Continu et discret en sémantique lexicale

Bernard Victorri

Cahiers de praxématique, 42, 2004 – Version préliminaire

Continu et discret en sémantique lexicale

Bernard Victorri

Introduction

Quel type d'outils mathématiques doit-on utiliser pour modéliser le sens d'une unité linguistique dans un énoncé ? Faut-il choisir des représentations discrètes, de nature algébrique, permettant par exemple de représenter le sens sous forme d'ensembles de traits sémantiques ou de graphes de relations sémantiques ? Ou bien doit-on préférer au contraire des représentations continues, de nature géométrique, permettant de représenter le sens sous forme de régions ou de fonctions dans des espaces sémantiques ?

Avant de chercher à répondre à cette question, il est important de bien préciser que lorsque nous parlons de discret et de continu, c'est pour caractériser les modèles du sens et non le sens lui-même, quelle que soit la manière dont on le définisse. En effet, nous nous en tiendrons ici à la notion mathématique de discret et de continu, qui, bien sûr, s'applique aux modèles et non aux phénomènes eux-mêmes. Il ne s'agit pas là d'une remarque de pure forme. En fait, on peut même dire que les deux démarches n'ont rien à voir l'une avec l'autre. D'abord parce que l'on peut très bien concevoir le sens comme continu, et le modéliser de façon discrète, ou vice-versa (cf. le débat entre Kayser 1994 et Victorri 1994). Ensuite parce que le débat sur la nature discrète ou continue du sens lui-même est d'ordre théorique, voire métaphysique, puisqu'il faut à la fois définir précisément ce qu'est le sens et ce que l'on entend par discret et continu dans le cadre de cette définition du sens, hors du domaine des mathématiques. Tandis qu'en ce qui concerne les modèles, le débat ne porte pas sur leur caractérisation comme discrets ou continus, tout le monde s'accordant généralement facilement sur ce point, grâce à la définition rigoureuse que les mathématiques donnent de ces notions, mais sur l'intérêt plus ou moins grand de faire appel à un modèle discret ou à un modèle continu. Cette situation n'est d'ailleurs pas spécifique de la linguistique : le problème se pose de la même manière pour toutes les sciences. Ainsi, pour prendre l'exemple de la physique, savoir si la matière est discrète ou continue reste un problème métaphysique sur lequel les physiciens, comme les philosophes, peuvent s'opposer, alors qu'il y a un consensus très large sur l'intérêt respectif du discret et du continu pour modéliser la matière à différentes échelles.

C'est donc de l'intérêt des représentations discrètes et continues du sens des unités lexicales que nous discuterons ici, et, plus précisément, de leur capacité à rendre compte du comportement sémantique de ces unités, à la fois en synchronie et en diachronie.

La partition saussurienne

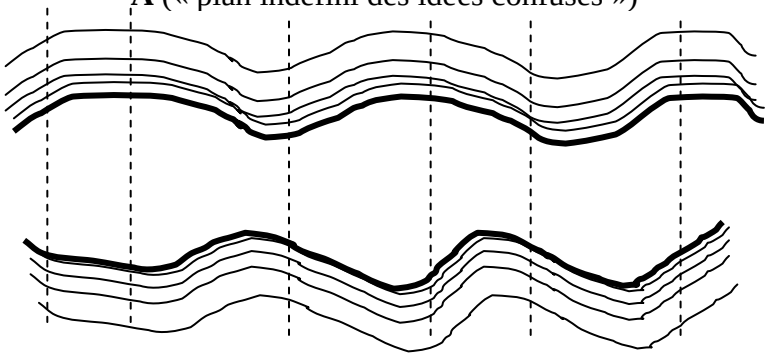
Notre point de départ sera le célèbre schéma du Cours de Saussure, présenté ici figure 1. Saussure pose très clairement la question du rapport entre le continuum de la pensée et les unités discrètes de la langue :

> « Nous pouvons donc représenter le fait linguistique dans son ensemble, c'est-à-dire la langue, comme une série de subdivisions contiguës dessinées à la fois sur le plan indéfini des idées confuses (A) et sur celui non moins indéterminé des sons (B) ; c'est ce que l'on peut figurer très approximativement par le schéma :
>
> [cf. fig. 1]
>
> Il n'y a donc ni matérialisation de la pensée, ni spiritualisation des sons, mais il s'agit de ce fait en quelque sorte mystérieux, que la « pensée-son » implique des divisions et que la langue élabore ses unités en se constituant entre deux masses amorphes. » (Saussure 1972 : 155-156).

Pour Saussure, il n'y a donc de structuration de la pensée que par le biais de la langue :

> « Psychologiquement, abstraction faite de son expression par les mots, notre pensée n'est qu'une masse amorphe et indistincte. Philosophes et linguistes se sont toujours accordés à reconnaître que, sans le secours des signes, nous serions incapables de distinguer deux idées de façon claire et constante. » (Saussure 1972 : 155)

A (« plan indéfini des idées confuses »)



B (« plan non moins indéterminé des sons »)

Figure 1 : La partition saussurienne

Si l'on adopte ce point de vue, le choix d'un modèle discret s'impose sans aucune équivoque. En effet, si le continuum de la pensée n'est qu'une « masse amorphe et indistincte », nous n'avons aucune raison de conserver dans le modèle les « points » de ce continuum. Seule compte la région à laquelle un point appartient dans la partition opérée par les unités discrètes de la langue, puisque cette partition est la seule source de différenciation au sein de ce continuum : deux points appartenant à une même région, c'est-à-dire correspondant à une même unité de la langue, ne sont tout simplement pas distinguables. Seules doivent être prises en compte les différences entre les unités, elles seules étant constitutives du système :

> « Un système linguistique est une série de différences de sons combinées avec une série de différences d'idées, mais cette mise en regard d'un certain nombre de signes acoustiques avec autant de découpures faites dans la masse de la pensée engendre un système de valeurs ; et c'est ce système qui constitue le lien effectif entre les éléments phoniques et psychiques à l'intérieur de chaque signe. » (Saussure 1972 : 166).

Ainsi l'approche structuraliste classique, tout en reconnaissant l'existence d'un continuum de la pensée, ne lui laisse aucune place dans son dispositif théorique : la correspondance entre les unités linguistiques et ce continuum est en quelque sorte à sens unique : c'est la langue qui impose sa structure discrète à la pensée.

La partition remise en question

Il est aujourd'hui difficile de défendre un point de vue aussi radical sur les relations entre pensée et langage. Les progrès de l'éthologie ont mis en évidence des capacités cognitives importantes chez diverses espèces animales : catégorisation, représentations de l'espace, des propriétés de l'environnement (végétaux, proies, prédateurs, etc.), des relations sociales, etc. (cf. par exemple Vauclair 1992 et Lestel 2001). Il est, à l'évidence, possible pour bien des animaux de distinguer deux idées sans avoir recours à des signes, contrairement à ce qu'affirmait Saussure. On peut donc penser que le langage s'est établi sur des formes de pensée sans doute frustes, mais déjà structurées, même si, une fois stabilisé, il a, en retour, profondément modifié et complexifié la pensée[1].

En conséquence, on ne peut plus soutenir que les unités discrètes de la langue opèrent des divisions sur un continuum amorphe. Ce continuum de la pensée est déjà structuré et il impose donc ses propres contraintes aux unités de la langue, limitant ainsi l'arbitraire de leurs découpages.

Prenons un exemple concret : le paradigme des couleurs. Hjelmslev, dans la lignée du passage de Saussure que nous avons commenté plus haut, utilise cet exemple, entre autres, pour expliquer comment la diversité des langues permet de « dégager » ce continuum amorphe :

> « On peut dire qu'un paradigme d'une langue et un paradigme correspondant d'une autre langue peuvent recouvrir une même zone de sens qui, détachée de ces langues, constitue un continuum amorphe et compact dans lequel les langues établissent des frontières. Derrière les paradigmes qui, dans les différentes langues,

[1] Voir aussi dans Carruthers 2002 §3 une argumentation allant dans le même sens dans une perspective cognitiviste.

> sont formés par les désignations des couleurs nous pouvons, par soustraction des différences, dégager ce continuum amorphe : le spectre des couleurs dans lequel chaque langue établit arbitrairement ses frontières. » (Hjelmslev 1968 : 76-77).

En réalité cette « zone de sens » recouverte par les paradigmes des couleurs dans différentes langues possède une structure indépendante de ces langues : l'organisation du spectre des couleurs est donnée par le système perceptif humain, qui lui confère une structure topologique spécifique (qui fait par exemple que la couleur orange se situe entre les couleurs jaune et rouge), et bien d'autres propriétés (ainsi certaines couleurs sont sans doute plus « saillantes » que d'autres)[2]. Toutes ces propriétés sont autant de contraintes qui s'appliquent à toutes les langues : aucune langue, par exemple, ne pourra posséder un terme qui désignerait à la fois le rouge et le jaune sans désigner aussi l'orange. Le spectre des couleurs, tout en étant effectivement un continuum, est muni d'une structure relativement riche, entièrement extra-linguistique, qui s'impose au langage. Autrement dit, et de manière générale, il faut cesser de qualifier systématiquement, comme le font Saussure et Hjelmslev, le continuum extra-linguistique d'amorphe et d'indistinct. La correspondance entre ce continuum et les unités discrètes de la langue est bien à double sens. L'arbitraire des frontières délimitant les régions occupées par les unités discrètes n'est pas absolu : il est contraint par la structure du continuum de sens dans lequel sont tracées ces frontières.

Prendre en compte cette structure sous-jacente de l'espace de sens permet aussi, et peut-être surtout, de rendre compte des phénomènes de polysémie et de synonymie partielle à l'intérieur d'une même langue. Les partitions saussuriennes ne laissent aucune place à ces phénomènes. En effet, reconnaître qu'une unité peut avoir des sens distincts dans différents énoncés suppose que des distinctions de sens soient possibles dans la région occupée par cette unité. De même, pour représenter le fait que deux unités sont interchangeables dans un énoncé sans en modifier sensiblement le sens, il faut que les deux unités partagent une sous-région de sens, autrement dit que les unités ne définissent pas une partition, au sens mathématique du mot, mais plutôt un recouvrement avec chevauchement des frontières. Cela n'est pas envisageable dans un modèle dans lequel les unités discrètes constituent la seule source de différenciation au sein du continuum. En revanche, comme on va le voir, c'est tout à fait possible si l'on munit ce continuum de sens d'une structure suffisamment riche sur laquelle les unités linguistiques opèrent leur discrétisation.

Prenons encore un exemple utilisé par Hjelmslev dans le même passage (Hjelmslev, 1968 : 78). Il s'agit du « champ sémantique » constitué par les mots français {*arbre*, *bois*, *forêt*}, qu'il met en correspondance avec les ensembles {*Baum*, *Holz*, *Wald*} de l'allemand et {*træ*, *skov*} du danois (cf. fig. 2). On peut douter de l'adéquation de la représentation unidimensionnelle qu'il donne pour chacun de ces ensembles, et qui est, bien sûr, la seule possible dans son cadre théorique. Le problème principal provient du fait que ce découpage regroupe deux dimensions différentes de sens : la dimension 'ensemble plus ou moins important d'arbres' (de l'arbre isolé à la forêt la plus étendue) et la dimension 'matière ligneuse' (du végétal au matériau de construction ou de chauffage).

[2] Pour un aperçu de la littérature très foisonnante qu'a suscitée le problème de la catégorisation des couleurs et de leur expression lexicale, voir Berlin et Kay 1969, Gellatly 1995, Saunders et van Brakel 1997, Hardin et Maffi 1997, Lucy 2002.

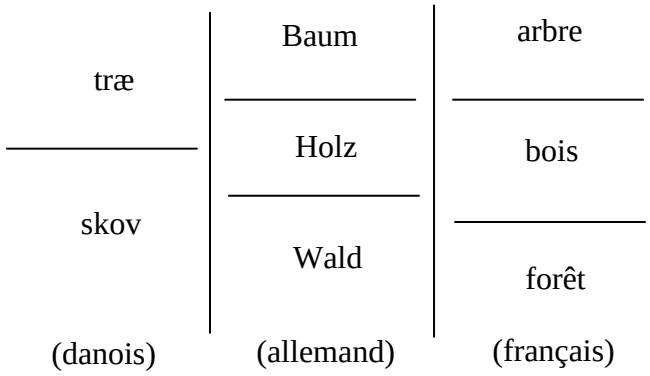


Figure 2 : le schéma de Hjelmslev

Du coup, le schéma de Hjelmslev ne permet pas vraiment de rendre compte de la polysémie des unités lexicales considérées, qui est différente dans les trois langues : on ne peut pas y deviner, par exemple, que le mot danois *træ* peut évoquer l'arbre ou la matière ligneuse, mais en aucun cas la collection d'arbres, et qu'à l'inverse *skov* ne peut pas, lui, être utilisé pour parler de la matière ligneuse. Il faut donc structurer le continuum de sens en en faisant un espace sémantique bidimensionnel pour que ces polysémies soient correctement représentées[3].

De plus, ce schéma ne montre pas que *bois* est partiellement synonyme de *forêt* dans la dimension 'ensemble d'arbres' (contrairement à d'autres unités du même paradigme comme *bosquet*) : il aurait fallu, pour indiquer cela, faire se chevaucher les espaces occupés par ces deux termes, ce qui est bien sûr impossible pour une théorie dans laquelle la seule source de différenciation provient des unités discrètes elles-mêmes.

On conçoit donc tout l'intérêt d'un modèle continu du sens : en se donnant un espace sémantique continu, multidimensionnel, muni d'une structure topologique, on peut rendre compte de propriétés essentielles des unités lexicales, comme la polysémie et la synonymie partielle. Comme nous allons le voir, ce type de modèle permet aussi de rendre compte de l'évolution diachronique des unités lexicales, et donc d'inscrire ces propriétés dans une dynamique évolutive.

Modélisation d'un système lexical

Comme dans le schéma saussurien, nous avons besoin pour construire notre modèle de deux espaces *A* et *B*, que nous appellerons *espace sémantique* et *espace phonétique*, et d'un ensemble d'unités discrètes u_i, que nous appellerons *unités lexicales*, qui se projettent d'une certaine manière dans les espaces *A* et *B*. Mais à l'inverse du schéma saussurien, les espaces *A* et *B* sont des espaces multidimensionnels munis d'une structure géométrique riche : ce sont des variétés différentielles[4]. De plus, nous ne ferons pas correspondre à chaque u_i des régions dans les espaces *A* et *B*, mais plutôt des fonctions, comme on l'a illustré figure 3, de manière à rendre compte de l'aspect graduel de la correspondance : une unité sera plus ou moins pertinente pour évoquer tel sens, et telle réalisation phonétique sera plus ou moins efficace pour énoncer cette unité.

[3] En fait, il faut bien plus de dimensions si l'on prend en compte toute la polysémie de ces mots : ainsi, pour *bois*, il faut aussi pouvoir représenter le sens de ramure de cervidés, d'instruments de musique (au pluriel), etc.

[4] Pour une justification du choix de cette structure mathématique pour l'espace sémantique, voir Victorri et Fuchs (1996 : 67-79). En ce qui concerne l'espace phonétique, qui n'est pas au centre de nos préoccupations ici, nous renvoyons à l'excellent travail de modélisation de Petitot (1985), qui montre bien l'intérêt théorique de ce type de structure mathématique pour l'interface phonétique-phonologie.

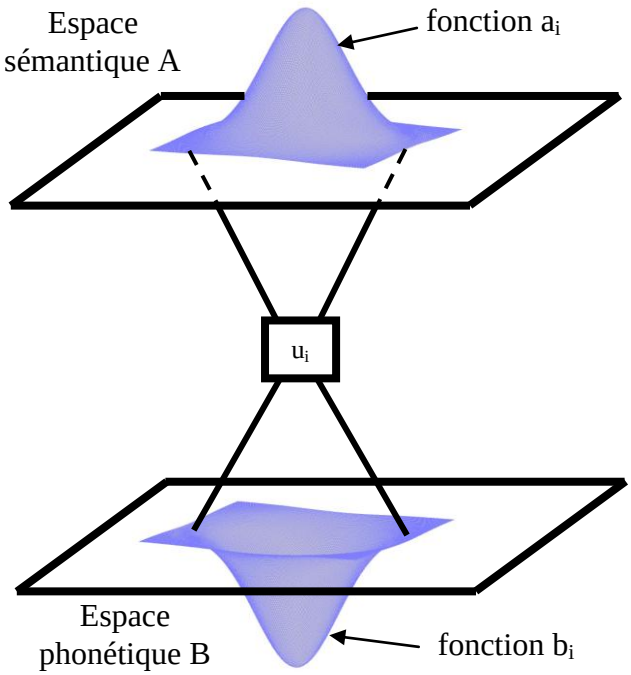


Figure 3 : Les deux composantes d'une unité lexicale

Plus rigoureusement, nous définissons un *système lexical* par la donnée de deux variétés différentielles *A* et *B*, et de deux ensembles de *n* densités de probabilité[5] $\{a_i\}$ et $\{b_i\}$ définies respectivement sur *A* et sur *B*. Une unité lexicale est un couple $u_i = (a_i, b_i)$. La fonction a_i est appelée la *composante sémantique* de l'unité u_i et la fonction b_i sa *composante phonétique*.

Un tel système lexical peut être plus ou moins grand et complexe, suivant le nombre d'unités lexicales qu'il contient et surtout le nombre de dimensions des espaces *A* et *B*. Dans les systèmes les plus simples, comme celui représenté figure 4, les espaces *A* et *B* n'ont qu'une dimension. Cela signifie que les quatre unités de ce système ne diffèrent au plan phonétique que par les valeurs d'un paramètre (par exemple, le degré d'aperture d'une voyelle), et qu'elles forment au plan sémantique un paradigme très simple, représentable par les variations d'une seule variable, comme par exemple le paradigme de la température extérieure (*il fait froid* / *frais* / *doux* / *chaud*).

Mais on peut bien entendu, en augmentant les dimensions des espaces *A* et *B*, représenter des paradigmes beaucoup plus complexes et des réalisations phonétiques plus réalistes. De fait, on pourrait théoriquement modéliser tout le lexique d'une langue par un seul système lexical de ce type, si l'on était capable de spécifier l'ensemble des dimensions sémantiques pertinentes.

[5] On appelle densité de probabilité sur une variété différentielle une fonction intégrable réelle non-négative de norme égale à 1, la norme étant définie comme la somme (intégrale) de la fonction sur la variété.

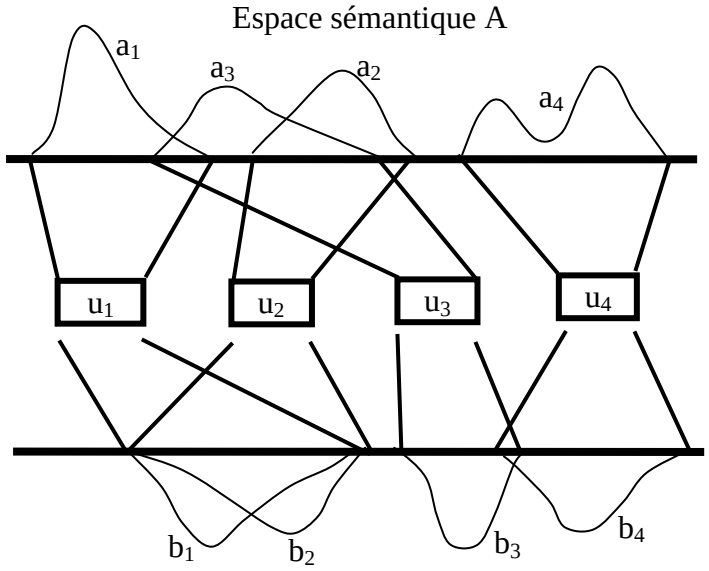


Figure 4 : Représentation d'un système lexical

On peut facilement représenter différents phénomènes sémantiques dans ce modèle. Ainsi, deux unités u_i et u_j sont homophones si leurs composantes phonétiques b_i et b_j se superposent alors que leurs composantes sémantiques a_i et a_j sont disjointes[6] (c'est le cas pour les unités u_1 et u_2 sur la figure 4). De même, on aura synonymie totale ou partielle si les composantes sémantiques a_i et a_j se recouvrent totalement ou partiellement (exemple : u_2 et u_3 sur la figure). Enfin une unité u_i sera polysémique si sa composante sémantique a_i possède plusieurs maxima et un support connexe (comme u_4 sur la figure).

Dynamique évolutive

L'un des intérêts essentiels d'un tel modèle, c'est qu'on peut le rendre dynamique, et donc tester la stabilité de différentes configurations dans diverses conditions. En effet, on peut, en faisant quelques hypothèses assez simples sur la manière dont se modifient les fonctions a_i et b_i en fonction de l'usage des unités lexicales, modéliser leur évolution dans le temps, et donc observer si telle ou telle propriété du système, comme par exemple l'existence de polysémie ou de synonymie partielle, émerge, reste stable ou au contraire disparaît.

Classiquement, on modélise une telle dynamique évolutive en se plaçant dans le cadre des systèmes multi-agents (cf., entre autres, Cangelosi et Parisi 1998, Steels 1996, Kaplan 2001, Oudeyer 1997, Kirby 1999, 2000) : ce sont des agents, munis de capacités cognitives plus ou moins importantes, qui utilisent le système de communication, et qui le transforment progressivement en fonction des succès et échecs de leurs échanges. Plus précisément, chaque agent possède une représentation interne du système qui lui est propre, et c'est cette représentation qu'il modifie en fonction de son appréciation du succès de la communication. Pour juger de l'évolution du système dans toute la communauté des agents, on fait en quelque sorte la « moyenne » de toutes ces représentations individuelles à un instant donné : c'est cette moyenne qui est considérée comme l'état du système de communication à cet instant. Pour notre part, nous allons travailler d'emblée sur cette moyenne, en faisant l'hypothèse que chaque échange vient modifier directement l'état du système lexical. Ainsi, nous éviterons le passage par un système multi-agents, tout en conservant l'essentiel de l'esprit de ce type de modélisation : le système lexical évolue

[6] Plus précisément, ce sont les *supports* de ces fonctions (région de l'espace où la fonction est non nulle) qui doivent se superposer ou être disjoints.

parce qu'il se produit des événements de parole qui utilisent les unités lexicales et qui, suivant qu'ils ont été des succès ou des échecs, modifient ces unités dans leurs composantes phonique et sémantique. En d'autres termes, il s'agit, en nous limitant au niveau lexical, de mettre en œuvre de la manière la plus simple possible le principe suivant lequel la langue produit la parole qui en retour transforme la langue.

Nous définissons donc des événements que nous appellerons *énonciations*, et qui constitueront la partie dynamique de notre modèle. Chaque énonciation est un processus qui se déroule en quatre étapes que nous allons détailler : choix d'un sens visé, production, réception, et adaptation du système.

La première étape consiste donc à définir le sens visé par l'énonciation. Si un locuteur est amené à utiliser le système lexical, c'est parce que le sens qu'il veut évoquer est couvert par ce système, autrement dit parce que ce sens visé peut être représenté dans l'espace sémantique *A*. Comme pour les composantes sémantiques des unités lexicales, nous modélisons le sens visé par une densité de probabilité sur l'espace *A*, que nous notons *s*.

La deuxième étape modélise le processus de production par un locuteur. En fonction du sens visé, il s'agit d'abord de sélectionner une unité lexicale susceptible d'évoquer ce sens. Pour ce faire, on compare le sens visé avec la composante sémantique de chacune des unités du système. L'unité u_i a d'autant plus de chance d'être prise que sa composante sémantique est plus proche du sens visé *s*. Techniquement, on choisit cette unité, appelons-la u_p, par un tirage aléatoire au sein des u_i, chaque unité étant pondérée par le degré de recouvrement[7] de *s* avec sa composante sémantique a_i.

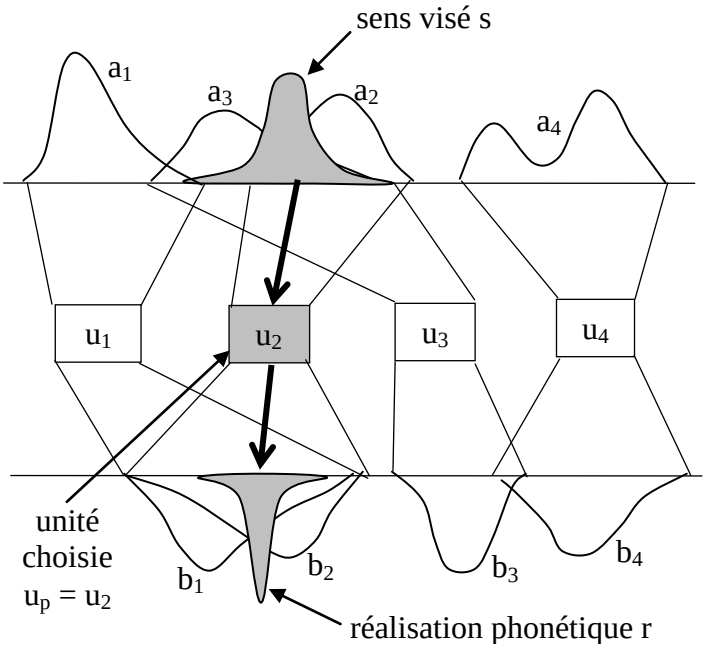


Figure 5 : Modélisation du processus de production

Une fois l'unité u_p sélectionnée, le locuteur doit choisir une réalisation phonétique de cette unité. Pour cela on utilise la fonction b_p (composante phonétique de u_p). Une valeur phonétique a d'autant plus de chance d'être choisie qu'elle correspond à une valeur forte de la fonction b_p. Techniquement, une réalisation phonétique est modélisée elle aussi par une densité sur *B* : une gaussienne de très faible écart-type dont le centre est donné par un tirage aléatoire avec la distribution b_p.

[7] Le degré de recouvrement de deux densités est défini comme la somme (l'intégrale) du produit de ces deux fonctions sur la variété.

Appelons *r* la réalisation phonétique ainsi obtenue. Ainsi le processus de production consiste à passer d'un sens visé *s* à une réalisation phonétique *r*, comme l'illustre la figure 5. La troisième étape, le processus de réception, illustré figure 6, consiste en l'opération inverse : passage de la réalisation phonétique *r* à un sens compris *s'*, suivant les mêmes principes : choix d'une unité u_q en comparant la réalisation phonétique *r* avec la composante phonétique de chaque unité u_i du système, puis choix d'une gaussienne *s'* en utilisant la distribution de probabilité a_q (composante sémantique de l'unité u_q).

Il faut noter que u_p et u_q peuvent être une seule et même unité, mais que cela n'est pas obligatoire, à cause des recouvrements partiels possibles des composantes phonétiques : c'est ce que l'on a illustré sur les figures 5 et 6, où l'unité u_1 a été choisie en réception, alors que c'est l'unité u_2 qui avait été choisie en production.

D'une manière générale, les phénomènes d'homophonie, totale ou partielle, de même que la polysémie, peuvent provoquer des discordances entre le sens compris et le sens visé. La dernière étape de l'énonciation consiste justement à adapter le système en fonction des résultats obtenus par les étapes précédentes. On compare donc *s'*, le sens compris, avec *s*, le sens visé, pour juger du succès ou de l'échec de l'énonciation[8], et pour modifier les unités impliquées en conséquence.

Les modifications touchent d'une part la composante phonétique b_p de l'unité choisie en production et d'autre part la composante sémantique a_q de l'unité choisie en réception.

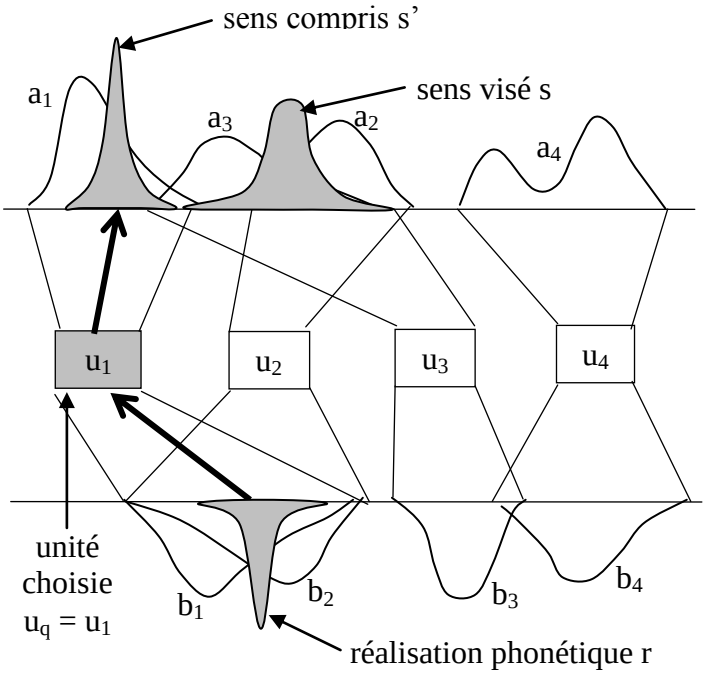


Figure 6 : Modélisation du processus de réception

En ce qui concerne la composante phonétique b_p, la règle d'adaptation est la suivante : suivant qu'il y a eu succès ou échec de l'énonciation, on renforce[9] positivement ou négativement la zone *r* de la fonction b_p.

[8] C'est le degré de recouvrement de *s* et de *s'* qui sert à calculer le résultat : il y a succès s'il est supérieur à un certain seuil, et échec dans le cas contraire.

[9] La nouvelle fonction b_p est obtenue à partir d'une somme pondérée de l'ancienne fonction b_p et de la fonction *r*, cette dernière étant affublée d'un poids positif ou négatif suivant que le renforcement est lui-même positif ou négatif.

En ce qui concerne la composante sémantique a_q, la règle est différente : il y a toujours renforcement positif de la zone s de la fonction a_q, mais ce renforcement est faible s'il y a eu échec alors qu'il est plus fort s'il y a eu succès.

Les principes qui ont inspirés ces règles d'adaptation sont simples. Du point de vue de l'énonciateur, le fait d'avoir réussi ou non à se faire comprendre va l'encourager ou non à réutiliser la même réalisation phonétique pour l'unité lexicale u_p qu'il a choisie. Quant au co-énonciateur, il enregistre de toute façon qu'un locuteur a pu utiliser l'unité u_q pour évoquer le sens visé s, mais il va accorder plus ou moins de poids à cette information, suivant qu'elle corrobore ou non ses connaissances antérieures sur le sens de l'unité en question.

Le modèle muni de la dynamique conférée par ce mécanisme d'énonciation-adaptation permet donc de représenter un système lexical en constante évolution. Bien entendu, c'est un modèle extrêmement simpliste par rapport à la réalité de l'évolution diachronique du lexique d'une langue. En particulier, il n'accorde aucune place aux relations syntagmatiques, qui sont en quelque sorte rejetées dans les conditions d'énonciation, puisqu'il est centré sur la production d'une seule unité à la fois. Il faudrait aussi le complexifier considérablement pour qu'il puisse rendre compte de phénomènes diachroniques fondamentaux comme la grammaticalisation, l'innovation, etc. Mais l'objectif essentiel de ce type de travail n'est pas de modéliser de façon « réaliste » ces phénomènes hautement complexes, mais de rechercher les conditions minimales dans lesquelles peuvent émerger les propriétés linguistiques auxquelles on s'intéresse.

Le statut de ce modèle est donc essentiellement épistémologique : il permet de valider (ou d'invalider) une réflexion théorique, en fournissant des exemples de systèmes dans lesquels émergent ou non les propriétés étudiées. De ce point de vue, il faut que ces systèmes restent les plus simples possibles : en effet, il est indispensable de bien maîtriser leurs différentes composantes pour pouvoir interpréter à bon escient les résultats obtenus. Il est toujours possible de rajouter des paramètres et des mécanismes supplémentaires pour que le modèle « ressemble » plus aux objets que l'on étudie, mais cela ne fait que brouiller son utilisation, car on ne sait plus à quel facteur imputer l'émergence de tel ou tel de ses comportements.

Ainsi cette activité de modélisation peut se révéler fructueuse dans les recherches sur l'évolution des langues, à condition de bien mesurer ses limites : une fois de plus, ces modèles n'ont pas l'ambition de reconstruire les conditions exactes dans lesquelles les langues évoluent, mais, plus modestement, d'aider à évaluer le rôle potentiel de tel ou tel facteur, dans le cadre d'un réflexion théorique sur l'émergence et la stabilisation de certaines propriétés spécifiques du langage. C'est en tout cas dans cet esprit que nous avons conçu ce modèle d'évolution « lexicale ».

Simulations informatiques

Nous avons réalisé un logiciel, EVOLEX[10], qui implémente fidèlement le modèle mathématique présenté ci-dessus. On peut y spécifier un système lexical comportant un nombre quelconque d'éléments sur des espaces sémantique et phonétique de dimension quelconque, du moins en principe : en pratique, pour des raisons de temps de calcul et d'espace mémoire, on doit se restreindre dans la version actuelle à des systèmes de moins de cinquante unités et à des espaces de très petites dimensions (deux au maximum). On peut étudier l'évolution dans le temps d'un tel système en simulant une suite d'énonciations : il suffit pour cela de spécifier une suite de fonctions densités sur l'espace sémantique qui correspondent chacune au sens visé de chaque énonciation de la séquence. Une interface conviviale permet de choisir facilement les différents paramètres et de visualiser[11] l'évolution du système au cours du temps à l'aide de représentations graphiques du type de celles qui ont été présentées ci-dessus (fig. 4-6).

Nous avons fait deux séries de simulations à l'aide de ce logiciel. Pour chacune de ces séries, nous avons varié le nombre d'unités du système et, dans une moindre mesure, la dimension des espaces

[10] EVOLEX est écrit dans le langage de programmation MATLAB.
[11] Dans la version actuelle, la visualisation n'est possible que pour des espaces sémantique et phonétique unidimensionnels.

sémantique et phonétique[12]. Au démarrage de chaque simulation, l'état initial était choisi au hasard : les composantes sémantique et phonétique de chaque unité étaient des gaussiennes de moyenne et d'écart type aléatoires. L'objectif de cette expérience était de découvrir dans quelles conditions le système se stabilisait au cours du temps et quelles propriétés des unités restaient elles-mêmes stables.

Les deux séries de simulations ne différaient que par une condition essentielle, portant sur les caractéristiques des sens visés. Dans la première série, tous les sens visés étaient des gaussiennes de moyenne aléatoire, mais d'écart-type fixe et très faible. Dans la deuxième série, les sens visés étaient aussi des gaussiennes de moyenne aléatoire, mais cette fois d'écart-type variable, lui aussi choisi aléatoirement. Autrement dit, ce qui change entre les deux séries, c'est le caractère vague ou non de ce qu'ont à dire les locuteurs. Dans la première condition, le sens visé est toujours très précis, alors que dans la deuxième il est souvent beaucoup plus large. Pour être concret, reprenons l'exemple du paradigme de la température extérieure. La première condition correspond à des locuteurs qui voudraient toujours indiquer la température exacte, à quelques degrés près : il s'agit de dire qu'il fait autour de 10°, -5°, 25°, etc. En revanche, dans la deuxième, l'intervalle de température évoqué peut être très variable : un locuteur peut vouloir dire qu'il fait plutôt frais, pas très chaud, assez froid, etc. aussi bien que vraiment glacial, très chaud, ou autour de 10°.

Les résultats de ces simulations sont très tranchés. Dans toutes les simulations de la première série, le système se stabilise très vite et n'évolue pratiquement plus. Les composantes sémantiques des unités forment une partition de l'espace sémantique sans aucun recouvrement, chaque unité occupant une région de même taille. Dans l'espace phonétique, les régions occupées sont beaucoup plus étroites et très espacées les unes des autres : on observe que ces petites régions peuvent se déplacer quelque peu sur le long terme, mais sans jamais se rejoindre. Il n'y a donc pas d'homonymie, ni de polysémie, ni de synonymie totale ou partielle : les unités sont toutes parfaitement distinctes dans leur composante phonétique, et elles se partagent équitablement, si l'on peut dire, les sens à exprimer : plus le nombre d'unités est grand, plus le sens de chacune est précis.

Dans la deuxième série au contraire, le système ne se stabilise jamais : les unité continuent perpétuellement de se modifier, tant dans leur composante sémantique que phonétique. Dans l'espace sémantique, on observe en permanence des recouvrements partiels (synonymie partielle) et des fonctions à plusieurs maxima (polysémie). Les déplacements sont continuels, ce qui fait que l'on observe des changements complets de région occupée par chaque unité sur des périodes suffisamment longues. Dans l'espace phonétique, les composantes occupent des régions beaucoup plus grandes que dans le cas précédent, avec assez fréquemment des recouvrements partiels (sorte « d'homonymie partielle »). Les déplacements sont continuels, et comme pour la composante sémantique ils aboutissent sur le long terme à des changements radicaux de région occupée par une unité donnée.

Ainsi, notre modèle, pourtant minimal, peut exhiber un nombre non négligeable de propriétés lexicales des langues : évolution continue, présence permanente de polysémie et de synonymie partielle. La condition qui semble cruciale pour l'apparition de ces phénomènes se situe du côté du « l'intenté » des locuteurs : elle est la conséquence, du moins dans notre modèle, du fait que ce que les locuteurs cherchent à exprimer présente une grande variabilité, notamment dans le degré de précision des catégorisations et des qualifications des entités (et des procès) qu'ils veulent évoquer. Bien entendu, comme nous l'avons déjà signalé, ce type d'expérimentation ne permet pas d'affirmer qu'il en va de même pour les systèmes lexicaux de nos langues. Ce n'est donc pas une preuve, mais un élément de réflexion qui peut servir à orienter les recherches sur les raisons profondes de l'existence et la permanence de la polysémie et de la synonymie partielle. Très clairement, les résultats obtenus sur ce modèle simplifié pousse à rechercher ces causes du côté de l'utilisation des langues, c'est-à-dire dans les phénomènes d'énonciation.

Conclusion

[12] La plupart des simulations ont été réalisées avec un système d'une dizaine d'unités ou moins sur des espaces unidimensionnels. Seul un petit nombre d'entre elles ont utilisé des espaces bidimensionnels et un nombre plus important d'unités (de l'ordre de la cinquantaine).

Au delà de ces résultats, ce modèle montre tout l'intérêt de l'utilisation du continu en sémantique lexicale. En effet, si des propriétés essentielles comme la polysémie réclament pour être comprises et explicables l'étude de phénomènes de parole, on ne peut pas éviter la modélisation de ces espaces extra-linguistiques cognitifs que Saussure appelait « le plan indéfini des idées confuses ». C'est au contraire en en représentant la structure et les propriétés (notamment topologiques) que l'on peut construire des modèles non réducteurs, capables de nous éclairer sur les relations sémantiques lexicales, d'abord au sein du lexique d'une même langue, mais aussi entre lexiques de différentes langues.

C'est aussi le cadre qui permet de mieux appréhender l'évolution dans le temps de ces relations, et donc de traiter des phénomènes diachroniques, sans pour autant faire l'erreur dénoncée par Saussure d'y voir la cause, ou l'explication, des propriétés synchroniques. La prise en compte de ces espaces cognitifs permet au contraire de faire coexister dans un même cadre de modélisation propriétés synchroniques et diachroniques, résultant les unes et les autres des interactions entre langue et parole.

Bibliographie

Berlin B., Kay P.
1969, *Basic Color Terms: Their Universality and Evolution*, University of California Press.

Cangelosi A., Parisi D.
1998, The Emergence of a "Language" in an Evolving Population of Neural Networks, *Connection Science*, 10(2), p. 83-97.

Carruthers P. 2002, The cognitive functions of language, *Behavioral and Brain Sciences*, 25:6.

Gellatly A. 1995, Colourful whorfian ideas: Linguistic and cultural influences on the perception and cognition of colour, and on the investigation of them, *Mind and Language*, 10(3), p. 199-225.

Hardin C, Maffi L. (éds.)
1997, *Color categories in thought and language*, Cambridge University Press.

Hjelmslev L. 1968, *Prolégomènes à une théorie du langage*, Paris, Minuit.

Kaplan, F. 2001, *La naissance d'une langue chez les robots*, Paris, Hermès.

Kayser D. 1994, What kind of models do we need for the simulation of understanding?, in C. Fuchs et B. Victorri, *Continuity in linguistic semantics*, Amsterdam, John Benjamins, p. 241-250.

Kirby S. 1999, *Function, Selection and Innateness: The Emergence of Language Universals*, Oxford University Press.

Kirby S. 2000, Syntax without Natural Selection: How compositionality emerges from vocabulary in a population of learners, in C. Knight, M. Studdert-Kennedy, J.R. Hurford, *The Evolutionary Emergence of Language*, Cambridge University Press, p. 303-323.

Lestel D. 2001, *Les origines animales de la culture*, Paris, Flammarion.

Lucy J. A. 2002, *Language Diversity and Thought. A Reformulation of the Linguistic Relativity Hypothesis*, Cambridge University Press, chap. 5.

Oudeyer P-Y 1997, Self-organization of a lexicon in a structured society of agents, in P. Husbands, I. Harvey (eds.), *Proceedings of the Fourth European Conference on Artificial Life*, MIT Press, p. 726-729.

Petitot J. 1985, *Les catastrophes de la parole*, Paris, Maloine.

Saunders B., van Brakel J.
1997, Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences*, 20(2), p. 167-228.

Saussure F.de 1972, *Cours de linguistique générale*, Paris, Payot.

Steels L. 1996, Self-organizing vocabularies, in C. Langton, T. Shimohara (éds.) *Artificial Life V*, MIT Press.

Vauclair J. 1992, *L'intelligence de l'animal*, Paris, Seuil.

Victorri B. 1994, The use of continuity in modelling semantic phenomena, in C. Fuchs et B. Victorri, *Continuity in linguistic semantics*, Amsterdam, John Benjamins, p. 241-250.

Victorri B., Fuchs C.
1996, *La polysémie. Construction dynamique du sens*, Paris, Hermès.