail » a une portée illimitée, mais ce n'est pas le cas du mot « x » si on l'introduit dans une démonstration mathématique, par la définition « $x = y + 4$ ». Une fois cette démonstration achevée, il n'est plus possible d'utiliser le mot « x » pour désigner la valeur « $y + 4$ ».

De même, dans le langage juridique, on débute le texte d'un contrat en donnant l'identité des contractants et en indiquant un mot par lequel ils seront désignés dans la suite du contrat, par exemple « Contrat de location entre les sous-signés, Monsieur Dupond, demeurant 1, rue Durand, à Grenoble, ci-après désigné le loueur et Monsieur Durand, demeurant 1, rue Dupond, à Grenoble, ci-après désigné le locataire ... ». Bien entendu cette manière de désigner Monsieur Dupond et Monsieur Durand comme « le loueur » et « le locataire » a une portée limitée au texte de ce contrat.

La rançon de l'universalité des langues semble donc être leur incapacité à exprimer précisément les choses, dès que l'on s'intéresse à un domaine précis : les sciences, le droit, la musique, etc. Et cela explique que les scientifiques, les juristes, les musiciens, etc. se soient éloignés des langues, parfois par de petits écarts, parfois en créant de véritables langages. Cela nous mène à relativiser quelque peu l'idée que nous avons suggérée ci-avant que l'utilisation de langages de programmation pour exprimer des algorithmes était motivée par la double contrainte d'écrire des textes compréhensibles à la fois par un être humain et par une machine. La manière d'exprimer les équations en algèbre, par exemple, s'est éloignée des langues, bien avant que l'on ait des machines, simplement parce que le langage algébrique est plus facile à utiliser qu'une langue pour exprimer des équations et les résoudre. De même, certains langages sont peut-être plus faciles à utiliser qu'une langue pour exprimer des algorithmes, même si ces algorithmes ne sont pas destinés à être exécutés par une machine.

Même si l'informatique a complètement renouvelé la notion de langage, et même si la musique et le droit ne sont pas des sciences, il semble y avoir un lien essentiel entre la notion de langage et les sciences : les sciences construisent sans cesse de nouveaux moyens pour exprimer des choses que les langues ne savent pas exprimer précisément. Cela semble être un trait caractéristique de la pensée scientifique.

IV. Langages et écriture

Comme nous ne pouvons pas donner un nom différent à l'infinité des nombres : un, deux, trois, ... nous avons décidé de nous arrêter à neuf et de décomposer les nombres plus grands, par exemple deux cent trente sept, en sept unités, trois dizaines et deux dizaines de dizaines. Cela peut mener à plusieurs manières d'écrire les nombres : CCXXXIIIIIIII, 2C 3X 7I ou 237, ainsi qu'à d'autres plus complexes, comme CCXXXVII. Le premier de ces langages ne demande des symboles que pour les puissances de dix : unités, dizaines, centaines, milliers, etc. : I, X, C, M, etc. Le dernier, la numération à position, ne demande que des symboles que pour les nombres compris entre zéro et neuf : 0, 1, 2, ..., 9. Le deuxième, enfin, à la fois des symboles pour les puissances de dix et pour les nombres compris entre 0 et 9.

Avec un nombre fini de symboles, le premier de ces langages ne permet jamais d'exprimer tous les nombres : avec les symboles I, X, C et M, par exemple, il ne permet d'exprimer que les nombres jusqu'à neuf mille neuf cent quatre vingt dix neuf : MMMMMMMMCCCCCCCCXXXXXXXXXIIIIIIII. Il en est de même pour le deuxième : 9M 9C 9X 9I. Seule la numération à position, qui ne demande pas de symboles pour les puissances de dix, permet d'exprimer tous les nombres, en particulier le nombre dix mille : 10000. Cela explique que nous ayons finalement abandonné les deux premiers langages, les chiffres romains, etc., au profit de la numération à position.

Pourtant cette victoire de la numération à position n'est que partielle : par oral, nous continuons à employer le deuxième langage : « deux cent trente sept », où le fait que le chiffre deux soit un nombre de centaines est indiqué par le mot « cent » qui suit ce chiffre et non par sa position dans l'écriture du nombre comme dans 237. Le fait que nous utilisions des langages différents par oral et par écrit est à l'origine de bien des difficultés dans l'apprentissage de l'écriture des nombres : il faut comprendre que « deux – cent – trente – sept » s'écrit « 237 » et non « 2 – 100 – 30 – 7 ». Nous pouvons alors nous demander pourquoi nous continuons à utiliser un langage archaïque par oral, et non le même langage que par écrit, pourquoi nous disons « deux cent trente sept » et non « deux trois sept ». La raison est que la numération à position manque de redondance et est de ce fait trop sensible aux perturbations, pour être utilisée par oral. Par exemple, une seule syllabe sépare « deux trois sept » et « deux trois sept un » qui expriment pourtant des nombres très différents. Alors que « deux cent trente sept » et « deux mille trois cent soixante et onze » indiquent, dès le début, l'ordre de grandeur des nombres : « deux cents » ou « deux mille ».

La numération à position est donc un langage qui ne peut être utilisé que par écrit, ce qui montre le rôle joué par l'écriture dans l'invention de ce langage. Et la numération à position n'est pas un cas isolé. Lire à haute voix une expression comme « nord nord nord est est est sud sud sud ouest ouest », un programme informatique, une grammaire, une partition, le mot « aminométhylpyrimidinylhydroxyéthylméthythiazolium », une formule chimique, etc. est souvent inutile et parfois impossible : certains langages comme celui des plans d'architectes, des diagrammes de Feynman, les langages de programmation graphiques, etc. sont bidimensionnels, ce qui rend impossible leur expression à haute voix.

Nous avons appris à parler avant d'apprendre à écrire. Et les premiers mots que nous avons appris à

écrire étaient des mots que nous savions dire et lire à haute voix. Cela explique peut-être que, comme Platon, nous ayons conçu l'idée que l'écriture sert à retranscrire la parole. Mais, de la numération à position aux langages de programmation, les langages nous donnent de nombreux exemples de textes qui peuvent s'écrire, mais qui ne peuvent se dire à haute voix : contrairement à une pièce de théâtre, la forme développée d'une molécule n'est pas écrite dans le but d'être dite.

L'écriture ne nous a donc pas seulement permis de transmettre plus sûrement, ou de conserver plus longtemps, des idées que nous aurions pu exprimer oralement, elle nous a aussi ouvert d'immenses possibilités de nous exprimer autrement, non dans des langues, mais dans des langages. Cela nous mène à nous interroger sur le lien entre les langages, les langues et l'écriture. L'écriture a-t-elle été inventée pour écrire les langues ou les langages ?

Pour répondre à cette question, nous pouvons nous pencher sur les plus anciennes traces écrites qui nous sont parvenues. Ces traces sont des tablettes sumériennes, qui ont un peu plus de cinq mille ans. Elles semblent provenir d'une transformation d'un système de communication antérieur utilisé par les Sumériens : les bulles-enveloppes. Pour enregistrer leurs opérations comptables, les Sumériens utilisaient en effet un système de jetons d'argile, de tailles et de formes différentes, qui étaient glissés dans une bulle, sphère en argile creuse et scellée. Si une bulle contenait le dénombrement d'un troupeau confié à un berger, par exemple, lorsque celui-ci ramenait le troupeau, il suffisait de casser la bulle pour vérifier qu'aucune bête ne manquait. Plus tard, on a eu l'idée d'apposer sur la bulle elle-même un résumé de son contenu, sous la forme d'encoches dont la forme évoquait celle des jetons. Les jetons sont alors devenus inutiles, les sphères se sont aplaties et transformées en tablettes. Ce n'est que dans une troisième étape, que ce système s'est complexifié de manière à permettre de transcrire la langue sumérienne.

Ce système de jetons et de bulles a toutes les caractéristiques d'un langage : il est très spécialisé puisqu'il ne permet d'exprimer essentiellement que des nombres, il utilise un vocabulaire restreint à quelques types de jetons, il a une grammaire simple à l'extrême puisque n'importe quel ensemble de jetons forme un texte grammaticalement correct, les « mots » ne sont pas encore disposés dans un certain ordre, et il est exclusivement « écrit » puisqu'un texte n'existe qu'incarné dans l'argile.

Cette comparaison des langues et des langages semble donc suggérer une certaine autonomie de l'écriture par rapport à la parole. L'écriture semble avoir été inventée pour transcrire, non une langue, mais un langage, et appartenir davantage à l'univers des langages qu'à celui des langues, même si elle a finalement été, en quelque sorte, détournée de sa vocation première, afin de transcrire aussi les langues.