De l'analogie rendant compte de la commutation en linguistique

Yves Lepage

▶ **To cite this version:**

Yves Lepage. De l'analogie rendant compte de la commutation en linguistique. Interface homme-machine [cs.HC]. Université Joseph-Fourier - Grenoble I, 2003. tel-00004372

HAL Id: tel-00004372
https://theses.hal.science/tel-00004372

Submitted on 29 Jan 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

YVES LEPAGE

DE L'ANALOGIE
rendant compte de la commutation en linguistique

Mémoire d'**habilitation à diriger des recherches**
défendu publiquement le 23 mai 2003
à l'université Joseph Fourier – Grenoble I
devant le jury composé de :

M. Jean CAELEN,	directeur de recherche au CNRS,	président,
M. Christian BOITET,	professeur des Universités,	parrain,
M. Hassan AÏT-KACI,	docteur d'État,	rapporteur,
M. Fathi DEBILI,	directeur de recherche au CNRS,	rapporteur,
M. Éric WEHRLI,	professeur des Universités,	rapporteur,
M. Marc DYMETMAN,	docteur d'État,	examinateur.

Table des matières

Liste des tableaux

Liste des figures

Liste des algorithmes

Remerciements

Je tiens à exprimer ma gratitude envers les directeurs successifs du laboratoire d'interprétation automatique d'ATR, MM. Kurematu, Yamazaki et Yamamoto. En plus, je remercie le plus chaudement du monde M. Sirai pour le soutien constant qu'il apporte aux chercheurs de son département et donc à notre groupe, le groupe ALEPH. Cet acronyme signifie Analogie, Langue Et Procédés par Homomorphismes et il rend compte, comme on le verra en lisant ce mémoire, des axes principaux de notre recherche. J'adresse donc aussi mes remerciements à tous les membres du groupe ALEPH qui se sont succédés au cours des dernières années : Jean-François Morreeuw, Eddy Taillefer, Laurent Knoll, Juliette-Izoumi L'Hériteau et Nicolas Auclerc, qui y a travaillé deux ans et demie à temps plein.

Je tiens aussi à exprimer ma reconnaissance envers Christian Boitet qui m'a proposé de me lancer dans l'entreprise ardue qu'a été la rédaction de ce mémoire. Il m'a soutenu, aiguillé, aiguillonné, lu, relu, conseillé et encouragé tout au long des derniers mois.

Avant-propos

Les analogies entre chaînes de symboles, notées $A : B = C : D$, mettent quatre chaînes de symboles en proportion. Elles rendent compte, par exemple, de *exact* : *inexact* = *fini* : *infini* ou *fable* : *fabuleux* = *miracle* : *miraculeux*, au niveau des chaînes de symboles. Elle ne sont pas faites pour rendre compte directement, par exemple, de *je prends* : *prendre* = *je viens* : *venir*, ni de *oiseau* : *ailes* = *poisson* : *nageoires* qui supposent des connaissances sur la langue ou sur le monde. Les analogies peuvent être lues comme des égalités, comme dans l'exemple arabe suivant :

$$aslama : muslimun = arsala : mursilun \quad [1]$$

ou comme des équations à résoudre, comme dans :

$$aslama : muslimun = arsala : x \quad \Rightarrow \quad x = mursilun$$

Les fondateurs de la linguistique moderne, de Baudouin de Courtenay à Saussure, en passant par Kuryłowicz, ont tous mentionné le phénomène, sans chercher à l'expliquer. Mais il a été profondément ressenti comme crucial dans la langue. Cela est clair pour la morphologie où l'effet synchronique de l'analogie ($actorem : actor = honorem : x \quad \Rightarrow \quad x = honor$) peut être distingué de l'effet diachronique des changements phonétiques (/honosem/ → /honorem/, par rhotacisme /s/ → /r/ entre voyelles), mais aussi pour la syntaxe selon les hypothèses d'Hermann Paul, de Bloomfield ou d'Itkonen. De nos jours, alors que la modélisation de la langue est un choix légitimé en linguistique[2], il est frappant de constater qu'un petit nombre seulement de propositions ont été faites pour la modélisation de l'analogie, les rares exceptions étant Itkonen et aussi le groupe de Hofstadter hors de la linguistique. Il faut peut-être en voir la cause dans ce que le courant dominant (et dominateur) en linguistique pendant de nombreuses années, le courant générativiste, rejetait explicitement l'analogie comme objet possible de recherche, ce qui en fait d'ailleurs une exception dans l'histoire de la linguistique[3]. Le but ultime

[1] *Arsala* (il envoya) et *aslama* (il se convertit [à l'Islam]) sont des verbes au passé, 3^{e} personne du singulier ; *mursilun* (un envoyé) et *muslimun* (un converti [à l'Islam], c'est-à-dire un musulman) sont des noms.

[2] Voir Mel'čuk, *Vers une linguistique Sens-Texte. Leçon inaugurale*, 10 janvier 1997, p. 2–4.

[3] Voir Itkonen & Haukioja, *A rehabilitation of analogy in syntax (and elsewhere)*, 1997, p. 132 et 136 pour des citations de Chomsky.

du présent travail, que nous n'avons pas encore atteint, est de proposer un algorithme pour la vérification et la résolution de l'analogie entre chaînes de symboles qui respecte toutes les propriétés que le bon sens attribue à l'analogie, et seulement celles-là, et qui permette aussi d'entrevoir une généralisation par exemple aux images et au signal sonore. Ce que nous cherchons à atteindre, ce ne sont ni des règles, ni des lois, mais la formalisation d'une opération que tout le monde reconnaît être à l'œuvre dans la langue. Ici, il nous faut ajouter que si certains ont refusé de considérer le phénomène scientifiquement, c'est, à notre avis, parce que, précisément, il lui manquait une formalisation.

Pour faire un parallèle avec le monde des nombres, si notre travail se situait en arithmétique, son objet ne serait pas, par exemple, de chercher une loi sur les nombres premiers, ou une règle de réécriture de certains nombres en facteurs particuliers. Il s'agirait simplement de proposer un algorithme permettant d'additionner n'importe quel couple d'entiers ! Donc notre but est à la fois humble et ambitieux. Humble, parce que l'opération que nous cherchons à formaliser est des plus intuitives : $\mathit{dire} : \mathit{je\ dis} = \mathit{faire} : x \quad \Rightarrow \quad x = \mathit{je\ fais}$ ou $\mathit{un\ cheval} : \mathit{des\ chevaux} = \mathit{un\ amiral} : x \quad \Rightarrow \quad x = \mathit{des\ amiraux}$. Ambitieux, parce que, une fois cette formalisation obtenue, nous voulons montrer que la résolution dès lors possible des analogies facilitera la tâche du traitement automatique des langues. Mais attention, si l'on ne saurait se passer de l'addition pour calculer une transformée de Fourier, savoir faire une addition n'implique pas la connaissance immédiate de la transformée de Fourier. De même, ce n'est pas parce que, à l'issue de notre travail nous saurions résoudre des analogies, que les nombreuses problématiques du traitement automatique des langues s'en trouveront du même coup toutes simplifiées au point où elles seraient quasiment résolues.

Notre regret est de présenter ce document alors que nous n'avons toujours pas atteint notre but. Il s'agit d'une recherche en cours. On verra plus bas que deux volets apparaissent clairement dans l'analogie : la similitude et la contiguïté. Si le volet de la similitude est maintenant bien compris et se présente de façon relativement élégante, le volet de la contiguïté marque les limites de nos travaux. Ainsi donc, comme la cathédrale de la Vierge à Cracovie exhibe deux tours de hauteurs inégales, de même notre manuscrit montre deux volets de recherche inégaux. Et plus, comme de la plus haute tour de cette cathédrale retentit tous les jours à midi le *hejnał*, signal de trompette qui annonça l'arrivée pleine de menaces des Mongols dans les plaines polonaises, de même, la date fatidique du 1$^{\text{er}}$ novembre 2001 a sonné pour nous du timbre du regret : nous ne pourrons mettre sous les yeux de nos lecteurs qu'une formalisation partielle. Il y a un an, nous avons eu l'imprudence de croire qu'il nous serait possible de trouver ce qui caractérise la contiguïté dans l'analogie entre chaînes de caractères avant le délai de soumission[4] de ce document. Nous avons appris à nos dépens que la recherche ne se planifie pas.

[4]Que nous n'avons même pas pu tenir !

Introduction

Question simple : qu'est-ce qui fait qu'un locuteur de la langue française admette sans mal et rie de bon cœur au type de blagues éculées que l'on trouve par exemple chez San Antonio ?

> La menace le distrait de sa peine.
> Il froncelesourcile[a] et sa bouche s'écarte pour une muette interrogation.
> [...]

[a] Du verbe froncelesourciler (premier groupe).

La réponse est simple, c'est l'analogie : *froncelesourciler* est à *il froncelesourcile* ce que *manger* est à *il mange*. Ce que l'on note :

$$froncelesourciler : il\ froncelesourcile = manger : il\ mange$$

On a ici une analogie entre mots, dans le cadre de la morphologie, c'est-à-dire, de façon générale, entre chaînes de symboles. Cet exemple a l'avantage de nous donner la mesure de la puissance de l'analogie. Premièrement, l'analogie est **universelle**. Le fonctionnement de l'analogie est supposé connu de tous, il peut donc être implicite dans l'énoncé de la blague. Deuxièmement, l'analogie est **créatrice** comme cela est révélé par cet exemple. Elle autorise la création d'un verbe nouveau, jamais rencontré auparavant. Troisièmement, l'analogie est **aveugle**. Ce qui fait rire, c'est le caractère absurde de la création, du fait de l'inexactitude de la découpe. Or, si le comique est selon Bergson du mécanique plaqué sur du vivant, l'application d'une procédure mécanique, l'analogie, dans l'une des expressions du vivant, la langue, fait bien la *vis comica* de la blague. En résumé cet exemple nous livre un condensé des caractères de l'analogie : c'est une opération universelle et créatrice, mais aveugle.

Caractère universel de l'analogie

L'analogie possède une propriété remarquable : elle est indépendante des symboles utilisées. Que l'on écrive

$$froncelesourciler : il\ froncelesourcile = manger : il\ mange$$

ou

[5] San Antonio, *Si, signore !*, 1974, p. 145.

$$\texttt{froncelesourciler} : \texttt{il froncelesourcile} = \texttt{manger} : \texttt{il mange}$$

ou encore

$$\text{FRONCELESOURCILER} : \text{IL FRONCELESOURCILE} = \text{MANGER} : \text{IL MANGE}$$

et voire même

$$\phi\rho o\nu\kappa\varepsilon\lambda\varepsilon\varsigma o\upsilon\rho\kappa\iota\lambda\varepsilon\rho : \iota\lambda\ \phi\rho o\nu\kappa\varepsilon\lambda\varepsilon\varsigma o\upsilon\rho\kappa\iota\lambda\varepsilon = \mu\alpha\nu\gamma\varepsilon\rho : \iota\lambda\ \mu\alpha\nu\gamma\varepsilon$$

ne change rien à cette analogie, qui reste la même. Derrière cette indifférence au codage, se cache donc nécessairement une opération fondamentale du même ordre que l'addition, elle aussi indifférente aux représentations des chiffres et des nombres. Fondamentalement donc, l'analogie induit une structure sur les chaînes de symboles de la même façon que l'addition induit une structure sur les nombres entiers. Et c'est cette structure qui serait universelle.

L'universalité dont nous parlons est une universalité humaine : tout homme serait capable de comprendre des analogies et d'en former sur n'importe quel ensemble de symboles. Mais dans aucun des textes que nous avons pu lire [6], cette assertion de l'universalité de l'analogie n'est clairement justifiée. Peut-être nous trompons-nous, mais nous n'avons pas trouvé de références à des travaux de psychologie expérimentale qui poseraient explicitement cette assertion comme une hypothèse, et chercheraient à l'invalider ou à la confirmer.

La tentation d'extérioriser une faculté humaine et de la projeter dans le monde est grande[7]. Les Encyclopédistes y succombent puisqu'ils n'hésitent pas à écrire :

> On fait en Physique des raisonnemens très-solides par analogie. Ce sont ceux qui sont fondés sur l'uniformité connue, qu'on observe dans les opérations de la nature ;[8]

Nous sommes tentés d'attaquer le raisonnement même. Il repose en effet, selon nous sur une conception trop longue à discuter ici qui nous vient de la science grecque. D'aucuns pensent en effet que l'immense découverte de la science grecque, c'est précisément la proportion. De là, pour les mêmes, le fait que toute la science grecque se réduirait à une science des proportions : recherche de celles-ci dans la Nature, et explicitation de lois proportionnelles. Dès lors, la conséquence extrême serait l'incapacité des Grecs anciens à penser les lois de la Nature autrement que comme proportions, c'est-à-dire, en gros, à dépasser l'équation linéaire. Ces considérations débordent largement de l'objet de nos travaux, mais on lira avec bénéfice certains développements de Bergson à ce sujet[9]. Toujours est-il que le caractère aveugle de l'analogie vient sans doute

[6] Même chez ITKONEN, *Iconicity, analogy, and universal grammar*, 1994, p. 46 à 50, où le sujet, explicitement abordé, n'est pas prouvé.

[7] Lors d'un exercice de probabilités au cours duquel, de nouveau, la courbe de Gauss collait parfaitement à je ne sais quel phénomène naturel, notre professeur de probabilités s'exclama un jour : « Encore une courbe de Gauss ! Est-elle dans la nature, ou plutôt — pointant l'index vers son front — ne serait-elle pas ici, que nous la voyons partout ? »

[8] DIDEROT & d'ALEMBERT, *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, entre 1751 et 1774, article SENS. Nous conservons l'orthographe originale.

[9] Voir BERGSON, *L'évolution créatrice*, 1998 1e ed 1941, p. 311, note de bas de page.

de notre difficulté à distinguer ce qui est dans notre esprit de ce qui est dans la Nature. Selon nous, cette difficulté vient sans doute de ce que l'analogie s'applique indifféremment à toute collection de symboles. Nous y reviendrons plus bas.

Caractère créateur de l'analogie

Passons au deuxième caractère de l'analogie, son pouvoir créateur. Nous verrons (p. 52) que, de façon typique, l'analogie, non mentionnée explicitement, sous-tendait chez les grammairiens du XVII^{e} et XVIII^{e} siècles les discussions sur les formes correctes et leur création : par exemple, doit-on dire *sers-je* ou *servé-je* ? Puisque notre sujet principal est l'analogie entre chaînes de symboles, nous nous pencherons à loisir sur le caractère créateur de l'analogie en morphologie ou en syntaxe (voir plus bas, p. 59, 60, 62 ou p. 68 à 76, et du point de vue formel, p. 167 et 183).

Aussi d'un point de vue plus large, donnons un exemple de création d'idée nouvelle par analogie pris chez un représentant de l'Ecole française de philosophie, Bergson (1859–1951). Cet exemple absolument remarquable d'utilisation de l'analogie à des fins conceptuelles se trouve dans une chute de paragraphe du *Rire*. Il s'agit pour ce philosophe au style particulièrement élégant de définir le comique, c'est-à-dire de proposer un terme approprié à la caractérisation du comique. Cette caractérisation est littéralement créée par le cadre analogique :

> Si donc on voulait définir ici le comique en le rapprochant de son contraire, il faudrait l'opposer à la grâce plus encore qu'à la beauté. Il est plutôt raideur que laideur.[10]

Passons sur le rythme[11]. On assiste ici à la construction d'une équation analogique et à sa résolution au fur et à mesure du déroulement des propositions. Analysons les procédures mises en œuvre par Bergson dans son raisonnement, car elles nous permettent d'introduire dès à présent certains des mots-clés de notre travail sur l'analogie. Nous faisons figurer ces mots-clés en gras.

On pose d'abord que l'inconnue est le comique, dont on cherche un **synonyme** (*si donc on voulait définir ici le comique*) :

$$x = comique$$

Par **contiguïté**, tout à fait habituelle, de la chose à son contraire[12], un premier **rapport** est naturellement recherché (*en le rapprochant de son contraire*) :

[10] Bergson, *Le rire*, 1999 1e ed 1940, p. 22.

[11] La cadence est de 22 syllabes, puis 16, et 9 pour finir, ces 9 syllabes se décomposant elles-mêmes en 6 et 3. L'effet même de la réduction est une marche harmonieuse vers la fin heureuse : $22 : 16 = 1{,}373 \simeq 16 : 9 = 1{,}777 \simeq 9 : 6 = 1{,}666$. Visait-il le nombre d'or, 1,61803399 ? Déliré-je ?

[12] Rigoureusement, son contradictoire. *Prodigue* est le contraire d'*avare*, *libéral* est son contradictoire. La négation logique est de l'ordre du contradictoire. Voir Ducrot & Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, 1978, figure de la page 281.

$$\neg\ comique : x$$

Ce rapport est caractérisé en nommant le contraire, c'est-à-dire en trouvant l'**antonyme** (*il faudrait l'opposer à la grâce*). Bien sûr, ce terme se justifie par un développement antérieur soigneusement argumenté dans le texte de Bergson.

$$\begin{cases} \neg\ comique = grâce \\ x = comique \end{cases} \quad \Rightarrow \quad \neg\ x : x = grâce : comique$$

Par **similarité** entre la grâce et la beauté (*plus encore qu'à la beauté*), on accole au rapport précédent un même rapport de contraire. Par définition, l'**égalité des rapports** est une analogie. Ici, comme elle comprend une inconnue, le comique, il s'agit d'une **équation analogique** :

$$beauté : \neg\ beauté = grâce : x\ \ (comique)$$

qui peut s'écrire aussi, de façon équivalente (voir plus bas, pages 56, et 114)

$$beauté : grâce = \neg\ beauté : x\ \ (comique)$$

Le contraire de la beauté est évidemment la laideur. Mais le génie de Bergson est de résoudre la devinette par l'introduction simultanée, dans le second membre de l'égalité, de deux termes (*il est plutôt raideur que laideur*) non seulement dans le rapport de sens désiré, mais aussi en rapport de similarité du point de vue de la forme (les deux termes diffèrent par une seule lettre). Ce qui emporte la conviction et provoque l'admiration.

$$beauté : grâce = laideur : x \quad \Rightarrow \quad x = comique = raideur$$

Pour finir cette analyse, et comme le lecteur nous aura sans doute précédé à dire qu'il n'y a rien d'extraordinaire à trouver parfois la beauté « glâçante[13] » parce que :

$$raideur : laideur = grâce : y \quad \Rightarrow \quad y = glâce = beauté$$

reconnaissons lui de l'esprit, puisque selon les mots mêmes de Bergson dans le même ouvrage :

> [...] l'esprit consiste souvent à prolonger l'idée d'un interlocuteur jusqu'au point [...] où il viendrait se faire prendre lui-même, pour ainsi dire, au piège de son discours.[14]

[13] Avec un accent circonflexe ! C'est la solution exacte de l'analogie entre chaînes de symboles produite par l'algorithme décrit page 213. On remplace ici seulement un r par un l.

[14] Bergson, *Le rire*, 1999 1e ed 1940, p. 89.

Caractère aveugle de l'analogie

L'analogie est suspecte parce qu'aveugle. La force de conviction que nous avons mentionnée plus haut, si elle ne s'appuie pas sur une analogie qui existe dans la nature elle-même, devenant pure rhétorique, devient trompeuse. Cet aspect est à l'origine du statut ambigu de l'analogie en logique.

Or, à qui se lance dans la rédaction d'un mémoire sur l'analogie, qui plus est pour un document soutenu dans une matière scientifique, se pose le problème de la scientificité de l'objet. Pour l'analogie, cette question est pertinente. Car c'est précisément, pensons-nous, en raison de sa force étrange que l'analogie a été rejetée par certains philosophes comme les Encyclopédistes, par certains linguistes comme les Générativistes, et surtout par certains épistémologues, comme Bachelard. Illustrons, encore une fois sur le mode humoristique, la force terrible du raisonnement analogique :

> [...] Gaston Bachelard [examinait] jadis en Sorbonne un candidat au certificat d'histoire des sciences. Voici l'anecdote (que je verse au dossier des topiques de la science grammaticale) : « Où se lève le soleil ?
> – Je ne comprends pas...
> – ... Vous êtes sans doute un travailleur de la nuit, mais enfin, raisonnez, et vous trouverez où se situe le lever du soleil !
> – (encore hésitant) le soleil se lève à l'est.
> – (enchaînant) vous voyez, mon ami, comme c'est simple, mais s'il vous plaît, répondez par des phrases entières : dans l'hémisphère septentrional, le soleil se lève à l'est, dans l'hémisphère méridional...
> – ... le soleil se lève à l'ouest !
> (Bachelard, heureux d'avoir une nouvelle fois démontré que l'on pouvait faire dire n'importe quoi au candidat le plus intelligent, ajouta simplement :) une dernière question, mon ami, où le soleil se lève à l'équateur ?[15]

Cette anecdote illustre tout simplement que les trois termes étant posés, nous en venons tout simplement à déduire, par analogie, le quatrième, dût-il nous faire émettre une absurdité.

$$\textit{hémisphère nord} : \textit{est} = \textit{hémisphère sud} : \textit{ouest}$$

Bachelard se fait un malin plaisir de la mettre en lumière en proposant un terme brisant les relations d'opposition, parce que le terme correspondant, dans son énormité, nous retient même de l'énoncer : à l'équateur, le soleil se lèverait au zénith !

Nous verrons que ni Aristote, ni les Encyclopédistes ne reconnaissent l'analogie comme un outil de raisonnement logique valable. Ce jugement n'est d'ailleurs pas justifié chez eux. Citons ainsi ce coup de patte donné à l'astrologie dans l'*Encyclopédie* en raison même de la fausseté du raisonnement analogique.

[15] ZEMB, *Vergleichende Grammatik Französisch-Deutsch – Comparaison de deux systèmes – Teil 2*, 1984, p. 95.

> Par exemple, il y a dans le ciel une constellation qu'on appelle lion ; l'analogie qu'il y a entre ce mot & le nom de l'animal, qu'on nomme aussi lion, a donné lieu à quelques Astrologues de s'imaginer que les enfans qui naissoient sous cette constellation étoient d'humeur martiale : c'est une erreur.[16]

Mais, ô paradoxe!, l'utilité de l'analogie comme fondement de raisonnement n'est déniée ni par Aristote, qui s'en sert pour raisonner sur le juste et l'injuste dans l'Éthique à Nicomaque, ni par les Encyclopédistes. Voici l'alinéa, déjà en partie cité plus haut, qui suit immédiatement le passage précédent :

> On fait en Physique des raisonnemens très-solides par analogie. Ce sont ceux qui sont fondés sur l'uniformité connue, qu'on observe dans les opérations de la nature ; & c'est par cette analogie que l'on détruit les erreurs populaires sur le phénix, le rémora, la pierre philosophale & autres.[17]

Bien étrange chose donc que l'analogie, qui fait la force du raisonnement, mais n'est tout de même pas logiquement valable. Que faire donc sinon mettre en garde contre elle, tout en soulignant qu'elle possède une force indéniable pour soutenir l'argumentation ? À moins comme Bachelard (1884–1962), dans *La Formation de l'esprit scientifique*, de la considérer définitivement comme un obstacle à la connaissance générale :

> Il s'agit des généralisations effectuées sur la base d'analogies non fondées entre des phénomènes qui ne sont aucunement apparentés. Une conceptualisation adroite des ressemblances indûment constatées vient alors conférer une pseudo-scientificité au pseudo-constat [...] Là encore Bachelard « psychanalyse » les analogies. À propos de comparaison entre la digestion et la fermentation, il souligne l'importance attribuée à l'époque par les restes du dernier repas[18] : « Ces restes font office d'un véritable levain, jouant le même rôle, d'une digestion à une autre, que la réserve de pâte, gardée par la ménagère au coin du pétrin pour porter d'une cuisson à une autre, les vertus de la panification. » [...] il voit surtout dans les généralisations absurdes qui viennent d'être évoquées la marque néfaste de la pseudo-méthode inductive de Francis Bacon (1561–1626) – cette conception de la science[19] « (...) qui prétend qu'il faut d'abord établir des faits et les collationner avant de pouvoir dégager une loi. »[20]

---

[16] Diderot & d'Alembert, *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, entre 1751 et 1774, article sens. Nous conservons l'orthographe originale.

[17] Ibidem. En graphie originale.

[18] Bachelard, *La formation de l'esprit scientifique*, 1996 1e ed 1938, p. 54

[19] Bartholy *et al.*, *La science : epistémologie générale*, 1978, p. 83.

[20] Acot, *L'histoire des sciences*, 1999, p. 96.

Figure de style et structure de pensée

Nonobstant les problèmes de vérité logique, les caractères universel, créateur et aveugle de l'analogie y fusionnent pour en faire une figure de style douée d'une force argumentative certaine. Nous n'étudierons pas l'analogie en tant que figure de style. Pour cela, contentons nous de renvoyer aux travaux sur les analogies cognitivement bien formées[21]. Mais afin de convaincre le lecteur de la place non négligeable de l'analogie en tant que figure de style nous allons ici quelque peu illustrer cet aspect.

La figure de style a été particulièrement prisée des Classiques et l'on peut la voir chez le parangon des fins esprits du XVII^{e} siècle, Blaise Pascal (1623–1662). Sa double identité de mathématicien et de philosophe fait précisément, selon nous, que ses *Pensées* pullulent de constructions analogiques.

> Dieu veut plus disposer la volonté que l'esprit, la clarté parfaite servirait à l'esprit et nuirait à la volonté.[22]

L'analogie présente ici

$$servir : esprit = nuire : volonté$$

force et renforce l'opposition entre *esprit* et *volonté*, qui est l'une des distinctions fondamentales chez Pascal entre le domaine du naturel, dont ressort l'esprit, et le domaine du surnaturel avec la volonté.

$$servir : nuire\ (= \neg\ servir) = esprit : volonté\ (= \neg\ esprit)$$

Pour prendre un autre exemple fameux, la très célèbre phrase :

> Le silence éternel de ces espaces infinis m'effraie.[23]

est elle-même fondée sur une analogie $silence : espaces = éternel : infinis$ qui joue sur les axes espace-temps en s'appuyant sur les classes grammaticales $nom_1 : adj_1 = nom_2 : adj_2$. En creusant un peu, on peut proposer une interprétation, peut-être un peu hardie, de cette analogie par une décomposition de chacun des termes en notions élémentaires complémentaires.

$$rien\text{-}temps : rien\text{-}espace = tout\text{-}temps : tout\text{-}espace$$

Cette analogie très particulière, que l'on peut représenter par $00 : 01 = 10 : 11$, vient d'être aperçue chez Bergson. (p. 25). Nous la retrouverons dans nos développements ultérieurs (p. 64 et 142). La réponse que Paul Valéry (1871–1945) adresse à Pascal par delà les siècles joue elle aussi sur une série analogique. La voici citée, légèrement modifiée, par Milner :

[21] Douay, *La contre-analogie – Réflexion sur la récusation de certaines analogies pourtant bien formées cognitivement*, 1985.

[22] Pascal, *Œuvres complètes*, 1963, p. 531, 236–*578*.

[23] Ibidem, p. 528, 201–*206*.

> Le bavardage intermittent de nos petites sociétés me rassurent.[24]

On a ici les analogies $\mathit{silence} : \mathit{bavardage} = \mathit{éternel} : \mathit{intermittent} = \mathit{espaces} : \mathit{sociétés} = \mathit{infinis} : \mathit{petites} = \mathit{effrayer} : \mathit{rassurer}$, qui utilisent elles aussi des rapports par contiguïté d'une chose à son contraire. Et de tels exemples sont nombreux chez Pascal lui-même.

> Il y a assez de clarté pour éclairer les élus et assez de d'obscurité pour les humilier. Il y a assez d'obscurité pour aveugler les réprouvés et assez de clarté pour les condamner et les rendre inexcusables.[25]

Cet extrait peut s'interpréter au moyen des analogies suivantes :

$$\begin{gathered} \mathit{clarté} : \mathit{obscurité} = \mathit{éclairer} : \mathit{humilier} \\ \mathit{obscurité} : \mathit{clarté} = \mathit{aveugler} : \mathit{condamner} \\ \mathit{clarté} : \mathit{obscurité} = \mathit{éclairer} : \mathit{aveugler} \quad = \quad \mathit{élus} : \mathit{réprouvés} \end{gathered}$$

qui reposent sur la série de contraires suivante :

$$\begin{aligned} \mathit{clarté} &= \neg\ \mathit{obscurité} \\ \mathit{éclairer} &= \neg\ \mathit{aveugler} \\ \mathit{élus} &= \neg\ \mathit{réprouvés} \end{aligned}$$

Mais le rapport de contiguïté par antonymie n'est pas le seul possible. La contiguïté peut se faire aussi par renversement ou inversion.

> La connaissance de Dieu sans celle de sa misère fait l'orgueil.
> La connaissance de sa misère sans celle de Dieu fait le désespoir.
> La connaissance de J.-C. fait le milieu parce que nous y trouvons, et Dieu, et notre misère.[26]

On observe ici d'une part, l'inversion *Dieu \ notre misère* en *notre misère \ Dieu*, et d'autre part l'inversion d'implication *connaissance de Dieu ou de notre misère seulement* $\Rightarrow$ *orgueil ou désespoir* en *connaissance de J.-C.* $\Rightarrow$ *Dieu et notre misère*, avec le renversement logique du *ou* en *et*. On peut donc, en résumé, voir l'analogie :

$$\begin{gathered} \mathit{Dieu} \setminus \mathit{notre\ misère} : \mathit{orgueil} = \mathit{notre\ misère} \setminus \mathit{Dieu} : \mathit{désespoir} \quad = \\ \mathit{Dieu\ et\ notre\ misère} : \mathit{J.\text{-}C.} \end{gathered}$$

Évidemment, l'analogie en tant que figure de rhétorique n'est pas seulement propre aux Classiques, et nous pensons qu'elle joue un rôle important dans la structuration de la pensée de maints auteurs. Ou, à tout le moins, un rôle pédagogique important dans la présentation de leur pensée car en structurant l'énoncé, elle structure aussi l'objet de la connaissance. Ainsi, par exemple, chez Deleuze (1925–1987), on trouve quelque part, en guise d'explication du passage de Leibniz (1646–1716) à Kant (1724–1804), le raccourci frappant et particulièrement éclairant suivant :

[24] MILNER, *Introduction à une science du langage*, 1989, p. 286.
[25] Ibidem, p. 531, 236–*578*.
[26] Ibidem, p. 525, 192–*527*.

l'idée synthétique du moi fini remplace l'idée analytique infinie de Dieu. [...] Leibniz, c'est l'analyse infinie ; Kant c'est la grande synthèse de la finitude.[27]

$$\begin{array}{rl} \textit{idée synthétique} : \textit{idée analytique} = & \textit{moi} : \textit{Dieu} = \textit{fini} : \textit{infini} \\ \textit{Kant} : \textit{Leibniz} = \textit{analyse} : \textit{synthèse} & = \textit{infinie} : \textit{de la finitude} \end{array}$$

On pourrait donc multiplier les exemples, et dans bien d'autres cultures que la française. Ainsi, de façon tout à fait exemplaire, dans *La structure de l'*îki[28], le schéma explicatif de la figure 1 (p. 31) éclaire la position de l'objet d'étude, l'*iki*, 意気, éponyme du livre, en le plaçant dans un faisceau d'analogies qui font de ce schéma une représentation structuraliste avant la lettre[29].

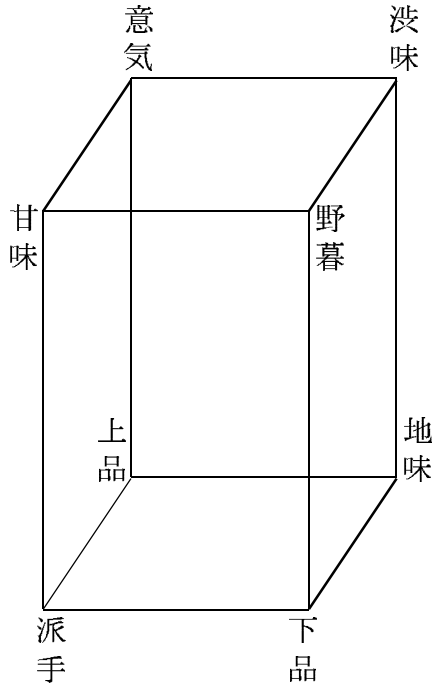


Figure 1: La structure de l'*iki*.

意気 : 渋味 = 甘味 : 野暮
意気 : 甘味 = 上品 : 派手
意気 : 渋味 = 上品 : 地味

[27] Deleuze, *Sur Leibniz*, 1980.

[28] 九鬼 周造 (Kuki Syuuzou), 「いき」の構造, 1999 1e ed 1930, p. 44.

[29] 意気 /iki/ (le chic, l'élégant, le raffiné, essence du dandysme), 渋味 /sibumi/ (âpreté, raffinement discret), 甘味 /amami/ (le sucré, saveur douce), 野暮 /yabo/ (grossièreté, inélégance), 上品 /zyouhin/ (élégance, raffinement), 地味 /zimi/ (sobriété, discrétion), 派手 /hade/ (caractère voyant, tapageur), 下品 /gehin/ (vulgarité, trivialité, grossièreté).

Ce schéma nous intéresse particulièrement, parce que nous y voyons une préfiguration de ce que nous proposerons pour la formalisation de la métaphore par correspondance entre espaces analogiques. Nous serons en effet amené à dessiner des schémas semblables, où deux analogies, par exemple ici le rectangle du haut et celui du bas, seront mises en correspondance par leurs éléments, ici selon les quatre traits verticaux. Tous les cas précédents, dans lesquels la structuration de la pensée est reflétée par une analogie dans l'énoncé, participent selon nous de notre schéma général de correspondance entre plusieurs domaines structurés par l'analogie (voir p. 202 et suivantes). Cette configuration rend compte d'une espèce restreinte de la métaphore, dont nous donnerons une formalisation générale afin d'en proposer une application particulière au traitement automatique des langues.

Présentation du plan

Notre étude de l'analogie entre chaînes de symboles s'articulera en quatre grandes parties.

D'abord, une **histoire** de la notion d'analogie en grammaire et en linguistique nous permettra de dégager les définitions et les notions inhérentes à l'analogie, telles que la similarité et la contiguïté. En même temps, cette histoire nous permettra de situer l'analogie par rapport à d'autres notions comme celles de métaphore et de métonymie.

Ensuite une partie de **formalisation** nous permettra de proposer une expression mathématique (partielle) de l'analogie entre chaînes de symboles. À partir de cette formalisation, nous montrerons comment on peut définir des langages formels particuliers, appelés langages de chaînes analogiques. La mise en correspondance de tels langages nous conduira à proposer une formalisation d'un type restreint de métaphore, grâce à la notion d'homomorphisme.

Une partie consacrée aux **algorithmes** reprendra les résultats de la partie de formalisation pour les utiliser dans des algorithmes destinés à être utilisés en traitement automatique des langues.

Enfin une partie d'**expérimentations** montrera l'utilisation de ces algorithmes ou de formes anciennes de ces mêmes algorithmes dans un certain nombre d'expériences qui se sont étalées sur une période de plus de cinq ans.

Dans chacune des parties de formalisation, d'algorithmes et d'expérimentations, nous adopterons une progression semblable. Trois sous-parties parleront successivement de l'analogie proprement dite, des langages définis à partir de l'analogie, et enfin de notre modèle d'homomorphismes entre espaces analogiques.

Partie I

Histoire de l'analogie

Chapitre 1

Historique de la notion

Nous essayons ici de faire un historique de la notion d'analogie. Ce n'est pas très simple. En effet, si elle est ancienne et bien définie, elle a connu quelques vicissitudes.

Première difficulté : si une caissière passe son temps à faire des additions, le mot d'addition n'apparaît pas nécessairement fréquemment dans son discours. De même, alors que l'usage de l'analogie dans la discipline qui nous intéresse, la linguistique, a été constant au cours des âges, sa mention ne l'est pas. On constate ainsi dans l'histoire de l'analogie un mouvement de balancier entre sa mention explicite et son usage implicite. Inversement, contrastant avec le fait que nombre de linguistes font usage de l'analogie sans la mentionner, le courant dominant de la fin du XX^{e} siècle, le générativisme, en a fait explicitement mention uniquement pour la rejeter avec force dans la géhenne du non-scientifique.

Deuxième difficulté: alors que la définition de l'analogie est clairement posée depuis l'Antiquité, la notion a été détournée de sa vraie signification par l'usage courant, et nombre de travaux actuels, essentiellement anglo-saxons, font un usage erroné du terme.

Troisième difficulté: au cours de l'histoire de la linguistique, l'analogie s'est vu assigner tantôt un rôle de fauteur de troubles, tantôt, au contraire, de facteur d'ordre. Elle a été impliquée dans des débats ou des révolutions scientifiques avec une étiquette tantôt positive tantôt négative. Le mot même d'analogie est donc chargé d'affectivité.

Nous allons tout de même essayer de présenter les avatars par lesquels est passée l'analogie à travers les âges. Puisqu'il s'agit d'histoire, et plus particulièrement d'une histoire de l'analogie en Occident, commençons avec l'Antiquité grecque.

1.1 Les Anciens : de la géométrie à la théologie

1.1.1 Les philosophes grecs : des définitions claires

Rappelons qu'Euclide (III^{e} s. av. J.-C.) est ce mathématicien qui fonda l'École d'Alexandrie, célèbre pour son approche axiomatisée de la géométrie plane, mais dont on n'a pas conservé d'écrits exacts. Son œuvre nous est connue à partir de commentaires seulement. Elle est divisée en quinze « éléments », dont le cinquième, tel qu'on peut le lire par exemple à travers les commentaires d'Henrion[1], constitue un traité des raisons et des proportions, principalement entre longueurs.

Citons les troisième et quatrième définitions de cette partie, en amputant certains commentaires d'Henrion.

> 3. Raison, est vne habitude de deux grandeurs de mesme genre, comparee l'vne à l'autre selon la quantité.
> *C'est à dire que quand deux quantitez de mesme genre, comme deux nombres, deux lignes, deux superficies, deux solides, &c. sont comparez entr'eux selon la quantité, c'est à dire selon que l'vne est plus grande que l'autre, ou moindre, ou egale, telle comparaison est appellée raison, & par quelqu'uns proportion : Parquoy on ne peut pas dire, qu'il y ait quelque raison d'vne ligne à vne superficie ; ou d'un nombre à vne ligne, puis que ny la ligne & la superficie, ny le nombre & la ligne, ne sont pas quantitez de mesme genre. Semblablement si on confere vne ligne avec vne ligne selon la qualité, c'est à dire, selon que l'vne est blanche, & l'autre noire ; ou bien que l'vne est chaude, & l'autre est froide, &c. encore que l'vne & l'autre soient de mesme genre, cette comparaison n'est pas dicte raison, pource qu'elle n'est pas faicte selon la quantité.*
>
> *Or iaçoit que la raison se trouue proprement és seules quantitez, si est-ce toutesfois que toutes autres choses, qui en quelque maniere prennent la nature de la quantité, comme sont les temps, les sons, les voix, les lieux, les monumens, les pois, & les puissances, sont aussi dictes avoir raison, si leur habitude est consideree selon la quantité, comme quand nous disons vn temps estre plus grand qu'un autre temps, ou moindre ; ou deux temps estre egaux, &c. telle habitude sera dite raison, pource qu'alors les temps sont considerez ainsi que certaines quantitez.*
>
> *En toute raison ceste quantité-là, qui est referee à vne autre, est dicte par Euclide, & d'autres Geometres, antecedant de la raison ; & celle-là à laquelle elle est referee, est dicte consequent d'icelle raison : Comme en la raison de A à B ; A est dict antecedant de la raison, & B consequent : Que si au contraire B est comparé à A ; B sera appellé antecedant, &*

[1] EUCLIDE, *Les quinze livres des éléments géométriques d'Euclide : plus le livre des donnez... trad. en françois*, 1632, Élément cinquiesme, p. 166-209. C'est bien aussi la cinquième partie dans la traduction japonaise à notre disposition, EUCLIDE, ユークリッド原論, 1996, traduction basée sur une édition grecque et latine parue en Allemagne, EUCLIDE, *Euclidis Elementa*, 1883. Nous n'avons pas eu cette dernière édition entre nos mains.

A consequent. [...]

4. Proportion, est vne similitude de raisons.

Tout ainsi que la comparaison de deux quantitez entr'elles est dicte raison, ainsi la comparaison & ressemblance de deux ou plusieurs raisons entr'elles, est dicte proportion : comme si la raison de A à B, est semblable à la raison de C à D, l'habitude d'entre ces raisons sera dite proportion. Et c'est ce que les Grecs appellent analogie, *& quelques Latins* proportionalité *: selon Boëtius & Iordanus il y en a de plusieurs sortes, dont les principales qu'ils appellent* Medietez, *sont la proportion Arithmetique, la Geometrique et l'Harmonique : Mais Euclide ne traite icy que de la Geometrique, laquelle est ou continue ou discrete : la proportion continuë, est celle de laquelle les grandeurs entre moyennes sont prises deux fois, tellement qu'il ne se faict nulle interruption de raisons, ains chaque quantité entre moyenne est antecedant & consequent, sçavoir antecedant de la quantité subsequente, mais consequent à la quantité antecedante : comme si on dit que telle qu'est la raison de A à B, telle est celle de B à C, où la quantité B est antecedant de la quantité C, & consequent de la quantité A. Mais la proportion discrette ou non continuë, est celle en laquelle chaque quantité entremoyenne est prinse seulement une fois, tellement qu'il se fait interruption de raisons, & aucune quantité n'est antecedant & consequent : mais seulement antecedant ou consequent : comme quand on dit que la raison A à B, est comme celle de C à D.* [2]

Qu'ajouter de plus ? Les concepts sont extrêmement clairs. Ils posent qu'une raison est une mesure quantifiée et que la proportion, aussi appelée analogie, est une similitude de raisons, ce qu'en termes modernes nous reformulerons par égalité de rapports.

Nous reviendrons bientôt (voir p. 42) sur la digression à propos de l'opposition entre proportions discrète et continue. Sans aucun doute le commentaire d'Henrion est à attribuer à la lecture d'Aristote.

À une époque où le savoir entier est transmis par les philosophes, il n'est pas étonnant de voir la notion géométrique d'analogie passer dans l'argumentation du plus célèbre d'entre eux, Platon (428-348 av. J.-C.). Dans une liste d'énumération des modes d'enseignement qu'il emploie, entre l'inspiration divine, l'exemple, l'induction et la géométrie, Olympiodore (VI^{e} s.) mentionne l'analogie géométrique :

> Il [Platon] use de l'analogie dans le *Gorgias* où il soutient que ce que la médecine est à la cuisine, la justice l'est à la sophistique (= rhétorique). [...] il démontre ce qu'il veut dire au moyen d'une analogie géométrique '*A* est à *B* comme...' [3]

L'analogie ou proportion entre figures géométriques est donc ici employée sur un autre terrain, celui du raisonnement, comme un moyen de persuasion.

[2] Ibidem, p. 167-171. Nous respectons la graphie originale. Ces définitions sont numérotées 2 et 6 dans la version japonaise, EUCLIDE, ユークリッド原論, 1996, 第5巻, p. 93.

[3] OLYMPIODORE, *Prolégomènes à la philosophie de Platon*, 1990, XI, § 27, 10, p. 43

Soulignons qu'il s'agit plutôt ici d'un emploi comme figure de style, comme nous en avions montré des exemples dans notre introduction, car la comparaison entre la justice, la sophistique, la médecine et la cuisine est affaire de notions qui se mesurent difficilement, et la notion de rapport est donc assez peu rigoureuse ici. Ainsi, suivant le commentaire d'Henrion, on n'a pas vraiment de raison ou rapport ici, mais de simples comparaisons, car on ne saurait voir de quantités mesurables ici.

Il n'en va pas de même de l'usage que fait Aristote (384-322 av. J.-C.). Le grand homme parle en effet de l'analogie de deux façons différentes dans deux de ses écrits. D'une part, dans la Poétique, à la rubrique où il est traité de la métaphore, il la mentionne comme un cas particulier de métaphore :

> La métaphore est le transport à une chose d'un nom qui en désigne une autre, transport ou du genre à l'espèce, ou de l'espèce au genre, ou de l'espèce à l'espèce, ou d'après le rapport d'analogie.[4]

Rappelons, sans faire aucun jeu de mots, que dans les Catégories d'Aristote, le genre est plus général que l'espèce, qui est elle, plus spécifique. Étendons-nous sur la notion de métaphore, car la méconnaissance de ces textes explique la confusion des termes qui règne en intelligence artificielle (voir plus bas, p. 83 et 96). Pour Aristote, donc, il en existe quatre types. Le premier type de métaphore est l'utilisation de la partie pour le tout [5], comme dans « une voile » pour un bateau. Le second est l'utilisation du tout pour la partie[6], comme dans « de l'herbe » pour du haschisch. Le troisième type consiste à dire « sucrer les fraises » pour trembler des mains. Enfin, seul le quatrième type fait appel à l'analogie, et prendre la métaphore pour l'analogie est donc une erreur... par utilisation abusive de la partie pour le tout !

Venons-en alors à la définition de l'analogie, accompagnée d'exemples :

> J'entends par « rapport d'analogie » tous les cas où le second terme est au premier comme le quatrième au troisième, car le poète emploiera le quatrième au lieu du second, ou le second au lieu du quatrième ; et quelquefois aussi, on ajoute le terme auquel se rapporte le mot remplacé par la métaphore. Pour m'expliquer par des exemples, il y a le même rapport entre la coupe et Dionysos qu'entre le bouclier et Arès ; le poète dira donc de la coupe qu'elle est « le bouclier de Dionysos » et du bouclier qu'il est « la coupe d'Arès ». De même : il y a le même rapport entre la vieillesse et la vie qu'entre le soir et le jour ; le poète dira donc du

[4] ARISTOTE, *Poétique*, 1996, p. 119.

[5] Souvent appelée synecdoque. Mais, à mon grand étonnement, les dictionnaires ne sont pas d'accord entre eux sur les définitions ! Les uns opposent métonymie (remplacement d'un terme par un autre ayant des traits contigus ou un rapport logique) à métaphore (remplacement d'un terme par un autre qui lui est similaire ou partage des traits de sens). Les autres opposent métonymie (le tout pour la partie) à synecdoque (la partie pour le tout). D'autres encore comprennent synecdoque aussi bien comme prendre la partie pour le tout que comme prendre le tout pour la partie.

[6] Souvent appelée métonymie. Voir note précédente.

> soir, avec Empédocle, que c'est « la vieillesse du jour », de la vieillesse que c'est « le soir de la vie » ou « le couchant de la vie ».[7]

Dans l'Éthique à Nicomaque, Aristote répète la définition de l'analogie, avant de s'en servir comme outil de raisonnement. Dans ce raisonnement, il s'agit réellement de mesurer des quantités, puisque le problème est de définir le juste, nous dirions la justice, dans la rétribution de deux personnes. Notons qu'Aristote insiste lourdement sur le fait que quatre termes interviennent dans une analogie.

> [...] la proportion étant une égalité des rapports et supposant quatre termes au moins. – Que la proportion discontinue implique quatre termes cela est évident, mais il en est de même aussi pour la proportion continue, puisqu'elle emploie un seul terme comme s'il y en avait deux et qu'elle le mentionne deux fois ; par exemple, ce que la ligne A est à la ligne B, la ligne B l'est à la ligne C, la ligne B est donc mentionnée deux fois, de sorte que si l'on pose B deux fois, il y aura quatre termes proportionnels. – Et le juste, donc, implique quatre termes au moins et le rapport [entre la première paire de termes] est le même [que celui qui existe entre la seconde paire], car la division s'effectue d'une manière semblable entre les personnes et les choses.[8]

Par lignes, il faut entendre des segments de droite dont la mesure est la longueur. C'est un simple rappel d'Euclide. Selon la note en bas de page donnée par le traducteur de l'édition citée, la raison pour laquelle Aristote insiste sur le fait que le cas auquel il s'intéresse est une proportion discontinue proviendrait du fait que, le raisonnement n'étant pas un syllogisme (A est B, B est C, donc A est C), il n'y a pas ici de partie commune (B) qui serve de moyen. Dans le cas particulier auquel Aristote s'intéresse, qui est en fait le cas général de l'analogie, le moyen est différent d'un membre à l'autre[9]. Il s'agit de la distinction faite par Henrion entre proportion continue et proportion discrète (voir plus haut, p. 39) qui constitue d'ailleurs chez Euclide la définition 9 :

> 9. Proportion ne peut estre constituée sur moins de trois termes.
> *Puisqu'il a esté dit en la 3. def. que raison est l'habitude de deux quantitez, & que proportion par la 4. def. est une similitude de deux ou*

[7] ARISTOTE, *Poétique*, 1996, p. 120.

[8] ARISTOTE, *Éthique à Nicomaque*, 1997 1er tirage 1990, V, 6, p. 228–229. Nous avons remplacé Γ par C.

[9] En notes de bas de page, le traducteur J. Tricot pose explicitement les égalités :

$$\frac{A}{B} = \frac{C}{D},\ \frac{A}{C} = \frac{B}{D} \text{ et } \frac{A+C}{B+D} = \frac{A}{B}$$

où nous avons remplacé Γ par C et Δ par D. La troisième égalité peut sembler étrange, mais elle exprime le fait que, selon Aristote, la rétribution de A par la quantité C et de B par la quantité D est juste seulement si elle laisse le rapport inchangé.

> *plusieurs raisons: il s'ensuit qu'il ne peut y auoir moins de trois quantitez ou termes en vne proportion, si elle est proportion continue, mais il en faut quatre au moins, si elle est proportion discrette.* [10]

En résumé de tout ce qui précède, l'analogie ne saurait être confondue avec la métaphore, chose évidente pour un Grec, étant donné le sens des éléments suivants: $\mu\varepsilon\tau\acute{\alpha}-$, /meta-/ (entre, parmi, avec ; aussi, après) et $\varphi\varepsilon\rho\varepsilon\tilde{\imath}\nu$, /pherein/ (porter) ; métaphore, composé des éléments précédents, signifie transfert de quelque chose à autre chose. Par contre, $\alpha\nu\alpha-$ /ana-/ (de nouveau, à nouveau ; aussi, de bas en haut) et $\lambda o\gamma\acute{o}\varsigma$ /logos/, $-\lambda o\gamma\acute{\imath}\alpha$ /-logia/ (rapport, proportion, relation ; par extension, discours, raison) forment analogie, qui signifie donc même relation, rapport égal (voir plus bas p. 46 pour sa traduction latine).

Pour ce qui est de l'application de l'analogie, même chez Aristote, l'analogie n'est pas un outil du raisonnement logique au même titre que le *modus ponens* ou le *modus tollens*. En effet, on n'est pas assuré de déduire toujours des propositions vraies par analogie. Mais, indéniablement l'analogie est un moyen d'emporter la conviction. Cette idée sera reprise par les Encyclopédistes, d'ailleurs sans être justifiée.

> L'analogie est aussi un des motifs de nos raisonnemens ; je veux dire qu'elle nous donne souvent lieu de faire certains raisonnemens, qui d'ailleurs ne prouvent rien, s'ils ne sont fondés que sur l'analogie. [...] Les raisonnemens par analogie peuvent servir à expliquer & à éclaircir certaines choses, mais non pas à les démontrer.[11]

Évidemment, ce défaut de l'analogie constitue un argument à son encontre, argument qui, transposé en syntaxe, sera utilisé par les génerativistes (voir plus bas, p. 76).

Pour donner une synthèse des notions dégagées, la lecture d'Aristote nous a donc enseigné quelques définitions fondamentales :

1. l'analogie et la métaphore ne sont pas la même chose ;
2. en revanche, l'analogie et la proportion sont la même chose ;
3. quatre termes interviennent nécessairement dans une analogie ; nous nous efforcerons de toujours les noter A, B, C et D par la suite ;
4. l'analogie est l'égalité de deux rapports ; si on note par $A : B$ le rapport de deux termes, une analogie s'écrit donc : $A : B = C : D$.

[10] EUCLIDE, *Les quinze livres des éléments géométriques d'Euclide : plus le livre des donnez... trad. en françois*, 1632, Élément cinquiesme, p. 174. Cela correspond à la définition 8 dans EUCLIDE, ユークリッド原論, 1996, 第 5 巻, p.93.

[11] DIDEROT & d'ALEMBERT, *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, entre 1751 et 1774, article ANALOGIE. Nous avons déjà donné p. 27 le passage que nous censurons ici entre crochets. Attention : nous respectons l'orthographe originale.

1.1.2 Les grammairiens latins : opposition avec l'anomalie

Le passage de l'analogie à la grammaire sera faite par les Grecs eux-mêmes. Leur tradition est reprise par les grammairiens latins, et Varron (116-27 av. J.-C.), illustre encyclopédiste, fait de l'analogie le pôle d'une opposition avec l'anomalie. Le *De lingua latine* fixe les définitions de cette opposition.

> Dans la conjugaison des verbes, beaucoup ont demandé si ce que l'on doit dire devait suivre la dissemblance ou la ressemblance. Relativement à cela, le rapport qui a pour origine la ressemblance sera appelé analogie, le reste étant appelé anomalie : [...]
> Je parlerai de quatre choses qui concernent la conjugaison des verbes : de ce qu'est le semblable et le dissemblable, de la raison que l'on appelle « logon », de la proportion que l'on appelle « analogon » et de leur manière d'être, et une fois ces choses expliquées, on définira l'analogie et l'anomalie, où on les trouve, ce qu'elles sont, et de quelle manière elles sont.[12]

> [...] de *Roma* on tire *Romanus*, de même de *Capua* on tire *Capuanus*, ce qui naît de la force de l'habitude, mais, comme je l'ai dit, la conjugaison des verbes pèche dans l'usage commun, ce qui vient de la multiplicité et de l'ignorance des peuples. C'est pourquoi de cette manière, on obtient plus d'anomalie que d'analogie dans la conversation.

Selon Mounin, « [Varron] fait des observations (à la vérité guidées par le désir de comprendre et défendre l'*analogia*) qui attendront deux mille ans pour être bien analysées sous le nom d'*aspects* du verbe. »

> L'analogie, dit-on, n'est pas observée dans les temps de certains verbes, comme *legi* (j'ai lu), *lego* (je lis) et *legam* (je lirai), dont le premier appartient au parfait [*perfectum*] et les deux autres à l'imparfait [*inchoatum*]. Pour faire tomber ce grief, il suffit de rétablir l'ordre dans la classification des temps de ce verbe, qui présentera alors une division tout à fait conforme à l'analogie, comme *discebam*, *disco*, *discam*[13] pour les temps imparfaits, et *didiceram*, *didico*, *didicero*[14] pour les temps parfaits. On voit que ce ne sont pas les verbes qui pèchent contre l'analogie, et que s'il y a anomalie, c'est du fait de ceux qui confondent sciemment les trois temps.[15]

> On confond encore les temps parfaits [*perfecti*] avec les temps imparfaits [*infecti*], en mettant sur la même ligne *fui*, *sum*, *ero*[16]. Le parfait *fui* est conforme dans toute sa conjugaison, et dans la parenté qui

[12] Varro, *De lingua latina*, 1954, livre X, I et II.

[13] *Discebam* (j'apprendrais), *disco*, (j'apprends, j'étudie, je suis informé de), *discam* (j'apprendrai).

[14] *Didiceram* (j'avais appris, j'avais étudié), *didico* (j'ai appris, j'ai étudié), *didicero* (j'aurai appris).

[15] Cité par Mounin, *Histoire de la linguistique – Des origines au XX^e siècle*, 1967, p. 100.

[16] *Fui* (j'ai été), *sum* (je suis), *ero* (je serai).

> l'unit à *fueram* et *fuero*[17]. Les temps imparfaits [*infecti*] offrent la même régularité: *sum* (autrefois *esum*), *es*, *est*; *eram*, *eras*, *erat*; *ero*, *eris*, *erit*[18]. En classant les temps ainsi dans leur ordre, on retrouve partout l'analogie.[19]

Si l'anomalie est l'absence de régularité[20], l'absence de règle, au sens de Varron, l'analogie apparaît bien comme l'expression de la régularité. Notons tout de suite que ce rôle lui sera retiré en linguistique comparative au début du XIX^{e} siècle, lorsque les lois phonétiques seront considérées comme des lois dont l'application régulière est troublée par l'analogie. On peut donc attribuer à Varron le début de cette oscillation qui ballotera l'analogie du camp de l'ordre à celui du désordre alternativement au cours des siècles. Mais soulignons que pour Varron anomalie et analogie ne peuvent toutefois aller l'une sans l'autre :

> Pour la raison que l'usage donne des mots et des déclinaisons à la fois semblables et différents, on ne saurait rejeter ni l'analogie ni l'anomalie. Ceci pour la même raison que l'homme n'est pas qu'une âme, mais à la fois un corps et une âme.[21]

Évidemment, les professionnels de la parole et les hommes politiques se serviront de la classification proposée par Varron pour étudier la langue latine, l'enrichir et l'illustrer. Jules César lui-même aurait écrit un traité sur l'analogie des mots[22] et il aurait même échangé de la correspondance avec Auguste à ce sujet, selon Quintilien (v. 30–v. 100). Ce dernier, maître de rhétorique, prônait l'analogie, mais fort prudemment, toujours dans les limites imposées par « le maître le plus sûr », l'usage, dont il faut cependant se servir « avec beaucoup de discernement » !

> On découvre aussi quelquefois l'indicatif à l'aide des temps obliques. Je me souviens d'avoir ramené à mon avis des personnes qui me reprenaient de m'être servi du prétérit *pepigi*[23]. Ils convenaient bien que de grands écrivains l'avaient employé; mais ils prétendaient que la règle ne le

[17] *Fueram* (j'avais été), *fuero* (j'aurai été).

[18] *Sum, es, est* (je suis, tu es, il est), *eram, eras, erat* (j'étais, tu étais, il était), *ero, eris, erit* (je serai, tu seras, il sera).

[19] Ibidem, p. 100.

[20] Selon l'étymologie maintenant admise, plus que l'absence de loi ou de règle, sens que semble bien avoir aussi le latin *anomalia*, l'anomalie serait l'absence de conformité, d'égalité ou d'unité. L'adjectif correspondant devrait s'analyser $\alpha\nu + \delta\mu\acute{\alpha}\lambda\varsigma$, de $\delta\mu\acute{\alpha}\lambda\varsigma$ /homálos/ (conforme, égal, plat, uni) sans rapport avec $\delta\ \nu\acute{o}\mu\varsigma$ /ho nómos/ (la règle, la loi).

[21] VARRO, *De lingua latina*, 1954, livre IX, paragraphe I.

[22] D'après DUCKETT, *Dictionnaire de la conversation et de la lecture – Dictionnaire raisonné des notions générales les plus indispensables à tous par une société de savants et de gens de lettres*, 1864, article GRAMMAIRE, signé de CHAMPAGNAC. Nous n'avons hélas pas pu mettre la main sur ce traité.

[23] *Pepigi* (j'ai conclu).

> permettait pas, parce que le présent de l'indicatif ayant la nature de la voix passive[24], devait faire au prétérit *pactus sum*[25] ;
>
> Et moi, outre l'autorité des orateurs, je me fondais encore sur l'analogie. En effet, en lisant dans les XII Tables *ni ita pagunt*[26], j'étais conduit par son analogue *cadunt*[27] à reconnaître que l'indicatif, tombé depuis en désuétude, était *pago*[28] comme *cado*[29], et qu'ainsi il n'était pas douteux qu'en disant *pepigi* je suivais la même règle que pour *cecidi*[30].[31]

Quintilien est aussi l'auteur d'une remarque assez intéressante sur l'analogie du point de vue épistémologique.

> C'est que l'analogie n'est pas descendue du ciel au moment de la formation de l'homme, pour lui apprendre à parler ; mais elle a été découverte après la parole, et après que le langage eut donné lieu à des remarques sur les désinences de certains mots. Ce n'est donc pas sur la raison que se fonde l'analogie, mais sur l'exemple; elle n'est donc pas la loi du langage, mais le résultat de l'observation; de sorte que l'analogie n'a d'autre origine que l'usage.[32]

Il prétend donc que l'analogie n'est qu'une conséquence du système de la langue, et que son origine ne saurait être cherchée dans la raison humaine. Cet avis est l'inverse de celui des Encyclopédistes (voir plus haut, p. 28), ou de celle que l'on peut lire dans le *Grand dictionnaire universel*[33]. Là, en effet,

[24]Il ne peut s'agir ici que du verbe *paciscor, eris, i, pactus sum* (conclure un pacte, un traité, convenir de, s'engager à).

[25]*Pactus sum* (j'ai conclu). En fait, il semble qu'il y ait confusion du verbe *paciscor* avec *pango, is, ere, pepigi, pactum* (enfoncer, ficher ; conclure un marché, convenir de, stipuler). La forme *pactus sum* est bien ambiguë entre les deux verbes, mais le sens devrait être différent, à cause du mode, qui ne peut être le même dans les deux cas. Si c'est *papiscor*, il s'agit du parfait actif, si c'est *pango*, il s'agit du parfait passif.

[26]Il s'agit des fragments 6 et 7 de la première table des douze lois romaines. *6. Rem ubi pacunt, orato. 7. Ni pacunt in comitio aut in foro ante meridiem causam coiciunto.* (Si l'affaire est réglée, que cela soit annoncé. Si le cas n'est pas réglé, alors qu'il soit présenté avant midi à un comice ou au forum.) Voir par exemple, `http://www.tu-berlin.de-/fb1/AGiW/Auditorium/RomRecht/SO3/LXIITab.htm` ou `http://www.filodiritto.com-/diritto/romano/12tavole.htm`. Cependant, le *ita* n'apparaît dans aucune de ces versions. La lecture *pagunt* de Quintilien provient sans doute de ce que le vieux latin ne notait pas la différence entre *C* et *G*. La lettre *G* est une invention tardive. Par exemple, on sait que *Caius* doit se lire *Gaius*. Cela expliquerait peut-être *pacunt* dans les tables.

[27]*Cadunt* (ils tombent).

[28]Pour *pango*? Quintilien reconstitue ici une forme inexistante à l'aide d'une équation analogique. $cadunt : cado = pagunt : x \quad \Rightarrow \quad x = pago.$

[29]*Cado* (je tombe).

[30]*Cecidi* (je suis tombé). Annonçons ici que cette solution à l'équation analogique $cado : cecidi = pago : x \quad \Rightarrow \quad x = pepigi$ est hors de portée de la formalisation que nous proposerons plus bas. En effet, celle-ci ne rend compte ni de la répétition ni du redoublement. Dans notre formalisation, nous n'obtenons que : $cado : cecidi = pago : x \Rightarrow x = pecigi$ ou $cepigi$ (voir p. 162).

[31]Quintilien, *L'institution oratoire*, 2000, chapitre 6, § 10 et 11.

[32]Ibidem, § 16.

[33]Larousse, *Grand dictionnaire universel*, 1865 a 1876, p. 313, article analogie.

on prétend que l'analogie fonde la pensée humaine et permet les découvertes scientifiques. Cette opinion sera contestée par Bachelard (voir p. 28).

À la suite de Varron, Aulu-Gelle (135–165) reprendra simplement la distinction opérée par lui :

> L'analogie est la similarité de déclinaison[34], elle est appelée proportion en latin. L'anomalie est le caractère inégal des déclinaisons conformément à l'usage.[35]

Notons que si Aulu-Gelle utilise le calque le plus fréquent *proportio* pour traduire le mot grec, il existe d'autres traductions en latin du mot grec analogie, par exemple, *aequalitas* (égalité), *comparatio* (comparaison), *similitudo* (ressemblance, similitude)[36].

Chez Donat (IV^{e} siècle), autre grand grammairien latin, le mot même d'analogie est passé sous silence. Mais la définition des parties du discours (*pars orationis* en latin) à la suite des auteurs Grecs (*μέρη τοῦ λόγου* en grec) se fait chez lui essentiellement selon les déclinaisons ou conjugaisons[37]. Le nom est ce qui prend des cas, le verbe est ce qui prend des temps. Si c'est parce que *album, albi, albo, albo* sont à *dominum, domini, domino, domino* ce que *albus* est à *dominus*, que l'on pose que *albus* est un nom comme *dominus*, ne peut-on dire que l'analogie sert ici de critère de définition ? Cette utilisation forte de l'analogie a comme conséquence, comme nous venons de l'illustrer, que tout au long de l'Antiquité, la différence entre substantif et adjectif n'est qu'une différence de qualité.

> Qu'est-ce qu'un nom ? La partie du discours avec un cas, désignant en propre ou en commun un corps ou une chose. Combien de déclinaisons a un nom ? Six. Lesquelles ? La qualité, la comparaison, le genre, le nombre, la figure et le cas. Qu'est-ce que la qualité d'un nom ? Elle est double : soit un nom est dit en propre d'une seule chose, soit il est dit de plusieurs et appellatif.[38]

La définition des parties du discours est donc plus paradigmatique (par sélection de flexions) que syntagmatique (par combinaison de mots). Ce que Salamanca, parlant des grammariens latins, résume par :

> Pour les grammairiens, l'analogie est à l'œuvre dans les formes semblables, dans ce que, à l'époque moderne, selon la dénomination grecque, on a appelé *paradigme*.

[34]Nous avons traduit par un hypallage. Mot-à-mot : « déclinaison semblable. »

[35]Cité par SALAMANCA, *La tradición histórica de la analogía lingüística*, 1984, p. 372 et 373.

[36]SALAMANCA, op. cit., p. 372.

[37]Le latin utilise d'ailleurs le même terme de *declinatio* pour désigner à la fois les conjugaison et les déclinaisons.

[38]DONATIUS, *De partibus orationis – Ars minor*, 1994, De nomine.

Il faudrait, pour être exact, souligner que chez Varron, des critères de distribution (donc syntagmatiques) sont aussi pris en compte pour définir les parties du discours[39].

Après l'époque latine, au Moyen-Âge, l'analogie s'inscrit dans une histoire des idées particulièrement mouvementée. À cette époque, alors que la grammaire traite de la morphologie et de la syntaxe selon les grammairiens latins que nous venons de voir, c'est ce que nous appellerions aujourd'hui la sémantique qui est le théâtre des passions. À cette période, en effet, on n'étudie pas tant la langue pour la langue que pour découvrir ce qui peut se dire à des fins métaphysiques, et plus particulièrement de ce qui peut se dire de l'être. Ce point est fort important, et nous y reviendrons, ainsi que sur la place de l'analogie dans ce contexte, dans un développement ultérieur sur les notions voisines d'univocité et d'équivocité (p. 48).

1.1.3 Les penseurs arabes : usage en droit et en grammaire

Ce que nous venons de dire anticipait. Cela dépend en réalité de ce que nous allons maintenant dire : en effet, puisqu'il s'agit pour nous de faire l'histoire de l'analogie, et qui plus est l'histoire de l'analogie en Occident, nous ne pouvons passer sous silence l'influence de la civilisation qui a porté haut le flambeau de la science durant plus de cinq siècles du VIII^{e} au XIV^{e}, la civilisation arabo-musulmane dont l'apport à notre Moyen-Âge a été si important.

Dans la tradition arabe, il semble que l'analogie trouve son origine dans les pratiques juridiques qui imposent d'obtenir l'unanimité des docteurs, et non pas seulement la majorité, pour l'interprétation de situations nouvelles par rapport au Coran ou aux Hāditha. Cette unanimité est appelée *'iğmāʒ*, de la racine *ğmʒ* (rassembler). Les raisonnements juridiques aboutissant à cette unanimité reposent généralement sur des « raisonnements par analogie », appelés *qiyāṣ*, de la racine *qyṣ* (mesurer, prendre une mesure). Nous retrouvons donc ici la force argumentative de l'analogie que nous avions illustrée dans notre introduction (p. 27 et 29). Il est difficile de savoir si cette pratique trouve sa source dans les écrits d'Aristote, et la question semble controversée[40].

Notons que cet usage de l'analogie dans le droit est reconnu dans toute la tradition juridique européenne. Ainsi l'encyclopédie polonaise WIEM donne la définition suivante :

> Analogie : méthode de raisonnement juridique – établissement des conséquences juridiques d'un fait non réglé par le droit (lacune du droit) sur la base d'une norme établissant les conséquences juridiques de faits semblables sous « des rapports réels » (analogia legis) ou bien sur la base de principes généraux ou selon des « valeurs » attribuées au droit

[39] Voir aussi DUCROT & SCHAEFFER, *Nouveau dictionnaire encyclopédique des sciences du langage*, 1978, p. 442.

[40] G. BOHAS, *The Arabic Linguistic Tradition*, 1990.

> faisant obligation (analogia iuris). Dans nombre de pays, l'analogie n'est pas reconnue dans le droit pénal.[41]

Ajoutons pour compléter la dernière phrase de cette définition que le principe qui s'oppose à l'*analogia iuris* s'appelle *argumentum a contrario*. Ce principe-là affirme que les conséquences juridiques d'un fait ne peuvent s'appliquer qu'aux seuls faits énumérés dans le code, et qu'elles ne sauraient s'appliquer à des faits autres, même semblables.

Du monde juridique, l'analogie a été transposée à la grammaire et elle structure toute la description de la morphologie dérivationnelle et conjugationnelle de l'arabe. Elle est la base de la pratique quotidienne, puisque les dictionnaires arabes classent leurs entrées par les racines. Les modèles permettent la reconstitution des paradigmes dérivationnels, conjugationnels ou inflexionnels.

Mais l'intérêt que nous portons à l'arabe n'est pas qu'historique, il est aussi théorique. En effet, le grec et le latin font grand usage de préfixes et de suffixes. Les phénomènes comme le redoublement, par exemple dans la formation des aoristes figés du grec, ou les alternances vocaliques internes, comme dans les opposition du genre *facere*/{*con-, suf-, etc.*}*ficere*, apparaissent comme des irrégularités à première vue. L'infixation régulière n'existe donc pas dans ces langues. Il faudra, en Europe, attendre l'étude des langues germaniques et slaves pour que l'on s'intéresse réellement aux alternances vocaliques des radicaux dans les langues indo-européennes[42].

L'arabe, et les langues sémitiques en général sont, c'est bien connu, construites sur un tout autre système. L'infixation multiple y est la règle. Or, du point de vue de la théorie formelle des langages, cette propriété serait, dans les vues des génératives plus compliquée que la simple préfixation ou suffixation. Nombrilisme indo-européen ! Nous montrerons que l'analogie permet de remettre les choses en place (p. 252) en n'octroyant aucune difficulté supplémentaire à l'infixation multiple.

1.1.4 Les théologiens médiévaux : entre univocité et équivocité

Pour en revenir au Moyen-Âge, le débat sur la séparation des domaines respectifs de la théologie et de la métaphysique, qui se soldera par les condamnations de 1277, affirmant en gros que les philosophes ne sont pas les seuls à pouvoir accéder à la connaissance de Dieu, repose essentiellement sur la possibilité du transfert à Dieu des qualités de l'être, et donc sur la nature de l'être. La question centrale était de caractériser ce transfert. Or, trois vues différentes s'opposaient, deux extrêmes, et une intermédiaire.

Deleuze explique d'abord les thèses extrêmes, caractérisées par les termes d'équivoque et d'univoque :

[41] WIEM, *Wielka Encyklopedia Internetowa Multimedialna*, 1996 2001, article ANALOGIA.
[42] C'est-à-dire à l'apophonie puisqu'il s'agit en fait de morphologie (voir plus bas, p. 65).

> Ceux que l'on appelait partisans de l'équivocité, ça importe peu qui c'était, ils discutaient une chose très simple : que ces différents sens du mot être étaient sans commune mesure [...] Alors le point d'hérésie de l'équivocité c'est que ceux qui disaient que l'être se dit en plusieurs sens et que ces différents sens n'ont aucune commune mesure, comprenez qu'à la limite ils préféraient dire : « Dieu n'est pas », plutôt que dire « il est », dans la mesure où « il est » était un énoncé qui se disait de la table ou de la chaise. Ou alors il est d'une tellement autre manière, d'une manière tellement équivoque, tellement différente et sans commune mesure avec l'être de la chaise, avec l'être de l'homme, etc. que, à tout bien considérer, il vaut mieux encore dire : il n'est pas, ce qui veut dire : il est supérieur à l'être. Mais s'ils avaient le sens des jeux de mots ça devenait très dangereux, [...]
>
> Puis il y en avait qui étaient partisans de l'univocité de l'être. Ils risquaient encore plus parce que qu'est-ce que ça veut dire par opposition à l'équivocité de l'être, l'univocité ? [...] Ça voulait dire : l'être n'a qu'un sens et se dit en un seul et même sens de tout ce dont il se dit. Là on sent que si les équivocistes avaient déjà comme le péché possible en eux, c'était, les univocistes, des penseurs qui nous disaient : de tout ce qui est, l'être se dit en un seul et même sens, il se dit en un seul et même sens d'une chaise, d'un animal, d'un homme ou de Dieu. Encore une fois, je simplifie tout [...][43] là.

Puis, Deleuze rappelle que les tenants de la position intermédiaire, sont les tenants de l'analogue. Ils soutiennent eux que :

> L'être qui est analogue, ça voulait dire : oui, l'être se dit en plusieurs sens de ce dont il se dit. Seulement ces sens ne sont pas sans commune mesure : ces sens sont régis par des rapports d'analogie.[44]

Ici, nous n'hésitons pas à citer trois longs passages transcrits[45], non seulement par plaisir, puisque la parole de Deleuze est savoureuse, mais aussi par intérêt polémique, parce que Deleuze insiste sur le fait que l'analogie peut être entendue en un sens vulgaire, et dans un sens technique, scientifique, mathématique (qui se divise à nouveau en deux). Nous aussi contestons le sens vulgaire comme étant impropre, chose que nous avons dite précédemment afin d'écarter certains travaux non pertinents (voir p. 83)[46]. Et bien sûr, nous concentrerons notre travail sur le deuxième sens. Laissons parler Deleuze :

[43] DELEUZE, *Anti-œdipe et Mille plateaux*, 1974, § 5 et 6.

[44] Ibidem, § 8.

[45] Ibidem, § 10, 11 et 13.

[46] Voir aussi MILNER, *Introduction à une science du langage*, 1989, p. 631, note 3 de bas de page, où il est dit que : ..., l'ensemble du chapitre IV de la troisième partie du *Cours* [de Saussure], intitulé « L'analogie », représente une tentative remarquable et réussie, visant à retrouver, par delà l'usage moderne et imprécis du terme *analogie*, son usage ancien et précis. Fin de citation.

> Alors qu'est-ce que ça veut dire : l'être se dit en plusieurs sens de ce dont il se dit et ces sens ne sont pas sans commune mesure, ils ont une mesure analogique ? Éh bien ! dans les thèses de Saint-Thomas, que je simplifie beaucoup, ça veut dire deux choses car l'analogie qui est ici prise en un sens technique ou scientifique, l'analogie était double, de toute manière prise dans un sens technique ou scientifique, c'est-à-dire qu'il ne s'agissait pas de l'analogie vulgaire. L'analogie vulgaire c'est la simple similitude de la perception : quelque chose est analogue à quelque chose d'autre. Si vous voulez c'est la similitude de la perception ou l'analogie de l'imagination, en gros ça se tient. L'analogie scientifique ou technique, l'analogie des concepts, elle est double : la première était nommée par Saint-Thomas analogie des proportions et la seconde était nommée par Saint-Thomas analogie de proportionnalité.
>
> L'analogie de proportion c'était ceci : l'être se dit en plusieurs sens et ces sens ne sont pas sans commune mesure, ils ont une mesure intérieure, ils ont une mesure conceptuelle, ils ont une mesure dans le concept. Pourquoi ? Éh bien !, au premier sens de l'analogie de proportion, ça voulait dire – parce qu'il y a un sens premier du mot être et puis des sens dérivés –, le sens premier du mot être c'était ce que l'on traduit souvent sous le terme « substance » ou parfois sous le terme « essence ». Les autres sens du mot être c'était des sens différents du mot être qui dérivaient suivant une loi de proportion du premier sens. Donc l'être se disait en plusieurs sens mais il y avait un sens premier dont les autres dérivaient.
>
> Et puis la seconde forme d'analogie scientifique, qui ne s'opposait pas à la première c'était l'analogie de proportionnalité qui consistait cette fois dans une figure bien proche de son équivalent, l'analogie mathématique : A est à B ce que C est à D. Exemple donné par Saint-Thomas : Dieu est bon. Suivant l'analogie de proportion : Dieu est bon et l'homme est bon ; suivant l'analogie de proportion Dieu est formellement bon, c'est-à-dire possède en soi la bonté dans la plénitude de cette qualité, et l'homme n'est bon que par dérivation en tant que créature de Dieu, donc l'homme est secondairement bon. C'est l'analogie de proportion. L'analogie de proportionnalité c'est le même exemple, mais vous devez sentir que ça change. Ce que la bonté infinie est à Dieu, la bonté finie l'est à l'homme.

Salamanca résume lui aussi le résultat des controverses médiévales de la façon suivante :

> Les attributs de Dieu ne peuvent être connus, ou plutôt, reconnus que par analogie avec les êtres finis, et seulement d'une certaine manière. Les êtres finis participent à l'essence de Dieu, et ceci d'une manière analogique. Pour autant, dans la tradition médiévale platonico-aristotélicienne, l'analogie implique une ressemblance exprimée normalement par une relation entre quatre termes, ou proportion : « L'œil est au corps comme

> l'intellect est à l'âme » ou bien implique une participation en un sens platonique. Selon la conception scholastique, qui prend sa source dans la métaphysique d'Aristote, il y a trois modes de connaissance :
>
> - l'univocité ou synonymie de l'être : « l'être est un » ;
> - l'équivocité ou homonymie de l'être : « l'être peut être de multiples manières » ;
> - et l'analogie : « les êtres sont semblables et participent les uns des autres ; l'être du chant est aussi dans celui qui chante. »[47]

Et citons encore le *Trésor de la langue française* pour donner clairement la position intermédiaire de l'analogie dans ce débat :

> *Analogie de l'être.* Thèse centrale de la philosophie scolastique d'après laquelle la notion d'être et les autres notions transcendentales (Un, Bien, Vrai) ne sont ni univoques ni équivoques mais analogiques. L'analogie se situe entre l'univocité (ou ressemblance pure, identité) et l'équivocité (ou pure dissemblance). L'analogie tend à exprimer ce qu'il y a de semblable et de différent entre Dieu et les créatures.[48]

Dans ce débat, on sait que la pensée arabe influencera fortement l'Occident médiéval. Ici se placent les grandes figures d'Ibn Rochd dit Averroès (1126-1198) et d'Ibn Sīnā dit Abu Sīnā ou Avicenne (980-1037). Pratiquement, il s'agissait, pour des raisons liées aussi à la politique, de séparer la métaphysique aussi bien de la physique que de la théologie. De la physique, à la suite d'Aristote lui-même (la Métaphysique est, rappelons-le, appelée ainsi pour l'unique raison que c'est le livre qui suit celui de la Physique). De la théologie, à la suite du débat entre Avicenne et Al Gazālī ou Algazel (1058–1111) puis Averroès. Pour bien montrer que la question de savoir où s'arrête la philosophie et où commence la théologie tourne bien autour de la notion d'analogie, citons le résumé qu'en fait Libera :

> La controverse médiévale sur « Avicenne et Averroès » porte, d'ailleurs, autant sur leurs conceptions respectives de l'analogie que sur leur découpage du sujet de la métaphysique.[49]

[47] SALAMANCA, *La tradición histórica de la analogía lingüística*, 1984, p. 373–374.

[48] Institut National de la Langue Française, *Trésor de la langue française informatisé*, 2000, article ANALOGIE.

[49] de LIBERA, *La philosophie médiévale*, 1992, p. 74.

1.2 Les Modernes : du fauteur de trouble au facteur d'ordre

Nous sautons par-dessus la Renaissance et les Classiques. D'une part, la Renaissance peut être vue comme une redécouverte des auteurs antiques. Et d'autre part, aucune de nos lectures sur l'analogie ne fait allusion au XVII^{e} ni au XVIII^{e} siècles. Et nous ne pensons pas qu'il y ait mention de l'analogie ou de problématique voisine dans la Grammaire de Port-Royal, pour ne citer que cet ouvrage pourtant particulièrement important de l'histoire de la grammaire. Mais l'analogie faisait sans doute partie du bagage de l'honnête homme, puisque le dictionnaire de l'Académie française de 1694 donne :

> ANALOGIE. s.f. Terme dogmatique, Rapport, ressemblance, conformité, proportion d'une chose à une autre. *La partie basse de la montagne s'appelle pied par analogie au pied de l'homme. analogie geometrique. analogie grammaticale. le mot ambitionner est formé par analogie d'ambition, comme passionné* (sic) *est formé de passion. regles d'analogie. par raison d'analogie.* [50]

On remarquera tout particulièrement l'exemple du mot *ambitionner* qui est une dérivation morphologique par analogie, du type de celles qui intéresseront Saussure à deux siècles de distance (voir p. 62). Dans la cinquième édition du dictionnaire, la définition du terme se fait plus précise, en quatre parties : sens commun, en histoire, en morale et en grammaire :

> ANALOGIE. s.f. Rapport, ressemblance, proportion. Il s'emploie un peu diversement en Mathématiques et en Philosophie. Dans les premières, il signifie, Rapport exact et rigoureux. *Il y a la même analogie de deux à trois, que de six à neuf. La solution de ce problème dépend de l'analogie, de plusieurs analogies.* En philosophie, il se dit Des rapports plus ou moins éloignés, même de similitude. *L'analogie du fer avec l'aimant. La partie basse d'une montagne s'appelle* le pied de la montagne, *par analogie avec* le pied de l'homme. *Raisonner par analogie. Foible analogie. Analogie frappante. Il ne faut pas toujours conclure par analogie.*
>
> [...]
>
> Il [le terme ANALOGIE] se dit aussi en termes de Grammaire, pour marquer Le rapport que divers mots d'une Langue ont ensemble pour leur formation. *Le mot passionné est formé de passion, par la même analogie qu'affectionné l'est d'affection.* [51]

Aussi, a posteriori, selon Salamanca, Alexandre von Humboldt (1767–1835) n'aurait vu dans ce que les grammairiens français appelaient « le génie de la langue », rien d'autre que l'analogie[52]. On peut en effet en déceler l'usage

[50] Académie française, *Dictionnaire de l'Académie française*, 1694, vol. 1, p. 38.

[51] Académie française, *Dictionnaire de l'Académie française*, an VI de la Republique 1798, vol 1., p. 56. Nous respectons la graphie originale, et en particulier l'usage des majuscules.

[52] SALAMANCA, *La tradición histórica de la analogía lingüística*, 1984, p. 379.

implicite dans la résolution de beaucoup de problèmes du « beau parler », typiques de l'époque : doit-on dire *sers-je* ou *servé-je*[53] ? Ou encore : faut-il dire *je vais* ou *je vas* ? qui opposent le peuple à la cour, certains auteurs entre eux. Et les professionnels de la diction, c'est-à-dire les comédiens, se déchirent au XVII^{e} siècle sur la prononciation française des noms latins : dit-on *Vergille* ou *Virgile* ? *Aule-Gelle* ou *Agelle* ? etc. Pour toutes ces questions, le choix est clairement à faire entre des formes héritées et des formes reconstruites par comparaisons avec d'autres séries, comparaisons qui reposent sur des égalités de rapports, c'est-à-dire sur des analogies. Il faudrait dire *sers-je* parce que l'on dit *dois-je* : $\quad je\ dois : dois\text{-}je\ ? = je\ sers : x \quad \Rightarrow \quad x = sers\text{-}je\ ?$.

Cependant c'est plutôt en tant que procédé de raisonnement que l'on trouve mention de l'analogie à cette époque, ce que la définition du *Dictionnaire de l'Académie* donnée plus haut mentionne par *raison d'analogie*. Ainsi, Beauzée (1717–1789) écrit dans l'Encyclopédie à l'article SENS :

> Si au témoignage des sens, nous ajoutons l'analogie, nous y trouverons une nouvelle preuve de la vérité des choses. L'analogie a pour fondement ce principe extrêmement simple, que l'univers est gouverné par des lois générales & constantes. C'est en vertu de ce raisonnement que nous admettons la règle suivante, que des effets semblables ont les mêmes causes.
> L'utilité de l'analogie consiste en ce qu'elle nous épargne mille discussions inutiles, que nous serions obligés de répéter sur chaque corps en particulier. Il suffit que nous sachions que tout est gouverné par des lois générales & constantes, pour être bien fondés à croire, que les corps qui nous paroissent semblables ont les mêmes propriétés, que les fruits d'un même arbre ont le meme goût, &c. La certitude qui accompagne l'analogie retombe sur les sens mêmes, qui lui prêtent tous les raisonnemens qu'elle déduit.[54]

Nous avons déjà vu cet aspect de l'analogie qui, quoique procédé rhétorique apprécié (voir p. 29), pose un problème de connaissance scientifique (voir p. 27). Ce problème est d'ailleurs épinglé dans le dernier exemple de la cinquième édition du *Dictionnaire de l'Académie* cité plus haut : *il ne faut pas toujours conclure par analogie* !

1.2.1 Les comparatistes : fausseté de l'analogie

En entrant dans le XIX^{e} siècle, l'analogie confirme son statut de notion importante de la linguistique, mais pas de façon directe. C'est parce qu'elle s'oppose aux lois phonétiques qu'elle se retrouve au coeur de la controverse majeure qui a présidé à la formation de la linguistique en tant que science.

[53] À l'époque, le e accent aigu se prononce bien comme il s'écrit.

[54] DIDEROT & d'ALEMBERT, *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, entre 1751 et 1774, article SENS, rédigé par Beauzée. Accès par le formulaire de recherche du site `http://encyclopedie.inalf.fr/`. De nouveau, nous respectons l'orthographe originale.

On peut dire, avec le recul, que la linguistique de la fin du XIX^{e} siècle, c'est-à-dire l'ensemble des travaux des linguistes comparatistes et des Néogrammairiens (*Junggrammatiker* en allemand), apparaît comme traversée par la recherche des invariants et des variantes : comment ramener ce qui change entre différentes langues à des prototypes communs historiques ? Par exemple, comment ramener le grec ἵππος au latin *equus* ou au sanscrit *ašvas*[55], etc. ? On sait que la réponse a été apportée par la « découverte » des lois phonétiques.

Le travail sur des langues européennes apparentées permettra la mise en correspondance des sons d'une langue à une autre. Les analogies régulières d'un état passé, c'est-à-dire les paradigmes de conjugaison ou de déclinaison apparaissent donc à un état ultérieur comme troublées par des changements de sons. Historiquement, ces observations remontent à loin, et les premières traces de l'utilisation de l'analogie comparative remonteraient même à la Renaissance selon Mounin.

> On voit poindre aussi [à la Renaissance] les premières règles de correspondance phonétique (*x* latin = *ss* italien, *i* latin de *litera* = *e* fermé de *lettera*, etc.) ; par exemple, chez Tolomei, ou Castelvetro en Italie, qui commencent aussi à se servir de l'analogie comparative : si *habeo* > *haggio*, en italien, *creggio* doit s'expliquer par l'existence d'un *credeo*[56].[57]

Mais toute conjugaison ou déclinaison ne nous apparaît pas troublée. Et nombreuses sont les formes qui, malgré les changements phonétiques qui auraient précisément dû les troubler restent régulières. Elles ne peuvent donc s'expliquer par changement phonétique. C'est donc qu'elles ont été alignées sur le reste du système, et elles se voient donc appelées « fausses analogies ». Fausses, parce qu'elles ne sont pas vraies comme les changements phonétiques. Dans une telle vue, la fausse analogie n'a plus le même statut épistémologique que l'analogie régulière des Grecs et des Latins.

L'idée que les lois phonétiques seraient justifiées par des considérations physiologiques finira d'ériger les changements phonétiques en lois naturelles, qui deviennent donc « les lois phonétiques. »[58] Parallèlement, se dégagent entre autres deux principes méthodologiques fondamentaux pour la linguistique naissante :

- la langue est un objet historique d'étude, auquel on peut s'intéresser en faisant fi des considérations sociologiques ou psychologiques ;
- les lois phonétiques, principal objet de cette science naturelle, puisqu'elles sont considérées comme lois naturelles, ne doivent souffrir aucune exception, et doivent donc s'appliquer partout[59].

[55] Ces trois mots désignent le cheval.

[56] Rappelons que la forme classique est *credo* (je crois). *Habeo* (j'ai). *Litera* (la lettre).

[57] MOUNIN, *Histoire de la linguistique – Des origines au XX^{e} siècle*, 1967, p. 127.

[58] Voir aussi MOUNIN, op. cit., p. 204.

[59] BOLTANSKI, *La linguistique diachronique*, 1995, p. 60, dernier paragraphe, ou encore,

L'adhésion de tous à cette vue sera emportée, on le sait, après la publication par Verner (1846–1893) de son explication à l'une des dernières exceptions qui subsistaient encore aux lois de Grimm (1785–1863), explication qui fait intervenir l'accent de mot[60]. Au royaume des lois phonétiques universelles triomphantes, la création de formes analogiques régulières est considérée finalement comme une perturbation anti-naturelle des lois naturelles du changement phonétique.

En passant, on pourrait s'interroger sur l'ironie qui a voulu que les changements phonétiques aient été en fait découverts essentiellement par l'usage d'une forme particulière d'analogie jouant sur plusieurs domaines :

$$\text{ὑπήρ} : super = \text{ἡμι-} : semi- \quad \Rightarrow \quad \text{/h/ en grec : /s/ en latin}$$

Cette forme particulière, avec correspondance entre deux domaines, c'est-à-dire entre deux langues ou deux systèmes, était en effet la clé des études comparatives.

En un siècle d'histoire de linguistique comparée et historique, la régularité passe donc du camp de l'analogie à celui des changements phonétiques. Le retournement est donc complet par rapport à Varron (p. 44). Jusqu'au début du XX^{e} siècle, les lois phonétiques vont seules avoir droit au statut d'objets scientifiques. Au contraire, les créations par analogie, ne suivant pas ces lois naturelles que sont les lois phonétiques, vont elles être considérées comme des écarts aberrants[61]. La meilleure preuve étant que ces créations sont souvent des barbarismes ou des fautes, donc à ranger dans le faux. Un exemple fameux en est donné par Saussure lui-même[62], avec la forme *viendre*, solution de :

$$\textit{éteindrai} : \textit{éteindre} = \textit{viendrai} : x \quad \Rightarrow \quad x = \textit{viendre}.$$

On pourrait citer des centaines de tels exemples, et si l'on devait se convaincre du caractère universel du phénomène, donnons en passant un exemple dans le dialecte japonais du Touhoku :

> [...] de même que l'on a *mita* : *minai*, de même on a, pour le changement de *kita* à la forme négative, la création *kinai*. Ici, dès l'apparition d'une alternance grammaticale irrégulière par changement phonétique, il y a eu travail de création d'une nouvelle régularité (système de conjugaison régulier).[63]

---

cité par MOUNIN, *Histoire de la linguistique – Des origines au XX^{e} siècle*, 1967, p. 209, la phrase suivante de Scherer (1841–1886) « Les changements phonétiques que nous pouvons observer dans l'histoire linguistique fondée sur des documents, procèdent selon des lois déterminées qui ne souffrent aucune dérogation, excepté en accord avec d'autres lois. »

[60] VERNER, *Eine Ausnahme der ersten Lautverschiebung*, 1875. Le problème venait de la double correspondance : sanscrit *t*, gothique *θ*, allemand *d* dans *bhráta*, *broθar*, *Bruder* (frère), et sanscrit *t*, gothique *d*, allemand *t* dans *pitá*, *fadar*, *Vater* (père). Il est levé en tenant compte de l'accent des mots sanscrits.

[61] Les Anglo-saxons appellent cela le paradoxe de Sturtevant : les changements phonétiques seraient réguliers dans leur application mais créeraient de l'irrégularité. Les changements analogiques seraient irréguliers dans leur application mais créeraient de l'ordre.

[62] de SAUSSURE, *Cours de linguistique générale*, 1995 1e ed 1916, p. 231.

[63] *Kinai* (ne pas venir), à la place de la forme *konai* (même sens) de la langue standard et

1.2.2 Les néogrammairiens : remise à l'étude progressive

Évidemment, écarter l'analogie sous prétexte d'anti-naturalité, ne résoud rien, et il faut bien trouver une place et une explication au phénomène. Hermann Paul (1846–1921), justement l'un des chefs de file des Néogrammairiens, ceux-là mêmes dont le programme initial était de donner un statut définitif de lois physiques aux changements phonétiques, se penchera sur la question et rétablira l'analogie comme sujet d'étude de plein droit. Il reconnaît en effet à l'analogie un plein statut d'opération sous-jacente de l'activité langagière, au point de lui consacrer un chapitre entier de son célèbre livre *Prinzipien der Sprachgeschichte*[64]. Reprenant les conceptions grecques et latines, il voit l'analogie dans ce qu'on appelle maintenant paradigmes et qu'il nomme *Proportionengruppen* (groupes de proportions). En premier lieu, l'analogie fait intervenir deux dimensions dans les groupes de mots : le sens et la forme sonore. Mais il faut bien constater qu'un nombre important de tels groupes s'opposant par le sens ne s'opposent pas par la forme, comme $taureau : vache = homme : femme \quad = \quad frère : sœur = moine : nonne$, ou encore $bon : meilleur = mauvais : pire$, etc. La similarité de forme dans les groupes de proportions apparaît donc soit selon les deux dimensions, par exemple $il\ mange : il\ mangea = il\ dérange : il\ dérangea$, soit selon seulement une seule, par exemple $il\ mange : il\ mangea = il\ voit : il\ vit$, soit selon aucune, par exemple $il\ va : il\ alla = il\ voit : il\ vit$.

Dans le premier des cas, les proportions sont telles que la permutation des moyens, transformation possible avec toutes les proportions, selon les propres termes d'Hermann Paul (voir plus bas, page 114, pour la citation exacte), s'applique aussi : $il\ mange : il\ dérange = il\ mangea : il\ dérangea$ [65].

Plusieurs points nous intéressent tout particulièrement dans les propos d'Hermann Paul sur les groupes de proportions.

D'abord, il mentionne que c'est sans doute par contiguïté qu'à partir de

$$pater\ mortuus \ : \ filia\ pulchra : caput\ magnum$$

on déduit $pater : filia : caput = mortuus : pulchra : magnum$ [66]. Nous reparlerons de la contiguïté en général dans notre développement sur les notions voisines de l'analogie (voir plus bas, p. 100). On pressent bien que Paul a mis là le doigt sur quelque chose de très important et de sans doute très prometteur formellement. Malheureusement, la contiguïté restant pour nous encore un mystère, nous sommes bien en peine d'exploiter formellement cette idée. Ce à notre plus grand regret.

---

en face de l'accompli *kita* (être venu), est aligné sur la série régulière de la langue standard *minai* (ne pas voir), *mita* (avoir vu). $Mita : minai = kita : x \quad \Rightarrow \quad x = kinai$ au lieu de *konai*. 井上 史雄 (INOUE Humio), 言語の構造の変遷・東北方言音韻史を例として, 1999, p. 278.

[64] PAUL, *Prinzipien der Sprachgeschichte*, 1920 5e ed 1e ed 1880, chap. 5 Analogie, p. 106–120.

[65] Ibidem, p 106–107.

[66] Ibidem, p 109.

Ensuite, Hermann Paul fait explicitement de l'analogie un moyen de création en mentionnant ce qu'il appelle les équations analogiques (*Proportionengleichungen* en allemand). Il pose clairement que :

> Les mots ou groupes de mots que nous employons dans le discours apparaissent seulement en partie comme simple reproduction mnémonique du déjà reçu. Une part aussi importante y est prise par une faculté combinatoire fondée sur l'existence des groupes de proportions. La combinaison consiste là dans une certaine mesure en la résolution d'une équation analogique lors de laquelle un deuxième membre de proportion est librement créé, selon un modèle de proportions analogiques devenues courantes, pour un autre mot également courant. Nous appelons ce fait une création analogique. Il est indubitable qu'un grand nombre de formes morphologiques et de relations syntaxiques, qui n'ont jamais été auparavant introduites du monde extérieur dans l'intellect, est obtenu à l'aide des groupes de proportions, non seulement de la plus simple des façons, mais aussi sans cesse et en toute confiance sans que le locuteur ait jamais le sentiment de quitter le terrain sûr de l'acquis.[67]

Enfin, il souligne longuement que c'est l'analogie qui est le moteur de l'apprentissage par modèles. L'apprenant d'une langue a une faculté de généralisation qui lui permet

> d'extraire des règles inconsciemment à partir de modèles.[68]

C'est la répétition et le nombre de modèles qui selon lui permettent la création des règles. On voit donc ici poindre une notion de fréquence sur laquelle nous reviendrons avec le travail de Mańczak (page 68), et que nous essaierons de faire nôtre dans nos expérimentations avec des espaces analogiques (page 296).

1.2.3 L'École de Kazan : caractère subjectif et individuel

À la même époque, le travail de Baudouin de Courtenay (1845–1929) se situe sur le même terreau. Ce linguiste, ainsi que son disciple et concurrent Kruszewski (1851–1887) sont « *ignorés de la généralité des savants occidentaux* » (Godel, *Sources manuscrites du Cours de linguistique générale de F. de Saussure*), alors qu'ils « *ont été plus près que personne d'une vue théorique de la langue.* » Il a, semble-t-il, manqué un Bally (1865–1947) et un Sechehaye à Baudouin de Courtenay, car, selon les propres termes de ses élèves,

> le *Cours* de Saussure ne contenait rien qu'ils n'aient déjà appris de Baudouin.[69]

[67] Ibidem, p 110.
[68] Ibidem, p 111.
[69] Stankiewicz, *Baudouin de Courtenay i podstawy współczesnego językoznawstwa*, 1986, p. 7.

Pour ajouter à la difficulté, l'accès aux textes de Baudouin de Courtenay est un peu difficile et l'effort n'est peut-être pas payant, pour la raison que :

> En vrai « philosophe de la langue », Baudouin s'est toujours intéressé profondément aux problèmes de linguistique générale, mais ses opinions sur ces thèmes sont dispersés dans de nombreux livres, articles et même critiques d'ouvrages, de peu d'importance.[70]

Revenons aux problème de l'analogie, ce fauteur de trouble dans l'ordonnancement qui devrait normalement résulter de l'application partout des changements phonétiques. Le souci de Baudouin de Courtenay est de trouver la place de l'analogie dans la science linguistique en gestation à l'époque. Pour lui, les faits sont incontournables, avec ces exceptions aux lois phonétiques que sont les fausses analogies, ces créations populaires, et ces barbarismes qui deviennent la règle. Son étude des contacts entre langues (contacts entre le slovène d'une part et l'allemand ou l'italien d'autre part) inclinent finalement Baudouin à écrire que les changements phonétiques

> ne tiennent pas devant certains faits réels de l'activité langagière ni devant l'aspect psycho-social de la langue.[71]

Baudouin de Courtenay sera conduit, à l'instar d'autres linguistes qui lui sont contemporains, à la remise en cause des deux principes précédemment énoncés (p. 54). D'une part, les exceptions aux règles phonétiques doivent être interprétées de façon statistique, chose sur laquelle nous reviendrons avec Mańczak (p. 68), et Baudoin est donc amené, a contrario, à examiner ce qui *reste toujours stable dans la langue,* et qui en constitue donc *les propriétés caractéristiques invariables.* D'autre part, l'idée d'un déterminisme de l'évolution des langues est à abandonner par la prise en compte des productions volontaires des locuteurs. Cette discussion est évidemment à rattacher à l'opposition saussurienne de *langue* et de *parole.* Ainsi, prenant finalement l'exact contrepied des principes que nous avions énoncés en parlant des comparatistes, Baudouin se penchera sur l'analogie en tant qu'activité psychologique du locuteur.

> Les phénomènes d'étymologie populaire et de comparaison analogique (que Baudouin appelle « assimilation morphologique », les fautes grammaticales commises par les enfants et les réinterprétations de termes étrangers ne doivent pas être traités comme des manifestations de « pathologie linguistique », mais bien plutôt comme des « manifestations de la capacité créatrice » du locuteur « à regrouper des mots isolés dans des moules sémantiquement déterminés. » Dans les regroupements de ce type, selon Baudouin, deux principes d'association, formulés d'abord par Kruszewski (qui les a repris à son tour des philosophes anglais) jouent un rôle : l'association par ressemblance et l'association

[70] Ibidem, p. 20–21.

[71] BAUDOUIN, cité par STANKIEWICZ, op.cit., p. 22.

> par voisinage. « Ces deux lois d'association transforment une masse de mots en un tout harmonieux de systèmes coordonnés (ou nids) et de séries ordonnées (ou rangs). » A la différence de Kruszewski, qui n'applique ces principes qu'à la morphologie, Baudouin voyait leur action aussi dans le domaine de la phonologie, parce que, comme il l'affirmait, les lois de « regroupement » et « d'équilibre » s'appliquent à tous les éléments et à tous les niveaux de la langue.[72]

Il ne nous reste qu'à ajouter que les philosophes qu'avait lus Kruszewski transposaient sans aucun doute la distinction faite par les Anciens entre métaphore et métonymie, c'est-à-dire entre transfert soit par traits « semblables » soit par traits « contigus », à l'analogie (voir plus haut, p. 40, note de bas de page), ainsi que les écrits d'Aristote sur la mémoire. Plus près de nous, cette idée de l'analogie comme opération à la fois métaphorique et métonymique sera illustrée par Itkonen (voir plus bas, p. 100).

Pour ce qui est de la phonologie, Baudouin de Courtenay est considéré comme l'un des précurseurs de la formalisation de la notion de phonème[73]. Et il a, semble-t-il, travaillé cette notion émergente en relation avec l'analogie. C'est pourquoi, l'encyclopédie WIEM note en raccourci à l'article concernant Baudouin :

> Il a travaillé le concept d'analogie linguistique et de phonème, et a analysé le rôle de l'alternance phonétique.[74]

Le lien entre la notion de phonème et celle d'analogie peut d'ailleurs être tracée plus avant dans le temps, puisque l'on trouve déjà dans le dictionnaire de l'Académie française de 1835, à l'article analogie :

> Analogie, se dit particulièrement, en Grammaire, Du rapport qu'ont entre elles les consonnes qui se prononcent avec la même partie de l'organe vocal. *Il y a de l'analogie entre le B et le P, consonnes labiables, le D et le T, consonnes dentales, etc.* [75]

1.2.4 Les structuralistes : le côté synchronique

C'est à Saussure (1857–1913) que reviendra, dans le *Mémoire sur les voyelles de l'indo-européen*, de renverser vraiment la vision qui fait de la création analogique une résistance aux changements dus aux lois phonétiques. Il peut lui attribuer la place qui lui faisait défaut dans la science linguistique grâce aux deux notions fondamentales de synchronie et de diachronie. Dans le *Mémoire*, Saussure traite en effet les changements vocaliques comme de simples alternances morphologiques. Il y expose un point de vue qui eut comme conséquence que :

[72] STANKIEWICZ, op. cit., p. 26–27. Les fragments cités sont de Baudouin de Courtenay.

[73] Voir ce qu'en dit MOUNIN, *Histoire de la linguistique – Des origines au XX^e siècle*, 1967, p. 223.

[74] WIEM, *Wielka Encyklopedia Internetowa Multimedialna*, 1996 2001, article BAUDOUIN DE COURTENAY.

[75] Académie française, *Dictionnaire de l'Académie française*, 1835, vol 1, p. 71.

> Il apparaissait maintenant que c'était plutôt les changements phonétiques qui dérangeaient les relations régulières des formes grammaticales, et que les uniformisations analogiques contrebalancent l'action des changements phonétiques « aveugles ».[76]

Saussure va même plus loin dans le *Cours*, où il énonce carrément que:

> [...] le phénomène phonétique est un facteur de trouble. Partout où il ne crée pas des alternances, il contribue à relâcher les liens grammaticaux qui unissent les mots entre eux.[77]

De nouveau, voici l'analogie renvoyée d'un extrême à l'autre, après Varron et les Néogrammairiens. Remarquons que tant que le changement phonétique crée des alternances, il n'est pas facteur de trouble. C'est cette conception, partagée par Baudouin et Kruszewski qui permet l'émergence de l'autonomie de l'alternance grammaticale comme « régularité », qu'elle provienne ou non du changement phonétique. Nous verrons d'ailleurs la distinction opérée à ce sujet par Kuryłowicz (p. 65). Ainsi, par exemple, que le /z/ et le /r/ alternant dans la conjugaison du verbe allemand *sein* avec les formes *waren* (fûmes, furent) et *gewesen* (été) soit difficilement explicable par un changement phonétique ne le rend pas plus irrégulier que l'alternance /s/ et /z/ apparaissant dans *Haus* et *Häuser*, qui, elle, s'explique par le fait qu'un /s/ entre deux voyelles devienne sonore. Et même plus, il devient faux de dire que le *a* se change en *ä* dans ce deuxième exemple, car aucun des deux n'est premier par rapport à l'autre. Nous reviendrons sur cette question de la primauté des formes lorsque nous commenterons les travaux de Kuryłowicz.

Dans le *Cours*, Saussure illustre son propos par un exemple célèbre que nous commentons maintenant. Il s'agit de l'explication selon laquelle la forme *honor* du latin classique est une création par analogie. Les faits sont les suivants: en latin, on constate les états de langue successifs suivants, pour les nominatifs et accusatifs des mots donnés dans le tableau 1.1,

Tableau 1.1: Action du rhotacisme puis de l'analogie

VII^{e} av. J.-C.	**II^{e} av. J.-C.**	**période classique**
actor, actorem	*actor, actorem*	*actor, actorem*
orator, oratorem	*orator, oratorem*	*orator, oratorem*
honos, honosem	*honos, honorem*	*honor, honorem*

Le passage de la première à la seconde colonne se fait par rhotacisme. Il s'agit d'un phénomène phonétique qui affecte la consonne /s/ :

entre deux voyelles, /s/ passe à /r/.

[76] Stankiewicz, op. cit., p. 31.

[77] de Saussure, *Cours de linguistique générale*, 1995 1e ed 1916, p. 221.

Par exemple :

$$/\text{honosem}/ \rightarrow /\text{honorem}/$$

En revanche, dans le mot *honos*, le /s/, final, ne se trouvant pas entre deux voyelles, aucun changement ne prend place. La forme *honos* devrait rester telle quelle. Le rhotacisme, changement insensible prenant place durant tout un intervalle de temps historique, est un phénomène **diachronique**, dont les locuteurs de la langue ne sont pas conscients, sauf par l'écriture, et seulement une fois le passage consommé.

Le passage de la deuxième à la troisième colonne, c'est-à-dire d'*honos* à *honor*, ne saurait s'expliquer par le rhotacisme qui ne s'applique qu'entre deux voyelles comme nous venons de le voir. La thèse de Saussure est que le rétablissement de la régularité des formes n'est pas obtenu par un phénomène phonétique : *honor* n'est pas l'aboutissement d'une transformation phonétique de *honos*. Selon lui, la nouvelle forme, *honor*, est une création consciente des locuteurs : *honor* est à *honorem* ce que *orator* est à *oratorem*. Il s'agit donc là d'une analogie au sens d'Euclide et d'Aristote, que l'on peut noter :

$$oratorem : orator = honorem : honor$$

ou mieux, pour expliciter le processus de création, sous forme d'équation à résoudre :

$$oratorem : orator = honorem : x \quad \Rightarrow \quad x = honor$$

Pour laisser la parole à Saussure lui-même, dans le *Cours de linguistique générale*, il énonce que la forme latine *honos* est remplacée

> [...] par la forme nouvelle *honor*, créée sur le modèle $\bar{o}r\bar{a}tor : \bar{o}r\bar{a}t\bar{o}rem$, *etc.*, par un procédé que nous étudierons plus bas et que nous ramenons dès maintenant au calcul de la quatrième proportionnelle :
>
> $$\bar{o}r\bar{a}t\bar{o}rem : \bar{o}r\bar{a}tor = hon\bar{o}rem : x$$
>
> $$x = honor$$
>
> On voit donc que, pour contrebalancer l'action diversifiante du changement phonétique ($hon\bar{o}s : hon\bar{o}rem$), l'analogie a de nouveau unifié les formes et rétabli la régularité ($hon\bar{o}r : hon\bar{o}rem$).[78]

Saussure reprend clairement la définition d'Aristote en l'appliquant aux mots ou formes, tout comme Hermann Paul. Cette définition fait bien intervenir quatre termes. Chose intéressante, elle met aussi clairement en évidence, par une notation inspirée des mathématiques, que l'analogie est ici une équation qui, étant donnés trois termes, fournit une solution, la quatrième forme manquante. Cette vue était aussi déjà présente chez Hermann Paul (voir plus haut, p. 57).

[78] Ibidem, p. 221–222.

> On le voit, il s'agit d'un « paraplasme », de l'installation d'un concurrent à côté d'une forme traditionnelle, d'une création enfin. Tandis que le changement phonétique n'introduit rien de nouveau sans annuler ce qui a précédé (*honōrem* remplace *honōsem*), la forme analogique n'entraîne pas nécessairement la disparition de celle qu'elle vient doubler. *Honor* et *honōs* ont coexisté pendant un temps et ont pu être employés l'un pour l'autre. Cependant, comme la langue répugne à maintenir deux signifiants pour une seule idée, le plus souvent la forme primitive, moins régulière, tombe en désuétude et disparaît.

Il faudra interroger cette dernière affirmation, tempérée par le « le plus souvent » par Saussure lui-même et interroger cette disparition. C'est précisément à cette tâche que se sont attelés Kuryłowicz et Mańczak qui montreront qu'il n'en va pas toujours ainsi, et que plusieurs facteurs, de longueur de mots et de fréquence, semblent intervenir dans la concurrence entre formes historiques et formes analogiques (voir plus bas, p. 68).

L'analogie est donc un phénomène **synchronique**, dont les locuteurs de la langue sont pleinement conscients. Cela explique que, pendant toute une période, les deux formes existent simultanément dans la langue et sont en concurrence. Il est facile d'illustrer en français par un grand nombre de telles formes concurrençant d'autres formes considérées comme correctes. Par exemple, les formes erronées du type *chausse-trappe* (pour *chausse-trape*) ou bien incorrectes par barbarisme du type *viendre* (pour *venir*), ou encore fausses par hypercorrection, du type *contredites* (pour *contredisez*), etc.

En résumé, pour Saussure, qui précise donc l'idée des Néogrammairiens, l'analogie est un phénomène synchronique qui contrebalance le désordre en morphologie introduit par les changements phonétiques diachroniques.

> L'analogie suppose un modèle, et son imitation régulière. Une forme analogique est une forme faite à l'image d'une ou plusieurs autres d'après une règle déterminée.[79]

Concluant le débat sur ce qui serait facteur d'ordre et ce qui serait faiseur de trouble dans la langue, il lui revient de replacer alors l'analogie sur un pied d'égalité avec le changement phonétique, à la suite d'Hermann Paul, pour en faire l'autre cause de l'évolution des langues.

> Les premiers linguistes n'ont pas compris la nature du phénomène de l'analogie, qu'ils appelaient « fausse analogie ». Ils croyaient qu'en inventant *honor* le latin « s'était trompé » sur le prototype *honōs*. Pour eux, tout ce qui s'écarte de l'ordre donné est une irrégularité, une infraction à une forme idéale. C'est que, par une illusion très caractéristique de l'époque, on voyait dans l'état originel de la langue quelque chose de supérieur et de parfait, sans même se demander si cet état n'avait pas été précédé d'un autre. Toute liberté prise à son égard était donc

[79] Ibidem, p. 221 et 224.

> une anomalie. C'est l'école néogrammairienne qui a pour la première fois assigné à l'analogie sa vraie place en montrant qu'elle est, avec les changements phonétiques, le grand facteur de l'évolution des langues, le procédé par lequel elles passent d'un état d'organisation à un autre.[80]

1.2.5 Les fondateurs de la phonologie : l'opposition

Nous avons vu plus haut (p. 59) que, déjà, Baudouin de Courtenay voyait l'analogie à l'œuvre en phonologie. Au XX^{e} siècle, l'attention se déplace de l'analogie vers l'une de ses articulations constitutives, le rapport. Ce n'est donc plus tant la notion d'analogie qui est première, mais celle de rapport, notion appelée opposition chez Saussure. Historiquement, on constate que ce n'est qu'accessoirement que certains rapports peuvent être égaux, pour former alors une véritable proportion (ou analogie). Il en est ainsi dans la phonologie de Troubetzkoy (1890–1938), selon Ducrot et Schaeffer :

> Reprenant Troubetzkoy, on qualifie une opposition de *proportionnelle* si le rapport existant entre ces deux termes se retrouve dans au moins une autre opposition. Ainsi en français, l'opposition /p-b/ est proportionnelle puisqu'il existe également les oppositions /t-d/, /f-v/, etc. Plus importante est la notion de **corrélation** : il y a une série de 6 consonnes sourdes, /ptkfsʃ/, qui s'opposent à une série de 6 consonnes voisées, /bdgvzž/ ayant respectivement le même lieu d'articulation.[81]

La corrélation est donc tout simplement une série de proportions. Par exemple, en français, la série d'oppositions proportionnelles de sourde à sonore peut bien se noter comme suit :

$$/p/ : /b/ = /t/ : /d/ \quad = \quad /k/ : /g/ = /f/ : /v/ \quad = \quad /s/ : /z/ = /ʃ/ : /ž/$$

L'énoncé d'une telle série ne fait que conforter l'idée de sa réalité. Il est clair qu'il s'agit là d'analogies au sens exact du terme. Proportions, égalités de rapports et analogies étant, rappelons-le, synonymes.

Bien évidemment, un esprit épris de formalisation, et c'est le cas de Jakobson (1896–1982), inclinera à pousser au maximum la représentation, et voudra voir d'autres oppositions, en français, par exemple celles de sourdes à sonores ou d'avant à arrière, ou encore de compact à diffus et de grave à aigu (voir le tableau 1.2).

En poussant encore plus loin, on sera alors tenté de faire des oppositions dégagées des oppositions binaires, c'est-à-dire telles qu'elles opposent des phonèmes possédant un trait (par exemple, la sonorité), à ceux ne le possédant pas. On retrouve ici un rapport de contiguïté du type de l'antonymie, c'est-à-dire de la chose à son contraire, comme nous en avions vus dans notre

[80] Ibidem, p. 223.

[81] Ducrot & Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, 1978, p. 397.

Tableau 1.2: Tableau de consonnes françaises

/p/	/t/	/b/	/d/
/k/	/f/	/g/	/v/
/s/	/ʃ/	/z/	/ž/

introduction (voir p. 25). Ainsi, toujours en français, on pourra décrire les huit consonnes /pkszbgʃž/ à l'aide de trois traits binaires seulement puisque $2^3 = 8$: sonore, sibilante, arrière. Par exemple, /p/ devient +arrière, − sonore et − sibilante, /s/ devient − arrière, − sonore et +sibilante, /ʃ/ +arrière, − sonore et +sibilante, etc.

Pour ce qui est de l'interprétation analogique, puisque, par définition, une analogie fait intervenir quatre objets, elle mettra en jeu deux traits binaires. Par exemple, pour les consonnes /szpb/, en notant d'abord le trait arrière puis le trait sonore, on aura l'explication suivante :

$$\begin{aligned} /s/ : /z/ = /p/ : /b/ \quad &\Leftrightarrow \quad -- : -+ = +- : ++ \\ &\Leftrightarrow \quad 00 : 01 = 10 : 11 \\ &\Leftrightarrow \quad aa : ab = ba : bb \end{aligned}$$

En plus d'une notation avec les signes plus et moins, et avec les symboles binaires 0 et 1, nous avons donné une notation avec les deux symboles a et b, car lors de notre formalisation de l'analogie, nous nous occuperons d'analogies entre chaînes définies sur un ensemble de symboles binaires, classiquement notés a et b en théorie des langages (voir page 142). Mais il faut bien voir que ces analogies sont d'un type fort particulier. En fait, elles sont lues comme étant du type :

$$p \wedge q : \neg\, p \wedge q = p \wedge \neg\, q : \neg\, p \wedge \neg\, q$$

Pour l'exemple précédent, p est la propriété d'avoir une articulation d'avant (qui s'oppose à arrière), et q celle d'avoir le trait sourde (pas sonore). Ces analogies suggèrent que la solution de l'équation analogique serait « plus contrainte » que la simple contradiction logique (qui serait $\neg\, p \vee \neg\, q$). Dans le carré analogique, l'« opposé » de $p \wedge q$ est en effet $\neg\, p \wedge \neg\, q$.

Mais déduire, à partir d'une analogie quelconque $A : B = C : D$ que l'on pourrait toujours réexprimer les quatre objets y intervenant de la façon suivante :

$$A \quad / \quad p(A) \wedge q(A) \tag{1.1}$$
$$B \quad / \quad \neg p(B) \wedge q(B) \tag{1.2}$$
$$C \quad / \quad p(C) \wedge \neg q(C) \tag{1.3}$$
$$D \quad / \quad \neg p(D) \wedge \neg q(D) \tag{1.4}$$

est parfois un travail risqué d'abstraction et de codage. Et le caractère nécessaire du codage obtenu pour parvenir à ce système minimal n'est pas toujours démontré (revoir notre interprétation risquée de la phrase de Pascal, p. 29).

Le débat célèbre entre Martinet (1908–1999) et Jakobson a porté précisément sur la nécessité et la réalité de ce type de représentation. Tout d'abord, il est clair que les langues ne possèdent pas nécessairement un nombre de consonnes ou de voyelles qui soit une puissance de 2. Se pose donc le problème des lacunes dans les systèmes. Faut-il se satisfaire de dire qu'une opposition peut n'être pas réalisée partout[82] ? Ensuite, l'universalité des oppositions binaires, voulue par Jakobson, est questionnable. L'opposition française /p-b/, /t-d/ du français en termes de sourde / sonore est bien difficile à trouver en chinois où l'opposition notée par le pinyin p-b t-d (et par p'-p et t'-t dans la notation de l'Ecole française d'Extrême-Orient) est une opposition entre aspirée et non-aspirée, tandis que l'opposition sourde / sonore n'existe pas dans cette langue, pas du moins entre deux phonèmes ayant une articulation par ailleurs identique[83].

En général, et pour souligner cette différence fondamentale d'avec le structuralisme tel que pratiqué par Jakobson sur les systèmes phonétiques, il faut bien reconnaître que, dans une analogie générale, les rapports entre objets ne sont pas nécessairement de l'ordre de la négation.

1.2.6 Les linguistes de la diachronie : le côté diachronique

Nous avons vu plus haut que Saussure avait rétabli de plein droit l'analogie en linguistique en montrant qu'il s'agit d'un phénomène *synchronique*. Conséquemment, son étude serait du ressort de la linguistique synchronique. D'autre part, en tant que phénomène de *parole*, elle devrait relever de la psycholinguistique plutôt que de la linguistique pure. En synchronie, en effet, l'analogie sous-tend l'ensemble des paradigmes flexionnels réguliers, de la conjugaison à la déclinaison, et aussi peut-être une large part de la productivité syntaxique langagière.

Mais un problème surgit : si une forme analogique se remarque dans un état donné de langue, en faisant doublon avec une forme irrégulière, c'est souvent parce qu'elle rétablit l'ordre des associations perturbé par l'ordre des changements phonétiques. Un état donné de langue reflète donc l'histoire selon deux dimensions : aussi bien l'effet des changements phonétiques que l'effet des changements par analogie peut y être constaté. Or ces deux effets ne sont interprétables que si on a la connaissance de deux moments donnés de la langue. Cette mise en évidence est donc du ressort de la linguistique diachronique.

Le travail du grand spécialiste des langues sémitiques du XX^{e} siècle, Kuryłowicz (1895–1978) se place justement dans cette problématique. Alors

[82] À ce propos voir aussi Mańczak, *Z zagadnień językoznawstwa ogólnego*, 1970, p. 166 à 170.

[83] Voir Institut des Langues de Pékin, *Manuel de chinois pratique*, 1995, p.26 et 113.

que, contrairement à tout ce que nous venions dire, on s'attendrait à une étude de l'analogie qui soit synchronique et psycholinguistique, Kuryłowicz, puis Mańczak (??–??), l'abordent du point de vue de la langue et de la diachronie. Kuryłowicz étudie en effet les langues sémitiques[84] dans une perspective historique, afin de dégager l'histoire et les règles régissant les changements analogiques dans ces langues. En particulier, il se consacre au phénomène de l'apophonie dans diverses langues sémitiques, akkadien, arabe et hébreu.

> Il faut rigoureusement distinguer entre apophonie qui relève de la morphologie, et alternance, phénomène purement phonologique. Malgré la transposition de l'alternance en apophonie, elles peuvent exister l'une à côté de l'autre. La disparition de l'alternance n'est pas une condition nécessaire de la genèse de l'apophonie. Ainsi l'apophonie $u : i$ (ar. *iaqtulu* : *ianqatilu*) n'exclut pas en arabe classique l'existence de l'alternance $u : i$, p. ex. *ḥumru*^{n} "rouges" : *bīḍu*^{n} (pour **buiḍu*^{n}) "blancs" . De même, à côté de l'apophonie voyelle brève *(ă^x)* : voyelle longue *(ā^x)*, l'arabe classique continue l'alternance respective en abrégeant $\bar{a}^x$ en $\breve{a}^x$ en syllabe entravée, p. ex. *iaqūlu* : *iaqŭl.* [85]

Sa vision des choses est donc que, si à l'origine de l'apophonie, on a bien l'alternance, celle-là, une fois réalisée, jouit d'une autonomie dans le cadre de la morphologie. Les oppositions existant entre formes ressenties comme participant d'un même paradigme acquièrent une dynamique propre qui s'exprime souvent par l'expansion de l'opposition et donc par création d'un nouveau paradigme. De différents exemples dans les langues indo-européennes, telles que l'allemand, le vieux-slave, le sanscrit ou le latin, il tire plusieurs règles importantes, concernant cette dynamique.

> (I) Un morphème bipartite tend à s'assimiler un morphème isofonctionnel consistant uniquement en un des deux éléments, c'est-à-dire le morphème composé remplace le morphème simple.
>
> (II) Les actions dites « analogiques » suivent la direction : formes de fondation → formes fondées, dont le rapport découle de leurs sphères d'emploi.
>
> (III) Une structure consistant en membre constitutif plus membre subordonné forme le fondement du membre constitutif isolé, mais isofonctionnel.
>
> (IV) Quand à la suite d'une transformation morphologique une forme subit la différenciation, la forme nouvelle correspond à sa fonction primaire (de formation), la forme ancienne est réservée pour la fonction secondaire (fondée).

[84] Et indo-européennes. Le dernier paragraphe de l'article MAŃCZAK, *Tendances générales des changements analogiques*, 1958, p. 420, fait référence à l'ouvrage de Kuryłowicz intitulé *L'apophonie en indo-européen*, 1956, ouvrage effectivement listé dans les publications linguistiques de l'Académie des sciences de Pologne. Nous n'avons hélas pas pu nous procurer cet ouvrage.

[85] KURYŁOWICZ, *L'apophonie en sémitique*, 1961, p.195.

(V) Pour établir une différence d'ordre central la langue abandonne une différence d'ordre plus marginal.

(VI) Le premier et le second terme d'une proportion appartiennent à l'origine à des systèmes différents : l'un appartient au parler imité, l'autre au parler imitant.[86]

Mais dans l'esprit de Kuryłowicz, les règles dégagées sont prises entre le caractère social de la langue, et le caractère individuel de la parole.

> Retournant à notre point de départ nous constatons que l'étendue d'une action « analogique » ne peut être prévue d'avance... L'extension de changements morphologiques est en même temps *externe* (à l'intérieur d'une communauté linguistique) et *interne* (à l'intérieur d'un système grammatical). Car d'une part un système défini est propre à un grand nombre d'individus, d'autre part l'individu représente un point de croisement de plusieurs systèmes (de parlers, dialectes, langues).[87]

Ces réflexions s'inscrivent dans le cadre de l'obsession de la linguistique à affirmer un statut scientifique semblable à celui de la physique ou de la chimie. D'où cette interrogation sur les types de prédictions qu'elle pourrait émettre. Les changements phonétiques, lois naturelles donc tout à fait prédictibles, suivent la pente indiquée par l'économie de la langue, combinaison de facteurs sociologiques imprévisibles par nature. Or, Kuryłowicz propose pour l'analogie un statut comparable du point de vue de la prédiction à celui des changements phonétiques. L'analogie suit des règles strictes, mais leur application est conditionnée par le facteur social :

> Somme toute les choses se présentent de la façon suivante : Il résulte d'un système grammatical concret quelles transformations « analogiques » sont possibles (formules I – V). Mais c'est le facteur social (formule VI) qui décide si et dans quelle mesure ces possibilités se réalisent. Il en est comme de l'eau de pluie qui doit prendre un chemin prévu (gouttières, égouts, conduits) *une fois qu'il pleut.* Mais la pluie n'est pas une nécessité. De même les actions prévues de l'« analogie » ne sont pas des nécessités. Étant obligée à compter avec ces deux facteurs différents la linguistique ne peut jamais prévoir les changements à venir. À côté de la dépendance mutuelle et de la hiérarchie d'éléments linguistiques à l'intérieur d'un système donné elle a affaire à la contingence historique de la structure sociale. Et bien que la linguistique générale penche plutôt vers l'analyse du système comme tel, les problèmes historiques concrets ne trouvent une solution satisfaisante que si l'on tient compte des deux facteurs simultanément.[88]

[86] KURYŁOWICZ, *La nature des procès dits « analogiques »*, 1949, p. 20, 23, 25, 30, 31 et 36, respectivement pour chaque règle. Article repris dans KURYŁOWICZ, *Esquisses linguistiques*, 1961, p. 66 à 86.

[87] KURYŁOWICZ, *La nature des procès dits « analogiques »*, 1949, p. 36.

[88] Ibidem, p. 37.

Certains des travaux de Mańczak se situent à la suite immédiate de cette problématique. Un certain nombre de ses travaux sur les langues indo-européennes trouvent leur origine explicitement dans le type de lois proposées par Zipf[89].

> Une telle loi [valable pour toutes les langues du monde et toutes les époques], a été formulée par Zipf [...], qui a découvert que les phonèmes moins composés étaient plus souvent utilisés que les phonèmes plus composés (par exemple, les consonnes sonores sont dans les différentes langues utilisées à peu près deux fois plus que les sourdes). De cette loi synchronique, Zipf a dérivé une loi diachronique selon laquelle, du moment que la fréquence d'un phonème donné croît de façon significative au-delà de sa fréquence normale, le phonème en question doit évoluer jusqu'à finalement atteindre à nouveau sa fréquence normale.[90]

Mańczak, par l'étude du vocabulaire de plusieurs langues, se propose donc de replacer l'analogie dans le cadre de lois statistiques d'expression plus générales. Et il parvient à énoncer quatre lois de « développement analogique ».

1. Les morphèmes, mots ou groupes de mots ayant le même sens que d'autres morphèmes, mots ou groupes de mots, disparaissent plus vite qu'ils n'apparaissent ;[91]

2. en ce qui concerne a) les morphèmes plus courts et les morphèmes plus longs, b) les mots plus courts et les mots plus longs, c) les groupes de mots plus courts et les groupes de mots plus longs, les seconds remplacent plus souvent les premiers, que le contraire ;[92]

3. en ce qui concerne a) les morphèmes plus courts et les morphèmes plus longs, b) les mots plus courts et les mots plus longs, c) les groupes de mots plus courts et les groupes de mots plus longs, les premiers se conservent plus souvent que les seconds, les premiers conservent un caractère archaïque plus souvent que les seconds, les premiers provoquent plus souvent des changements chez les seconds, que le contraire ;[93]

4. en ce qui concerne les formes plus souvent utilisées et les formes moins souvent utilisées, les premières se conservent plus souvent que les secondes, les premières conservent un caractère archaïque plus souvent que les secondes, les premières provoquent plus souvent des changements chez les secondes.[94]

[89] ZIPF, *Human behavior and the principle of least effort*, 1949. Voir aussi MANDELBROT, *Les constantes chiffrées du discours*, 1968.

[90] MAŃCZAK, *Z zagadnień językoznawstwa ogólnego*, 1970, p. 55.

[91] Ibidem, p. 115.

[92] Ibidem, p. 117.

[93] Ibidem, p. 119.

[94] Ibidem, p. 120.

Quoique l'expression de ces lois semble vague du fait de l'utilisation de termes relatifs de comparaison, Mańczak les appuie sur de multiples comptages et mesures de fréquences réalisés sur de nombreuses langues. Une conséquence directe des règles précédentes est une justification formelle de ce que l'on appelle le « supplétivisme », c'est-à-dire l'utilisation de plusieurs racines dans la flexion de certains mots[95]. Un exemple typique est le verbe français *aller* dont la conjugaison s'appuie sur trois radicaux, *all-*, *v-* et *ir-*, provenant de trois verbes historiques différents, *allāre* <? *ambulāre* (se promener), *vādere* (s'avancer), et *eo, īs, īre* (aller)[96].

De Mańczak, nous retenons pour notre part l'idée particulièrement intéressante que l'analogie entretient certaines relations avec la fréquence d'occurrences de certaines formes. Notre vue sera en fait beaucoup plus simpliste et moins linguistique, car elle provient de considérations pratiques et expérimentales. Pratiques, parce que les équations analogiques produisant des solutions multiples, il nous faudra parfois devoir choisir la « meilleure » solution. Être meilleur signifie répondre à l'impératif de la langue. Or, nos résultats expérimentaux nous inclinerons à retenir simplement, pour effectuer ce choix, une règle élémentaire de sélection de la solution la plus fréquente (p. 249 et 296).

[95]Ibidem, p. 150 à 165.

[96]Le Goffic, *Les formes conjugées du verbe français – oral et écrit*, 1997, p. 39–40.

1.3 Les Contemporains : de la condamnation à la réhabilitation ?

1.3.1 Les structuralistes américains : usage en syntaxe

Nous avons vu l'application du concept d'analogie à la phonologie et à la morphologie. Son action en syntaxe avait été aussi mentionnée par Hermann Paul. Plus tard dans le XX^{e} siècle, Bloomfield (1887–1949) eut aussi l'idée que, de même que l'analogie agit en morphologie, de même, elle pourrait être à l'œuvre en syntaxe. Pour le structuraliste américain, cité par Itkonen :

> ... un auditeur, connaissant les constituants et le modèle grammatical, peut prononcer des [paroles] sans les avoir jamais entendues ; ... Un modèle grammatical est souvent appelé une analogie.[97]

Ce que nous pourrions illustrer par :

$$\textit{il donne un livre à Marie} : \textit{il lui donne un livre} = \textit{elle écrit à Jean} : x$$
$$\Rightarrow \quad x = \textit{elle lui écrit}$$

On peut voir dans la notion qui sous-tend celle, fameuse de commutation, chez Bloomfield, rien d'autre que la notion d'analogie. Et l'on sait le succès remporté par la méthode des commutations dans l'étude des langues amérindiennes. Dans ce type de linguistique structurale, l'analogie a en effet été utilisée avec une efficacité remarquable pour révéler, par des méthodes presque aveugles, c'est-à-dire utilisant à rebours le caractère aveugle de l'analogie pour dégager la structure, les oppositions syntaxiques, et donc les modèles grammaticaux de langues amérindiennes.

1.3.2 Les génératives tes : le rejet

L'histoire que nous avons jusqu'à présent tracée de l'analogie en grammaire puis en linguistique montre clairement que toute la tradition grammaticale utilise de façon implicite ou explicite l'analogie. En particulier, les grands hommes qui ont présidé à la naissance de la science linguistique la considère tous comme une notion fondamentale. De façon exactement opposée, le second courant dominant du XX^{e} siècle, le génération isme, est connu pour récuser explicitement l'analogie. Historiquement, donc, c'est un fait surprenant, car si tous jusqu'à présent s'accordaient sur les limites de l'analogie, personne n'avait pensé à l'éliminer totalement[98]. L'argumentation des génératives tes contre l'analogie tient en deux points. Le premier repose sur des hypothèses qui seront plus tard ruinées, et la seconde repose sur une interprétation, à notre avis trop rapide, des faits linguistiques. Nous allons maintenant nous étendre longuement sur cet argumentaire.

[97] Bloomfield, *Language*, 1933, p. 275.
[98] Voir Itkonen, *Iconicity, analogy, and universal grammar*, 1994.

Hypothèse de l'inné Le courant générativiste pose bien des questions linguistiques en des termes fortement influencés par les résultats obtenus en théorie du calcul dans le début de la seconde moitié du XX^{e} siècle. Ainsi les auteurs de ce courant se focalisent-ils en particulier sur le nombre potentiellement infini d'énoncés productibles par un locuteur. En réponse à cet impératif d'infinitude, Chomsky propose une conception de la langue comme système formel de production et postule l'existence de symboles intermédiaires, dérivés de l'analyse en constituants[99] d'abord proposée par Zellig Harris (1909–1992), puis retravaillés pour la théorie standard. Certains linguistes s'interrogent à la base sur la réalité psychologique de ces catégories, voire sur leur définissabilité. Le génértivisme sera lui-même confronté au problème du raffinement nécessaire de ces classes morpho-syntaxiques et réintroduira par la fenêtre la θ-théorie, après avoir évacué par la porte les bons vieux cadres argumentaires. Nous reviendrons plus bas sur cette question des classes morpho-syntaxiques lorsque nous proposerons un système formel sans symboles intermédiaires (p. 171).

La conception même de la langue comme système formel est bien sûr mise en doute par les contradicteurs du génértivisme. D'aucuns pensent que Chomsky est tout simplement passé à côté d'une opération psychologique, expliquant la productivité, qui serait fondée sur un processus créateur, et non sur un processus fondamentalement répétitif, les grammaires formelles. Ainsi Mounin :

> [...] le problème de savoir pourquoi, et d'abord surtout comment, tout locuteur est apte à produire et décoder un nombre infini de phrases qu'il n'a jamais entendues auparavant, s'éclaire si, au lieu de s'enfermer dans une conception de l'apprentissage en termes étroitement skinnériens de répétition, on se demande : qu'est-ce, en linguistique, qu'une forme jamais entendue auparavant ? Là-dessus, Saussure a bien montré comment l'enfant fabrique la forme **viendre* qu'il n'a jamais entendue : par l'application de ce que le maître de Genève appelait modestement, et peut-être très suggestivement la règle de la quatrième proportionnelle :
>
> peindrai : peindre :: viendrai : *viendre
>
> que l'on doit lire ainsi (quand on a fait les vieilles mathématiques en classe de quatrième) : « peindrai » [que j'ai entendu] est à « peindre » [que j'ai entendu] comme « viendrai » [que j'ai entendu] est à « viendre » [que je n'ai jamais entendu encore, mais dont le système de la langue me permet de postuler l'existence]. Très suggestivement, avons-nous dit, parce que c'est peut-être toute une épistémologie génétique du type de celle de Piaget, par exemple (et non pas un innéisme trop commode) qui pourra nous aider à comprendre comment l'enfant acquiert l'aptitude à de telles opérations mentales *logiques*.[100]

Itkonen attaque, lui, les conceptions de Chomsky sous l'angle de la procédure :

[99] Voir MOUNIN, *Clefs pour la linguistique*, 1968, p. 119–120.
[100] Ibidem, p. 129.

> [...] selon Chomsky, le seul moyen de rendre compte de l'analogie est de montrer que « c'est, en première instance, un concept inapproprié ». Mais l'argument devrait être renversé. L'analogie est une procédure ; les grammaires chomskiennes sont des descriptions de la compétence, donc des descriptions de la structure, pas des procédures ; par conséquent les grammaires chomskyennes sont, en permière instance, inappropriées pour rendre compte de l'analogie.[101]

D'autres, comme Pullum ont une attitude plus radicale : la question de l'infinitude des énoncés n'est pas, pour eux, une question linguistiquement pertinente :

> Une réponse à (ii) [(i) les gens savent tacitement (et apprennent dans leur enfance) quelles phrases sont grammaticales et font sens dans leur langue ; (ii) ils possèdent (et aquièrent) une telle connaissance même à propos de nouvelles phrases, ...] est que, tout simplement, les locuteurs généralisent ou analogisent à partir de cas familiers.[102]

L'hypothèse de la langue comme système formel implique pour les génératives tes l'existence d'une faculté grammaticale innée chez l'Homme qui ne serait que paramétrée lors de l'apprentissage de la langue maternelle. Ce paramétrage ne pourrait se faire par induction. Itkonen résume cet aspect de la pensée de Chomsky dans les termes suivants.

> Chomsky (1957) ayant déclaré qu'il n'y avait pas de procédure de découverte en grammaire, c'est-à-dire pas de méthode permettant de dériver des grammaires à partir des données de langue ; et parce qu'une telle méthode aurait dû avoir un caractère analogique (ou inductif), il s'ensuivait (ou plutôt il semblait s'ensuivre) que l'analogie (ou l'induction) n'était d'aucune utilité en linguistique.[103]

Fondamentalement, c'est la quantité réduite de données qui ferait qu'aucune procédure analogique ne pourrait être mise en œuvre. Nous essaierons, mais sans conclusion claire, d'examiner cette hypothèse lors de nos expérimentations (voir plus bas, p. 276).

Du point de vue argumentaire, l'hypothèse de l'absence d'induction est longuement réfutée par Itkonen dans un article sur l'iconicité[104]. Ses arguments découlent d'un grand nombre d'expériences de psychologie qui montrent que l'iconicité, ou l'appariement, ou l'analogie, jouent bien un rôle lors de processus d'apprentissage.

Un autre type de réfutation est donné par toute une série d'expérimentations effectuées avec des robots équipés d'organes de vision (des

[101] ITKONEN & HAUKIOJA, *A rehabilitation of analogy in syntax (and elsewhere)*, 1997, p. 136.

[102] PULLUM, *Generative grammar*, 1999. Voir encore PULLUM, *Model-theoretical syntax*, 2001.

[103] ITKONEN & HAUKIOJA, *A rehabilitation of analogy in syntax (and elsewhere)*, 1997.

[104] ITKONEN, *Iconicity, analogy, and universal grammar*, 1994

caméras), et d'un processus de catégorisation (classification des images et extraction des traits saillants). Les robots se livrent à un « jeu de langue » inspiré de Wittgenstein : ils observent des scènes se déroulant sous leurs yeux et transmettent des signaux à un ou plusieurs autres robots similaires[105]. Le robot locuteur est compris du robot auditeur si les traits saillants de la scène décrite par le locuteur sont bien identifiés par l'auditeur ; en cas d'échec, l'auditeur fait un effort de collaboration pour s'adapter. La seule contrainte sur les messages est qu'ils sont de simples concaténations de symboles, sans plus. La conclusion de telles expériences est :

- qu'il y a création d'un ordre spécifique (aléatoire) dans l'énoncé des traits saillants dans les messages ;
- que cet ordre est adopté par les autres robots, afin de réussir dans le « jeu de langue » ;
- qu'il y a création de nouveaux symboles abrégeant des situations déjà vues ;
- qu'il y a disparition progressive (érosion) des schémas trop utilisés.

Du point de vue de la syntaxe, ces expériences montrent donc clairement qu'il n'est nul besoin de supposer de quelconques paramètres qu'il faudrait prérégler pour qu'un robot adopte la syntaxe du groupe dans lequel il est placé, à l'exact opposé des théories professées par Chomsky.

Hypothèse du hors-contexte De l'idée des paramètres découle l'hypothèse du hors-contexte. Parmi tous les systèmes formels possibles seuls un certain type serait adéquat pour formaliser la langue. Chomsky fait l'hypothèse que les grammaires hors-contexte seraient les meilleurs candidats. La conséquence directe est que le modèle hors-contexte deviendrait universel, propre au genre humain. Mais l'hypothèse semble bien biaisée par l'influence de l'anglais. De nouveau citons Itkonen dans un passage où il discute des différentes conceptions de l'universel en linguistique :

> Chomsky pratique, lui, une sorte d'« universalité de la grammaire de l'anglais », en prenant la syntaxe de sa langue maternelle pour un composant inné du cerveau humain.[106]

Ce reproche est aussi formulé par Mel'čuk dans l'introduction de son livre sur la syntaxe en dépendance dans un paragraphe intitulé de façon significative *l'anglais, langue maternelle des pères fondateurs de la syntaxe moderne*. Il y souligne que :

[105] Steels, *Origin of syntax in visually grounded robotic agents*, 1985. Pour ce qui est de l'adoption du vocabulaire par un groupe de robots, d'autres expériences reproduisent des situations de « bilinguisme » ou de fusion de vocabulaires, voir Steels, *Language learning and language contact*, 1997.

[106] Itkonen, *Iconicity, analogy, and universal grammar*, 1994

> Cependant l'anglais est très exotique dans son utilisation des constituants comme unique moyen d'expression en syntaxe, c'est-à-dire comme unique moyen de codage des structures syntaxiques en phrases.[107]

Nous devons donc nous arrêter ici pour présenter la classification des langages formels selon Chomsky-Schützenberger. Nous y reviendrons pour la contester sous l'angle de la « naturalité ». Elle est illustrée dans la figure 1.3. Nous n'avons mentionné ici que les trois grandes familles de langages formels qui occupent les linguistes et les informaticiens linguistes. Cette classification repose sur une répartition en fonction du système de dérivation utilisé. Plus précisément, elle repose sur une complexité croissante des chaînes apparaissant dans les dérivations. En ce sens, certains la jugent comme peu pertinente pour parler de la langue.

Tableau 1.3: Classification des langages formels

type de règles	$\begin{cases} A \to Ba \\ A \to a \end{cases}$	$\begin{cases} A \to BC \\ A \to a \end{cases}$	$w \to w' \ / \ \lvert w\rvert \leq \lvert w'\rvert$
dénomination	réguliers $\subset$	hors-contexte $\subset$	sous-contexte
	rationnels $\subset$	algébriques	
exemples	$\{a^n / n > 0\}$	$\{a^n b^n / n > 0\}$	$\{a^n b^n c^n / n > 0\}$ $\{a^n b^n c^n d^n / n > 0\}$ $\{a^n b^m c^n d^m / n, m > 0\}$

Une première remarque. Cette classification repose sur un présupposé très fort, celui de l'existence de symboles qui, n'apparaissant pas dans les chaînes terminales, sont appelés pour cette raison même *symboles non-terminaux.* Pour les linguistes génératives, ces symboles sont les parties du discours ou classes syntaxiques, comme `N` pour nom, `V` pour verbe, etc., et les symboles des constituants tels que `PH` pour phrase, `GN` pour groupe nominal, etc. Le problème de la définition des parties du discours, de leur nombre, de leur universalité, c'est-à-dire de leur pertinence pour toutes les langues, etc., se pose donc. On sait le débat non clos et l'on pourrait le rapprocher de celui sur l'universalité des oppositions en phonologie (p. 64). Notre intention est de montrer que l'on peut éviter le débat sur les parties du discours, comme le font par ailleurs certaines grammaires de dépendance, en les éliminant tout simplement ! Nous proposerons en effet, nous l'avons déjà annoncé, un système formel capable d'engendrer des phrases de la langue sans faire appel à elles (p. 171).

Dans un deuxième temps, nous en appelons au bon sens, et remarquons simplement que cette classification n'est pas très intuitive. Considérons en effet les exemples classiques illustrant les trois grandes familles, à savoir les

[107] MEL'ČUK, *Dependency syntax: Theory and practice*, 1988, p. 4.

trois langages formels $\{a^n/n>0\}$, $\{a^nb^n/n>0\}$ et $\{a^nb^nc^n/n>0\}$. Certes on peut concevoir que le premier et les deux derniers soient classés différemment, car l'utilisation d'une pluralité de symboles engendrerait nécessairement un changement de « difficulté ». Cependant, pourquoi les deuxième et troisième seraient-ils de « difficulté » différente ? Notons de plus que la troisième famille est en fait illustrée par l'infinité des langages du type :

$$\{a_1^n a_2^n \dots a_m^n / n \geq 1\}$$

avec $m \geq 2$. La coupure en trois (l'élémentaire, le simple et le complexe ?) peut donc sembler hautement anti-intuitive. Pour nous, c'est l'angle d'attaque de la classification par type des règles qui est fautif. Existe-t-il un autre angle sous lequel tous ces langages seraient rangés dans la même famille, et, qui plus est, une famille de langages simples ? La réponse que nous apporterons sera évidemment positive, et ce bien sûr, grâce à l'analogie (théorème 22, p. 185 et discussion, p. 186).

On sait que l'hypothèse du hors-contexte a été critiquée « par les deux bouts ». Par le « bout » des langages réguliers, dans le fameux article de Gross de 1979 où il plaide qu'une description d'une complexité supérieure, le hors-contexte, ne saurait avoir la finesse nécessaire pour la couverture en grandeur réelle de phénomènes linguistiques réels comme par exemple, la rection des verbes français, dont la description se satisfait fort bien d'une complexité moindre, le régulier. Certes, la contrepartie à payer est la taille énorme des descriptions nécessaires[108]. Cette critique a trouvé un écho pratique dans tous ces travaux sur la génération, par limitation, à partir de grammaires hors-contexte, d'automates réguliers gigantesques mais exploitables grâce à la puissance des ordinateurs actuels[109], et par la mesure de leur couverture de corpus de très grande taille, qui se révèle acceptable.

Par le « bout » du sous-contexte par la découverte de deux structures similaires, l'une dans la morphologie du Bambara[110], l'autre dans la syntaxe de la variante zurichoise du suisse-allemand[111].

Dans ces deux exemples, le cœur de la démonstration repose sur l'apparition du schéma $a^nb^mc^nd^m$, représentatif d'un langage formel sous-contexte : $\{\, a^nb^mc^nd^m / \;\; n \in \mathbb{N}, m \in \mathbb{N} \,\}$. Ce schéma peut être vu comme l'intersection de la langue en question, bambara ou variante zurichoise du suisse-allemand, avec un langage formel régulier, $\{\, a^pb^qc^rd^s / \;\; (p,q,r,s) \in \mathbb{N}^4 \,\}$, couramment noté $a^*b^*c^*d^*$. Comme l'intersection d'un langage formel régulier avec un langage formel régulier (*resp.* hors-contexte) est un langage formel régulier (*resp.* hors-contexte), on en conclut que ni le bambara, ni le dialecte en question du suisse-allemand ne peuvent être réguliers ou hors-contexte. Il sont au moins sous-contexte.

$$\mathcal{L} \cap a^*b^*c^*d^* = \{\, a^nb^mc^nd^m, \;\; n \in \mathbb{N}, m \in \mathbb{N} \,\} \quad \text{est sous-contexte}$$
$$\Rightarrow \;\; \mathcal{L} \quad \text{est au moins sous-contexte.}$$

[108] Gross, *On the failure of generative grammar*, 1979.
[109] Voir par exemple Black, *Finite state machines from feature grammars*, 1989.
[110] Culy, *The complexity of the vocabulary of Bambara*, 1985
[111] Shieber, *Evidence against the context-freeness of natural language*, 1985

wulu	*nyinina*^{n}	*filèla*^{m}	*o wulu(nyinina)*^{n} *(filèla)*^{m}
chien	qqn cherchant	qqn regardant	quiconque

« quiconque (regarde quelqu'un qui)^m (cherche quelqu'un qui)$^{n-1}$ cherche un chien »

(d'chind)^n	*(em Hans)*^{m}	*es huus haend wele*	*laa*^{n}	*hälfe*^{m}	*aastriche*
les enfants-ACC	Hans-DAT	la maison-ACC avoir voulu	laisser	aider	peindre

« [...que nous] ayons voulu (laisser les enfants)^n (aider Hans à)^m peindre la maison »

Figure 1.1: Réfutation de l'hypothèse du hors-contexte : exemples linguistiques

Nous ne discutons pas la validité d'un tel argument[112], mais nous nous resservirons du langage formel $\{ a^n b^m c^n d^m / \ n \in \mathbb{N}, m \in \mathbb{N} \}$ dans notre plaidoyer en faveur de l'analogie. En effet, nous montrerons que, vu comme un langage formel reposant sur l'analogie, le langage $\{ a^n b^m c^n d^m / \ n \in \mathbb{N}, m \in \mathbb{N} \}$ est relativement simple (p. 186). Ce que nous prétendrons alors, c'est que le retour à l'analogie constitue un retour au bon sens :

- il supprime des barrières anti-intuitives dans la mesure où les exemples de langages formels illustrant les trois grandes familles de langages formels de la classification de Chomsky-Schützenberger ne sont plus artificiellement séparés ;
- le bambara ou la variante zurichoise du suisse-allemand n'ont pas à apparaître comme des langues extrêmes ou fort particulières dans leur structure, puisque le langage formel qui les montrait au moins sous-contexte n'est qu'un langage de chaînes analogiques relativement simple.

Surproduction Le second point de l'argumentation de Chomsky repose sur un argument pour le moins spécieux. Il se borne à constater, chose que personne n'a jamais contestée, qu'il est fort facile de produire des phrases grammaticalement incorrectes par analogie. Certes, on a :

$$\textit{il donne un livre à Marie} : \textit{il lui donne un livre} = \textit{elle écrit à Jean} : x$$
$$\Rightarrow \quad x = \textit{elle lui écrit}$$

Mais on a aussi :

[112] Donnant un exposé à l'université Humboldt à Berlin, c'est sans grande surprise que nous nous sommes vu critiquer cet exemple par une Suissesse de Zürich ! Que dire aussi de la « preuve » donnée dans MICHAELIS & KRACHT, *Logical aspects of computational linguistics*, 1997, que les langues humaines ne sauraient être semi-linéaires, car le vieux-géorgien, langue que personne ne parle plus, ne le serait pas ?

$$il\ donne\ un\ livre\ à\ Marie : il\ lui\ donne\ un\ livre = elle\ vit\ à\ la\ ville : x$$
$$\Rightarrow \quad x = {}^{*}elle\ lui\ vit$$

Dès lors, l'obtention de phrases agrammaticales étant possible par analogie, cette opération ne saurait constituer un outil de jugement pour la grammaticalité. Chomsky va aussi plus loin en montrant qu'une analogie valable grammaticalement sur le plan formel peut livrer des sens qui ne sont pas analogiques :

$$\begin{matrix} Max\ peint \\ le\ mur\ en : \\ rouge \end{matrix} \begin{matrix} Max\ peint\ le \\ mur\ rouge \end{matrix} = \begin{matrix} Max\ \ voit \\ le\ mur\ en : \\ rouge \end{matrix} \begin{matrix} Max\ voit\ le \\ mur\ rouge \end{matrix}$$

À notre avis, dire que, puisque l'on peut obtenir des phrases incorrectes par analogie, c'est que l'analogie n'est pas à l'œuvre en syntaxe et qu'elle ne peut donc servir de critère de grammaticalité, est une transposition de la logique à la syntaxe. Car si l'analogie n'est pas une méthode de raisonnement logique sûre au même titre que, par exemple, le *modus tollens*, son champ d'application n'est pas nécessairement en relation avec la logique. On peut se gausser avec Zemb de l'aveuglement de Chomsky à refuser son aristotélisme[113], qui peut se représenter, ô ironie !, par l'analogie suivante :

$$agrammatical : grammatical = faux : vrai$$

qui énonce que : « l'agrammatical est au grammatical ce que le faux est au vrai ». Il ne s'agit bien sûr de rien d'autre que du programme aristotélicien d'interprétation logique des énoncés de langue :

> La démarche [des générativistes] trouve l'une de ses sources dans les travaux du logicien Post qui, reprenant la conception "générative" de la logique, emprunte l'ensemble du vocabulaire grammatical (règle, mot, phrase, langage) pour exposer diverses interprétations possibles du système formel de départ. Au lieu de renoncer aux conceptions les plus banales de la grammaire, la théorie linguistique reprend l'analogie qui va de la grammaire à la logique pour postuler que, parmi les interprétations possibles d'un système formel, figure l'articulation grammaticale des langues que, par anglicisme, on appelle « naturelles ». A cette différence près que, si chez Post le rapprochement entre logique et langage ne dépasse pas l'analogie utile à la description des opérations logiques, chez Chomsky, au contraire, il y a identité structurale entre les deux systèmes. La notion de « correction » grammaticale en découle directement, non par souci normalisateur mais en vertu du principe de générativité logique (qui définit le vrai et le faux), donc grammaticale (qui distingue le correct de l'incorrect). On y reconnaîtra la bivalence qui se situe au fondement de la linguistique chomskyenne comme

[113] ZEMB, *Vergleichende Grammatik Französisch-Deutsch – Comparaison de deux systèmes – Teil 1*, 1978, p. ??.

étant celle de la logique classique puisque, comme l'affirme Chomsky, la logique des langues "est" la logique classique (avec des variables, c'est-à-dire le calcul des prédicats), et non une quelconque logique intentionnelle (sans variables).[114]

De nouveau, que l'on puisse produire des énoncés agrammaticaux par analogie n'a jamais été contesté par personne ; tout le monde s'accorde sur ce point, et c'était précisément notre point lorsque nous avons parlé du caractère aveugle de l'analogie dans notre introduction (p. 27). Notre propos sera de renverser l'argument, et de faire de la surproduction une bénédiction. En effet, dans notre conception du fonctionnement de la langue, la surproduction due à l'analogie est la condition nécessaire pour que l'adéquation au monde existe. Sans elle, on ne saurait construire des intersections d'espaces analogiques, où l'un de ces espaces seulement est celui de la langue. Cette vue rejoint celle d'Itkonen qui plaide que l'argument de la surproduction s'écroule si l'on fait fonctionner l'analogie simultanément sur les structures syntaxiques et les structures sémantiques[115].

1.3.3 Retour de l'analogie ?

Heureusement, avec l'effondrement scientifique (annoncé dès 1979, comme nous l'avons vu, par Maurice Gross) et la dégénérescence sociologique (Milner *dixit*[116]) du « programme générativiste » à la suite de la réfutation de ses principaux piliers idéologiques qu'étaient l'hypothèse de l'inné et l'hypothèse du hors-contexte, battue en brèche par des données linguistiques, l'analogie peut revenir sur le devant de la scène.

Notre espoir est d'apporter une humble contribution à l'étude de ce processus psychologique qu'est l'analogie, en le formalisant, tout au moins pour ce qui est de son application aux chaînes de symboles, puis de l'utiliser pour définir une famille de langages formels que nous appellerons langages de chaînes analogiques, et enfin d'appliquer tous ces résultats au traitement automatique des langues.

[114] Ghils, *Langage et contradiction*, 1998, § 4 (Les paradoxes des théories linguistiques), deuxième alinéa.

[115] Itkonen & Haukioja, *A rehabilitation of analogy in syntax (and elsewhere)*, 1997, p. 150–156.

[116] Milner, *Introduction à une science du langage*, 1989, p. 17.

Chapitre 2

L'analogie dans les sciences et techniques

2.1 En mathématiques

Nous avions mentionné dans notre introduction que l'un des actes révolutionnaires de la science grecque serait la découverte des proportions. Nous avions aussi mentionné que certains vont jusqu'à dire que la science grecque serait essentiellement une science des proportions, qui plus est des proportions égales, en un mot, des analogies. Nous allons ici reparler brièvement de ces concepts.

2.1.1 Règle de trois

Dans la tradition scolaire, la règle de trois constitue l'exemple par excellence de l'analogie et le vocabulaire de l'analogie, ou proportion, avec ses égalités et ses rapports, est tiré d'elle. Ce n'est pas une surprise, puisque, historiquement comme on l'a vu avec Euclide, l'analogie n'est guère plus que le nom de la règle de trois pour les mathématiciens grecs. La règle de trois permet de définir la division, opération complexe, à partir de la multiplication, opération plus simple.

$$\forall (A,B,C,D) \in \mathbb{Q}^{*4},\ \ A \times D = B \times C \ \Leftrightarrow\ \frac{A}{B} = \frac{C}{D}$$

Mais la règle de trois nous intéresse particulièrement à trois titres. À un premier titre, parce qu'elle est, nous semble-t-il, la seule réalisation scientifique de l'analogie, scientifique au sens où elle énonce le vrai. Nous avons déjà vu plus haut (p. 76) que cette incapacité à dériver le vrai était l'un des principaux handicaps de l'analogie.

À un deuxième titre, parce que la règle de trois énonce une équivalence d'expression entre division et multiplication. Un parallèle imposé aux chaînes de symboles, induit le désir de reformuler les rapports entre chaînes de symboles à l'aide d'une opération[1], notée par exemple $\star$, telle que :

[1] Une telle opération exprimerait en fait la contiguïté entre chaînes.

$$\forall (A,B,C,D) \in \mathcal{V}^{*4},\ \ A \star D = B \star C \ \Leftrightarrow \quad A : B = C : D$$

Au cours de notre recherche, idéalement, nous avons longtemps cherché si l'analogie entre chaînes de symboles ne pourrait pas s'exprimer de cette même manière[2]. On verra plus bas (p. 257) que notre première intuition avait été de penser que, à tout le moins, une implication de la droite vers la gauche était donnée par la distance munie des trois opérations d'insertion, de suppression et d'échange avec un même poids. En fait, des exemples linguistiques ont montré qu'il n'en était rien. En l'état actuel de nos travaux, nous ne sommes qu'en mesure de présenter trois résultats allant dans la direction d'une telle expression (p. 149, 144 et 156). Mais cela n'épuise pas la question.

À un troisième titre, parce que la règle de trois n'est pas seulement l'énoncé d'une proportion, elle est comprise habituellement comme la résolution d'une équation analogique, c'est-à-dire de la façon suivante :

$$\forall (A,B,C) \in \mathbb{Q}^{*3},\ \ \frac{A}{B} = \frac{C}{x} \ \Leftrightarrow\ x = \frac{B \times C}{A}$$

Il est à noter que nous avions rencontré cette idée chez Hermann Paul (p. 57) et chez Saussure (p. 61).

2.1.2 Nombre d'or

Nous avions vu qu'Aristote insistait sur la différence entre proportion continue et proportion discrète (voir p. 39), distinction reprise par Henrion dans son commentaire d'Euclide (voir p. 42). Nous avions justifié cette distinction par rapport au syllogisme, où un seul moyen existait, contrairement à l'analogie dans laquelle deux moyens interviennent, B et C dans $A : B = C : D$. La proportion continue donne lieu à une notion bien connue, celle de nombre d'or. Ce nombre est défini comme la grandeur d'un tout tel que

> le rapport du tout à la plus grande des parties est égal au rapport de la plus grande à la plus petite des parties.

C'est-à-dire, en posant la plus petite des parties comme l'unité et en appelant x la plus grande partie :

$$\frac{1+x}{x} = \frac{x}{1}$$

[2]Remarque due à Christian Boitet : la seule découverte d'une opération $\star$ possible ne dira pas forcément tout de l'analogie entre chaînes. Sur les fractions entre nombres, on sait bien que l'équivalence donnée plus haut ne saurait être écrite à la légère : $3 \times 0 = 0 \times 2$ ne permet pas d'écrire $\frac{3}{2} = \frac{0}{0}$. Avant de passer aux rapports, il faut bien vérifier certaines propriétés. De même, sur les chaînes de symboles, la vérification d'une égalité $A \star D = B \star C$, pour une opération $\star$ même bien choisie, ne permettra peut-être pas forcément de passer directement à l'analogie $A : B = C : D$. Certaines propriétés devront peut-être encore être énoncées et vérifiées avant de pouvoir écrire l'analogie.

On reconnaît là une règle de trois en tous points semblable à celle vue plus haut avec $A = 1 + x$, $B = x$, $C = x$ et $D = 1$. Mais il s'agit aussi d'une analogie continue car on a $C = B$. La résolution de l'équation $x^2 = x + 1$ donne les deux solutions $\frac{1}{2} \times (1 + \sqrt{5}) = 1,61903399\ldots$ et $\frac{1}{2} \times (1 - \sqrt{5}) = -0,61903399\ldots$ Euclide appelle *partage en moyenne et extrême raison* la première de ces solutions[3]. On l'appelle maintenant *nombre d'or*, car il a été admis qu'elle possède des valeurs esthétiques certaines. Rappelons pêle-mêle l'étude de ces proportions par Vitruve à des fins d'architecture, la fameuse analyse du célèbre tableau des *Bergers d'Arcadie* de Poussin ou encore la remise en vogue du nombre d'or en architecture par Le Corbusier dans un système de proportions appelé *modulor*.

2.1.3 Suites arithmétiques et géométriques

On sait que la suite de Fibonacci entretient un lien particulier avec le nombre d'or. Il s'agit d'une suite de nombres tels que chacun soit la somme des deux précédents. En commençant avec 0 et 1, on obtient :

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots$$

La forme générale d'un élément u_{n+2} de cette suite est donc :

$$u_{n+2} = u_n + u_{n+1}$$

et la limite du quotient $\frac{u_{n+1}}{u_n}$ est le nombre d'or.

Dans la règle de trois, le rapport est confondu avec la division et c'est aussi de division dont nous avons parlé à propos de la suite de Fibonacci. Or il est important de remarquer que si Euclide ne parle que de division, ses commentateurs, dont Henrion, ont bien compris qu'on pouvait généraliser et qu'il pouvait y avoir diversité des rapports c'est-à-dire plusieurs sortes de proportionalités. Henrion énumère l'arithmétique, la géometrique et l'harmonique (voir citation, p. 39). Évidemment vient tout de suite à l'esprit le lien avec les suites respectives, arithmétiques, géométriques et harmoniques :

$$\begin{aligned} u_{n+1} &= u_n + a \\ u_{n+1} &= u_n \times a \\ u_{n+2} &= u_{n+1}/u_n \end{aligned}$$

Laissons de côté les suites harmoniques. De façon remarquable, l'expression générale des suites arithmétiques et géométriques prend la forme analogique suivante :

$$u_1 : u_0 = u_{n+1} : u_n$$

Il est alors intéressant d'examiner la notion de rapport pour chacun de ces types de suites :

[3]Voir à l'adresse URL : `http://chronomath.irem.univ-mrs.fr/chronomath/Fibonacci.html` (consultée le 28 août 2001).

- suite arithmétique : le rapport est la différence $u_1 - u_0 = u_{n+1} - u_n = a$, où a est appelée la raison arithmétique. On a : $u_n = a \times n + u_0$;
- suite géométrique : le rapport est la division $\frac{u_1}{u_0} = \frac{u_{n+1}}{u_n} = a$ où a est appelée la raison géométrique. On a : $u_n = a^n \times u_0$;

On remarquera, et c'est historiquement vrai, que le vocabulaire des suites utilise le mot raison dans son sens latin de rapport.

La forme $u_1 : u_0 = u_{n+1} : u_n$, ou plutôt, sa forme équivalente $u_0 : u_1 = u_n : u_{n+1}$, donne, en quelque sorte, une impulsion de récurrence. Cette seconde forme est caracterisée depuis longtemps, puisque ce n'est rien d'autre qu'une analogie continue selon Aristote. On a là un rapprochement avec la suite de Fibonacci où le terme u_{n+2} est obtenu à partir des termes u_n et u_{n+1}. Dans une analogie continue aussi, le terme u_{n+2} est obtenu à partir de u_n et u_{n+1} par l'équation analogique $u_n : u_{n+1} = u_{n+1} : x$. On a donc bien un schéma analogique de récurrence dont le patron est l'analogie continue.

La généralisation se fait par l'ajout automatique du rapport qui vient logiquement ensuite.

$$u_0 : u_1 \;=\; u_1 : u_2 \;=\; u_2 : u_3 \;= \ldots = \; u_n : u_{n+1} \;=\; u_{n+1} : u_{n+2} \;= \ldots$$

Il s'agit en fait d'une succession d'analogies continues qui se chevauchent. Bien sûr, pour rendre compte de cette écriture, il est nécessaire d'introduire une quantification sur n. Mais on voit bien que l'analogie continue contient en germe la récurrence pour la raison structurelle que l'on peut faire se chevaucher deux analogies. Dès lors, dans la succession de rapports précédents, en retenant le premier rapport et en en prenant un autre plus loin dans la succession, on peut écrire $u_0 : u_1 = u_n : u_{n+1}$ quel que soit n. Ce schéma est identique à celui des suites arithmétiques ou géométriques.

Tous ces rapprochements entre analogie, c'est-à-dire proportion, et suites de nombres, ne sont pas gratuits. Le lien avec les suites dévoile le lien que l'analogie continue entretient avec la notion d'induction, qui rejoint le caractère créateur de l'analogie (p. 25). Lors de notre tentative de formalisation, et parce que nous nous intéresserons aux langages illustrant la classification de Chomsky-Schützenberger, nous serons amené à tester nos hypothèses sur des exemples formels (voir p. 142) tels que :

$ab : aabb = aaaaaaabbbbbbb : aaaaaaaabbbbbbbb$
$abc : aabbcc = aaaaaaabbbbbbbccccccc : aaaaaaaabbbbbbbbccccccccc$
$ab : abab = abababababab : abababababababab$

qui peuvent se généraliser et se noter :

$a^1 b^1 : a^2 b^2 = a^n b^n : a^{n+1} b^{n+1}$
$a^1 b^1 c^1 : a^2 b^2 c^2 = a^n b^n c^n : a^{n+1} b^{n+1} c^{n+1}$
$(ab)^1 : (ab)^2 = (ab)^n : (ab)^{n+1}$

Toutes ces expressions relèvent de la forme générale :

$$u_1 : u_2 = u_n : u_{n+1}$$

que nous venons de voir pour les suites arithmétiques et géométriques. C'est précisément de ce rapprochement que nous ferons naître la notion de dérivation analogique (p. 169), et que nous tirerons la définition de langage de chaînes analogiques (p. 169), dont le type élémentaire (p. 172) et paresseux (p. 174) revêtira précisément la forme d'une suite. Dans ce cas particulier (p. 174), on se donnera en effet seulement l'élément de départ u_0 et la « raison » sous la forme $u_0 : u_1$, pour engendrer une infinité éventuelle d'objets de la même manière qu'une suite arithmétique ou géométrique définit une infinité éventuelle de nombres. On peut voir dans le nombre éventuellement infini des objets produits une conséquence du caractère aveugle de l'analogie (p. 27), parce que les objets produits le sont par des moyens purement formels.

2.2 En psychologie

Nous nous tournons maintenant vers les recherches effectuées en intelligence artificielle à propos de l'analogie. En fait, notre propos sera relativement court, ce qui étonnera pourtant quiconque ferait une recherche documentaire dans ce domaine avec le mot-clé analogie. L'explication en est simple : sous l'influence de l'anglais, l'analogie dont il est question en intelligence artificielle n'est pas celle dont nous parlons. Le terme y est utilisé dans son acception « vulgaire » (Deleuze *dixit*, voir citation, p. 49). La mécompréhension du sens originel induit une acception trop large et plus psychologique[4]. Cependant, le domaine d'application de ce travail étant le traitement automatique des langues, nous ne pouvions passer totalement sous silence les travaux sur l'analogie dans ce domaine particulier de l'informatique qu'est l'intelligence artificielle.

Commençons par les travaux à cheval sur la psychologie et l'intelligence artificielle. Un article de 1983[5] fait date. Il s'agit d'une tentative de modélisation de phrases du genre « *un atome est comme le système solaire* », à des fins d'application à l'intelligence artificielle. Dans cet article, et à sa suite cette acception sera reprise dans tout le domaine de l'intelligence artificielle, de telles phrases sont appelées « analogies ». Dans ces phrases, deux domaines sont mis en correspondance, d'où la nécessité d'une modélisation des domaines.

domaine astronomique		**domaine atomique**
`soleil`	$\xrightarrow{f}$	`noyau`
`planète`	$\xrightarrow{f}$	`électron`

De plus, des propriétés représentées par des propositions logiques, des formules, etc. sont transférées d'un domaine à l'autre, et leur nombre détermine en quelque sorte la qualité de l'analogie.

[4]Voir Hoffman, *Monster analogies*, 1995.

[5]Gentner, *Structure mapping: A theoretical model for analogy*, 1983

domaine astronomique		**domaine atomique**
`attire(soleil, planète)`	$\xrightarrow{f}$	`attire(noyau, électron)`
`plus_lourd(soleil, planète)`	$\xrightarrow{f}$	`plus_lourd(noyau, électron)`

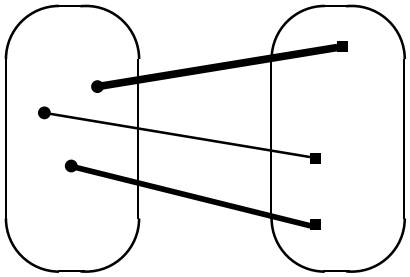


Figure 2.1: Modèle de la métaphore selon Gentner : plusieurs liens existent, certains plus forts que d'autres

Une représentation succinte du modèle de Gentner est donnée dans la figure 2.1. Cependant, appeler analogies, comme Gentner le fait, des phrases du genre « *un A est comme un B* », est critiquable. Rappelons seulement que les définitions d'Euclide et d'Aristote requièrent quatre objets pour une analogie (p. 41 et 39). Ici, on en n'a que deux. D'autres[6] heureusement reconnaissent bien dans de telles phrases, à la suite des définitions d'Aristote (voir p. 40), des *métaphores* dont la validité repose sur des phrases du genre : « *A est à B ce que C est à D* », phrases auxquelles on réserve précisément le nom d'*analogie*. Ce qui caractérise les premières phrases, c'est que des propriétés sont transférées. Étymologiquement, comme nous l'avons déjà vu (p. 42) ce sont donc bien des métaphores. Dans les phrases du second type, on a bien une égalité de rapport entre quatre termes, et donc une analogie. Pour reprendre l'exemple précédent, il est en effet clair que la métaphore « *un atome est comme le système solaire* » repose sur l'analogie « *les électrons sont au noyau atomique ce que les planètes sont au soleil* ».

À la suite de Gentner, la réponse de la communauté de l'intelligence artificielle dans son ensemble, au problème de l'analogie est donc complexe, voire alambiquée, car ses membres se sont précipités en premier lieu sur des problèmes difficiles et rendus difficiles par la confusion des deux notions d'analogie et de métaphore.

[6]Par exemple, STEINHART, *Truth conditions for metaphors*, 1994.

Mais la complexité provient aussi des buts que se donnent les chercheurs en intelligence artificielle. D'une part, l'un des buts affichés de ce domaine est la construction de connaissances. Les systèmes-experts, représentatifs des travaux de l'intelligence artificielle, sont tels que la base de connaissances a autant d'importance que le moteur d'inférence[7]. D'autre part, l'une de ses préoccupations majeures est d'établir la signification des symboles pour travailler sur ces significations. Celles-ci s'appuient sur le contexte du problème, justement donné par la base de connaissances avec sa modélisation.

Pour toutes ces raisons, l'intelligence artificielle propose un traitement des analogies, confondues avec les métaphores, qui peut se caractériser de la façon suivante[8] :

- deux domaines différents interviennent ;
- pour chacun de ces deux domaines, il faut une base de connaissances ;
- la mise en correspondance des objets et le transfert des propriétés sont deux opérations de natures distinctes ;
- la qualité des analogies doit être évaluée en fonction du poids des propriétés transférées : nombre, valeur de vérité, etc.

Pour reprendre un exemple que nous avions déjà cité dans notre introduction (p. 29), celui de la réponse de Valéry à Blaise Pascal, l'analyse qu'en propose Milner[9] relève d'une telle approche. Si un carré analogique rend bien compte des positions des termes de l'analogie *silence* : *éternel* = *bavardage* : *intermittent*, les relations qui existent entre les termes sont relativement complexes (trois types différents selon cette explication).

Pour notre part, nous ressentons que les réponses complexes données à ces problèmes complexes cachent des problèmes élémentaires importants dont nous pensons que la résolution se devrait d'être d'abord recherchée. Nous voulons donc réduire de façon draconienne tout l'attirail précédent, et viser à énoncer un problème beaucoup plus simple, dont la résolution ne le sera pas nécessairement. Nous le ferons en simplifiant d'abord les types de données, en nous limitant aux seules chaînes de symboles, avec comme conséquence une réduction des caractéristiques du problème.

On pourrait alors penser que les travaux sur le programme Copycat[10], qui ressortissent précisément à l'intelligence artificielle, répondraient à nos préoccupations. Le programme Copycat a été conçu pour la résolution de devinettes analogiques du genre :

[7] Voir aussi les travaux se concentrant sur la formalisation des connaissances à des fins d'automatisation. Par exemple, Bachimont, *Herméneutique matérielle et artéfacture : des machines qui pensent aux machines qui donnent à penser*, 1991.

[8] Hall, *Computational approaches to analogical reasoning: A comparative analysis*, 1989. Il existe pléthore d'articles consacrés à la résolution de problèmes par « analogie ». Voir par exemple la bibliographie du site `http://www.nbu.bg/cogs/events/wana98.html`.

[9] Milner, *Introduction à une science du langage*, 1989, p. 291–292.

[10] Hofstadter & the Fluid Analogies Research Group, *Fluid concepts and creative analogies*, 1994, p. 205–265.

> En supposant que la chaîne de lettres *aabc* a été changée en *aabd*, comment de la même façon changeriez-vous la chaîne de lettres *ijkk*?[11]

À notre sens, cet énoncé biaise l'équation analogique de plusieurs manières. D'abord, il élimine implicitement la propriété que nous avons déjà vue chez Hermann Paul (p. 56) et que nous appellerons plus loin la permutation des moyens (p. 114), c'est-à-dire le fait que les deux termes du milieu, *aabd* et *ijkk*, devraient être permutables. Il le fait en focalisant l'attention sur le changement de *aabc* en *aabd* et en éliminant celui qui aurait pu avoir lieu de *aabc* à *ijkk*. La conséquence en est que toutes les symétries que nous dégagerons dans une analogie (p. 116) sont ici brisées. De façon révélatrice, la notation formelle utilisée dans la représentation de l'énoncé précédent n'est pas, sous la plume de ses auteurs la notation classique de l'analogie :

$$aabc : aabd = ijkk : x$$

mais une notation où la direction est explicitement imposée :

$$aabc \Rightarrow aabd;\ ijkk \Rightarrow\ ?$$

Ensuite, cet énoncé focalise l'attention sur les transformations (*de la même façon changeriez-vous*), et déplace donc l'attention de la relation globale qui existe entre chaînes, aux transformations à appliquer à ces chaînes. L'intérêt du problème n'est donc pas tant, semble-t-il, pour les auteurs du programme, l'obtention d'une réponse, c'est-à-dire une chaîne de lettres, que l'examen des procédures de transformations mises en œuvre pour obtenir une réponse. Dans nos travaux, au contraire, nous prendrons un soin extrême à ne pas expliciter les transformations. L'une des raisons en est que l'explicitation de transformations a généralement un coût de calcul supplémentaire par rapport à l'obtention directe de solutions par utilisation de propriétés intrinsèques. En ce sens, nous pouvons donc dire que nous ne nous plaçons pas dans la même optique que les travaux sur Copycat, ni même que d'autres travaux en traitement automatiques des langues qui cherchent, eux aussi, à expliciter des transformations, des commutations ou des patrons avec variables[12]. Pour en revenir à Copycat, tous ces biais font qu'il nous est difficile d'y voir de vraies équations analogiques à résoudre.

Enfin, à l'instar des approches d'intelligence artificielle, le programme Copycat utilise une modélisation du domaine à l'aide de fonctions comme : précédent dans l'alphabet, numéro d'ordre dans l'alphabet, etc. Ce type de connaissances, ne serait-ce que l'ordre d'un alphabet, n'a pas, à notre sens, à intervenir dans la résolution générale d'analogies considérées du seul point de vue linguistique. On en veut pour meilleure preuve qu'il n'est pas nécessaire de connaître l'ordre de l'alphabet ou des caractères dans lequel est transcrit une langue pour la parler. Et que dire de celles pour lesquelles un tel ordre n'existe pas, comme le chinois ?

[11] Ibidem, p. 206.

[12] Par exemple, Hathout, *Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes*, 2001.

2.3 En traitement automatique des langues

2.3.1 Traduction automatique

Puisque nous venons de parler d'analogies entre chaînes de lettres, nous pouvons maintenant passer du domaine de l'intelligence artificielle à celui du traitement automatique des langues, où la structure des données de base est précisément celle de chaînes de caractères.

Abordons d'emblée l'un des problèmes les plus ambitieux du traitement automatique des langues, celui de la traduction automatique. La traduction entre langues peut être posée dans des termes semblables à ceux des devinettes proposées au programme Copycat.

> En supposant que la phrase A se traduise par B, comment de la même façon traduiriez-vous la phrase C?

Cette vue est celle offerte par un article célèbre du domaine qui propose d'utiliser ce schéma comme moteur de base des systèmes de traduction automatique. La méthode se schématise de la façon suivante[13] :

$$\begin{array}{ccc} \text{langue 1} & & \text{langue 2} \\ \\ \alpha X \beta & \Longleftrightarrow & \alpha' X' \beta' \\ \Downarrow & & \Downarrow \\ \alpha Y \beta & \Longleftrightarrow & \alpha' Y' \beta' \end{array}$$

En fait, si l'on peut voir dans ce schéma l'équation analogique

$$\alpha X \beta : \alpha Y \beta = \alpha' X' \beta' : x \quad \Rightarrow \quad x = \alpha' Y' \beta'$$

l'auteur se concentre sur l'extraction des informations :

$$\begin{array}{rcl} \alpha - - \beta & \sim & \alpha' - - \beta' \\ X & \sim & X' \\ Y & \sim & Y' \end{array}$$

À notre avis, de même qu'avec Copycat, on se focalise de nouveau sur les transformations à effectuer plutôt que sur la relation globale exprimée par l'analogie. La réflexion sur le processus d'extraction des informations a conduit à la formation du paradigme de traduction par l'exemple, c'est-à-dire par patrons et dictionnaires[14].

Nous proposons de visualiser la méthode proposée par Nagao dans la figure 2.2 (p. 89). Elle est plus proche de l'image que l'on se fait d'une analogie

[13] 長尾 誠 (Nagao Makoto), *A framework of a mechanical translation between Japanese and English by analogy principle*, 1984, p. 174.

[14] Cette approche est chère aux laboratoires ATR qui l'ont illustrée et popularisée. Voir les articles sur la traduction automatique guidée par le transfert (acronyme anglais : TDMT) de Furuse et sur la traduction automatique par l'exemple (acronyme anglais : EBMT) de Sumita.

que la figure 2.1 (p. 84) qui donnait plutôt une image de la métaphore. Dans cette nouvelle figure on voit en effet se dessiner un quadrilatère avec quatre termes, nécessaires pour avoir une analogie. Mais cette figure indique clairement qu'un tel modèle ne saurait encore faire l'économie d'une représentation des connaissances. En effet, si une analogie exprime une similitude ou un égalité de rapports, le problème se pose de comparer les deux valeurs Δ et Δ' qui appartiennent à deux domaines différents. Dans cet exemple particulier, on ne saurait faire autrement que de représenter explicitement la négation (*il est jeune* par opposition à *il n'est pas jeune*). Mais plus encore, il faut que la représentation de la négation dans chacun des deux domaines soit identique ou à tout le moins comparable. On en revient donc à des conceptions proches de celles de l'intelligence artificielle, puisqu'une représentation des connaissances est nécessaire à l'automatisation d'une telle méthode.

La dernière critique adressée au modèle de Nagao provient du type des connaissances extraites. On a vu plus haut que ce type de connaissances est en fait un patron à une seule variable. Les travaux ultérieurs[15] sur ce modèle ne dépassent pas la puissance de la transduction entre langages réguliers. Or cette puissance a été reconnue comme insuffisante pour le traitement de certains problèmes linguistiques relevant de la seule morphologie. Ainsi, la morphologie de l'arabe ne peut se satisfaire de ce modèle dans sa forme simple. Des extensions y sont nécessaires. Pour remédier à ce problème, une réalisation pratique comme celle du système de traduction automatique TDMT introduit, d'une part des patrons à plusieurs variables et d'autre part la récursivité dans l'application de certains patrons, ce qui la place dans le hors-contexte[16] (voir p. 74).

2.3.2 Prononciation par analogie

Passons d'un extrême à l'autre, et considérons maintenant un problème qui, a priori, semble beaucoup plus simple et bien moins ambitieux que celui de la traduction automatique, la prononciation automatique. Il ne s'agit pas de synthèse de la parole, où le but est de produire un signal sonore, mais seulement, à partir d'une forme orthographique, d'obtenir une représentation de sa prononciation, par exemple dans l'alphabet phonétique international. Pour cette raison, le problème est baptisé *transcription graphémique-phonémique*. Il joue donc seulement sur des formes écrites et sur des symboles. Pour nous, il s'agit là d'une simplification souhaitée par rapport à la généralité des problèmes que se posait l'intelligence artificielle. Les structures de données deviennent élémentaires : il s'agit seulement de chaînes de symboles.

Le problème de la transcription graphémique-phonémique est trivial pour des langues utilisant un alphabet avec ses vertus inhérentes, c'est-à-dire faisant presque correspondre un seul son à un seul graphème ou graphème composé. C'est le cas de langues comme l'espagnol, le portugais, l'italien, le polonais ou

[15] 佐藤 理史 (Satou Satosi), *Example-based Machine Translation*, 1991.

[16] Boitet, *On the automatic transformation of a set of EBMT Constituent Boundary Patterns into a Context-Free Grammar, and associated bottom-up algorithms*, 1994.

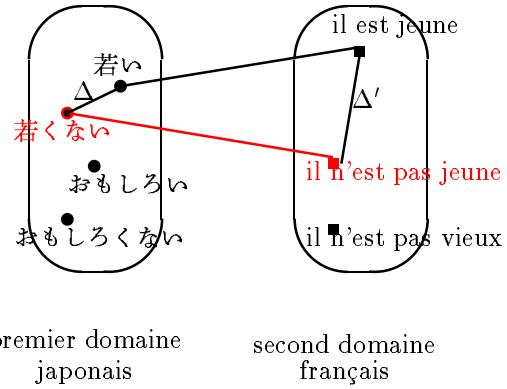


Figure 2.2: Modèle de traduction proposé par Nagao : une analogie chevauche deux domaines

encore le tchèque. Et la liste ne s'arrête pas là, et n'est pas non plus limitée aux langues faisant usage de l'alphabet latin. L'arabe voyellé serait un autre exemple. Il n'en va pas de même de l'anglais. L'orthographe anglaise est en effet un ratage de l'application du système alphabétique du fait de l'échec de la symbiose de deux orthographes, germanique et française. Il s'agit en effet de

> l'une des applications les moins réussies de l'alphabet romain.[17]

Pour cette raison il existe quantité de travaux consacrés à la prononciation de l'anglais adoptant différentes approches[18]. Certains de ces travaux essaient même de reproduire le comportement humain[19]. Un nombre important de ces approches utilisent l'analogie. Un résumé de la tâche dans sa conception analogique est donné par Pirelli et Federici[20] :

$$\begin{array}{ccc} \textit{vain} & \xrightarrow{f} & /\text{vejn}/ \\ \downarrow g & & \downarrow h \\ \textit{sane} & \xrightarrow{f} & x= /\text{sejn}/ \end{array}$$

[17] Abercrombie, *Extending the Roman alphabet: Some orthographic experiments of the past four centuries*, 1981, p. 196, à propos de l'orthographe anglaise. Admirez l'euphémisme. Rappelons aussi la blague légèrement malhonnête qui veut que *fish* /fɪʃ/ (poisson) puisse s'écrire *ghoti* avec *gh* /f/ comme dans *laugh* /lʌf/ (rire), *o* /ɪ/comme dans *women* /wɪmɪn/ (femmes), et *ti* /ʃ/ comme dans *nation* /neiʃ(e)n/ (nation).

[18] Certains mélangent évidemment ces différentes approches. Par exemple, Marchand & Damper, *A multistrategy approach to improving pronounciation by analogy*, 2000.

[19] Voir, par exemple, Damper & Eastman, *Pronouncing text by analogy*, 1996.

[20] Anglais : *sane* (sain), *vain* (vain), adjectifs. Pirelli & Federici, *"Derivational" paradigms in morphonology*, 1994, p. ??.

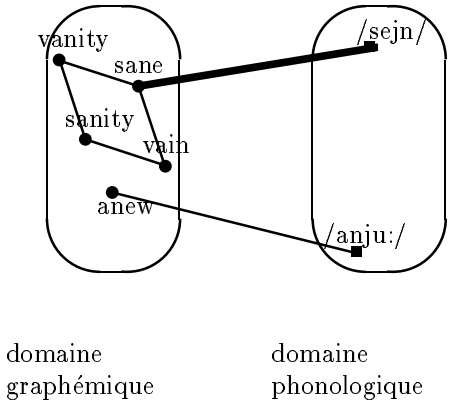


Figure 2.3: Modèle d'Yvon : les liens reposant sur une relation paradigmatique sont plus forts

Multiplicité ou unicité des domaines

De même qu'en intelligence artificielle, deux domaines distincts apparaissent nettement, graphémique et phonémique. Par conséquent, les fonctions f, g et h sont de types différents parce que leurs ensembles sources et buts sont de types différents.

Encore une fois, de même qu'en intelligence artificielle, de tels systèmes utilisent des bases de données de formes écrites ou phonétiques. A propos de son propre modèle, Yvon[21] note que :

> « Le [...] modèle repose de façon cruciale sur l'existence de nombreuses *relations paradigmatiques* existant dans la base lexicale. »

Les relations paradigmatiques étant des relations dans lesquelles quatre mots interviennent, ce sont en fait des analogies au niveau morphologique : « réaction *est à* réacteur, *ce que* faction *est à* facteur. »

$$\begin{array}{ccc} \mathit{réacteur} & \xrightarrow{f} & \mathit{réaction} \\ \downarrow g & & \downarrow g \\ \mathit{facteur} & \xrightarrow{f} & \mathit{faction} \end{array}$$

Répétons que, bien que la prononciation par analogie soit un exercice intéressant pour l'anglais, du fait de l'incohérence de son système orthographique, c'est un problème dénué d'intérêt pour bien d'autres langues. En bref, la transcription graphème-phonème n'est nullement un problème universel. Cependant, la résolution d'analogies au niveau morphologique semble

[21] Yvon, *Paradigmatic cascades: a linguistically sound model of pronunciation by analogy*, 1994.

bien lui, universel, même pour des langues comme le chinois, où bien des mots de plus de deux caractères s'inscrivent indéniablement dans des séries analogiques :

科学 : 科学家 = 政治 : 政治家 [22]
我 : 我們 = 他 : 他們 [23]
今年 : 今天 = 明年 : 明天 [24]

Par opposition aux approches de l'intelligence artificielle, les analogies au niveau morphologique jouent dans un seul domaine, celui des mots. Par conséquent, le nombre de relations entre les termes de l'analogie diminue de trois (f, g et h) à deux (f et g). De plus, comme les quatre termes intervenant dans une analogie sont du même domaine, les ensembles sources et buts de f et g sont identiques ; f et g sont donc de même type. Enfin, les analogies au niveau morphologique peuvent être considérées comme des équations indépendantes de quelque connaissance que ce soit sur la langue des mots. Cela élimine le recours à des dictionnaires ou à des bases de connaissances et conséquemment à des notions telles que l'ordre dans l'alphabet, comme dans Copycat (p. 85).

$$\begin{array}{ccc} \textit{réacteur} & \xrightarrow{f} & \textit{réaction} \\ \downarrow g & & \downarrow g \\ \textit{facteur} & \xrightarrow{f} & x? \end{array}$$

Toutes les remarques précédentes font que le problème des analogies entres mots marque une réduction importante par rapport aux problèmes généraux et ouverts de l'intelligence artificielle.

Unicité ou multiplicité des modifications

La résolution des analogies au niveau morphologique reste cependant un problème difficile parce que, vues sous l'angle des modifications à appliquer, plusieurs modifications simultanées peuvent être nécessaires dans la transformation d'un mot en un autre (par exemple *faiseur* → *défaire* requiert le remplacement du suffixe *-eur* par *-re* et l'insertion du préfixe *dé-*). Ce problème n'a toujours pas été résolu de façon satisfaisante. Par exemple, Yvon[25] n'autorise qu'une seule modification à la fois. Pour lui, plusieurs modifications s'obtiennent par l'application successive d'analogies morphologiques à une seule modification, d'où le nom donné à son modèle de *modèle en cascades*. Cependant, il y a des langues pour lesquelles les modifications simultanées sont nécessaires, comme les langues sémitiques. Rappelons que nous

[22] Prononciation donnée en pinyin. 科学 /kēxué/ (la science), 科学家 /kēxuéjiā/ (un scientifique), 政治 /zhèngzhì/ (la politique), 政治家 /zhèngzhìjiā/ (un homme politique).

[23] 我 /wǒ/ (je), 他 /tā/ (il), 我們 /wǒmen/ (nous), 他們 /tāmen/ (ils).

[24] 今年 /jīnnián/ (cette année), 今天 /jīntīan/ (aujourd'hui), 明年 /míngnián/ (l'année prochaine), 明天 /míngtīan/ (demain).

[25] Yvon, *Paradigmatic cascades: a linguistically sound model of pronunciation by analogy*, 1994.

avions déjà mentionné notre intérêt envers l'arabe (voir page 48) pour ces raisons formelles.

Nagao[26] se limite aussi à « une seule modification à la fois » pour sa méthode de traduction automatique. On voit clairement que la traduction d'une phrase source dans son modèle est l'adaptation de traductions de phrases similaires extraites d'une base de données, selon le même schéma théorique que celui appliqué à la transcription graphèmes-phonèmes de l'anglais. La difficulté de traiter plusieurs modifications à la fois est palliée, lors de la création de la base, par la fourniture de phrases ne différant entre elles que par une seule modification. Bien que la réponse de Sadler[27] soit plus satisfaisante que celle de Nagao, avec l'utilisation d'une algèbre sur les arbres (addition et soustraction de sous-arbres), l'usage de base de connaissances et la multiplicité des domaines refont malheureusement surface. Tout cela provient, croyons-nous, de la focalisation sur les opérations à effectuer pour modifier les chaînes, impliquée par une vision dynamique du problème, alors que l'analogie, fondamentalement, se contente d'énoncer des rapports entre ces chaînes, c'est-à-dire des relations statiques.

2.4 Introduction à notre propos : analogies entre chaînes de symboles

Nous avons fini notre tour d'horizon du problème en traitement automatique des langues, et nous venons d'aboutir sur ce qui a constitué en fait le départ de notre recherche : l'*analogie* telle que vue par Saussure. C'est-à-dire l'opération par laquelle, étant données deux formes d'un même mot, et seulement une forme d'un second mot, on crée la forme manquante : « *honor* est à *honōrem* ce que *ōrātor* est à *ōrātōrem* » notée $\bar{o}r\bar{a}t\bar{o}rem : \bar{o}r\bar{a}tor = hon\bar{o}rem : honor$. C'est tout simplement là la définition d'Aristote : « A est à B ce que C est à D », en posant en plus, soulignons-le, l'identité des domaine de A, B, C et D comme chez Euclide. Cependant, bien que ce type d'analogie ait été mentionné et utilisé durant des siècles, aucun moyen algorithmique n'a, à notre connaissance, jamais été donné pour résoudre des équations analogiques, peut-être parce que l'opération semble si intuitive.

Notre but initial était de construire un véritable solveur d'analogies, c'est-à-dire un algorithme qui, étant données trois chaînes de symboles, fournisse une chaîne de symboles, analogue aux données d'entrée. Nous[28] avions essayé de donner une première caractérisation qui n'était hélas pas toujours valide, parce qu'elle englobait aussi des cas qui ne sont pas des analogies (voir plus bas, le système du rectangle, p. 258). Cependant, cette première tentative possédait déjà certaines caractéristiques importantes selon nous, puisqu'il s'agissait :

[26] 長尾 誠 (Nagao Makoto), *A framework of a mechanical translation between Japanese and English by analogy principle*, 1984 et à sa suite 佐藤 理史 (Satou Satosi), *Example-based Machine Translation*, 1991.

[27] Sadler & Vendelmans, *Pilot implementation of a bilingual knowledge bank*, 1990.

[28] Lepage & Ando, *Saussurian analogy: a theoretical account and its application*, 1996

- d'être capable de réaliser plusieurs modifications à la fois ;
- d'appliquer ces modifications sur une unique structure de données ;
- de faire cela sans aucun dictionnaire ni connaissances externes.

En effet, tout ce que nous venons de voir et qui nous a permis petit à petit de dégager le problème particulier de l'analogie entre chaînes de symboles d'un même domaine, nous permet de dégager les points suivants.

- Dans l'analogie entre chaînes de symboles seulement, tous les éléments appartiennent au même domaine. C'est une différence énorme d'avec l'analogie en intelligence artificielle. Là, en effet on recherche, ou on cherche à établir, des relations entre domaines différents.
- Plusieurs opérations peuvent être nécessaires pour expliquer le passage entre mots intervenant dans une analogie. Par exemple, $d\acute{e}corer : ind\acute{e}corable = vider : invidable$ requiert deux transformations. Cela est encore une différence significative d'avec les recherches courantes[29], où, nous l'avons vu, le problème de l'application simultanée de plusieurs opérations n'étant pas résolu, la transformation d'un mot (ou phrase) en un autre ne se fait que par étapes successives.
- Plus que de mettre en jeu des transformations, les analogies expriment des rapports. Il s'agit ici d'effectuer une changement de perspective et de passer d'une vue dynamique ou procédurale à une vue statique ou déclarative.

En conclusion, avant de proposer un algorithme (partiel) de résolution d'équations analogiques ou de vérifications d'analogies entre chaînes de symboles d'un même domaine, nous allons nous consacrer à la *caractérisation* de ces analogies. Dans un premier temps, nous allons refuser toute connaissance, et donc toute base de connaissances. En plus, nous allons aussi refuser de rechercher quelque signification que ce soit pour les symboles. Ce ne sera que dans un deuxième temps que nous pourrons éventuellement aborder une (très humble) théorisation de l'émergence du sens, ou plutôt de la création de nouvelles correspondances entre domaines munis d'interprétations différentes, au moyen des homomorphismes d'analogies (p. 200 et 202).

Notre travail doit être compris comme une première tentative de caractérisation des analogies entre chaînes de symboles. Nous essaierons en effet d'en tirer un algorithme qui pourrait être comparé à l'algorithme d'addition des nombres entiers que nous apprenons tous à l'école primaire. Or, si des enfants en bas âge sont capables d'additionner des nombres entiers petits, ils ne sont pas capables de le faire pour certains nombres entiers grands sans connaître l'algorithme d'addition appris en classe. De même, alors que nous

[29]長尾 誠 (Nagao Makoto), *A framework of a mechanical translation between Japanese and English by analogy principle*, 1984 ou Yvon, *Paradigmatic cascades: a linguistically sound model of pronunciation by analogy*, 1994.

sommes capables de résoudre des analogies entre chaînes simples, voire entre certaines chaînes de symboles complexes particulières, il nous est impossible de résoudre certaines analogies complexes. Et jusqu'à présent, on ne possède pas de tel algorithme explicite universel.

Un tel algorithme, si nous l'obtenions réaliserait ce que nous appelons l'espoir analogique qui est de faire plus à effort moindre d'une part, et d'autre part de reproduire l'activité humaine d'apprentissage. Nous illustrons nos recherches dans la première direction en montrant le côté très pratique des motivations. Il s'agit de construire des systèmes de traitement automatique des langues qui minimisent l'effort humain nécessaire à leur mise en place. Le but est d'utiliser les connaissances humaines sous leur forme brute, non structurée et non traitée. Nous sommes déjà en mesure d'illustrer notre propos par une maquette de système de traduction automatique (p. 314). Cette maquette utilise la donnée d'une centaine de milliers de phrases dans deux langues différentes faisant d'ailleurs usage de systèmes d'écriture différents, sous leur forme brute. L'avantage d'un tel système est l'absence de règles explicites, ou de patrons, et l'inutilité de l'écriture de telles règles à la main, voire de la construction de tels patrons par machine. Évidemment, disons tout de suite que nous ne prétendons pas avoir résolu le problème de la traduction automatique. Les résultats ne sont pas encore probants, mais ils ont au moins l'avantage d'ouvrir une voie. Dans la deuxième direction, celle de la reproduction de l'activité humaine, nous sommes aussi déjà en mesure de montrer une application de conjugaison automatique des verbes français ou de déclinaison des noms allemands. Dans une telle application, la faculté de conjuguer un verbe jamais vu en l'alignant sur un modèle de conjugaison connu reflête bien l'activité humaine qui consiste à se servir de la « quatrième proportionnelle » comme le disait Saussure (p. 61) et le répétait Mounin (p. 71), c'est-à-dire à analogiser, comme le disait Pullum (p. 72).

Chapitre 3

Synthèse des principales notions

3.1 L'analogie comme pôle d'opposition

Nous essayons maintenant de faire une synthèse des principales notions concernant l'analogie. Nous avons vu que l'analogie, dans l'histoire de la grammaire et de la linguistique, s'est vu attribuer tantôt un rôle de facteur d'ordre, tantôt de fauteur de troubles. Elle a donc été un pôle d'opposition dans le panorama de la langue, aspect que nous esquissons dans le tableau 3.1.

Tableau 3.1: L'analogie comme pôle d'opposition

	pôle de l'analogie	**pôle opposé**
Grecs et Latins	–	anomalie
	régularité	irrégularité
Néogrammairiens	reconstructions analogiques	changements phonétiques
	trouble	lois naturelles
Structuralistes	synchronie	diachronie
	réalignement par analogie	changements phonétiques
	ordre	désordre
Kuryłowicz	diachronie	synchronie
Mańczak	effet diachronique	effet synchronique
	formes secondaires	formes primaires
	faibles fréquences	hautes fréquences
Générativistes	acquis ?	inné
	induction ?	paramètres
	surproduction	adéquation
Bachelard	représentations analogiques	mathématisation
	pseudo-scientificité	rigueur scientifique

3.2 L'analogie comme position intermédiaire

Mais l'analogie n'a pas été qu'un pôle d'opposition. Elle s'est aussi vu attribuer une place intermédiaire entre différents concepts. Nous allons montrer ci-après qu'ils se rattachent tous à deux notions fondamentales que nous pensons être constitutives de l'analogie.

3.2.1 Synonymie et homonymie

Tout d'abord, dans le débat médiéval sur le caractère de l'être, le problème central était l'évaluation de la qualité des rapports entre Dieu et sa création, problème qui débouchait sur celui de savoir comment s'effectuaient les transferts de sens de l'être. Nous avons vu avec Salamanca (p. 50) que ce que l'on peut dire de l'être peut l'être de façon synonymique « l'être est un » ou de façon homonymique « l'être peut être de multiples manières ». En d'autres termes, les tenants de l'univocité pensait l'être de façon identique, c'est-à-dire totalement semblable, tandis que ceux de l'équivocité le pensait dans la diversité, c'est-à-dire le totalement dissemblable. La thèse analogique, intermédiaire de l'univoque et de l'équivoque, est caractérisée en deux termes par Salamanca[1] : « les êtres sont semblables et participent les uns des autres ». Cette position médiane fait donc usage de deux notions, le semblable et la participation ou coexistence.

3.2.2 Métaphore et métonymie

En langue, la manifestation du transfert de sens est le remplacement d'un terme par un autre. Or, un tel remplacement est possible selon deux modes, *métaphorique* ou *métonymique*, caractérisés chacun par une relation différente, du type de la similarité ou de la contiguïté. C'est là la définition de plusieurs dictionnaires. Cette définition est centrale chez Jakobson pour qui la métaphore s'oppose à la métonymie de cette façon précise.

> Au contraire de la métonymie qui associe deux termes (idées) en fonction de leur contiguïté, la métaphore substitue un terme (idée) par un autre terme similaire.[2]

Ainsi donc, on peut voir dans les notions avancées par Salamanca, le semblable et la participation ou coexistence, et qui correspondent à la similarité et à la contiguïté, comme des correspondants, par transitivité, de la métaphore et de la métonymie dans l'activité langagière. Qui plus est, Jakobson pousse les termes de métaphore et de métonymie jusqu'à en faire carrément des synonymes de syntagmatique et paradigmatique.

[1] SALAMANCA, *La tradición histórica de la analogía lingüística*, 1984, p. 373, voir citation complète, p. 50.

[2] Cité par VERHAEGEN, *Image, diagramme, métaphore*, 1994.

> Cette dualité [unités semblables et unités coexistantes] a, pour Jakobson, une grande généralité. Elle serait à la base des figures de rhétorique les plus employées par « le langage littéraire » ; la métaphore (un objet est désigné par le nom d'un objet semblable) et la métonymie (un objet est désigné par le nom d'un objet qui lui est associé dans l'expérience) relèveraient respectivement de l'interprétation paradigmatique et syntagmatique, si bien que Jakobson prend parfois pour synonymes *syntagmatique* et **métonymique**, *paradigmatique* et **métaphorique**.[3]

Pour Jakobson, ces deux notions caractérisent la poésie en ce qu'elles se superposent dans la comparaison poétique.

> La superposition de la similarité sur la contiguïté confère à la poésie son essence de part en part symbolique, complexe, polysémique. [...] En poésie, où la similarité est projetée sur la contiguité, toute métonymie est légèrement métaphorique, toute métaphore a une teinte métonymique.[4]

Les notions de métaphore et de métonymie se rattachent donc aux deux notions de similarité et de contiguïté. On ne peut que constater la continuité depuis l'Antiquité. En effet, Jakobson ne fait que reprendre les termes de Kruszewski, et nous avons déjà vu plus haut (p. 59) que ce linguiste reprenait les mêmes notions de philosophes anglais qui, cela nous semble évident, devaient les reprendre des philosophes médiévaux qui, à leur tour, les reprenaient certainement d'Aristote qu'aucun d'eux n'ignorait :

> [...] [Jakobson] écrit dès 1963, en hommage au linguiste polonais Kruszewski : « il y a deux opérations fondamentales sous-jacentes au comportement verbal : la sélection et la combinaison. Dans son *Aperçu de la science du langage*, publié il y a 80 ans mais toujours capital, Kruszewski relie ces deux opérations à deux modèles de relations : il fonde la sélection sur la *similarité*, la combinaison sur la *contiguïté* ».[5]

3.2.3 Sélection et combinaison

Jakobson considère que l'articulation du discours est semblable au transfert de sens, en ce qu'ils s'articulent tous deux autour des deux mêmes opérations de l'esprit distinctes que sont la *sélection* et la *combinaison* :

> Une autre orientation de la poétique structuraliste a consisté dans sa lecture du discours littéraire à partir du croisement de la **métaphore** et de la **métonymie** (que Jakobson rapporte respectivement au processus linguistiques de la *sélection* et de la *combinaison*). Tout signe

[3] Ducrot & Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, 1978, p. 275.

[4] Jakobson, *Question de poétique*, 1973, p. 235-236, cité par Ghils, *Langage et contradiction*, 1998, note 19.

[5] Cité par Morenon, *Roman Jakobson*, 1997, §L'apport de Kruszewski.

> linguistique implique deux modes d'arrangement : la combinaison et la sélection ou substitution. Le discours se déroule selon deux axes : celui de la similarité (c'est le processus métaphorique), celui de la contiguïté (c'est le processus métonymique).[6]

Puisque cette explication repose sur des opérations de l'esprit, il n'est pas étonnant de la voir commenté par des psycho-linguistes :

> On choisit les mots et on construit les phrases, ce double comportement de toute personne qui parle permet de définir, pour Jakobson, les deux termes cités :
>
> - la sélection est un choix et une comparaison : elle implique, entre deux termes alternatifs, la possibilité de substituer l'un à l'autre, équivalent du premier sous un aspect, différent sous un autre ;
> - mais c'est en combinaison avec d'autres signes qu'apparaît nécessairement un signe linguistique, lui-même composé d'éléments constituants.
>
> Dans l'article de 1956[7] l'auteur s'attache à montrer que ces simples opérations ressortissent à des attitudes psychiques d'une portée beaucoup plus vaste, conformément à l'intuition de Kruszewski.
>
> Il relève d'abord que la similarité peut prendre des formes et des degrés variés. Lorsque, entre des entités distinctes, l'esprit reconnaît "la similitude, l'équivalence, la ressemblance, l'"être comme", l'analogie(...), le contraste", c'est la capacité de sélection qui est en jeu.
>
> La combinaison implique, de son côté, "sous différentes formes et degrés, la relation externe de contiguïté : voisinage, proximité et éloignement ; la subordination et coordination". Autrement dit, l'esprit humain, par la faculté de combinaison
>
> - reconnaît que des entités, ayant des éléments distincts, font partie d'un même ensemble ;
> - il peut les associer dans un ensemble commun.
>
> La combinaison est congruente à la métonymie, la sélection est congruente à la métaphore.[8]

3.2.4 Comparativité et connectivité

Chez le psychologue Morier, ces termes sont encore requalifiés dans le cadre de l'étude de l'aphasie par les termes de *comparativité* et de *connectivité*. Dans une présentation fort claire due à Martine Morenon, toutes ces notions sont regroupées de la façon suivante[9] :

[6] Ducrot & Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, 1978, p. 584.

[7] *Deux aspects du langage et deux types d'aphasie*, repris dans Jakobson, *Essais de linguistique générale*, 1963 comme deuxième chapitre.

[8] Morenon, *Roman Jakobson*, 1997, §L'apport de Kruszewski.

[9] Morenon, *Roman Jakobson*, 1997, §Résumé.

« La comparativité gouverne :

- La *similarité* qui est un état ou un rapport.
- La *sélection* qui est l'opération par laquelle l'esprit choisit et discerne les ressemblances et dissemblances. Tout système codé lui est assujetti.
- La *métaphore* qui est création de sens par abstraction de caractères communs. La forme schématisante en est l'intersection.

La connectivité gouverne :

- La *contiguïté* qui est un état ou un rapport ;
- La *combinaison* qui est une opération de composition-décomposition d'une unité fonctionnelle, éventuellement extra-linguistique.
- La *métonymie* qui rend compte, au plan du langage, d'un ensemble compréhensif (parties d'un même ensemble). Elle traduit dans le langage les rapports extra-linguistiques. La forme schématisante est la figure d'inclusion. »

3.2.5 Similarité et contiguïté

En résumé, dans tout ce que nous venons de voir, les deux notions centrales sont celles de de *similarité* et de *contiguïté*. Elles se trouvent à la croisée de notions polaires entre lesquelles se situe l'analogie. On peut donc en conclure que l'analogie, en tant que position intermédiaire, doit nécessairement mettre en jeu à la fois la similarité et la contiguïté.

On ne sera pas surpris d'apprendre que les deux notions de similarité et de contiguïté ont elles aussi une histoire. Elles se rattachent toutes deux au travail de la pensée médiévale sur la notion de proximité dont la conclusion est que la proximité se découpe seulement en deux espèces, justement, la similarité et la contiguïté. Aristote donnait déjà ces deux notions lorsqu'il parlait de l'association d'idées, et elles se retrouvent chez Leibniz :

> Au sommet de la classification, comme j'ai écrit dans ma jeunesse, je distingue entre relations de comparaison et de conjonction. Une relation de comparaison existe dans le même et le divers, dans le semblable et le dissemblable, dans l'égal et l'inégal. Une relation de conjonction à son tour, est ou bien simple (comme entre tout et partie, entre partie et co-partie, et dans les compléments de lieu, de temps, ou d'autres de ce genre) ou bien une connexion, dans laquelle interviennent une certaine influence et un enchaînement (comme entre cause et effet, entre signe et ce dont il est le signe).[10]

Finalement, on ne sera pas surpris de voir apparaître explicitement, dans le schéma de l'analogie « vulgaire » (voir page 1.1.4) donné par Itkonen, à la suite de bien d'autres[11] le similaire et le contigu comme étant les deux notions

[10] Leibniz, *Nova methodus pro maximis et minimis*, 1646 1716.
[11] Itkonen, *Iconicity, analogy, and universal grammar*, 1994, p. 44.

orthogonales constitutives de l'analogie. L'orthogonalité s'explique par le fait que la relation de contiguïté est de l'ordre de la connaissance, alors que la relation de similarité est d'ordre perceptif (« similitude de la perception » dans les termes de Deleuze). Ainsi, dans une telle vue, l'analogie est à la fois métaphore et métonymie, c'est-à-dire selon Jakobson, poétique.

Tableau 3.2: Le contigu et le semblable

	← similarité →	
↑ contiguïté ↓	oiseau ailes poumons plumes	poisson nageoires branchies écailles

	← similarité →	
↑ contiguïté ↓	A	B
	C	D

3.2.6 Tableaux des notions dégagées

Nous pouvons maintenant donner un second tableau récapitulatif des notions dégagées par l'histoire de l'analogie. Dans ce tableau, l'analogie n'est plus elle-même un pôle d'opposition, mais la position intermédiaire entre deux pôles. En résumé ce tableau délivre l'idée que, fondamentalement, l'analogie est structurée par les deux notions de similarité et de contiguïté.

Tableau 3.3: L'analogie comme position intermédiaire

	premier pôle	**place de l'analogie**	**second pôle**
sens de l'être	univocité	–	équivocité
classes des mots	synonymie	paronymie	homonymie
emplois des mots	métaphore	–	métonymie
axes linguistiques	paradigme	–	syntagme
opérations de l'esprit	sélection	–	combinaison
ordre de l'intellect	perception	–	connaissance
opérations ensemblistes	intersection	–	inclusion
relations	similarité	–	contiguïté

3.3 Les notions constitutives de l'analogie

Il nous faut maintenant exploiter les notions dégagées ci-dessus pour les appliquer au problème dégagé plus haut (p. 92), à savoir celui des analogies entre chaînes de symboles. Répétons que notre but premier est une *caractérisation*

formelle de l'analogie, afin, dans un deuxième temps de proposer un algorithme de vérification d'analogies et de résolution d'équations analogiques entre chaînes de symboles. Or, dans tous les cas, ce qui semble faire cruellement défaut dans toutes nos lectures historiques, c'est une mesure quelconque, une quantification, des similarités comme des contiguïtés, condition *sine qua non* pour pouvoir traiter des analogies selon Henrion (p. 38). Il nous faut donc nous interroger pour savoir s'il en est de même en ce qui concerne le domaine des chaînes de symboles. En vue de cette caractérisation formelle, la question qui se pose donc est la suivante: n'existerait-il pas déjà des mesures de la similarité et des mesures de la contiguïté que nous pourrions utiliser ?

3.3.1 Similarité et distance

Fort heureusement, pour le domaine qui nous intéresse, à savoir celui des chaînes de symboles, on dispose déjà d'une mesure de similarité ou de ressemblance. La *similitude*[12] entre deux chaînes de symboles A et B, que nous noterons $\sigma(A, B)$, est définie comme la longueur de leur plus long sous-mot[13] commun[14].

La notion « complémentaire » de celle de similitude est celle de la mesure de la dissimilarité ou d'incompatibilité, ou d'incertitude ou encore de désaccord. Notons au passage que les notions d'incertitude et de désaccord sont utilisées directement par certaines approches de l'analogie en raisonnement automatique[15]. Cette notion trouve son expression mathématique dans celle de distance.

Une distance se caractérise par trois propriétés fondamentales[16] :

DÉFINITION **1 (Distance)** *Soit $\mathcal{E}$ un ensemble, δ une fonction de $\mathcal{E} \times \mathcal{E}$ dans $\mathbb{R}^+$, l'ensemble des réels positifs, δ est une distance sur $\mathcal{E}$ si et seulement si*

- $\forall (A, B) \in \mathcal{E}^2,\ \delta(A, B) = 0 \Leftrightarrow A = B$ *(égalité)*
- $\forall (A, B) \in \mathcal{E}^2,\ \delta(A, B) = \delta(B, A)$ *(symétrie)*
- $\forall (A, B, C) \in \mathcal{E}^3,\ \delta(A, C) \leq \delta(A, B) + \delta(B, C)$ *(inégalité triangulaire)*

Pour le domaine qui nous occupe, à savoir celui des chaînes de symboles, on dispose aussi déjà d'un ensemble de distances « intuitives », appelées distances d'édition. Ces distances donnent le nombre minimal d'opérations d'édition, insertion, suppression, remplacement, voire aussi permutation, nécessaires

[12]Nous utilisons *similitude* pour réserver le suffixe *-ité* de *similarité* à une valeur réelle dans [0; 1] comme dans *probabilité*. La similarité entre deux chaînes de symboles A et B est définie par : $2 \times \sigma(A, B)/(|A| + |B|)$.

[13]Nous rappelons qu'une sous-chaîne est connexe, alors qu'un sous-mot ne l'est pas nécessairement. Par exemple, *angle* est une sous-chaîne de *triangle*, alors que *tringle* en est seulement un sous-mot.

[14]Voir HIRSCHBERG, *Algorithms for the longest common subsequence problem*, 1977.

[15]Par exemple, SKOUSEN, *Analogical modeling of language*, 1989, p. 23, § 2.

[16]BARTHÉLEMY & GUÉNOCHE, *Les arbres et les représentations des proximités*, 1988, p. 44 à 47.

pour transformer une chaîne de symboles en une autre. En annexe (p. 349), nous montrons que les distances d'édition peuvent être des distances au sens mathématique du terme: il est nécessaire et suffisant pour cela d'avoir une distance entre symboles. Elle se généralise alors naturellement aux chaînes de symboles.

Le lien entre la ressemblance et la dissemblance est fait, dans le domaine des chaînes de symboles par un résultat remarquable. Il existe une distance d'édition bien particulière qui entretient un rapport privilégié avec la similitude. Il s'agit de la distance équipée seulement des deux opérations d'insertion et de suppression[17]. Pour une telle distance, le remplacement est l'application de deux opérations l'une à la suite de l'autre : une suppression et une insertion, au même endroit dans n'importe quel ordre.

DÉFINITION **2 (Distance d'édition canonique)** *Soit $\mathcal{V}$ un alphabet. On appelle* distance d'édition canonique *la distance d'édition particulière telle que les insertions et les suppressions valent 1.*

De façon équivalente, on peut considérer que les remplacements valent donc 2, ce qui permet à ceux qui tiennent à définir les distances d'édition nécessairement avec les trois opérations d'insertion, suppression et remplacement, d'être satisfaits.

La distance d'édition canonique vérifie bien l'inégalité triangulaire. Il s'agit donc bien d'une distance au sens mathématique.

La relation remarquable avec la similitude est donnée par la proposition suivante:

PROPOSITION **1** *Soit $\mathcal{V}$ un alphabet, soit δ la distance d'édition canonique. Alors*

$$\forall (A,B) \in (\mathcal{V}^*)^2, \quad \delta(A,B) = |A| + |B| - 2 \times \sigma(A,B)$$

En faisant apparaître la similarité, terme que nous avions réservé à un réel de l'intervalle $[0;1]$,

$$\frac{\delta(A,B)}{|A|+|B|} = 1 - \frac{2 \times \sigma(A,B)}{|A|+|B|}$$

on obtient une expression qui rejoint le sens commun qui veut qu'une mesure de la dissemblance (liée à la distance) soit égale à l'unité moins la mesure de la similarité.

3.3.2 Contiguïté et probabilité d'occurrence

Pour ce qui est de la seconde notion constitutive de l'analogie, malheureusement, contrairement à ce que nous venons de voir pour la similarité, nous ne voyons pas quelle notion connue sur les chaînes de symboles serait une expression mathématique de celle de contiguïté.

[17] LEVENSHTEIN, *Binary codes capable of correcting deletions, insertions and reversals*, 1966.

La notion de contiguïté caractérise la relation de la partie au tout, du tout à la partie ou de la conjonction de deux parties dans l'expérience selon les définitions que nous avons vues à propos de la métonymie, de la conjonction et de la connectivité (p. 99). Si nous transposons ces définitions aux chaînes, la contiguïté devrait donc caractériser le fait qu'une sous-chaîne (ou un symbole) appartienne à une sous-chaîne, ou qu'une sous-chaîne contienne une sous-chaîne (ou un symbole), ou que deux sous-chaînes se suivent (connectivité). Ce que nous recherchons donc ce serait une mesure de la qualité du fait qu'un symbole appartienne à une chaîne dans les conditions de l'analogie. De nouveau, nous ne pouvons que regretter de n'être pas en mesure de répondre à cette question en l'état actuel de nos recherches.

Au cours de nos travaux, nous avons essayé, comme tout ce que nous venons de dire nous incitait à le faire, de cristalliser les deux dimensions de similarité et de contiguïté en une notion unique. Mais en vain. Le versant de la contiguïté nous échappe encore.

3.4 Les articulations constitutives de l'analogie

Les deux notions de similarité et de contiguïté sont des notions constitutives de l'essence de l'analogie. Elles lui sont en quelque sorte intérieures, voire elles y sont intériorisées. Revenons maintenant à l'aspect extérieur de l'analogie. À la suite des définitions des Anciens qui énonçaient que l'analogie ou proportion est l'égalité de deux rapports, on note $A : B = C : D$. Immédiatement, on obtient donc deux articulations: l'égalité et le rapport, qui, elles, sont extériorisées dans la notation.

3.4.1 Égalité ou plutôt conformité

Par choix, afin de limiter notre recherche, nous ne nous sommes pas intéressé à des définitions de l'analogie où l'on s'interrogerait sur une définition de l'égalité différente de l'ordinaire. Nous sommes parti d'une vision dure de l'égalité. Cela ne signifie pas que nous rejetons comme non pertinente toute recherche qui s'interrogerait sur l'égalité. Notre souci a été pratique : il faut bien commencer par quelque chose pour y voir plus clair, et nous avons simplement choisi de nous concentrer sur la signification du rapport dans l'analogie entre chaînes de symboles plutôt que sur des significations alternatives de l'égalité. Il semble bien que l'intuition que des linguistes comme Paul, Saussure, Bloomfield ou Mounin, ont de l'analodgie est que l'articulation entre A et B d'une part et C et D d'autre part est une égalité avec toutes les propriétés qu'elle possède.

En fait, si dans un premier temps cette attitude se défend pour l'étude des analogies isolées, la confrontation avec les analogies de la langue montre que cette position doit être révisée. Dans le cas des analogies entre chaînes de symboles des exemples formels montreront qu'il faudra renoncer à la transitivité de l'égalité (p. 113). Autrement dit, l'égalité dans une analogie entre

chaînes de symboles conforme à l'intuition de Paul ou de Saussure, n'est pas une égalité au sens mathématique du terme !

Un de nos rapporteurs ayant pointé le caractère fort ennuyeux de l'usage d'un signe par trop bien défini en mathématiques, nous nous sommes résigné à abandonner ce signe malgré notre réticence à nous écarter de la notation usuelle en linguistique. À partir de maintenant, nous utiliserons le signe $\doteq$. Les deux lectures classiques d'une analogie entre quatre objets A, B, C et D, que nous noterons donc désormais $A : B \doteq C : D$, à savoir « A est à B comme C est à D » et « A est à B ce que C est à D » nous indiquent qu'il faut lire ce signe « comme » ou « ce que ». La première lecture, « comme », plus couramment utilisée dans la métaphore est à désigner par le terme de comparaison. La seconde lecture, « ce que », la seule que nous retiendrons pour l'analogie, nous incite à ne plus parler d'égalité mais de conformité[18].

Redisons qu'une égalité est une relation d'équivalence, c'est-à-dire une relation réflexive, symétrique et transitive. Dès lors, on peut se poser la question de savoir à quel type de relation correspond notre nouveau signe de conformité, $\doteq$, puisqu'il dénote une relation réflexive et symétrique, mais pas transitive. Schreider parle de relation de tolérance[19]. Nous reprendrons son terme. Il part de la notion intuitive de ressemblance pour aboutir à la formalisation de relation de tolérance. Toute chose se ressemble nécessairement, et si A ressemble à B, alors B ressemble aussi à A. Mais, si A ressemble à B et B ressemble à C, A ne ressemble pas nécessairement à C. À partir de là, il pose donc la définition suivante :

DÉFINITION **3 (Relation de tolérance)** *Soit $\mathcal{E}$ un ensemble, et $\sim$ une relation dans $\mathcal{E}$. $\sim$ est une relation de tolérance si et seulement si*

- $\forall A \in \mathcal{E},\ A \sim A$ *(réflexivité)*
- $\forall (A, B) \in \mathcal{E}^2,\ A \sim B \Leftrightarrow B \sim A$ *(symétrie)*

3.4.2 Rapport ou raison

En linguistique, le rapport est noté par deux points. Il en est ainsi en phonétique où les oppositions ou mutations sont notées à l'aide de ce symbole. Il en est ainsi aussi en morphologie où les formes d'un même paradigme qui s'opposent sont notées de la même façon[20].

Or, nous avons vu qu'il pouvait éventuellement y avoir une notion de causalité dans l'opposition. C'est le cas des changements phonétiques où l'opposition exprime parfois deux états d'un même phonème, l'un ayant entraîné l'autre. La commutation, l'opposition ou le rapport entretiennent donc un lien avec la notion de cause ou de raison. Ces deux notions sont précises en philosophie. Nous donnons ici les explications lumineuses de Deleuze :

[18] Le terme *identité* semblait un bon choix, car c'est une notion distincte de l'égalité. Malheureusement, il a un sens précis aussi bien en mathématique qu'en philosophie..

[19] Шрейдер (You. A. SCHREIDER), Равенство, сходство, порядок, 1975, p. 81 à 103.

[20] Voir par exemple tous les écrits de KURYŁOWICZ cités en bibliographie.

> Quelle différence y-a-t-il entre la raison suffisante et la cause ? On comprend très bien. La cause n'est jamais suffisante. Il faut dire que le principe de causalité pose une cause nécessaire, mais pas suffisante. Il faut distinguer la cause nécessaire et la raison suffisante. Qu'est-ce qui les distingue de toute évidence ? C'est que la cause d'une chose c'est toujours autre chose. La cause de A c'est B, la cause de B c'est C, etc. Série indéfinie des causes. La raison suffisante, ce n'est pas du tout autre chose que la chose. La raison suffisante d'une chose, c'est la notion de la chose. Donc la raison suffisante exprime le rapport de la chose avec sa propre notion tandis que la cause exprime le rapport de la chose avec autre chose. C'est limpide.[21]

Pour reprendre ces définitions, dans une équation analogique $A : B \doteq C : D$ d'inconnue D on peut voir que la cause de D sera B (et aussi C) car D sera construit à partir de B et de C. Mais en plus, D sera aussi par raison avec B (et aussi C). Pour nous, comme nous le soulignions plus haut (p. 92) aucune connaissance extérieure aux seules données d'une équation analogique entre chaînes de symboles ne devra intervenir dans sa résolution. Le rapport de D avec sa notion est donc entièrement exprimé par son rapport avec B (et aussi C), au travers de la conformité donnée par l'analogie. Cela puisque le rapport de A à C est égal au rapport de B à D, et que le rapport de A à B est égal au rapport de C à D. Sans aucun jeu de mots donc, la raison de D sera aussi son rapport à B (et à C).

3.5 Nouvelle introduction à notre propos : notions constitutives de l'analogie entre chaînes de symboles

Pour faire suite à ce que nous annoncions plus haut (p. 92), notre souhait aurait été de présenter une formalisation de l'analogie entre chaînes de symboles, ainsi qu'une règle de résolution sur les chaînes de symboles, qui aurait été aussi simple que la règle de trois. Nous aurions voulu réaliser ce but en montrant que les rapports peuvent se caractériser par la distance, notion « duale » de la similarité, tandis que les produits auraient été caractérisés par une répartition ou un ordre, notion qui aurait été « duale » de la contiguïté. Malheureusement, nous ne sommes en mesure de ne remplir que la première partie de ce souhait.

Le versant de la contiguïté reste un mystère pour nous. Nous nous interrogeons sur la relation qui pourrait exister entre cette notion et celle de produit ou de multiplication. En effet, nous avons rappelé que la règle de trois possédait une expression équivalente en terme de multiplications. Nous montrerons (p. 148) que le rapport entre deux chaînes de symboles peut être exprimé par la distance. Le parallèle avec la règle de trois tendrait à nous faire dire que la contiguïté, ou une notion duale ou opposée, comme la distance l'est

[21] Deleuze, *Sur Leibniz*, 1980.

à la similarité, devrait sans doute refléter une sorte de « multiplication » entre chaînes de symboles. Hélas, nous sommes là dans le domaine du spéculatif.

Qui plus est, nous ne savons toujours pas si les deux notions précédentes de similarité et de contiguïté ne pourraient pas être en quelque sorte cristallisées dans une distance particulière qui capturerait à la fois la distance et la répartition de deux chaînes de symboles, et qui nous permettrait de réaliser notre vœu d'échanger similarité et contiguïté (p. 116) comme tous nos schémas nous invitent à le faire.

Partie II

Formalisation

Chapitre 4

Relation d'analogie

Cette partie est le cœur de notre travail. Il s'agit de formaliser l'analogie entre chaînes de symboles. Rappelons qu'évidemment nous nous plaçons dans le cas

> où « analogie » a le sens authentique de proportion mathématique à quatre termes (a:b::c:d).[1]

c'est-à-dire dans l'acception donnée comme emploi spécial en mathématiques dans le *Trésor de la langue française* :

> **B.** *Emplois spéc.*. (Correspond à *analogique*).
>
> **1.** *MATH.* Proportion, identité de deux rapports ; d'où en scolastique *analogie de proportionnalité*.[2]

Soulignons de nouveau que, même si nous avons un souci de généralité, nous nous intéressons essentiellement aux chaînes de symboles. On verra que nous nous intéresserons aussi aux ensembles et aux multi-ensembles, mais il en sera ainsi pour l'unique raison que cela nous servira pour les chaînes de symboles.

Nous nous proposons de faire passer l'analogie entre chaînes de symboles d'un statut intuitif mais inutilisable automatiquement, à celui d'une opération désormais réapplicable aveuglément et donc aussi reproductible. Ce passage permettra, à notre avis, deux changements de grandeurs :

- premièrement, en quantité d'analogies calculables. On sait que la puissance actuelle des ordinateurs, et corrélativement, la mise à disposition d'importants corpus de textes, a rendu la linguistique plus semblable que jamais à une autre science de terrain. Notre espoir est que la possibilité d'application de l'analogie à un grand nombre de cas contribue à la découverte de phénomènes linguistiques nouveaux ;

[1] de LIBERA, *La philosophie médiévale*, 1992, p. 91.

[2] Institut National de la Langue Française, *Trésor de la langue française informatisé*, 2000, article ANALOGIE. Selon ce dictionnaire, l'adjectif *analogique* ne se rapporte qu'à cette acception.

- deuxièmement, en longueur de chaînes. Alors que tout le monde sait de tête résoudre des analogies sur des mots, la formalisation et son automatisation permettront son application à des morceaux de langues plus importants. Il deviendra possible d'envisager son application aux phrases, voire aux textes. L'application à des symboles moins intuitifs (comme par exemple les oppositions phonétiques) devient elle aussi possible.

Une possiblité immédiate d'application est la contestation de la prémisse de Chomsky (voir page 76) sur la pauvreté, grammaticalement parlant, de l'analogie. Nous pourrons compter massivement le nombre d'analogies, entre phrases, présentes dans un corpus (voir page 279). Cela est d'un autre ordre que la simple application de l'analogie entre mots à des fins de morphologie. Ainsi, nous pourrons chiffrer l'intuition d'Hermann Paul et de Bloomfield sur la pertinence de l'analogie pour la syntaxe.

Pour parvenir à nos fins, il nous faut d'abord suivre seulement les conséquences logiques d'idées sûres qui se présentent à notre esprit de façon claire. Mais, au cours de ce travail, il nous faudra bien avoir recours à l'expérience pour conforter les idées premières que nous énoncerons à propos de l'analogie entre chaînes de symboles.

Pour ce faire, parce que notre intuition de l'analogie entre chaînes de symboles est souvent justifiée par son usage entre mots (et aussi pour des raisons bassement matérielles de place !), nous illustrerons souvent notre travail en morphologie. Là aussi les difficultés sont nombreuses ; l'une des non moindres est de réussir à rendre compte simultanément de la morphologie flexionnelle de langues comme le latin (essentiellement par suffixation) et de la morphologie dérivationnelle de langues sémitiques comme l'arabe (qui se fait par infixation multiple).

Tout au cours de cet essai de formalisation, nous garderons en mémoire les grandes notions mises en évidence lors de notre exposé historique. Tout particulièrement, nous montrerons comment la notion du semblable trouve son expression mathématique dans la fonction de distance. Nous ne pouvons hélas pas en dire autant de la contiguïté. Mais, nécessairement, un jour, ces deux notions devront se confondre ou se recouvrir afin que la métaphore et la métonymie fusionnent en un seule notion : l'analogie. Il nous faudra donc, d'après Aristote lui-même, quelque part dans le *Traité des songes* être comme les fous, puisque « les fous disent et pensent le contigu dans le semblable ».

4.1 Esquisse d'une formalisation et conséquences

L'esquisse de formalisation de l'analogie que nous allons maintenant proposer s'appuie sur les résultats de notre étude historique. Nous avons en effet dégagé des **articulations constitutives** à l'analogie (p. 103) et des **notions constitutives** (p. 100). Les premières se rapportent à la notion même d'analogie. Elles constituent donc en quelque sorte une **vue générale** sur l'analogie. Les secondes intéressent les objets qui interviennent dans une analogie. Elles constituent donc en quelque sorte une **vue spécifique** sur l'analogie. Pour reprendre plus en détail les résultats de notre étude historique, rappelons que, selon les articulations et les notions dégagées, qui découlent des définitions d'Aristote et de la tradition historique, une analogie fait toujours intervenir quatre objets, que nous avions décidé de noter A, B, C et D, et que l'analogie ou proportion est la conformité de leurs rapports, notée $A : B \doteq C : D$. Par définition, une analogie est *une conformité de rapports entre objets de même type*.

Notre exposé d'une esquisse de formalisation de l'analogie procédera donc logiquement du général au spécifique, c'est-à-dire de l'expression générale de l'analogie en tant que conformité, jusqu'à considérer les objets qui y interviennent, puis l'identité de leur type, en passant par les rapports. Disons déjà que plusieurs hypothèses fondamentales que nous allons énoncer maintenant sous forme de postulats tournent autour de la notion de symétrie ou d'inversion.

4.1.1 Articulations constitutives

Les propriétés de la vue générale concernent les articulations constitutives de l'analogie, en tant qu'elles sont indépendantes des objets intervenant dans les analogies. En conséquence, elles seront nécessairement des propriétés valables quel que soit le type de ces objets. C'est pourquoi nous allons commencer par elles. Nous avons vu que proportions et analogies étaient la même chose. Comme nombre de propriétés sont bien connues sur les proportions entre nombres, nous reprendrons ici les termes usuels[3].

Réflexivité de la conformité

L'analogie serait en premier lieu une *conformité*. Remarquons que nous ne disons pas entre quels objets cette conformité s'appliquerait. En fait, il faut garder à l'esprit le fait que l'analogie est une relation quaternaire et que l'introduction de la conformité dans la définition que nous avons esquissée plus haut est pour le moment une facilité de langage et une convention de notation.

Le premier postulat que nous poserons est une propriété que l'on souhaite avoir quel que soit le domaine des objets A, B, C et D intervenant dans une

[3] *Cf.* par exemple, l'encyclopédie *WebEncyclo*, édition Atlas, 1999, à la rubrique : proportion (mathématiques).

analogie.

> POSTULAT **1 (Réflexivité de la conformité)** *L'analogie* $A : B \doteq A : B$ *est vraie.*

Comme justification, il est simplement difficile d'imaginer qu'une telle analogie puisse ne pas être vérifiée toujours. On pourrait certes imaginer que cela soit le cas dans un univers où les objets changeant par exemple au cours du temps, certains de leurs rapports s'en trouveraient changés. Mais se poserait alors le problème de la permanence des objets A et A dans chacun des deux termes et donc de leur identité. Nous nous en tenons à une vue fort classique des choses.

Symétrie de la conformité

Relativement à la symétrie, une première hypothèse de bon sens est introduite implicitement par Aristote dans son utilisation de l'analogie. En français, il s'agirait de la symétrie du mot « comme » : si l'on peut dire que A est à B comme C est à D, alors on doit pouvoir dire que C est à D comme A est à B. Cette propriété exprime la symétrie de la conformité.

Par opposition, dans le commentaire d'Euclide par Henrion, nous n'avons pas trouvé mention de cette propriété. Peut-être n'est-ce pas si étrange que cela, puisque, dans le cas des proportions sur les nombres, on peut penser que cette propriété ne découle pas de l'analogie elle-même, mais de l'égalité utilisée dans ce cas.

> POSTULAT **2 (Symétrie de la conformité)** *Soit l'analogie* $A : B \doteq C : D$. *Alors, de façon équivalente, l'analogie suivante est aussi vraie :*
>
> $C : D \doteq A : B$ *par symétrie de la conformité*

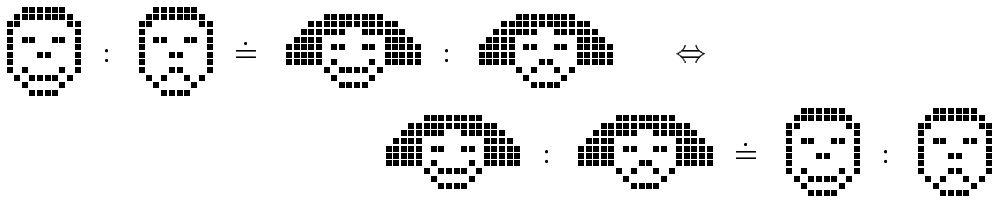
Si l'on reprend encore la phrase *les électrons sont au noyau atomique ce que les planètes sont au soleil*, on obtient : *les planètes sont au soleil ce que les électrons sont au noyau atomique*. Ici, le sens semble plus difficile, sans doute parce que cette analogie n'est déjà pas une analogie pure. Elle est déjà de l'ordre de la métaphore et il y a un sens dans le transfert effectué par une métaphore. La symétrie du « ce que » n'est donc pas forcément assuré. Ce qui est énoncé dans cette métaphore, c'est précisément le sens du transfert : « vous connaissiez déjà le système solaire, éh bien !, je vais vous apprendre quelque chose sur l'atome ».

Ce problème de sens (direction) n'apparaît pas dans des exemples entre chaînes de symboles ou entre images noir et blanc. Là, les analogies sont indépendantes de quelque métaphore que ce soit, et l'on admet sans mal toutes les équivalences suivantes :

$\square\bigcirc : \triangle\bigcirc \doteq \square\bullet : \triangle\bullet \quad \Leftrightarrow \quad \square\bullet : \triangle\bullet \doteq \square\bigcirc : \triangle\bigcirc$

prendre : *prends* $\doteq$ *vendre* : *vends* $\quad \Leftrightarrow \quad$ *vendre* : *vends* $\doteq$ *prendre* : *prends*

$$muslim : aslama \doteq mursil : arsala \quad \Leftrightarrow \quad mursil : arsala \doteq muslim : aslama$$



嫁 : 妙 ≒ 稼 : 秒 ⇔ 稼 : 秒 ≒ 嫁 : 妙 [4]

Transitivité de la conformité

On peut à juste titre se demander pourquoi il nous semble nécessaire de poser les deux postulats précédents. Si nous avions suivi l'intuition de la notation utilisée par les linguistes, à savoir avec le signe égal, il aurait en effet suffi de les faire dériver des propriétés de l'égalité. Dans une telle vue, où l'analogie noterait une vraie égalité, la réflexivité, la symétrie et la transitivité en serait aussi nécessairement trois propriétés découlant de ce que l'égalité est une relation d'équivalence. C'est bien là le cas pour l'analogie appliquée aux nombres, c'est-à-dire la règle de trois, où l'on a bien la transitivité.

$$\forall (A,B,C,D,E,F) \in \mathbb{Q}^{*6}, \quad \frac{A}{B} = \frac{C}{D} \ \wedge \ \frac{C}{D} = \frac{E}{F} \ \Rightarrow \ \frac{A}{B} = \frac{E}{F}$$

Malheureusement, il n'en va pas de même sur les chaînes de symboles. La transitivité n'y est pas assurée, du moins pour la formalisation que nous proposons ici[5]. Donnons un exemple de ce paradoxe. On a bien :

$$bab : b \doteq abab : ab \qquad \text{et} \qquad abab : ab \doteq aba : a$$

mais l'analogie suivante ne saurait être tenue pour vraie selon les mêmes critères.

$$bab : b \neq aba : a$$

[4] Japonais : 嫁 /yome/ (épouse ; bru ; mariée), 妙 /myou/ (subtile, étrange), 稼 /kase/ radical de 稼ぎ /kasegi/ (revenu, salaire), 稼ぐ /kasegu/ (gagner sa vie) et 秒 /byou/ (seconde [division de la minute de temps]). Ces kanjis sont composés des quatre éléments graphiques suivants : 女 /onna/ (femme) et 禾 /nogi/ caractère inutilisé actuellement en japonais, en partie gauche, qui commutent avec, en partie droite : 家 /ie/ (maison), 少 /syou/ (un peu). L'analogie est ici donnée en tant qu'analogie entre images composée d'éléments d'image carrés noirs ou blancs, comme pour l'analogie immédiatement au-dessus entre les dessins de visages. Les eléments d'image sont seulement beaucoup plus petits.

[5] Notre formalisation ne couvre pas toutes les interprétations possibles de l'analogie entre chaînes de symboles. Par exemple, nous verrons plus loin qu'elle écarte le phénomène de redoublement que l'on peut pourtant à juste titre considérer comme relevant de l'analogie en général.

La formalisation à laquelle nous sommes arrivé jusqu'à présent, si elle rend bien compte des deux premières analogies ne peut rendre compte de la troisième. Les deux premières analogies illustrent des jeux sur les préfixes et les suffixes, c'est-à-dire des commutations. La troisième analogie n'est pas de même espèce. On pourrait dire, si l'on cherchait à expliquer, qu'il s'agit de la substitution des a par des b et réciproquement.

Les propriétés de l'analogie particulière entre chaînes de symboles dont nous traitons, se trouvent donc à mi-chemin entre celles de la proportion entre nombres, avec réflexivité, symétrie et transitivité pour l'égalité et celle de la métaphore où, comme nous le montrions plus haut, l'opérateur « comme » n'étant plus ni symétrique, ni transitif, n'est plus guère que réflexif.

Tout ceci explique pourquoi, dans la notation $A : B \doteq C : D$, nous avons été obligé d'abandonner le signe égal de la tradition linguistique. En effet, il aurait été trompeur dans le cas des chaînes de symboles. La relation que nous avions désignée par le terme de conformité n'est que réflexive et symétrique. Il ne s'agit donc pas d'une relation d'équivalence, mais d'une relation de tolérance[6].

Permutations des moyens

Redisons que l'analogie est une conformité de *rapports*. Elle fait donc intervenir en second lieu des rapports.

Dans l'Éthique à Nicomaque, au livre V, au cours du raisonnement définissant le juste et qui suit le rappel de la définition de l'analogie, Aristote utilise la propriété suivante :

> la proportion étant une égalité des rapports et supposant quatre termes au moins. [...] Ce que le terme A, alors, est à B, le terme C le sera à D, et de là, par interversion[7], ce que A est à C, B l'est à D ;[8]

Il s'agit d'une propriété de permutation des moyens sur les proportions. Le terme est employé par Hermann Paul lorsqu'il l'applique aux mots, c'est-à-dire à des chaînes de symboles :

> Il peut aussi y avoir unité sonore dans les deux directions, par ex. *Tag* : *Tages* : *Tage* = *Arm* : *Armes* : *Arme* = *Fisch* : *Fisches* : *Fische* ; [...] soit, par permutation des moyens, possible avec toutes les proportions *Tag* : *Arm* : *Fisch* = *Tages* : *Armes* : *Fisches* etc.[9]

Cette propriété existait déjà pour les proportions entre longueurs chez Euclide :

---

[6] Шрейдер (You. A. Schreider), Равенство, сходство, порядок, 1975, p. 83.

[7] *τὸ ανὰλογον εναλλάξ*. Nous disons permutation des moyens.

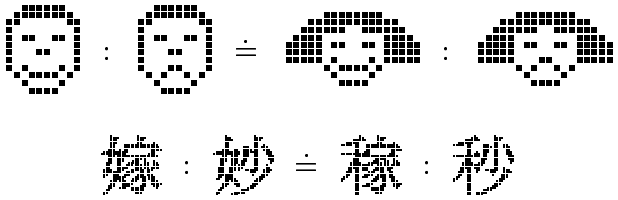
[8] Aristote, *Éthique à Nicomaque*, 1997 1er tirage 1990, V, 6, p. 229.

[9] Paul, *Prinzipien der Sprachgeschichte*, 1920 5e ed 1e ed 1880, chap. 5 Analogie, § 76, p. 107.

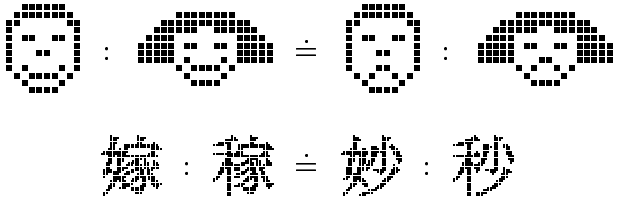
12. Raison alterne, est prendre l'antecedant comparé à l'antecedant, & le consequent au consequent.
Euclide explique en ceste def. & és suiuantes, aucuns moyens d'argumenter és proportions, desquels l'vsage est fort frequent en la Geometrie. Il dit donc icy que la raison alterne ou permutee, est quand de quatre grandeurs proportionnelles proposees, comme A, B, C, D, sçavoir que comme A est à B, ainsi C est à D; on vient à conclurre qu'il y a mesme raison de l'antecedant A à l'antecedant C, que du consequent B au consequent D: & ceste maniere d'argumenter (laquelle est demonstree à la 10. p. de ce liu.) se couche ordinairement ainsi: comme A est à B, ainsi C est à D: Donc en permutant comme A sera à C, ainsi B sera à D. Et est à noter qu'en ceste maniere d'argumenter, les quatre grandeurs doivent estre de mesme genre. [10]

À un niveau d'interprétation élevé, on peut reprendre la phrase fameuse *les électrons sont au noyau atomique ce que les planètes sont au soleil*, pour obtenir la phrase: *les électrons sont aux planètes ce que le noyau atomique est au soleil.* Cette permutation ne heurte pas le sens, mais elle semble de compréhension un peu plus difficile, chose qui tient au fait que cette analogie joue sur deux domaines, et que donc les objets qui y interviennent ne sont pas « de même genre » comme le dit justement Henrion. En fait, dans cet exemple particulier, la fonction de l'analogie est précisément de forcer, d'imposer l'homogénéité entre l'atome et le système solaire.

Pour ce qui est d'autres objets, comme les images noir et blanc, la permutation est tout-à-fait naturelle. Ainsi, les analogies



peuvent se reformuler de la façon suivante sans heurter le moins du monde le sens commun.



En conclusion, donc, de même qu'Aristote ou Paul tiennent pour évidente cette propriété, de même, nous l'admettrons, et par conséquent nous l'érigeons en postulat.

[10] EUCLIDE, *Les quinze livres des éléments géométriques d'Euclide : plus le livre des donnez... trad. en françois*, 1632, Élément cinquiesme, p. 175. Nous respectons toujours la graphie originale. C'est bien aussi la définition 12 dans EUCLIDE, ユークリッド原論, 1996, 第5巻, p. 93.

POSTULAT **3 (Permutation des moyens)** *Soit l'analogie* $A : B \doteq C : D$. *Alors, de façon équivalente, l'analogie suivante est aussi vraie:*

$$A : C \doteq B : D \qquad \textit{par permutation des moyens}$$

Ce postulat a pour conséquence immédiate que, dans le tableau 3.2 de la page 100, les notions de similarité et de contiguïté doivent pouvoir s'échanger. Pour ce faire, il nous faudrait soit dégager une notion qui les recouvre toutes deux, soit les faire intervenir simultanément.

4.1.2 Conséquences et reformulation

Formes équivalentes

On dérive les cinq formes équivalentes suivantes de l'analogie en utilisant les deux postulats précédents de permutation des moyens et de symétrie de la conformité.

LEMME **1** *Soit l'analogie* $A : B \doteq C : D$. *Alors, de façon équivalente, les cinq analogies suivantes sont vraies:*

$B : A \doteq D : C$	*par inversion des rapports;*
$B : D \doteq A : C$	*par inversion des rapports et permutation des moyens;*
$C : A \doteq D : B$	*par permutation des moyens et inversion des rapports;*
$D : B \doteq C : A$	*par permutation des extrêmes;*
$D : C \doteq B : A$	*par inversion des rapports et symétrie de la conformité.*

DÉMONSTRATION : Toutes ces formes équivalentes de l'analogie dérivent d'une seule forme de l'analogie par l'application d'une séquence de permutations des moyens et de symétrie de la conformité. L'inversion des rapports se dérive de la façon suivante : $A : B \doteq C : D \Leftrightarrow$ (permutation des moyens) $A : C \doteq B : D \Leftrightarrow$ (symétrie de la conformité) $B : D \doteq A : C \Leftrightarrow$ (permutation des moyens) $B : A \doteq D : C$. La permutation des extrêmes se dérive de la façon suivante : $A : B \doteq C : D \Leftrightarrow$ (permutation des moyens) $A : C \doteq B : D \Leftrightarrow$ (symétrie de la conformité) $B : D \doteq A : C \Leftrightarrow$ (inversion des rapports) $D : B \doteq C : A$. La dérivation des autres formes de l'analogie est donnée dans l'énoncé du lemme. CQFD

Au total, on peut donc lister huit formes équivalentes de la même analogie, quelle qu'elle soit. En respectant l'ordre alphabétique sur les noms des variables de termes A, B, C et D, on a :

THÉORÈME **1 (Formes équivalentes)** *Les huit analogies suivantes sont équivalentes :*

$A : B \doteq C : D$
$A : C \doteq B : D$ *(permutation des moyens)*
$B : A \doteq D : C$ *(inversion des rapports)*
$B : D \doteq A : C$
$C : A \doteq D : B$
$C : D \doteq A : B$ *(symétrie de la conformité)*
$D : B \doteq C : A$ *(permutation des extrêmes)*
$D : C \doteq B : A$ *(symétrie de lecture)*

Ces huit analogies peuvent chacune être placées au sommet d'un même cube (voir p. 118). En chacun des sommets, on peut s'arranger pour que l'inversion des rapports, la symétrie de la conformité et la permutation des extrêmes soient chacune portées par un axe différent[11]. Par exemple, à partir de $A : B \doteq C : D$, les axes horizontal, vertical et de profondeur respectivement. On obtient ainsi, au sommet de chacune des faces du cube, quatre analogies qui commencent chacune par un terme différent.

Remarquons que la propriété d'inversion des rapports se trouve déjà dans les définitions sur les proportions entre longueurs chez Euclide :

> 13. Raison inuerse ou transposee, est lors qu'on prend le consequent comme antecedent pour le comparer à l'antecedant, comme si c'estoit le consequent.
> *Comme si A est à B, ainsi que C est à D, nous infererons que par raison inuerse, comme B est à A, ainsi D est à C, c'est à dire les consequens aux antedeans. En cette sorte d'argumenter les autheurs parlent presque toujours ainsi : comme A est à B, ainsi C est à D : donc en changeant, ou au contraire, B sera à A, comme D à C. Cette maniere d'argumenter sera demonstrée au Corrolaire de la 4. proposition de ce livre.* [12]

Une dernière remarque. Les formes équivalentes de l'analogie permettent de comprendre pourquoi nous avions rejeté la transitivité du signe égal (voir p. 113). Elle entrainerait en effet des paradoxes. Tout d'abord, il semble impossible de refuser que certaines équations analogiques aient bien plusieurs solutions. Par exemple : $a : aa \doteq b : ba$ ou ab ou bien : $xy : xaby \doteq xaby : xababy$ ou $xaabby$. Une fois admis ce point, la transitivité impliquerait pour toute équation analogique ayant deux solutions D_1 et D_2 différentes par application des formes équivalentes de l'analogie, la dernière analogie suivante.

$$\left\{\begin{array}{l} A : B \doteq C : D_1 \\ A : B \doteq C : D_2 \end{array}\right. \Rightarrow \left\{\begin{array}{l} C : D_1 \doteq A : B \\ A : B \doteq C : D_2 \end{array}\right. \Rightarrow \quad C : D_1 \doteq C : D_2$$

$$\Rightarrow \quad C : C \doteq D_1 : D_2$$

Or, il semble intuitivement difficile d'admettre une telle analogie où, redisons-le, D_1 et D_2 sont des chaînes de symboles différentes. Elle signifierait que l'opération d'analogie pourrait, en quelque sorte, créer du désordre.

[11] Attention, les arêtes parallèles du cube ne représentent donc pas forcément les mêmes transformations.

[12] EUCLIDE, *Les quinze livres des éléments géométriques d'Euclide : plus le livre des donnez... trad. en françois*, 1632, Élément cinquiesme, p. 175. En graphie originale toujours. Aussi la définition 13 dans EUCLIDE, ユークリッド原論, 1996, 第5巻, p. 93.

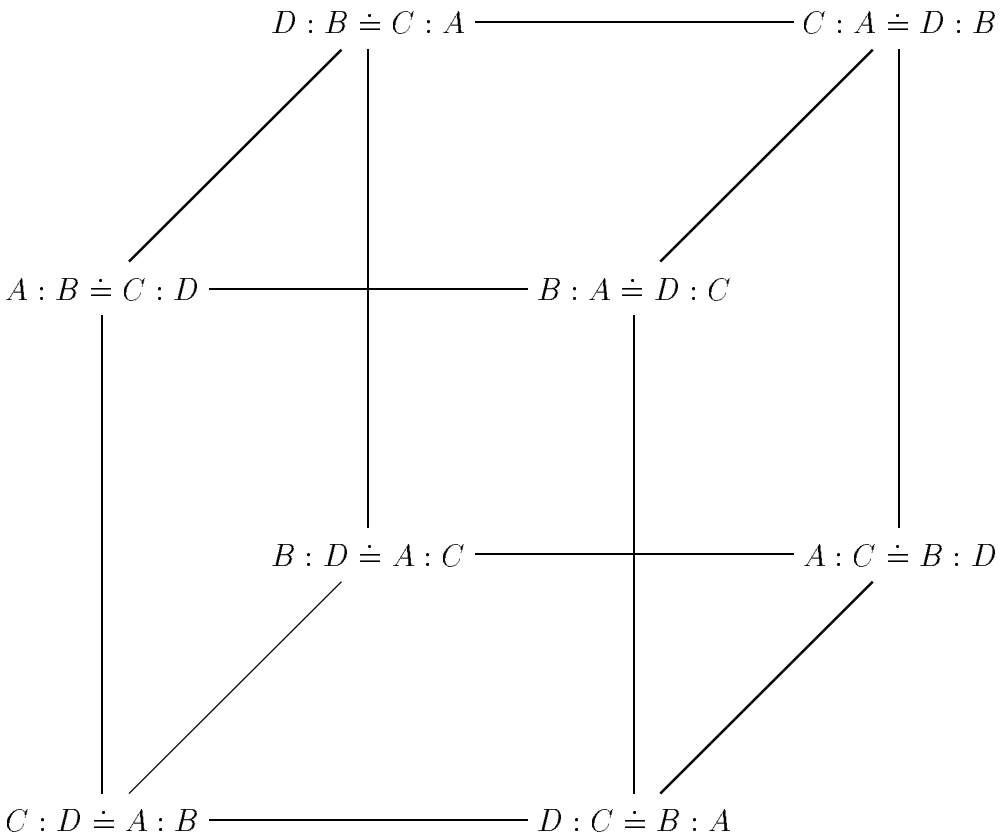


Figure 4.1: Le cube des analogies équivalentes

Structure induite et classes

La structure induite naturellement sur un ensemble quelconque par l'analogie peut être regardée à partir des deux points de vue naturels sur l'analogie. Le premier considère les analogies en tant que relation entre quadruplets. Le second part des équations analogiques définies, rappelons-le, par des triplets. On peut donc poser les deux définitions suivantes.

DÉFINITION **4** *Soit $\mathcal{E}$ un ensemble quelconque. On appelle structure induite par l'analogie sur $\mathcal{E}$, notée $(\mathcal{E}, \alpha)$, le sous-ensemble de $\mathcal{E}^4$ défini trivialement de la manière suivante.*

$$(A, B, C, D) \in (\mathcal{E}, \alpha) \Leftrightarrow \quad A : B \doteq C : D$$

DÉFINITION **5** *Soit $\mathcal{E}$ un ensemble quelconque. On appelle application induite*

par l'analogie, l'application, notée α, définie de la manière suivante[13].

$$\begin{array}{rccl} \alpha : & \mathcal{E}^3 & \to & \wp(\mathcal{E}) \\ & (A,B,C) & \mapsto & \{\, D \in \mathcal{E} \;/\quad A : B \doteq C : D \;\} \end{array}$$

Dès lors, nous pouvons passer à l'étude d'une telle structure. En particulier, on peut se demander quels types de quadruplets existent dans la structure. Par exemple, comme conséquence du théorème 1 (p. 116), on a :

LEMME **2** *Soit (A,B,C,D) un quadruplet d'éléments de quelqu'ensemble que ce soit, il n'existe que trois classes d'analogies possibles :*

$$A : B \doteq C : D \qquad ou \qquad A : C \doteq D : B \qquad ou \qquad A : D \doteq B : C$$

DÉMONSTRATION : Étant donnés quatre termes A, B, C et D, il y a $4! = 24$ permutations possibles de ces termes pour remplir les positions dans le moule $\sqcup : \sqcup \doteq \sqcup : \sqcup$. Le théorème des formes équivalentes sépare ces 24 permutations possibles en 3 ensembles de 8 analogies equivalentes. Les analogies suivantes peuvent être choisies comme représentantes de ces trois ensembles :

$$\begin{array}{c} A : B \doteq C : D \\ A : C \doteq D : B \\ A : D \doteq B : C \end{array}$$

Pour montrer que ces trois analogies ne sont pas équivalentes en général, il suffit d'exhiber un quadruplet de n'importe quel ensemble pour lequel une analogie seulement est vraie alors que les deux autres ne le sont pas. Par exemple, sur les chaînes de symboles, le quadruplet (*donner*, *donnons*, *marcher*, *marchons*) est tel que la première analogie est vraie, alors que les deux autres ne le sont pas. CQFD

Un cas particulier est le cas où tous les rapports sont symétriques sans entrainer nécessairement la conformité des facteurs. Dans ce cas, si une analogie est vraie pour un quadruplet donné, alors l'une quelconque des 24 permutations possibles permet de former une analogie vraie, et donc, les rapports sont égaux pour les éléments qui interviennent dans ces analogies.

THÉORÈME **2** *Soit $\mathcal{E}$ un ensemble quelconque.*

$$(i) \quad \forall (A,B) \in \mathcal{E}^2, \quad A : B = B : A$$

est équivalent à :

$$(ii) \quad \forall (A,B,C,D) \in \mathcal{E}^4, \quad A : B \doteq C : D \quad \Leftrightarrow \quad A : C \doteq D : B \quad \Leftrightarrow \quad A : D \doteq B : C$$

[13] On note par $\wp(\mathcal{E})$ l'ensemble des parties de $\mathcal{E}$.

DÉMONSTRATION : $(i) \Rightarrow (ii)$.
$A : B \doteq C : D \quad \Leftrightarrow \quad A : C \doteq B : D$ (permutation des moyens) $\Leftrightarrow \quad A : C \doteq D : B$ (hypothèse de la définition des rapports). Cette dernière analogie est la représentante de la deuxième classe d'analogies.
De façon similaire, $A : B \doteq C : D \quad \Leftrightarrow \quad A : B \doteq D : C$ (hypothèse de définition des rapports) $\Leftrightarrow \quad A : D \doteq B : C$ (permutation des moyens). Cette dernière analogie est la représentante de la troisième classe d'analogies.
$(i) \Leftarrow (ii)$.
Trivialement, pour tout A et B de n'importe quel ensemble, $A : B \doteq A : B$.
Par application de $A : B \doteq D : C \quad \Leftrightarrow \quad A : B \doteq C : D$, avec $C = B$ et $D = A$, on obtient : $A : B \doteq B : A \quad \Leftrightarrow \quad A : B \doteq A : B$. CQFD

Révision des deux postulats

De façon à donner une certaine logique à notre système de postulats, nous proposons de remplacer le postulat de permutation des moyens par celui d'inversion des rapports. De cette façon, deux des postulats généraux que nous imposons à l'analogie, *conformité* de *rapports* entre objets de même type, se révèlent être des postulats de conservation de l'analogie par inversion des articulations constitutives de l'analogie :

- inversion du sens de *la conformité* : $A : B \doteq C : D \quad \Leftrightarrow \quad C : D \doteq A : B$
- inversion des *rapports* : $A : B \doteq C : D \quad \Leftrightarrow \quad B : A \doteq D : C$

Si cette formulation permet de gagner en élégance, ces deux postulats à eux seuls ne permettent pas de réobtenir les huit formes équivalentes de l'analogie. Dans la figure 4.1 (p. 118), en partant de $A : B \doteq C : D$, on reste coincé sur la face avant du cube. Pour en sortir, et atteindre la face arrière du cube, il est nécessaire de rajouter la permutation des moyens ou la permutation des extrêmes. Cela est évidemment critiquable car, généralement, on cherche à avoir un nombre minimal de postulats.

- permutation des extrêmes : $A : B \doteq C : D \quad \Leftrightarrow \quad D : B \doteq C : A$

4.1.3 Notions constitutives

Nous nous tournons maintenant vers les notions constitutives de l'analogie. L'esquisse de formalisation que nous allons proposer a trait aux propriétés spécifiques qui jouent sur les objets apparaissant dans une analogie. Mais nous allons rester général dans le spécifique. En effet, comme le remarquait Euclide, pour pouvoir mesurer les rapports, il est nécessaire que les objets intervenant dans une analogie soit de même type. Donc, les postulats que nous allons proposer maintenant se fonderont sur cette hypothèse, mais ne la dépasseront pas. Ce ne sera qu'ultérieurement, dans les développements

spécifiques, que nous spécialiserons ces deux postulats aux types particuliers des objets auxquels nous nous intéresserons, à savoir les ensembles, les multi-ensembles et les chaînes de symboles.

Contiguïté et inversion des objets

Ayant vu que les deux premiers postulats généraux jouent tous deux sur la notion d'inversion, la tentation est grande de suivre les définitions mot-à-mot pour pousser l'idée plus avant. Puisque l'analogie est conformité de rapports entre *objets* de même type, il est en effet tentant de rechercher un troisième postulat qui serait une inversion des objets.

Or, chez Euclide, commenté par Henrion, une forme spécifique d'une telle hypothèse pour les segments de droite existe. C'est la proposition 19 :

> 19. Si le tout est au tout, comme le retranché au retranché; le reste sera aussi au reste, comme le tout est au tout.
> Soit la toute AB à la toute DE, comme le retranché AC, au retranché DF. Ie dis que la reste CB est aussi au reste FE, comme la toute AB à la toute DE.
> [14]

Il s'agit de ce que l'on peut retrancher un segment d'un segment plus grand et que l'on peut considérer son reste, c'est-à-dire son complément dans le plus grand. Si l'on transposait aux ensembles, en termes ensemblistes, on pourrait écrire que, si $A : B \doteq A \setminus E : B \setminus F$, alors $E : F \doteq A : B$. Nous abstrayons cette idée dans une expression plus générale où nous proposons de prendre les restes des objets par rapport à une référence plus grande. Si, pour tout objet, le reste par rapport à cette référence peut être défini, on a affaire à une application bijective dans le domaine de référence.

POSTULAT **4 (Conservation par inversion des objets)** *Soit $\mathcal{E}$ un ensemble quelconque muni d'une bijection de lui-même dans lui-même, notée par l'exposant -1, telle que, pour tout A de $\mathcal{E}$, A^{-1} est l'élément de $\mathcal{E}$, différent de A, appelé inverse de A et le plus contigu à A. Soient A, B, C et D des éléments de $\mathcal{E}$. Soit l'analogie $A : B \doteq C : D$. Alors, de façon équivalente, l'analogie suivante est aussi vraie:*

$A^{-1} : B^{-1} \doteq C^{-1} : D^{-1}$ *par conservation par inversion des objets*

Nous sommes bien conscient que l'expression de ce postulat est défectueux puisque la bijection en question n'est pas vraiment définie. En effet, que signifie « le plus contigu » ? À notre plus grand regret nous ne pouvons que confesser ici les limites de nos travaux. Pour le moment, nous sommes incapables d'exprimer de façon précise les contraintes à imposer sur cette bijection. C'est que, comme nous l'avons déjà dit, notre formalisation du versant

[14] EUCLIDE, *Les quinze livres des éléments géométriques d'Euclide : plus le livre des donnez... trad. en françois*, 1632, Élément cinquiesme, p. 190. Aussi la proposition 19 dans EUCLIDE, ユークリッド原論, 1996, 第5巻, p. 110.

de la contiguïté de l'analogie n'est pas achevée. Mais notre intuition est que cette bijection exprime la contiguïté par antonymie ou contradictoire, conformément à tous les exemples que nous avons mentionnés dans notre introduction (p. 25 et 29). Autrement dit, pour nous, ce postulat devrait exprimer le cas extrême de la contiguïté. Dans la formalisation de l'analogie entre ensembles que nous proposons plus bas (voir p. 125) nous verrons dans cette bijection la complémentation au sens des ensembles. En tout cas, nous avons ici un postulat dont l'expression est semblable à celle des deux postulats généraux d'inversion de la conformité et d'inversion des rapports :

- « inversion » des *objets* : $A : B \doteq C : D \quad \Leftrightarrow \quad A^{-1} : B^{-1} \doteq C^{-1} : D^{-1}$

Similarité

Après avoir vu le versant malheureusement imparfait de la contiguïté, tournons-nous vers la seconde notion constitutive de l'analogie, la similarité. Nous avions déjà vu qu'Henrion insistait sur le fait qu'un rapport ne pouvait être établi qu'entre objets considérés sous un angle quantifiable (voir p. 38 pour la citation). Dès lors, il y a nécessairement, dans les objets intervenant dans une analogie, une dimension qui est comparable, et cette comparaison est faite en terme de rapports. Cela reflète le fait que nous abordons finalement ici le quatrième constituant de la définition d'une analogie : une conformité de rapports entre objets *de même type*. Dans tout objet intervenant dans une analogie, on doit donc pouvoir voir des propriétés dans la dimension selon laquelle les rapports seront établis. Nécessairement, pour que les rapports existent, il faut que les propriétés selon cette dimension soient partagés par les objets d'une façon ou d'une autre. C'est en ce sens qu'il y a similarité entre objets et c'est de cette façon qu'il peut exister une relation de cause entre les objets. Pour établir la façon de se retrouver des propriétés dans les objets d'une analogie, considérons les analogies sous leur aspect d'équations analogiques. Étant donné une équation analogique $A : B \doteq C : D$ d'inconnue D, nous désirons obtenir tous les D qui vérifient complètement l'équation sans rien laisser de côté, et seulement ceux-là. Autrement dit, ces D doivent épuiser les rapports de similarité que l'analogie exprime. Par conséquent, pour pouvoir former un D comme solution à l'équation analogique, sans avoir recours à aucune connaissance extérieure comme nous nous l'imposons, il est nécessaire que la similarité de A avec B et C soit elle aussi épuisée dans les deux rapports $A : B$ et $A : C$. En d'autres termes, pour nous, nécessairement, une analogie épuise complètement les rapports de similitude que A entretient avec B et C. Le postulat que nous proposons maintenant est la formulation de cette idée.

> POSTULAT **5 (Distribution)** *Soit l'analogie* $A : B \doteq C : D$, *toute propriété de* A *se retrouve dans* B *ou dans* C.

De nouveau, il faut bien comprendre que les propriétés dont il est question dans ce postulat sont celles qui font sens pour la relation d'analogie considérée, c'est-à-dire celles qui seront interprétables au sens des rapports entre objets

sous l'angle quantifié considéré. Les autres angles peuvent être passées sous silence. En cela, nous rejoignons une remarque de Cournot sur le jugement analogique :

> [...] la vue de l'esprit, dans le *jugement* analogique, porte uniquement sur les *rapports* et sur les raisons des *ressemblances*: les *ressemblances* sont de nulle valeur dès lors qu'elles n'accusent pas des *rapports* dans l'ordre des faits où l'*analogie* s'applique.[15]

En conclusion de l'exposé des postulats sur les notions constitutives de l'analogie, nous pouvons dire que les deux postulats considèrent les deux notions de contiguïté et de similarité chacune poussée à son extrême. De la même façon que le postulat que nous avions posé pour la contiguïté est la situation extrême de la contiguïté, celle de l'antonymie ou du contradictoire, de même, nous envisageons la similarité dans la situation extrême où elle est maximale dans les rapports existant entre les objets. En d'autres termes, non seulement, on peut tout dire de la contiguïté jusqu'à même considérer l'inversion des objets, mais en plus, il faut tout dire de la similarité jusqu'à ne plus pouvoir en dire rien, par épuisement.

4.1.4 Synthèse de l'esquisse d'une formalisation

Nous avons donc posé six postulats pour l'analogie. Mettons de côté le premier, qui correspondait à la réflexivité de la conformité, et la permutation des extrêmes que nous avions été contraint de rajouter pour atteindre les huit formes équivalentes d'une analogie. Chacun des quatre autres restants se concentre sur l'un des constituants de la définition d'une analogie : une *conformité* (I) de *rapports* (II) entre *objets* (III) de *même type* (IV). Deux de ces postulats (I et II) concernent les articulations constitutives de l'analogie. Ce sont des postulats généraux. Les deux autres (III et IV) concernent les notions constitutives de l'analogie. Ils sont donc spécifiques, au sens où ils concernent les objets. En plus, ils sont extrêmes au sens des deux notions impliquées. Trois de ces postulats (I et II et III) s'expriment sous forme d'une inversion soit de la conformité (I), soit des rapports (II), soit des objets (III). Nous rappelons tous ces postulats pour mémoire dans le tableau suivant.

[15] Cournot, *Essais sur les fondements de nos connaissances et sur les caractères de la critique philosophique*, 1851, p. 68, cité dans Institut National de la Langue Française, *Trésor de la langue française informatisé*, 2000, article ANALOGIQUE.

Tableau 4.1: Esquisse d'une formalisation de l'analogie

(O)	reflexivité de *la conformité*	$A : B \doteq A : B$
		$A : B \doteq C : D$
(I)	inversion du sens de *la conformité*	$C : D \doteq A : B$
(II)	inversion des *rapports*	$B : A \doteq D : C$
(III)	inversion des *objets*	$A^{-1} : B^{-1} \doteq C^{-1} : D^{-1}$
(IV)	distribution dans les *objets*	toute propriété de A se retrouve dans B ou dans C
(V)	permutation des extrêmes	$D : B \doteq C : A$

4.2 Théorèmes sur les ensembles

Après avoir posé les postulats sur l'analogie générale, et donné quelques résultats préliminaires, nous allons, dans cette section, spécialiser notre propos au cas où les termes de l'analogie sont des ensembles.

4.2.1 Distribution

Pour des ensembles, les « propriétés » sont simplement les éléments de l'ensemble, et « se retrouver » est simplement le fait d'appartenir au sens ensembliste du terme. On peut alors donner, sur les ensembles, l'expression suivante au postulat de distribution.

LEMME **3 (Distribution sur les ensembles)** *Soient quatre ensembles A, B, C et D.*

$$A : B \doteq C : D \quad \Rightarrow \; A \subset B \cup C$$

Le lemme de distribution implique la propriété remarquable suivante.

LEMME **4 (Égalité des unions des moyens et des extrêmes)** *Soient quatre ensembles A, B, C et D.*

$$A : B \doteq C : D \quad \Rightarrow \; A \cup D \; = \; B \cup C$$

DÉMONSTRATION : Les huit formes équivalentes de l'analogie permettent d'obtenir huit formes de l'hypothèse de distribution sur les ensembles. Au total, on a donc huit inclusions, dont seulement quatre, par commutativité de l'union, sont distinctes. Ces quatre inclusions impliquent et sont impliquées par deux inclusions réciproques qui donnent l'égalité.

$$\left\{ \begin{array}{l} A \subset B \cup C \\ B \subset A \cup D \\ C \subset A \cup D \\ D \subset B \cup C \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} A \cup D \subset B \cup C \\ B \cup C \subset A \cup D \end{array} \right. \Leftrightarrow \; A \cup D \; = \; B \cup C$$

CQFD

4.2.2 Inversion des objets

Nous exploitons maintenant le postulat d'inversion des objets.

Pour des ensembles, l'inversion peut se voir comme la prise du complémentaire. La question est donc de définir l'ensemble de référence dans lequel on prendra ce complémentaire. Comme une analogie est entièrement définie par la donnée de quatre objets, c'est-à-dire de quatre ensembles pour le cas présent, quels que soient les ensembles considérés, on peut toujours prendre leur réunion pour ensemble de référence. Tout cela permet d'énoncer la forme suivante du postulat d'inversion des objets.

LEMME **5 (Conservation par complémentation à l'union)** *Soient quatre ensembles A, B, C et D. On pose $E = A \cup B \cup C \cup D$.*

$$A : B \doteq C : D \quad \Leftrightarrow \quad (E \setminus A) : (E \setminus B) \doteq (E \setminus C) : (E \setminus D)$$

Cela implique facilement la propriété suivante.

LEMME **6 (Égalité des intersections des moyens et des extrêmes)** *Soient quatre ensembles A, B, C et D.*

$$A : B \doteq C : D \quad \Rightarrow \; A \cap D \; = \; B \cap C$$

DÉMONSTRATION : On pose $E = A \cup B \cup C \cup D$, et on applique l'égalité des unions des moyens et des extrêmes à l'analogie $(E \setminus A) : (E \setminus B) \doteq (E \setminus C) : (E \setminus D)$. Cela donne l'égalité :

$$(E \setminus A) \cup (E \setminus D) = (E \setminus B) \cup (E \setminus C)$$

équivalente à :

$$E \setminus (A \cap D) \; = \; E \setminus (B \cap C)$$

et, parce que E est un sur-ensemble de A, B, C et D, et donc de $A \cap D$ et $B \cap C$, on a :

$$A \cap D \; = \; B \cap C$$

CQFD

Au passage, et pour mieux faire comprendre la forme que prend une analogie entre ensembles, notons les deux corollaires suivants dont les démonstrations sont triviales étant donné le lemme précédent.

COROLLAIRE **1** *Soient quatre ensembles A, B, C et D.*

$$A : B \doteq C : D \quad \Rightarrow \; A \supset B \cap C$$

COROLLAIRE **2** *Soient quatre ensembles A, B, C et D.*

$$A : B \doteq C : D \quad \Rightarrow \; (A \cap D) \setminus (A \cap B \cap C \cap D) = \emptyset$$

4.2.3 Résolution d'équations analogiques

Avec les deux lemmes d'égalité précédents, on est en mesure de résoudre le problème de l'analogie entre ensembles. Cherchons donc, étant données les propriétés vues ci-dessus, à construire l'ensemble D solution de l'équation analogique $A : B \doteq C : D$.

On a déjà vu que le lemme 3 (p. 125) impliquait le lemme d'égalité des unions des moyens et des extrêmes $A \cup D = B \cup C$. On peut donc écrire :

$$(A \cup D) \setminus A = (B \cup C) \setminus A$$

Comme, d'autre part, on peut écrire :

$$(A \cup D) \setminus A = D \setminus A = D \setminus (A \cap D)$$

On a donc : $D = ((B \cup C) \setminus A) \cup (A \cap D)$. Il est aussi facile de montrer que le corollaire 1 (p. 126) conjugué aux formes équivalentes de l'analogie entraine le lemme d'égalité des intersections des moyens et des extrêmes $A \cap D = B \cap C$. Cela permet d'obtenir :

$$D = ((B \cup C) \setminus A) \cup (B \cap C)$$

Comme la formule précédente est toujours possible quels que soient les ensembles A, B et C, l'ensemble D peut toujours être construit.

De cette construction, on tire le théorème suivant qui résout le problème de l'analogie entre ensembles.

THÉORÈME **3** *Soient trois ensembles A, B et C. L'équation analogique $A : B \doteq C : D$ d'inconnue D a une solution si et seulement si $A \subset B \cup C$ et $A \supset B \cap C$. La solution, unique, est alors :*

$$D = ((B \cup C) \setminus A) \cup (B \cap C)$$

Faisons une interprétation des deux conditions d'inclusion. Nous essayons de nous rattacher aux conceptions de Kruszewski, de Morier et d'Itkonen presentées plus haut (voir 59, p. 99 et 100). Que A soit inclus dans $B \cup C$ est en une condition de ressemblance. Elle est même l'expression extrême du postulat 4.1.3 (p. 122) de distribution puisque tout A se retrouve dans B et dans C. Elle devrait donc se rattacher à la métaphore. Or, selon Morier, la forme schématisante de la métaphore devrait être l'intersection. Maintenant, que A contienne $B \cap C$ est une condition de contiguïté, c'est-à-dire de l'ordre de la métonymie, dont la forme schématisante selon Morier est l'inclusion. Alors que dans les conceptions de tous les auteurs cités, la métaphore et la métonymie formaient deux axes perpendiculaires (voir les tabeaux de la page 100), ici, selon nos vœux (voir p. 110), nous les avons fusionnées pour les appliquer simultanément sur deux axes orthogonaux grâce à l'inversion des objets.

Tableau 4.2: Cristallisation du contigu et du semblable

	← similarité → & contiguïté	
↑ similarité & contiguïté ↓	A	B
	C	D

Revenons à la démonstration du théorème précédent.

DÉMONSTRATION : On pose (i) $A : B \doteq C : D$, (ii) le système d'inclusions, et (iii) l'égalité $D = ((B \cup C) \setminus A) \cup (B \cap C)$. On va en fait montrer $(i) \Rightarrow (ii) \wedge (iii)$ et $(ii) \wedge (iii) \Rightarrow (i)$.

$(i) \Rightarrow (ii)$ On a déjà vu que le lemme 3 (p. 125) de distribution sur les ensembles et le corollaire 1 (p. 126) donnent cette implication. On vient aussi de donner la construction de D.

$(ii) \wedge (iii) \Rightarrow (i)$ Étant donnés trois ensembles A, B et C, on peut toujours construire l'ensemble :

$$D = ((B \cup C) \setminus A) \cup (B \cap C)$$

On va maintenant montrer que la relation $A : B \doteq C : D$ vérifie les cinq postulats fondamentaux de l'analogie.

Postulat de permutation des moyens : Ce postulat est trivialement vérifiée puisque B et C jouent le même rôle dans les deux conditions et dans la définition de D.

Postulat de réflexivité de la conformité : Il s'agit de vérifier que $A : B \doteq A : B$ quels que soient les ensembles A et B. Cela est vrai car on vérifie bien les deux conditions $A \subset B \cup A$ et $A \supset B \cap A$. D'autre part, B peut bien s'écrire comme la solution de l'équation analogique $A : B \doteq A : x$.

$$B = ((B \cup A) \setminus A) \cup (B \cap A)$$

Postulat de symétrie de la conformité : Il s'agit de vérifier que $C : D \doteq A : B$, c'est-à-dire que B peut bien s'écrire comme la solution de cette analogie.

$$B = ((D \cup A) \setminus C) \cup (D \cap A)$$

Tout d'abord, il faut vérifier que les conditions pour l'équation analogique $A : B \doteq C : D$ d'inconnue D impliquent bien celles pour l'équation analogique $C : D \doteq A : B$ d'inconnue B.

Donnons le tableau des valeurs de vérité de $D = ((B \cup C) \setminus A) \cup (B \cap C)$. Les lignes marquées par un triangle ne peuvent être retenues, car elles ne vérifient pas les deux conditions $A \subset B \cup C$ et $A \supset B \cap C$. Pour toutes les autres lignes, les deux conditions $C \subset D \cup A$ et $C \supset D \cap A$ sont bien vérifiées.

A	B	C	$B \cup C$	$(B \cup C) \setminus A$	$B \cap C$	$D =$ $((B \cup C) \setminus A) \cup (B \cap C)$
0	0	0	0	0	0	0
0	0	1	1	1	0	1
0	1	0	1	1	0	1
0	1	1	1	1	1	1 △
1	0	0	0	0	0	0 △
1	0	1	1	0	0	0
1	1	0	1	0	0	0
1	1	1	1	0	1	1

Grâce aux valeurs de D données par le tableau précédent, on peut alors calculer le tableau des valeurs de vérités de $((D \cup A) \setminus C) \cup (D \cap A)$. On vérifie dans ce tableau que, en dehors des lignes marquées, les valeurs de B et de $((D \cup A) \setminus C) \cup (D \cap A)$ sont bien les mêmes.

A	B	C	D	$D \cup A$	$(D \cup A) \setminus C$	$D \cap A$	$((D \cup A) \setminus C) \cup (D \cap A)$
0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0
0	1	0	1	1	1	0	1
0	1	1	1	1	0	0	0 △
1	0	0	0	1	1	0	1 △
1	0	1	0	1	0	0	0
1	1	0	0	1	1	0	1
1	1	1	1	1	0	1	1

Postulat de distribution : Il s'agit précisément, comme on l'a exposé au début de cette section, de la première inclusion de (ii).
Postulat d'inversion des objets : Il s'agit de vérifier que, si l'on a $A : B \doteq C : D$, alors, en posant $E = A \cup B \cup C \cup D$, on a bien $E \setminus A : E \setminus B \doteq E \setminus C : E \setminus D$. Autrement dit, il faut prouver la proposition suivante :

$$E \setminus (((B \cup C) \setminus A) \cup (B \cap C)) = ((E \setminus B \cup E \setminus C) \setminus (E \setminus A)) \cup (E \setminus B \cap E \setminus C)$$

Appelons D' le second terme de cette égalité. Ici aussi, il suffit de dresser le tableau des valeurs de vérité. La colonne $E \setminus D$ est obtenue à partir de la colonne D du premier tableau précédent. La colonne D' est obtenue en reportant à partir du premier tableau précédent les lignes de D correspondant aux valeurs des ensembles $E \setminus A$, $E \setminus B$, $E \setminus C$, et $E \setminus D$. On constate l'égalité dans les colonnes $E \setminus D$ et D'.

A	B	C	D	$E \setminus A$	$E \setminus B$	$E \setminus C$	$E \setminus D$	D'
0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0
0	1	0	1	1	0	1	0	0
0	1	1	1 △	1	0	0	0	0 △
1	0	0	0 △	0	1	1	1	1 △
1	0	1	0	0	1	0	1	1
1	1	0	0	0	0	1	1	1
1	1	1	1	0	0	0	0	0

CQFD

Donnons une expression du résultat de la résolution d'analogie entre ensembles sous forme de phrase :

> Dans une analogie entre ensembles, l'intersection des moyens est nécessairement l'intersection des quatre termes. De plus, un extrême contient tous les éléments des moyens qui n'appartiennent pas à l'autre extrême.

4.2.4 Représentations graphiques

D'après tous les résultats précédents, on peut figurer la résolution d'une équation analogique entre ensembles au moyen du diagramme 4.2. On peut aussi représenter une analogie par le diagramme 4.3. Dans le cas $A = B$, on a évidemment $C = D$, et dans ce deuxième diagramme, les ensembles $A \cap C$ et $B \cap D$ deviennent vides.

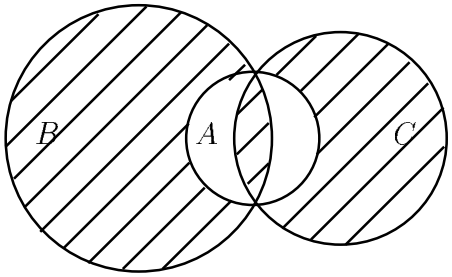


Figure 4.2: Diagramme de Venn de la résolution d'équations analogiques entre ensembles. La solution D est en hachuré

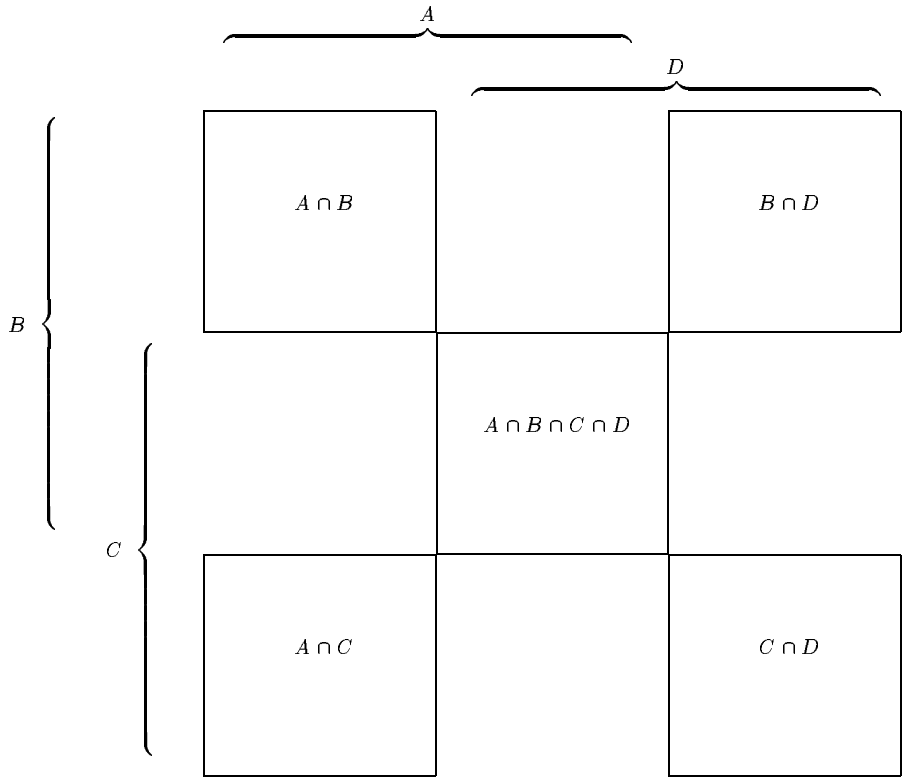


Figure 4.3: Diagramme à la Carroll de l'analogie entre ensembles

4.2.5 Égalité sur les cardinaux

Nous établissons un résultat remarquable sur l'analogie entre ensembles finis. La raison de notre intérêt est qu'il aura son pendant sur les chaînes de symboles.

THÉORÈME **4 (Égalité des sommes des cardinaux)**

$$A : B \doteq C : D \quad \Rightarrow \quad |A| + |D| = |B| + |C|$$

DÉMONSTRATION : On a déjà vu que $(A \cap D) \setminus (A \cap B \cap C \cap D) = \phi$ et $(B \cap C) \setminus (A \cap B \cap C \cap D) = \phi$. On peut donc écrire, comme le montre la figure 4.3 (p. 130) :

$$\begin{cases} A = (A \cap B) \vartriangle (A \cap C) \vartriangle (A \cap B \cap C \cap D) \\ B = (B \cap A) \vartriangle (B \cap D) \vartriangle (A \cap B \cap C \cap D) \\ C = (C \cap A) \vartriangle (C \cap D) \vartriangle (A \cap B \cap C \cap D) \\ D = (D \cap B) \vartriangle (D \cap C) \vartriangle (A \cap B \cap C \cap D) \end{cases}$$

Comme on a affaire a des différences symétriques, on peut écrire en prenant les cardinaux : $|A| = |A \cap B| + |A \cap C| + |A \cap B \cap C \cap D|$ et de même pour les trois autres ensembles B, C et D. D'où, par commutativité de l'intersection sur les ensembles et de l'addition sur les entiers :

$$\begin{array}{rcl} |A| + |D| & = & |A \cap B| + |A \cap C| + |A \cap B \cap C \cap D| \\ & & +|D \cap B| + |D \cap C| + |A \cap B \cap C \cap D| \\ & = & |B \cap A| + |B \cap D| + |A \cap B \cap C \cap D| \\ & & +|C \cap A| + |C \cap D| + |A \cap B \cap C \cap D| \\ & = & |B| + |C| \end{array}$$

CQFD

4.2.6 Treillis ensembliste

Nous mentionnons maintenant un résultat qui a trait aux treillis ensemblistes. En fait, nous ne nous servirons pas du tout de ce résultat dans nos applications, mais nous pensons qu'il est important pour tout le courant du traitement automatique des langues qui se fonde sur la notion d'unification. Il découle d'un théorème remarquable sur les ensembles, qui est en fait une forme élémentaire d'analogie.

THÉORÈME **5** *Soient B et C deux ensembles quelconques,*

$$B \cap C : B \doteq C : B \cup C$$

DÉMONSTRATION : On applique simplement le théorème 3 (p. 127) donnant la solution pour les analogies ensemblistes. Les deux conditions $B \cup C \subset B \cup C$ et

$B \cup C \supset B \cap C$ étant trivialement vérifiées, la solution de l'équation analogique $B \cap C : B \doteq C : D$ d'inconnue D est donnée par :

$$\begin{aligned} D &= (B \cup C \setminus (B \cap C)) \cup (B \cap C) \\ &= B \cup C \end{aligned}$$

CQFD

Comme l'on sait, une structure de traits partielle, où seules certains traits sont mentionnés, est en réalité une notation pour l'ensemble des structures de traits complètes avec tous les traits possibles réalisés[16]. L'unification de structures de traits partielles est par définition, et paradoxalement, l'intersection ensembliste des ensembles de structures de traits complètes représentés par ces notations. Ce résultat permet de montrer que l'unification entre structures de traits, opération fondamentale pour tout un courant du traitement automatique des langues, repose en fait sur une analogie. Dans cette analogie, intervient la généralisation, qui est la seconde opération donnant aux structures de traits la structure de treillis au sens mathématique du terme[17]. Si on note par $\wedge$ la généralisation de structures de traits et par $\vee$ leur unification, on a le résultat remarquable suivant qui découle directement du théorème précédent :

THÉORÈME **6** *Soient f_1 et f_2 deux structures de traits, alors*

$$f_1 \vee f_2 : f_1 \doteq f_2 : f_1 \wedge f_2$$

4.2.7 Synthèse des résultats

Il faut remarquer que dans tout le développement précédent sur les ensembles, nous avons implicitement utilisé la notion de rapport, mais sans jamais l'expliciter. Arrivé en fin de ce développement, nous ne savons toujours pas ce qu'est le rapport entre deux ensembles intervenant dans une analogie, alors même que nous savons maintenant vérifier une analogie entre ensembles, et résoudre une équation analogique entre ensembles.

Le diagramme de Caroll donne une représentation intéressante de l'analogie entre ensembles. En effet, il met bien en évidence que, pour qu'une analogie existe entre ensembles, la partie commune à deux des ensembles l'est forcément aussi aux autres. Ce ne sera plus le cas pour les chaînes de symboles. En effet, dans *muslim* : *aslama* $\doteq$ *mursil* : *arsala*, les deux premières chaînes ont les symboles *s*, *l* et *m* en commun, mais le symbole *m* n'apparaît pas dans *arsala*.

[16] KAY, *Functional unification grammar: a formalism for machine translation*, 1984.

[17] AÏT-KACI, *A lattice theoretic approach to computation based on a calculus of partially ordered type structures*, 1984.

4.3 Théorèmes sur les multi-ensembles

La présente section fait en quelque sorte le lien entre les ensembles et les objets qui nous intéressent vraiment, à savoir les chaînes de symboles. En effet, les multi-ensembles peuvent être vus comme un intermédiaire entre ces deux types d'objets. Nous allons donc étudier ici la spécialisation au cas où les termes de l'analogie sont des multi-ensembles. Il faut tout d'abord donner la définition d'un multi-ensemble.

DÉFINITION **6 (Multi-ensemble)** *On appelle* multi-ensemble *toute application* f *d'un ensemble fini* $\mathcal{E}$ *quelconque dans* $\mathbb{N}^*$. *On appelle* $\mathcal{E}$ *l'ensemble de départ et on le note* $\mathrm{D}(\mathrm{f})$.

Comme toute application, un multi-ensemble f est donc entièrement défini par le sous-ensemble de $\mathcal{E} \times \mathbb{N}^*$ des éléments de la forme $(a, \mathrm{f}(a))$ pour tout $a \in \mathcal{E}$. À titre d'exemple, le multi-ensemble

$$\{(i,1),(l,1),(m,1),(n,1),(r,1),(s,1),(u,2)\}$$

représente les occurrences des lettres dans le mot *mursilun*. Dans ce cas, l'ensemble $\mathcal{E}$ est l'ensemble $\{i,l,m,n,r,s,u\}$ de lettres ou symboles.

4.3.1 Distribution

Dans une analogie entre multi-ensembles, les propriétés sont les éléments des ensembles, mais « pondérés » par leur image par l'application définie par le multi-ensemble. Pour une analogie $A : B \doteq C : D$ entre multi-ensembles, le postulat de distribution exprime donc le fait que les éléments de l'ensemble de départ de A se retrouvent soit dans l'ensemble de départ de B, soit dans celui de C, soit dans les deux à la fois, mais pondérés par leur image dans les applications respectives. On désire donc écrire :

$$\forall a \in \mathrm{D}(A),\ \ A(a) \leq B(a) + C(a)$$

Dans cette écriture, $B(a)$ ou $C(a)$ peuvent ne pas être définis, c'est pourquoi il est nécessaire d'étendre tout multi-ensemble A à une application de l'ensemble de tous les éléments possibles de l'univers $\mathcal{U}$ dans $\mathbb{N}$, c'est-à-dire $\mathbb{N}^* \cup \{0\}$.

DÉFINITION **7 (Multi-ensembles étendus)** *Soit* A *un multi-ensemble. Il est naturellement étendu à l'application de* $\mathcal{U}$ *dans* $\mathbb{N}$ *de la façon suivante.*

$$\forall a \in (\mathcal{U} \setminus \mathrm{D}(A)),\ \ A(a) = 0$$

Avec cette écriture, on a :

$$A(a) \neq 0 \quad \Leftrightarrow \quad a \in \mathrm{D}(A)$$

Grâce à cela, l'écriture de l'inégalité précédente devient possible, et l'on peut donc donner le lemme suivant comme expression du postulat de distribution sur les multi-ensembles.

LEMME **7 (Distribution sur les multi-ensembles)** *Soient quatre multi-ensembles A, B, C et D.*

$$A : B \doteq C : D \quad \Rightarrow \ \forall a \in \mathrm{D}(A),\ A(a) \ \leq \ B(a) + C(a)$$

4.3.2 Inversion des objets

Le postulat d'inversion des objets est difficile à exprimer sur les multi-ensembles, car, au contraire des ensembles, pour lesquels un élément appartient simplement ou n'appartient pas à un ensemble, dans le cas des multi-ensembles, l'appartenance est pondérée par un entier non nul. C'est pourquoi, nous laissons ouverte la question de la formulation de ce postulat aux multi-ensembles, et nous redisons encore une fois notre regret du fait que le versant de la contiguïté n'est pas achevé.

4.3.3 Résolution d'équations analogiques

En revanche, la notion de rapports entre multi-ensembles peut être définie intuitivement comme la différence des pondérations entre images de chaque élément de départ. Cela peut se définir de la façon suivante :

DÉFINITION **8** *Soient A et B deux multi-ensembles. On définit le rapport entre deux multi-ensembles comme l'application de* $\mathrm{D}(A) \cup \mathrm{D}(B)$ *dans* $\mathbb{Z}$ *définie de la façon suivante.*

$$\forall a \in \mathrm{D}(A) \cup \mathrm{D}(B),\ \ (A : B)(a) = A(a) - B(a)$$

À partir de cette définition, il est facile d'expliciter l'analogie entre multi-ensembles de la façon suivante.

THÉORÈME **7 (Analogie entre multi-ensembles)** *Soient A, B, C et D quatre multi-ensembles. L'analogie* $A : B \doteq C : D$ *est vraie si et seulement si*

$$\forall a \in (\mathrm{D}(A) \cup \mathrm{D}(B) \cup \mathrm{D}(C) \cup \mathrm{D}(D)), A(a) - B(a) = C(a) - D(a)$$

Ce théorème peut se réexprimer très simplement sous une forme qui aura son pendant sur les chaînes de symboles. En reprenant la terminologie déjà vue pour les proportions, et que nous avions aussi rencontrée chez Henrion, on peut parler des moyens et des extrêmes. On a alors l'expression suivante.

THÉORÈME **8 (Égalité des sommes sur les moyens et les extrêmes)** *Soient A, B, C et D quatre multi-ensembles. L'analogie* $A : B \doteq C : D$ *est vraie si et seulement si*

$$\forall a \in (\mathrm{D}(A) \cup \mathrm{D}(B) \cup \mathrm{D}(C) \cup \mathrm{D}(D)),\quad A(a) + D(a) = B(a) + C(a)$$

Cette forme a l'avantage sur la précédente de rester en quelque sorte dans le monde des multi-ensembles, puisque la somme de deux entiers est un entier. On peut donc bien voir dans $A + D$ et dans $B + C$ deux multi-ensembles. Dans son expression, cette forme rappelle aussi le théorème 4 (p. 131) sur les ensembles, qui énonçait que, s'il y a analogie, alors les sommes des cardinaux des moyens et des extrêmes sont égales.

Le théorème précédent nous permet de résoudre très facilement des équations analogiques entre multi-ensembles. En effet, on a immédiatement.

THÉORÈME **9** *Soient quatre multi-ensembles A, B, C et D. L'équation analogique $A : B \doteq C : D$ a une solution si et seulement si*

$$\forall a \in (\mathrm{D}(A) \cup \mathrm{D}(B) \cup \mathrm{D}(C)), \quad A(a) \leq B(a) + C(a)$$

La solution, unique, est alors le multi-ensemble dont l'ensemble de départ est :

$$\mathrm{D}(D) = \{a \in (\mathrm{D}(A) \cup \mathrm{D}(B) \cup \mathrm{D}(C)) \ / \ A(a) < B(a) + C(a)\}$$

et dont les images sont données par :

$$\forall a \in \mathrm{D}(D), \ \ D(a) = B(a) + C(a) - A(a)$$

4.3.4 Synthèse des résultats

Contrairement au cas des ensembles, où nous n'avions pas explicité les rapports, c'est l'explicitation directe de ceux-ci qui nous a permis de résoudre le problème de l'analogie entre multi-ensembles. Dans le cas des chaînes de symboles qui fait l'objet de la partie suivante, nous nous retrouverons dans le cas des ensembles, où la notion de rapport ne s'offre pas directement.

Il est intéressant de remarquer que le cas des ensembles apparaît heureusement comme un cas particulier des multi-ensembles. En effet, un ensemble est un multi-ensemble pour lequel l'ensemble des images possibles par l'application est réduit au singleton $\{1\} \subset \mathbb{N}^+$. On peut alors réécrire le tableau des valeurs de D en appliquant la formule $D(a) = B(a) + C(a) - A(a)$. Les seules résultats admissibles pour ce calcul sont 0 ou 1. Tout autre valeur, négative ou supérieure strictement à 1, signifie que l'équation analogique *entre ensembles* n'a pas de solution. Pour un élément a, on a, dans dans la table 4.3 (p. 136), huit possibilités seulement énumérées dans les colonnes représentant les ensembles A, B et C.

Dans cette table, on a marqué les deux seules lignes ne remplissant pas la condition. Comme nous avons vu plus haut (p. 128), chacune est équivalente à une proposition logique sur les ensembles. La ligne (i) est équivalente à $\neg(B \cap C \subset A)$. Toutes les autres lignes vérifient bien $B \cap C \subset A$. La ligne (ii), elle, est équivalente à $\neg(A \subset B \cup C)$. Ce sont là les deux conditions donnée dans le théorème 3 (p. 127) pour qu'une équation analogique entre ensembles ait une solution. Ainsi donc, les deux conditions pour qu'une solution à une équation analogique entre ensembles existe ont été fusionnées dans la condition unique sur les multi-ensembles.

Tableau 4.3: Valeurs données par la formule de résolution d'équations analogiques entre ensembles et valeurs obtenues si les ensembles sont considérés comme des multi-ensembles

A	B	C	$((B \cup C) \setminus A) \cup (B \cap C)$	$D(a) = B(a) + C(a) - A(a)$	
0	0	0	0	0 + 0 - 0 = 0	
0	0	1	1	0 + 1 - 0 = 1	
0	1	0	1	1 + 0 - 0 = 1	
0	1	1	1 △	1 + 1 - 0 = 2	(i)
1	0	0	0 △	0 + 0 - 1 = -1	(ii)
1	0	1	0	0 + 1 - 1 = 0	
1	1	0	0	1 + 0 - 1 = 0	
1	1	1	1	1 + 1 - 1 = 1	

On vérifie aussi que les valeurs obtenues par le calcul sur les multi-ensembles sont bien les mêmes que celles obtenues par la formule directe de résolution d'une équation analogique entre ensembles : $D = (B \cap C) \cup (B \cup C \setminus A)$. Le fait que le cas des ensembles soit un cas particulier du cas des multi-ensembles se retrouvera lorsque nous donnerons des algorithmes pour la résolution des équations analogiques (voir p. 211 et suivantes).

4.4 Théorèmes sur les chaînes de symboles

Nous spécialisons maintenant au cas où les termes de l'analogie appartiennent à un ensemble de chaînes de symboles.

4.4.1 Observations

Versant de la similarité

Notre première observation a été d'essence négative, puisque nous nous sommes intéressé à ce qui ferait que certaines équations analogiques ne peuvent être résolues. En linguistique, on connaît des exemples où, à un état antérieur de la langue, il existait une analogie entre des formes de certains mots, analogie devenue impossible par action du changement phonétique. Le phénomène inverse existe aussi, qui aligne au moyen de la résolution d'une équation analogique les formes d'un certain mot sur les formes d'un autre mot. Ces deux aspects de l'analogie, d'une part **égalité donnée** $honos : honosem \doteq orator : oratorem$, mais également **équation à résoudre** $oratorem : orator \doteq honorem : x \quad \Rightarrow \quad x = honor$, sont les deux volets constitutifs de l'explication de l'analogie par Saussure que nous avons vue plus haut (page 61).

Différentes langues, l'allemand, l'arabe (en transcription latine), le français, le japonais, le latin et le malais, nous ont fourni des exemples d'équations analogiques insolubles dont certaines relevaient de l'un des deux types mentionnés. Après examen, nous avons émis la conjecture que certaines équations ne pouvaient être résolues pour la raison qu'un symbole au moins de la première chaîne n'apparaissait ni dans la deuxième ni dans la troisième chaîne. Nous avons alors pu élargir nos exemples d'équations analogiques insolubles à des exemples sans rapport avec la phonétique.

Nous donnons ci-après un certain nombre de ces exemples. Afin de faire apparaître clairement notre conjecture, nous y distinguons[18] les symboles de la première chaîne qui n'appartiennent ni à la deuxième ni à la troisième.

$$\text{Allemand :} \quad \ddot{u}bersetzen : setzte \doteq lachen : x \quad [19]$$

Ce premier exemple n'a rien à voir avec un quelconque changement phonétique. Il est en revanche particulièrement clair, parce que trivial, comme illustration de notre conjecture. L'envie est grande de retrancher *über* à *lachen*, mais comment ? C'est bien là ce qui bloque la résolution de l'équation. En revanche, on ne voit aucune difficulté à faire : $\ddot{u}bersetzen : setzte \doteq \ddot{u}berlachen : x \quad \Rightarrow \quad x = lachte.$ [20]

$$\text{Arabe marocain :} \quad bab : bwiyeb \doteq kelb : x \quad [21]$$

[18] Par la couleur rouge.

[19] *Übersetzen* (traduire), *setzte* (il ou elle a posé), *lachen* (rire).

[20] *Überlachen* (rire exagérément), *lachte* (il ou elle a ri). L'analogie est ici purement formelle, elle n'existe pas en sens.

[21] *Bwiyeb* (une petite porte) est le diminutif de *bab* (une porte). *Kelb* (un chien).

Kelb (un chien) a bien un diminutif *kliyeb*, qui ne peut cependant être considéré comme la solution de l'équation analogique donnée ici, à cause de la différence de voyelle entre *a* et *e*. Pour résoudre le problème, on serait tenté de revenir à la forme de l'arabe classique *kalb*, en reposant l'équation analogique comme nous l'autorisent les formes équivalentes de l'analogie vues plus haut (page 116), *bwiyeb* : *bab* ≐ *kliyeb* : *x*. Mais là encore, un *w* en trop interdit la résolution.

Français : *cheval* : *chevaux* ≐ *étau* : *x*

Les deux premiers termes illustrent la formation du pluriel des noms masculins en *al*. Le pluriel *étaux* ne saurait s'expliquer par ce modèle, car *étau* ne se termine pas par *al*. Plus précisément, et c'est ce qui nous intéresse ici, *l* n'apparaît pas dans ce mot.

Japonais : 飛びます : 飛べます ≐ 飲みます : *x* [22]

Les hiraganas, caractères japonais, notent des syllabes du type consonne-voyelle. Le caractère び /bi/ n'est présent ni dans 飛べます, ni dans 飲みます. L'analogie ne saurait donc être résolue **à la seule vue** des caractères. Il est nécessaire de connaître la **prononciation** de ces mots pour former le quatrième terme. En effet, la transposition phonétique de cette équation analogique est résolue sans problème : /tobimasu/ : /tobemasu/ ≐ /nomimasu/ : *x* ⇒ *x* = /nomemasu/, parce que /b/ et /i/ apparaissent bien tous deux dans /tobemasu/ et /nomimasu/.

Latin : *honos* : *honorem* ≐ *orator* : *x* [23]

Honorem est la forme obtenue par effet du rhotacisme (voir plus haut, p. 60). Alors que l'on avait bien, avant rhotacisme *honos* : *honosem* ≐ *orator* : *oratorem*, la présence du *s* de *honos*, insensible au rhotacisme, car en position finale, brise le parallèle avec la déclinaison d'*orator*.

Malais : *sewa* : *penyewa* ≐ *main* : *x* [24]

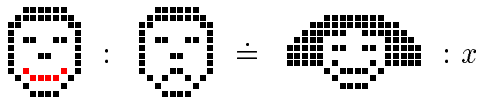
Un joueur se dit *pemain* en malais, mais cette forme ne peut s'expliquer par dérivation selon le modèle *sewa* : *penyewa* car *s* n'est présent ni dans *penyewa* ni dans *main*. Les descriptions du malais postulent donc couramment un archiphonème noté *N* à la fin du préfixe *peN* qui forme les noms d'agent, et posent que cet archiphonème donne *ny* en s'agglutinant avec *s*, *ng* en s'agglutinant avec *k*, etc., et disparaît devant *m*, *n*, *r*, etc. (voir 347). De cette façon, la formation des noms d'agent devient analogiquement régulière : *sewa* : *peNsewa* (> *penyewa*) ≐ *main* : *x* ⇒ *x* = *peNmain* (> *pemain*). On observera que l'introduction de l'archiphonème provoque la réapparition de *s* dans *peNsewa*, ce qui autorise par là-même la résolution.

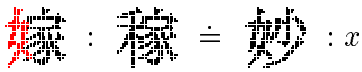
[22] 飛びます /tobimasu/ (voler) construit son potentiel 飛べます /tobemasu/ (pouvoir voler) en changeant la voyelle /i/ du radical en /e/. De même pour 飲みます /nomimasu/ (boire).

[23] *Honos* (honneur) au nominatif fait son accusatif en *honorem*, *orator* (un orateur).

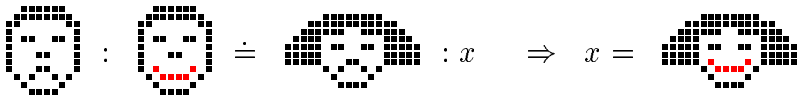
[24] *Sewa* ([prendre à] louer), *penyawa* (locataire). *Main* (jouer).

On peut encore observer que notre conjecture s'applique au cas des images. En effet, si pour un « alphabet » d'éléments d'images contenant trois éléments $\{$, , $\}$, on a les équations analogiques suivantes, dans lesquelles une partie seulement de la première image est rouge, la bouche pour la première équation, la partie gauche du kanji dans la seconde :





aucune solution n'est possible, car les éléments d'image n'apparaissent ni dans la seconde ni dans la troisième image. Mais les équations analogiques suivantes peuvent, elles, être résolues.



Une forme plus contrainte de notre conjecture énoncerait que, si tous les caractères de A apparaissent dans B et C, mais pas dans le même ordre, alors, là non plus, il n'y a pas de solution à l'équation analogique $A : B \doteq C : x$. On peut le vérifier simplement sur l'exemple suivant :

$$\textit{velours} : \textit{élevé} \doteq \textit{sourd} : x$$

Un cas extrême, et problématique pour nous, est celui où tous les caractères de la première chaîne apparaissent dans l'ordre inverse pour former la seconde chaîne, si bien que la seconde chaîne est le miroir de la première. La question est de savoir si, par exemple,

$$\textit{écart} : \textit{tracé} \doteq \textit{un} : x \quad \Rightarrow \quad x = \textit{nu}$$

est une analogie acceptable. Autrement dit, de façon générale, en notant μ l'opération miroir sur les chaînes admet-on toujours :

$$A : \mu(A) \doteq B : x \quad \Rightarrow \quad x = \mu(B) \,?$$

Le miroir des chaînes est une transformation « régulière » de l'ordre. On sent donc bien que notre conjecture serait peut-être bien valide à toute transformation régulière de l'ordre près. Mais comme nous sommes bien en peine pour le moment de préciser de quelle régularité il pourrait s'agir, nous sommes forcé d'abandonner cette piste pour nous limiter à la transformation « la plus régulière » possible qui est l'absence de transformation, c'est-à-dire la simple conservation de l'ordre.

Revenons à l'expression de notre conjecture. La contraposée de l'observation précédente, qui disait qu'aucune solution à une équation analogique $A : B \doteq C : x$ n'existait si des symboles de A n'apparaissaient ni dans B ni dans C, est que tout symbole de A doit apparaître dans B ou dans C. Nous nous proposons donc maintenant de vérifier cette hypothèse sur des exemples.

Nous listons donc ci-après des exemples d'analogies valides dans les mêmes langues que précédemment. Mais contrairement à ce qui précède, nous distinguons maintenant les apparitions, dans l'ordre, des symboles de la première chaîne dans les deuxième et troisième chaînes.

setzen : *setzte* ≐ *lachen* : *lachte* [25]
lang : *längste* ≐ *scharf* : *schärfste* [26]
fliehen : *er floh* ≐ *schließen* : *er schloß* [27]
sprechen : *ihr aussprächet* ≐ *nehmen* : *ihr ausnähmet* [28]
kennen : *gekannt* ≐ *brennen* : *gebrannt* [29]

aslama : *arsala* ≐ *muslimun* : *mursilun* [30]
kataba : *kātib* ≐ *sakana* : *sākin* [31]
huzila : *huzāl* ≐ *ṣudi'a* : *ṣudā'* [32]
kalb : *kulaib* ≐ *masjid* : *musaijid* [33]
yaṣilu : *yaṣala* ≐ *yasimu* : *yasama* [34]

科学 : 科学家 ≐ 政治 : 政治家 [35]

[25] *Setzen* (poser, asseoir), *lachen* (rire) sont des infinitifs. *Setzte* (il ou elle a posé), *lachte* (il ou elle a ri) sont au passé.

[26] *Lang* (long), *scharf* (aiguisé) sont des adjectifs dont les superlatifs sont les formes correspondantes *längste* (le plus long), *schärfste* (le plus aiguisé).

[27] *Fliehen* (s'enfuir), *schließen* (fermer) sont des infinitifs. *Floh* (il ou elle s'est enfuit), *schloß* (il ou elle a fermé) sont au passé.

[28] *Sprechen* (parler), *nehmen* (prendre) sont des infinitifs. *Ihr aussprächet* (que vous exprimiez), *ihr ausnähmet* (que vous mettiez à part) sont des dérivés des verbes précédents conjugués à la deuxième personne du pluriel, au subjonctif II.

[29] *Kennen* (savoir), *brennen* (brûler) sont des infinitifs de deux verbes dits faibles irréguliers dont le participe passé prend bien une terminaison régulière en *t* selon la conjugaison faible, mais dont la voyelle *e* du radical est infléchie en *a*.

[30] *Aslama* (il se convertit [à l'Islam]) et *arsala* (il envoya) sont des verbes au passé, 3ème personne du singulier. *Muslimun* (un converti [à l'Islam], c'est-à-dire un musulman) et *mursilun* (un envoyé) sont des noms.

[31] *Kataba* (il écrivit), *kātib* (un écrivain), *sakana* (il habita), *sākin* (un habitant).

[32] *Huzāl* (amaigrissement, affaiblissement), *ṣudā'* (mal de tête, migraine). Les deux autres formes sont les formes accomplies passives des verbes correspondants, construits sur les racines trilitères *hzl* et *ṣd'*. On a donc : *huzila* (a été amaigri, a été affaibli) et *ṣudi'a* (a attrapé un mal de tête, et, par extension ou peut-être contamination par une autre racine ?, a été pendu).

[33] *Kalb* (un chien), *masjid* (une mosquée). Les deux autres formes sont leur pluriel.

[34] *Yaṣilu* (il arrivait), *yasimu* (il marquait) sont des formes de l'imperfectif, alors que *yaṣala* (il arriva), *yasama* (il marqua) sont les formes correspondantes du perfectif.

[35] 科学 /kēxué/ (la science), 政治 /zhèngzhì/ (la politique). Le suffixe 家 /jiā/ construit un nom de profession : 科学家 /kēxuéjiā/ (un scientifique), 政治家 /zhèngzhìjiā/ (un homme politique).

我 : 我們 ≐ 他 : 他們 [36]
今年 : 今天 ≐ 明年 : 明天 [37]
我是中国人 : 我是学生 ≐ 他不是中国人 : 他不是学生 [38]
他們是很好的朋友 : 他們不是很好的朋友 ≐ 我去法国 : 我不去法国 [39]

inné : *nées* ≐ *indu* : *dues*
réaction : *réactionnaire* ≐ *répression* : *répressionnaire*
aimer : *ils aimaient* ≐ *marcher* : *ils marchaient*
joindre : *je joins* ≐ *rejoindre* : *je rejoins*
logique : *logiciel* ≐ *ludique* : *ludiciel*

食べます : 食べる ≐ 決めます : 決める [40]
痛い : 痛む ≐ 親しい : 親しむ [41]
あれ : これ ≐ あっち : こっち [42]
乗る : 乗せる ≐ 寄る : 寄せる [43]
自由 : 不自由な ≐ 用意 : 不用意な [44]

oratorem : *orator* ≐ *honorem* : *honor* [45]
facio : *conficio* ≐ *capio* : *concipio* [46]
amo : *amas* ≐ *oro* : *oras* [47]

---

[36]我 /wǒ/ (je) et 他 /tā/ (il) sont des pronoms personnels au singulier. 我們 /wǒmen/ (nous) et 他們 /tāmen/ (ils) sont les formes du pluriel. (們 est la forme traditionnelle du caractère utilisée à Taïwan.)

[37]今年 /jīnnián/ (cette année) et 今天 /jīntīan/ (aujourd'hui). 明年 /míngnián/ (l'année prochaine) et 明天 /míngtīan/ (demain).

[38]我是中国人 /wǒ shì zhōngguóren/ (je suis chinois). 我是学生 /wǒ shì xuésheng/ (je suis étudiant). 他不是中国人 /tā bù shì zhōngguóren/ (il n'est pas chinois). 他不是学生 /tā bù shì xuésheng/ (il n'est pas étudiant).

[39]他們是很好的朋友 /tāmen shì hěn hǎo de péngyou/ (ils sont de bons amis). 他們是很好的朋友 /tāmen bù shì hěn hǎo de péngyou/ (ils ne sont pas bons amis). 我去法国 /wǒ qù Fǎguó/ (je vais en France). 我不去法国 /wǒ bù qù Fǎguó/ (je ne vais pas en France).

[40]食べます /tabemasu/ (manger) et 決めます /kimemasu/ (décider) sont à la forme de politesse marquée. Leurs correspondants 食べる /taberu/ (manger) et 決める /kimeru/ (décider) sont les formes de politesse non marquée.

[41]痛い /itai/ (douloureux, qui fait mal) et 親しい /sitasii/ (familier, proche) sont des adjectifs (dits en *-i*). 痛む /itamu/ (être douloureux, faire mal) et 親しむ /sitasimu/ (être familier, être proche) sont des verbes.

[42]あれ /are/ (celui-là), これ /kore/ (celui-ci), あっち /atti/ (là-bas) et こっち /kotti/ (ici) font partie des démonstratifs de la série *ko-so-a*. On a aussi それ /sore/ et そっち /sotti/, qui réfèrent à un objet ou à un lieu partagé par le locuteur et l'interlocuteur.

[43]乗る /noru/ (monter [en voiture]), 寄る /yoru/ (tirer [à soi]). 乗せる /noseru/ (faire monter) et 寄せる /yoseru/ (faire s'approcher, attirer) sont des formes factitives.

[44]自由 /ziyuu/ (liberté), 用意 /youi/ (préparation) sont des noms à partir desquels sont construits les adjectifs suivants (en な /na/) exprimant la négation (préfixe 不 /hu/) de la possession de l'idée. 不自由な /huziyuuna/ (géné, privé de, handicapé), 不用意な /huyouina/ (imprévoyant, imprudent).

[45]*Orator* (un orateur), *honor* (l'honneur) sont au nominatif. *Oratorem* et *honorem* sont à l'accusatif.

[46]Les verbes *facio* (faire) et *capio* (prendre) présentent, dans leurs composés *conficio* (composer ; achever, accomplir) et *concipio* (recevoir, contenir), une inflexion de la voyelle *a* de leur radical en *i*.

[47]*Amo*, *amas* sont la première et deuxième personnes du singulier du présent de l'indicatif

amo : *amat* ≐ *oro* : *orat* [48]
amo : *amamus* ≐ *oro* : *oramus* [49]

tinggal : *ketinggalan* ≐ *duduk* : *kedudukan* [50]
pekerja : *kerja* ≐ *pelawat* : *lawat* [51]
kawan : *mengawani* ≐ *keliling* : *mengelilingi* [52]
isteri : *beristeri* ≐ *ilmu* : *berilmu* [53]
keras : *mengeraskan* ≐ *kena* : *mengenakan* [54]

biorąc : *bierzesz* ≐ *piorąc* : *pierzesz* [55]
ubezpieczony : *ubezpieczeni* ≐ *obrażony* : *obrażeni* [56]
śpiewać : *śpiewaczka* ≐ *łechtać* : *łechtaczka* [57]
wyszedłem : *wyszłaś* ≐ *poszedłem* : *poszłaś* [58]
rozproszyć : *rozpraszać* ≐ *rozmnożyć się* : *rozmnażać się* [59]

Nous pouvons lister aussi un certain nombre d'analogies formelles.

a : *aa* ≐ *aaa* : *aaaa*
aa : *ab* ≐ *ba* : *bb*
b : *ab* ≐ *aab* : *aaab*

---

du verbe *amare* (aimer). Le verbe *orare* (prier) se conjugue de la même façon.

[48] Les mêmes verbes à la troisième personne du singulier de l'indicatif présent.

[49] Les mêmes verbes à la première personne du pluriel de l'indicatif présent. Et il en va ainsi pour toute la conjugaison de ces verbes qui peut être mise en relation analogique.

[50] *Tinggal* (demeurer, habiter), *duduk* (asseoir, installer). *Ketinggalan* (rater, manquer), *kedudukan* (situation, grade).

[51] *Kerja* (un travail), *lawat* (visiter, examiner). Le préfixe *peN* forme des noms d'agents à partir d'un nom d'action ou d'un verbe : *pekerja* (un travailleur, un ouvrier), *pelawat* (un visiteur).

[52] *Kawan* (un ami ; un troupeau (ou groupe d'animaux)), *keliling* (le pourtour, les environs). Le patron *meN_i* forme des $\mathbf{V}_0$ dans la nomenclature de Mel'čuk *et al.*, *Dictionnaire explicatif et combinatoire du français contemporain. Recherche lexico-sémantiques I*, 1984, p. 7, c'est-à-dire des dérivés syntaxiques verbaux des noms de départ ayant le même sens : *mengawani* (accompagner), *mengelilingi* (entourer).

[53] *Isteri* (femme, épouse), *ilmu* (savoir, connaissances). L'une des fonctions du préfixe *ber* est de former des dérivés adjectivaux indiquant une relation du type « avoir, être muni de » : *beristeri* (marié [pour un homme]), *berilmu* (érudit, docte).

[54] *Keras* (raide, dur), *kena* (toucher, être en contact, être attaché) indiquent l'état. *Mengeraskan* (raidir, durcir), *mengenakan* (mettre [un vêtement], attacher, infliger) indiquent le procès.

[55] *Biorąc* (portant), *piorąc* (lavant) sont des participes présents. *Bierzesz* (tu portes) et *pierzesz* (tu laves) sont les deuxièmes personnes du singulier du présent.

[56] *Ubezpieczony* (assuré), *obrażony* (vexé, blessé) sont des adjectifs au masculin singulier. *Ubezpieczeni* (assurés), *obrażeni* (vexés, blessés) sont leurs correspondants au masculin animé pluriel.

[57] *Śpiewać* (chanter), *łechtać* (chatouiller) sont des verbes à l'infinitif. Le suffixe *aczka* construit le nom d'agent féminin correspondant : *śpiewaczka* (une chanteuse, une cantatrice).

[58] *Wyszedłem* (je sortis, je suis sorti (m.)), *poszedłem* (j'allai, je suis allé (m.)) sont des verbes perfectifs à la première personne du masculin singulier. *Wyszłaś* (tu sortis (f.), tu es sortie), *poszłaś* (tu allas (f.), tu es allée) sont les formes des mêmes verbes à la deuxième personne du féminin singulier.

[59] *Rozproszyć* (réduire en poudre), et *rozmnożyć się* (se multiplier, se reproduire) sont des verbes perfectifs. Ils ont pour correspondants imperfectifs *rozpraszać* et *rozmnażać się*.

$ab : aabb \doteq aaabbb : aaaabbbb$
$ab : aabb \doteq aaaaaaabbbbbbb : aaaaaaaabbbbbbbb$
$abc : aabbcc \doteq aaabbbccc : aaaabbbbcccc$
$abc : aabbcc \doteq aaaaaaabbbbbbbccccccc : aaaaaaaabbbbbbbbccccccccc$
$aab : aaaabb \doteq aaaaaabbb : aaaaaaaabbbb$
$aba : aabbaa \doteq aaabbbaaa : aaaabbbbaaaa$
$aab : aaaabb \doteq aaaaaaaaaaaaaabbbbbb : aaaaaaaaaaaaaaaabbbbbbbb$
$aba : aabbaa \doteq aaaaaaabbbbbbbaaaaaa : aaaaaaaabbbbbbbbaaaaaaaa$
$baa : bbaaaa \doteq bbbbbbbaaaaaaaaaaaaa : bbbbbbbbaaaaaaaaaaaaaaaa$
$ab : abab \doteq abababababab : abababababababab$

Dans tous les cas, on observera que **tous les symboles** de la première chaîne se retrouvent **dans le même ordre** dans les deuxième et troisième chaînes, même si parfois il peut y avoir ambiguïté sur la position dans ces chaînes. Par exemple, le *j* de *joindre* se retrouve-t-il en première position dans *je joins* ou en seconde *je joins* dans l'analogie *joindre* : *je joins* ≐ *rejoindre* : *je rejoins* ?

Notre conjecture est donc confortée par l'inspection de ces exemples : pour qu'une analogie soit vraie, tout symbole de A doit apparaître soit dans B, soit dans C, soit dans les deux à la fois. On peut dire aussi, que dans une analogie entre chaînes de symboles, la première chaîne est couverte, au sens des chaînes, c'est-à-dire au sens du nombre d'occurrences et de l'ordre d'apparition des symboles, par les seconde et troisième chaînes de symboles. Cela est en accord avec ce que nous avons déjà dit dans nos commentaires du postulat de distribution (p. 122). La couverture doit être complète, et peut être redondante. Dans les exemples précédents, on constate en effet qu'un même symbole de A peut provenir à la fois de B et de C.

Versant de la contiguïté

Nous nous tournons maintenant vers notre seconde observation. Elle part elle aussi d'un constat négatif. Observons tout d'abord qu'il semble bien que l'on ait les inégalités suivantes qui sont autant d'analogies non admissibles :

$ab : aabb \neq aaaaabbbbb : aaabaabbbabb$
$ab : aabb \neq aaaaabbbbb : aaaaabababbbb$
$ab : aabb \neq aaaaabbbbb : baaaaabbbbba$

Pour ce qui est de la première inégalité, elle semble provenir du fait que la façon dont *ab* est inclus dans *aabb* n'est pas la même que celle dont *aaaaabbbbb* est inclus dans *aaabaabbbabb* au sens de la répartition. Nous ne parlons plus directement de l'ordre dans lesquelles les symboles d'une chaîne apparaissent dans l'autre, mais bien de la forme que prend cette inclusion. Nous pouvons marquer[60] les façons d'inclure *ab* dans *aabb*. Il en existe quatre : *aabb*, *aabb*, *aabb*, et *aabb*. De même pour la seule façon d'inclure *aaaaabbbbb* dans *aaabaabbbabb* : *aaabaabbbabb*. Si on considère un symbole qui apparaît comme

[60] Toujours par la couleur rouge.

l'objet d'une alternance ou d'un échange, on constate que ses contextes gauches et droits forment des analogies. Cela doit pouvoir se généraliser à n'importe quel symbole des chaînes. L'hypothèse que nous sommes tenté d'émettre est que les préfixes des quatre chaînes intervenant dans une analogie respectent une propriété de répartition exprimée par une analogie entre multi-ensembles (p. 156).

Pour le moment, les considérations précédentes sont de l'ordre de l'intuition. Contrairement à ce que nous avons été capables de réaliser sur le versant de la similarité, pour ce qui est du versant de la contiguïté, nous sommes encore en pleine recherche d'une expression convaincante de notre intuition et surtout d'une formalisation satisfaisante.

$ab : aabb \neq aaaaabbbbb : aaaaababbbbb$
$aabb : ab \neq aaaaababbbbb : aaaaabbbbb$
$ab : aaaaabbbbb \neq aabb : aaaaababbbbb$
$aaaaabbbbb : ab \neq aaaaababbbbb : aabb$

4.4.2 Similitude

La première observation nous permet donc de formuler la conjecture énoncée plus haut (p. 143), sous la forme du lemme suivant. Il s'agit en fait d'une expression faible de l'hypothèse de distribution. Nous notons par $\overline{A}$ l'ensemble des symboles d'une chaîne A que l'on appelle *ensemble caractéristique* de A.

LEMME **8 (Inclusion des ensembles de symboles)** *Soit $\mathcal{V}$ un alphabet,*

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad A : B \doteq C : D \quad \Rightarrow \quad \overline{A} \subset \overline{B} \cup \overline{C}$$

La spécialisation de l'égalité ensembliste des moyens et des extrêmes aux chaînes de symboles s'écrit sous la forme du théorème suivant.

THÉORÈME **10** *Soit $\mathcal{V}$ un alphabet,*

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad A : B \doteq C : D \quad \Rightarrow \quad \overline{A} \cup \overline{D} = \overline{B} \cup \overline{C}$$

Cependant, l'expression précédente du postulat de distribution, et donc le lemme d'inclusion des symboles, n'épuisent pas les propriétés des chaînes des symboles dans lesquelles une notion d'ordre existe aussi. On peut en effet encore ajouter que le fait pour les symboles d'une chaîne de « se retrouver » dans une autre chaîne est donné par la similitude entre ces deux chaînes. Les « distributions » peuvent donc être vues comme la similitude, et le nombre de « propriétés » d'une chaîne comme sa longueur.

Pour l'analogie $A : B \doteq C : D$, la forme forte que prend l'hypothèse de distribution est donc que la longueur de A est inférieure ou égale à la somme de ses similitudes avec B et C :

LEMME **9 (Inclusion sur les chaînes de symboles)** *Soit $\mathcal{V}$ un alphabet, soit σ la similitude entre deux chaînes.*

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad |A| \quad \leq \quad \sigma(A,B) + \sigma(A,C)$$

Il est intéressant de remarquer que, puisque pour toute chaîne A, $\sigma(A,A) = |A|$, l'hypothèse de distribution sur les chaînes de symboles prend la forme particulièrement élégante suivante :

$$\sigma(A,A) \quad \leq \quad \sigma(A,B) + \sigma(A,C)$$

Cet énoncé a une conséquence très importante pour les chaînes réduites à un seul symbole :

LEMME **10 (Rapports inverses)** *Soit $\mathcal{V}$ un alphabet,*

$$\forall (a,b) \in \mathcal{V}^2, \qquad a : b \doteq b : a \quad \Leftrightarrow \; a = b$$

DÉMONSTRATION : Si l'on a $a : b \doteq b : a$, l'hypothèse d'inclusion implique : $|a| \leq 2 \times \sigma(a,b)$. Si a est différent de b, on a $1 \leq 2 \times 0 = 0$, ce qui est impossible. Nécessairement donc, $a = b$ avec : $1 \leq 2 \times 1 = 2$. CQFD

Bien que nous ne donnions pas de signification à la notation $A : B$, le lemme précédent peut trivialement s'interpréter par le fait que les rapports inverses de deux symboles différents sont nécessairement différents $a \neq b \;\Leftrightarrow\; a : b \neq b : a$.

Si la longueur de A est strictement inférieure à la somme des similitudes, alors, certains symboles de A sont communs aux trois chaînes A, B et C dans le même ordre. Comme ces symboles devront forcément être recopiés dans la solution D, ils apparaissent aussi dans D dans le même ordre. Appelons $\gamma(A,B,C,D)$ le nombre de tels symboles. On peut alors écrire
$\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B \doteq C : D \quad \Rightarrow \quad |A| = \sigma(A,B) + \sigma(A,C) - \gamma(A,B,C,D)$$

Par application du théorème des formes équivalentes, on a aussi :

$$\begin{aligned}
|A| &= \sigma(A,C) + \sigma(A,B) - \gamma(A,C,B,D) \\
|B| &= \sigma(B,A) + \sigma(B,D) - \gamma(B,A,D,C) \\
|B| &= \sigma(B,D) + \sigma(B,A) - \gamma(B,D,A,C) \\
|C| &= \sigma(C,A) + \sigma(C,D) - \gamma(C,A,B,D) \\
|C| &= \sigma(C,D) + \sigma(C,A) - \gamma(C,D,A,B) \\
|D| &= \sigma(D,B) + \sigma(D,C) - \gamma(D,B,C,A) \\
|D| &= \sigma(D,C) + \sigma(D,B) - \gamma(D,C,B,A)
\end{aligned}$$

De façon complémentaire au fait que la longueur de A est inférieure à la somme de ses similitudes avec B et C, on peut énoncer le résultat suivant : si on retire de B et de C ce qu'ils ont de commun avec A, doit rester un peu moins que ce qui permet de construire D. On a donc :

$$|D| \quad \geq \quad |B| - \sigma(A,B) + |C| - \sigma(A,C)$$

c'est-à-dire :

$$|B| + |C| - |D| \quad \leq \quad \sigma(A,B) + \sigma(A,C)$$

La partie qui manque pour construire D entièrement n'est autre que ce qui est commun à la fois aux quatres termes de l'analogie. On posera donc que :

$$\begin{aligned}
\gamma(A,B,C,D) &= \gamma(A,C,B,D) \\
&= \gamma(B,A,D,C) \\
&= \gamma(B,D,A,B) \\
&= \gamma(C,A,D,B) \\
&= \gamma(C,D,A,B) \\
&= \gamma(D,B,C,A) \\
&= \gamma(D,C,B,A)
\end{aligned}$$

Sur les exemples précédents, le calcul de $\gamma(A,B,C,D)$ donne les résultats suivants. Nous marquons les symboles en commun dans le même ordre dans les quatre chaînes. Nous ne donnons qu'une seule des possibilités, mais il peut en exister plusieurs. Par exemple, pour la troisième analogie, on aurait pu choisir le *e* au lieu du *l* comme symbole commun et écrire *fliehen* : *er floh* ≐ *schließen* : *er schloß* ou *fliehen* : *er floh* ≐ *schließen* : *er schloß* ou encore *fliehen* : *er floh* ≐ *schließen* : *er schloß* ou enfin *fliehen* : *er floh* ≐ *schließen* : *er schloß*.

setzen : *setzte* ≐ *lachen* : *lachte*	$\gamma(A,B,C,D) = 1$
lang : *längste* ≐ *scharf* : *schärfste*	$\gamma(A,B,C,D) = 0$
fliehen : *er floh* ≐ *schließen* : *er schloß*	$\gamma(A,B,C,D) = 1$
sprechen : *ihr aussprächet* ≐ *nehmen* : *ihr ausnähmet*	$\gamma(A,B,C,D) = 2$
kennen : *gekannt* ≐ *brennen* : *gebrannt*	$\gamma(A,B,C,D) = 3$
aslama : *arsala* ≐ *muslimun* : *mursilun*	$\gamma(A,B,C,D) = 2$
kataba : *kātib* ≐ *sakana* : *sākin*	$\gamma(A,B,C,D) = 2$
huzila : *huzāl* ≐ *ṣudi'a* : *ṣudā'*	$\gamma(A,B,C,D) = 1$
kalb : *kulaib* ≐ *masjid* : *musaijid*	$\gamma(A,B,C,D) = 1$
yaṣilu : *yaṣala* ≐ *yasimu* : *yasama*	$\gamma(A,B,C,D) = 2$
科学 : 科学家 ≐ 政治 : 政治家	$\gamma(A,B,C,D) = 0$
我 : 我們 ≐ 他 : 他們	$\gamma(A,B,C,D) = 0$
今年 : 今天 ≐ 明年 : 明天	$\gamma(A,B,C,D) = 0$
我是中国人 : 我是学生 ≐ 他不是中国人 : 他不是学生	$\gamma(A,B,C,D) = 1$
他們是很好的朋友 : 他們不是很好的朋友 ≐ 我去法国 : 我不去法国	$\gamma(A,B,C,D) = 0$
inné : *nées* ≐ *indu* : *dues*	$\gamma(A,B,C,D) = 0$
réaction : *réactionnaire* ≐ *répression* : *répressionnaire*	$\gamma(A,B,C,D) = 5$
aimer : *ils aimaient* ≐ *marcher* : *ils marchaient*	$\gamma(A,B,C,D) = 2$
joindre : *je joins* ≐ *oindre* : *je oins*	$\gamma(A,B,C,D) = 3$
logique : *logiciel* ≐ *ludique* : *ludiciel*	$\gamma(A,B,C,D) = 2$
食べます : 食べる ≐ 決めます : 決める	$\gamma(A,B,C,D) = 0$
痛い : 痛む ≐ 親しい : 親しむ	$\gamma(A,B,C,D) = 0$
あれ : これ ≐ あっち : こっち	$\gamma(A,B,C,D) = 0$
乗る : 乗せる ≐ 寄る : 寄せる	$\gamma(A,B,C,D) = 1$

自由 : 不自由な ≐ 用意 : 不用意な $\gamma(A,B,C,D) = 0$

oratorem : *orator* ≐ *honorem* : *honor* $\gamma(A,B,C,D) = 3$

facio : *conficio* ≐ *capio* : *concipio* $\gamma(A,B,C,D) = 3$

amo : *amas* ≐ *oro* : *oras* $\gamma(A,B,C,D) = 0$

amo : *amat* ≐ *oro* : *orat* $\gamma(A,B,C,D) = 0$

amo : *amamus* ≐ *oro* : *oramus* $\gamma(A,B,C,D) = 0$

tinggal : *ketinggalan* ≐ *duduk* : *kedudukan* $\gamma(A,B,C,D) = 0$

pekerja : *kerja* ≐ *pelawat* : *lawat* $\gamma(A,B,C,D) = 1$

kawan : *mengawani* ≐ *keliling* : *mengelilingi* $\gamma(A,B,C,D) = 1$

isteri : *beristeri* ≐ *ilmu* : *berilmu* $\gamma(A,B,C,D) = 1$

keras : *mengeraskan* ≐ *kena* : *mengenakan* $\gamma(A,B,C,D) = 2$

biorąc : *bierzesz* ≐ *piorąc* : *pierzesz* $\gamma(A,B,C,D) = 2$

ubezpieczony : *ubezpieczeni* ≐ *obrażony* : *obrażeni* $\gamma(A,B,C,D) = 2$

śpiewać : *śpiewaczka* ≐ *łechtać* : *łechtaczka* $\gamma(A,B,C,D) = 2$

wyszedłem : *wyszłaś* ≐ *poszedłem* : *poszłaś* $\gamma(A,B,C,D) = 3$

rozproszyć : *rozpraszać* ≐ *rozmnożyć się* : *rozmnażać się* $\gamma(A,B,C,D) = 1$

aa : *ab* ≐ *ba* : *bb* $\gamma(A,B,C,D) = 0$

a : *aa* ≐ *aaa* : *aaaa* $\gamma(A,B,C,D) = 1$

b : *ab* ≐ *aab* : *aaab* $\gamma(A,B,C,D) = 1$

ab : *aabb* ≐ *aaabbb* : *aaaabbbb* $\gamma(A,B,C,D) = 2$

abc : *aabbcc* ≐ *aaabbbccc* : *aaaabbbbcccc* $\gamma(A,B,C,D) = 3$

aab : *aaaabb* ≐ *aaaaaabbb* : *aaaaaaaabbbb* $\gamma(A,B,C,D) = 3$

aba : *aabbaa* ≐ *aaabbbaaa* : *aaaabbbbaaaa* $\gamma(A,B,C,D) = 3$

baa : *bbaaaa* ≐ *bbbaaaaaa* : *bbbbaaaaaaaa* $\gamma(A,B,C,D) = 3$

a : *aa* ≐ *aaaaaaa* : *aaaaaaaa* $\gamma(A,B,C,D) = 1$

ab : *aabb* ≐ *aaaaaaabbbbbb* : *aaaaaaaabbbbbbbb* $\gamma(A,B,C,D) = 2$

abc : *aabbcc* ≐ *aaaaaaabbbbbbccccccc* : *aaaaaaaabbbbbbbccccccc* $\gamma(A,B,C,D) = 3$

aab : *aaaabb* ≐ *aaaaaaaaaaaaaabbbbbb* : *aaaaaaaaaaaaaaaabbbbbbb* $\gamma(A,B,C,D) = 3$

aba : *aabbaa* ≐ *aaaaaaabbbbbbaaaaaaa* : *aaaaaaaabbbbbbbaaaaaaaa* $\gamma(A,B,C,D) = 3$

baa : *bbaaaa* ≐ *bbbbbbbaaaaaaaaaaaaa* : *bbbbbbbbaaaaaaaaaaaaaaa* $\gamma(A,B,C,D) = 3$

ab : *abab* ≐ *abababaabbabab* : *abababababababab* $\gamma(A,B,C,D) = 2$

Grâce à cela, et par symétrie de la similitude, la liste d'égalité sur les longueurs se réduit à seulement quatre égalités, avec $|A|$, $|B|$, $|C|$ et $|D|$ comme premier membre.

$$\begin{cases} |A| = \sigma(A,B) + \sigma(A,C) - \gamma(A,B,C,D) \\ |B| = \sigma(B,A) + \sigma(B,D) - \gamma(A,B,C,D) \\ |C| = \sigma(C,A) + \sigma(C,D) - \gamma(A,B,C,D) \\ |D| = \sigma(D,B) + \sigma(D,C) - \gamma(A,B,C,D) \end{cases}$$

D'où le théorème suivant.

THÉORÈME **11 (Contrainte de similitude)** *Soit $\mathcal{V}$ un alphabet,* $\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B \doteq C : D \quad \Rightarrow \quad \begin{cases} \sigma(A,B) + \sigma(A,C) - |A| &= \sigma(A,B) + \sigma(B,D) - |B| \\ &= \sigma(A,C) + \sigma(C,D) - |C| \\ &= \sigma(B,D) + \sigma(C,D) - |D| \end{cases}$$

Du premier théorème précédent, on peut tirer deux propriétés. La première est une forme équivalente du théorème.

LEMME **11** *Soit $\mathcal{V}$ un alphabet,* $\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B \doteq C : D \quad \Rightarrow \quad \begin{cases} |A| - \sigma(A,B) = |C| - \sigma(C,D) & (1) \\ |B| - \sigma(B,D) = |A| - \sigma(A,C) & (2) \\ |C| - \sigma(C,A) = |D| - \sigma(D,B) & (3) \\ |D| - \sigma(D,C) = |B| - \sigma(B,A) & (4) \end{cases}$$

DÉMONSTRATION : Selon la contrainte d'analogie :

$$|A| + \gamma(A,B,C,D) = \sigma(A,B) + \sigma(A,C) \tag{4.1}$$

De même, par permutation des moyens,

$$|C| + \gamma(A,B,C,D) = \sigma(C,A) + \sigma(C,D)$$

Parce que $\sigma(A,B) = \sigma(B,A)$

$$|C| + \gamma(A,B,C,D) = \sigma(A,C) + \sigma(C,D) \tag{4.2}$$

Par soustraction des égalités (4.1) et (4.2) :

$$|A| - \sigma(A,B) = |C| - \sigma(C,D)$$

On permute A, B, C et D grâce au théorème des formes équivalentes de l'analogie pour obtenir les trois autres égalités.

Réciproquement, la première égalité du théorème de contrainte d'analogie est obtenue en faisant ligne (1) - ligne (2) + ligne (3). Et de même, *mutatis mutandis*, pour les deux autres égalités. CQFD

Égalité sur les distances

Du lemme donnant la relation entre distance canonique et similitude et du lemme 11 (p. 148), on tire facilement le théorème suivant :

THÉORÈME **12 (Égalité des distances)** *Soit $\mathcal{V}$ un alphabet et soit δ la distance d'édition canonique.*

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad A : B \doteq C : D \quad \Rightarrow \quad \begin{cases} \delta(A,B) = \delta(C,D) \\ \delta(A,C) = \delta(B,D) \end{cases}$$

En termes géométriques, la distance des paires de côtés adjacents opposées est égale.
DÉMONSTRATION : Par soustraction des lignes (1) et (4) et des lignes (2) et (3) du lemme 11 (p. 148), on obtient le système :

$$\begin{cases} |A| + |B| - 2 \times \sigma(A,B) = |C| + |D| - 2 \times \sigma(C,D) \\ |A| + |C| - 2 \times \sigma(A,C) = |B| + |D| - 2 \times \sigma(B,D) \end{cases}$$

La proposition 1 donnant la relation entre distance canonique et similitude, permet de réécrire :

$$\begin{cases} \delta(A,B) = \delta(C,D) \\ \delta(A,C) = \delta(B,D) \end{cases}$$

CQFD

Ce résultat se représente graphiquement par le parallélogramme de la figure 4.4.

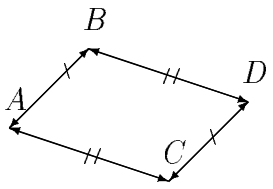


Figure 4.4: L'analogie comme un parallélogramme

Égalité sur les longueurs

Un second théorème est impliqué par le théorème de contrainte de similitude.

THÉORÈME **13 (Égalité des sommes des longueurs)** *Soit $\mathcal{V}$ un alphabet,*

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad A : B \doteq C : D \quad \Rightarrow \quad |A| + |D| = |B| + |C|$$

En termes géométriques, on peut énoncer que dans une analogie entre chaînes de symboles, les sommes des longueurs des côtés opposés sont égales.
DÉMONSTRATION : Par addition des lignes (1) et (4) (ou des lignes (2) et (3)) du lemme 11 (p. 148), et par commutativité de la similitude. CQFD

Nous avions annoncé ce théorème lors de la section sur l'analogie entre ensembles (voir p. 131). Dans le cas où les variables désignaient des ensembles, la notation $|A|$ désignait le cardinal de l'ensemble. Ici, les variables désignent des chaînes de symboles, et la notation désigne la longueur de la chaîne.

Inégalité sur les diagonales

Le théorème suivant caractérise les distances entre extrêmes et moyens.

THÉORÈME **14 (Inégalité sur les diagonales)** *Soit* $\mathcal{V}$ *un alphabet,* $\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B \doteq C : D \quad \Rightarrow \quad |\delta(A,D) - \delta(B,C)| \leq 2 \times \min(\delta(A,B), \delta(A,C))$$

DÉMONSTRATION : En utilisant deux fois l'inégalité triangulaire sur les distances pour réécrire $\delta(A,D)$, on a : $\delta(A,D) \leq \delta(A,B) + \delta(B,C) + \delta(C,D)$. Comme $\delta(A,B) = \delta(C,D)$ d'après le théorème 12, on obtient l'inégalité :

$$\delta(A,D) - \delta(B,C) \leq 2 \times \delta(A,B)$$

Pour la différence opposée, on obtient l'inégalité similaire suivante par la même méthode à partir de $\delta(B,C)$:

$$\delta(B,C) - \delta(A,D) \leq 2 \times \delta(A,B)$$

L'échange des moyens donne les deux inégalités :

$$\delta(A,D) - \delta(B,C) \leq 2 \times \delta(A,C)$$
$$\delta(B,C) - \delta(A,D) \leq 2 \times \delta(A,C)$$

Les quatre inégalités précédentes impliquent l'inégalité du théorème :

$$|\delta(A,D) - \delta(B,C)| \leq 2 \times \min(\delta(A,B), \delta(A,C))$$

CQFD

Partie commune et périmètres

En reprenant les quatre égalités de la page 147 faisant intervenir $\gamma(A,B,C,D,)$ on peut établir le théorème suivant, dont l'expression géométrique est : dans une analogie entre chaînes de symboles, la longueur de la partie commune est égale au quart de la différence du périmètre et de toutes les distances entre côtés adjacents.

THÉORÈME **15 (Longueur de la partie commune)** *Soit* $\mathcal{V}$ *un alphabet,* $\forall (A,B,C,D) \in (\mathcal{V}^*)^4$, $\quad A : B \doteq C : D \quad \Rightarrow$

$$\gamma(A,B,C,D) \;=\; \tfrac{1}{4}\Big[\;\; |A| + |B| + |C| + |D| \\ - \delta(A,B) - \delta(A,C) - \delta(B,D) - \delta(C,D) \;\;\Big]$$

DÉMONSTRATION : Par addition des quatre égalités, on obtient :

$$\begin{aligned} 4 \times \gamma(A,B,C,D) &= 2 \times (\sigma(A,B) + \sigma(A,C) + \sigma(B,D) + \sigma(C,D)) \\ &\quad - (|A| + |B| + |C| + |D|) \end{aligned}$$

La relation entre distance canonique, similitude et longueur permet d'écrire :

$$\begin{aligned} 4 \times \gamma(A,B,C,D) &= 2 \times (\sigma(A,B) + \sigma(A,C) + \sigma(B,D) + \sigma(C,D)) \\ &\quad - 2 \times (|A| + |B| + |C| + |D|) \\ &\quad + (|A| + |B| + |C| + |D|) \\ &= - |A| - |B| + 2 \times \sigma(A,B) \\ &\quad - |A| - |C| + 2 \times \sigma(A,C) \\ &\quad - |B| - |D| + 2 \times \sigma(B,D) \\ &\quad - |C| - |D| + 2 \times \sigma(C,D) \\ &\quad + (|A| + |B| + |C| + |D|) \end{aligned}$$

C'est-à-dire l'égalité du théorème. CQFD

Ce théorème permet de dériver facilement :

LEMME **12** *Soit $\mathcal{V}$ un alphabet,*
$\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B \doteq C : D \quad \Rightarrow \quad \begin{aligned} \gamma(A,B,C,D) &= \tfrac{1}{2} \times (|B| + |C| - \delta(A,B) - \delta(A,C)) \\ &= \tfrac{1}{2} \times (|A| + |D| - \delta(B,A) - \delta(B,D)) \\ &= \tfrac{1}{2} \times (|D| + |A| - \delta(C,D) - \delta(C,A)) \\ &= \tfrac{1}{2} \times (|C| + |B| - \delta(D,C) - \delta(D,B)) \end{aligned}$$

DÉMONSTRATION : Par les théorèmes 12 et 13, on a les égalités : $\delta(A,B) = \delta(C,D)$, $\delta(B,D) = \delta(A,C)$ et $|A| + |D| = |B| + |C|$. En reportant ces égalités dans le théorème 15 (p. 150), on obtient les égalités du lemme. CQFD

Système analogique

Nous énonçons maintenant notre résultat final sur les analogies entre chaînes de symboles.

THÉORÈME **16** *Soit $\mathcal{V}$ un alphabet, soit δ la distance d'édition canonique.*
$\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B \doteq C : D \quad \Rightarrow \quad \left\{ \begin{aligned} \delta(A,B) &= \delta(C,D) \\ \delta(A,C) &= \delta(B,D) \\ |A| + |D| &= |B| + |C| \\ \gamma(A,B,C,D) &= \tfrac{1}{4} \times (|A| + |B| + |C| + |D| \\ &\qquad - \delta(A,B) - \delta(A,C) \\ &\qquad - \delta(B,D) - \delta(C,D)) \end{aligned} \right.$$

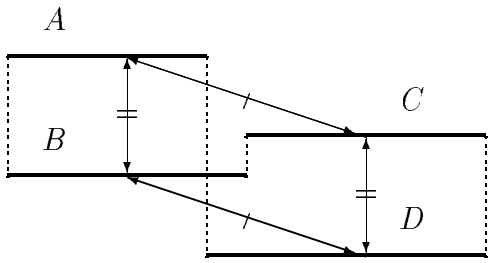


Figure 4.5: Visualisation des relations de distances dans l'analogie entre chaînes de symboles

Nous appellerons S 16 le système d'égalités de ce théorème. Les trois premières égalités du système précédent peuvent se représenter par la figure 4.5. La quatrième égalité difficilement visualisable dit que la différence des périmètres des chaînes à leur distances est égale à la longueur de la partie commune aux chaînes.

La forme précédente du système d'égalités est relativement parlante car visualisable. En remplaçant les distances par des similitudes dans le théorème 16 (p. 151) ci-dessus, on obtient une seconde caractérisation équivalente de l'analogie entre chaînes de symboles.

THÉORÈME **17** *Soit $\mathcal{V}$ un alphabet, soit σ la similitude entre chaînes de symboles.*
$\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B \doteq C : D \quad \Rightarrow \quad \left\{ \begin{array}{rcl} \sigma(B,D) & = & -|A| + |B| + \sigma(A,C) \\ \sigma(C,D) & = & -|A| + |C| + \sigma(A,B) \\ |D| & = & -|A| + |B| + |C| \\ \gamma(A,B,C,D) & = & -|A| + \sigma(A,B) + \sigma(A,C) \end{array} \right.$$

Cette caractérisation, où nous avons fait passer D du côté gauche du signe égal et où le côté droit ne contient que A, B et C, est plus intéressante que la première du point de vue calculatoire, car elle elle constitue un pas vers la conception et la réalisation d'un algorithme de résolution d'équations analogiques.

Conservation par transformations de chaînes

La caractérisation de l'analogie donnée par le théorème 16 (p. 151) permet de tirer la conjecture suivante :

Conjecture 1 *Soit $\mathcal{V}$ un alphabet, soit* f *un application de $\mathcal{S} \subset V^*$ dans V^*, telle que*

$$\forall (A,B) \in \mathcal{S}^2,\ \delta(\mathrm{f}(A),\mathrm{f}(B)) = \lambda.\delta(A,B) \quad et \quad \mathrm{f}(\varepsilon) = \varepsilon$$

alors

$$\forall (A,B,C,D) \in \mathcal{S}^4,\ \ S\ 16(A,B,C,D) \Rightarrow S\ 16(\mathrm{f}(A),\mathrm{f}(B),\mathrm{f}(C),\mathrm{f}(D))$$

DÉMONSTRATION : Par définition de la distance canonique, on a : $|A| = \delta(\varepsilon, A)$, Comme ε a pour image lui-même par f, on a : $\forall A \in \mathcal{S},\ |\mathrm{f}(A)| = \lambda.|A|$.

Toutes les égalités du système du thèorème 16 sont donc conservées au λ près.

$$\left\{\begin{array}{ccccccc}
\delta(\mathrm{f}(A),\mathrm{f}(B)) & = & \lambda.\delta(A,B) & = & \lambda.\delta(C,D) & = & \delta(\mathrm{f}(C),\mathrm{f}(D)) \\
\delta(\mathrm{f}(A),\mathrm{f}(C)) & = & \lambda.\delta(A,C) & = & \lambda.\delta(B,D) & = & \delta(\mathrm{f}(B),\mathrm{f}(D)) \\
|\mathrm{f}(A)| + |\mathrm{f}(D)| & = & \lambda.(|A| + |D|) & = & \lambda.(|B| + |C|) & = & |\mathrm{f}(B)| + |\mathrm{f}(C)|
\end{array}\right.$$

$$\left\{\begin{array}{rcl}
\gamma(\mathrm{f}(A),\mathrm{f}(B),\mathrm{f}(C),\mathrm{f}(D)) & = & \frac{1}{4} \times (|\mathrm{f}(A)| + |\mathrm{f}(B)| + |\mathrm{f}(C)| + |\mathrm{f}(D)| \\
 & & \quad - \delta(\mathrm{f}(A),\mathrm{f}(B)) - \delta(\mathrm{f}(A),\mathrm{f}(C)) \\
 & & \quad - \delta(\mathrm{f}(B),\mathrm{f}(D)) - \delta(\mathrm{f}(C),\mathrm{f}(D))) \\
 & = & \frac{1}{4} \times (\lambda.|A| + \lambda.|B| + \lambda.|C| + \lambda.|D| \\
 & & \quad - \lambda.\delta(A,B) - \lambda.\delta(A,C) \\
 & & \quad - \lambda.\delta(B,D) - \lambda.\delta(C,D)) \\
 & = & \gamma(A,B,C,D)
\end{array}\right.$$

CQFD

Comme nous n'avons toujours pas le versant de la contiguïté, le système S 16 n'est pas équivalent à l'analogie. Cependant, nous avons bon espoir que la relation obtenue sur le versant de la contiguïté soit neutre vis-à-vis de ce système, et que donc, il possède les mêmes propriétés d'invariance par les applications remplissant les conditions de la conjecture 1. Dans le reste de cette section, nous faisons cette hypothèse. De cette façon, on obtient tous les corollaires suivants, qui restent tout de même pour le moment des conjectures. Le corollaire suivant s'obtient en contraignant la fonction f du théorème précédent.

COROLLAIRE **3 (Conservation par isométrie)** *Soit $\mathcal{V}$ un alphabet, l'analogie entre chaînes sur $\mathcal{V}^*$ est conservée par isométrie de $\mathcal{V}^*$ dans $\mathcal{V}^*$.*

DÉMONSTRATION : En faisant $\lambda = 1$. CQFD

Ce corollaire énonce que l'analogie serait insensible au système de notation utilisé. En effet considérons deux alphabets $\mathcal{V}$ et $\mathcal{V}'$ en bijection. On peut

définir, sur $(\mathcal{V} \cup \mathcal{V}')^*$ la symétrie qui remplace chaque symbole de $\mathcal{V}$ par son correspondant dans $\mathcal{V}'$ et réciproquement. Les analogies sur $\mathcal{V}^*$ sont valables sur leurs correspondants dans $(\mathcal{V}')^*$. Autrement dit, les translittérations (bijectives) conservent les analogies.

On peut illustrer cela avec la romaine et l'italique. L'application de la transformation de romain en italique et d'italique en romain, utilisée pour la mise en relief, est bien neutre vis-à-vis de l'analogie. On peut écrire indifféremment *croire* : *je crois* $\doteq$ *boire* : *je bois*, ou croire : je crois $\doteq$ boire : je bois. Nous retrouvons ici la caractéristique que nous annoncions dans notre introduction. Il s'agit du caractère universel de l'analogie, et de son indifférence au codage (voir un autre exemple p. 23).

Rappelons maintenant que la chaîne miroir de $a_1.a_2.\ldots.a_n$ est $a_n.\ldots.a_2.a_1$. L'opération miroir sur les chaînes, notée μ, étant une isométrie, on a donc :

$$\forall (A,B,C,D) \in \mathcal{V}^{*4}, \quad A : B \doteq C : D \quad \Leftrightarrow \quad \mu(A) : \mu(B) \doteq \mu(C) : \mu(D)$$

COROLLAIRE **4 (Conservation par miroir)** *Soit $\mathcal{V}$ un alphabet, l'analogie entre chaînes sur $\mathcal{V}^*$ est conservée par miroir.*

DÉMONSTRATION : Le miroir de chaînes étant une isométrie, le corollaire 3 premet de conclure. CQFD

Ce corollaire énonce que le sens d'écriture, de gauche à droite ou de droite à gauche, n'a pas d'importance pour l'analogie. Les premiers Grecs qui écrivaient en boustrophédon, c'est-à-dire de gauche à droite comme de droite à gauche ne détruisaient pas les analogies de leur langue par ce changement de sens. Quant aux Turcs, qui sont passés en 1928 d'une transcription de leur langue utilisant des caractères arabes écrits de droite à gauche, à une transcription par des caractères latins écrits de gauche à droite, ils n'ont pas perdu pour autant les analogies morphologiques ou syntaxiques de leur langue dans l'opération : *ve* : *relve* $\doteq$ *icnerğö* : *relicnerğö*, ou *ev* : *evler* $\doteq$ *öğrenci* : *öğrenciler* [61]. L'analogie y est insensible.

Il est intéressant de noter que, en considérant le symboles deux-points et le symbole de conformité comme faisant partie d'une chaîne, et donc en considérant l'équation de l'analogie comme une chaîne elle-même, on peut écrire :

$$\mu(\ \ A : B \doteq C : D\ \ ) = \ \ \mu(D) : \mu(C) \doteq \mu(B) : \mu(A)$$

Soit $A = a_1 a_2 \ldots a_m$ une chaîne, soit u une chaîne n'ayant pas de symboles en commun avec A, on note $A \downarrow u$ la chaîne : $a_1 u a_2 u \ldots u a_m$. On pose que $A \downarrow u = A$ si A n'a qu'un élément ou est égale à la chaîne vide ε. Nous allons montrer le corollaire suivant.

COROLLAIRE **5 (Conservation par incrustation)** *Soit $\mathcal{V}$ un alphabet, l'analogie entre chaînes sur $\mathcal{V}^*$ est conservée par incrustation de toute chaîne de symboles qui ne sont pas dans $\mathcal{V}$.*

[61] *Dev* (la maison), *öğrenci* (l'élève) sont au singulier. Les formes en *-ler* sont les mêmes mots au pluriel.

Ce corollaire a pour synonyme le théorème suivant.

THÉORÈME **18** *Soit $\mathcal{V}$ un alphabet,* $\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \forall u \in \mathcal{V}^* \;/\; \overline{u} \cap (\overline{A} \cup \overline{B} \cup \overline{C} \cup \overline{D}) = \phi$,

$$A : B \doteq C : D \quad \Leftrightarrow \quad A \downarrow u : B \downarrow u \doteq C \downarrow u : D \downarrow u$$

DÉMONSTRATION : La transformation décrite ne conserve pas la distance, mais on peut démontrer le résultat suivant :

$$\overline{u} \cap (\overline{A} \cup \overline{B}) = \phi \Rightarrow \delta(A \downarrow u, B \downarrow u) = \delta(A,B) + |u| \times \Big||A| - |B|\Big|$$

Soient donc A, B, C et D quatre chaînes de $\mathcal{V}$, tel que $A : B \doteq C : D$ et soit u une chaîne n'ayant de symboles en commun avec ni A, ni B, ni C, ni D.

Supposons que $|A| \geq |B|$. La troisième égalité du théorème 16 (p. 151), énonçant que $|A| + |D| = |B| + |C|$, on a donc : $|C| \geq |D|$. On a donc le résultat : $\Big||A| - |B|\Big| = \Big||C| - |D|\Big|$. Avec ce qui précède, ceci implique que $\delta(A \downarrow u, B \downarrow u) = \delta(C \downarrow u, D \downarrow u)$. On aura aussi le même résultat si $|A| < |B|$. Et de même avec les couples A et C et B et D.

Les deux premières égalités du système du théorème 16 (p. 151) sont donc conservées.

$$\begin{cases} \delta(A,B) = \delta(C,D) \\ \delta(A,C) = \delta(B,D) \end{cases} \Rightarrow \begin{array}{l} \delta(A \downarrow u, B \downarrow u) = \delta(C \downarrow u, D \downarrow u) \\ \delta(A \downarrow u, C \downarrow u) = \delta(B \downarrow u, D \downarrow u) \end{array}$$

On a aussi $|A \downarrow u| = |A| + |u| \times (|A| - 1)$ si $A \neq \varepsilon$. Ce qui s'écrit aussi : $|A \downarrow | = |A| \times (|u| + 1) - |u|$. Avec cette égalité, il est trivial de vérifier que

$$|A| + |D| = |B| + |C| \;\Rightarrow\; |A \downarrow u| + |D \downarrow u| = |B \downarrow u| + |C \downarrow u|$$

Les cas où l'une ou plusieurs des chaînes sont vides sont aussi facilement vérifiés. CQFD

Il existerait donc une certaine neutralité des éléments intermédiaires. On pourrait toujours « plâtrer » les interstices entre les symboles intervenant dans les chaînes sans changer la validité de l'analogie. Redisons que le versant de la contiguïté n'étant pas encore achevé, ce corollaire reste un conjecture.

Mentionnons pour illustrer qu'il était de tradition dans les livres imprimés en Pologne jusque dans les années soixante de mettre en relief un mot ou une expression en insérant une espace entre chacune de ses lettres. Cette convention est maintenant peu à peu remplacée par l'usage de l'italique. Appliquée a l'exemple *śpiewać* : *śpiewaczka* ≐ *działać* : *działaczka*, elle donnait : *ś p i e w a ć* : *ś p i e w a c z k a* ≐ *d z i a ł a ć* : *d z i a ł a c z k a*. [62]

[62] *Śpiewać* (chanter), *śpiewaczka* (une chanteuse, une cantatrice). *Działać* (agir, être actif), *działaczka* (une activiste).

4.4.3 Contiguïté

Nous présentons ici les derniers développements de notre recherche sur le versant de la contiguïté. Nous ne pouvons ici qu'aborder la direction dans laquelle nous nous sommes engagé. Il nous manque encore une justification simple, ainsi qu'une mise en forme qui soit élégante.

Tout d'abord, et à titre d'exemple, considérons l'équation analogique $ab : aabb \doteq a^n b^n : x$. Si l'on énumère l'ensemble des chaînes de symboles sur $\{a, b\}$ vérifiant le système du théorème 16 (p. 151), on obtient la chaîne $a^{n+1}b^{n+1}$, mais aussi toutes les chaînes de la forme $a^i b a^{n-i} b^j a b^{n-j}$ avec i et j deux entiers compris entre 0 et n. Cela fait $(n+1)^2 + 1$ solutions. Or le sentiment veut qu'il y ait une unique solution à cette équation analogique, $a^{n+1}b^{n+1}$.

L'intuition sur la contiguïté que nous avons est que, à partir des début et fin de chaînes, les symboles apparaissant dans D sont fonctions des possibilités au sens des multi-ensembles. Ainsi, pour une équation analogique $A : B \doteq C : D$ d'inconnue D, le premier symbole de la chaîne D doit être tel que l'analogie au sens des multi-ensembles est vraie avec les débuts des trois chaînes de symboles A, B et C. La sous-chaîne ne contenant que le premier symbole a une longueur de 1. De plus, selon le théorème 13 (p. 149) d'égalité des sommes des longueurs des extrêmes et des moyens, on a la relation $|A|+|D| = |B|+|C|$. Il suffit donc d'examiner tous les multi-ensembles possibles. Par exemple, pour l'équation analogique $ab : aabb \doteq aaaabbbb : x$, ces possibilités sont listés dans le tableau 4.4.

Tableau 4.4: Sous-chaînes de $A = ab$, $B = aabb$ et $C = aaaabbbb$ à examiner pour obtenir le premier symbole de D solution de l'équation analogique

$$ab : aabb \doteq aaaabbbb : x$$

sous-chaînes					
A	B	C	$\lvert D\rvert = \lvert B\rvert + \lvert C\rvert - \lvert A\rvert$	$\overline{\overline{D}}$	
a	a	a	1	{ (a,1) }	bon
a	a	aa	2		impossible
a	a	$\vdots$	> 2		impossible
a	aa	a	2		impossible
a	$\vdots$	a	> 2		impossible
ab	a	a	-1		impossible
ab	aa	a	1	{ (a,2), (b,-1) }	impossible
ab	aab	a	2	{ (a,2), (b,0) }	impossible
ab	$\vdots$	$\vdots$	> 2		impossible
ab	a	aa	1	{ (a,2), (b,-1) }	impossible
ab	a	aaa	1	{ (a,3), (b,-1) }	impossible
ab	$\vdots$	$\vdots$	> 2		impossible

On peut construire de cette façon un tableau complet des poids des symboles pour chaque position dans l'inconnue x (voir figure 4.5). Ce tableau est indicé sur les positions possibles dans x, et sur ses symboles possibles. En divisant chaque case du tableau par la somme des éléments de la colonne à laquelle elle appartient, on obtient une matrice des probabilités d'occurrence de chaque symbole en chaque position de D. Nous donnons aussi la matrice des probabilités d'occurrence des symboles dans la chaîne D obtenue en divisant chaque case de chaque colonne par la somme des valeurs (table 4.5).

Tableau 4.5: Poids des occurrences des symboles dans la chaîne D solution de l'équation analogique

$$ab : aabb \doteq aaaabbbb : x$$

a	1	16	69	204	448	630	789	872	805	570
b	0	0	9	28	62	162	345	536	644	570

Tableau 4.6: Matrice des probabilités d'occurrence des symboles dans D

a	1	1	0,88	0,88	0,88	0,80	0,70	0,62	0,56	0,5
b	0	0	0,12	0,12	0,12	0,20	0,30	0,38	0,44	0,5

Dans cet exemple, les deux premiers symboles de la chaîne sont imposés car la probabilité d'occurrence de l'un des symboles est de 1 pour ces deux positions. La solution de l'équation analogique $ab : aabb \doteq aaaabbbb : x$ commence donc par le préfixe aa. Elle se termine aussi par le préfixe bb pour des raisons de symétrie. Rien que ces observations nous permettent d'éliminer un grand nombre de solutions superfétatoires. Nous venons de réduire l'ensemble des solutions à a^5b^5 et aux seules chaînes de la forme $a^2a^iba^{2-i}b^jab^{2-j}b^2$ avec i et j compris entre 0 et 2, On n'a donc plus que 10 solutions au lieu 26 auparavant. Notre but d'arriver à la seule solution, a^5b^5, n'est donc pas encore atteint, mais en observant le tableau 4.6, on sent bien que la seule solution admissible a^5b^5 sera sans doute obtenue en maximisant la probabilité globale de placement des symboles d'une façon qui reste à trouver. On pourrait penser que déjà, en suivant les maximums locaux, on pourrait construire la solution désirée. En effet, de cette manière, sur le même exemple, on construit tout de suite le préfixe $aaaaa$ et comme l'on connaît les symboles de D, a et b, et leur nombre, 5 chacun, on est alors forcé de compléter par des b. On obtiendrait bien ainsi la seule solution tolérable a^5b^5. Malheureusement, d'autres exemples montrent que tout n'est pas aussi simple. Par exemple, pour l'équation analogique $aslama : muslim \doteq arsala : x$, le tableau 4.7 donne les symboles de probabilité non nulle pour la première position dans la solution. On constate que le symbole m, qui correspond au premier symbole de la seule solution admissible, a une probabilité très faible par rapport à r.

Tableau 4.7: Probabilités d'occurrence des symboles en première position dans la chaîne D solution de l'équation analogique

$$aslama : muslim \doteq arsala : x$$

m	0,08
r	0,92

Calculer toutes les probabilités à l'avance aurait été économique. Il semble cependant que l'on obtienne de meilleurs résultats si l'on tient compte pour calculer les probabilités du symbole suivant à produire de l'occurrence des symboles précédents. Avec cette technique, on peut ordonner les symboles possibles pour une position donnée. On peut penser que cette manière de faire augmente la fiabilité de la probabilité du symbole suivant à produire. Nous avons effectué ce calcul sur les données précédentes pour un caractère de la chaîne solution à produire, marqué dans la quatrième chaîne. On considère donc que tous les symboles précédents sont connus. Les symboles possibles apparaissent dans la colonne de droite avec un facteur. Plus ce facteur est élevé plus le symbole a de chances d'être bien le symbole suivant. On notera que dans la plupart des cas, il y a bien égalité entre ce symbole et le symbole de la chaîne solution.

setzen : *setzte* ≐ *lachen* : *lachte* (*e*,7), (*h*,13), (*n*,2), (*t*,9), (*z*,3)
lang : *längste* ≐ *scharf* : *schärfste*
(*ä*,8), (*a*,2), (*f*,1), (*g*,1), (*n*,2), (*r*,5), (*s*,1)
fliehen : *er floh* ≐ *schließen* : *er schloß* (*c*,3), (*l*,1), (*o*,1), (*s*,5)
sprechen : *ihr aussprächet* ≐ *nehmen* : *ihr ausnähmet*
(*ä*,10), (*e*,3), (*h*,2), (*m*,1), (*n*,13), (*p*,1), (*r*,1), (*t*,1)
brennen : *gebrannt* ≐ *kennen* : *gekannt* (*a*,11), (*e*,2), (*n*,4), (*r*,15)

aslama : *arsala* ≐ *muslimun* : *mursilun* (*a*,8), (*i*,5), (*l*,5), (*m*,1), (*s*,13)
kataba : *kātib* ≐ *sakana* : *sākin* (*a*,8), (*b*,2), (*i*,7), (*n*,6), (*t*,3)
huzila : *huzāl* ≐ *ṣudi'a* : *ṣudā'* (*'*,7), (*ā*,7), (*a*,1), (*i*,6), (*l*,2), (*u*,2), (*z*,3)
kalb : *kulaib* ≐ *masjid* : *musaijid* (*a*,6), (*i*,4), (*j*,5), (*l*,3)
yaṣilu : *yaṣala* ≐ *yasimu* : *yasama* (*ṣ*,3), (*a*,11), (*i*,6), (*l*,2), (*m*,7), (*u*,1)

科学 : 科学家 ≐ 政治 : 政治家 (学,1), (家,1)
我 : 我們 ≐ 他 : 他們 (們,1)
今年 : 今天 ≐ 明年 : 明天 (年,1), (天,1)
我是中国人 : 我是学生 ≐ 他不是中国人 : 他不是学生
(学,9), (生,5), (国,4), (人,2), (是,2), (中,4)
他們是很好的朋友 : 他們不是很好的朋友 ≐ 我去法国 : 我不去法国
(好,3), (的,3), (法,8), (国,1), (很,3), (們,2), (是,17), (朋,3), (友,2)

inné : *nées* ≐ *indu* : *dues* (*é*,2), (*n*,1), (*s*,1)
réaction : *réactionnaire* ≐ *répression* : *répressionnaire*

(*E*,2), (*a*,4), (*c*,3), (*e*,8), (*i*,3), (*n*,6), (*o*,3), (*r*,16), (*t*,3)

aimer : *ils aimaient* ≐ *marcher* : *ils marchaient*

(espace,5), (*a*,1), (*m*,5), (*r*,1)

joindre : *je joins* ≐ *oindre* : *je oins*

(*d*,3), (*e*,2), (*i*,2), (*j*,13), (*n*,3), (*o*,8), (*r*,3)

logique : *logiciel* ≐ *ludique* : *ludiciel*

(*c*,6), (*e*,2), (*g*,3), (*i*,17), (*o*,2), (*q*,5), (*u*,3)

食べます : 食べる ≐ 決めます : 決める (べ,2), (ま,4), (る,3), (す,3)

痛い : 痛む ≐ 親しい : 親しむ (い,1), (む,1)

あれ : これ ≐ あっち : こっち (ち,1), (れ,1)

乗る : 乗せる ≐ 寄る : 寄せる (る,2)

自由 : 不自由な ≐ 用意 : 不用意な (な,1), (由,1)

oratorem : *orator* ≐ *honorem* : *honor*

(*a*,3), (*e*,4), (*m*,2), (*o*,19), (*r*,13), (*t*,3)

facio : *conficio* ≐ *capio* : *concipio* (*a*,1), (*c*,8), (*i*,9), (*o*,1), (*p*,7)

amo : *amas* ≐ *oro* : *oras* (*m*,2), (*o*,2), (*s*,2)

amo : *amat* ≐ *oro* : *orat* (*m*,2), (*o*,2), (*t*,2)

amo : *amamus* ≐ *oro* : *oramus* (*m*,6), (*o*,2), (*u*,2)

tinggal : *ketinggalan* ≐ *duduk* : *kedudukan*

(*a*,6), (*d*,7), (*g*,6), (*i*,2), (*l*,3), (*n*,4), (*u*,14)

pekerja : *kerja* ≐ *pelawat* : *lawat* (*a*,12), (*e*,4), (*j*,3), (*k*,3), (*r*,3), (*t*,5)

kawan : *mengawani* ≐ *keliling* : *mengelilingi*

(*a*,2), (*e*,6), (*g*,10), (*i*,1), (*l*,1), (*n*,1), (*w*,1)

isteri : *beristeri* ≐ *ilmu* : *berilmu*

(*e*,1), (*i*,12), (*l*,8), (*m*,2), (*r*,1), (*s*,1), (*t*,1)

keras : *mengeraskan* ≐ *kena* : *mengenakan*

(*a*,3), (*e*,8), (*g*,11), (*k*,1), (*n*,3), (*r*,1), (*s*,1)

biorąc : *bierzesz* ≐ *piorąc* : *pierzesz*

(*A*,3), (*c*,2), (*e*,1), (*i*,2), (*o*,3), (*r*,14), (*z*,7)

ubezpieczony : *ubezpieczeni* ≐ *obrażony* : *obrażeni*

(*Z*,12), (*a*,21), (*b*,2), (*c*,3), (*e*,9), (*i*,3), (*n*,1), (*p*,3), (*z*,6)

śpiewać : *śpiewaczka* ≐ *łechtać* : *łechtaczka* (*a*,6), (*c*,3), (*e*,3), (*h*,14), (*i*,3), (*k*,1), (*p*,2), (*t*,8), (*w*,5), (*z*,3)

wyszedłem : *wyszłaś* ≐ *poszedłem* : *poszłaś*

(*a*,2), (*d*,3), (*e*,5), (*l*,9), (*s*,3), (*y*,3), (*z*,19)

rozproszć : *rozpraszać* ≐ *rozmnożyć się* : *rozmnażać się*

(*a*,4), (*m*,7), (*n*,4), (*r*,2), (*t*,2), (*w*,3)

On peut aussi tester l'hypothèse sur les analogies formelles, précédentes.

aa : *ab* ≐ *ba* : *bb* (*a*,1), (*b*,1)

ab : *aabb* ≐ *aaabbb* : *aaaabbbb* (*a*,5), (*b*,11)

abc : *aabbcc* ≐ *aaabbbccc* : *aaaabbbbcccc* (*a*,6), (*b*,16), (*c*,2)

ab : *aabb* ≐ *aaaaaaabbbbbbb* : *aaaaaaaabbbbbbbb* (*a*,8), (*b*,4)

$aab : aaaabb \doteq aaaaaabbb : aaaaaaaabbbb$ (a,23), (b,5)

$aba : aabbaa \doteq aaabbbaaa : aaaabbbbaaaa$ (a,9), (b,25)

$aab : aaaabb \doteq aaaaaaaaaaaaaabbbbbbb : aaaaaaaaaaaaaaabbbbbbbb$ (a,21), (b,5)

$aba : aabbaa \doteq aaaaaaabbbbbbbaaaaaaa : aaaaaaaabbbbbbbbaaaaaaaa$ (a,15), (b,11)

$baa : bbaaaa \doteq bbbbbbbaaaaaaaaaaaaa : bbbbbbbbaaaaaaaaaaaaaaa$ (a,9), (b,11)

$ab : abab \doteq abababababab : ababababababab$ (a,6), (b,10)

Examinons l'un des cas où le symbole de poids le plus élevé n'est pas le symbole désiré, celui de l'équation analogique $brennen : gebrannt \doteq kennen : gekannt$ avec le préfixe gek de x connu. Les symboles suivants sont proposés avec leur poids associés : (a,11) (e,2) (n,4) et (r,15). Ici, le symbole de poids le plus élevé est r. Or cela est impossible car, pour un symbole donné, la somme de ses occurrences dans les extrêmes doit être égale à la somme de ses occurrences dans les moyens. Le symbole r apparaissant une fois dans *brennen* et une fois dans *gebrannt*, ne saurait apparaître à nouveau dans x. Il est donc à exclure. Et vient ensuite le symbole a qui est bien le symbole suivant dans la solution. Malheureusement, encore une fois tout n'est pas réglé comme le montre le cas de certaines analogies formelles comme $abc : aabbcc \doteq aaabbbccc : aaaabbbbcccc$.

Bien que nous ne puissions en dire plus pour le moment, nos recherches s'orientent dans la direction que nous venons de montrer. Mais bien sûr, nous sommes à la recherche d'une formulation qui serait du genre $\kappa(A, D) = \kappa(B, C)$, où nous pressentons que κ serait une sorte de mesure de la distance entre multi-ensembles. Nous ne pouvons en dire plus pour le moment.

Analogies disjointes

Une idée intuitive à propos des analogies entre chaînes de symboles est que deux analogies devraient pouvoir toujours être concaténées. Savoir si cette intuition est juste reste un problème ouvert.

Cependant, il existe un cas particulier où l'intuition est vérifiée. C'est celui où les deux analogies à concaténer ne partagent aucun symbole. L'ensemble de leurs symboles sont alors des ensembles disjoints, et, pour cette raison, on appellera de telles analogies *analogies disjointes.* L'intuition veut que des analogies disjointes puissent s'appliquer l'une à la suite de l'autre sans problème. Donc, si (A_1, B_1, C_1, D_1) et (A_2, B_2, C_2, D_2) n'ont aucun symbole en commun, et si $A_1 : B_1 \doteq C_1 : D_1$ et $A_2 : B_2 \doteq C_2 : D_2$ sont vraies, alors $A_1A_2 : B_1B_2 \doteq C_1C_2 : D_1D_2$ est aussi vraie.

Mais l'analogie précédente, par concaténation dans le même ordre n'est pas la seule analogie que l'on puisse constituer en utilisant les termes disjoints.

Conjecture 2 *Soient $\mathcal{V}$ un alphabet, et $\mathcal{V}_1 \subset \mathcal{V}$, $\mathcal{V}_2 \subset \mathcal{V}$, tels que $\mathcal{V}_1 \cap \mathcal{V}_2 = \phi$,*

$\forall (A_1, B_1, C_1, D_1) \in (\mathcal{V}_1{}^*)^4, \forall (A_2, B_2, C_2, D_2) \in (\mathcal{V}_2{}^*)^4,$

$$\left.\begin{array}{l} A_1 : B_1 \doteq C_1 : D_1 \\ A_2 : B_2 \doteq C_2 : D_2 \end{array}\right\} \Rightarrow \left\{\begin{array}{l} A_1A_2 : B_1B_2 \doteq C_1C_2 : D_1D_2 \\ A_1A_2 : B_1B_2 \doteq C_2C_1 : D_2D_1 \\ A_1A_2 : B_2B_1 \doteq C_2C_1 : D_1D_2 \end{array}\right.$$

DÉMONSTRATION : partielle du fait qu'il n'existe que trois telles analogies au plus. On obtient $2^4 = 16$ analogies en énumérant toutes les possibilités de permuter ou non les sous-chaînes d'indices 1 et 2 dans les termes. On numérote ces analogies en utilisant une numérotation binaire qui reflète la place de la permutation. Avec cette notation, deux analogies dont les représentations sont des compléments binaires sont équivalentes, et l'on peut donc éliminer l'une des deux de la liste. En ne retenant que celles dont le premier terme est A_1A_2, on obtient :

$$\begin{array}{ll} (0000) & A_1A_2 : B_1B_2 \doteq C_1C_2 : D_1D_2 \\ (0001) & A_1A_2 : B_1B_2 \doteq C_2C_1 : D_2D_1 \\ (0010) & A_1A_2 : B_1B_2 \doteq C_2C_1 : D_1D_2 \\ (0011) & A_1A_2 : B_1B_2 \doteq C_2C_1 : D_2D_1 \\ (0100) & A_1A_2 : B_2B_1 \doteq C_1C_2 : D_1D_2 \\ (0101) & A_1A_2 : B_2B_1 \doteq C_1C_2 : D_2D_1 \\ (0110) & A_1A_2 : B_2B_1 \doteq C_2C_1 : D_1D_2 \\ (0111) & A_1A_2 : B_2B_1 \doteq C_2C_1 : D_2D_1 \end{array}$$

Le nombre de cas différents est réduit par la propriété de symétrie de lecture des formes équivalentes de l'analogie. Du point de vue formel, (0001) $\Leftrightarrow$ (1000) $\Leftrightarrow$ (0111) et (0010) $\Leftrightarrow$ (0100). Les analogies restantes sont : $\{$ (0000), (0001), (0010), (0011), (0101), (0110) $\}$.

De façon similaire, la propriété de permutation des moyens donne les équivalences (0010) $\Leftrightarrow$ (0100) et (0011) $\Leftrightarrow$ (0101). Les analogies restantes sont alors : $\{$ (0000), (0001), (0011), (0110) $\}$.

De ces quatre analogies possibles, la seconde (0001) où une seule permutation est opérée n'est pas vraie en général. Par exemple, $ay : az \doteq by : x$ n'est pas vraie pour $x = zb$. Les trois autres analogies semblent intuitives. CQFD

Maintenant, parce que l'essence véritable de la contiguïté nous échappe encore, et afin d'obtenir un certain nombre de résultats importants sur des exemples de langages de chaînes analogiques (voir plus bas, p. 169), nous avons choisi d'admettre un résultat qu'il nous faudra sans doute remttre en cause ultérieurement. Il s'agit d'affirmer que les seules possibilités de combinaison d'analogies disjointes sont les trois seules possibilités de concaténations données plus haut. Cette conjecture très importante nous sera fort utile par la suite.

Conjecture 3 (Concaténation d'analogies disjointes) *Soient $\mathcal{V}$ un alphabet, et $\mathcal{V}_1 \subset \mathcal{V}$, $\mathcal{V}_2 \subset \mathcal{V}$, tels que $\mathcal{V}_1 \cap \mathcal{V}_2 = \phi$,*

$\forall (A_1, B_1, C_1, D_1) \in (\mathcal{V}_1{}^*)^4, \forall (A_2, B_2, C_2, D_2) \in (\mathcal{V}_2{}^*)^4,$

$$\left.\begin{array}{l} A_1 : B_1 \doteq C_1 : D_1 \\ A_2 : B_2 \doteq C_2 : D_2 \end{array}\right\} \Rightarrow \left\{\begin{array}{lll} A_1A_2 : B_1B_2 \doteq C_1C_2 : x & \Rightarrow & x = D_1D_2 \\ A_1A_2 : B_1B_2 \doteq C_2C_1 : x & \Rightarrow & x = D_2D_1 \\ A_1A_2 : B_2B_1 \doteq C_2C_1 : x & \Rightarrow & x = D_1D_2 \end{array}\right.$$

Redoublement

C'est aussi dans le cadre de la contiguïté qu'il nous faut mentionner une intuition importante à propos de l'analogie. Il est en effet courant de penser que le redoublement devrait être toujours possible par analogie. Exprimé sous forme de postulat, cette intuition s'exprime de la façon suivante.

POSTULAT **6 (Redoublement)** *Soit $\mathcal{V}$ un alphabet,*

$$\forall (A, B) \in (\mathcal{V}^*)^2, \quad A : B \doteq AA : BB$$

Il est particulièrement intéressant de constater que cette intuition, et donc ce postulat sont en contradiction avec toute la formalisation que nous avons élaborée jusqu'à maintenant. En effet, on a le théorème suivant.

THÉORÈME **19** *Les hypothèses d'inclusion des symboles et de redoublement sont contradictoires.*

DÉMONSTRATION : L'hypothèse du redoublement contredit le théorème d'égalité des sommes des longueurs qui dérive du théorème de contrainte de similitude. Par exemple, pour l'analogie $a^n : b^m \doteq a^{2n} : b^{2m}$, avec n'importe quels n et m, l'égalité des sommes de longueurs implique $n + 2m = 2n + m$, ce qui est logiquement équivalent à $n = m$. Donc, avec l'hypothèse du redoublement, $|A| + |BB| = |AA| + |B|$ n'est pas vraie en général, quelles que soient les chaînes A et B. CQFD

Le rejet de cette hypothèse permet d'éliminer certaines solutions éventuelles à certaines équations analogiques. Par exemple, l'analogie $a : b \doteq aa : x$ n'a que deux solutions : $x = ab$ et $x = ba$, où a est soit interprété comme un préfixe, soit comme un suffixe. Si le redoublement était autorisé, on aurait, en plus, une troisième solution : $x = bb$.

D'autre part, ce résultat est intéressant parce qu'il met en évidence le fait que l'analogie et le redoublement sont en quelle que sorte orthogonaux. Ils constituent deux espèces distinctes d'analogie. Dans la première, celle de l'analogie proprement dite, on a quatre postulats, d'inversion du sens de la conformité, d'inversion des rapports, d'inversion des objets et de distribution. Dans la seconde, on aurait les postulats d'inversion du sens de la conformité, d'inversion des rapports, d'inversion des objets et le postulat du redoublement. Cette deuxième espce d'analogie est bien moins riche que la première, car en fait de résultats, on n'obtient guère que des redoublements. Du point de vue linguistique, cependant, le redoublement est une opération que l'on ne

saurait rejeter. Elle est en effet nécessaire pour traiter un certain nombre de phénomènes. Par exemple, le pluriel marqué du malais-indonésien repose sur ce phénomène:

$$meja : meja\text{-}meja \doteq pelajar : pelajar\text{-}pelajar \quad [63]$$

Cependant, dans le même ordre d'idées, il reste un corollaire de la conjecture 1 sur la validité de l'itération d'analogies vraies.

COROLLAIRE **6** *Soit $\mathcal{V}$ un alphabet,*

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \forall n \in \mathbb{N}, \quad A : B \doteq C : D \quad \Leftrightarrow \quad A^n : B^n \doteq C^n : D^n$$

DÉMONSTRATION : La fonction qui à une chaîne A fait correspondre A^n et qui associe ε à ε, remplit les critères de la conjecture 1. On devrait donc avoir conservation de l'analogie par cette fonction. CQFD

La répétition locale des symboles intervenant dans les chaînes serait aussi possible. Il s'agit d'une sorte de « bégaiement » sur chacun des symboles. Soit $A = a_1 a_2 \ldots a_m$ une chaîne, on note $A^{\hat{n}}$ la chaîne: $a_1^n a_2^n \ldots a_m^n$.

À titre d'exemple, l'analogie *croire* : *je crois* $\doteq$ *boire* : *je bois* peut-être modifiée en *ccrrooiirree* : *jjee ccrrooiiss* $\doteq$ *bbooiirree* : *jjee bbooiiss* sans changer la validité.

L'énoncé formel de cette propriété est la suivante.

COROLLAIRE **7** *Soit $\mathcal{V}$ un alphabet,*

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad A : B \doteq C : D \quad \Leftrightarrow \quad A^{\hat{n}} : B^{\hat{n}} \doteq C^{\hat{n}} : D^{\hat{n}}$$

DÉMONSTRATION : La fonction qui à une chaîne $a_1.a_2 \ldots a_m$ fait correspondre $a_1^n.a_2^n \ldots a_m^n$ et qui associe ε à ε, remplit les critères de la conjecture 1. On a donc conservation de l'analogie par cette fonction. CQFD

4.4.4 Structure induite sur un vocabulaire

On peut particulariser la structure générale induite sur un ensemble par l'analogie au cas où cet ensemble est un ensemble de chaînes de symboles. On obtient alors la définition suivante.

DÉFINITION **9** *Soit $\mathcal{V}$ un ensemble fini de symboles. Soit $\mathcal{V}^*$ le monoïde libre sur $\mathcal{V}$. On appelle monoïde libre analogique le monoïde libre $\mathcal{V}^*$ sur $\mathcal{V}$ équipé de la structure $(\mathcal{V}^*, \alpha)$.*

[63] *Meja* (une table), *pelajar* (un étudiant). Les formes redoublées sont les formes marquées du pluriel. Si le contexte laisse à comprendre que l'on parle de plusieurs tables ou de plusieurs étudiants, on emploie alors les formes simples. Ce sont aussi les formes simples qui apparaissent après les adjectifs numéraux suivis normalement des quantifieurs dont le malais-indonésien fait usage à l'instar de nombre de langues asiatiques.

Nous n'étudierons pas cette structure générale. Mais nous nous consacrerons plus bas à un cas très particulier où les trois premiers éléments des quadruplets doivent appartenir à des ensembles prédéfinis. Ce cas sera justifié par notre souci d'étudier l'analogie en relation avec la langue (p. 169 et suivantes). En plus, nous nous intéresserons au cas où des monoïdes libres analogiques peuvent être mis en relation pas des homomorphismes. Cela nous permettra, d'une part théoriquement de proposer une formalisation de la notion de métaphore (p. 196), et d'autre part, pratiquement, de concevoir des applications en traitement automatique des langues (p. 289 et suivantes).

4.5 Extensions souhaitables

4.5.1 Le continu : la parole

Puisque notre domaine est le traitement automatique des langues, nous souhaiterions naturellement étendre notre formalisation de l'analogie au domaine de la parole, si possible. Qui dit parole dit signal sonore. Mathématiquement donc, une extension de la formalisation est nécessaire. Il s'agit de passer de représentations symboliques à des représentations numériques. Du point de vue mathématique, il s'agit d'une extension aux fonctions de $\mathbb{R}$ dans $\mathbb{R}$. Pour la partie dont nous possédons la clé, celle de la similitude, il est donc nécessaire d'étendre les notions de similarité ou de distance d'édition au continu. Pour cette deuxième notion, celle de distance d'édition, l'équivalent dans le continu ne semble pas encore vraiment compris par les chercheurs du traitement automatique de la parole. Il semble exister plusieurs propositions, mais aucune ne semble être une extension naturelle ou mathématiquement justifiée[64]. Du point de vue pratique, il nous semble cependant possible de nous ramener au cas des chaînes de symboles. Notre proposition est le suivante.

Considérons le problème de la résolution d'une équation analogique qui serait définie par trois fichiers de son, ou de parole. Premièrement, du point de vue informatique, les fichiers de son sont toujours échantillonnés à une fréquence donnée. Nous nous placerons dans le cas où les trois fichiers A, B et C sont échantillonnés à la même fréquence donnée. Le fichier résultat D sera interprété comme un fichier de signal sonore échantillonné à cette même fréquence. De cette façon, chaque fichier peut être vu comme une chaîne de valeurs entières. Ces valeurs donnent l'amplitude du signal sonore.

Deuxièmement, afin de passer d'une représentation numérique à une représentation symbolique, on normalisera les trois fichiers à une amplitude donnée que nous appelons amplitude de normalisation. Pour cela, chaque valeur est simplement divisée par l'amplitude maximale du fichier (que nous noterons Max) auquel elle appartient puis multipliée par la valeur de l'amplitude de normalisation. On obtient donc des chaînes de valeurs réelles comprises entre 0 et 1.

[64] Voir Kruskal & Liberman, *The symmetric time warping problem: from continuous to discrete*, 1999 pour une synthèse des différentes propositions, et pour un essai de formalisation.

Troisièmement, on calcule l'amplitude maximale de la solution D selon la formule maintenant classique :

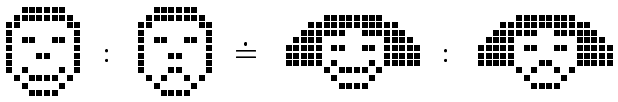
$$\mathrm{Max}_D = \mathrm{Max}_B + \mathrm{Max}_C - \mathrm{Max}_A$$

Quatrièmement, on se ramène à des chaînes de symboles en considérant des sortes de filtres définis par un seuil, et qui mettent à 0 toute valeur inférieure à ce seuil, et à 1 toute valeur supérieure à ce seuil. On peut appliquer n'importe quel nombre de tels filtres. On propose de considèrer les n filtres pour les seuils $\frac{1}{n}$, $\frac{2}{n}$, ...jusqu'à $\frac{n}{n} = 1$. On obtient ainsi n représentations de chaque fichier sous forme de chaînes de symboles binaires. On résoud alors chaque équation analogique obtenue pour chaque filtre.

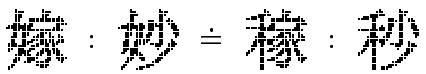
Cinquièmement et finalement, les résultats obtenus pour chaque filtre peuvent être réunis dans un même fichier, puis lissés par des procédés de traitement de la parole, et réétalonnés en divisant par l'amplitude de normalisation et en multipliant par l'amplitude Max_D.

4.5.2 Deux dimensions : les images

La seconde extension que nous désirerions proposer est une extension aux images. Dans un premier temps, afin que le passage soit modéré, et réalisé de façon sûre, nous ne considérerons que les images en noir et blanc. On obtient ainsi un domaine à deux symboles seulement, l'élément d'image noir et l'élément d'image blanc. Le but est alors de résoudre des images du type que nous avions déjà montré (voir p. 113, 115, et 139).



ou



Pour ce qui est de la partie formellement achevée du versant de la similitude, une extension de la distance à la dimension deux est requise. Or, cela a déjà été proposé[65]. De la même façon, il semble relativement facile d'étendre l'algorithme rapide de calcul des similitudes entre chaînes de symboles par opérations binaires[66] à deux dimensions. Si donc les images traitées sont en noir et blanc, de tels algorithmes peuvent jouer directement sur la représentation de ces images en nombres binaires.

La formalisation sur lequel nous devrions aboutir devrait posséder une propriété remarquable. En effet, les images n'ont pas nécessairement à être de même taille. Les hauteurs et les largeurs de A, B et C peuvent être quelconques. Comme conséquence de la formalisation, l'image D, elle, devra vérifier la condition suivante imposée par l'analogie, condition qui reprend simplement le théorème 13 (p. 149) sur les longueurs de chaînes de symboles.

[65] Moore, *A dynamic programming algorithm for the distance between two finite areas*, 1979.

[66] Allison & Dix, *A bit string longest common subsequence algorithm*, 1986.

$$\begin{cases} l_A + l_D &= l_B + l_C \\ h_A + h_D &= h_B + h_C \end{cases}$$

Pour ce qui est du versant de la contiguïté, nous sommes encore incertain de ce qu'il sera. En tout cas, la considération de cette extension à deux dimensions nous aide déjà, par réaction, dans la compréhension de cette notion sur les chaînes de symboles.

Après les images en noir et blanc, il nous faudra envisager les images en couleur. Dans de telles images, chaque élément d'image est en fait un triplet de trois valeurs réelles ou entières dans un espace à trois dimensions de couleur. Les axes sont le plus souvent les couleurs bleu, rouge et vert, mais il existe aussi d'autres représentations. En particulier, pour une qualité d'impression meilleure, la quadrichromie définit des espaces à quatre dimensions sur trois couleurs, magenta, cyan et jaune, plus une dimension de noir. Notre idée actuelle est d'utiliser une technique semblable à celle proposée pour le son, par filtres.

Pour le moment, cependant, toutes ces propositions restent du domaine du projet. Nous avons cependant bon espoir qu'à un horizon de quelques années ces extensions seront tout de même réalisées.

Chapitre 5

Langages de chaînes analogiques

Nous nous proposons maintenant, grâce à l'analogie, de jeter un éclairage nouveau sur les langages formels. En fait, il nous semble que l'utilisation d'une notion « naturelle » comme l'analogie dans la langue permettrait d'introduire justement un côté plus « naturel » dans les langages formels, côté qui, personnellement, nous semblait y faire défaut. Nous montrerons que les langages formels que nous définirons grâce à l'analogie collent tout de suite mieux à la langue parce qu'ils semblent se placer d'emblée dans le domaine du modérément sous-contexte (voir plus haut p. 75 et plus bas p. 176). Les Générativistes ont exhibé les langages formels comme argument péremptoire en faveur de la prétendue scientificité, et donc de la prétendue supériorité, de leur approche de la linguistique. En particulier, ils ont reproché aux Structuralistes leur manque d'outils formels. L'introduction de l'analogie, notion éminemment structuraliste, dans le royaume des langages formels nous tenait à cœur pour contester cette vue des choses.

Nous procéderons en plusieurs étapes. Dans un premier temps, nous nous inspirerons des présentations classiques des langages formels pour introduire une notion de dérivation analogique. Nous en tirerons immédiatement la définition des langages de chaînes analogiques. Nous pourrons alors, à partir de cette définition, proposer une classification purement formelle des langages ainsi définis.

Cette classification nous sera utile dans un deuxième temps, lorsque nous démontrerons certaines propriétés des langages de chaînes analogiques. En particulier, nous montrerons que ces langages possèdent une bonne propriété des langages modérément sous-contexte, celle de croissance constante des longueurs.

Dans un troisième temps, nous illustrerons les langages de chaînes analogiques par des exemples. En particulier, nous montrerons que le langage formel clé de la preuve contre la nature prétendument hors-contexte de la langue (voir p. 75) est un langage de chaînes analogiques relativement simple.

Enfin, dans un quatrième temps nous essaierons de définir formellememt dans le cadre d'une approche par l'analogie, la notion de représentativité au moyen de celles d'ensembles libres et générateurs, notions bien connues en algèbre, en attendant d'aborder le problème de la surproduction que nous

avions déjà vu lors de l'histoire de l'analogie que nous avons tracée (p. 76).

5.1 Définitions

5.1.1 Dérivation

Avant de montrer comment certains langages, c'est-à-dire certains ensembles de chaînes de symboles peuvent être engendrés par une procédure mécanique fondée sur l'analogie, nous introduisons la notion de dérivation analogique immédiate. Nous utilisons ce terme de façon à établir un parallèle avec le vocabulaire des systèmes formels.

DÉFINITION **10 (Dérivation analogique immédiate)** *Soit $\mathcal{V}$ un alphabet. Soit $\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$ dont les éléments (A, B) sont notés $A \rightarrow B$. La dérivation analogique immédiate modulo $\mathcal{M}$, notée $\vdash_{\overline{\mathcal{M}}}$, est définie de la façon suivante.*

$$\forall (C, D) \in \mathcal{V}^* \times \mathcal{V}^*, \quad C \vdash_{\overline{\mathcal{M}}} D \quad \Leftrightarrow \quad \exists A \rightarrow B \in \mathcal{M} \; / \quad A : B \doteq C : D$$

Cette définition

- ancre deux des termes de l'analogie dans un ensemble prédéfini ;
- interprète l'analogie comme une équation à résoudre ;
- définit la direction de la dérivation, c'est-à-dire celle de l'équation à résoudre en suivant l'ordre donné des couples de $\mathcal{M}$.

Bien que nous utilisions la notation $\rightarrow$, elle ne doit pas être interprétée dans le sens habituel des systèmes de réécriture. Cette notation se veut juste un parallèle avec les présentations classiques des grammaires formelles, où les éléments de $\mathcal{M}$ sont appelés règles. Avec des règles habituelles, C est comparé exactement avec A pour produire, dans une seconde étape, D. Donc, D dépend fondamentalement de la façon dont C et A sont comparés. Ici, c'est différent. Le résultat dépend de la façon dont A est comparé *simultanément à C et à B*. De fait, il serait préférable de noter $A : B$ les éléments de $\mathcal{M}$ pour conserver le lien avec l'analogie. Or, dans une analogie, ce rapport, $A : B$, peut être compris comme une fonction de $\mathcal{V}^*$ dans l'ensemble des parties de $\mathcal{V}^*$, puisque zéro, une ou plusieurs solutions sont possibles pour une équation analogique. La dérivation analogique à partir de C est donc simplement l'application de la fonction $A : B$ à C. En notation fonctionnelle, on aurait donc : $\{D\} = (\ A : B\ )(C)$.

5.1.2 Langages

DÉFINITION **11 (Langages de chaînes analogiques)** *Soit $\mathcal{V}$ un alphabet. Soient $\mathcal{A} \subset \mathcal{V}^*$ et $\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$, tous deux finis et non vides. Alors, $\Lambda(\mathcal{A}, \mathcal{M}) = < \mathcal{A}, \{\vdash_{\overline{\mathcal{M}}}\} >$ est le langage de chaînes analogiques défini de la façon suivante*

$$\Lambda(\mathcal{A}, \mathcal{M}) = \mathcal{A} \cup \{\ D \in \mathcal{V}^* \; / \; \exists C \in \mathcal{A},\ C \vdash^{+}_{\overline{\mathcal{M}}} D\ \}$$

avec $\vdash^{+}_{\overline{\mathcal{M}}}$, la fermeture transitive de la dérivation analogique immédiate $\vdash_{\overline{\mathcal{M}}}$. $\mathcal{A}$ est appelé l'ensemble des chaînes attestées et $\mathcal{M}$ l'ensemble des modèles.

La définition précédente suit les présentations habituelles des langages formels. Son but est la *production* d'éléments du langage. Par conséquent, et comme d'habitude, on utilise l'induction structurale pour produire tous les éléments d'un langage de chaînes analogiques. En partant des éléments de $\mathcal{A}$, on applique toutes les analogies possibles avec les éléments de $\mathcal{M}$ comme modèles. On peut donc aussi écrire, en posant

$$\begin{array}{rcl}
\Lambda_0 & = & \mathcal{A} \\
\Lambda_1 & = & \{D \in \mathcal{V}^* \;/\; \exists C \in \Lambda_0,\ C \vdash_{\overline{\mathcal{M}}} D\} \\
 & = & \{D \in \mathcal{V}^* \;/\; \exists C \in \Lambda_0,\ \exists A \to B \in \mathcal{M}, \quad A : B \doteq C : D \ \} \\
 & \vdots & \\
\Lambda_{i+1} & = & \{D \in \mathcal{V}^* \;/\; \exists C \in \Lambda_i,\ C \vdash_{\overline{\mathcal{M}}} D\} \\
 & = & \{D \in \mathcal{V}^* \;/\; \exists C \in \Lambda_i,\ \exists A \to B \in \mathcal{M}, \quad A : B \doteq C : D \ \} \\
 & \vdots &
\end{array}$$

que le langage $\Lambda(\mathcal{A}, \mathcal{M})$ est défini par

$$\Lambda(\mathcal{A}, \mathcal{M}) = \bigcup_{i=0}^{\infty} \Lambda_i$$

On appellera chaque Λ_i la i^{e} couche du langage de chaînes analogiques. Si on impose que $\mathcal{A}$ et $\mathcal{M}$ soient finis, c'est parce qu'évidemment on souhaite construire de l'infini à partir du fini. Considérer ces deux ensembles comme infinis n'aurait pas grand intérêt théorique.

Le problème réciproque de la production est celui de la *reconnaissance*. Avec un système par analogie, la grammaticalité d'une chaîne donnée, c'est-à-dire son appartenance au langage, est testée par comparaison avec les chaînes attestées de ce langage, après réduction de la chaîne donnée, par analogie, en utilisant l'ensemble des modèles. En reconnaissance, les chaînes qui apparaissent dans les couples de $\mathcal{M}$ sont utilisées en sens inverse de celui dans lequel elles apparaissent dans $\mathcal{M}$, et les analogies sont résolues dans le sens opposé à celui de la production. Cela est formellement possible grâce au théorème 1 (p. 116) des formes équivalentes de l'analogie. La figure 5.1 donne un schéma du processus de la reconnaissance.

L'interprétation linguistique de ce qui précède est la suivante: $\mathcal{A}$ est l'ensemble des chaînes attestées, c'est-à-dire l'ensemble des chaînes de symboles par rapport auxquelles tout élément du langage est testé *in fine*, et $\mathcal{M}$ est l'ensemble des modèles utilisé pour réduire[1] toute chaîne du langage, par analogie. Cette approche, par réduction à des formes attestées, en utilisant des paradigmes de déclinaisons, conjugaisons, dérivations morphologiques ou transformations syntaxiques, semble intuitivement proche de la façon dont nous, êtres humains, vérifions la grammaticalité de phrases nouvelles. Elle peut aussi être reliée à l'intuition fondamentale de la syntaxe transformationnelle.

[1] Le terme *réduire* est à prendre au sens de réduire à une forme canonique. Pas au sens où ces chaînes de symboles deviendraient nécessairement plus courtes.

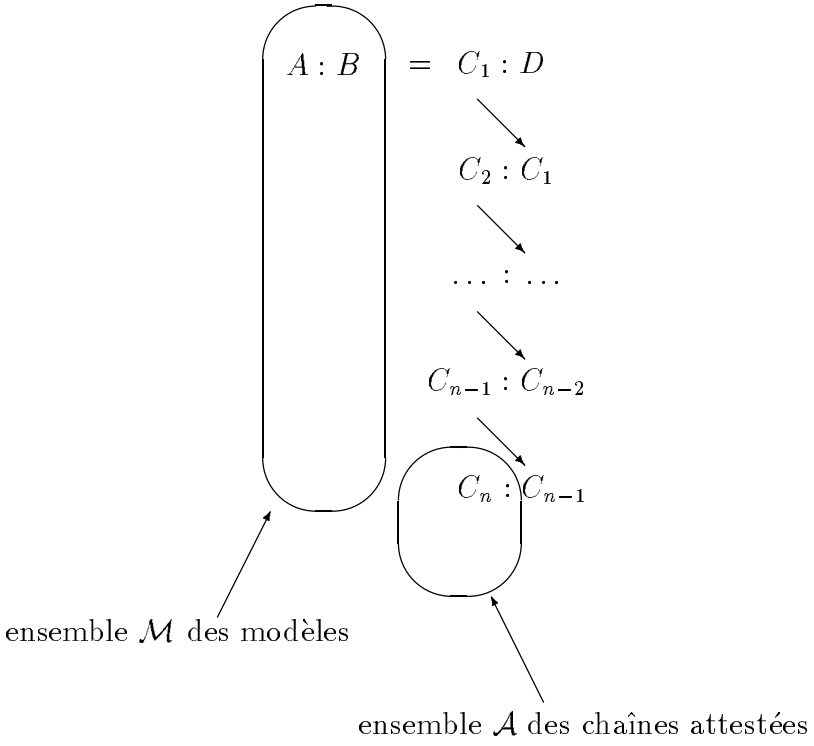


Figure 5.1: Reconnaissance de D comme élément du langage $\Lambda(\mathcal{A}, \mathcal{M})$

> Deux phrases ayant la même structure distributionnelle [...] n'ont pas la même structure syntaxique lorsqu'elles n'acceptent pas les mêmes transformations. [...] De cette technique Harris et Chomsky (première manière) ont tiré une théorie syntaxique selon laquelle la syntaxe d'une langue est composée d'un stock relativement réduit de phrases de base ou **phrases-noyaux** [...] dont toutes les autres phrases de la langue peuvent ou doivent être dérivées par des opérations dites **transformations** [...].[2]

Cette intuition pourrait être vue comme une justification par un impératif théorique du fait que nous imposions à $\mathcal{A}$ et $\mathcal{M}$ d'être de cardinal fini.

Une autre remarque. La présentation précédente ne fait en aucun cas intervenir la notion de non-terminal. Ce n'est pas une chose nouvelle en traitement automatique des langues[3] ni non plus dans le cadre des grammaires formelles, puisque d'autres systèmes de production sont ainsi. Dans le domaine de la linguistique mathématique, en particulier, les langages contextuels de Solomon Marcus présentent aussi cette particularité.

> Les grammaires contextuelles, introduites dans MARCUS, *Contextual grammars*, 1969, constituent un mécanisme de production n'utilisant

[2] MOUNIN, *Clefs pour la linguistique*, 1968, p. 125.
[3] Par exemple, SALKOFF, *Une grammaire en chaîne du français*, 1973.

pas de symboles auxiliaires et fondé sur une opération linguistique fondamentale, à savoir l'acceptation d'un contexte par un mot.[4]

Pour notre part, nous voyons dans l'absence de non-terminaux un avantage important : cela permet d'éviter les débats sur les parties du discours auxquels nous avons déjà fait allusion plus haut (p. 74). En plus, notre espoir est que l'analogie sous sa forme algorithmique puisse contribuer, par inspection automatique des commutations possibles, à la constitution automatique de classifications des parties du discours à la manière des syntaxes distributionnelles[5].

5.1.3 Classification

Après avoir posé leur définition, nous pouvons maintenant proposer une classification des langages de chaînes analogiques qui repose directement sur les éléments constitutifs de ces langages, l'ensemble des chaînes attestées et l'ensemble des modèles.

DÉFINITION **12 (Langage de chaînes analogiques simple)** *Un langage de chaînes analogiques* $\Lambda(\mathcal{A},\mathcal{M})$ *est dit* simple *si et seulement si* $\mathcal{A}$ *est un singleton.*

En d'autres termes, un langage de chaînes analogiques est simple quand tout élément du langage de chaînes analogiques est reconnu par comparaison avec une et un seule chaîne de ce langage en dernier ressort.

DÉFINITION **13 (Langage de chaînes analogiques monotone)** *Un langage de chaînes analogiques* $\Lambda(\mathcal{A},\mathcal{M})$ *est dit* monotone *si et seulement si* $\mathcal{M}$ *est un singleton.*

En d'autres termes, un langage de chaînes analogiques est monotone quand il n'y a qu'une façon, toujours la même, de réduire les éléments de ce langage.

DÉFINITION **14 (Langage de chaînes analogiques élémentaire)** *Un langage de chaînes analogiques* $\Lambda(\mathcal{A},\mathcal{M})$ *est dit* élémentaire *si et seulement si* $\mathcal{A}$ *et* $\mathcal{M}$ *sont des singletons.*

En d'autres termes, pour un langage de chaînes analogiques élémentaire, il n'y a qu'une seule chaîne attestée et qu'un seul modèle, ou encore il est à la fois monotone et simple. Bien que nous ne démontrions de telles égalités que plus tard (p. 183, 183 et 184), donnons dès maintenant des exemples frappants de langages de chaînes analogiques élémentaires.

$$\begin{array}{rcll} \Lambda(\{a\},\{a\rightarrow aa\}) & = & \{a^n\ /n\geq 1\} & \text{(langage régulier)} \\ \Lambda(\{ab\},\{ab\rightarrow aabb\}) & = & \{a^nb^n\ /n\geq 1\} & \text{(langage hors-contexte)} \\ \Lambda(\{abc\},\{abc\rightarrow aabbcc\}) & = & \{a^nb^nc^n\ /n\geq 1\} & \text{(langage sous-contexte)} \end{array}$$

[4]Première phrase du premier article du livre édité par MARTÍN-VIDE, *Mathematical and computational analysis of natural language*, 1998, qui a pour objet les grammaires contextuelles.

[5]MOUNIN, *Clefs pour la linguistique*, 1968, p. 118–119.

La classification précédente repose uniquement sur le cardinal de l'ensembles de chaînes attestées et de l'ensemble des modèles. Nous nous penchons maintenant sur la forme que peuvent prendre les modèles d'un langage de chaînes analogiques. La première définition proposée nous sera utile dans l'obtention de résultats sur la complexité d'analyse des langages de chaînes analogiques. Il s'agit en fait de retrouver le sens intuitif de réduction par analogie contre lequel nous avions pris soin de mettre en garde dans une note de bas de page (p. 170).

DÉFINITION **15 (Modèle décroissant)** *Soit* $(A, B) \in (\mathcal{V}^*)^2$, *noté* $A \to B$. $A \to B$ *est dit* décroissant[6] *si et seulement si*

$$|A| < |B|$$

Le lien avec l'analogie est fait par le thórème 13 (p. 149) qui énonce l'égalité de la somme des longueurs des moyens et des extrêmes. Rappelons qu'en reconnaissance, à partir d'un élément dont on veut prouver l'appartenance au langage $\Lambda(\mathcal{A}, \mathcal{M})$, on applique les modèles en sens inverse de leur donnée dans $\mathcal{M}$. Ainsi, si on veut tester l'appartenance de D, et pour le modèle $A \to B$, on cherchera à résoudre l'équation analogique $B : A \doteq D : C$ d'inconnue C. Si des solutions C à une telle équation existent, d'après le théorème 13 (p. 149) elles vérifieront l'égalité

$$|A| + |D| = |B| + |C|$$

c'est-à-dire

$$|D| - |C| = |B| - |A|$$

Cette différence étant positive puisque $A \to B$ est un modèle décroissant, on aura aussi nécessairement $|C| < |D|$. Autrement dit, en reconnaissance, et c'est ce que nous voulions, l'application d'un modèle décroissant réduit les longueurs. Pour un langage donné, afin d'obtenir cela à tout coup, nous posons la définition suivante.

DÉFINITION **16 (Langage de chaînes analogiques décroissant)** *Un langage de chaînes analogiques* $\Lambda(\mathcal{A}, \mathcal{M})$ *est dit* décroissant *si et seulement si tous les éléments de* $\mathcal{M}$ *sont des modèles décroissants.*

Les trois langages que nous venons de donner en exemples (p. 172) sont des langages de chaînes analogiques décroissants.

Voyons maintenant une seconde définition dont la justification est pratique. En effet, un modèle grossier d'apprentissage des langues comme langages de chaînes analogiques considérerait que l'ensemble des modèles est construit à partir des chaînes attestées. Or, celles-ci, et celles-ci seulement, constituent les données immédiates de l'expérience. À partir de là, tous les modèles possibles

[6]Oui, et pas croissant ! Car ce qui nous intéresse, c'est le rapport $B : A$ qui reflète l'application du modèle en reconnaissance.

et imaginables construits à partir de ces données peuvent être tentés pour reconnaître des données nouvelles. On entrevoit donc un mode d'apprentissage paresseux où les modèles sont simplement tous les couples possibles formés avec les chaînes attestées. Dans une tel mode d'apprentissage, il n'y a aucun travail d'élaboration sur les modèles. Nous justifions ainsi le fait que cette représentation de la langue peut être qualifée de *paresseuse* et nous posons la définition suivante.

DÉFINITION **17 (Langage de chaînes analogiques paresseux)** *Un langage de chaînes analogiques* $\Lambda(\mathcal{A},\mathcal{M})$ *est dit* paresseux *si et seulement si* $\mathcal{M}=\mathcal{A}\times\mathcal{A}$.

Pour de telles langages, la suite des langages Λ_i donnée plus haut (p. 170) peut se réécrire de la façon suivante en faisant l'abus de langage qui consiste à étendre l'analogie entre chaînes de symboles à des ensembles de telles chaînes de symboles, en notant l'ensemble

$$\mathcal{D}=\{D\in\mathcal{V}^*\ /\ \exists(A,B,C)\in\mathcal{A}\times\mathcal{B}\times\mathcal{C},\quad A:B\doteq C:D\ \}$$

par l'analogie $\mathcal{A}:\mathcal{B}\doteq\mathcal{C}:\mathcal{D}$. On obtient de cette façon l'expression suivante. Nous notons par $\wp(\mathcal{E})$ l'ensemble des parties de $\mathcal{E}$.

$$\begin{array}{l}\Lambda_0=\mathcal{A}\\ \Lambda_1\subset\wp(\mathcal{V}^*)\ /\quad \mathcal{A}:\mathcal{A}\doteq\Lambda_0:\Lambda_1\\ \quad\vdots\\ \Lambda_{i+1}\subset\wp(\mathcal{V}^*)\ /\quad \mathcal{A}:\mathcal{A}\doteq\Lambda_i:\Lambda_{i+1}\\ \quad\vdots\end{array}$$

c'est-à-dire

$$\Lambda(\mathcal{A},\mathcal{M})=\bigcup_{i=0}^{\infty}\Lambda_i\quad/\quad\Lambda_0=\mathcal{A}\wedge\forall i\in\mathbb{N},\Lambda_{i+1}\subset\wp(\mathcal{V}^*)\ /\quad\mathcal{A}:\mathcal{A}\doteq\Lambda_i:\Lambda_{i+1}$$

De cette façon, le langage $\Lambda(\mathcal{A},\mathcal{M})$ prend la forme de la limite d'une sorte de série (la sommation entre ensembles étant l'union) correspondant à une sorte de suite arithmétique ou géométrique, d'élément de départ $\mathcal{A}$, et de raison $\mathcal{A}:\mathcal{A}$.

DÉFINITION **18 (Langage de chaînes analogiques paresseux décroissant)** *Un langage de chaînes analogiques* $\Lambda(\mathcal{A},\mathcal{M})$ *est dit* paresseux décroissant *si et seulement si* $\mathcal{M}$ *est l'ensemble des modèles décroissants de* $\mathcal{A}\times\mathcal{A}$.

Les langages que nous avons donnés plus haut (p. 172) comme exemples de langages de chaînes analogiques élémentaires vérifient la définition donnée ci-dessus si l'on inclut *aa*, *aabb* ou *aabbcc* dans les ensembles de chaînes attestées. Ce sont donc aussi des langages de chaînes analogiques paresseux décroissants. La paresse de ce type de langage recouvre en fait une certaine immédiateté formelle.

Poursuivons le parallèle entre langage de chaînes analogiques paresseux et suites arithmétiques ou géométriques. Nous avions entr'aperçu le parallèle entre langage de chaînes analogiques élémentaire paresseux et suite arithmétique ou géométrique (p. 81). Il est bien visible maintenant dans le cas des exemples précédents de langages de chaînes analogiques élémentaires paresseux. L'ensemble de chaînes attestées $\mathcal{A}$ contient en effet les deux premiers éléments de la suite, u_0 et u_1 c'est-à-dire a et aa, ab et $aabb$ ou abc et $aabbcc$ selon le cas. Quant au singleton $\mathcal{M}$, il donne la raison de la suite, c'est-à-dire le rapport des deux premiers éléments, $u_0 : u_1$, c'est-à-dire $a : aa$, $ab : aabb$ ou $abc : aabbcc$ selon le cas.

Enfin, la paresse de ce type de langage reflète aussi une certaine simplicité de réalisation et dans toutes nos expériences, aussi bien d'analyse par analogie (p. 294) que de traduction directe par analogie (p. 314), nous utiliserons des langages de chaînes analogiques paresseux décroissants.

5.2 Propriétés

5.2.1 Décomposition en langages simples

Nous donnons maintenant une vue différente des langages de chaînes analogiques qui établit un lien entre leur définition et leur classification. Pour ce faire, nous décomposons simplement un langage de chaînes analogiques, non pas par étapes de production, mais par chaîne attestée du langage. On obtient la formulation suivante :

LEMME **13** *Soit* $\Lambda(\mathcal{A},\mathcal{M})$ *un langage de chaînes analogiques.*

$$\Lambda(\mathcal{A},\mathcal{M}) = \bigcup_{a\in\mathcal{A}} \Lambda(\{a\},\mathcal{M})$$

DÉMONSTRATION : Quel que soit D, un élément de $\Lambda(\mathcal{A},\mathcal{M})$, il existe une suite $C_0, C_1, \ldots, C_{n+1} = D$ commençant par C_0, élément de $\mathcal{A}$, telle que

$$\forall i \in \mathbb{N} \;/\; 0 < i \leq n,\;\; \exists A \to B \in \mathcal{M},\quad A : B \doteq C_i : C_{i+1}$$

On a donc :

$$\forall D \in \Lambda(\mathcal{A},\mathcal{M}),\;\; \exists a \in \mathcal{A},\;\; D \in \Lambda(\{a\},\mathcal{M})$$

ce qui suffit à établir le résultat. CQFD

Autrement dit, en utilisant la classification des langages de chaînes analogiques que nous avons introduite plus haut (p. 172), tout langage de chaînes analogiques peut s'exprimer comme une réunion finie de langages de chaînes analogiques simples puisque, par définition de ces langages (p. 169), $\mathcal{A}$ doit être un ensemble fini. Cette propriété va nous être utile pour étudier l'adéquation des langages de chaînes analogiques aux langues (démonstration du théorème 20, p. 180).

5.2.2 Croissance constante des longueurs

Nous nous posons maintenant la question de l'adéquation des langages de chaînes analogiques à la description des langues humaines[7]. De façon général, le problème de l'adéquation des modèles de langages formels à la description des langues est un problème qui tire son origine de l'idée proposée par les Générativistes que les langues seraient hors-contexte. On a vu plus haut que cette hypothèse a été réfutée. La recherche s'est alors orientée vers une caractérisation à partir de certains modèles du traitement automatique des langues.

Ainsi, le modérément sous-contexte a été introduit par Joshi[8] pour tenter de caractériser la famille des langages formels qui recouvrirait exactement les langues, à partir des grammaires par adjonction d'arbres qu'il a lui-même

[7] Cette partie reprend et développe LEPAGE, *Analogy and formal languages*, 2001, p. 8 à 11.

[8] JOSHI, *Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural description?*, 1985.

proposées. D'après ce que nous avons déjà vu (page 75), cette famille devrait nécessairement déborder du hors-contexte, mais devrait être strictement incluse dans le sous-contexte. Il existe une caractérisation à l'aide de quatre propriétés :

> À partir des propriétés formelles des grammaires par adjonction d'arbres, Joshi (1985) a proposé que la classe des grammaires nécessaires à la description des langues soit caractérisée comme la classe des grammaires *modérément sous-contexte* (langages modérément sous-contexte pour les langages correspondants) possédant au moins les propriétés suivantes :
>
> - (1) la famille des langages hors-contexte est incluse dans la famille des langages modérément sous-contexte ;
> - (2) les langages modérément sous-contexte peuvent être analysés en temps polynomial ;
> - (3) les grammaires modérément sous-contexte rendent seulement compte de certains types de dépendance, tels que les dépendances enchâssées et certains types restreints de structures de dépendances croisées (par exemple, telles que l'on en trouve en néerlandais dans la construction des propositions subordonnées ou de leurs variantes, mais peut-être pas le langage MIX (ou langage de Bach), qui consiste en un nombre égal de a, de b et de c dans n'importe quel ordre) ;
> - (4) les langages modérément sous-contexte ont la propriété de croissance constante.
>
> Cette dernière propriété signifie que si les chaînes du langage sont rangées par ordre croissant de longueur, alors deux longueurs consécutives ne diffèrent pas par une quantité arbitrairement grande. En fait, chaque longueur donnée peut être décrite comme combinaison linéaire d'un ensemble fini de longueurs fixées. Cette propriété est légèrement plus faible que la propriété de semi-linéarité. Elle est donnée comme une approximation de l'intuition linguistique que les phrases d'une langue sont construites à l'aide de certaines opérations linéaires simples à partir d'un ensemble fini de propositions de structure bornée. La caractérisation de cette intuition par la propriété de croissance constante est une approximation, parce qu'elle fait appel à la croissance des longueurs et non pas à la croissance des structures.[9]

Une thèse[10] veut que le point clé du « modérément sous-contexte » soit en fait la propriété de croissance des longueurs plutôt que les deux propriétés de semi-linéarité et d'analyse en temps polynomial. À l'appui, partiel, de cela,

[9] Joshi *et al.*, *The convergence of mildly context-sensitive grammar formalisms*, 1991, p. 32. Nous énonçons chaque propriété dans un alinéa différent pour plus de clarté.

[10] Marcus *et al.*, *Contextual grammars versus natural languages*, 1996.

un récent travail montre que certaines langues ne possèdent pas la propriété de semi-linéarité [11].

L'idée de rendre compte de l'intuition que toute phrase d'une langue donnée serait obtenue par combinaison linéaire d'un nombre fini de structures élémentaires est à rapprocher de ce que nous avons vu plus haut sur les langages de chaînes analogiques (voir page 176), à savoir que tout langage de chaînes analogiques peut s'exprimer comme réunion fini de langages de chaînes analogiques simples. Il semblerait donc logique que ces langages possédassent la propriété de croissance constante des longueurs. Donnons donc de cette propriété une expression formelle[12].

DÉFINITION **19 (Croissance constante des longueurs)** *Soit $\mathcal{L}$ un langage. On note $L(\mathcal{L}) \subset \mathbb{N}$ l'ensemble des longueurs des chaînes de $\mathcal{L}$. Soit l un élément de $L(\mathcal{L})$. On note par $\mathrm{s}(l)$ le plus petit élément supérieur strictement à l dans $L(\mathcal{L})$ quand il existe. Un langage $\mathcal{L}$ vérifie la propriété de croissance constante des longueurs si et seulement si*

$$L(\mathcal{L}) \text{ est fini} \quad \vee \quad \exists k \in \mathbb{N} \;/\; \forall l \in L(\mathcal{L}),\ \mathrm{s}(l) - l \leq k$$

Dans un premier temps, on établit une série de petits lemmes qui montrent que la propriété est conservée par réunion finie d'ensembles la vérifiant. On décompose étape par étape, selon que les langages sont finis ou infinis.

LEMME **14** *Une réunion finie de langages finis possède la propriété de croissance constante des longueurs.*

DÉMONSTRATION : Comme une réunion finie d'ensembles finis est finie, la propriété de croissance constante des longueurs est vérifiée par définition de cette propriété. CQFD

LEMME **15** *Une réunion finie de langages infinis possédant chacun la propriété de croissance constante des longueurs possède cette propriété.*

DÉMONSTRATION : Soit $\mathcal{L} = \bigcup_{0 \leq i \leq n} \mathcal{L}_i$ où chaque $\mathcal{L}_i$, langage infini, vérifie la propriété de croissance constante des longueurs. On note par $\mathrm{s}_{\mathcal{L}}$ le successeur de l dans $L(\mathcal{L})$. Pour chaque $\mathcal{L}_i$, soit un k_i, pas nécessairement le plus petit, pour lequel la propriété en question est vérifiée. On prend :

$$k = \max_{0 \leq i \leq n} k_i$$

[11] MICHAELIS & KRACHT, *Logical aspects of computational linguistics*, 1997.

[12] Cette définition sent, de façon évidente, sa notion de continuité, voire de dérivation. Est-il possible de rapprocher cela de la notion de voisinage que nous avions introduite dans LEPAGE, *Regular languages have regular neighbourhoods for the Wagner and Fischer distance*, 1994 ? Puisque nous allons mettre en relation l'analogie avec la croissance constante des longueurs, et puisque l'analogie entretient un lien étroit avec les distances d'éditions, ces voisinages pourraient peut-être être mis en rapport avec la croissance constante des longueurs. On tomberait peut-être sur une expression qui rappellerait plus clairement la continuité.

On peut alors écrire :

$$\forall l \in \bigcup_{0 \leq i \leq n} \mathcal{L}_i, \quad \mathrm{s}_{\left(\bigcup_{0 \leq i \leq n} \mathcal{L}_i\right)}(l) - l \ \leq \ \mathrm{s}_{\mathcal{L}_i}(l) - l \ \leq \ k_i \ \leq \ k$$

Ce qui prouve la propriété de croissance constante des longueurs pour $\mathcal{L}$. CQFD

LEMME **16** *La réunion de deux langages, l'un fini, l'autre infini, possédant la propriété de croissance constante des longueurs possède cette propriété.*

DÉMONSTRATION : Appelons $\mathcal{L}_F$ le langage fini et $\mathcal{L}_I$ le langage infini. On pose $\mathcal{L} = \mathcal{L}_F \cup \mathcal{L}_I$. D'une part, posons $k_F = \max(\mathrm{s}(l) - l)$ pour tout l de $\mathcal{L}_F$ pour lequel $\mathrm{s}(l)$ est défini. D'autre part, pour le langage infini, soit k_I un entier pour lequel la propriété est vérifiée. On a :

$$\forall l \in L(\mathcal{L}_I), \quad \mathrm{s}_{(\mathcal{L}_I \cup \mathcal{L}_F)}(l) - l \ \leq \ \mathrm{s}_{\mathcal{L}_I}(l) - l \ \leq \ k_I \ \leq \ \max(k_I, k_F)$$

Comme $\mathcal{L}_I$ est infini, toute longueur dans $L(\mathcal{L}_F)$ a un successeur dans $L(\mathcal{L}_I \cup \mathcal{L}_F)$. Nécessairement, on a :

$$\forall l \in L(\mathcal{L}_F), \quad \mathrm{s}_{(\mathcal{L}_I \cup \mathcal{L}_F)}(l) - l \leq \max(k_I, k_F)$$

Les deux propositions précédentes impliquent que :

$$\forall l \in L(\mathcal{L}_I \cup \mathcal{L}_F), \ \ \mathrm{s}_{(\mathcal{L}_I \cup \mathcal{L}_F)}(l) - l \ \leq \ \max(k_I, k_F)$$

CQFD

LEMME **17** *Une réunion finie de langages possédant la propriété de croissance constante des longueurs possède cette propriété.*

DÉMONSTRATION : Soit $\mathcal{L} = \bigcup_{i=0}^{n} \mathcal{L}_I$, où chaque $\mathcal{L}_I$ possède la propriété de croissance constante des longueurs. On peut séparer la réunion en deux réunions : les langages finis d'un côté, et les langages infinis de l'autre. Le lemme 14 (p. 178) permet de conclure que la réunion (finie) des langages finis dont $\mathcal{L}$ est constitué est un langage fini vérifiant la propriété. Le lemme 15 (p. 178) permet de conclure que la réunion (finie) des langages infinis dont $\mathcal{L}$ est constitué est un langage infini vérifiant la propriété. Le lemme 16 (p. 179) permet de conclure que la réunion des deux réunions précédentes, c'est-à-dire $\mathcal{L}$, est un langage vérifiant la propriété. CQFD

Le théorème que nous nous proposons maintenant de démontrer énonce que les langages auxquels nous nous intéressons, et qui sont définis par l'analogie, possèdent cette propriété du modérément sous-contexte. Cela est intéressant, car nous allons ensuite donner des exemples de langages de chaînes analogiques qui sont des langages sous-contexte (voir plus bas, page 184 et 186). Le théorème qui suit énonce donc que les langages de chaînes analogiques n'iraient pas trop loin dans le sous-contexte. La question restant en suspens étant évidemment : juste assez loin ?

THÉORÈME **20 (Croissance constante des longueurs)** *Tout langage de chaînes analogiques vérifie la propriété de croissance constante des longueurs.*

Nous allons prouver ce théorème en deux étapes. Tout d'abord, nous établissons que :

LEMME **18** *Tout langage de chaînes analogiques simple vérifie la propriété de croissance constante des longueurs.*

DÉMONSTRATION : Soit $\Lambda(\mathcal{A},\mathcal{M})$ un langage de chaînes analogiques simple. Par définition des langages de chaînes analogiques simples, $\mathcal{A}$ est un singleton. Notons C_0 son unique élément.

Posons $k_{\mathcal{M}} = \max_{A\to B\in\mathcal{M}}(|B|-|A|)$. $k_{\mathcal{M}}$ existe car, par définition des langages de chaînes analogiques, $\mathcal{M}$ est fini et non vide.

Quel que soit un élément D de $\Lambda(\mathcal{A},\mathcal{M})$, il existe une suite $C_0, C_1, \ldots, C_{n+1} = D$ commençant par C_0, l'unique élément de $\mathcal{A}$, et telle que

$$\forall i \in \mathbb{N} \;/\; 0 < i \leq n,\;\; \exists A \to B \in \mathcal{M},\quad A : B \doteq C_i : C_{i+1}$$

Le théorème 13 (p. 149) d'égalité des longueurs permet d'écrire : $|A|+|C_{i+1}| = |B|+|C_i|$, et donc : $|C_{i+1}|-|C_i| = |B|-|A| \leq k_{\mathcal{M}}$. La chaîne C_{i+1} est donc située dans dans un « disque » de centre C_i et de rayon $k_{\mathcal{M}}$, soit avant, soit après C_i.

Réciproquement, quelle que soit l une longueur de $L(\Lambda(\mathcal{A},\mathcal{M}))$, son successeur dans $L(\Lambda(\mathcal{A},\mathcal{M}))$, s'il existe, est donc nécessairement à une distance inférieure à $k_{\mathcal{M}}$, ce qui s'écrit :

$$\forall l \in L(\Lambda(\mathcal{A},\;\mathcal{M})),\; s(l) - l \leq k_{\mathcal{M}}$$

Cela établit que $\Lambda(\mathcal{A},\mathcal{M})$, avec $\mathcal{A}$ singleton, vérifie la propriété de croissance constante des longueurs. CQFD

La démonstration du théorème 20 (p. 180) est alors simple par utilisation des lemmes précédents.

DÉMONSTRATION : Soit $\Lambda(\mathcal{A},\mathcal{M})$ un langage de chaînes analogiques quelconque. Par le lemme 13 (p. 176), c'est une réunion finie de langages de chaînes analogiques simples qui, par le lemme 18 (p. 180) que nous venons d'établir, vérifient chacun la propriété de croissance constante des longueurs. Par le lemme 17 (p. 179), $\Lambda(\mathcal{A},\mathcal{M})$ vérifie donc la propriété de croissance constante des longueurs. CQFD

Une démonstration directe du théorème 20 (p. 180) aurait évidemment été possible. Mais il aurait alors fallu faire intervenir les différences de longueurs entre éléments de $\mathcal{A}$. La décomposition par langages simples évite ces lourdeurs techniques.

Une conséquence du théorème 20 (p. 180) est qu'un langage comme $\{a^{2^n}/n \in \mathbb{N}\}$ n'est pas un langage de chaînes analogiques, puisqu'il ne possède

pas la propriété de croissance constante des longueurs. Heureusement donc, certains langages « peu naturels »[13] se trouvent rejetés hors de la portée des langages de chaînes analogiques.

5.2.3 Analyse polynomiale

Lorsque nous avons mentionné l'interprétation linguistique des langages de chaînes analogiques (p. 170), nous avons précisé que le mot réduction était à prendre au sens de réduction à une forme normale, et non pas au sens de raccourcissement des chaînes. Cependant, si $|A| < |B|$ pour tous les $A \rightarrow B$ de $\mathcal{M}$, alors, les chaînes raccourcissent bien lors de la reconnaissance, et l'on peut énoncer le théorème suivant.

THÉORÈME **21** *Soit un langage de chaînes analogiques monotone décroissant, alors, pour ce qui est du versant de la similarité, toute chaîne D est reconnue comme appartenant (ou n'appartenant pas) à ce langage en $O(|D|^2)$.*

DÉMONSTRATION : On essaie d'abord d'obtenir une formule générale avant de la spécialiser aux langages monotones. Posons $\mathcal{L} = \Lambda(\mathcal{A}, \mathcal{M})$ le langage de chaînes analogiques décroissant auquel on s'intéresse. On suppose donc dans un premier temps que $\mathcal{M}$ possède au moins un élément, mais toujours que $\forall A \rightarrow B \in \mathcal{M}, |A| < |B|$. Appelons L la longueur maximale des chaînes apparaissant dans $\mathcal{M}$, et l la longueur minimale des chaînes de $\mathcal{A}$. Supposons qu'un algorithme de résolution de l'équation analogique $B : A \doteq D : C$ d'inconnue C, équivalente à $A : B \doteq C : D$, en $O(|B| \times (|A| + |D|))$ existe[14] pour le versant de la similarité.

La reconnaissance de la chaîne D s'effectue en un certain nombre d'étapes (voir figure 5.1, p. 171). A chaque étape on tente toutes les analogies possibles $B : A \doteq D : x$ pour tous les éléments (A, B) de $\mathcal{M}$. Lors de l'étape i, l'ensemble des C_i vaut toutes les solutions obtenues à l'étape précédente. Comme la taille des C_i diminue d'au moins un symbole à chaque étape, à l'étape n, le nombre d'opérations est majoré par $L \times (L + |D| - n + 1) \times |\mathcal{M}|^n$.

Le test de reconnaissance se faisant sur les C_i de longueur égale à la longueur d'un élément de $\mathcal{A}$, et la taille de C_i décroissant strictement à chaque étape, le nombre d'étapes est majoré par $|D| - l$. Ce résultat est valable pour l'appartenance comme pour la non-appartenance, puisque, pour des $|C_i| \leq l$, on est sûr de ne plus obtenir d'éléments de $\mathcal{A}$ par raccourcissement de C.

[13] MARCUS *et al.*, *Contextual grammars versus natural languages*, 1996, p. 4.

[14] Un résultat dû à UKKONEN, *Algorithms for approximate string matching*, 1985 montre que, pour calculer la distance exacte entre deux chaînes, il est suffisant de ne calculer qu'une bande diagonale plus deux bandes supplémentaires de chacun de ses côtés dans la matrice de distance d'édition, si l'on sait à l'avance que la distance totale est inférieure à un seuil donné. La largeur de ces bandes supplémentaires croît linéairement avec le seuil. Comme nous l'avons mentionné ailleurs (LEPAGE, *Solving analogies on words: an algorithm*, 1998), ce résultat est utilisé dans l'une de nos implémentations pour réduire le coût du calcul des matrices de distance. En pratique, donc, le coût du calcul sur le versant de la similarité est strictement moins que quadratique.

En résumé, la reconnaissance s'effectue en $\Sigma_{n=1}^{|D|-l} L \times (L+|D|-n+1) \times |\mathcal{M}|^n$ opérations.

Dans le cas d'un langage monotone, $\mathcal{M} = 1$. La reconnaissance s'effectue donc en $\Sigma_{n=1}^{|D|-l} L \times (L + |D| - n + 1)$ opérations. c'est-à-dire, comme L et l sont des constantes pour le langage donné, en $O(|D|^2)$. CQFD

L'idée qui se profile derrière ce théorème est que tout langage de chaînes analogiques puisse se ramener à un langage de chaînes analogiques dont tous les modèles seraient décroissants, c'est-à-dire à un langage de chaînes analogiques décroissant (p. 173). Nous émettons donc la conjecture suivante, qui reste à prouver.

Conjecture 4 *Soit* $\Lambda(\mathcal{A}, \mathcal{M})$ *un langage de chaînes analogiques. Il existe* $\Lambda(\mathcal{A}', \mathcal{M}')$ *un langage de chaînes analogiques décroissant tel que* $\Lambda(\mathcal{A}, \mathcal{M}) = \Lambda(\mathcal{A}', \mathcal{M}')$

En pratique, dans nos expériences de traduction automatique par analogie (p. 314), nous supposerons cette conjecture car nous nous limiterons aux modèles décroissants. Ils ont en effet l'immense avantage d'assurer la terminaison de l'analyse après un seuil égal à la longueur de la chaîne d'entrée, puisque, à chaque étape, les chaînes à analyser diminuent de la longueur d'au moins un symbole.

5.3 Exemples

5.3.1 Langages $\{a_1^n a_2^n \dots a_m^n / m \geq 1, n \geq 1\}$

Nous démontrons que le langage régulier ou rationnel[15] $\{a^n\}$ est un langage de chaînes analogiques élémentaire.

LEMME **19** $\Lambda(\{a\}, \{a \to aa\}) = \{a^n / n \geq 1\}$

DÉMONSTRATION : En deux étapes, en montrant l'inclusion dans les deux sens.

Complétude: $\Lambda(\{a\}, \{a \to aa\}) \subset \{a^n / n \geq 1\}$. Rappelons que $\overline{D}$ est l'ensemble des symboles de D. Supposons que $D \in \Lambda(\{a\}, \{a \to aa\})$, ce qui est équivalent à $a \overset{*}{\vdash\!\!-} D$. Donc il existe une suite de chaînes C_1, C_2, ..., C_n telle que les analogies de la première colonne du tableau de relations suivantes sont vérifiées :

$$\begin{array}{ccccc}
a : C_1 \doteq a : aa & \Leftrightarrow & C_1 : a \doteq aa : a & \Rightarrow & \overline{C_1} \subset \overline{a} \cup \overline{aa} = \{a\} \\
C_1 : C_2 \doteq a : aa & \Leftrightarrow & C_2 : C_1 \doteq aa : a & \Rightarrow & \overline{C_2} \subset \overline{C_1} \cup \overline{aa} \\
\vdots & & \vdots & & \vdots \\
C_n : D \doteq a : aa & \Leftrightarrow & D : C_n \doteq aa : a & \Rightarrow & \overline{D} \subset \overline{C_n} \cup \overline{aa}
\end{array}$$

La deuxième colonne est l'application de l'échange des moyens à la première colonne; la troisième colonne est l'application de la propriété d'inclusion des symboles à la deuxième colonne. La troisième colonne implique $\overline{D} \subset \{a\}$, qui montre que D est du type a^n (on n'a pas de chaîne vide ici).

Consistance: $\{a^n / n \geq 1\} \subset \Lambda(\{a\}, \{a \to aa\})$. Par induction sur n. Supposons que a^{n+1} est un élément de $\Lambda(\{a\}, \{a \to aa\})$. Comme $a^{n+1} : a^n \doteq aa : a$ est une analogie vérifiée, la solution x de l'analogie $a^{n+1} : x \doteq a : aa$ est $a^n \in \{a^n / n \geq 1\}$.

Les deux inclusions précédentes donnent $\Lambda(\{a\}, \{a \to aa\}) = \{a^n / n \geq 1\}$. CQFD

Le langage hors-contexte $\{a^n b^n\}$ est aussi un langage de chaînes analogiques élémentaire.

LEMME **20** $\Lambda(\{ab\}, \{ab \to aabb\}) = \{a^n b^n / n \geq 1\}$

Pour la démonstration de ce lemme et de ceux qui suivent, nous remplaçons l'axiome d'inversion des objets par celui de concaténation d'analogies

[15] Rappelons que, pour un langage $\mathcal{L}$, on peut noter par a_n le nombre d'éléments de $\mathcal{L}$ de longueur n, et l'on peut poser :

$$\begin{array}{rl} f\colon & \mathbb{C} \to \mathbb{C} \\ & z \mapsto \sum_{n=0}^{+\infty} a_n z^n \end{array}$$

f est appelée *série génératrice* de $\mathcal{L}$. $\mathcal{L}$ est un langage rationnel si et seulement si f est rationnelle, c'est-à-dire si f est le quotient de deux polynômes.

disjointes (p. 161). Le résultat de formalisation du versant de la contiguïté, qui doit constituer une expression de l'axiome d'inversion des objets, devra aussi nécessairement impliquer la propriété de concaténation des analogies disjointes dans les cas des analogies particulières avec $a^n b^n$ ou $a^n b^n c^n$ qui nous intéressent. La raison en est que, les seules solutions admissibles aux équations analogiques $ab : a^2b^2 \doteq a^n b^n : x$ et $abc : a^2b^2c^2 \doteq a^n b^n c^n : x$ sont bien $a^{n+1}b^{n+1}$ et $a^{n+1}b^{n+1}c^{n+1}$. À l'heure actuelle, l'utilisation du seul versant de la similarité pour la résolution de ces équations analogiques provoque la production de solutions superfétatoires (et heureusement, selon nos études, n'élimine aucune solution désirée). Si l'analogie n'est bien constituée que des notions de similarité et de contiguïté, c'est donc cette dernière qui doit éliminer les solutions en trop. En fait, nous pensons que la concaténation d'analogies disjointes dans les cas particuliers cités plus haut[16] devra provenir d'une propriété plus générale qui impliquera certainement aussi que les langages de parenthèses, ou langages de Dyck, sont des langages de chaînes analogiques. Tout cela reste malheureusement de l'ordre de l'intuition, même si c'est de l'intuition forte.

DÉMONSTRATION :

Complétude: $\Lambda(\{ab\}, \{ab \rightarrow aabb\}) \subset \{a^n b^n / n \geq 1\}$. Un raisonnement similaire au précédent donne $D \in \{a^n b^n / n \geq 1\} \Rightarrow \overline{C} \subset \{a, b\}$. Par induction, par concaténation d'analogies disjointes, tous les a sont avant les b, ce qui implique $D = a^n b^m$ avec n nécessairement égal à m.

Consistance: $\{a^n b^n / n \geq 1\} \subset \Lambda(\{ab\}, \{ab \rightarrow aabb\})$. Par induction sur n. Supposons que $a^{n+1}b^{n+1}$ est un élément de $\Lambda(\{ab\}, \{ab \rightarrow aabb\})$. Comme $a^{n+1} : a^n \doteq aa : a$ et $b^{n+1} : b^n \doteq bb : b$ sont des analogies vérifiées, et par concaténation d'analogies disjointes, la solution x de l'analogie $a^{n+1}b^{n+1} : x \doteq ab : aabb$ est $a^n b^n \in \{a^n b^n / n \geq 1\}$.

Les deux inclusions précédentes donnent $\Lambda(\{ab\}, \{ab \rightarrow aabb\}) = \{a^n b^n / n \geq 1\}$. CQFD

Au tour maintenant du langage sous-contexte $\{a^n b^n c^n\}$. C'est aussi un langage de chaînes analogiques élémentaire.

LEMME **21** $\Lambda(\{abc\}, \{abc \rightarrow aabbcc\}) = \{a^n b^n c^n / n \geq 1\}$

DÉMONSTRATION : La démonstration est la même que pour $\{a^n b^n / n \geq 1\}$, en décomposant

$$a^{n+1}b^{n+1}c^{n+1} : a^n b^n c^n \doteq aabbcc : abc$$

en $a^{n+1}b^{n+1} : a^n b^n \doteq aabb : ab$ et $c^{n+1} : c^n \doteq cc : c$ qui sont toutes deux vérifiées. CQFD

[16] La conjecture de concaténation d'analogies disjointes est certainement fausse dans le cas général, mais elle est bien vérifiée dans les cas particuliers cités. Elle est en quelque sorte « de guingois » par rapport à la contiguïté. Nous ne pouvons proposer mieux, car si nous étions en mesure de proposer une conjecture qui implique l'expression recherchée sur la contiguïté, notre formalisation serait déjà complète ! Et nous ne lancerions pas dans toutes ces précautions oratoires !

Ce langage où intervient non seulement deux phénomènes que l'on pourrait considérer comme la préfixation et la suffixation, par l'ajout de symboles en début et en fin de chaînes, montre aussi l'ajout d'un symbole en milieu de chaîne. Cela est de l'ordre de l'infixation. Ce langage laisse donc pressentir que l'analogie pourrait traiter les phénomènes d'infixation de l'arabe, langue pour laquelle nous avions exprimé notre intérêt théorique (p. 48).

De façon plus générale, le même type de preuve s'applique pour démontrer le théorème suivant.

THÉORÈME **22** $\Lambda(\{a_1a_2\ldots a_m\}, \{a_1a_2\ldots a_m \rightarrow a_1^2a_2^2\ldots a_m^2\}) = \{a_1^na_2^n\ldots a_m^n \ / \ n \geq 1\}$

Dans la même veine que la remarque précédente, on voit ici, puisque le nombre de symboles différents en milieu de chaînes est libre, que l'analogie pourrait non seulement expliquer les infixations uniques mais encore les phénomènes d'infixation multiple.

Après avoir vu que tous les langages qui illustrent classiquement la classification de Chomsky-Schützenberger sont des langages de chaînes analogiques, qui plus est élémentaires et voire même paresseux (pour peu qu'on rajoute à l'ensemble des chaînes attestées le second élément du modèle), nous pouvons nous intéresser à la complexité de leur analyse.

THÉORÈME **23 (Complexité polynomiale)** *La reconnaissance de* $a_1^na_2^n\ldots a_m^n$, $n \geq 1$ *comme élément du langage de chaînes analogiques*

$$\Lambda(\{a_1a_2\ldots a_m\}, \{a_1a_2\ldots a_m \rightarrow a_1^2a_2^2\ldots a_m^2\})$$

se fait en $O(n^p)$ *pour le versant de la similarité.*

DÉMONSTRATION : L'inégalité suivante est trivialement vraie.

$$|a_1a_2\ldots a_m| = m \ < \ |a_1^2a_2^2\ldots a_m^2| = 2 \times m$$

Le modèle $a_1a_2\ldots a_m \rightarrow a_1^2a_2^2\ldots a_m^2$ est donc décroissant. Comme c'est le seul modèle du langage $\Lambda(\{a_1a_2\ldots a_m\}, \{a_1a_2\ldots a_m \rightarrow a_1^2a_2^2\ldots a_m^2\})$, ce langage est donc, par définition, un langage monotone décroissant. On a donc le résultat par application directe du théorème 21 (p. 181). CQFD

Comme résultat remarquable, nous venons d'obtenir un même comportement asymptotique :

- pour tout langage sous-contexte de la forme générale $a_1^na_2^n\ldots a_m^n$;
- pour le langage hors-contexte a^nb^n ;
- pour le langage régulier a^n.

Nous considérons comme un bon point du point de vue linguistique que la complexité d'analyse des langages précédents lorsque vus en tant que langages de chaînes analogiques, soit la même. La raison en est que l'on ne voit pas intuitivement la nécessité d'une différence de complexité pour la reconnaissance de ces langages, si ce n'est précisément pour le nombre de symboles y intervenant. Les langages de chaînes analogiques ont ainsi l'avantage d'abolir la distinction contre nature entre les langages $\{a^n b^n / n \geq 1\}$ et $\{a^n b^n c^n / n \geq 1\}$.

Bien sûr, un comportement plus que quadratique est excessif pour l'analyse du langage régulier $\{a^n / n \geq 1\}$. Mais il faut bien voir que la même complexité est suffisante pour tous les langages du type $\{a^n / n \geq 1 \wedge p(n)\}$, avec $p(n)$ n'importe quelle proposition en n telle que être pair, impair, un multiple d'un entier donné, *etc.*

En résumé, l'analogie permet de contourner la classification de Chomsky-Schützenberger. Nous avions déjà plus ou moins annoncé cela comme étant l'une de nos préoccupations lors de notre critique de la classification de Chomsky-Schützenberger (p. 74). Il s'agit d'un exemple supplémentaire, quoique converse, d'un phénomène déjà rencontré avec les grammaires contextuelles[17], où deux langages de la même complexité lorsque produits par des grammaires chomskyennes, sont de types différents lorsque produits par un procédé différent, en l'espèce, les grammaires contextuelles.

5.3.2 Langage sous-contexte $\{a^m b^n c^m d^n / m \geq 1, n \geq 1\}$

Le langage $\{a^m b^n c^m d^n\}$ est un langage de chaînes analogiques simple.

THÉORÈME **24** $\Lambda(\{abcd\}, \{abcd \rightarrow abbcdd, abcd \rightarrow aabccd\}) = \{a^m b^n c^m d^n \ / \ n \geq 1 \wedge m \geq 1\}$

La démonstration est aisée par induction et en utilisant la concaténation d'analogies disjointes.

Ce langage est le fameux langage, au cœur des deux arguments par l'exemple contre l'hypothèse du hors-contexte des langues naturelles que nous avons vu plus haut (p. 75). Rappelons que le premier exemple est en morphologie du bambara et que le second utilise la syntaxe d'un dialecte zurichois du suisse-allemand. Voici reproduit ce que Shieber disait à la fin de son article :

> Il faut aussi garder en mémoire ce qui n'a *pas* été montré par cet argument. En prouvant qu'une grammaire de compétence du suisse-allemand ne pouvait être hors-contexte, nous n'avons pas démontré qu'il était impossible voire difficile d'analyser les langues humaines. Les constructions du néerlandais comme du suisse-allemand sont analysables en temps linéaire, et s'il n'en était pas ainsi en théorie, des contraintes de performance pourraient bien les rendre telles.[18]

[17] Voir MARCUS *et al.*, *Contextual grammars versus natural languages*, 1996, p. 12.
[18] SHIEBER, *Evidence against the context-freeness of natural language*, 1985.

Nous n'assurons pas la linéarité de la reconnaissance (pourquoi devrait-il en être ainsi, d'ailleurs ?) mais nous avons montré que si la compétence repose sur l'analogie, un langage, considéré comme langage de chaînes analogiques peut très bien être analysé pas une grammaire de compétence, sans avoir recours à quelque contrainte de performance que ce soit.

5.4 Représentativité

Le problème de la surproduction, qui constituait l'une des critiques des Générativistes envers l'analogie, n'est aucunement résolu par la formalisation que nous venons de proposer des langages formels analogiques. Le chapitre suivant traitera ce problème en proposant de le résoudre par l'utilisation d'homomorphismes entre langages de chaînes analogiques. Cependant, nous pouvons dès à présent aborder de façon formelle un problème qui est en quelque sorte réciproque, celui de la représentativité.

Notre formalisation de la représentativité s'appuie sur la définition formelle des langages de chaînes analogiques que nous avons donnée plus haut (voir pages 167 et suivantes). Elle passe aussi par certaines notions simples d'algèbre[19].

Soit un ensemble Λ de chaînes de symboles. Habituellement les chaînes de symboles forment des phrases, et Λ est donc simplement un corpus. Nous désirons représenter Λ de façon plus économique, au sens de l'analogie, par un ensemble plus petit $\mathcal{B}$. En algèbre, c'est la notion de générateur.

DÉFINITION **20 (Générateur)** *Un ensemble de chaînes de symboles $\mathcal{G}$ est un* générateur *pour Λ si et seulement si*

$$\forall D \in \Lambda,\ \exists (A,B,C) \in \mathcal{G}^3 \ /\quad A : B \doteq C : D$$

On peut reformuler la définition précédente, en faisant ressortir l'ensemble $\mathcal{G}^2$.

$$\forall D \in \Lambda,\ \exists (A,B) \in \mathcal{G}^2, \exists C \in \mathcal{G},\ /\quad A : B \doteq C : D$$

Un langage de chaînes analogiques paresseux est tel que l'ensemble des modèles est le produit cartésien de l'ensemble des chaînes attestées : $\Lambda(\mathcal{A}, \mathcal{A}^2)$ (p. 174). Dans la reformulation précédente de la définition d'un générateur, l'ensemble Λ est donc égal au Λ_1 du langage de chaînes analogiques paresseux $\Lambda(\mathcal{G}, \mathcal{G}^2)$ (p. 174). Ce lien entre générateur et Λ_1 expliquera les expériences présentées plus bas sur la représentativité et la productivité (p. 256).

De plus, l'information contenue dans $\mathcal{B}$ ne devrait pas être redondante au sens de l'analogie, c'est-à-dire qu'aucune phrase de $\mathcal{B}$ ne devrait pouvoir être obtenue par analogie avec d'autres phrases de $\mathcal{B}$. En algèbre, c'est la notion de libre.

DÉFINITION **21 (Libre)** *Un ensemble de chaînes de symboles $\mathcal{L}$ est* libre *si et seulement si*

$$\forall D \in \mathcal{L},\ \neg\ (\exists (A,B,C) \in (\mathcal{L} \setminus \{D\})^3 \ /\quad A : B \doteq C : D\ \ )$$

Pour un ensemble donné Λ de chaînes de symboles, un ensemble représentatif idéal devrait vérifier les deux définitions précédentes. Un tel ensemble est appelée une base en algèbre.

[19] Cette partie reprend partiellement LEPAGE, *Corpus contraction by sentence extraction using analogy*, 1997, p. 458–459.

DÉFINITION **22 (Base)** *Un ensemble $\mathcal{B}$ est une* base *pour Λ si et seulement si c'est un libre et un générateur pour Λ.*

Une fois posées ces définitions, nous pouvons proposer une définition rigoureuse de la représentativité.

DÉFINITION **23 (Représentativité au sens de l'analogie)** *Un ensemble de chaînes de symboles $\mathcal{B}$ est représentatif d'un autre ensemble de chaînes de symboles Λ si et seulement si c'est une base pour Λ.*

Les générateurs (*resp.* les libres) peuvent fort bien être construits à partir de chaînes en dehors de Λ. Nous nous restreignons cependant à des ensembles contruits à partir de chaînes de Λ seulement, c'est-à-dire à des sous-ensembles de Λ. Avec cette restriction, nous rappelons les définitions suivantes.

DÉFINITION **24 (Générateur minimal)** *Un générateur $\mathcal{G}$ pour Λ est un* générateur minimal *si et seulement si*

$$\forall D \in \mathcal{G},\ \mathcal{G} \setminus \{D\} \textit{ n'est pas générateur pour } \Lambda.$$

DÉFINITION **25 (Libre maximal)** *Un libre $\mathcal{L}$ pour Λ est un* libre maximal *si et seulement si*

$$\forall D \in \Lambda \setminus \mathcal{L},\ \mathcal{L} \cup \{D\} \textit{ n'est pas libre.}$$

Toutes les notions et définitions introduites ici nous seront utiles lors de nos expériences sur des corpus. En effet, à partir d'un corpus donné, nous essaierons pratiquement de trouver l'ensemble des phrases représentatives par des procédures automatiques fondées sur l'analogie. Du point de vue morphologique, ces notions et définitions peuvent contribuer à la formalisation de phénomènes bien connus des gens apprenant les langues étrangères. Par exemple, pourquoi est-il nécessaire et suffisant de ne connaître que cinq formes d'un verbe latin pour le conjuguer entièrement ? La réponse passe par la notion de base, libre et générateur. Ce résultat pourrait être retrouvé de façon automatique grâce aux notions et aux notations introduites plus haut. La conjugaison entière de deux verbes semblables constitue un Λ_0 dont une base, c'est-à-dire un générateur et libre, est la conjugaison entière de l'un de ces verbes, plus cinq seulement des formes du second verbe. Ce résultat est purement formel, il ne fait aucunement intervenir le sens des verbes, ni aucune autre interprétation de ces verbes. La partie que nous allons maintenant aborder va nous permettre de répondre à la critique éventuelle qui pourrait nous être adressée de rester encore et toujours au seul niveau des symboles.

Chapitre 6

Homomorphismes entre structures analogiques

Dans cette partie, nous allons revenir sur la notion de métaphore que nous avions soigneusement isolée de celle d'analogie. Rappelons que l'on confond souvent improprement des phrases comme « l'atome est comme un système solaire » qui sont des métaphores avec des phrases du genre « les électrons sont au noyau atomique comme les planètes sont au soleil » qui sont bien des analogies. Nous avons vu que les phrases du premier type entrent bien dans la description de la métaphore par Aristote et constituent la quatrième espèce de métaphore selon lui (voir p. 40). Si elles méritent le nom d'analogies vulgaires selon Deleuze (voir p. 49), c'est parce qu'elles sont sous-tendues par de vraies analogies. Pour les phrases données en exemples ci-dessus, la première sous-tend bien la seconde. Quoique notre sujet principal soit l'analogie, nous allons ici nous intéresser à cette espèce particulière de métaphore. Plus exactement, nous allons proposer une formalisation d'un phénomène plus restreint. Cela vient de ce que notre formalisation de l'analogie a été restreinte aux analogies entre objets de même domaine.

Bien que cette formalisation soit restreinte, elle a une généralité qui lui permet de couvrir un nombre non négligeable d'applications possibles au traitement automatique des langues. Rappelons ici ce mot attribué à Le Lionnais à propos des mathématiques :

> En tendant vers un degré nul d'application, elles tendent vers un degré infini d'applicabilité.

Nous n'avons nous-même appliqué cette formalisation qu'à un certain nombre de tâches du traitement automatique des langues qui nous intéressaient plus particulièrement, l'analyse structurale et la traduction automatique.

6.1 Rappel des modèles précédents

Revenons sur les différents modèles proposés dans le cadre de l'informatique pour la formalisation de la métaphore et de l'analogie (p. 83 et suivantes). Dans un premier temps, nous refaisons la liste des critiques que nous avons adressées à toutes ces formalisations. Dans un deuxième temps, nous soulignons leurs aspects positifs. Il est bien évident que nous désirons conserver ceux-ci dans nos travaux.

Commençons avec le modèle de Gentner (p. 84). On y voyait apparaître des relations entre domaines. Ces relations étaient dues à des modélisations de domaines. Nous avons dit que nous refusions cette approche. Dans les formalisations de l'analogie que nous venons de proposer (p. 125 et suivantes), nous sommes parti d'un nombre restreint de postulats. En définitive, dans le domaine des chaînes de symboles, la seule opération que nous avons autorisée est celle de la comparaison de deux symboles au sens de l'égalité. Notre vue de la modélisation est donc minimale. Du modèle de Gentner, nous retenons cependant deux points positifs. Premièrement, à partir de certaines correspondances entre données des deux domaines, on en établit d'autres par calcul. Pour nous, le seul calcul toléré sera un calcul fondé sur l'analogie. Deuxièmement, chez Gentner, la qualité des correspondances peut être mesurée. Chez Gentner et ses continuateurs en intelligence artificielle, toutes sortes de critères sont utilisés. Pour nous, le premier des critères est celui de l'existence de relations d'analogies. Nous y reviendrons plus bas. En fait plus que d'un critère, il s'agit pour nous de la condition nécessaire d'établissement de la correspondance. Mais encore plus, de même que nous l'avions vu en linguistique à propos des recherches sur les effets diachroniques de l'analogie, la fréquence avec laquelle les correspondances sont établies devra être un critère de force. Dans nos expériences sur l'analyse structurale (p. 303) et en traduction automatique (p. 316), nous aurons l'occasion de montrer que la qualité des analyses ou des traductions obtenues par notre méthode d'homomorphisme est sans doute intimement liée à la fréquence avec laquelle une correspondance nouvelle est établie.

Le point positif du modèle de Nagao (p. 89) était que quatre objets étaient bien présents. Cela est nécessaire pour pouvoir commencer à parler d'analogies. Malheureusement, ils chevauchaient deux domaines. Aussi, la vérification de l'égalité des proportions semblait difficile à établir directement. Un autre point négatif était que, pratiquement, à cause de la restriction à l'application d'une seule transformation à la fois sur les chaînes de symboles, la méthode était limitée. Le point positif que nous voulons sauvegarder était que l'on y voyait nettement apparaître un sorte de parallélogramme analogique.

Le modèle d'Yvon (p. 90), baptisé méthode par cascades, offrait une réponse à la limitation de la méthode consécutive aux transformations uniques sur les chaînes. Par application récursive de la méthode, on pouvait espérer obtenir plusieurs transformations. Comme avec le modèle de Gentner, le but était l'établissement de correspondances entre deux domaines, dont la qualité fût mesurable, Nous en retenons une autre particularité fondamentalement

positive: les correspondances sont jugées meilleures si elles sont sous-tendues par des carrés paradigmatiques, et d'autant meilleures que le nombre de carrés paradigmatiques est grand. Après tout notre effort de formalisation, nous ne pouvons nous empêcher de voir dans les carrés paradigmatiques de simples analogies entre chaînes de symboles du même domaine. Mais rappelons que nous refusons toute modélisation de domaine et tout recours à des bases de connaissances contrairement à ce modèle. Dès lors, pour nous, l'existence de relations analogiques à l'intérieur d'un même domaine devient la condition *sine qua non* de l'établissement d'une correspondance. Pas d'analogie dans le domaine de départ, pas de correspondance possible avec le second domaine. Poussons plus loin l'idée d'Yvon en redisant que nous ne sommes capables de traiter que des analogies entre objets d'un même domaine. Nous disposons ici de deux domaines. Pourquoi ne pas voir aussi des analogies dans le second domaine ? Si, comme dans le cas du problème de la transcription graphémique-phonémique le second domaine est aussi un domaine de chaînes de symboles, alors cette idée en est d'autant plus naturelle pour nous.

6.2 Vue statique

Nous venons de mettre en place tous les éléments nécessaires à notre vue statique d'un certain type de métaphore, celui restreint au cas où elle est sous-tendue par une analogie. Rappelons qu'une métaphore est un transfert d'un domaine à un autre. Précisons aussi tout de suite que notre modèle sera moins large que celui de Gentner ou que la définition d'Aristote. Mais nous prétendons qu'elle repose sur des bases plus formelles. De plus, elle possède un avantage éminent, puisqu'elle peut faire l'objet de calculs avec toutes les bonnes restrictions que nous nous étions déjà imposées lorsque nous avions écarté les approches de l'intelligence artificielle : aucune connaissance externe ni modélisation des domaines. En fait, pour être exact, il vaut mieux dire que la modélisation que nous imposons aux domaines est minimale. Toutes les applications que nous donnerons pour illustrer notre propos n'utiliseront que des domaines de chaînes de symboles où la seule opération élémentaire autorisée se réduit au test de l'égalité entre deux symboles.

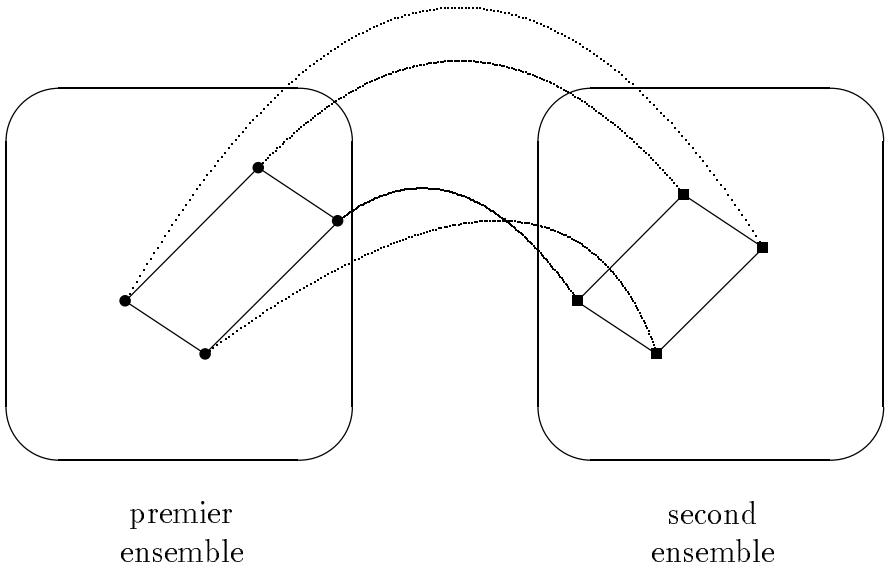


Figure 6.1: Vue statique des homomorphismes

Bases Le tableau que nous proposons est donc le suivant. On dispose de deux domaines. Les seules correspondances possibles entre les deux domaines sont celles qui existent entre des objets vérifiant des analogies dans les deux domaines. La figure 6.1 montre une telle correspondance. Analysons plus finement. Il existe tout d'abord les relations d'analogies à l'intérieur des domaines. Cela nous permet de reprendre à notre compte ce que proposait Yvon, mais nous avons été plus loin en imposant des analogies dans les deux domaines. En

plus, notre définition de l'analogie rend compte de plusieurs modifications à la fois. Elle est donc immédiatement plus puissante. Les deux analogies qui sous-tendent cette vue des choses sont donc les bases de notre figure (figure 6.2).

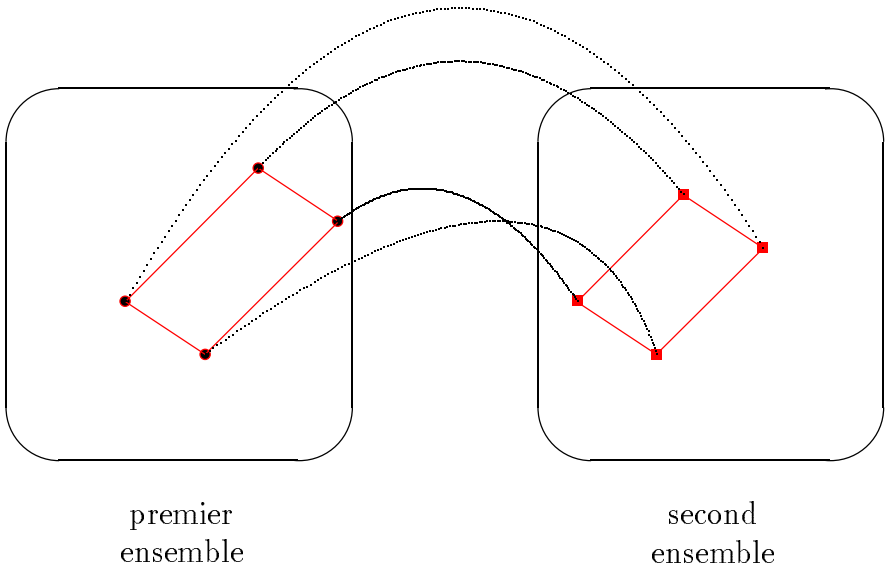


Figure 6.2: Bases d'une analogie conservée par homomorphisme entre structures analogiques

Examinons maintenant les différentes relations qui chevauchent les deux domaines. Dans ce schéma, on peut en distinguer trois sortes.

Volume La première est globale. Elle donne la correspondance entre des analogies des deux domaines (une seule dans la figure 6.1). Ce type de relation est de l'ordre du volume. Il est visualisé par une sorte de parallélépipède difforme dans la figure 6.3 (p. 196). On pourrait aussi dire que nous nous plaçons dans une vue où les analogies sont conservées d'un domaine à l'autre. Cette remarque justifie l'usage que nous faisons du terme d'homomorphisme. Un homomorphisme est une application conservant la structure. Ici, la structure est celle induite par les analogies présentes dans un domaine.

Arêtes La seconde sorte de relations est celle qui existe entre objets de deux domaines. La force de ces correspondances locales entre objets est commandée par l'existence de la correspondance globale entre analogies. Brisez l'analogie, vous brisez certaines de ces correspondances. Cela nous permet de reprendre l'idée des tenants de l'intelligence artificielle selon laquelle on devrait être capable de qualifier ou de quantifier la force des transferts de domaine à domaine. Ici, nous somme carrément restrictifs en proclamant que les seules relations à

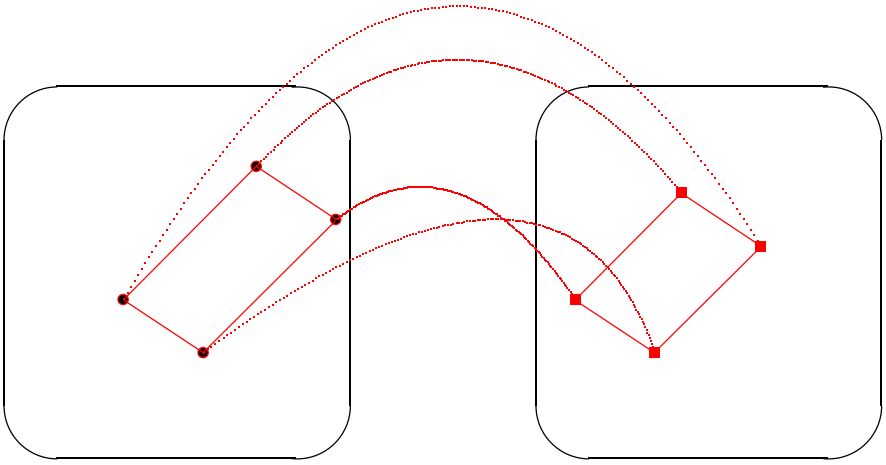


Figure 6.3: Vue statique des homomorphismes. Le parallélépipède déformé en entier établit la correspondance

retenir doivent être fortes car inscrites dans une relation globale soutenue par deux analogies. Graphiquement, les correspondances locales sont de l'ordre de la droite ou de la courbe, puisque ce sont les arêtes des parallélépipèdes difformes qui chevauchent les domaines. Il y en a quatre par analogie conservée d'un domaine à l'autre (voir la figure 6.4, p. 197).

Faces Enfin, il existe des correspondances entre des couples des deux domaines. Nous retrouvons ici ce que nous trouvions de positif dans le modèle de Nagao. En effet, ce sont, dans ce modèle-là, des analogies qui chevauchent les domaines. Si le premier domaine était le français avec les deux phrases (*il est jeune* et *il n'est pas jeune*) et si le second domaine était le japonais avec les phrases 若い et 若くない, on retrouverait ici le type d'analogies qui fondait cette idée de la traduction automatique:

$$\textit{il est jeune} : \textit{il n'est pas jeune} \doteq \text{若い} : \text{若くない}$$

Nous avions dit que nous ne savions pas calculer ni vérifier de telles analogies directement. Dans le schéma que nous proposons ici, nous sommes maintenant en mesure de les établir automatiquement, mais indirectement. Cette sorte de relations entre couples des deux domaines est de l'ordre de la surface, puisqu'il s'agit des côtés du parallélépipède chevauchant les deux domaines. Il en existe six par analogie conservée entre domaines. En effet, on a non seulement les quatre faces du parallélépipède difforme mais aussi les deux

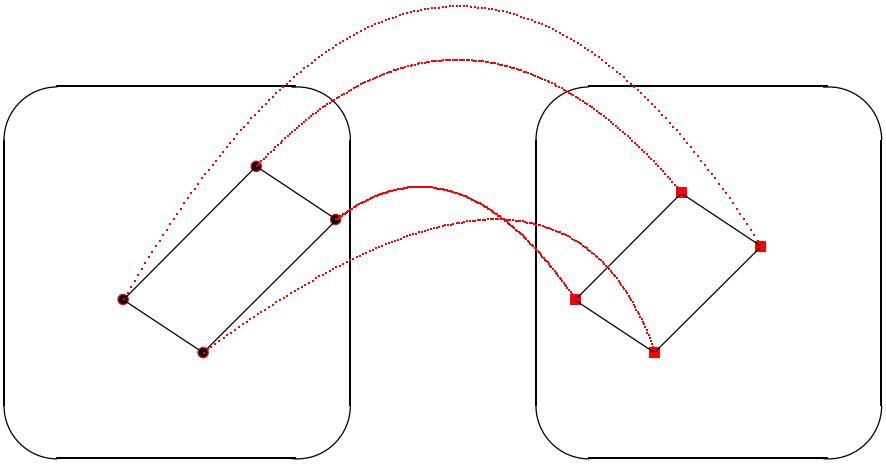


Figure 6.4: Arêtes dans une analogie conservée par homomorphisme entre structures analogiques

rectangles difformes en diagonale. La figure 6.5 (p. 198) n'en montre qu'une seule.

Au total, nous retrouvons un schéma tracé dans notre introduction (voir p. 31). Ce n'est pas par hasard que nous avions reproduit ce schéma trouvé dans un livre de philosophie[1]. De la même façon que dans notre vue statique de la métaphore restreinte sous-tendue par l'analogie et que nous venons de présenter, de même, dans ce schéma existait deux domaines, le goût individuel ou l'apparence extérieure, et le jugement social, entre lesquels deux analogies préexistantes étaient mises en correspondance. En réalité, dans la pensée de l'auteur, ce schéma n'impose pas de partage entre premier et second domaine. Puisque l'homomorphisme joue donc dans un seul ensemble, on voit ici se profiler la notion d'isomorphisme dans une structure analogique.

goût individuel : 意気 : 渋味 ≒ 甘味 : 野暮

↑ ↑ ↑ ↑

jugement social : 上品 : 地味 ≒ 派手 : 下品

Redisons que la présentation de la métaphore que nous venons de faire est limitée. Premièrement, il ne s'agit que de la quatrième espèce de métaphores selon Aristote, celles sous-tendues par l'analogie. Deuxièmement, nous nous

[1] 九鬼 周造 (Kuki Syuuzou), 「いき」の構造, 1999 1e ed 1930, p. 44.

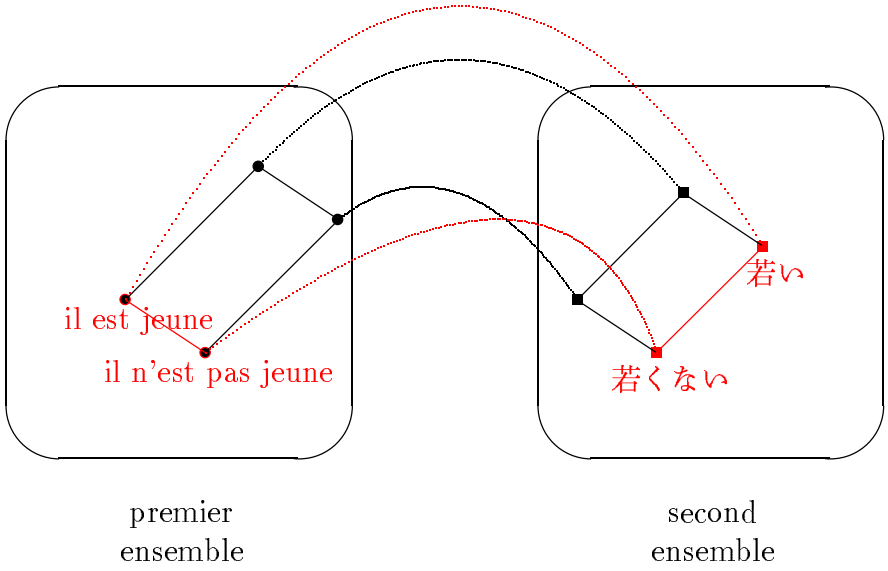


Figure 6.5: Face dans une analogie conservée par homomorphisme entre structures analogiques. Exemple de la traduction automatique

sommes restreint au cas où il existe des analogies dans chacun des deux domaines en présence. Bien que cette vue puisse donc sembler réductrice, elle autorise cependant bien des calculs. C'est précisément l'objet de la partie suivante de montrer que cette vue statique fonde en fait une vision dynamique des choses.

6.3 Définitions

Nous formalisons maintenant simplement ce que nous venons de dire de notre modèle restreint de la métaphore. Mathématiquement, la formalisation repose sur la notion d'homomorphisme. Afin de préparer une expression simple des définitions que nous allons donner, il nous faut rappeler d'abord quelques notions sur les extensions naturelles des fonctions. En effet, nous allons un peu tirer sur la définition d'homomorphisme. Rigoureusement, un homomorphisme est une *application* d'un ensemble muni d'une loi de composition interne dans une autre ensemble muni d'une autre loi de composition interne. L'application α que nous avions vue pour définir les structures analogiques (p. 118) existe entre un ensemble et l'ensemble des parties d'un autre ensemble. Elle ne saurait donc être considérée comme une loi de composition interne.

La première extension naturelle des fonctions que nous allons rappeler permet précisément de passer à l'ensemble des parties de l'ensemble de départ.

DÉFINITION **26** *Soit* h *une fonction d'un ensemble* $\mathcal{E}$ *dans un autre* $\widehat{\mathcal{E}}$. *On étend* h *à l'ensemble* $\wp(\mathcal{E})$ *des parties de* $\mathcal{E}$ *de la façon suivante.*

$$\forall \mathcal{A} \in \wp(\mathcal{E}),\ \mathrm{h}(\mathcal{A}) = \{\ \mathrm{h}(A)\ /\ A \in \mathcal{A}\} = \bigcup_{A \in \mathcal{A}} \{\mathrm{h}(A)\}$$

La seconde extension porte sur les équations analogiques, c'est-à-dire sur l'application α. Rappelons que cette application avait comme ensemble de départ les triplets d'un ensemble quelconque $\mathcal{E}$ et comme ensemble d'arrivée l'ensemble des parties de $\mathcal{E}$, noté $\wp(\mathcal{E})$ (voir p. 118). On étend cette application de la façon suivante. En fait, formellement, cette deuxième extension peut se dériver de la première par composition de fonctions.

DÉFINITION **27** *L'extension de l'application* α *de* $\mathcal{E}^3$ *dans* $\wp(\mathcal{E})$ *est définie de la façon suivante.*

$$\forall (\mathcal{A},\mathcal{B},\mathcal{C}) \in \wp(\mathcal{E})^3,\ \alpha(\mathcal{A},\mathcal{B},\mathcal{C}) = \bigcup_{(A,B,C) \in \mathcal{A}\times\mathcal{B}\times\mathcal{C}} \alpha(A,B,C)$$

Avec ces deux extensions, on peut poser de façon simple la définition suivante.

DÉFINITION **28** *Soient* $(\mathcal{E},\alpha)$ *et* $(\widehat{\mathcal{E}},\alpha)$ *deux structures analogiques induites sur deux ensembles quelconques* $\mathcal{E}$ *et* $\widehat{\mathcal{E}}$. *Une fonction* h *de* $(\mathcal{E},\alpha)$ *dans* $(\widehat{\mathcal{E}},\alpha)$ *est un homomorphisme si et seulement si*

$$\forall (A,B,C) \in \mathcal{E}^3,\ \mathrm{h}(\alpha(A,B,C)) = \alpha(\mathrm{h}(A),\mathrm{h}(B),\mathrm{h}(C))$$

Dans cette définition, la première occurrence de h se réfère à son extension à l'ensemble des parties $\wp(\mathcal{E})$ de l'ensemble de départ $\mathcal{E}$. La seconde occurrence de α est l'extension à $\wp(\mathcal{E})^3$. On peut naturellement appliquer cette définition au cas où les deux ensembles $\mathcal{E}$ et $\widehat{\mathcal{E}}$ sont en fait des monoïdes libres $\mathcal{V}^*$ et $\widehat{\mathcal{V}}^*$ construits sur deux ensembles finis de symboles $\mathcal{V}$ et $\widehat{\mathcal{V}}$.

Cette définition correspond bien à la description de notre modèle restreint de la métaphore. Il impose la conservation des analogies d'un domaine à l'autre, c'est-à-dire d'une structure analogique à l'autre. En fait, ce qui sera important dans nos applications, ce ne sera plus les deux structures analogiques nécessaires à la fondation des homomorphismes, mais bien les correspondances établies par les homomorphismes. Car la vue statique des homomorphismes sous-tend une vision dynamique que nous mettrons justement en œuvre dans nos applications. Nous allons la décrire maintenant.

6.4 Vision dynamique

On part d'une situation où deux domaines différents coexistent. Ils sont baptisés domaines de départ et d'arrivée. On y a délimité deux ensembles. Le premier, dans le domaine de départ, est appelé ensemble de départ. Le second, dans le domaine d'arrivée, est l'ensemble d'arrivée. Des correspondances sont données à l'avance entre ces deux ensembles (figure 6.6, p. 202). La forme que prend globalement la correspondance est libre. Ce n'est surtout pas nécessairement une bijection.

En entrée, on propose un nouvel élément dans le domaine de départ, en dehors de l'ensemble de départ. Cet élément est appelé nouvelle donnée. Il est représenté en rouge dans la figure 6.7 (p. 202). Le but est de trouver, dans le deuxième domaine, que ce soit dans l'ensemble d'arrivée ou en dehors, les ou des éléments correspondants à cette nouvelle donnée. Le nombre d'éléments du domaine d'arrivée en correspondance avec la nouvelle donnée est libre. Il peut en exister zéro, un ou plusieurs.

Les domaines sont évidemment structurés par l'analogie. Aussi, avec la nouvelle donnée, on va essayer de former des analogies avec des éléments de l'ensemble de départ. Cela est possible si on dispose d'une formalisation des analogies dans le domaine de départ. Dans nos applications au traitement automatique des langues, nous considérons des chaînes des symboles, et nous disposons (presque) d'un algorithme de vérification. Par commodité, nous n'avons représenté qu'une seule telle analogie dans la figure 6.8 (p. 203). Mais il peut en exister plusieurs. Il peut aussi n'en exister aucune. Dans ce cas, le processus s'arrête là. On conclut qu'il n'existe aucun élément du second domaine correspondant à la donnée nouvelle.

Dans le cas où une ou plusieurs analogies peuvent être formées dans l'ensemble de départ, on passe à l'ensemble d'arrivée en suivant simplement les correspondances. Un élément de l'ensemble de départ peut avoir zéro, un ou plusieurs correspondants dans l'ensemble d'arrivée. Lors de cette étape donc, si un élément au moins de l'ensemble de départ apparaissant dans une analogie n'a pas de correspondant, l'analogie en question est perdue. Si plusieurs éléments de l'ensemble d'arrivée correspondent à un élément de l'ensemble de départ apparaissant dans une analogie, on forme autant de triplets d'éléments de l'ensemble d'arrivée que possible. Cette combinatoire correspond aux extensions naturelles de la notion de fonction aux parties des ensembles que nous venons de voir dans notre formalisation des homomorphismes analogiques (définition 26, p. 198). En résumé, à partir d'une seule relation d'analogie entre la nouvelle donnée et des éléments de l'ensemble de départ, on forme donc zéro, un ou plusieurs triplets avec des éléments de l'ensemble d'arrivée. Et globalement, à partir de toutes les correspondances possibles, on forme zéro, un ou plusieurs tels triplets, dont certains peuvent même apparaître plusieurs fois selon les possibles configurations. Par commodité, nous n'avons représenté qu'un seul tel triplet d'éléments de l'ensemble d'arrivée dans la figure 6.9 (p. 203). Or, nous avons vu que la donnée d'un triplet équivaut, du point de vue de l'analogie, à la donnée d'une équation analogique. C'est

donc zéro, une ou plusieurs équations analogiques qui viennent d'être formées dans le domaine d'arrivée. Si aucune équation analogique n'a été formée, le processus échoue évidemment à cette étape.

Dans le cas contraire, et si on dispose d'une formalisation des analogies dans le second domaine, l'étape suivante est de résoudre ces équations analogiques. Dans nos applications au traitement automatique des langues, puisque nous nous plaçons dans des domaines de chaînes des symboles, nous disposons (presque) d'un algorithme de résolution. Lors de ces résolutions, pour chaque équation analogique, zéro, une ou plusieurs solutions peuvent être obtenues. Globalement donc, sur toutes les équations analogiques à cette étape, zéro, une ou plusieurs solutions sont obtenues. Dans le cas où aucune solution n'est obtenue, le processus s'arrête ici. Dans le cas contraire, si plusieurs solutions sont obtenues, certaines peuvent l'être plusieurs fois selon les configurations possibles. Nous retrouvons donc ici, au final, la notion de fréquence d'apparition. Les solutions obtenues peuvent appartenir à l'ensemble d'arrivée mais elles peuvent aussi être en dehors de cet ensemble. Dans la figure 6.10 (p. 204), de nouveau, nous n'avons fait apparaître qu'une seule solution, en dehors de l'ensemble d'arrivée.

Naturellement, puisque c'était le but de l'opération, les solutions obtenues sont considérées comme correspondant à la donnée nouvelle de départ. On établit donc les correspondances reliant ces solutions à la donnée nouvelle (figure 6.11, p. 204).

La dernière étape, facultative, s'apparente à un apprentissage. En cas de succès, c'est à dire, si des éléments du domaine final ont pu être mis en correspondance avec la nouvelle donnée, on peut faire trois ajouts. Le premier est l'ajout de la donnée nouvelle à l'ensemble de départ, ou plutôt, l'accroissement de l'ensemble de départ par la donnée nouvelle. Les éléments du domaine d'arrivée créés par les résolutions des multiples équations analogiques peuvent eux aussi être ajoutées à l'ensemble d'arrivée, pour ceux qui se trouvaient en dehors. Enfin, les correspondances obtenues peuvent elles aussi être ajoutées aux correspondances données. Nous avons fait figurer cette étape en 6.12 (p. 205), où la nouvelle donnée, son correspondant dans l'ensemble d'arrivée et la correspondance elle-même apparaissent de la même couleur que les données initiales. Plus haut, nous avons mentionné les fréquences avec lesquelles les nouvelles correspondances sont établies. Dans une application informatique, ces fréquences peuvent éventuellement être utilisées ultérieurement pour peser dans le calcul de nouvelles correspondances. Nous n'avons pas vraiment exploré ces possibilités, mais nos recherches s'orientent dans cette direction en parallèle avec le problème du traçage d'un tel processus.

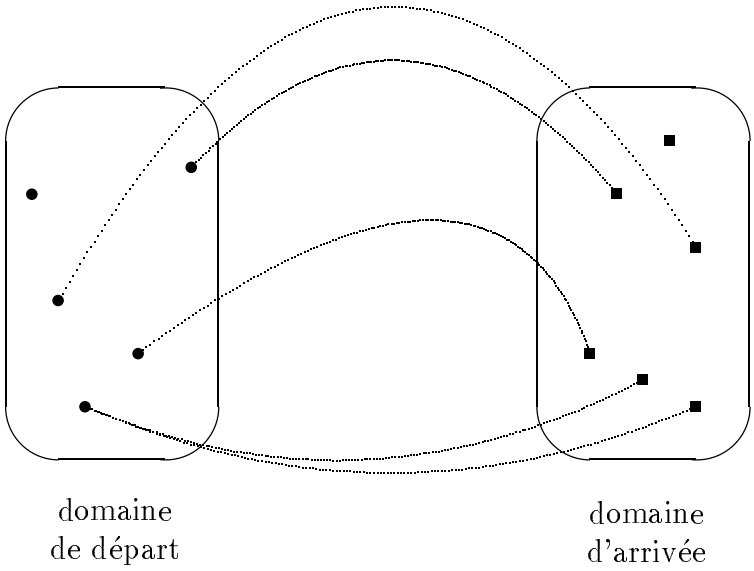


Figure 6.6: Correspondances données à la base entre éléments de l'ensemble de départ et éléments de l'ensemble d'arrivée

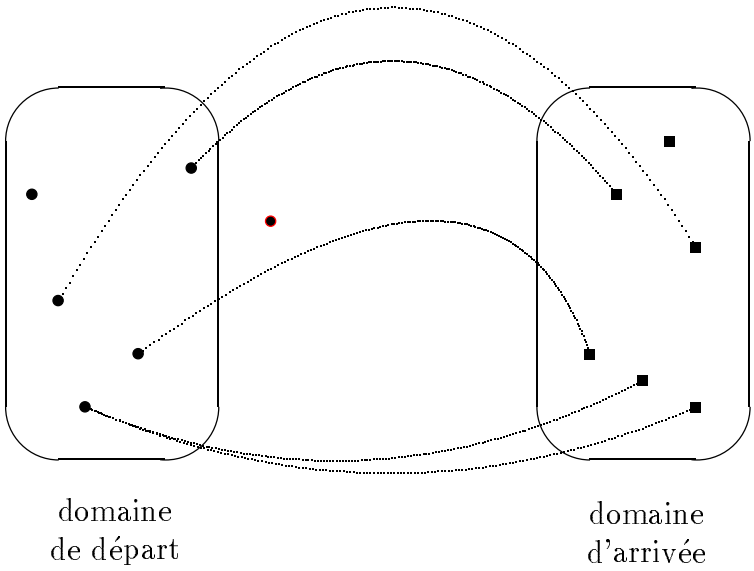


Figure 6.7: Donnée d'un nouvel élément homogène au domaine de départ

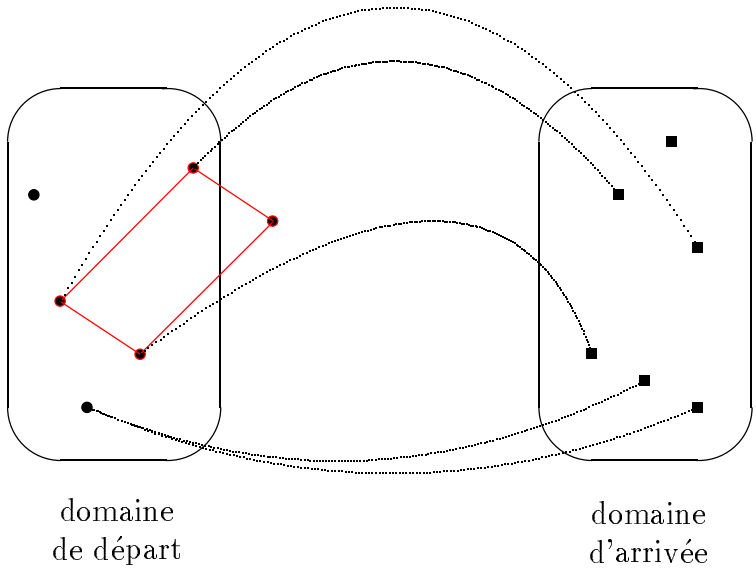


Figure 6.8: Vérification d'analogies dans le domaine de départ

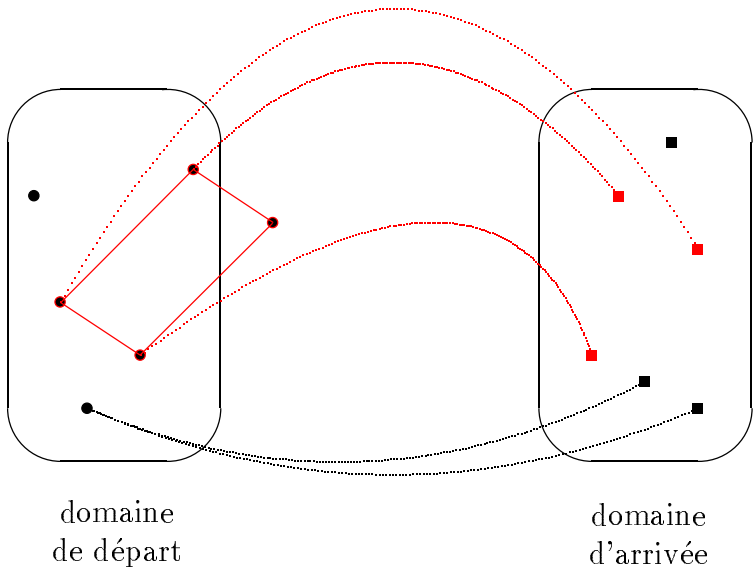


Figure 6.9: Suivi des correspondances entre domaines de départ et d'arrivée

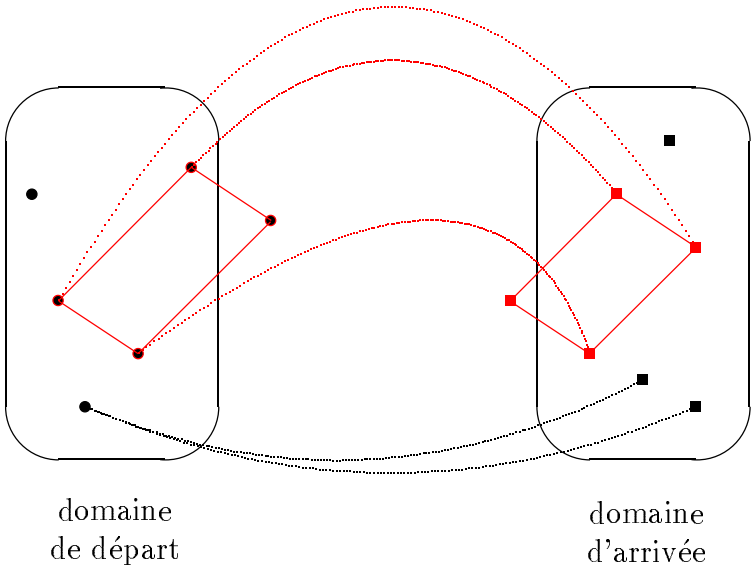


Figure 6.10: Création de nouveaux éléments homogènes au domaine d'arrivée par résolution d'équations analogiques dans ce domaine

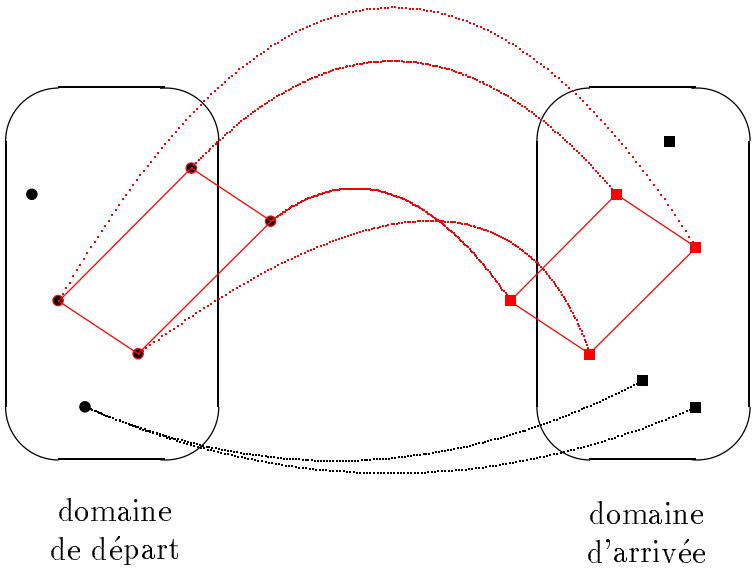


Figure 6.11: Établissement de nouvelles correspondances

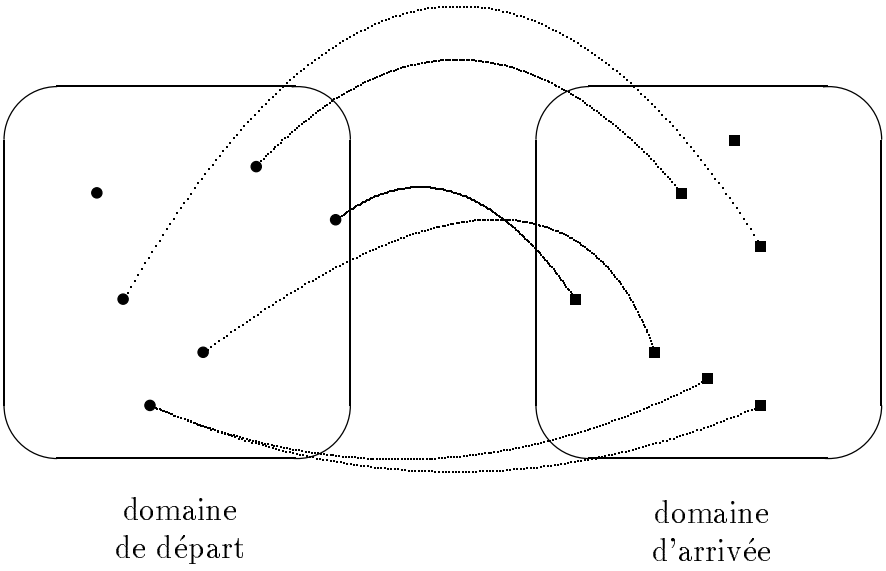


Figure 6.12: Extension des ensembles dans les deux domaines et affectation éventuelle des fréquences aux nouvelles correspondances

Partie III

Algorithmes

Chapitre 7

Résolution d'équations analogiques

Les algorithmes que nous présentons ici sont des algorithmes de résolution d'équations analogiques. Rappelons qu'une équation analogique est entièrement définie par la donnée de trois objets. Les algorithmes suivants prennent donc tous en entrées trois objets d'un même type. Une analogie, elle, est entièrement caractérisée par la donnée de quatre objets. Dans ce cas, donc, un objet supplémentaire apparaît par rapport au cas de l'équation analogique. On peut envisager de réaliser simplement la vérification d'analogies à partir des mêmes algorithmes que ceux de la résolution. Deux façons sont envisageables. La première, triviale et coûteuse en temps, consiste simplement à résoudre l'équation analogique correspondante et à vérifier que la quatrième donnée de l'analogie à vérifier est bien produite parmi les résultats. La seconde consiste à guider pas-à-pas les algorithmes que nous allons décrire maintenant par cette même quatrième donnée. Cette technique n'étant pas très difficile à mettre en œuvre, nous ne détaillerons pas les algorithmes que l'on obtient ainsi.

Nous allons examiner tour à tour les différents types d'objets pour lesquels nous avons proposé ou tenté de proposer une formalisation de l'analogie. Nous verrons donc d'abord le cas des ensembles, puis celui des multi-ensembles, et enfin celui des chaînes de symboles.

7.1 Ensembles finis

L'application directe du théorème 3 (p. 127) sur la résolution des équations analogiques entre ensembles permet d'écrire l'algorithme 7.1.

Figure 7.1: Résolution d'équations analogiques entre ensembles finis

Entrée : A, B, C, trois ensembles finis

Sortie : D, un ensemble tel que $A : B \doteq C : D$

1: si $B \cap C \subset A$ *et* $A \subset B \cup C$ alors
2: $\quad D \leftarrow (B \cap C) \cup (B \cup C \setminus A)$
3: $\quad$ retourner D
4: sinon
5: $\quad$ écrire « pas de solution »
6: fin si

Théoriquement, on peut représenter la réunion des trois ensembles A, B et C par une chaîne de bits de longueur $|A \cup B \cup C|$. Les trois ensembles A, B et C sont alors représentés chacun par une chaîne de bits de même longueur, où seuls les bits correspondant aux éléments de l'ensemble sont positionnés à 1. Avec cette représentation, n'importe quelle opération ensembliste, que ce soit l'union, l'intersection, la complémentation ou le test d'égalité, est trivialement réalisée en un temps linéaire en ce nombre de bits grâce aux opérations logiques ou, et, non et égale. L'algorithme de résolution d'analogie sur les ensembles est donc linéaire en le nombre d'éléments de $A \cup B \cup C$.

Une remarque. Le genre de résultat que nous venons d'énoncer ne prend jamais en compte le fait que deux étapes pour l'interfaçage avec l'être humain sont pourtant nécessaires. Si les ensembles A, B et C sont donnés sous formes de suites de symboles, une étape, légèrement plus coûteuse, en $|A| + |B| + |C|$ sera nécessaire pour leur codage sous forme de chaînes de bits. En revanche, le décodage de la solution, si elle existe, ne nécessitera que $|A \cup B \cup C|$ opérations.

En conclusion, on peut énoncer que l'algorithme de résolution des équations analogiques entre ensembles est linéaire en la somme de la taille des ensembles donnés. Pour être plus précis, et sans même faire de jeu de mots, en la taille de la donnée des ensembles.

7.2 Multi-ensembles finis

Un multi-ensemble est une application d'un ensemble dans $\mathbb{N}^*$. Il peut donc se décrire comme un ensemble de couples dont les premiers éléments appartiennent à l'ensemble en question, et dont les seconds éléments appartiennent à $\mathbb{N}^*$. Nous nous plaçons dans le cas où le premier ensemble est un ensemble de symboles contenu dans un alphabet fini. Cela implique que les multi-ensembles formés sont finis. De plus, dans un tel cas, du point de vue algorithmique, un multi-ensemble peut être simplement représenté par un tableau indicé sur l'alphabet. Les valeurs dans ce tableau sont données par le multi-ensemble si le symbole considéré y apparaît. Dans le cas contraire, la valeur est de 0. La résolution d'une équation analogique entre multi-ensembles se réduit donc au parcours en parallèle de trois tableaux indicés sur l'alphabet. Au cours de ce parcours, on vérifie que la valeur calculée pour la solution appartient bien à $\mathbb{N}^*$ ou est égale à 0. Autrement dit, on vérifie simplement si elle est dans $\mathbb{N}$. Le calcul des ces valeurs est effectué selon la formule générale de l'égalité de la somme des moyens et des extrêmes. Tout cela nous permet d'écrire l'algorithme 7.2.

Figure 7.2: Résolution d'équations analogiques entre multi-ensembles finis

Entrée : A, B, C, trois multi-ensembles finis,
c'est-à-dire trois tableaux à valeurs dans $\mathbb{N}$
indicés par un alphabet $\mathcal{V}$.

Sortie : D, un multi-ensemble tel que $A : B \doteq C : D$

1: pour chaque $a \in \mathcal{V}$ faire
2: $\quad D[a] \leftarrow B[a] + C[a] - A[a]$
3: $\quad$ si $D[a] < 0$ alors
4: $\quad\quad$ écrire « pas de solutions »
5: $\quad\quad$ quitter la fonction
6: $\quad$ fin si
7: fin pour
8: retourner D

Évidemment, de façon à accélérer l'exécution d'un tel algorithme, il est conseillé d'avoir recours à toute technique permettant de ne parcourir que les symboles appartenant à l'union des trois ensembles sous-jacents aux multi-ensembles A, B et C.

Il est intéressant de noter que l'algorithme précédent permet de reformuler l'algorithme de résolution d'équations analogiques entre ensembles proposé précédemment. Nous avons déjà vu (p. 136) qu'un ensemble n'est rien d'autre qu'un multi-ensemble restreint au cas où la seule valeur autorisée dans $\mathbb{N}^*$ est

1. Les seules valeurs autorisées pour les éléments de D sont donc aussi 0 et 1. Dès lors, une nouvelle forme de l'algorithme sur les ensembles peut utiliser l'algorithme sur les multi-ensembles. Elle constitue l'algorithme 7.3.

Figure 7.3: Résolution d'équations analogiques entre ensembles finis. Version par les multi-ensembles

Entrée : A, B, C, trois ensembles finis

Sortie : D, un ensemble tel que $A : B \doteq C : D$

1: si $B \cap C \subset A$ *et* $A \subset B \cup C$ alors
2: $\quad \overline{\overline{A}} \leftarrow$ le multi-ensemble correspondant à A
3: $\quad \overline{\overline{B}} \leftarrow$ le multi-ensemble correspondant à B
4: $\quad \overline{\overline{C}} \leftarrow$ le multi-ensemble correspondant à C
5: $\quad \overline{\overline{D}} \leftarrow$ résultat de l'appel de l'algorithme de résolution d'équations analogiques entre multi-ensembles avec $\overline{\overline{A}}$, $\overline{\overline{B}}$ et $\overline{\overline{C}}$
6: $\quad D \leftarrow$ l'ensemble correspondant à $\overline{\overline{D}}$
7: $\quad$ retourner D
8: sinon
9: $\quad$ écrire « pas de solution »
10: fin si

7.3 Chaînes de symboles

7.3.1 Prétraitement

Pour résoudre une équation analogique entre chaînes de symboles $A : B \doteq C : D$ d'inconnue D, nous nous servons du théorème le plus accompli que nous ayons pu donner. Il s'agit du théorème 17 (p. 152). Il permet de séparer clairement une étape de prétraitement de l'étape de résolution proprement dite, puisque, sous cette forme que nous rappelons ici :

$$A : B \doteq C : D \quad \Rightarrow \quad \left\{ \begin{array}{rcl} |D| & = & -|A| + |B| + |C| \\ \sigma(B, D) & = & -|A| + |B| + \sigma(A, C) \\ \sigma(C, D) & = & -|A| + |C| + \sigma(A, B) \\ \gamma(A, B, C, D) & = & -|A| + \sigma(A, B) + \sigma(A, C) \end{array} \right.$$

D n'apparaît qu'à gauche du signe égal, et les côtés droits des égalités ne contiennent que les données d'entrée A, B et C. Le calcul des valeurs trouvées en partie droite constituera donc le prétraitement, alors que la résolution proprement dite consistera à produire des chaînes de symboles tels que les résultats obtenus par calcul des valeurs de la partie gauche vérifient le système.

Longueurs des solutions D

Figure 7.4: Résolution d'équations analogiques entre chaînes de symboles. Première esquisse

Entrée : A, B, C, trois chaînes de symboles

Sortie : D, une ou plusieurs chaînes de symboles telles que $A : B \doteq C : D$

1: calculer $|A|$
2: calculer $|B|$
3: calculer $|C|$
4: $|D| \leftarrow |B| + |C| - |A|$
5: si $0 \leq |D|$ alors
6: continuer la résolution
7: sinon
8: écrire « pas de solution »
9: fin si

La première ligne du système mentionné ci-dessus, qui constitue aussi le théorème 13 (p. 149), permet de calculer à l'avance la longueur des solutions s'il en existe :

$$|D| = -|A| + |B| + |C|$$

Dans les cas où la valeur obtenue est négative, la résolution de l'équation analogique peut tout de suite s'arrêter sur un échec. Dans le cas où la valeur obtenue est zéro, la chaîne vide est une solution candidate de l'équation, mais les autres conditions restent à vérifier. On obtient l'algorithme 7.4.

Ensemble des symboles de D

Dans un deuxième temps, il est possible de calculer l'ensemble des symboles apparaissant dans chacune des solutions de l'équation, s'il en existe, grâce aux multi-ensembles associés aux chaînes. Ce calcul s'effectue en deux étapes :

- premièrement, on prétraite les chaînes A, B et C pour trouver les multi-ensemble de leurs symboles ;
- deuxièmement, on utilise l'égalité sur les multi-ensembles pour obtenir l'ensemble des symboles de D.

C'est ici que se place la vérification concernant la chaîne vide solution candidate : si le multi-ensemble obtenu est vide, alors, la chaîne vide est bien solution de l'équation analogique. Sinon, il n'y a pas de solution. Cet algorithme est donné en 7.5.

Précalcul d'informations sur les chaînes

Le calcul des longueurs de chaînes, des multi-ensembles associés aux chaînes et de leur représentation sous forme de chaînes binaires peut être regroupé en un calcul préparatoire aux opérations ultérieures de résolution.

Tableau 7.1: Tableau T des positions dans la chaîne *aslama*

n	1	2	3	4	5	6	7	8	9	10
$T[n]$	6	4	1	0	3	0	5	0	2	0

Tableau 7.2: Tableau I des indices intermédiaires pour la chaîne *aslama*

n	a	...	l	m	...	s	...
$I[n]$	1	0	5	7	0	9	0

Le but de la fonction suivante est de pouvoir accéder rapidement à toutes les positions d'un même symbole dans une chaîne donnée. Le codage obtenu ici

Figure 7.5: Résolution d'équations analogiques entre chaînes de symboles. Deuxième esquisse

Entrée: A, B, C, trois chaînes de symboles

Sortie: D, une ou plusieurs chaînes de symboles telles que $A : B \doteq C : D$

1: calculer $|A|$
2: calculer $|B|$
3: calculer $|C|$
4: $|D| \leftarrow |B| + |C| - |A|$
5: si $0 \leq |D|$ alors
6: $\quad \overline{\overline{A}} \leftarrow$ multi-ensemble associé à A
7: $\quad \overline{\overline{B}} \leftarrow$ multi-ensemble associé à B
8: $\quad \overline{\overline{C}} \leftarrow$ multi-ensemble associé à C
9: $\quad \overline{\overline{D}} \leftarrow$ résultat de l'appel à la résolution de l'équation analogique $\overline{\overline{A}} : \overline{\overline{B}} \doteq \overline{\overline{C}} : \overline{\overline{D}}$
10: $\quad$ si $\overline{\overline{D}}$ existe alors
11: $\quad\quad$ si $0 = |D|$ et $\phi \neq \overline{\overline{D}}$ alors
12: $\quad\quad\quad$ écrire « pas de solution »
13: $\quad\quad$ sinon
14: $\quad\quad\quad$ continuer la résolution
15: $\quad\quad$ fin si
16: $\quad$ fin si
17: sinon
18: $\quad$ écrire « pas de solution »
19: fin si

produit un ordre décroissant. Cela n'est pas important. Nous aurions très bien pu aussi le faire par ordre croissant. Afin d'avoir une représentation compacte, on choisit de mettre les informations sur les différents symboles dans un unique tableau, en les séparant par des zéros, qui sont utilisés comme bornes d'arrêt facilitant la programmation.

Afin d'accéder à ce premier tableau, un second tableau d'indices intermédiaires est nécessaire. Il est lui indicé sur l'alphabet, et donne, pour chaque symbole de l'alphabet, le début du sous-tableau des indices de ce symbole dans le premier tableau.

Par exemple, soit la chaîne $A = aslama$. Les tableaux produits par la fonction sont donnés par les tables 7.1 et 7.2 (p. 214). À titre d'illustration, pour le symbole s, le second tableau nous donne $I[s] = 9$ qui est un indice dans le premier tableau T. Le premier tableau donne $T[9] = 2$ car $aslama[2] = s$. En conjonction, ces deux tableaux permettent aussi d'obtenir de suite toutes

les positions d'un même symbole dans la chaîne. Par exemple, pour le symbole a, le second tableau nous donne $I[a] = 1$. À partir de cet indice, le tableau T donne l'une à la suite de l'autre toutes les positions des occurrences de a dans la chaîne *aslama* : 6, 4 et 1 jusqu'à la valeur d'arrêt, 0. L'algorithme 7.6 (p. 217) réalise le calcul de ces tableaux pour une chaîne donnée de symboles.

Figure 7.6: Calcul des tableaux des positions et des indices intermédiaires pour une chaîne de symboles

Entrée : A une chaîne de symboles

Sortie : T : un tableau donnant les positions des symboles de A
$\underline{I}$: un tableau d'indices intermédiaires
$\overline{\overline{A}}$: un tableau représentant le multi-ensemble associé à A

Registre : TMP : un tableau temporaire d'entiers initialisés à 0, indicé par les symboles de l'alphabet
Registre : DER : un entier représentant la taille du tableau T

{Comptage du nombre d'occurrences par symbole}
1: pour $i = 1$ à $|A|$ faire
2: $\quad TMP[A[i]] \leftarrow TMP[A[i]] + 1$
3: fin pour
{Remplissage du tableau intermédiaire I}
4: pour $i = 1$ à la taille de l'alphabet faire
5: $\quad \overline{\overline{A}}[i] \leftarrow TMP[i]$
6: $\quad$ si $0 \neq TMP[i]$ alors
7: $\quad\quad I[i] \leftarrow DER$
8: $\quad\quad DER \leftarrow DER + TMP[i] + 1$
9: $\quad$ fin si
10: fin pour
{Remplissage du tableau des positions T}
11: pour $i = 1$ à $|A|$ faire
12: $\quad$ si $0 \neq TMP[i]$ alors
13: $\quad\quad T[I[A[i]] + TMP[A[i]]] \leftarrow i$
14: $\quad\quad TMP[A[i]] \leftarrow TMP[A[i]] - 1$
15: $\quad$ fin si
16: fin pour

7.3.2 Calcul des multi-ensembles

Les tableaux précédents peuvent être calculés pour chacune des chaînes A, B et C des données d'une équation analogique entre chaînes de symboles. On obtient donc six tableaux, que l'on peut désigner par $T[A]$, $T[B]$, $T[C]$, et $I[A]$, $I[B]$, $I[C]$. Lors de la construction de ces tableaux, nous avons utilisé un tableau intermédiaire *TMP* qui contenait le nombre d'occurrences de chaque symbole dans la chaîne. Ce tableau est donc une représentation du multi-ensemble associé à la chaîne. On peut donc le mémoriser. C'est ce que nous avons fait dans l'algorithme 7.6. Grâce à cela, l'algorithme 7.5 peut être précisé par l'algorithme 7.7.

Figure 7.7: Résolution d'équations analogiques entre chaînes de symboles. Troisième esquisse

Entrée : A, B, C, trois chaînes de symboles

Sortie : D, une ou plusieurs chaînes de symboles telles que $A : B \doteq C : D$

{Calcul des représentations compactes des chaînes A, B et C}

pour $X = A, B, C$ faire
  calculer $|X|$, $T[X]$, $I[X]$ et $\overline{\overline{X}}$
fin pour

$|D| \leftarrow |B| + |C| - |A|$
$cardinal \leftarrow 0$
pour $i = 1$ à la taille de l'alphabet faire
  $\overline{\overline{D}}[i] \leftarrow \overline{\overline{B}}[i] + \overline{\overline{C}}[i] - \overline{\overline{A}}[i]$
  {Tester si le multi-ensemble de D est possible}
  si $0 > \overline{\overline{D}}[i]$ alors
    écrire « pas de solutions »
    quitter la fonction
  sinon
    si $0 < \overline{\overline{D}}[i]$ alors
      $cardinal \leftarrow cardinal + 1$
    fin si
  fin si
fin pour
{Tester si les deux cardinaux de D sont compatibles}
si $cardinal \neq |D|$ alors
  écrire « pas de solution »
  quitter la fonction
fin si
{Poursuite de la résolution}
continuer la résolution

Calcul des similitudes

Dans un troisième temps, on peut alors passer au calcul des parties droites des trois dernières égalités du système (p. 213) donné par le théorème 16. Cette forme du système permet en effet de calculer à l'avance les similitudes $\sigma(A, B)$ et $\sigma(A, C)$, ainsi que la valeur de $\gamma(A, B, C, x)$. Pour le calcul des similitudes, on peut adopter le calcul classique au moyen de la formule :

$$\sigma(a.A', b.B') = \max \begin{pmatrix} \sigma(A', B') + \sigma(a, b), \\ \sigma(a.A', B'), \\ \sigma(A', b.B') \end{pmatrix}$$

Cela se réalise facilement par une méthode de programmation dynamique en retenant les calculs intermédiaires dans une matrice. Nous avons choisi d'utiliser une autre méthode, plus rapide. Elle repose sur un algorithme travaillant sur une représentation en chaînes binaires[1] des chaînes d'entrée A, B et C. Cette représentation permet de diviser le calcul de la similitude entre deux chaînes de caractères par la longueur du mot-machine. Le comportement asymptotique n'est donc pas modifié. Mais si cette amélioration est insensible pour un seul calcul, elle ne l'est pas lorsque l'on a à effectuer des milliers voire des millions de résolutions d'équations analogiques. Or, dans nos applications à la traduction automatique ou à l'analyse, le nombre de résolutions d'équations analogiques à effectuer est justement de l'ordre du million. L'amélioration, même si elle ne change pas le comportement asymptotique de l'algorithme apporte donc un gain en temps important dans nos applications.

[1] Allison & Dix, *A bit string longest common subsequence algorithm*, 1986.

Figure 7.8: Résolution d'équations analogiques entre chaînes de symboles. Quatrième esquisse

Entrée : A, B, C, trois chaînes de symboles

Sortie : D, une ou plusieurs chaînes de symboles telles que $A : B \doteq C : D$

{Calcul des représentations compactes des chaînes A, B et C}
pour $X = A, B, C$ faire
  calculer $|X|$, $T[X]$, $I[X]$ et $\overline{\overline{X}}$
  calculer $|\overline{\overline{X}}|$
fin pour

$|\overline{\overline{D}}| \leftarrow |\overline{\overline{B}}| + |\overline{\overline{C}}| - |\overline{\overline{A}}|$
$cardinal \leftarrow 0$
pour $i = 1$ à la taille de l'alphabet faire
  $\overline{\overline{D}}[i] \leftarrow \overline{\overline{B}}[i] + \overline{\overline{C}}[i] - \overline{\overline{A}}[i]$
  {Tester si le multi-ensemble de D est possible}
  si $0 > \overline{\overline{D}}[i]$ alors
    écrire « pas de solutions »
    quitter la fonction
  sinon si $0 < \overline{\overline{D}}[i]$ alors
    $cardinal \leftarrow cardinal + \overline{\overline{D}}[i]$
  fin si
fin pour
{Tester si les deux cardinaux de D sont compatibles}
si $cardinal \neq |\overline{\overline{D}}|$ alors
  écrire « pas de solution »
  quitter la fonction
fin si
{Calcul des similitudes}
pour $X = A, B, C$ faire
  calculer les représentations binaires de X pour chaque symbole
fin pour
calculer $\sigma(A, B)$, $\sigma(A, C)$
$\sigma(B, D) \leftarrow -|A| + |B| + \sigma(A, C)$
$\sigma(C, D) \leftarrow -|A| + |C| + \sigma(A, B)$
$\gamma(A, B, C, D) \leftarrow -|A| + \sigma(A, B) + \sigma(A, C)$

Ordonnancement des symboles dans D

En résumé des étapes précédentes, nous avons mis en place presque tous les éléments pour pouvoir énumérer les solutions possibles d'une équation analogique. Nous savons quelle taille ont les solutions. Nous savons aussi quels symboles apparaissent dans les solutions, et en quel nombre. Seul nous reste à préciser leur ordonnancement. Nous pourrions dès à présent énumérer toutes les combinaisons possibles, calculer les similitudes avec les deuxième et troisième chaînes, B et C, et vérifier a posteriori qu'elles sont égales aux similitudes attendues. Les combinaisons pour lesquelles les égalités ne sont pas vérifiées peuvent être simplement éliminées. Nous avons déjà vu que, même ainsi, parmi les solutions obtenues, certaines ne correspondraient pas au sentiment de l'analogie (voir p. 156). Nous avions dit à cette occasion que nous attendons l'élimination des solutions superfétatoires des résultats sur la contiguïté. Les intuitions que nous avons exposées plus haut permettent déjà de concevoir l'algorithme 7.9 qui construit un tableau des poids des symboles pour chaque position de D (à titre d'exemple, voir plus haut, le tableau 4.5, p. 157). Un tel tableau est indicé sur les positions possibles dans D, et sur ses symboles. Ces derniers sont connus, comme nous venons de le voir. En divisant chaque case du tableau par la somme des éléments de la colonne à laquelle elle appartient, on obtient une matrice des probabilités d'occurrence de chaque symbole en chaque position de D. Afin d'obtenir ce tableau et cette matrice, un calcul préliminaire de tables donnant les nombres d'occurrences cumulées de chaque symbole dans chaque chaîne est nécessaire. Comme ce calcul est trivial, nous ne donnons pas l'algorithme, mais, à titre d'exemple, nous montrons ces tables pour les trois chaînes *ab*, *aabb* et *aaaabbbb* (voir tableau 7.3).

Tableau 7.3: Occurrences cumulées des symboles dans les chaînes *ab*, *aabb* et *aaaabbbb*

	a	*b*
a	1	1
b	0	1

	a	*a*	*b*	*b*
a	1	2	2	2
b	0	0	1	2

	a	*a*	*a*	*a*	*b*	*b*	*b*	*b*
a	1	2	3	4	4	4	4	4
b	0	0	0	0	1	2	3	4

L'algorithme 7.9 permet de faire le calcul des poids des symboles dans la position suivante de D étant donné un préfixe de D. Il est très lourd puisqu'il est en $O(|B| \times |C| \times |\overline{A} \cup \overline{B} \cup \overline{C}|)$. C'est cependant un comportement polynômial. Tant que les résultats formels ne seront pas obtenus pour la contiguïté, nous ne nous intéressons pas à l'optimisation de cet algorithme. Mais nous avons une intuition très forte que les résultats finals sur la contiguïté seront polynomiaux, ce qui conservera tous les résultats annoncés sur l'analyse polynomiale de langages de chaînes analogiques particuliers donnés plus haut (p. 181). En effet, nous ne savons pas encore exactement ce que nous recherchons, c'est-à-dire ce qui, dans ce calcul, est nécessaire et suffisant à la résolution des équations analogiques. Un exemple de résultat de calculs effectués par cet algorithme a

déjà été donné dans le tableau 4.5 (p. 157).

En appelant récursivement la fonction donnée par l'algorithme 7.9, nous pouvons construire les solutions D position par position en commençant par le début de la chaîne. Dès qu'une telle solution est obtenue, on peut alors calculer $\sigma(B, D)$ et $\sigma(C, D)$ pour les comparer aux résultats précalculés dans l'algorithme 7.8. Si les égalités sont vérifiées, la solution peut être conservée. Cet algorithme nous permet d'ores et déjà de proposer des solutions à des équations analogiques entre chaînes de symboles. Testé sur une liste d'exemples linguistiques et formels, il nous conforte dans l'idée que nous nous dirigeons dans la bonne direction. En effet, il semble se contenter de surgénérer en n'écartant aucune solution réelle.

Figure 7.9: Construction du tableau des poids des symboles en chaque position de D étant donné le préfixe de D jusqu'à cette position

Entrée : A, B, C, D', quatre chaînes de symboles

Sortie : V_D, le vecteur des poids de chaque symbole de D pour la position immédiatement après D'

```
pour X = A, B, C, D' faire
    X̄ ← ensemble des symboles de X
    M_X ← tableau des occurrences cumulées des symboles de X
fin pour
i_D ← |D'| + 1
pour i_B = 1 à |B| faire
    pour i_C = 1 à |C| faire
        i_A = i_B + i_C − i_D' − 1
        si 1 ≤ i_A ≤ |B| + |C| − |A| alors
            {La variable cohérent est une variable de contrôle}
            cohérent ← vrai
            {Sauvegarder les valeurs pour pouvoir les rétablir éventuellement}
            pour chaque a ∈ Ā ∪ B̄ ∪ C̄ faire
                s[a] ← M_D[i_D ; a]
            fin pour
            pour chaque a ∈ Ā ∪ B̄ ∪ C̄ et tant que cohérent = vrai faire
                v ← M_B[i_B ; a] + M_C[i_C ; a] − M_A[i_A ; a]
                si i_D > 1 alors
                    v = v − M_D'[i_D − 1 ; a]
                fin si
                {Tester la cohérence pour l'analogie entre multi-ensembles}
                si 0 ≤ v alors
                    M_D[i_D ; a] ← M_D[i_D ; a] + v
                sinon
                    cohérent ← faux
                fin si
            fin pour
            {Rétablissement de la valeur si une incohérence pour l'analogie entre multi-ensembles a été détectée}
            si cohérent = faux alors
                M_D[j_D ; a] ← s[a]
            fin si
        fin si
    fin pour
fin pour
```

Chapitre 8

Langages de chaînes analogiques

8.1 Production

Tout langage de chaînes analogiques est défini par la donnée d'un ensemble d'éléments attestés et d'un ensemble de modèles. Nous avions noté $\Lambda(\mathcal{A}, \mathcal{M})$ le langage de chaînes analogiques obtenu pour deux tels ensembles $\mathcal{A}$ et $\mathcal{M}$ (voir p. 169). Nous avions aussi dit que l'on obtient tous les éléments d'un langage de chaînes analogiques par induction. Du point de vue algorithmique, il est plus facile de construire le langage par couches successives (p. 170). L'algorithme composé des deux fonctions 8.1 et 8.2 fait simplement cela. Il commence par énumérer les éléments de $\mathcal{A}$, égal, rappelons-le, à la première couche du langage de chaînes analogiques, Λ_0. Puis, il appelle récursivement une fonction calculant les éléments de la couche suivante en appliquant simplement tous les modèles sur les éléments de la couche précédente. La construction de Λ_{n+1} passe évidemment par la résolution d'équations analogiques au moyen de l'algorithme donné plus haut. Rappelons que la résolution d'une telle équation délivre zéro, une ou plusieurs solutions.

Cet algorithme est général, dans le sens où il pourrait être utilisé avec n'importe quel autre type d'objets que les chaînes de symboles à condition, bien sûr, de disposer d'un algorithme de résolution d'équations analogiques pour le type d'objets en question.

Figure 8.1: Production des éléments d'un langage de chaînes analogiques par couches successives. Fonction d'appel

Entrée : $\mathcal{A}$, un ensemble de chaînes attestées,
$\mathcal{M}$, un ensemble de modèles

Sortie : les éléments du langage de chaînes analogiques $\Lambda(\mathcal{A}, \mathcal{M})$

si $\mathcal{A} = \phi$ alors
  le langage de chaînes analogiques est vide
sinon
  $\Lambda_0 \leftarrow \mathcal{A}$
  énumérer les éléments de Λ_0
  si $\mathcal{M} \neq \phi$ alors
    appel de la fonction suivante avec $\mathcal{A}$, $\mathcal{M}$ et Λ_0
  fin si
fin si

Figure 8.2: Production des éléments d'un langage de chaînes analogiques par couches successives. Fonction récursive

Entrée : $\mathcal{A}$, un ensemble de chaînes attestées,
$\mathcal{M}$, un ensemble de modèles,
Λ_n, un ensemble de chaînes de symboles

Sortie : les éléments du langage de chaînes analogiques $\Lambda(\mathcal{A}, \mathcal{M})$ à partir de la $(n+1)^{\text{e}}$ couche

Registre : Λ_{n+1} un ensemble de chaînes initialié à ϕ

pour chaque $C \in \Lambda_n$ faire
  pour chaque $A \rightarrow B \in \mathcal{M}$ faire
    $\Lambda_{n+1} \leftarrow \Lambda_{n+1} \cup \{D \in \mathcal{V}^* \,/\, A : B \doteq C : D\}$
  fin pour
fin pour
énumérer $\Lambda_{n+1} \setminus \bigcup_{i=0}^{n} \Lambda_i$
appel récursif avec $\Lambda_{n+1} \setminus \bigcup_{i=0}^{n} \Lambda_i$

8.2 Reconnaissance

8.2.1 Appartenance à un langage de chaînes analogiques

Le problème réciproque de la production est la reconnaissance. Suivant la définition des langages de chaînes analogiques, l'appartenance d'une chaîne de symboles à un tel langage se décide par la réduction de cette chaîne selon les modèles de $\mathcal{M}$, récursivement, jusqu'à obtenir un élément de $\mathcal{A}$. Cela conduit naturellement à l'algorithme donné en 8.3. De même que dans le cas de la production, l'algorithme est général, car il pourrait être utilisé pour n'importe quel type d'objets pour lequel on disposerait d'un algorithme de résolution d'équations analogiques.

Figure 8.3: Appartenance d'une chaîne à un langage de chaînes analogiques

Entrée : $\mathcal{A}$, un ensemble de chaînes attestées,
$\mathcal{M}$, un ensemble de modèles,
D, une chaîne de symboles

Sortie : vrai si $D \in \Lambda(\mathcal{A}, \mathcal{M})$,
faux sinon.

Registre : réponse, une variable logique initialisée à faux
$\mathcal{C}$, un ensemble de chaînes de symboles intialisé à ϕ

si $D \in \mathcal{A}$ alors
  réponse $\leftarrow$ vrai
sinon
  pour chaque $A \rightarrow B \in \mathcal{M}$ faire
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{C \in \mathcal{V}^* \ / \ A : B \doteq C : D \ \}$
  fin pour
  si $\mathcal{A} \cap \mathcal{C} \neq \phi$ alors
    réponse $\leftarrow$ vrai
  sinon
    si $\mathcal{C} \neq \phi$ alors
      pour chaque $C \in \mathcal{C}$ et tant que réponse = faux faire
        réponse $\leftarrow$ résultat de l'appel récursif avec $\mathcal{A}$, $\mathcal{M}$ et C
      fin pour
    fin si
  fin si
fin si
retourner la réponse

8.2.2 Test d'intersection non vide avec un langage de chaînes analogiques

L'algorithme précédent peut se réexprimer de façon plus élégante, en le décomposant en deux fonctions. L'introduction d'une fonction (p. 228) pour le test de l'intersection non vide d'un ensemble de chaînes de symboles avec un langage de chaînes analogiques généralise l'algorithme précédent. En particulier, il nous permettra de travailler plus facilement par la suite. Trivialement, pour résoudre le problème de l'appartenance d'une seule chaîne de symbole à un langage de chaînes analogiques, il suffit de construire le singleton correspondant à cette chaîne, et d'appeler cette fonction. C'est simplement ce que fait l'algorithme 8.5.

Figure 8.4: Test d'intersection non vide d'un ensemble de chaînes avec un langage de chaînes analogiques

Entrée : $\mathcal{A}$, un ensemble de chaînes attestées,
$\mathcal{M}$, un ensemble de modèles,
$\mathcal{D}$, un ensemble de chaînes de symboles

Sortie : vrai si l'un au moins des éléments de $\mathcal{D}$ appartient à $\Lambda(\mathcal{A}, \mathcal{M})$
c'est-à-dire si $\mathcal{D} \cap \Lambda(\mathcal{A}, \mathcal{M}) \neq \phi$,
faux sinon.

Registre : réponse, une variable logique initialisée à faux

1: si $\mathcal{A} \cap \mathcal{D} \neq \phi$ alors
2: réponse $\leftarrow$ vrai
3: sinon si $\mathcal{D} \neq \phi$ alors
4: pour chaque $D \in \mathcal{D}$ et tant que réponse = faux faire
5: pour chaque $A \rightarrow B \in \mathcal{M}$ faire
6: réponse $\leftarrow$ appel récursif avec $\mathcal{A}$, $\mathcal{M}$ et $\{C \in \mathcal{V}^* / \ A : B \doteq C : D \}$
7: fin pour
8: fin pour
9: fin si
10: retourner la réponse

8.2.3 Test d'appartenance à une couche

En production, la n^{e} couche d'un langage de chaînes analogiques (p. 170) est l'ensemble des chaînes obtenues par n applications successives de modèles à partir des chaînes attestées. En reconnaissance, l'appartenance d'une chaîne à Λ_n est prouvée si l'on peut montrer que la chaîne peut être réduite à une chaîne

Figure 8.5: Appartenance d'une chaîne à un langage de chaînes analogiques (deuxième algorithme)

Entrée : $\mathcal{A}$ un ensemble de chaînes attestées,
$\mathcal{M}$ un ensemble de modèles,
D une chaîne de symboles

Sortie : vrai si $D \in \Lambda(\mathcal{A}, \mathcal{M})$,
faux sinon.

Registre : réponse, une variable logique initialisée à faux

1: réponse $\leftarrow$ résultat de l'appel à la fonction pécédente avec $\{D\}$
2: retourner la réponse

attestée en un nombre de réductions inférieur à n. Dans l'algorithme 8.6, la ligne du test d'arrêt utilise le fait que $\Lambda_0 = \mathcal{A}$.

L'algorithme 8.6 correspond à une exploration en profondeur d'abord. Il est mieux de procéder en largeur d'abord pour maintenir l'exploration couche par couche. Cette façon de faire est donnée par l'algorithme 8.7. Elle a l'avantage qu'à chaque étape on peut éliminer du nouvel ensemble obtenu les éléments de l'ensemble précédent. Cela permet de gagner en temps de calcul et cela constitue une amélioration par rapport à tous les algorithmes précédents, où le risque existait de tester plusieurs fois, en vain, l'appartenance d'éléments en dehors du langage.

Dans nos programmes, toutes les fonctions ensemblistes, telles que l'union, l'intersection ou la différence entre ensembles de chaînes sont réalisées sur une structure de données imposant un ordre sur les chaînes et permettant le test rapide de l'appartenance d'une chaîne à l'ensemble. Il s'agit chez nous, d'une structure d'arbres semi-équilibrés à la Adel'son-Velskiĭ et Landis[1], mais nous aurions pu en adopter une autre.

[1] Adel'son-Velskiĭ & Landis, *An algorithm for the organization of information*, 1962.

Figure 8.6: Test d'intersection non vide d'un ensemble de chaînes avec un langage de chaînes analogiques jusquà sa n^{e} couche. En profondeur d'abord

Entrée : $\mathcal{A}$ un ensemble de chaînes attestées,
$\mathcal{M}$ un ensemble de modèles,
$\mathcal{D}$ un ensemble de chaînes de symboles,
n un entier

Sortie : vrai si un au moins des éléments de $\mathcal{D}$ appartient à Λ_n du $\Lambda(\mathcal{A}, \mathcal{M})$, c'est-à-dire si $\mathcal{D} \cap \Lambda_n \neq \phi$,
faux sinon.

Registre : réponse, une variable logique initialisée à faux

1: si $n < 0$ alors
2: erreur : n doit être positif ou nul
3: sinon si $n = 0$ alors
4: si $\mathcal{A} \cap \mathcal{D} \neq \phi$ alors
5: réponse $\leftarrow$ vrai
6: fin si
7: sinon
8: si $\mathcal{D} \neq \phi$ alors
9: pour chaque $D \in \mathcal{D}$ et tant que réponse = faux faire
10: pour chaque $A \rightarrow B \in \mathcal{M}$ faire
11: réponse $\leftarrow$ réponse à l'appel récursif avec $\mathcal{A}$, $\mathcal{M}$, $\{C \in \mathcal{V}^* / \quad A : B \doteq C : D \}$ et $n - 1$
12: fin pour
13: fin pour
14: fin si
15: fin si
16: retourner la réponse

Figure 8.7: Test d'intersection non vide d'un ensemble de chaînes avec un langage de chaînes analogiques jusquà sa n^{e} couche. En largeur d'abord

Entrée : $\mathcal{A}$ un ensemble de chaînes attestées,
$\mathcal{M}$ un ensemble de modèles,
$\mathcal{D}$ un ensemble de chaînes de symboles,
n un entier

Sortie : vrai si un au moins des éléments de $\mathcal{D}$ appartient à Λ_n du $\Lambda(\mathcal{A}, \mathcal{M})$, c'est-à-dire si $\mathcal{D} \cap \Lambda_n \neq \phi$,
faux sinon.

Registre : réponse, une variable logique initialisée à faux
$\mathcal{C}$, un ensemble de chaînes de symboles intialisé à ϕ

si $n < 0$ alors
  erreur : n doit être positif ou nul
sinon si $n = 0$ alors
  si $\mathcal{A} \cap \mathcal{D} \neq \phi$ alors
    réponse $\leftarrow$ vrai
  fin si
sinon
  si $\mathcal{D} \neq \phi$ alors
    pour chaque $D \in \mathcal{D}$ faire
      pour chaque $A \rightarrow B \in \mathcal{M}$ faire
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{C \in \mathcal{V}^* \ / \ A : B \doteq C : D \}$
      fin pour
    fin pour
    réponse $\leftarrow$ réponse à l'appel récursif avec $\mathcal{A}$, $\mathcal{M}$, $\mathcal{C} \setminus \mathcal{D}$ et $n - 1$
  fin si
fin si
retourner la réponse

8.3 Construction d'un ensemble des modèles

8.3.1 Restriction aux langages décroissants

La terminaison des algorithmes 8.3 et 8.4 n'est pas assurée dans le cas général. Mais si le langage de chaînes analogiques considéré est décroissant (voir plus haut p. 173 et plus bas p. 307), alors, la taille des C diminue strictement à chaque appel récursif et l'exécution se termine à coup sûr. Dans ce cas, on peut calculer le temps d'exécution en moyenne dans le pire des cas. Passons sous silence le temps nécessaire pour effectuer les opérations ensemblistes telles que tester l'intersection non vide de $\mathcal{A}$ et de $\mathcal{D}$ ou faire l'union de $\mathcal{C}$ avec un ensemble de chaînes. Le pire des cas est celui où, à chaque étape, on aurait un nombre maximal de solutions à l'équation analogique $B : A \doteq D : C$ d'inconnue C, mais où aucune des solutions obtenue n'appartiendrait jamais à $\mathcal{A}$. Supposons que le temps moyen de résolution soit de T. À chaque étape, on doit résoudre $|\mathcal{M}|$ équations analogiques et il faut réappeler l'algorithme récursivement autant de fois qu'il y a d'éléments dans $\mathcal{C}$. Dans notre développement sur la résolution d'équations analogiques, nous n'avons pas mentionné nos résultats préliminaires sur le calcul a priori du nombre de solutions à une équation analogique. Nous laissons donc ici $|\mathcal{C}|$ imprécisé. Supposons qu'il soit de cardinal N en moyenne. Pour un langage de chaînes analogiques décroissant, la profondeur des appels récursifs est borné par la taille de D. Au total, dans le pire des cas, la réponse faux est donc apportée en un temps proportionnel à $(|\mathcal{M}| \times T \times N)^{|D|}$.

8.3.2 Restriction aux langages paresseux

Dans nos expériences, nous sommes pratiquement confronté au cas où aucun langage de chaînes analogiques ne nous est explicitement donné. Il nous faut donc en construire un à partir des données brutes, généralement un corpus de phrases. Dans ce cas-là, l'attitude la plus simple est de considérer le langage de chaînes analogiques paresseux construit à partir de l'ensemble des données. Or, dans le cas des langages de chaînes analogiques paresseux (p. 174), $\mathcal{M}$ est égal à $\mathcal{A}^2$. On pourrait donc simplement appeler les algorithmes précédents avec $\mathcal{A}$ et $\mathcal{A}^2$. Mais du point de vue algorithmique, il est bon de supprimer de $\mathcal{M} = \mathcal{A}^2$ les couples du type (A, A). Cela évite des boucles qui empêcheraient la terminaison des algorithmes. Dans une étape préliminaire de construction de $\mathcal{M}$ à partir de $\mathcal{A}$, on rejettera donc ces couples. En plus, afin d'assurer la convergence des algorithmes, on ne considérera que les langages de chaînes analogiques paresseux décroissants.

8.3.3 Contrainte de proximité sur les modèles

Considérer, à partir de données bruts, le langage de chaînes analogiques paresseux décroissant peut quand même, en pratique, mener à construire un ensemble de modèles trop important. Typiquement, dans nos expériences de tra-

duction automatique, nos données de base comprennent plus de $|\mathcal{A}| = 150\ 000$ éléments. Avec les restrictions données plus haut, l'ensemble des modèles a tout de même une taille de

$$(|\mathcal{A}|^2 - |\mathcal{A}|)/2 = (15^2 \times 10^4 - 15 \times 10^3)/2 \approx 125 \times 10^8$$

On comprend bien que l'exploration, élément par élément, d'un ensemble de plus de 10^{10} éléments est illusoire. Par exemple, dans le pire des cas, le test de l'appartenance à un tel ensemble d'une chaîne de 10 symboles par l'algorithme 8.3 requerrait un facteur de 10^{100} résolutions d'équations analogiques ! Il est absolument nécessaire de limiter cette explosion. Cela est déjà possible en se contentant de tester l'appartenance jusqu'à une certaine couche seulement du langage de chaînes analogiques. Mais même avec cette restriction, la taille de $\mathcal{M}$ est invalidante. Or, on sent bien que la majorité des modèles de $\mathcal{M}$ ne serviront pas dans la pratique.

Une intuition raisonnable est de ne considérer que les modèles dont les chaînes vérifient une certaine contrainte de proximité. En d'autres termes, pour un modèle $A \to B$, on peut exiger que la similitude $\sigma(A, B)$ entre A et B soit supérieure ou égale à un seuil donné. Pour obtenir une contrainte d'ordre générale, ce seuil peut être exprimé comme un pourcentage, ce qui revient à une similarité (voir p. 101, et la note de bas de page). Nous proposons donc la contrainte suivante.

$$\frac{\sigma(A, B)}{|A| + |B|} \geq k$$

Rappelons que la similarité et la distance entretiennent une relation. La contrainte est donc logiquement équivalente à la contrainte suivante.

$$\frac{\delta(A, B)}{|A| + |B|} \leq 1 - k$$

Cela rentre bien dans le cadre de nos résultats de formalisation de l'analogie entre chaînes de symboles, et plus particulièrement de nos résultats sur le versant de la similarité. En effet, le résultat le plus important que nous ayons obtenu est le suivant.

$$A : B \doteq C : D \quad \Rightarrow \quad \delta(A, B) = \delta(C, D)$$

La contrainte proposée plus haut est donc sensée agir directement sur les analogies possibles. Nous montrerons plus bas (p. 308 et suivantes) quelques expériences menées afin d'étudier l'influence de cette contrainte sur la taille de l'ensemble des modèles.

8.3.4 Contrainte de proximité avec la chaîne testée

Toujours en nous plaçant dans une perspective de moindre effort en traitement automatique des langues, nous pouvons nous poser la question de reprendre l'idée intuitive, mais erronée !, qui voudrait que les analogies soient d'autant

meilleures que les objets y intervenant sont proches. Ainsi, lors de l'exécution des algorithmes précédents, dans un contexte de données réelles et de travail sur les homomorphismes, plutôt que de considérer tous les modèles possibles, on peut se restreindre à ne considérer que les modèles dont l'un des éléments est le plus proche possible de la chaîne examinée à un moment donné. Le critère du plus proche constitue une méthode de sélection des couples « intéressants » de $\mathcal{M}$ pour un D donné. Cette restriction réalise en pratique une réduction importante du calcul global. Pour une chaîne D donnée, l'algorithme 8.8 ne retient que les modèles dont le second élément est le plus similaire à la chaîne D. On pourrait aussi prendre les plus proches. Le résultat de ce calcul réduit considérablement le nombre de modèles à prendre en compte dans le cas général. Une telle technique appliquée pour l'analyse des phrase source dans le cadre de nos expériences de traduction automatique (voir p. 314 et suivantes) rend les temps d'exécution de nos programmes tout à fait raisonnables.

Figure 8.8: Calcul de l'ensemble des modèles « intéressants » pour une chaîne donnée

Entrée : $\mathcal{M}$, un ensemble de modèles,
D, une chaîne de symboles

Sortie : $\mathcal{M}(D)$, l'ensemble des modèles intéressants pour D, initialisé à ϕ

1: $\mathcal{C} \leftarrow \{\exists A \rightarrow B \in \mathcal{M} \ / \ \sigma(D, B) \text{ est maximale}\}$

L'algorithme 8.8 n'est possible que si un algorithme de filtrage tolérant[2] existe sur les objets considérés. C'est le cas pour les chaînes sur un ensemble fini de symboles. Même mieux, nous avons proposé un algorithme plus rapide en moyenne que l'algorithme prétendûment le plus rapide du moment. Dans une expérience de 250 000 recherches de mots dans un texte de 25 000 mots, notre algorithme[3] exhibe, en moyenne, un meilleur comportement asymptotique que l'algorithme de la commande *agrep*[4]. Pour une chaîne donnée, et un fichier donné, il permet plusieurs modes d'utilisation :

- recherche d'enregistrements contenant des sous-chaînes à une distance d'édition inférieure à un seuil donné ;

[2]Aussi appelé recherche approximative.

[3]Lepage, *String approximate pattern-matching*, 1997 en donne une présentation succinte. Il fait déjà l'objet de brevets japonais : Lepage & 安藤　真一 (Andou Sin-Iti), 類似検索装置, 1998, et américain : Lepage & Ando, *Similarity search apparatus for searching unit string based on similarity*, 1999. Une demande de brevet européen a aussi été déposée : Lepage, *Apparatus and method for producing analogically similar word based on pseudo-distances between words*, 1999.

[4]Wu & Manber, *Fast text searching allowing errors*, 1992.

- recherche des enregistrements contenant les chaînes du fichier les plus proches de la chaîne donnée au sens de la distance d'édition. Dans ce cas, le programme donne aussi la valeur de cette distance minimale ;
- recherche d'enregistrements contenant des sous-chaînes ayant une similitude supérieure à un seuil donné ;
- recherche des enregistrements contenant les chaînes du fichier les plus semblables à la chaîne donnée. Dans ce cas, le programme donne aussi la valeur de cette similitude maximale ;

Différents modes d'utilisation sont donc envisageables. Tester ces différents modes et déterminer lequel est le plus efficace pratiquement fait bien sûr partie de notre plan de recherche. Il est aussi important de comprendre quelles solutions sont passées sous silence par ces différents modes, car la réduction de $\mathcal{M}$ obtenue n'est qu'heuristique. Elle peut donc laisser de côté nombre de solutions dans l'algorithme 8.5 (voir aussi plus bas p. 276 et p. 232).

8.4 Construction de bases

Le problème de la représentativité est en quelque sorte un intermédiaire entre la production ou de la reconnaissance. Nous avons vu précédemment (p. 189) que la notion de base pouvait rendre compte de cette notion. Pour construire une base au fil de la lecture d'un corpus Λ donné, pour chaque phrase du corpus, il suffit d'ajouter cette phrase à la base si et seulement si aucune analogie ne peut être trouvée avec d'autres phrases de la base.

Figure 8.9: Calcul d'une base pour un corpus donné

Entrée : Λ, un ensemble de chaînes

Sortie : $\mathcal{B}$, un ensemble de chaînes qui est une base pour Λ, initialisé à ϕ

```
1: pour chaque  D ∈ Λ  faire
2:    si  |B| < 3  alors
3:       B ← B ∪ {D}
4:    sinon
5:       si ∃(A,B,C) ∈ B³ /   A : B ≐ C : D   alors
6:          rien
7:       sinon
8:          B ← B ∪ {D}
9:       fin si
10:   fin si
11: fin pour
```

L'algorithme 8.9 vérifie, au fil de la lecture d'un corpus, qu'un libre est bien construit qui est en plus un générateur pour Λ. Il s'agit d'un algorithme glouton, algorithme fréquent pour la résolution de ce genre de problèmes. Nous l'appliquerons au problème de la représentativité de corpus (p. 259) lié à celui de la représentativité des langages de chaînes analogiques et plus précisément au problème de la représentativité de $\mathcal{A}$ pour Λ_1 dans nos notations de la page 174.

Nos expériences nous ont montré que, plutôt que d'examiner tous les triplets de $\mathcal{B}$, il est plus rapide d'examiner les couples (A, B) de $\mathcal{B}^2$, et de résoudre l'équation analogique formée avec D, $B : A \doteq D : C$ d'inconnue C. C'est ce que nous avons réalisé en pratique. Cette manière de faire a l'avantage de rapprocher de la notion théorique de langage de chaînes analogiques paresseux. En effet, la base apparaît alors comme un ensemble $\mathcal{A}$, et le corpus, comme la première couche Λ_1 du langage de chaînes analogiques $\Lambda(\mathcal{A}, \mathcal{A}^2)$.

Chapitre 9

Homomorphismes entre espaces analogiques

Après avoir vu les algorithmes de résolution d'équations analogiques et ceux qui se rattachent aux langages de chaînes analogiques, nous abordons maintenant les algorithmes jouant sur la notion d'homomorphismes entre domaines structurés par l'analogie. Selon la vision statique des choses, un homomorphisme entre espaces analogiques établit une relation entre deux ensembles structurés chacun implicitement par l'analogie. La vision dynamique est mise en action par les algorithmes que nous allons proposer maintenant. Dans une telle vue, deux ensembles définissent deux domaines. Dans notre spécialisation aux chaînes de symboles, à des fins d'application au traitement automatique des langues, chaque ensemble, noté respectivement $\mathcal{A}$ et $\widehat{\mathcal{A}}$ est un ensemble de chaînes de symboles construites sur des vocabulaires a priori différents. On notera ces vocabulaires $\mathcal{V}$ et $\widehat{\mathcal{V}}$. Les domaines des ensembles $\mathcal{A}$ et $\widehat{\mathcal{A}}$ sont donc $\mathcal{V}^*$ et $\widehat{\mathcal{V}}^*$.

9.1 Calcul général

Il s'agit étant donné un élément D quelconque du premier domaine d'obtenir une pluralité éventuelle d'éléments $\widehat{D}$ correspondants dans le second domaine. La première tâche est de vérifier l'appartenance de D au langage de chaînes analogiques $\Lambda(\mathcal{A}, \mathcal{M})$. Nous venons de voir des algorithmes pour cela. Ce sont les algorithmes de reconnaissance (p. 227 à 231). La seconde tâche est de produire des éléments $\widehat{D}$ du langage de chaînes analogiques $\Lambda(\widehat{\mathcal{A}}, \widehat{\mathcal{M}})$ correspondant à D. Or nous avons aussi donné des algorithmes de production (p. 226). Mais ici, la condition de production est que les $\widehat{D}$ correspondent à D. Aussi, nous faut-il en quelque sorte imbriquer la production dans la reconnaissance.

9.1.1 Correspondances

Nous avons bien précisé ailleurs que la correspondance entre $\mathcal{A}$ et $\widehat{\mathcal{A}}$ peut prendre n'importe quelle forme. En général, ce n'est pas une bijection. Or, la forme de la correspondance entre les éléments de $\mathcal{A}$ et ceux de $\widehat{\mathcal{A}}$ contrôle aussi

la production. Si aucun lien ne peut être établi en cours d'exécution, aucune correspondance nouvelle ne saurait être établie.

Avant de nous consacrer au problème général, il est facile de concevoir tout de suite un algorithme élémentaire qui, étant donné $\mathcal{H}$, un ensemble de couples de $\mathcal{A}\times\widehat{\mathcal{A}}$, et D, une chaîne de symboles, produise l'ensemble des formes en correspondance avec D qui existe déjà dans la correspondance entre $\mathcal{A}$ et $\widehat{\mathcal{A}}$: $\{\widehat{D}/(D,\widehat{D})\in\mathcal{H}\}$. Cela impose que D appartienne à $\mathcal{A}$. Évidemment, cet algorithme se généralise facilement à une forme où les correspondances sont obtenues pour un ensemble de chaînes de symboles. C'est uniquement cette forme que nous utiliserons par la suite.

Figure 9.1: Correspondance

Entrée: $\mathcal{H}$, un ensemble de couples de $\mathcal{A}\times\widehat{\mathcal{A}}$,
$\mathcal{D}$, un ensemble de chaîne de symboles

Sortie: l'ensemble $\mathcal{H}'$ des couples $(D,\widehat{D})\in\mathcal{H}$ avec $D\in\mathcal{D}$

1: si $\mathcal{D}\neq\phi$ alors
2: pour chaque $D\in\mathcal{D}$ faire
3: $\mathcal{H}'\leftarrow\mathcal{H}'\cup\{(D,\widehat{D})/(D,\widehat{D})\in\mathcal{H}\}$
4: fin pour
5: fin si
6: retourner $\mathcal{H}'$

9.1.2 Prétraitement

En fait, nous exigeons que les ensembles $\mathcal{A}$ et $\widehat{\mathcal{A}}$ définissent chacun un langage de chaînes analogiques. Si nous ne disposons que de la donnée de ces ensembles, ces langages de chaînes analogiques seront paresseux. Sinon, la donnée supplémentaire d'un ensemble $\mathcal{M}$ de modèles est nécessaire. Afin de gagner en généralité, nous nous placerons dans cette seconde hypothèse.

De plus, comme les correspondances sont de forme quelconque, on peut se contenter de ne retenir dans $\mathcal{A}$ que les éléments apparaissant dans un couple au moins de $\mathcal{H}$. Et de même, pour les éléments de $\mathcal{M}$ on peut ne retenir que ceux dont les deux éléments apparaissent dans au moins un couple de $\mathcal{H}$. En faisant cette hypothèse sur les ensembles donnés en entrée à l'algorithme, à tout élément $A\rightarrow B$ de $\mathcal{M}$, on peut faire correspondre plusieurs couples $(\widehat{A},\widehat{B})$. On appelle l'ensemble de ces couples $\mathcal{H}(A\rightarrow B)$. Si un ensemble $\widehat{\mathcal{M}}$ de modèles est disponible pour le second modèle, alors, on peut imposer que $\mathcal{H}(A\rightarrow B)$ en soit un sous-ensemble.

Dans les algorithmes qui suivent, on supposera donc que tous les éléments de $\mathcal{A}$ et que tous les éléments des couples de $\mathcal{M}$ ont au moins un correspondant

par $\mathcal{H}$. Pour des ensembles $\mathcal{A}$, $\mathcal{M}$, et $\mathcal{H}$ quelconques, cette restriction peut s'obtenir par un prétraitement qui ne pose pas de difficulté à réaliser.

9.1.3 Correspondance pour un ensemble de chaînes de symboles

À partir des algorithmes 8.1 et 8.4, et de tous les impératifs précédents, nous pouvons écrire l'algorithme 9.2. Cet algorithme profite des accélérations déjà en place dans les algorithmes de reconnaissance. Ainsi, au lieu d'essayer tous les triplets possibles de $\mathcal{A}$, et de vérifier s'il existe une analogie avec l'entrée D, on considére les couples de l'ensemble $\mathcal{M}$ des modèles, on résoud l'analogie et on vérifie que la solution, si elle existe, appartient à $\mathcal{A}$ en utilisant n'importe quelle méthode d'indexage.

9.1.4 Correspondance pour une chaîne de symboles

De façon triviale, si on n'a qu'une seule chaîne en entrée, il suffit d'appeler l'algorithme précédent avec le singleton contenant cet unique élément. Bien que cela soit évident, rappelons que même si l'entrée est un singleton, plusieurs éléments du second domaine peuvent être produits comme correspondant à l'entrée. Cela provient bien sûr du fait que plusieurs triplets peuvent être en relation d'analogie avec l'entrée. Mais cela provient aussi du fait qu'une équation analogique peut avoir plusieurs solutions. En plus, une autre pluralité de solutions provient de la correspondance. Les éléments du premier domaine peuvent en effet avoir plusieurs correspondants dans le second domaine.

Figure 9.2: Homomorphisme de langages de chaînes analogiques. En profondeur d'abord

Entrée : $\mathcal{A} \subset \mathcal{V}^*$, un ensemble de chaînes attestées,
$\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$, un ensemble de modèles,
$\widehat{\mathcal{A}} \subset \widehat{\mathcal{V}}^*$, un ensemble de chaînes attestées,
$\widehat{\mathcal{M}} \subset \widehat{\mathcal{V}}^* \times \widehat{\mathcal{V}}^*$, un ensemble de modèles,
$\mathcal{H} \subset \mathcal{A} \times \widehat{\mathcal{A}}$, un ensemble de couples décrivant la correspondance,
$\mathcal{D}$, un ensemble de chaîne de symboles

Sortie : un ensemble $\mathcal{H}'$ de couples $(D, \widehat{D})$ tels que $D \in \mathcal{D}$ et $\widehat{D}$ soit élément du langage de chaînes analogiques $\Lambda(\widehat{\mathcal{A}}, \widehat{\mathcal{M}})$ et corresponde à D

Registre : $\mathcal{C} \subset \mathcal{V}^*$, initialisé à ϕ,
$\mathcal{H}'' \subset (\mathcal{V}^*)^2$, initialisé à ϕ

si $\mathcal{D} \cap \mathcal{A} \neq \phi$ alors
  pour chaque $D \in \mathcal{D} \cap \mathcal{A}$ faire
    $\mathcal{H}' \leftarrow \mathcal{H}' \cup \{(D, \widehat{D}) / (D, \widehat{D}) \in \mathcal{H}\}$
  fin pour
fin si
si $\mathcal{D} \neq \phi$ alors
  pour chaque $D \in \mathcal{D}$ faire
    pour chaque $A \rightarrow B \in \mathcal{M}$ faire
      $\mathcal{C} \leftarrow \{C \ / \ \ A : B \doteq C : D \ \}$
      $\mathcal{H}'' \leftarrow$ résultat de l'appel récursif avec $\mathcal{A}$, $\mathcal{M}$, $\widehat{\mathcal{A}}$, $\widehat{\mathcal{M}}$, $\mathcal{H}$ et $\mathcal{C} \setminus \mathcal{D}$
      pour chaque $(C, \widehat{C}) \in \mathcal{H}''$ faire
        $\widehat{\mathcal{C}} \leftarrow \widehat{\mathcal{C}} \cup \{\widehat{C}\}$
      fin pour
      pour chaque $\widehat{C} \in \widehat{\mathcal{C}}$ faire
        pour chaque $(\widehat{A}, \widehat{B}) \in \mathcal{H}(A \rightarrow B)$ faire
          si $\widehat{A} \rightarrow \widehat{B} \in \widehat{\mathcal{M}}$ alors
            $\mathcal{H}' \leftarrow \mathcal{H}' \cup \{(D, \widehat{D}) \ / \ \ \widehat{A} : \widehat{B} \doteq \widehat{C} : \widehat{D} \ \}$
          fin si
        fin pour
      fin pour
    fin pour
  fin pour
fin si
retourner $\mathcal{H}'$

9.2 Calcul jusqu'à une certaine couche seulement

La terminaison de l'algorithme 9.2 n'est pas assurée. Aussi, il serait bon de pouvoir en limiter l'exécution d'une manière ou d'une autre. Or, on peut remarquer que $\mathcal{H}''$ est un résultat homogène à $\mathcal{H}'$ et à $\mathcal{H}$. Il est obtenu par un appel récursif à la fonction. On entrevoit donc la notion de couches d'un langage de chaînes analogiques. De même que pour la reconnaissance, on peut donc modifier l'algorithme pour qu'il examine différents cas selon le paramètre entier passé en entrée. La nouvelle version de l'algorithme obtenue permet de limiter la récursion à un seuil passé en paramêtre. Ce seuil correspond au numéro de la couche du langage de chaînes analogiques jusqu'à laquelle on autorise une correspondance.

Figure 9.3: Homomorphisme de langages de chaînes analogiques. En largeur d'abord

Entrée : $\mathcal{A} \subset \mathcal{V}^*$, un ensemble de chaînes attestées,
$\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$, un ensemble de modèles,
$\widehat{\mathcal{A}} \subset \widehat{\mathcal{V}}^*$, un ensemble de chaînes attestées,
$\widehat{\mathcal{M}} \subset \widehat{\mathcal{V}}^* \times \widehat{\mathcal{V}}^*$, un ensemble de modèles,
$\mathcal{H} \subset \mathcal{A} \times \widehat{\mathcal{A}}$, un ensemble de couples décrivant la correspondance,
$\mathcal{D}$, un ensemble de chaîne de symboles,
n, un entier

Sortie : un ensemble $\mathcal{H}'$ de couples $(D, \widehat{D})$ tels que $D \in \mathcal{D}$ et $\widehat{D}$ soit élément du langage de chaînes analogiques $\Lambda(\widehat{\mathcal{A}}, \widehat{\mathcal{M}})$ et corresponde à D

Registre : $\mathcal{H}'' \subset \mathcal{V}^* \times \widehat{\mathcal{V}}^*$, initialisé à ϕ,
$\mathcal{C} \subset \mathcal{V}^*$, initialisé à ϕ,
$\mathcal{C}' \subset (\mathcal{V}^*)$, initialisé à ϕ,
$\mathcal{T} \subset (\mathcal{V}^*)^4$, initialisé à ϕ

1: si $n < 0$ alors
2: erreur : n doit être positif ou nul
3: sinon si $n = 0$ alors
4: si $\mathcal{D} \cap \mathcal{A} \neq \phi$ alors
5: pour chaque $D \in \mathcal{D} \cap \mathcal{A}$ faire
6: $\mathcal{H}' \leftarrow \mathcal{H}' \cup \{(D, \widehat{D}) / (D, \widehat{D}) \in \mathcal{H}\}$
7: fin pour
8: fin si
9: sinon
10: si $\mathcal{D} \neq \phi$ alors
11: pour chaque $D \in \mathcal{D}$ faire
12: pour chaque $A \rightarrow B \in \mathcal{M}$ faire
13: $\mathcal{C}' \leftarrow \{C \ / \ \ A : B \doteq C : D \ \}$
14: $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}'$
15: $\mathcal{T} \leftarrow \mathcal{T} \cup \{(A, B, C, D) \ / \ C \in \mathcal{C}'\}$
16: fin pour
17: fin pour
18: $\mathcal{H}'' \leftarrow$ résultat de l'appel récursif avec $\mathcal{A}$, $\mathcal{M}$, $\widehat{\mathcal{A}}$, $\widehat{\mathcal{M}}$, $\mathcal{H}$, $\mathcal{C} \setminus \mathcal{D}$ et $n - 1$
19: pour chaque $(C, \widehat{C}) \in \mathcal{H}''$ faire
20: $\widehat{\mathcal{C}} \leftarrow \widehat{\mathcal{C}} \cup \{\widehat{C}\}$
21: fin pour
22: pour chaque $\widehat{C} \in \widehat{\mathcal{C}}$ faire
23: pour chaque $A \rightarrow B \in \mathcal{M}$ faire
24: si $(A, B, C, D) \in \mathcal{T}$ alors
25: pour chaque $(\widehat{A}, \widehat{B}) \in \mathcal{H}(A \rightarrow B)$ faire
26: si $\widehat{A} \rightarrow \widehat{B} \in \widehat{\mathcal{M}}$ alors
27: $\mathcal{H}' \leftarrow \mathcal{H}' \cup \{(D, \widehat{D}) \ / \ \ \widehat{A} : \widehat{B} \doteq \widehat{C} : \widehat{D} \ \}$
28: fin si
29: fin pour
30: fin si
31: fin pour
32: fin pour
33: fin si
34: retourner $\mathcal{H}'$
35: fin si

Partie IV

Illustrations et expériences

Chapitre 10

Analogie seule

L'application pratique de l'algorithme de résolution d'équations analogiques que nous proposons concerne la morphologie flexionnelle. Ce n'est que juste retour des choses. Nous avons vu dans nos rappels historiques, que l'analogie en grammaire avait émergé de l'observation de la régularité de la morphologie grecque et latine (p. 43). Il s'agira ici de conjuguer les verbes français aux temps simples ou de décliner les substantifs allemands.

10.1 Conjugaison des verbes français

Les méthodes du traitement automatique des langues en morphologie, où des listages extensifs de toutes les formes possibles sont utilisés sous forme d'automates, ont notoirement négligé l'analogie. Or, notre algorithme pour la vérification et la résolution d'analogies entre mots nous permet de réaliser l'automatisation de la méthode d'enseignement des grammairiens médiévaux : la conjugaison par simple listage de tables[1].

La présentation de tables de conjugaison ou de déclinaison aux commençants dans une langue est maintenant une pratique plus que commune[2]. Ces tables ne sont pas seulement communes dans les grammaires pour étrangers, mais bien aussi dans les dictionnaires pour les locuteurs de la langue en question[3]. Par exemple, sur le français, intéressons-nous au seul paradigme d'inflexions sérieux, celui des verbes. La figure 10.1 (p. 246) montre un extrait de la page VII du *Petit Larousse en couleurs*, (1986), l'un des dictionnaires français les plus communément utilisés. Cet extrait est le commencement de la table de la conjugaison du verbe *aimer*. On sait que la conjugaison des verbes français est assez compliquée, par conséquent, un certain nombre d'autres tables semblables sont listées sur un certain nombre de pages du même dic-

[1] Nous reprenons ici en le modifiant l'article : Lepage, *Analogy + Tables = Conjugation*, 1999.

[2] Elle est toujours pratiquée dans l'enseignement du latin, du grec ou de l'allemand en France, et notre expérience personnelle nous l'a rappelée lors de notre apprentissage du japonais.

[3] Voir par exemple les douze pages de tables utilisées dans l'enseignement public japonais dressées dans 小川 芳男 (Ogawa Yosio) *et al.*, 日本語教育辞典, 1988, p. 241–254.

INDICATIF présent			SUBJONCTIF présent		
j'	*aim*	*e*	*que j'*	*aim*	*e*
tu	*aim*	*es*	*que tu*	*aim*	*es*
il	*aim*	*e*	*qu' il*	*aim*	*e*
nous	*aim*	*ons*	*que nous*	*aim*	*ions*
vous	*aim*	*ez*	*que vous*	*aim*	*iez*
ils	*aim*	*ent*	*qu' ils*	*aim*	*ent*
⋮			⋮		

Figure 10.1: Extrait de tables de conjugaison des verbes français

tionnaire. Chacune porte un numéro auquel il est fait référence dans les entrées du dictionnaire. Par exemple, l'entrée *percevoir* mentionne *conj. 29*, parce que ce verbe se conforme à la conjugaison de la table 29, celle du verbe *décevoir*. Il est aussi intéressant de noter que l'entrée d'un verbe comme *oublier* ne fait référence à aucune table, parce que ce verbe suit la conjugaison du verbe le plus régulier *aimer*. Il s'agit d'un paradigme par défaut.

Pour conjuguer un verbe, par exemple, *oublier*, il est donc nécessaire de se référer à une table et d'appliquer le raisonnement suivant :

> « les formes conjugées du verbe *oublier* sont à *oublier*
> ce que *j'aime*, *tu aimes*, ..., sont à *aimer*. »

Nous sommes tous capables de produire, de tête, les formes conjuguées *j'oublie*, *tu oublies*, ..., en résolvant des équations analogiques[4].

$$\begin{aligned} \textit{aimer} : \textit{j'aime} \doteq \textit{oublier} : x \quad &\Rightarrow \quad x = \textit{j'oublie} \\ \textit{aimer} : \textit{tu aimes} \doteq \textit{oublier} : x \quad &\Rightarrow \quad x = \textit{tu oublies} \\ &\vdots \end{aligned}$$

Puisque nous disposons (presque) d'un algorithme de résolution d'équations analogiques, nous sommes en mesure de réaliser l'opération précédente sur une machine[5]. À cette fin, nous avons dû recueillir des données de conjugaison. Les nôtres proviennent de la section de conjugaison du *Dictionnaire standard japonais-français* des éditions Taisyuukan[6]. Elles consistent en deux parties conformément à l'idée intuitive de la conjugaison par analogie :

[4]Rappelons que Pullum disait que nous « analogisons », (p. 72) et que Mounin parlait de la « quatrième proportionelle » à la suite de Saussure (p. 71).

[5]Ce programme datant un peu, il utilise un vieil algorithme pour la résolution des équations analogiques. Il s'agit de celui présenté dans Lepage, *Solving analogies on words: an algorithm*, 1998.

[6]鈴木 信太郎 (監修) (sous la dir. de Suzuki Sintarou), *Dictionnaire standard japonais-français* – スタンダード和仏辞典, 1991 1e ed 1970.

- une liste de tables exactement semblables à celle donnée dans la figure 10.1 ;
- et une liste de verbes avec leur modèle de conjugaison, c'est-à-dire, un autre verbe (voir la figure 10.2).

Un formulaire d'interface écrit en HTML permet à l'utilisateur d'entrer un verbe (zone d'entrée libre) et de choisir tous les paramètres pertinents pour la conjugaison des verbes français : la personne (première, seconde ou troisième), le nombre (singulier ou pluriel), le mode et le temps (voir figure 10.3, p. 248). Le verbe entré est recherché dans une liste de verbes avec leur modèle. S'il ne peut être trouvé dans cette liste, un modèle par défaut est sélectionné. Ici, pour le cas particulier de la conjugaison des verbes français, il y a deux modèles par défaut selon que l'infinitif se termine par *er* ou *ir*. Sur l'exemple donné par la figure 10.3 (p. 248), le modèle trouvé pour *percevoir* est *recevoir*. Les paramètres sélectionnent la forme conjuguée dans la table de conjugaison du modèle par un simple indexage dans cette table. Par exemple, les paramètres troisième personne singulier indicatif futur simple sélectionnent : *il ou elle recevra*.

⋮		⋮
pelleter	...	*jeter*
pendre	...	*entendre*
pénétrer	...	*préférer*
percer	...	*avancer*
percevoir	...	*recevoir*
perdre	...	*perdre*
périr	...	*finir*
permettre	...	*mettre*
persévérer	...	*préférer*
peser	...	*lever*
⋮		⋮

Figure 10.2: Extrait à la lettre *p* de la liste de verbes français avec leurs modèles

La forme conjuguée produite par le programme et retournée à l'utilisateur, est la solution de l'équation analogique obtenue par une simple application de l'algorithme de résolution d'équations analogiques :

$$recevoir : il\ ou\ elle\ recevra \doteq percevoir : x \quad \Rightarrow \quad x = il\ ou\ elle\ percevra$$

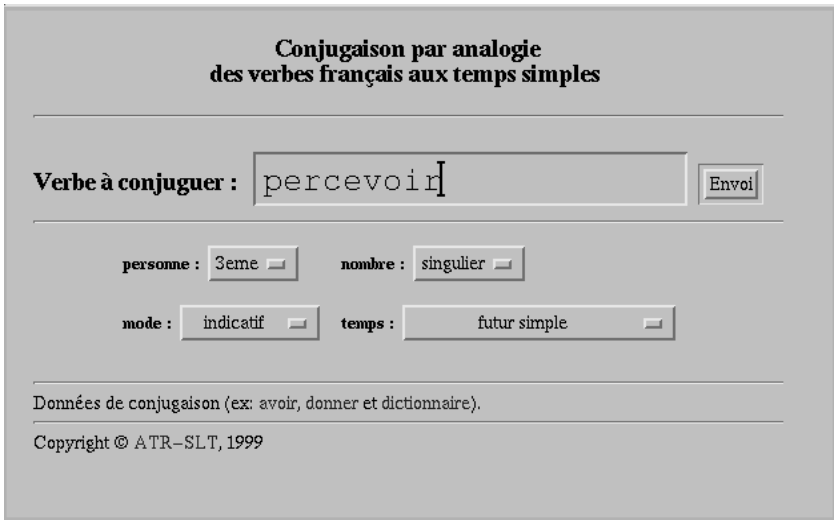


Figure 10.3: Formulaire de conjugaison par analogie des verbes français

Un tel système, s'il remplit bien sa tâche soulève cependant un problème particulier que nous commentons maintenant. Considérons l'équation analogique *zébrer* : *je zèbre* ≐ *préférer* : *x*. Formellement, deux solutions existent : *x* = *je prèfére* et *x* =*je préfère*. Dans le système du français, c'est la forme *je préfère* qui est correcte. On peut dire qu'ici le français « préfère la droite. » Or on peut trouver des exemples d'autres langues, où « la gauche sera préférée. » La question est donc, plus que de savoir pourquoi cela est possible, de trouver une simulation satisfaisante du phénomène. Nous pensons que ce n'est pas à l'algorithme de résolution d'analogies de décider quelle est la bonne solution. La réponse doit, selon nous, provenir du système de la langue. Notre conception de la langue est une conception où les formes attestées sont centrales. Or, il suffit d'avoir un nombre restreint de formes attestées pour résoudre correctement le problème. Supposons que le couple suivant soit connu : *célébrer* et *je célèbre*. Ce couple, dans un modèle d'apprentissage peut être interprété comme une forme entendue, donc attestée. En insérant cette connaissance dans le système, le nombre de fois où *je préfère* est obtenu sur l'ensemble des solutions obtenues lors de la résolution des deux analogies devient majoritaire sur l'ensemble des solutions possibles :

zébrer : *je zèbre* ≐ *préférer* : *je préfère* ou *je prèfére*
célébrer : *je célèbre* ≐ *préférer* : *je préfère*

Dans une telle vue, la décision de retenir une forme plutôt qu'une autre provient de la fréquence avec laquelle cette forme peut être produite dans le système de la langue reposant sur les formes attestées. Nous retrouvons ici notre préoccupation sur le rôle des fréquences à la suite des travaux de Mańczak (p. 68). Il est intéressant de noter que, dans le cas particulier de la forme *je préfère*, le résultat ne changera pas si l'on ajoute d'autres formes exactes contenant même éventuellement plus de deux *é*, aux formes attestées.

10.2 Déclinaison des substantifs allemands

Nous avons appliqué la méthode précédente à la déclinaison des substantifs allemands. Il s'agit exactement de la même procédure. Le formulaire d'interface seul a changé entre les deux applications. Dans cette nouvelle application, l'utilisateur se sert de la zone d'entrée libre pour saisir un substantif à décliner. Les paramètres sont évidemment différents, puisqu'il s'agit maintenant de choisir le cas et le nombre seulement (voir figure 10.6, p. 251). Chaque substantif allemand du dictionnaire s'est vu auparavant attribué un modèle de déclinaison et chaque modèle de déclinaison est listé dans des tables. Nous donnons quelques exemples de tables de déclinaisons (voir figure 10.5) ainsi que des exemples de substantifs avec leur modèle (voir figure 10.4). Pour le reste, les deux systèmes ne diffèrent en rien et utilisent les mêmes programmes.

⋮		⋮
Entschuldigung	...	*Präfektur*
Entwicklung	...	*Präfektur*
Entwurf	...	*Ratschlag*
Epidemie	...	*Tinte*
Epoche	...	*Tinte*
Eröffnung	...	*Präfektur*
Erbse	...	*Tinte*
Erdbeere	...	*Tinte*
Erdbeertorte	...	*Tinte*
Erdgeschoß	...	*Fangnetz*
Erdnuß	...	*Unterkunft*
Erfahrung	...	*Präfektur*
Erfolg	...	*Auslieferungstag*
Ergebnis	...	*Gedächtnis*
⋮		⋮

Figure 10.4: Extrait à la lettre E de la liste de substantifs allemands avec leurs modèles

	Singular	Plural
⋮	⋮	⋮
Nominativ	*Katholizismus*	–
Genitiv	*Katholizismus*	–
Dativ	*Katholizismus*	–
Akkusativ	*Katholizismus*	–
Nominativ	*Km/H*	*Km/H*
Genitiv	*Km/H*	*Km/H*
Dativ	*Km/H*	*Km/H*
Akkusativ	*Km/H*	*Km/H*
Nominativ	*Kompromiß*	*Kompromisse*
Genitiv	*Kompromisses*	*Kompromisse*
Dativ	*Kompromiß*	*Kompromissen*
Akkusativ	*Kompromiß*	*Kompromisse*
Nominativ	*Loch*	*Lochs*
Genitiv	*Loch*	*Lochs*
Dativ	*Loch*	*Lochs*
Akkusativ	*Loch*	*Lochs*
⋮	⋮	⋮

Figure 10.5: Tables des déclinaisons des substantifs allemands

Figure 10.6: Formulaire de conjugaison par analogie des substantifs allemands

10.3 Synthèse des avantages de la méthode proposée

Notre méthode de conjugaison ou de déclinaison par analogie suit l'idée intuitive de la conjugaison ou de la déclinaison par modèles et par tables. Elle se rapproche de la façon naturelle dont les gens décrivent leur propre langue, façon qui semble bien prendre appui sur la faculté de tous à résoudre des équations analogiques. Cette méthode intuitive ne pouvait donc être mise en œuvre sur machine tant que nous ne disposions pas au moins d'un début d'algorithme de résolution d'équations analogiques entre chaînes de symboles.

Si la méthode est intuitive dans sa conception, elle est aussi particulièrement simple dans sa réalisation. Elle permet en effet une exploitation directe de ressources usuelles, c'est-à-dire accessibles à tous et disponibles n'importe où. Par là-même, elle réduit donc considérablement la charge dans l'acquisition des données. En effet, l'implémentation d'un système de conjugaison des verbes français ou de déclinaison des substantifs allemands se réduit presque à la pure et simple recopie de tables de modèles et de listes de mots avec leur modèle associé. Cette tâche, facile jusqu'à l'ennui, est particulièrement économique, puisqu'elle est à la portée de tout le monde. Nous réalisons donc là l'un des espoirs que nous mettions dans l'analogie, à savoir celui du moindre effort en traitement automatique des langues (voir p. 94). Soulignons que la méthode proposée ici ne présente pas d'étapes de travail linguistique sur les données, étape qui est nécessaire et particulièrement coûteuse dans les approches utilisant des automates, par exemple.

Un autre avantage de notre méthode est qu'elle simule la façon de penser par règles et exceptions. La productivité de la langue est simulée par un mécanisme de cas par défaut. Pour la conjugaison des verbes français, les verbes qui suivent les modèles par défaut n'ont pas besoin d'être indexés dans la liste des verbes et modèles. Par conséquent, notre système conjuguera correctement le verbe populaire *biduler* qui se trouve rarement dans tous les dictionnaires.

En prenant deux exemples dans deux langues différentes, nous avons montré que la méthode est générale. Nous envisageons depuis longtemps de l'appliquer à la conjugaison des verbes arabes[7] et à leur système de dérivation morphologique. Le temps nous a jusqu'à présent manqué. L'intérêt n'est pas d'avoir un n-ième exemple d'application de l'analogie à la morphologie flexionnelle, mais bien de travailler l'une des ses propriété majeures, l'infixation multiple (voir p. 48). Ce phénomène n'a pu être pris en compte que récemment dans le cadre de la morphologie à deux niveaux, qui utilise des transducteurs à nombre d'états finis[8], moyennant une extension du modèle. Or, le phénomène d'infixation multiple fait partie fondamentalement des possibilités de l'analogie. Une telle application aurait donc constitué, une fois réalisée, une bonne démonstration de la puissance formelle de l'analogie.

[7]Les données nous ont été gracieusement fournies par Fathi Debili, auquel nous adressons nos remerciements.

[8]Voir Beesley, *Consonant spreading in Arabic stems*, 1998.

Enfin, si les deux exemples traités ci-dessus relevaient de la morphologie inflexionnelle, on peut envisager d'étendre les applications à la morphologie dérivationnelle d'une langue, si tant est qu'elle est régulière. Le finlandais ou l'espéranto pourraient peut-être constituer des candidats pour de tels exemples d'application. Et le désir est grand, alors, d'étendre l'application à la syntaxe en manipulant des modèles de phrases, dont on ferait jouer un certain nombre d'éléments variables, tels que le temps, le nombre du sujet, etc., voire même le verbe lui-même, le sujet, etc. renouant ainsi plus ou moins avec les vues de la syntaxe transformationnelle (voir p. 170).

En résumé de ces exemples d'application de l'analogie, permettons nous un rapprochement avec l'une des formules célèbres de l'histoire de la programmation en informatique. De la même façon que Niklaus WIRTH, concepteur du langage de programmation Pascal, avait pu écrire

$$\textit{algorithmes} + \textit{structures de données} = \textit{programmes},$$

nous venons de montrer que le traitement automatique des langues pourrait aussi écrire

$$\textit{algorithme d'analogie} + \textit{tables} = \textit{programmes de morphologie flexionnelle}.$$

Chapitre 11

Corpus et langages de chaînes analogiques

L'hypothèse selon laquelle les langues seraient représentables par des langages formels établit le lien entre théorie des langages et linguistique. Après avoir donc fait une proposition théorique, celle des langages de chaînes analogiques, il nous faut maintenant nous appliquer à étudier l'adéquation de la théorie à la pratique. Nous nous proposons, dans ce chapitre, de faire le lien entre ces nouveaux langages formels et des données de langue que sont les corpus.

Nous avons vu (p. 169) qu'un langage de chaînes analogiques était entièrement défini par la donnée d'un ensemble $\mathcal{A}$ de formes attestées et d'un ensemble $\mathcal{M}$ de modèles. Nous avons vu aussi que l'ensemble des éléments d'un langage de chaînes analogiques peut être produit par une sorte d'expansion successive de $\mathcal{A}$.

$$\begin{aligned}
&\Lambda_0 = \mathcal{A} \\
&\Lambda_1 = \{D \in \mathcal{V}^* \;/\; \exists C \in \Lambda_0,\ \exists A \to B \in \mathcal{M}, \quad A : B \doteq C : D \ \} \\
&\quad\vdots \\
&\Lambda_{i+1} = \{D \in \mathcal{V}^* \;/\; \exists C \in \Lambda_i,\ \exists A \to B \in \mathcal{M}, \quad A : B \doteq C : D \ \} \\
&\quad\vdots
\end{aligned}$$

Le langage $\Lambda(\mathcal{A}, \mathcal{M})$ est défini par :

$$\Lambda(\mathcal{A}, \mathcal{M}) = \bigcup_{i=0}^{\infty} \Lambda_i$$

Si l'on fait l'hypothèse qu'une langue ou, à tout le moins, un sous-langage d'une langue, est reflété par la donnée d'un corpus, il nous faudrait donc, étant donné un tel corpus, premièrement chercher à caractériser ce que serait l'ensemble $\mathcal{A}$ des formes attestées, deuxièmement chercher à caractériser l'ensemble $\mathcal{M}$ des modèles, et enfin montrer quels ensembles Λ_i peuvent être obtenus à partir des deux ensembles précédents. Nous ne présenterons ici que deux expériences reflétant très imparfaitement ce programme.

Dans la pratique, pour ce qui est de l'application aux langues, la situation se présente de la façon suivante. Il n'y a pas de séparation entre $\mathcal{A}$ et

$\mathcal{M}$. La donnée primordiale est un ensemble de phrases à considérer comme $\mathcal{A}$. L'ensemble des modèles est alors l'ensemble de tous les couples possibles d'éléments de $\mathcal{A}$. Autrement dit, pour les langues, on a $\mathcal{M} = \mathcal{A} \times \mathcal{A} = \mathcal{A}^2$. Ceci est la définition d'un langage de chaînes analogiques paresseux (p. 174).

Comme conséquence de cette vue, les deux premiers points concernant la détermination de $\mathcal{A}$ et de $\mathcal{M}$ seront donc abordés simultanément. Partant d'un corpus donné, nous essaierons d'obtenir un ensemble $\mathcal{A}$ à partir duquel on puisse produire à nouveau les éléments du corpus de départ comme éléments du langage formel $\Lambda(\mathcal{A}, \mathcal{A}^2)$. Par facilité, on s'intéressera seulement à produire à nouveau les phrases du corpus en une seule analogie avec des éléments de l'ensemble obtenu. Autrement dit, le corpus sera vu comme l'ensemble Λ_1. Nous retrouvons ici ce que nous annoncions plus haut (p. 188).

Pour ce qui est du troisième point, il s'agirait normalement de produire le langage en entier. Autrement dit, il faudrait produire tous les Λ_i, quel que soit i. Nous nous limiterons à ne produire que Λ_1. De cette façon, ce problème et le précédent apparaissent comme réciproques l'un de l'autre. Dans le premier cas, il s'agit de réduire un corpus en une passe, alors que dans le second cas, on l'augmente, en une seule passe aussi. Le premier cas opère donc une *contraction* du corpus, alors que le second en opère une *expansion*.

Toutes les notions précédentes se sont petit à petit dégagées de nos travaux à partir de la notion de repésentativité de corpus. Nous avions d'abord essayé de caractériser la notion par le voisinage de langages formels. Le problème est le suivant : étant donné un langage formel d'une famille donnée, celle des langages réguliers, hors-contexte ou sous-contexte, à quelle famille appartient le langage formé de l'ensemble des chaînes situées à une distance donnée des éléments du langage de départ[1] ? Nous avions ensuite reprécisé et adapté cette idée en la rattachant à la notion d'analogie, et nous avions alors proposé les deux mesures de *densité* et de *prégnance*[2] dont nous ferons mention plus bas.

[1] LEPAGE, *Regular languages have regular neighbourhoods for the Wagner and Fischer distance*, 1994, p. 19, § 4, représentativité de corpus.

[2] LEPAGE, *Un éditeur pour la construction de banques d'arbres*, 1996, p. 109.

11.1 Ensemble représentatif d'un corpus

Les deux opérations précédemment mentionnées de contraction et d'expansion sont liées à deux notions connues en linguistique. Pour ce qui est du premier cas, il s'agit de la *représentativité* d'un ensemble de phrases, fameux serpent de mer de la linguistique : comment extraire d'un corpus donné les phrases représentatives de l'ensemble du corpus ? On peut alors s'intéresser aux tailles comparées du corpus de départ et de l'ensemble obtenu. Cette comparaison caractérise la représentativité de l'ensemble obtenu. L'opération que nous allons réaliser consiste donc à extraire d'un corpus les phrases pertinentes du point de vue de l'analogie. Quelques auteurs n'hésitent pas à appeler *résumé automatique*, l'opération d'extraction de phrases pertinentes à partir d'un texte donné. Ils arguent du fait qu'effectivement, dans certaines tâches de résumé, on demande seulement d'extraire les phrases les plus significatives du texte. Nous nous contentons d'appeler cette opération de la *contraction de corpus par extraction de phrases.*

Les expériences que nous allons rapporter ici datent de plusieurs années. Elles reflètent un état encore primitif de notre recherche. À l'époque, nous utilisions une explication différente de l'analogie. Nous avions proposé, à partir de l'examen d'un certain nombre d'exemples seulement, l'explication suivante de l'analogie. Nous remplaçons le signe $\doteq$ par le signe $\approx$ pour bien marquer la différence.

$$A : B \approx C : D \quad \Leftrightarrow \quad \begin{cases} \delta'(A,B) & = & \delta'(C,D) \\ \delta'(A,C) & = & \delta'(B,D) \\ \delta'(A,D) & = & \delta'(B,C) \end{cases}$$

où $\delta'(A,B)$ désignait la distance de Wagner et Fischer, c'est-à-dire la distance d'édition équipée du remplacement, avec le même poids que l'insertion et la suppression. Par exemple, $\mathit{statisticien} : \mathit{statistique} \doteq \mathit{physicien} : \mathit{physique}$ serait justifié par le système suivant.

$$\begin{cases} \delta'(\mathit{statisticien}, \mathit{statistique}) & = & 3 & = & \delta'(\mathit{physicien}, \mathit{physique}) & = & 3 \\ \delta'(\mathit{statisticien}, \mathit{physicien}) & = & 6 & = & \delta'(\mathit{statistique}, \mathit{physique}) & = & 6 \\ \delta'(\mathit{statisticien}, \mathit{physique}) & = & 9 & = & \delta'(\mathit{statistique}, \mathit{physicien}) & = & 9 \end{cases}$$

En d'autres termes, à l'époque, notre vision a priori de l'analogie était celle d'un vrai rectangle (voir la figure 11.1). Nous appellerons donc cette explication imparfaite celle du système du rectangle. Comme nous savons que cette explication n'est pas valable, et afin de bien marquer la différence d'avec nos recherches sur la vraie explication de l'analogie entre chaînes de symboles, nous appellerons *quasi-analogies* des quadruplets vérifiant le système du rectangle.

Faisons ici une digression personnelle. Notre travail sur la formalisation de l'analogie a connu un rebondissement avec les expériences utilisant le système du rectangle. À l'origine, nous nous étions étonné de l'absence notoire de l'analogie dans le traitement automatique des langues, Nous avions alors proposé, après un certain temps de réflexion tout de même, l'explication simple du rectangle pour rendre compte de l'analogie. Mais les expériences que nous

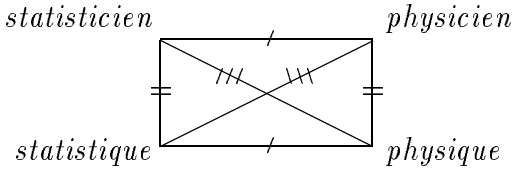


Figure 11.1: Rectangle de l'analogie (explication très imparfaite et obsolète)

allons maintenant rapporter nous ont laissé un goût de curiosité insatisfaite. D'un côté, nous pensions avoir mis la main sur quelque chose d'intéressant, mais d'un autre, nous découvrions que l'explication par le système du rectangle avait été posée un peu vite. Les expériences nous ont livré un trop grand nombre de quasi-analogies qui n'étaient pas des analogies au niveau des chaînes de symboles. Leur examen nous a permis de construire ou d'épingler des exemples typiques. L'analogie en arabe $aslama : muslim \doteq arsala : mursil$ [3], déjà cité en introduction, en est un. Elle ne vérifie pas le système du rectangle

$$\begin{cases} \delta'(aslama, muslim) = 4 = \delta'(arsala, mursil) = 4 \\ \delta'(aslama, arsala) = 3 = \delta'(muslim, mursil) = 3 \\ \delta'(aslama, mursil) = 6 \neq \delta'(muslim, arsala) = 5 \end{cases}$$

alors que, de façon étonnante, $aslama : muslimun \doteq arsala : mursilun$, très semblable, y répond.

$$\begin{cases} \delta'(aslama, muslim) = 5 = \delta'(arsala, mursil) = 5 \\ \delta'(aslama, arsala) = 3 = \delta'(muslim, mursil) = 3 \\ \delta'(aslama, mursil) = 7 = \delta'(muslim, arsala) = 7 \end{cases}$$

Tout cela nous a insensiblement poussé a examiner d'autres exemples, dans de nombreuses langues, puis à passer à l'examen des analogies formelles. Bref, nous nous laissions embarquer dans la recherche d'une explication formelle, la vraie, de l'analogie entre chaînes de symboles.

11.1.1 Densité d'un corpus

Pour revenir à la représentativité et à l'extraction, nous pouvons d'abord essayer de caractériser a priori et intrinséquemment un corpus en termes de nombre d'analogies ou de quasi-analogies existant en son sein. Les analogies faisant intervenir nécessairement quatre objets, nous proposons d'appeler *densité analogique* ou *quasi-analogique* d'un corpus le rapport entre le nombre de quadruplets de phrases en relation d'analogie ou de quasi-analogie et le nombre théorique total de quadruplets.

Dans une expérience préliminaire[4], nous avons calculé toutes les distan-

[3] *Arsala* (il envoya) et *aslama* (il se convertit [à l'Islam]) sont des verbes au passé, 3ème personne du singulier. *Mursil* (un envoyé) et *muslim* (un converti [à l'Islam], c'est-à-dire un musulman) sont les formes nominales usuelles correspondant à l'arabe classique *mursilun* et *muslimun*.

[4] LEPAGE, *Un éditeur pour la construction de banques d'arbres*, 1996, p. 109.

ces entre phrases (respectivement entre arbres) d'un même corpus arboré. Le corpus arboré utilisé était un extrait du corpus arboré de l'université de Pennsylvanie, comprenant 787 phrases, chacune avec son arbre correspondant. Théoriquement, il peut exister $787 \times 786 \times 785 \times 784$ quadruplets. Un certain nombre seulement d'entre eux vérifie la relation de quasi-analogie. Pour les phrases, nous avons considéré les mots comme les symboles élémentaires. Cela fait des phrases des chaînes de mots. Pour les arbres, nous avons calculé les distances sur la forme parenthésée des arbres, en considérant les étiquettes de nœuds comme des symboles élémentaires. Avec cela, il était facile d'appliquer une procédure automatique pour trouver tous les quadruplets vérifiant la relation de quasi-analogie.

Les densités obtenues sont de 0,1% pour les phrases comme pour les arbres. Il est difficile de savoir comment interpréter ces chiffres. À cette fin, il serait nécessaire de refaire ces mesures premièrement avec notre nouvelle définition, même incomplète, de l'analogie, et deuxièmement avec les nouveaux corpus, bien plus gros, dont nous disposons maintenant. En tout cas, nous pensons avoir proposé une mesure formelle intéressante.

11.1.2 Taille des bases ou cardinal de $\mathcal{A}$

Nous nous tournons maintenant vers une seconde mesure qui est le rapport des cardinaux de deux ensembles, l'ensemble des phrases représentatives du corpus de départ et ce corpus. Nous avons vu précédemment (p. 189) que la notion de base, c'est-à-dire de générateur et libre, pouvait rendre compte de la notion de représentativité. Des définitions, nous avions dérivé un algorithme de construction des bases par élimination de phrases au fur et à mesure de la lecture du corpus (p. 236). C'est cette algorithme que nous utilisons ici pour construire des ensembles représentatifs[5].

Conditions expérimentales

Les données de nos expériences sont une collection de textes extraits du corpus d'ATR–Lancaster qui rassemble des pages de l'internet, des rapports économiques ou médicaux, des descriptions géographiques, voire des envois à des forums de discussions sur l'internet, etc. Notre extrait, que nous appelons Λ_1 pour nous mettre en conformité avec les notations introduites plus haut (p. 256), représente 10 000 phrases dans lesquelles figurent environ 27 000 mots différents.

Contraintes sur la quasi-analogie

L'application de l'algorithme de construction de base par élimination de chaînes au fur et à mesure de la lecture permet de recueillir les tailles successives de la base obtenue, au fil du corpus. Ces tailles sont données dans la

[5]Cette partie reprend LEPAGE, *Corpus contraction by sentence extraction using analogy*, 1997.

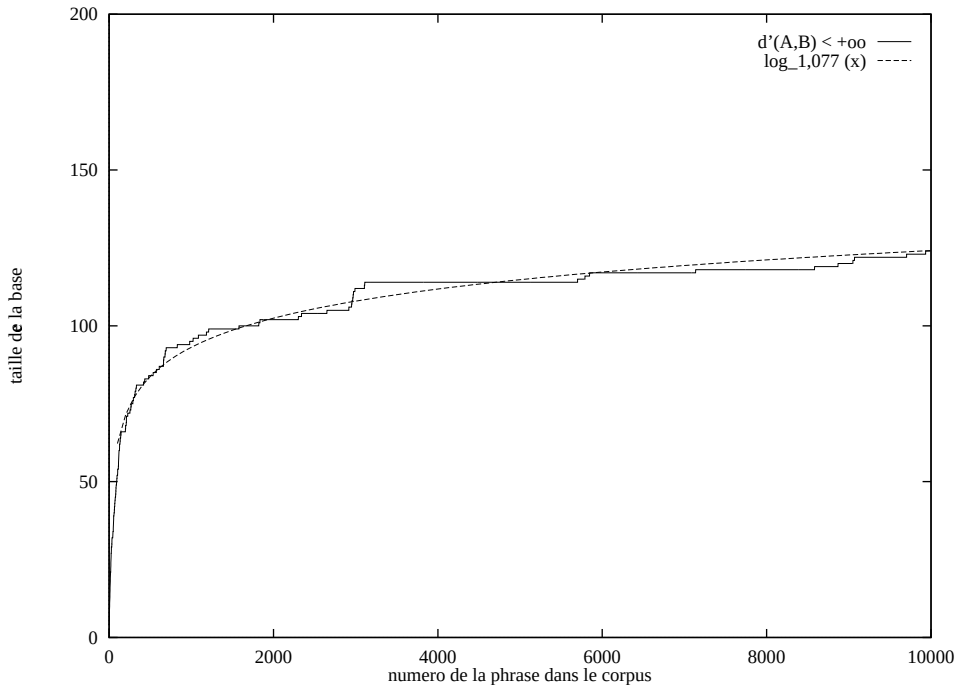


Figure 11.2: Taille de la base au fil de la lecture (quasi-analogies, aucune contrainte)

figure 11.2 (p. 260). La courbe peut être raisonnablement approximée par un logarithme en base $1,077$. On observe les mêmes courbes sur les textes bruts et sur une représentation de ces mêmes textes où les mots ont été remplacés par leur catégories morpho-syntaxiques fines, c'est-à-dire par des étiquettes choisies parmi un millier de catégories. Le graphe de la figure 11.3 (p. 261) donne l'évolution de la taille des bases pour la moitié du corpus seulement, c'est-à-dire 5 000 phrases. L'ordre des phrases ne semble pas non plus influencer sensiblement la taille des bases obtenues. Lorsque l'on applique des permutations au hasard sur les phrases du corpus, ces tailles restent dans l'intervalle de 110 à 130. Le graphe de la figure 11.4 (p. 261) donne différentes tailles de base obtenues pour différentes permutations de phrases. Bien sûr, quelques permutations ne représentent pas l'ensemble, gigantesque !, de toutes les permutations possibles, mais l'on peut toujours espérer que ces tests soient significatifs[6].

Tous ces résultats sembleraient encourageants puisqu'un corpus représentatif au sens de la quasi-analogie serait de taille logarithmique en la taille réelle du corpus. Malheureusement, l'examen des quasi-analogies montre que le système du rectangle est bien trop lâche. Un grand nombre des quasi-analogies detectées ne sont pas des analogies et ne sont donc pas

[6]Figures extraites de LEPAGE *et al.*, *The snow-ball effect of analogy*, 1997, p. 292.

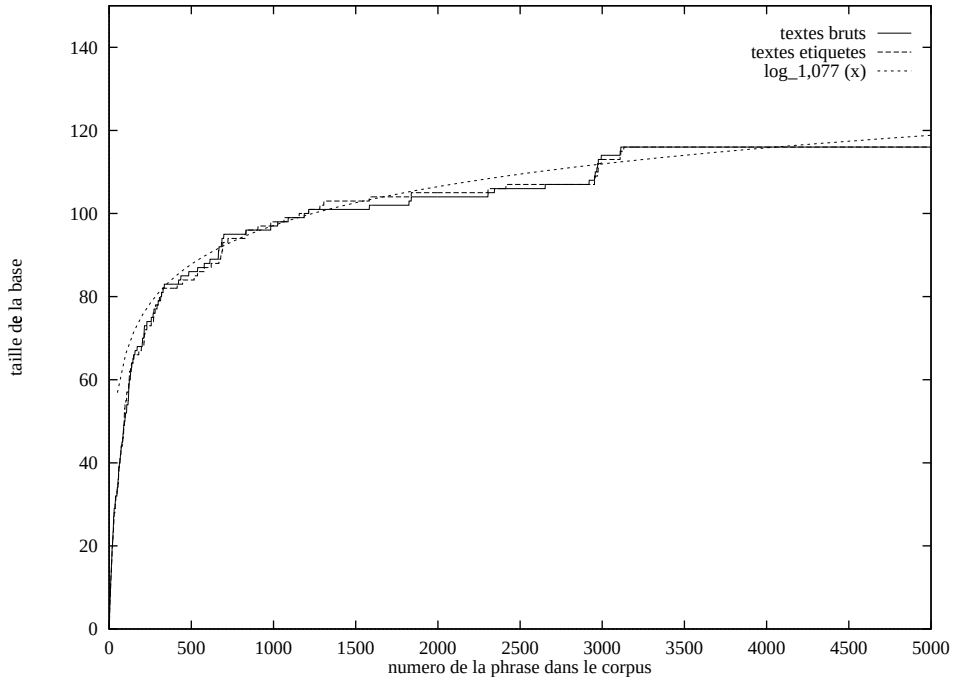


Figure 11.3: Taille de la base au fil de la lecture pour les textes bruts (mots) et pour les textes étiquetés

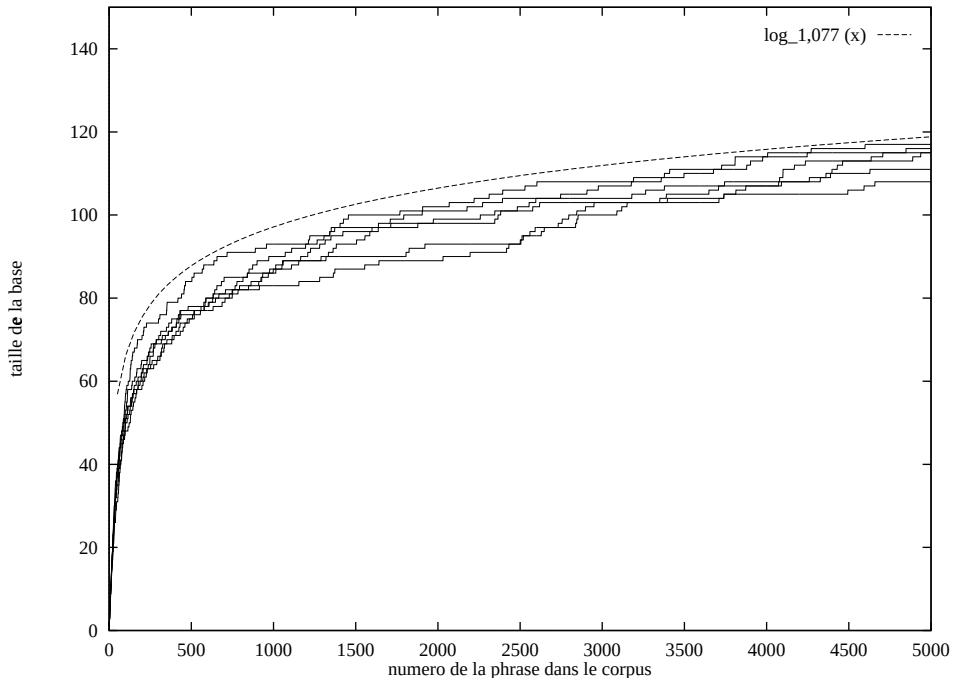


Figure 11.4: Taille de la base au fil de la lecture pour différentes permutations des phrases

pertinentes linguistiquement. Voici par exemple l'une des quasi-analogies détectées.

Getting where you want to go can often be one of the most difficult aspects of using networks.	:	*The variety of ways that places are named will probably leave a blank stare on your face at first.*	≈	*a person's email address on a computer: user@somewhere.domain*	:	*Likewise, researchers from all corners of the earth are finding that their work thrives in a networked environment.*

[7]

Afin de réduire le nombre de quasi-analogies non valides, nous sommes conduits à imposer des contraintes sur le rectangle quasi-analogique. La contrainte la plus sévère possible consisterait à limiter la distance dans le rectangle à 0, c'est-à-dire à réduire la relation de quasi-analogie à une égalité. Cette contrainte a pour seul effet de détecter les phrases répétées, et donc de les éliminer lors de l'application de l'algorithme de construction de base. Notre corpus en a été réduit à 8 271 phrases. Il contenait donc 1 729 répétées. Elles représentent en fait 602 phrases différentes seulement. En voici quelques unes[8] :

Michele Tepper
In article <xx@yy.zz.tt>, <xx@yy.zz.tt> wrote:
November 3, 1994
Date:
1.
2.

L'examen du corpus permet d'expliquer ces répétitions de deux façons. Premièrement, les signatures de courrier, les numérotations, les dates isolées, les titres habituels de chapitres ou sections, etc. sont bien évidemment sujets à répétition. Deuxièmement, le corpus contient des contributions à des forums de discussions dans lesquels, comme on sait, l'habitude d'inclure une contribution précédente est assez répandue.

À partir de la contrainte précédente, nous avons recherché d'autres contraintes qui la généraliseraient. Nous en avons défini trois.

(a) Les distances étant des entiers puisqu'il s'agit d'un nombre d'insertions, de suppressions et de remplacements, appliquer la contrainte précédente, c'est-à-dire imposer $\delta'(A,B) = 0$, revient à imposer $\delta'(A,B) < 1$ sur tous les côtés du rectangle. Au contraire, l'absence de contraintes peut être interprétée comme le fait d'imposer $\delta'(A,B) < +\infty$. Ces deux inégalités peuvent se généraliser naturellement en posant un entier K (donc compris entre 1 et $+\infty$) pour

[7](Arriver là où vous le souhaiter vraiment peut être l'un des aspects les plus pénibles de l'utilisation des réseaux.) : (La diversité des dénominations pour un même endroit va vous laisser pantois.) ≈ (adresse d'un individu sur un ordinateur : utilisateur@quelquepart.domaine) : (De la même façon, les chercheurs de tous les coins du monde découvrent que leur travaux bénéficient d'un environnement en réseau.)

[8](Dans son message <xx@yy.zz.tt>, <xx@yy.zz.tt> déclare :)
(Le 3 novembre 1994)
(Date :)

lequel on définirait la contrainte δ'(A,B) $< K$. On appellera cette contrainte *contrainte absolue.*

(b) Or, le distance maximale entre deux phrases est le maximum de leurs longueurs. On peut donc limiter la distance relativement à cette longueur maximale, c'est-à-dire considérer $k \in]\ 0\ ;\ 1\ ]$ pour lequel

$$\delta'(\mathrm{A,B}) < k \times \max(|A|, |B|)$$

On appellera cette contrainte *contrainte du maximum.*

(c) Parce que des analogies non-valides restent encore même en appliquant cette contrainte, on peut limiter les distances relativement, non pas au maximum, mais au minimum des longueurs des phrases c'est-à-dire considérer des $k \in]\ 0\ ;\ 1\ ]$ pour lesquels

$$\delta'(\mathrm{A,B}) < k \times \min(|A|, |B|)$$

On appellera cette contrainte *contrainte du minimum.*

Nous avons appliqué ces trois familles de contraintes à nos données. Les différentes tailles de bases obtenues au fil de la lecture du corpus, pour différentes valeurs de K et k, sont données dans les figures 11.5, 11.6 et 11.7 (p. 264 et 265).

Pour ce qui concerne la contrainte du maximum, mentionnons que des expériences conduites avec cette contrainte sur beaucoup plus de valeurs de k ont montré que pour $k \leq 0,25$, les courbes sont très similaires à celle obtenue pour $K = 1$. Le corpus n'est réduit de façon significative que pour $0,80 \leq k \leq 1$.

La contrainte du minimum donne de meilleurs résultats. Elle permet la détection d'un plus grand nombre de vraies analogies. Elle permet l'obtention d'une base d'une taille égale à un peu moins d'un tiers de celle du corpus pour $k = 1$, et de presque la moitié pour $k = 0,90$. Toutes les courbes obtenues croissent de façon moins que linéaire. Notons que, pour $k = 1$ et $k = 0,85$, les courbes sont relativement bien approchées par les fonctions $x^{2/3}$ et $x^{4/5}$.

Pour $k > 1$, des expériences complémentaires sur 10 000 phrases de départ exhibent des courbes très similaires. On obtient alors des bases de 200 phrases environ. Cette taille très faible de la base montre que la relation de quasi-analogie est bien trop lâche.

Lois de contraction

Nous venons de donner une vue de la taille de la base au fil de la lecture du corpus. Nous avons tracé un certain nombre de fois le même genre de courbes avec trois familles de contraintes. Ces contraintes étant paramétrées, on peut adopter ces paramètres, K et k, comme point de vue sur les résultats de contraction. Nous allons donc maintenant tracer ceux-ci en fonction de ces paramètres.

Pour la contrainte absolue, c'est le paramètre K qui joue. Dans nos expériences, la taille de la base décroît rapidement lorsque K augmente. La

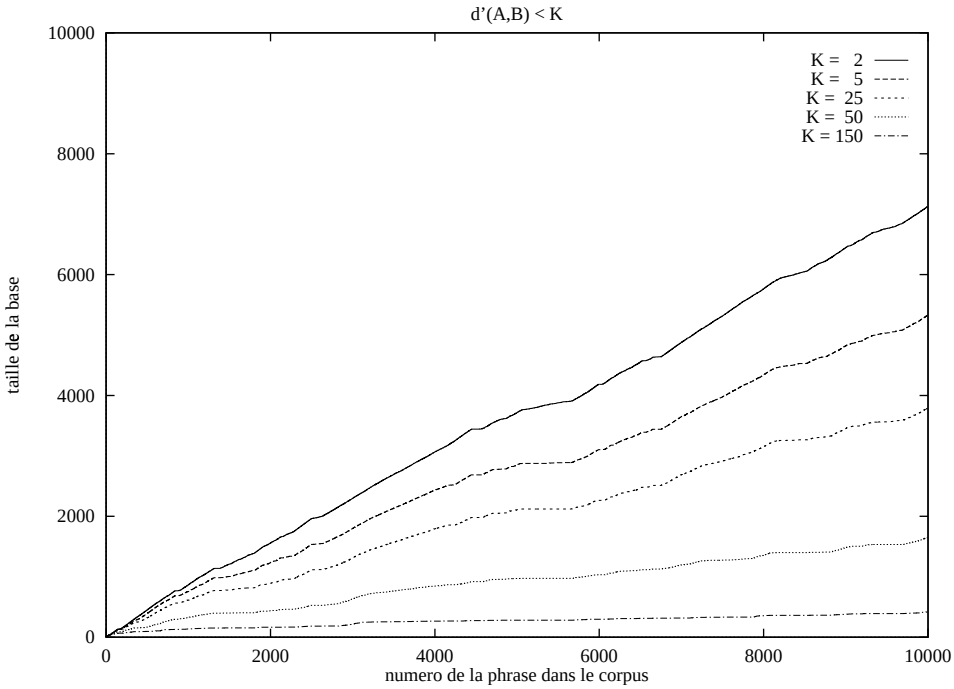


Figure 11.5: Contrainte absolue avec différentes valeurs de K

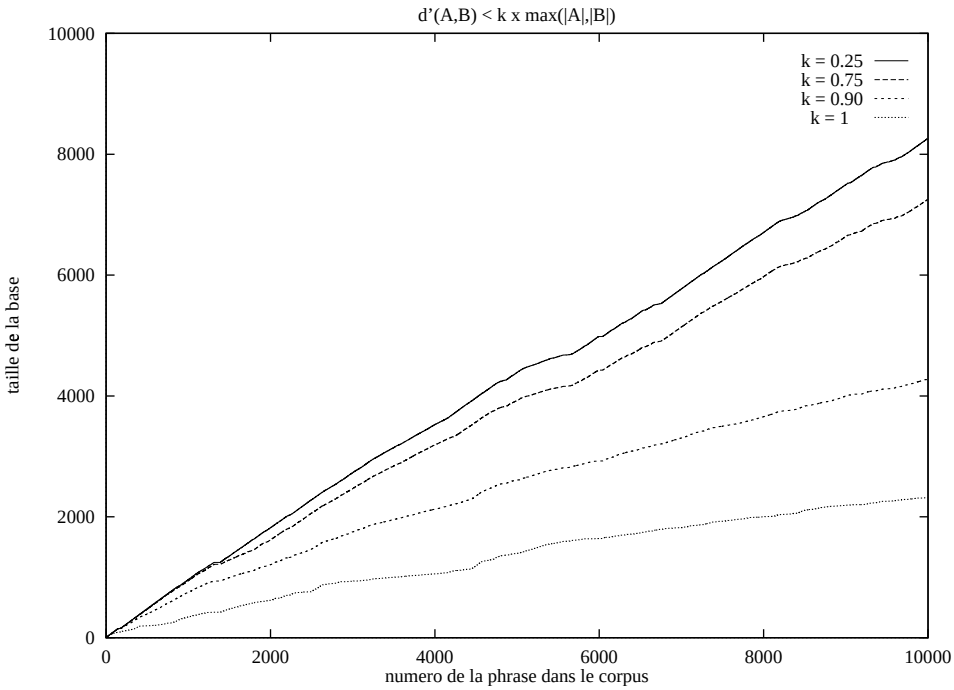


Figure 11.6: Contraintes du maximum avec différentes valeurs de k

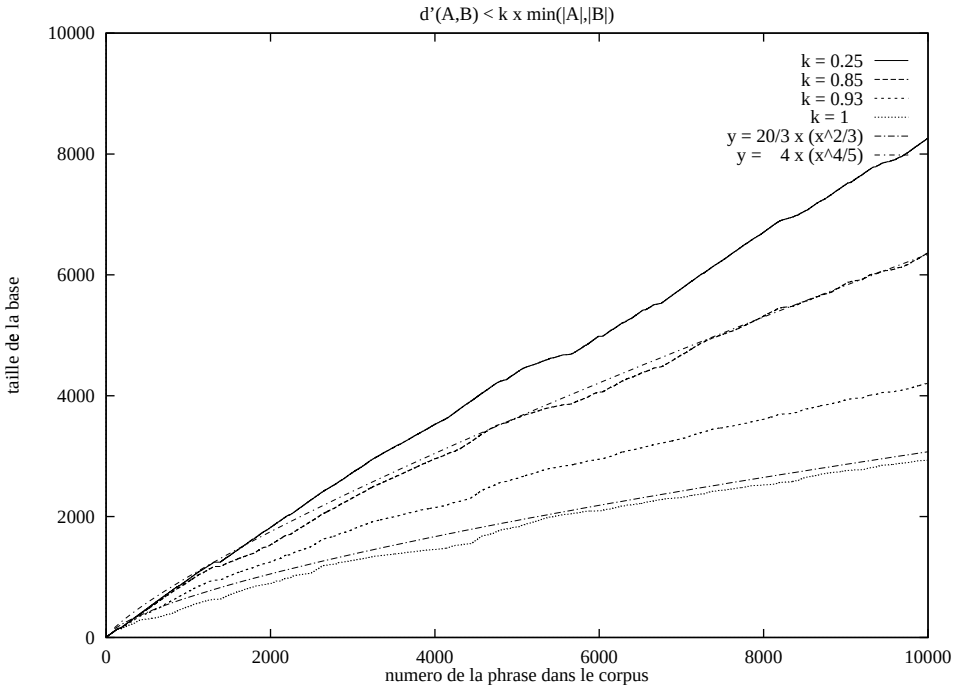


Figure 11.7: Contraintes du minimum avec différentes valeurs de k

figure 11.8 (p. 266) donne les courbes obtenues. En prenant en compte seulement les valeurs de K entre 3 et 33, on peut approcher la courbe par régression linéaire par un segment de droite d'équation général $y = a \times x + b$. Ces coefficients sont listés dans le tableau 11.1.

Tableau 11.1: Coefficients de la régression linéaire pour les valeurs de K comprises entre 3 et 33

numéro de la phrase	a	b
10 000	-139	5628
6 666	-88	3643
3 333	-50	2146
1 000	-17	792

Une deuxième régression linéaire donne la forme générale de a et b, en fonction du numéro r de la phrase dans le corpus : $a(r) = (-13 \times 10^{-3} \times r) - 4$ et $b(r) = 0,5 \times r - 300$. Ce numéro est en fait la taille du corpus lu jusqu'à ce point. On obtient donc une formule générale expérimentale qui approche la taille de la base obtenue avec la contrainte absolue sous les conditions $3 \leq K \leq 33$ et $1 < |\Lambda_1| \leq 10\,000$:

$$(13 \times 10^{-3} \times |\Lambda_1| \times K) - (4 \times K) + (0,5 \times |\Lambda_1|) - 300$$

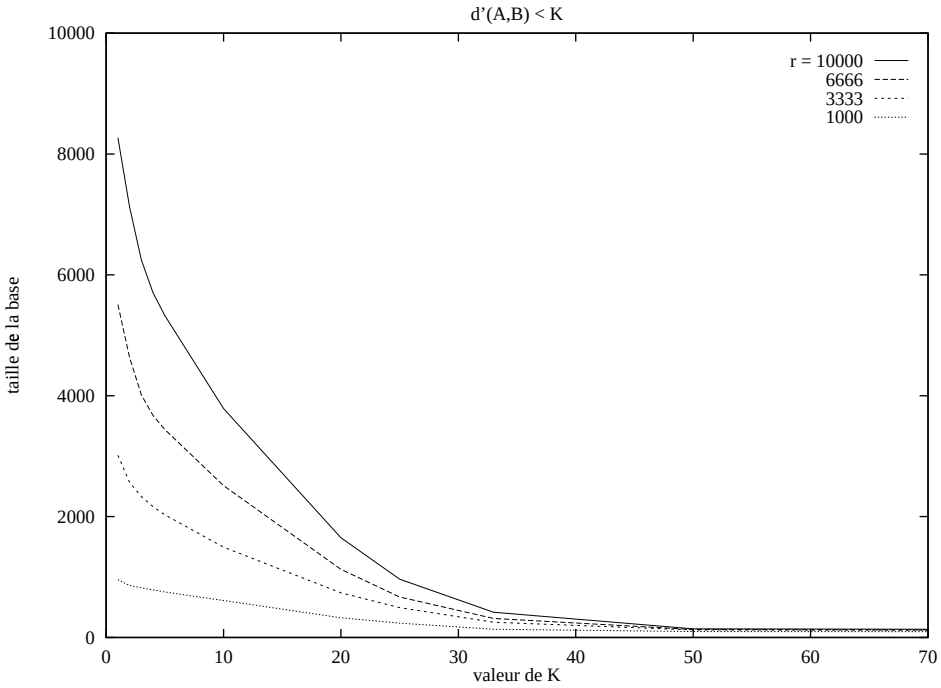


Figure 11.8: Taille de la base pour différents numéros r de phrases dans le corpus (contrainte absolue)

Les figures 11.9 et 11.10 (p. 267 et 267) donnent les courbes obtenues pour les contraintes du maximum et du minimum. Elles exhibent des formes différentes de celles obtenues pour la contrainte absolue. Mais elles sont toutes de forme « sigmoïdale ». Aucune contraction notable n'est obtenue avant $k = 0,75$. On observe un court plateau après $k = 0,95$. La contraction a lieu essentiellement entre 0,75 et 0,95. Pratiquement donc, si on se servait de cette contrainte pour construire une base représentative d'un corpus, il vaudrait mieux appliquer la contrainte du minimum ou du maximum avec une valeur proche de $k = 0,85$.

De même que précédemment on peut calculer une approximation linéaire de la taille de la base en fonction de la taille $|\Lambda_1|$ du corpus Λ_1 (de 0 à 10 000) et en fonction du facteur k (compris entre 0,75 et 0,95). Pour la contrainte du maximum, on a :

$$|\mathcal{A}| = (-2,3 \times |\Lambda_1| \times k) + (378 \times k) + (2,5 \times |\Lambda_1|) + 87$$

Pour la contrainte du minimum, on a :

$$|\mathcal{A}| = (-2 \times |\Lambda_1| \times k) + (1539 \times k) + (2,3 \times |\Lambda_1|) - 869$$

11.1.3 Analyse de la qualité des quasi-analogies detectées

L'algorithme utilisé dans toutes ces expériences (p. 236) rejette une phrase de la base dès qu'une quasi-analogie est détectée pour cette phrase. Or cette

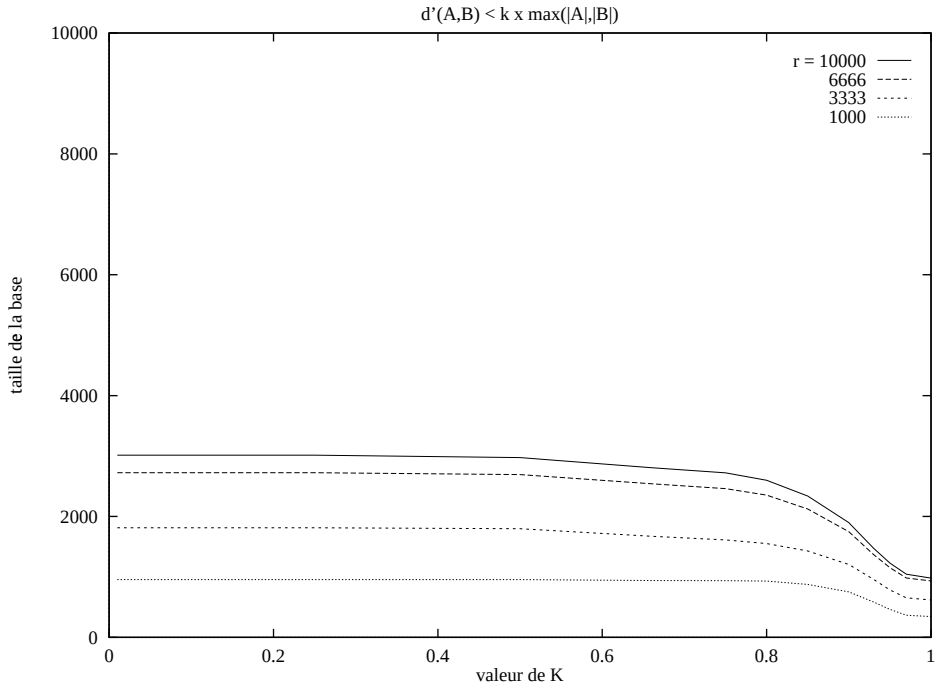


Figure 11.9: Taille de la base pour différents numéros r de phrases dans le corpus (contraintes du maximum)

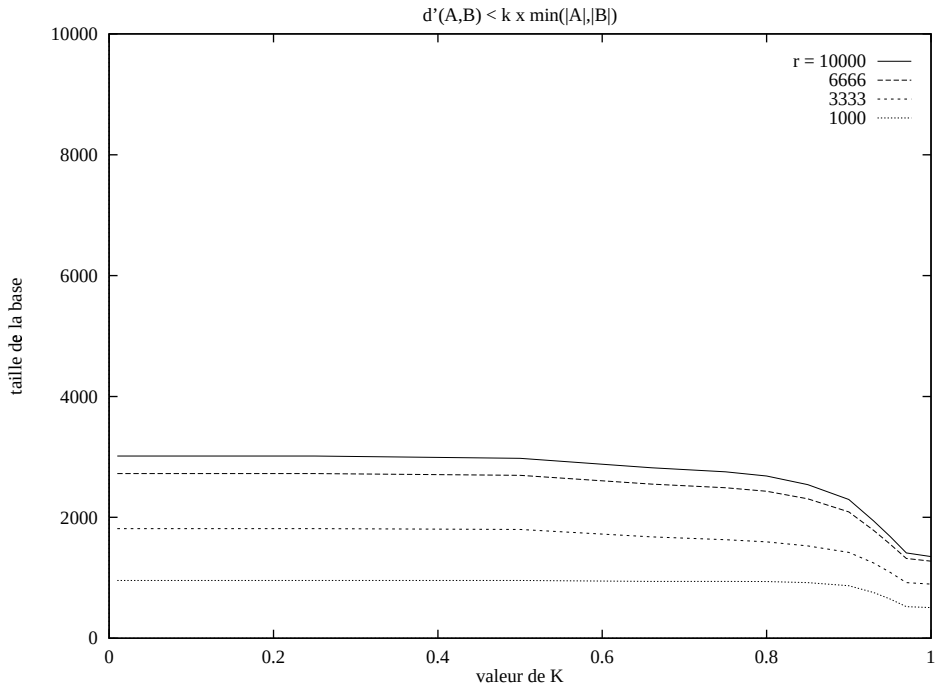


Figure 11.10: Taille de la base pour différents numéros r de phrases dans le corpus (contraintes du minimum)

quasi-analogie n'est pas forcément la meilleure possible, c'est-à-dire qu'elle n'est pas forcément analogique, au vrai sens du terme. En effet, d'autres quasi-analogies pourrient exister qui seraient des analogies. Nous nous sommes donc penché sur la qualité des quasi-analogies pour la contrainte du minimum pour différentes valeurs de k (voir le tableau 11.2).

La deuxième colonne de ce tableau donne le nombre de quasi-analogies détectées sur tout le corpus, c'est-à-dire le nombre de phrases rejetées de la base. Ce nombre se décompose en quatre selon la « qualité » des quasi-analogies.

- *Triviales*: phrases reconnues comme appartenant déjà à la base par l'algorithme. La quasi-analogie est trivialement une analogie. Dans le cas général, les phrases répétées ne sont pas forcément toutes detectées de cette façon. En effet, elles peuvent éventuellement avoir été detectées avant par d'autres quasi-analogies ;

- *Exactes*: phrases detectées par l'algorithme par des quasi-analogies qui sont des analogies ;

$$\begin{array}{l}\textit{A Directory}\\ \textit{With Thumbnails}\end{array} : \begin{array}{l}\textit{A Directory}\\ \textit{Without Thumbnails}\end{array} \doteq \begin{array}{l}\textit{A directory}\\ \textit{With Thumbnails}\end{array} : \begin{array}{l}\textit{A directory}\\ \textit{Without Thumbnails}\end{array} \quad [9]$$

- *Acceptables*:

  - *acceptables directement*: phrases rejetées de la base par une quasi-analogie qui peut être considérée comme analogie au niveau syntaxique ;

$$\begin{array}{l}\textit{Finance}\\ \textit{Committee}\end{array} : \begin{array}{l}\textit{Judiciary}\\ \textit{Committee}\end{array} \approx \begin{array}{l}\textit{Resources}\\ \textit{Committee}\end{array} : \begin{array}{l}\textit{Transportation}\\ \textit{Committee}\end{array} \quad [10]$$

  - *acceptables par transitivité*: phrases rejetées par une quasi-analogie ininterprétable syntaxiquement, mais qui sont des phrases répétées ou pour lesquelles existe une analogie avec une phrase syntaxiquement équivalente ;

$$\textit{Smoking is not forbidden.} : \textit{This is new terrain.} \approx \textit{HTML is a snap.} : \textit{User is really bummed.} \quad [11]$$

- *Erronées*: phrases pour lesquelles la quasi-analogie n'est pas interprétables comme analogie.

Pour $k = 0,25$, toutes les phrases répétées sont détectées et 11 analogies acceptables sont detectées, dont 7 correspondent à un remplacement de nombres, et 4 correspondent à un remplacement de noms. L'exemple suivant montre un tel remplacement de noms dans une structure syntaxique élémentaire.

[9](Un Répertoire avec des Vignettes.) Mêmes phrases avec *sans* et avec des changements de majuscules.

[10](Commission des finances) : (Commission des affaires judiciaires) ≈ (Commission des ressources) : (Commission des transports)

[11](Il n'est pas interdit de fumer.) : (C'est un nouvel endroit.) ≈ (HTML, c'est fastoche.) : (L'utilisateur est vraiment déçu.)

Tableau 11.2: Nombre de quasi-analogies détectées pour la contrainte du minimum avec différentes valeurs de k

	nombre total de quasi-analogies
0,25	1740
0,66	2213
0,80	2899
0,85	3640
0,93	5796

Tableau 11.3: Répartition par espèces des quasi-analogies détectées pour la contrainte du minimum avec différentes valeurs de k

	triviales	exactes	acceptables		erronées
			directement	par transitivité	
0,25	1729 = 99%	0	11 = 1%	pas de sens	0 = 0%
0,66	1071 = 48%	2	193 = 9%	835 = 38%	112 = 5%
0,80	932 = 32%	3	775 = 27%	544 = 19%	648 = 22%
0,85	844 = 23%	2	535 = 15%	1486 = 41%	974 = 27%
0,93	714 = 12%	1	82 = 1%	1738 = 30%	3371 = 57%

$$\textit{User is a robot.} : \textit{User is a vampire.} \approx \textit{User is a hosehead.} : \textit{User is a wizard.}$$ [12]

Quand k augmente, le nombre de quasi-analogies erronées augmente encore, mais il est limité à 22% autour de $k = 0,80$, ce qui semble raisonnable. Pour cette valeur de k, le nombre total de quasi-analogies acceptables est assez élevé : 46%, et le nombre d'analogies exactes, quoique peu significatif, atteint sa plus haute valeur. En tout, pour cette valeur de k, 78% des quasi-analogies detectées, c'est-à-dire plus des trois quarts, sont au moins acceptables par transitivité.

11.1.4 Synthèse sur la représentativité de corpus

Les expériences précédentes nous ont permis d'examiner la possibilité de contracter un grand corpus en extrayant seulement un certain nombre de phrases représentatives, en fait, en rejetant au fil de sa lecture les phrases déductibles de phrases précédentes par quasi-analogie. Pour réaliser ces expériences nous ne nous sommes servi que de notions élémentaires de linguistique avec l'analogie, d'algèbre moderne avec les ensembles libres, générateurs et bases, et d'algorithmique avec l'algorithme glouton.

[12] (L'utilisateur est un robot.) : (L'utilisateur est un vampire.) ≈ (L'utilisateur est un gros nul.) : (L'utilisateur est magicien.)

Il serait évidemment souhaitable de reprendre de telles expériences en remplaçant les mots du corpus par leur catégories morpho-syntaxiques, ou par leur codes sémantiques.

Nous avons vu que, étant donné un ensemble Λ_1 de taille $|\Lambda_1|$, le nombre de relations analogiques à examiner par les algorithmes 9.2 ou 9.3 (p. 240) sont en le cube de la taille de l'ensemble. Pour l'analyse par analogie ou la traduction directe par analogie, susceptibles d'utiliser directement ces algorithmes, l'analyse ou la traduction par analogie d'une phrase nouvelle prendrait donc un temps cubique en l'ensemble des données. Si cet ensemble de données était réduit en utilisant la méthode présentée ici avec la contrainte du minimum, le nombre de relations analogiques à examiner tomberait à $4 \times |\Lambda_1|^{3 \times 4/5} = 4 \times |\Lambda_1|^{2,4}$ pour $k = 0,85$ ou à $10 \times 2/3 \times |\Lambda_1|^{3 \times 2/3} = 6,66 \times |\Lambda_1|^2$ pour $k = 1$. Par conséquent, le comportement asymptotique de l'analyse ou de la traduction d'une phrase, originellement cubique deviendrait presque quadratique pour des résultats que l'on espère identiques. Cela représenterait un gain non négligeable.

11.2 Ensemble de phrases mémorisant la cohésion

Avant d'aborder la partie concernant la productivité, nous nous arrêtons pour une incursion dans le domaine de la reconnaissance de la cohésion des textes. Avec les caractérisations que nous avons données et que nous allons donner des deux notions de représentativité et de productivité, celle de cohésion locale peut être vue comme faisant un pont. Nous avons en effet caractérisé la représentativité par le nombre d'analogies présentes dans un corpus. La productivité sera caractérisée, dans la partie suivante, par le nombre de phrases qui peuvent être produites à l'aide des phrases d'un corpus. On peut spécialiser aussi bien la caractérisation de la représentativité que celle de la productivité en fixant l'une des phrases du corpus et en calculant le nombre de fois où elle peut être produite par des phrases du même corpus.

Considérons alors, dans une perspective d'apprentissage, la lecture phrase par phrase d'un corpus. À un moment donné, on dispose de l'ensemble des phrases lues jusqu'à cet instant. On peut considérer cela comme une mémoire. La phrase suivante à lire peut éventuellement être produite par analogie avec des phrases du corpus lues jusqu'à cet instant, c'est-à-dire à l'aide de cette mémoire. Combien de fois elle peut l'être peut caractériser à la fois la représentativité du corpus jusqu'à cet instant et sa productivité relativement à la phrase à lire.

Or aucune mémoire ne saurait être extensible à loisir. On pourrait envisager une mémoire qui s'arrête de croître dès qu'un certain volume est atteint. Un tel type de mémoire est peu vraisemblable: le passé lointain y est immuable, et le passé récent est oublié au fur et à mesure de la marche du présent. Nous préférons donc plutôt une mémoire de type immédiate, c'est-à-dire un type de mémoire où le passé ancien est oublié en faveur du passé récent.

Appliquée à la représentativité et à la productivité du point de vue de la phrase suivante à lire dans un corpus, la mémoire immédiate est le nombre de phrases à retenir à partir de la phrase à lire, et en direction du passé, pour pouvoir réobtenir celle-ci à tout coup par analogie. Si réobtenir une phrase par analogie, ou être capable de la faire entrer dans une quasi-analogie, est interprêté comme la comprendre, on obtient alors une sorte de modèle très primaire de la compréhension avec mémoire immédiate par analogie ou quasi-analogie.

Plus le nombre d'analogies ou de quasi-analogies est élevé, meilleure est la « compréhension ». Mais comme d'un point de vue pratique, toute application informatique a intêret à utiliser la plus petite mémoire possible, nous nous intéresserons à savoir comment évolue la taille de cette mémoire pour une compréhension plus ou moins sûre.

L'expérience que nous proposons consiste donc à déterminer localement, au fil de la lecture du corpus, le nombre de phrases déjà lues qui permettrait d'assurer au minimum un certain nombre de quasi-analogies, avec la phrase en

train d'être lue[13]. En termes plus formels, pour chaque r allant de 1 au nombre de phrases dans le corpus, on détermine le plus petit nombre de phrases $m(r,n)$ tel que la phrase de numéro $r+1$ entretienne n relations de quasi-analogie avec des triplets de phrases dont les numéros se trouvent dans l'intervalle $[\, r - m(r,n)\ ;\ r\,]$. Nous désignons par $\Lambda_1[i]$ la i^{e} phrase du corpus Λ_1. On pose

$$\lambda(i) = \frac{1}{6} \times \Big|\ \{(a,b,c) \in [\, r-i\ ;\ r\,]^3\ /\quad \Lambda_1[a] : \Lambda_1[b] \approx \Lambda_1[c] : \Lambda_1[r+1]\quad \}\ \Big|$$

La division par 6 provient du fait que, pour la quasi-analogie, expliquée par le système du rectangle (p. 258), les six triplets obtenus en faisant toutes les permutations possibles de a, b et c sont équivalents. On définit alors $m(n,r)$ comme suit.

$$m(n,r) = \min_{i \in \mathbb{N}} \{\lambda(i)\ /\ \lambda(i) \geq n\}$$

Nous avons effectué le calcul de $m(n,r)$ dans six cas sur le même corpus que celui des expériences sur la représentativité (p. 259). Les six cas correspondent au calcul pour les textes bruts et les textes étiquetés pour trois valeurs différentes de n : 1, 10 et 100. Soulignons que cette expérience utilise l'explication rectangulaire de l'analogie que nous savons imparfaite et que nous avons baptisée quasi-analogie (voir p. 258).

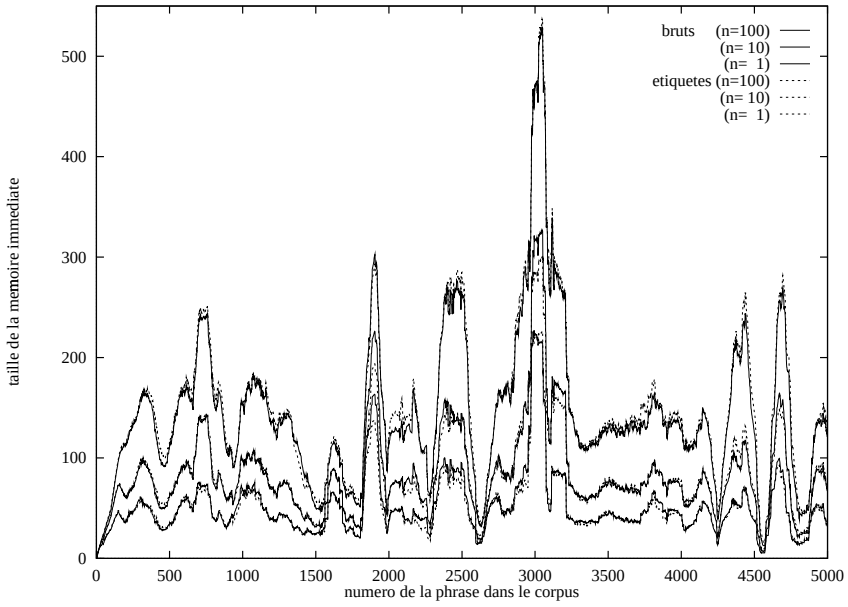


Figure 11.11: Taille de la mémoire nécessaire pour obtenir en moyenne un nombre fixé de quasi-analogies. Textes bruts et étiquetés, trois nombres différents de quasi-analogies exigées : 100, 10 et 1

La figure 11.11 donne les courbes lissées obtenues dans les six différents cas. La courbe la plus haute est celle obtenue pour les textes bruts et une mémoire

[13]Nous reprenons LEPAGE *et al.*, *The snow-ball effect of analogy*, 1997, p. 290 et 291.

de 100. La courbe en pointillé pour les textes étiquetés pour le même nombre de quasi-analogies lui est quasiment superposée. Les courbes intermédiaires sont celles obtenues pour 10 analogies, pour les textes bruts ou étiquetés. Enfin, les deux courbes les plus basses, solide et en pointillé, donnent la taille de la mémoire pour les textes bruts ou étiquetés pour une seule quasi-analogie.

On observe premièrement que les courbes des textes bruts et des textes étiquetés sont quasiment superposées pour le même nombre de quasi-analogies. Il n'y a pas de différence notable, pour ce calcul, entre mots et catégories syntaxiques. On en conclut que l'ensemble des catégories syntaxiques, qui comprend environ 960 catégories différentes, représente de façon très fine, voire trop fine, les commutations entre les 10 600 différents mots des textes, en tout cas en ce qui concerne la quasi-analogie. Deuxièmement, on observe que la forme des six courbes est quasiment la même, à ce qui semble des translations vers le haut près. La taille de la mémoire immédiate serait donc, à des constantes près, indépendante du nombre de quasi-analogies exigées. On en conclut qu'elle serait une caractéristique intrinsèque du corpus du point de vue de la quasi-analogie.

Interrogeons-nous sur la signification de cette caractéristique intrinsèque. Plus la taille de la mémoire immédiate est petite, plus les phrases dans un contexte restreint sont quasi-analogiques entre elles. Plus elle est grande, plus les phrases sont diverses. La taille de la mémoire reflète donc la cohésion locale du corpus du point de vue de la quasi-analogie.

Figure 11.12: Taille moyenne de la mémoire immédiate

nombre de quasi-analogies exigées	taille en nombre de phrases	taille en nombre de pages
100	250	5 ou 6
10	80	1 et demie
1	51	1

En moyenne, pour nos données d'expérience, un contexte d'environ 250 phrases en moyenne, est suffisant pour obtenir 100 quasi-analogies pour la phrase suivante. Cela représente environ 5 ou 6 pages. Un contexte plus petit de 80 phrases, ou, respectivement, de 51 phrases, est nécessaire pour obtenir 10 quasi-analogies, ou, respectivement, une seule quasi-analogie. Nous résumons ces résultats dans le tableau 11.12.

L'examen des courbes révèle des minimums locaux pour certains numéros de phrases. Si on pense que les quasi-analogies rendent compte d'une certaine cohésion des phrases, alors, il semble raisonnable de penser que les phrases ressortissant d'un même style de texte devrait partager une même mémoire. Par conséquent, pour un changement abrupt de style, les courbes obtenues devraient croître de façon abrupte, pour décroître ensuite. Elles devraient rester constante si le style reste le même pendant un temps supérieur à la taille de

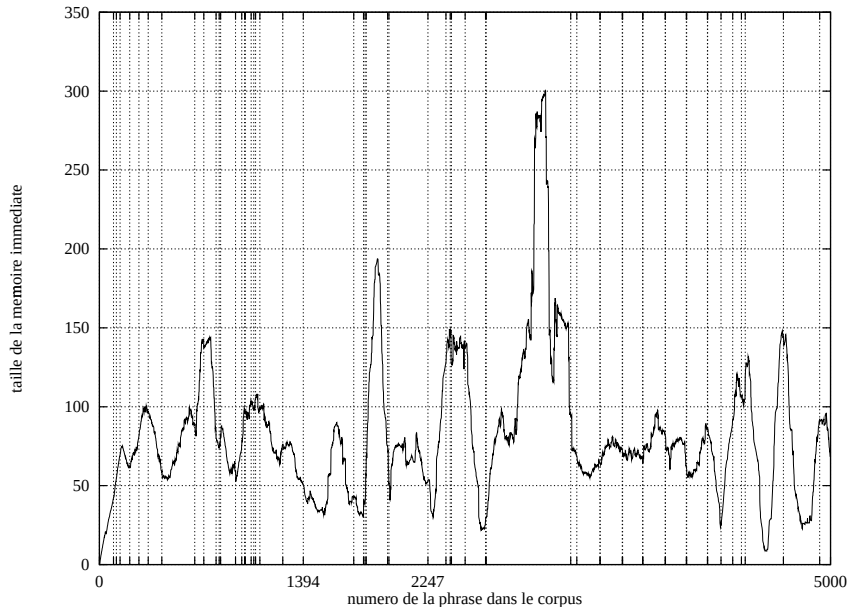


Figure 11.13: Découpage du corpus par textes en abscisses. Taille de la mémoire immédiate quasi-analogique en ordonnées

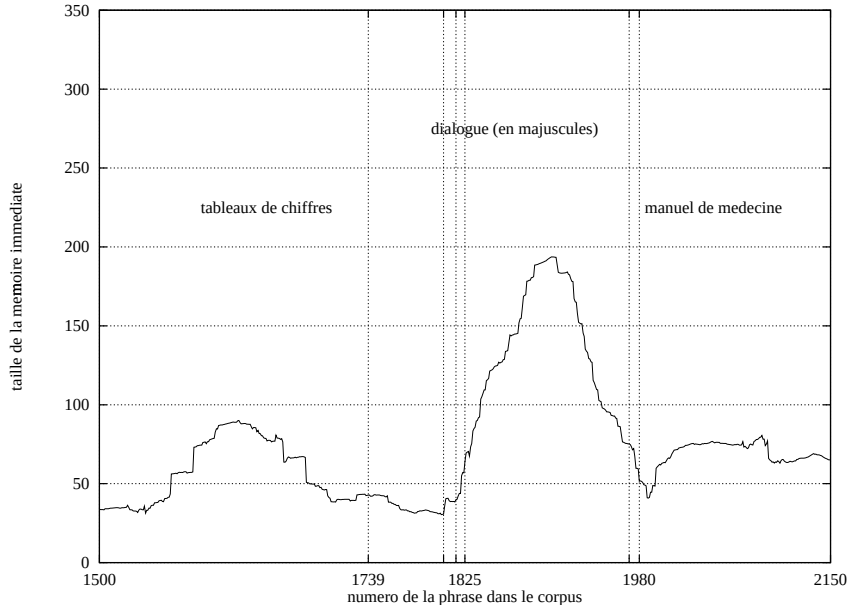


Figure 11.14: Découpage par textes en abscisses. Taille de la mémoire immédiate quasi-analogique en ordonnées (phrases comprises entre les numéros 1 500 et 2 150)

la mémoire immédiate. Par conséquent, les minimums locaux devraient correspondre à peu près à des frontières de textes. Mais toutes les frontières de textes ne correspondront pas nécessairement à des minimums locaux car deux textes qui se suivent peuvent partager le même style.

Sur l'ensemble du corpus, nous avons vérifié que les minimums locaux correspondaient bien à peu près aux frontières entre styles de texte. C'est ce que montre la figure 11.13 (p. 274). Sur cette figure, on a visualisé par des frontières verticales les limites entre textes les plus proches des minimums locaux. De façon plus précise, nous avons isolé dans la figure 11.14 (p. 274) la portion du corpus contenant les phrases numérotées de 1 500 à 2 150. Cela représente une centaine de pages. De la phrase numéro 1 500 à la phrase numéro 1 739, on trouve une liste de tableaux donnant les superficies, populations, etc. de pays africains, puis une table des pavillons maritimes internationaux et de leur signification, etc. Des numéros 1 826 à 1 980, on a un dialogue tout écrit en majuscules. La troisième partie est un extrait d'un manuel de médecine traitant du cancer.

Des expériences plus fines seraient évidemment nécessaires avant d'arriver à une application fiable, mais les conclusions de l'expérience que nous venons de rapporter permettent d'entrevoir la possibilité de faire du découpage automatique de textes fondé sur le style, à condition que deux textes qui se suivent ne partagent pas le même style.

11.3 Ensemble de phrases produites à partir d'un corpus

Le problème de la productivité a directement à voir avec le refus des Générativistes de considérer l'analogie comme pertinente pour la linguistique. On a vu que Chomsky rejetait l'idée que l'analogie pouvait servir de critère de grammaticalité sous prétexte qu'il est facile de produire des énoncés agrammaticaux par analogie (p. 76). Nous nous proposons ici de faire quelques mesures pour estimer dans quelle mesure l'application aveugle de l'analogie produit des énoncés agrammaticaux.

Nous désirons donc savoir quelle quantité de phrases nouvelles grammaticales peuvent être produites à partir d'un corpus donné. Reprenons les notations introduites avec les langages de chaînes analogiques. Par rapport aux deux précédents problèmes, nous changeons de perspective. En effet, là, le corpus était l'ensemble Λ_1. Ici, il s'agit de produire l'ensemble Λ_1 à partir de la donnée d'un ensemble $\mathcal{A}$. Le corpus est donc ici $\mathcal{A}$ et non Λ_1.

11.3.1 Prégnance d'un corpus

Pour caractériser la productivité, nous comparerons les tailles du corpus de départ et du résultat en calculant le rapport $|\Lambda_1|/|\mathcal{A}|$. Nous proposons d'appeler *prégnance théorique* d'un corpus ce rapport. Comme nous nous intéressons au nombre de phrases grammaticales produites, nous appellerons *prégnance utile* le rapport entre le nombre de phrases grammaticalement correctes produites par analogie et le nombre de phrases du corpus[14].

A partir d'un corpus donné, il nous est aisé de générer automatiquement par analogie un ensemble de nouvelles phrases. En effet, pour tout triplet de phrases du corpus considéré, la résolution de l'équation analogique peut éventuellement produire une ou plusieurs phrases nouvelles.

Dans une première expérience, nous avons utilisé un tout petit corpus exemple de 31 phrases, traduction des phrases japonaises du corpus artificiel de Satou[15] pour ses expériences d'apprentissage automatique de patrons en vue de la traduction par l'exemple. La figure 11.15 (p. 277) liste ce corpus. Dans cette expérience, l'analogie est résolue en considérant les phrases comme des chaînes de caractères, et non comme des chaînes de mots. Le résultat est surprenant et extrêmement prometteur, puisque 92 phrases nouvelles sont générées parmi lesquelles 41 s'avèrent grammaticalement correctes ! La figure 11.16 (p. 278) donne une petite moitié des phrases produites par application aveugle de l'analogie.

On pourrait alors envisager d'ajouter les phrases nouvelles grammaticales au corpus pour obtenir, dans un deuxième temps, l'ensemble appelé Λ_2 dans nos notations des langages de chaînes analogiques (p. 170). Et l'on peut envisager de recommencer indéfiniment pour calculer la fermeture transitive du

[14] Lepage, *Un éditeur pour la construction de banques d'arbres*, 1996, p. 109.

[15] 佐藤 理史 (Satou Satosi), *Example-based Machine Translation*, 1991, p. 23, figure 2.5.

je suis Takuma.
je suis Taro.
tu es Takuma.
tu es Hanako.
je suis un garçon.
tu es une fille.
tu es un garçon.
je suis un grand garçon.
tu es une petite fille.
ceci est un livre.
cela est un livre.
ceci est une pomme.
ceci est une orange.
ceci est ma pomme.
ceci est mon livre.
cela est ton livre.
ceci est un chien.
cela est le chien de Taro.
il est un garçon.
elle est une fille.
il est mon ami.
elle est professeur.
elle est le professeur de Hanako.
cela est une balle de tennis.
c'est ta balle de tennis.
je ne suis pas Taro.
elle est infirmière.
tu n'es pas infirmière.
il n'est pas mon professeur.
elle n'est pas la mère de Hanako.
ce n'est pas une balle de tennis.

Figure 11.15: Corpus de 31 phrases

tu es Taro.
cela est le chien de Takuma.
je suis Hanako.
**jtulle est le profesuiur de Hanako.*
**jtulle est le profesuiur de Takuma.*
elle n'est pas la mère de Takuma.
il est Hanako.
**tmo es Hanako.*
**o es Hanako.*
je suis une fille.
je suis une petite fille.
**jtulle est le profesuiur de Hanako.*
**elle est le pofesseund gar de Hanako.*
**je ne suis pas Trand garço.*
**tu n'es pas infimiènd garçe.*
je suis mon garçon.
je suis ton garçon.
**je sucesa un livre.*
**je suis mae garçon.*
**je sucesa ton livre.*
je suis mon ami.
**je sucesa une balle de tennis.*
**il ene st pas Taro.*
il est une fille.
elle est un garçon.
elle est une petite fille.
**tu es mone fille.*
**tu es tone fille.*
tu es ma fille.
tu es professeur.
tu es infirmière.
il est une petite fille.
tu es mon garçon.
tu es ton garçon.
**tu es mae garçon.*
tu es mon ami.
il n'est pas infirmière.
je suis mon grand garçon.
je suis ton grand garçon.
⋮

Figure 11.16: Exemples de phrases produites par application aveugle de l'analogie

corpus par analogie, c'est-à-dire le langage de chaînes analogiques paresseux $\Lambda(\mathcal{A}, \mathcal{A}^2)$, en utilisant uniquement les modèles construits sur le corpus de départ $\mathcal{A}$.

Pour en revenir à la caractérisation de la productivité, notre petit corpus artificiel a une prégnance théorique de 297%, et une prégnance utile de 135%.

11.3.2 Nombre de phrases produites ou cardinal de Λ_1

Dans une deuxième expérience, sur un corpus un peu plus grand, après avoir produit des phrases aveuglément par analogie, nous nous sommes penchés sur le problème de l'agrammaticalité des phrases produites[16].

Conditions expérimentales

Du corpus japonais arboré ATR-NEC[17], nous avons extrait toutes les phrases contenant le caractère ou symbole 持 /motu/. Il en existe 153. Cet ensemble constitue notre corpus de départ $\mathcal{A}$. La répartition des phrases par longueurs dans ce corpus est donnée dans la figure 11.17. Cette répartition est résumée dans le tableau 11.4.

Pour résoudre les équations analogiques, nous utilisons une ancienne version de l'algorithme[18]. Cet algorithme permet, par exemple, la production des phrases suivantes.

お持ちしませんか。 : いいえ，お持ちします。 ≐ 持っていませんか。 : x ⇒ $x =$ いいえ，持っています。 [19]

場所が彼に決められた。 : 場所を彼が決めた。 ≐ 柿が彼に食べられた。 : x ⇒ $x =$ 柿を彼が食べた。 [20]

Bien qu'il puisse exister plusieurs solutions à une équation analogique, dans de cette expérience, comme nous l'illustrons ci-dessous, nous n'avons retenu que la première solution produite par l'algorithme.

こっちこい : そっちいって ≐ ここへこい : x ⇒ $x =$ 1/* そいへこって 2/ そこへいって [21]

[16] Cette partie reprend en le modifiant l'article Lepage & 白井諭 (Sirai Satosi), 言語学的類推による生成文における非文法生の分析, 2000.

[17] Voir Lepage *et al.*, *An annotated corpus in japanese using Tesnière's structural syntax*, 1998.

[18] Lepage, *Solving analogies on words: an algorithm*, 1998.

[19] (Dois-je vous le porter ?) : (Non, je vous le porte.) = (En avez-vous ?) : (Non, je n'en ai pas ?) La première phrase est en japonais à la forme négative, ce qui est une manière plus polie de demander.

[20] (Le lieu a été choisi par lui.) : (Il a choisi le lieu.) = (La figue-caque a été mangée par lui.) : (Il a mangé la figue-caque.)

[21] (Viens ici.) : (Va là.) ≐ (Viens ici.) : (Va là.) Les formes こっち /kotti/ (ici (direction)) et ここへ /kokohe/ (prononcée /kokoe/) sont synonymes. De même pour les formes そっち /sotti/ et そこへ /sokohe/ (prononcée /sokoe/). Ces formes font parties de la série *ko-so-a* mentionnée plus haut (note 42, p. 141).

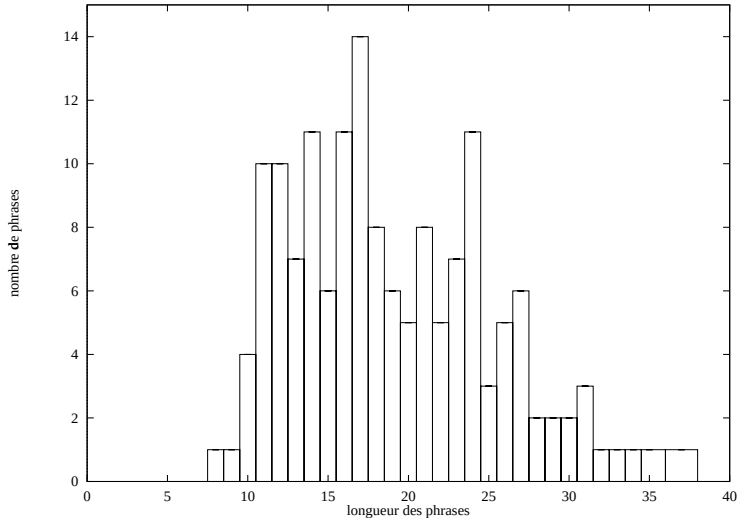


Figure 11.17: Distribution des longueurs de phrases en nombre de caractères

Tableau 11.4: Caractéristiques des données

taille en caractères			nombre de phrases
min	moyenne ± écart-type	max	
8	19,5 ± 6,2	37,5	153

Tableau 11.5: Distribution en fréquence des phrases produites

	phrases nouvelles	phrases déjà existantes produites		total
		une fois	deux fois	
nombre de phrases	1095	151	2	1248
fréquence	1-8	en tout 142	en tout 280	environ 25 000

Tableau 11.6: Classement grossier des phrases produites en termes de grammaticalité

phrases correctes		incorrectes		total
naturelles	possibles en contexte	agrammaticales	asémantiques	
453	2	769	24	1248
36,3%	0,2%	61,6%	1,9%	100%

Chiffres de l'expansion

Si l'on considère simplement l'ensemble des triplets possibles, pour notre corpus, il en existe $153^3/8 \simeq 3,6 \times 10^6/8 = 450\ 000$. En comparaison, seulement 25 000 cas produisent un résultat, lors de la résolution des équations analogiques correspondantes. Cela donne un taux de résolution d'environ 1/8. Le nombre de phrases différentes produites est de 1 248. Comparé à la taille du corpus, cela représente $1\ 248/153 = 8,16$ fois plus de phrases. La prégnance est donc de 816%. Parmi les phrases produites, les 153 phrases du corpus de départ sont à nouveau produites. Tous ces résultats sont résumés dans le tableau 11.5 (p. 280).

11.3.3 Analyse de la qualité des phrases produites par analogie

Nous donnons d'abord dans le tableau 11.6 un classement grossier des phrases produites en termes de grammaticalité. Dans plus d'un tiers des cas, des phrases correctes grammaticalement et en sens ont été obtenues.

En retirant des 425 phrases correctes grammaticalement et en sens, les 153 phrases égales aux phrases du corpus, on obtient 299 phrases entièrement nouvelles. Nous avions proposé plus haut la notion de « prégnance utile » comme quantité caractérisant la productivité. Ici, la prégnance utile est de $299/153 = 195\%$. Presque 2 fois plus de phrases nouvelles ont été produites aveuglément par l'analogie. Voici des exemples de phrases nouvelles produites grammaticalement correcte.

あとでテーブルにそこにあるもの何でも持ってきて下さい。
大きな荷物は船内にはそこにあるもの何でも持ち込めません。
小皿を二, 三枚そこにあるもの何でも持ってきて下さい。
確認書を持っていないのですが。
手持ちのドルがあります。[22]

Ces exemples montrent clairement qu'il y a eu transport de morceaux de phrases par échanges de morceaux entre phrases analogiques. Par exemple, l'insertion du morceau そこにあるもの何でも[23] dans les phrases suivantes du corpus あとでテーブルに持ってきて下さい。[24] et 大きな荷物は船内には持ち込めません。[25] a permis la production des phrases nouvelles données en exemples ci-dessus. Si dans ces exemples, un seul morceau seulement a été inséré, en théorie, plusieurs morceaux peuvent être simultanénement insérés.

[22](Ensuite, apportez tout ce qu'il y a là sur la table.)
(On ne peut emporter à bord des bagages volumineux et tout ce qu'il y a là.)
(Apportez deux ou trois petites assiettes et tout ce qu'il y a là, s'il vous plaît.)
(Vous n'avez pas de certificat ?)
(J'ai des dollars sur moi.)

[23]/Soko ni aru mono nan demo/ (tout ce qu'il y a là.)

[24]/Ato de teeburu ni motte kite kudasai./ (Ensuite, apportez[-le] sur la table.)

[25]/Ooki na nimotu ha sennai ni ha motikomemasen./ (On ne peut apporter à bord des bagages volumineux.)

La définition de l'analogie veut qu'une analogie ne puisse exister qu'avec quatre objets, et donc qu'une équation analogique ne puisse être constituée qu'avec trois objets. Par conséquent, si une relation d'opposition par échange de morceaux n'apparaît qu'une seule fois dans le corpus, alors la production de phrases nouvelles par utilisation de cette opposition s'avère impossible. Par exemple, dans le corpus utilisé dans cette expérience, la relation d'opposition entre les deux phrases suivantes (par échange de は別の[26] et 違う[27]) n'apparaissait que dans le seul couple suivant. Par conséquent, conformément à la définition de l'analogie nous n'avons pu observer[28] cette opposition entre des phrases du corpus et des phrases produites.

このワインとは別のものを持ってきてください。
このワインと違うものを持ってきてください。[29]

Analyse succinte des erreurs

Au contraire, l'utilisation incorrecte de relations d'opposition, typiquement dans un contexte non pertinent, autorise la production de phrases agrammaticales ou ne faisant pas sens. Bien des combinaisons sont possibles. Par exemple, des chaînes de symboles ne constituant pas des mots, ou des groupements grammaticalement correctes, mais apparaissant dans une opposition peuvent provoquer la production de phrases agrammaticales. En plus, des cas de groupements grammaticalement corrects peuvent aussi provoquer la production de phrases ne faisant pas sens. Ces différents cas, et leur effet négatif sur la grammaticalité et le sens sont possibles dans cette expérience puisque nous n'appliquons ici l'analogie qu'au niveau des symboles seulement.

Échange de « す » et « せん » Nous avons vu qu'il est possible de rendre compte de la régularité dans la conjugaison ou la déclinaison par la relation d'analogie (p. 245). Dans l'expérience rapportée ici, on a bien observé la production juste de phrases grâce à l'opposition dans la conjugaison japonaise des formes affirmatives en ます /masu/ et des formes négatives en ません /masen/. Cela est possible grâce à l'existence dans la base des deux phrases suivantes.

はい、持っています。
はい、持っていません。[30]

On a donc pu observer la production des phrases suivantes.

[26] /Ha betu no/ (à part, différent). Le は est en fait une postposition, 別 est un nom, et の correspond au français *de*.

[27] /Tigau/ (différent)

[28] Je tiens à exprimer ici mes plus chauds remerciements à Monsieur SIRAI Satosi pour son aide précieuse dans l'examen des résultats et leur classification.

[29] (Apportez autre chose que ce vin.)
(Apportez quelque chose de différent de ce vin.)

[30] /Hai, motteimasu./ (Oui, j'en ai.)
/Hai, motteimasen./ (Non, je n'en ai pas.)
Le mot はい ne marque pas une affirmation dans l'absolu, mais l'acquiescement en contexte. En conjonction avec un verbe à la forme négative, il se traduit donc par *non* en français.

生ものや植物の種などは持っておられますか。
キーホルダーを持っていません。[31]

à partir des phrases suivantes du corpus.

生ものや植物の種などは持っておられませんか。
キーホルダーを持っています。[32]

Mais, l'application excessive de cette relation d'opposition, par interprétation erronée comme opposition entre す /su/ et せん /sen/, à un élément étranger à cette conjugaison, la forme verbale です, a donné lieu à la production de phrases agrammaticales.

* 缶ジュースなら持ち込んでも宜しいでせんか。

à partir de la phrase suivante du corpus.

缶ジュースなら持ち込んでも宜しいですか。[33]

La forme * でせん /desen/ constitue un horrible barbarisme. La forme correcte serait ではありません /dehaarimasen/ (n'est pas). Le barbarisme a été clairement obtenu par la résolution de l'équation analogique

$$\text{おられます} : \text{おられません} \doteq \text{です} : x.$$

Ce type d'équation est tout à fait admissible pour toutes les formes verbales en ます /masu/, mais pas pour le verbe irrégulier です.

$$\text{おられます} : \text{おられません} \doteq \text{持っています} : x \quad \Rightarrow \quad x = \text{持っていません}$$

En français, un barbarisme équivalent serait, par exemple, **êtra* pour *sera*. Nous avons donc ici une bonne illustration du caractère aveugle de l'analogie (p. 27). Afin de remédier à cette situation, on entrevoit la nécessité de contraindre l'analogie, en notant par avance l'exception à です / ではありません afin de bloquer la formation de l'équation analogique donnée plus haut. Cela implique l'existence d'un autre domaine que celui des seules phrases, domaine qui serait une représentation abstraite du domaine des phrases. Ce genre de formalisation fait l'objet de la partie suivante où nous traiterons des homomorphismes entre espaces analogiques (p. 292).

[31] (Transportez-vous des êtres vivants ou des graines de plantes ?)
(Je n'ai pas de porte-clés.)

[32] (Ne transportez-vous pas des êtres vivants ou des graines de plantes ?)
(J'ai un porte-clés.)

[33] (Et si c'est du jus de fruit en canette, je peux le prendre avec moi ?)

Lieux des insertions Nous avons déjà donné des exemples d'insertion de séquences contiguës, mais, dans le cas général, suivant en cela les différents types d'opérations d'édition, les morceaux de chaîne introduits en plusieurs endroits simultanément peuvent être des remplacements, des insertions et des suppressions.

Voici un exemple où, à partir de la phrase du corpus 小皿を二, 三枚持ってきて下さい。[34], 小皿[35] et を二, 三枚持ってきて下さい。 ont été introduits sans aucun sens, avec en plus l'insertion en tête de phrase de 日[36].

* 日小皿のマンガを二, 三枚おみやげに持ってきていますが問題はありませんか。
* 日小皿の雑誌を二, 三枚たくさん持って行きたいのですが, 税関で問題はないですか。
* 日小皿の植木を二, 三枚アメリカに持って行きたいのですが。

Dans tous ces exemples, si on supprime 日, on obtient des phrases qui bien que grammaticales sont insatisfaisantes du point de vue du sens.

Types de morceaux introduits L'opposition mentionnée plus haut entre ます et ません existe bien dans la conjugaison des verbes japonais. Mais comme l'application que nous avons faite ici de l'analogie est aveugle, des oppositions qui n'ont pas de sens dans le système de la langue provoquent la production de phrases agrammaticales. La raison pour laquelle l'exemple ci-dessous a été produit n'est pas clair, car il est difficile de juger à partir de quelles phrases du corpus elle a été obtenue.

* グラスは [手荷物] おいくつ持 [って] いしましょう, アイスはボックス一つで宜しい [ま] すか。

Cependant, la phrase qui y ressemble le plus dans le corpus est : グラスはおいくつお持ちしましょう, アイスはボックス一つで宜しいですか。[37]. C'est pourquoi nous avons pu marqué les morceaux insérés en les mettant entre crochets. Les oppositions entre 持ち et 持って[38] d'une part, et entre ます et です d'autre part apparaissant dans le corpus, on peut penser qu'elles ont été utilisées. Une explication à l'insertion fautive de 手荷物 est plus difficile.

[34] /Kozara wo ni-sanmai motte kite kudasai./ (Apportez deux ou trois petites assiettes, s'il vous plaît.)

[35] /Kozara/ (petite assiette). Morphologiquement, on a la même construction en japonais et en français. Le caractère 小 /ko/ signfie *petit*, et 皿 /sara/ *assiette*. Dans ce type de composition, la consonne initiale du second mot devient sonore. Ce phénomène est appelé *rendaku*.

[36] /Niti/ ou /hi/ (le jour).

[37] /Gurasu ha o ikutu o moti simasyou, aisu ha bokksu hitotu de yorosii desu ka./ (Combien dois-je vous apporter de verres ? Un seul bac à glace, cela ira ?)

[38] /Moti/ est la forme nominale du verbe 持つ /motu/ (porter), obtenue par remplacement de la finale /u/ en /i/. /Motte/ (porté, portant) est une forme participiale du verbe.

Décalage de symboles Comme il a été dit plus haut, dans cette expérience, nous ne retenons que les premières solutions produites par l'algorithme. Pour cette raison, certains décalages de symboles sont observés, dues au fait que notre algorithme est imparfait. Rappelons que nous n'avons pas encore réussi à formaliser le versant de la contiguïté de l'analogie. La phrase * 石鹸をお持てきちくださし。 a peut-être été produite à partir des trois phrases suivantes du corpus. Entre la première et la deuxième phrase, on pourrait voir, entre autres, un remplacement du caractère い /i/ par し /si/.

$$\text{持っていません。} : \text{石鹸を持ってきてください。} \doteq \text{お持ちしません。} : x \quad \Rightarrow \quad x = \text{*石鹸をお持てきちくださし。} \quad [39]$$

Par rapport à cela, on trouve aussi dans l'ensemble des phrases produites par analogie la phrase * 石鹸をお持ちきてくださし。 qui, quoique n'étant pas du japonais, semble plus proche d'une phrase correcte.

Erreurs du point de vue de la naturalité Les erreurs du point de vue de la naturalité sont celles qui font d'une phrase pourtant grammaticale des phrases fautives en sens.

Même si les insertions de morceaux de phrases n'affectent pas la grammaticalité, selon que les morceaux insérés sont compatibles ou pas, on peut avoir des cas où le sens n'est pas affecté, et des cas où il l'est. Les cas où le jugement est difficile ne sont pas les moins nombreux. Par exemple, les phrases suivantes produites par analogie pourrait très bien être prononcée dans un contexte particulier.

> ? このバッグの大きさなら機内にそこにあるもの何でも持ち込むことが出来ますか。
> ? すぐお持ちしません。[40]

La phrase suivante est possible grammaticalement, mais elle difficile à interpréter, et donc à imaginer dans une situation réelle.

> ?? ホテルの看板をそこにあるもの何でも持ってるそうなので見つけて下さい。[41]

Dans cet exemple, l'expression ホテルの看板を持ってる semble faire difficilement sens, et paraît donc peu naturelle. Il faut souligner cependant qu'il n'est pas rare en linguistique d'être confronté à de tels exemples fabriqués pour lesquels le jugement en sens paraît difficile.

[39] /Motte imasen/ (Je n'en ai pas.) : /sekken wo motte kite kudasai/ (Apportez du savon, s'il vous plaît.) = /o moti simasen/ (Vous n'en avez pas.) : /sekken wo o mote tekiti kudasasi/ La quatrième phrase est agrammaticale. Mais on ≪ sent ≫ qu'elle voudrait dire *Apportez du savon, s'il vous plaît.*

[40] (?Si c'est de la taille de ce sac, à bord, toutes les choses qui sont là, vous pouvez les emporter avec vous.)
(?Vous ne l'avez pas à l'instant.) ou (?Je ne vous l'apporte pas immédiatement.)

[41] Une traduction tirée par les cheveux serait : *Trouvez[-le], parce que tout ce qu'il y a là a l'air d'avoir l'enseigne de l'hôtel.*

Dans l'exemple suivant, si une pause est marquée après 問題は[42], alors on peut interpréter 禁止されています[43] comme une interrogation, et la phrase suivante produite par analogie devient possible en langue parlée.

日本のマンガをおみやげに持ってきていますが問題は禁止されています。[44]

Lorsque les phrases produites semblent provenir du collage de deux parties de phrases indépendantes, la phrase produite, même si elle est tout-à-fait correcte grammaticalement, peut poser un problème de cohérence au niveau du sens. L'effet peut même être comique, quand on obtient une contradiction.

?すみません，ウォンは持ち合わせていますので日本円で支払ってもいいですか。[45]

11.3.4 Synthèse sur la productivité d'un corpus

Lors de l'expérience que nous venons de rapporter, à partir des 153 phrases d'un corpus, nous avons produit des phrases nouvelles par analogie, à l'aide d'un algorithme de résolution d'équations analogiques. Pour cela, nous avons appliqué l'algorithme sur toutes les combinaisons possible de trois phrases. Le résultat a été la production de 299 phrases tant grammaticalement que sémantiquement correctes. Pour les autres 708 phrases produites, nous avons tenté une classification des erreurs, classification que nous avons illustrée par des exemples. La majorité de ces erreurs est provoqué par les commutations, au niveau des chaînes de symboles, soit à cause d'un lieu d'insertion impropre, soit à cause d'une incompatibilité contextuelle.

Dans l'expérience relatée ici, nous n'avons en aucune manière essayé de contraindre l'analogie afin d'éviter la production des phrases agrammaticales. Maintenant, le problème de la suppression de telles erreurs est de trouver comment imposer des contraintes grammaticales ou sémantiques sur la production de phrases par analogie.

Pour contrebalancer le caractère aveugle de l'analogie nous allons imposer que les chaînes soient mises en correspondance avec d'autres objets qui contraindraient les analogies sur les chaînes de symboles. Cette mise en correspondance peut être vue comme l'affectation à ces chaînes d'informations syntactiques, sémantiques ou pragmatiques. Cela est tout ce qu'il y a de plus classique dans le domaine du traitement automatique des langues. On peut imposer que ces objets appartiennent eux aussi à un espace de chaînes de symboles, mais différents. En plus, et là réside l'originalité de notre proposition, nous allons imposer que cet espace soit lui aussi structuré par l'analogie. C'est le sujet du

[42] /Mondai ha/ (le problème).

[43] /Kinsi sarete imasu/ (est interdit).

[44] (Comme j'ai amené des bandes dessinées japonaises, le problème, là, est-ce que c'est interdit ?).

[45] (Excusez-moi, comme j'ai des wons [unité monétaire coréenne] sur moi, est-ce que je peux payer en yens [unité monétaire japonaise] ?).

chapitre suivant. Nous allons y montrer comment les homomorphismes introduits plus haut (p. 194) permettent l'établissement de correspondances entre espaces analogiques qui brident la force aveugle de l'analogie.

Chapitre 12

Homomorphismes d'espaces analogiques

Avant de passer à des applications sérieuses de la notion d'homomorphisme entre espaces analogiques, nous allons, dans un premier temps, à des fins pédagogiques, décrire l'application de cette notion à un exemple simplissime. Nous allons en effet reprendre la conjugaison des verbes français du premier groupe aux temps simples (voir p. 245) et montrer comment une telle application peut profiter de la notion d'homomorphisme. Dans cet exemple, seul apparaîtra la notion d'homomorphisme. Celle de langages de chaînes analogiques ne sera pas utilisée. En utilisant exactement la même technique et par un simple changement du domaine d'application, nous montrerons comment réaliser une des opérations cruciales du traitement automatique des langues, l'analyse structurale des phrases. Nous avons testé cette technique sur des langues différentes mais aussi avec des descriptions structurales différentes. Nous présenterons les conclusions que nous en avons tirées.

Dans un deuxième temps, après ces deux exemples préliminaires, nous introduirons la notion de langage de chaînes analogiques. On fera jouer ensemble les homomorphismes et les langages de chaînes analogiques par une extension naturelle du point de vue informatique de la technique précédente. On appliquera simplement la technique précédente de façon récursive. Cette nouvelle technique a été testée dans une optique d'amélioration tant du point de vue de la rapidité que de la qualité des résultats, de la technique d'analyse structurale précédente. À cette fin, nous avons procédé à un certain nombre de mesures en faisant varier un certain nombre de paramètres. Nous rapporterons nos conclusions.

Enfin, dans un troisième temps (p. 314), nous n'aurons pas peur de nous attaquer au rêve de tout automaticien des langues : la traduction automatique. Pour reprendre la progression précédente, nous rapporterons d'abord des essais préliminaires sur une maquette entre le français et le japonais, puis sur un prototype de démonstration entre le japonais et l'anglais. Nous proposerons ensuite l'ébauche d'un système utilisant les homomorphismes entre langages de chaînes analogiques, mais nous étendrons la technique proposée précédemment à plusieurs domaines simultanément. Cette ébauche possède déjà, en théorie,

bien des aspects positifs prometteurs dont nous ferons la liste. Ces avantages énumérés, nous pourrons faire part de notre projet de réalisation d'un système de traduction automatique plus élaboré.

12.1 Conjugaison des verbes français

Nous reprenons ici l'exemple de la conjugaison par analogie (p. 245). Ce programme de conjugaison automatique peut profiter de la notion d'homomorphisme entre espaces analogiques pour prendre une forme plus compacte. Cette forme rejoint la précoccupation linguistique du nombre de formes minimales à connaître pour pouvoir faire usage d'un paradigme de flexions. Cela est évidemment lié à la représentativité, notion que nous avons abordée plus haut (p. 257). Illustrons notre propos par un exemple très simple : la conjugaison des verbes français du premier groupe[1]. Nous partons d'une liste partielle des formes conjuguées des verbes donner et marcher.

donner	*marcher*
je donne	*je marche*
tu donnes	*tu marches*
il ou elle donne	*il ou elle marche*
nous donnons	*nous marchons*
vous donnez	*vous marchez*
ils ou elles donnent	*ils ou elles marchent*

On peut lister toutes les formes déductibles par analogie avec des formes les précédant dans la liste, et donc par complémentation, obtenir l'ensemble des formes non déductibles des autres par analogie. Mathématiquement, conformément à ce que nous avons vu plus haut (p. 188), l'ensemble des formes non déductibles est une base. En effet, cet ensemble est libre car aucun élément de cet ensemble ne peut être reconstruit à partir d'autres éléments de ce même ensemble. De plus, cet ensemble est générateur car tout élément de l'ensemble de départ peut être reconstruit par analogie avec des éléments de cet ensemble. Cette base peut être obtenue en utilisant les mêmes algorithmes (voir p. 236) que ceux utilisés dans les expériences sur la représentativité (p. 259). Pour notre exemple simplissime, le résultat est le suivant :

donner	*marcher*
je donne	*donner* : *je donne* $\doteq$ *marcher* : *je marche*
tu donnes	*donner* : *tu donnes* $\doteq$ *marcher* : *tu marches*
il ou elle donne	*donner* : *il ou elle donne* $\doteq$ *marcher* : *il ou elle marche*
nous donnons	*donner* : *nous donnons* $\doteq$ *marcher* : *nous marchons*
vous donnez	*donner* : *vous donnez* $\doteq$ *marcher* : *vous marchez*
ils ou elles donnent	*donner* : *ils ou elles donnent* $\doteq$ *marcher* : *ils ou elles marchent*

On retrouve donc mécaniquement que la seule donnée nécessaire à la conjugaison très partielle donnée ici du verbe marcher, en plus des formes du verbe

[1] Nous reprenons ici en le modifiant l'article LEPAGE, *Formalisation de l'analogie entre chaînes de symboles*, 2001, p 126 à 128.

modèle donner, est l'infinitif *marcher*. De cette manière, il serait possible de retrouver le résultat avancé par Le Goffic[2] selon lequel il suffit de connaître seulement six formes particulières d'un verbe français pour être capable de le conjuguer entièrement.

Montrons maintenant comment le modèle des homomorphismes entre espaces analogiques permet la conjugaison des verbes. Le problème est celui de la mise en correspondance des formes du verbe avec une description de leur analyse. Il s'agit là d'une modélisation d'un domaine. Pour cela, retenons les connaissances extraites précédemment. Elles constituent un premier domaine. Associons-les à leurs représentations abstraites qui constituent un second domaine. On obtient le panorama suivant.

donner	`donner-inf`
je donne	`donner-ind-pres-1-sg`
tu donnes	`donner-ind-pres-2-sg`
il ou elle donne	`donner-ind-pres-3-sg`
nous donnons	`donner-ind-pres-1-pl`
vous donnez	`donner-ind-pres-2-pl`
ils ou elles donnent	`donner-ind-pres-3-pl`
marcher	`marcher-inf`

Il est intéressant de noter que, dans un tel procédé, les exceptions sont obtenues simplement lors du calcul de la base. D'un autre point de vue, cela correspond à un simple ajout des exceptions dans la base de connaissances.

essayer	`essayer-inf`
j'essaie	`essayer-ind-pres-1-sg`

Le principe est que chacun des domaines, celui de gauche comme celui de droite, sont des espaces analogiques indépendants. Les approches classiques en intelligence artificielle (voir p. 85) chercheraient à établir des analogies chevauchant les deux domaines (voir aussi le modèle de Nagao, p. 89) du type suivant.

$$\textit{marcher} : \texttt{marcher-inf} \doteq \textit{nous marchons} : \texttt{marcher-ind-pres-1-pl}$$

Rappelons que nous avions refusé cette vue pour des raisons méthodologiques. Ici, si nous avons noté les objets des deux domaines dans des corps de lettres différents, c'est précisément pour bien marquer que ces objets doivent rester incomparables. Ils ne sauraient être comparés directement. On ne saurait donc effectuer des calculs à cheval sur les deux domaines. Notre approche ne s'autorise à vérifier des analogies ou à résoudre des équations analogiques que dans un seul domaine à la fois. Ainsi donc, si l'on propose au système une forme nouvelle, par exemple *nous marchons*, le système recherche, dans le même domaine, trois formes en relation d'analogie avec la forme d'entrée. Ici, il trouve : *donner*, *nous donnons* et *marcher* qui forment l'analogie

[2] Le Goffic, *Les formes conjugées du verbe français – oral et écrit*, 1997, p. 30.

$$donner : nous\ donnons \doteq marcher : nous\ marchons$$

Le système transpose alors cette analogie aux second domaine, faisant passer l'analogie du domaine de gauche au domaine de droite (voir notre modèle, p. 203). Sur les quatre formes, seules trois ont un correspondant dans le second domaine, le dernier étant précisément celui que nous recherchons. On obtient donc l'équation analogique :

$$\texttt{donner-inf} : \texttt{donner-ind-pres-1-pl} \doteq \texttt{marcher-inf} : x$$

La résolution de cette équation analogique, en tant que chaînes de symboles d'un même domaine selon l'algorithme décrit plus haut (p. 213), produit la représentation abstraite `marcher-ind-pres-1-pl` qui peut alors être proposée comme analyse de la forme d'entrée *nous marchons*.

Une application d'apprentissage pourrait proposer d'ajouter cette nouvelle connaissance à la base.

$$nous\ marchons \quad | \quad \texttt{marcher-ind-pres-1-pl}$$

Ce serait le cas si, contrairement à ce que nous avons exposé, on était par exemple parti de la donnée des correspondances entre domaines, sans être passé par la réduction par calcul de la base, et si le but était précisément d'obtenir l'analyse de toutes les formes conjuguées des verbes et de les retenir.

Nous venons de montrer l'utilisation d'un tel système en analyse. Or il est clair qu'un tel système est par essence bidirectionnel. On peut s'en servir également pour la conjugaison automatique, c'est-à-dire pour la production de formes conjuguées. On peut reprendre l'interface que nous avions présentée plus haut et son formulaire. Pour l'exemple de la figure 10.3 (p. 248), la saisie d'informations doit trouver la correspondance entre les deux domaines pour ce qui est des infinitifs :

$$percevoir \quad | \quad \texttt{recevoir-inf}$$

si cette correspondance n'existe pas, et produire par assemblage des informations choisies par l'utilisateur la nouvelle chaîne de symboles suivante du domaine des représentations abstraites : `percevoir-inf-futur-1-sg`. Pour conjuguer la verbe marcher au présent de l'indicatif à la première personne du pluriel, le travail du formulaire aurait consisté à trouver la correspondance

$$marcher \quad | \quad \texttt{marcher-inf}$$

et à construire la chaîne de symboles `marcher-ind-pres-1-pl` dans le second domaine. À l'aide de cette nouvelle chaîne, l'analogie suivante peut être formée dans le domaine des représentations abstraites.

$$\texttt{donner-inf} : \texttt{donner-ind-pres-1-pl} \doteq \texttt{marcher-inf} : \texttt{marcher-ind-pres-1-pl}$$

La transposition de cette analogie dans le premier domaine des formes conjuguées nous donne une équation analogique dont la solution est la forme conjuguée attendue.

$$donner : nous\ donnons \doteq marcher : x \quad \Rightarrow \quad x = nous\ marchons$$

12.2 Analyse par analogie

La méthode que nous venons de voir peut être transposée facilement en changeant seulement les domaines. Il suffit que les nouveaux domaines soient des ensembles de chaînes de symboles pour que la méthode s'applique sans rien changer. Par exemple, si le premier domaine est un ensemble de formes écrites, et si le second domaine est l'ensemble des prononciations correspondant à ces formes écrites, on peut réaliser de la transcription graphémique-phonologique. Un autre exemple, pour lequel nous allons maintenant rapporter un certain nombre d'expériences est celui de l'analyse structurale de phrases. Le premier domaine est alors celui des phrases, simplement transcrites. Le second domaine est celui des représentations structurales des phrases, données sous une forme parenthésée, c'est-à-dire sous forme de chaînes de symboles. L'avantage de notre méthode est donc que, du point de vue informatique, elle peut être implantée à l'aide d'un unique moteur universel, indépendant des applications. Et c'est bien là la façon dont nos implémentations ont été réalisées.

L'analyse de phrases étant l'une des tâches cruciales pour maintes applications du traitement automatique des langues, nous avons appliqué notre méthode à cette tâche. Nous avons essayé d'analyser plus de 1 500 phrases dans deux langues différentes, avec des représentations structurales différentes, en constituants et en dépendance. Nous avons donc directement utilisé des corpus arborés. Plus précisément, nous ne partons pas des formes brutes des textes, mais de leur forme étiquetée. Le premier domaine est donc en fait un ensemble de chaînes de symboles représentant des classes morpho-syntaxiques. Le second domaine est l'ensemble correspondant des arbres sous leur forme parenthésée. Notre but ici est d'étudier les possibilités de la méthode en présentant une analyse des résultats obtenus.

12.2.1 Principe et méthode

Réexpliquons ici le principe de la méthode en illustrant son application à l'analyse structurale[3]. L'idée repose, bien sûr, sur l'hypothèse émise par Hermann Paul ou Bloomfield que l'analogie serait à l'œuvre en syntaxe dans la création de phrases nouvelles (p. 70). C'est aussi l'opinion d'Itkonen qui pense que la grammaticalité des phrases produites n'est assurée que lorsque l'analogie joue à la fois sur la surface et sur les représentations structurales[4]. L'analyse par analogie repose donc sur le principe suivant. Pour des phrases données, par exemple :

Comment est-ce ?	« Comment allez-vous ? » demanda-t-il.	Y êtes-vous ?	« Où suis-je ? » questionna-t-il.

[3]Cette partie reprend l'article LEPAGE, *Open set experiments with direct analysis by analogy*, 1999.

[4]Voir aussi ITKONEN & HAUKIOJA, *A rehabilitation of analogy in syntax (and elsewhere)*, 1997.

s'il existe une analogie à un certain niveau de représentation par exemple celui des classes morpho-syntaxiques :

$$\begin{array}{l} mnr\ vb \\ pn\ ? \end{array} : \begin{array}{l} \ll\ mnr \quad vb \\ pn\ ?\ \gg \quad vb \\ liaison\ pn \end{array} \doteq lieu\ vb\ pn\ ? : \begin{array}{l} \ll lieu\ vb\ pn\ ?\ \gg\ vb \\ liaison\ pn \end{array}$$

alors, il devrait y avoir analogie à un niveau de représentation plus élevé comme celui des représentations structurales (figure 12.1, p. 295)[5].

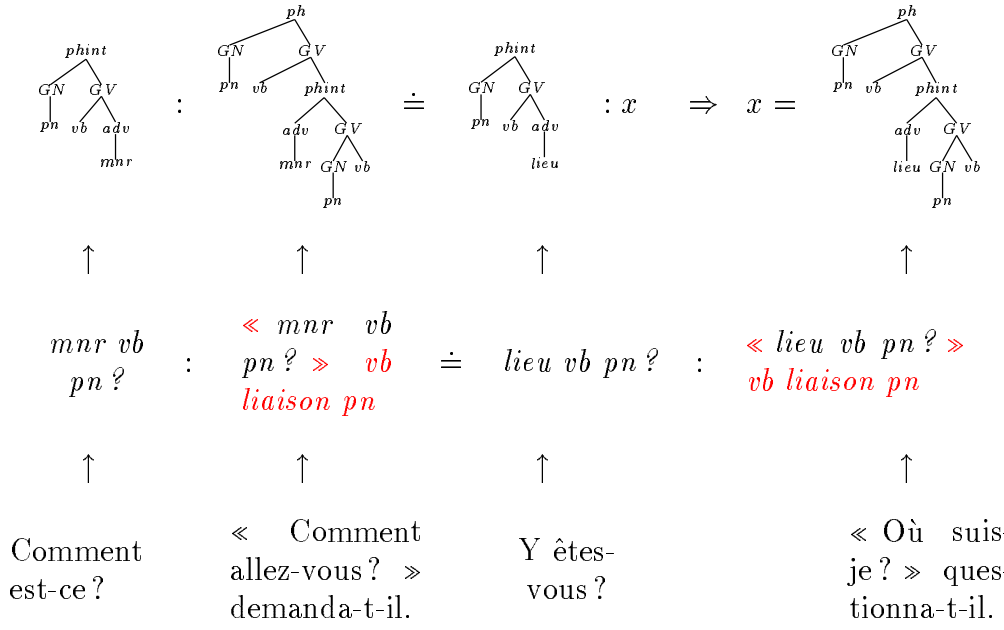


Figure 12.1: Analyse directe par analogie. Vue fondamentale

Analyse directe par analogie

La mise en application de ce principe implique l'utilisation d'un corpus arboré, c'est-à-dire d'un ensemble de phrases avec leur représentation linguistique associée[6].

Pour calculer l'analyse d'une phrase nouvelle, par exemple celle en haut et à gauche de la figure 12.2 (p. 296), on recherche dans l'ensemble des phrases du corpus arboré, que nous appelons $\mathcal{A}$, trois phrases qui soient en relation d'analogie avec la nouvelle phrase. Si de telles phrases existent, alors leurs structures linguistiques associées sont prises comme les termes d'une équation analogique dont la *résolution* fournit éventuellement une nouvelle structure linguistique. On posera alors que cette structure est l'analyse de

[5]Cette technique fait l'objet d'un brevet au Japon Lepage & 安藤　真一 (Andou Sin-Iti), 用例主導型言語構造解析装置, 2000.

[6]Historiquement, nous avons introduit cette méthode d'abord dans Lepage, *Un éditeur pour la construction de banques d'arbres*, 1996, p. 110, puis nous l'avons reprise dans Lepage & Ando, *Saussurian analogy: a theoretical account and its application*, 1996, p. 721,

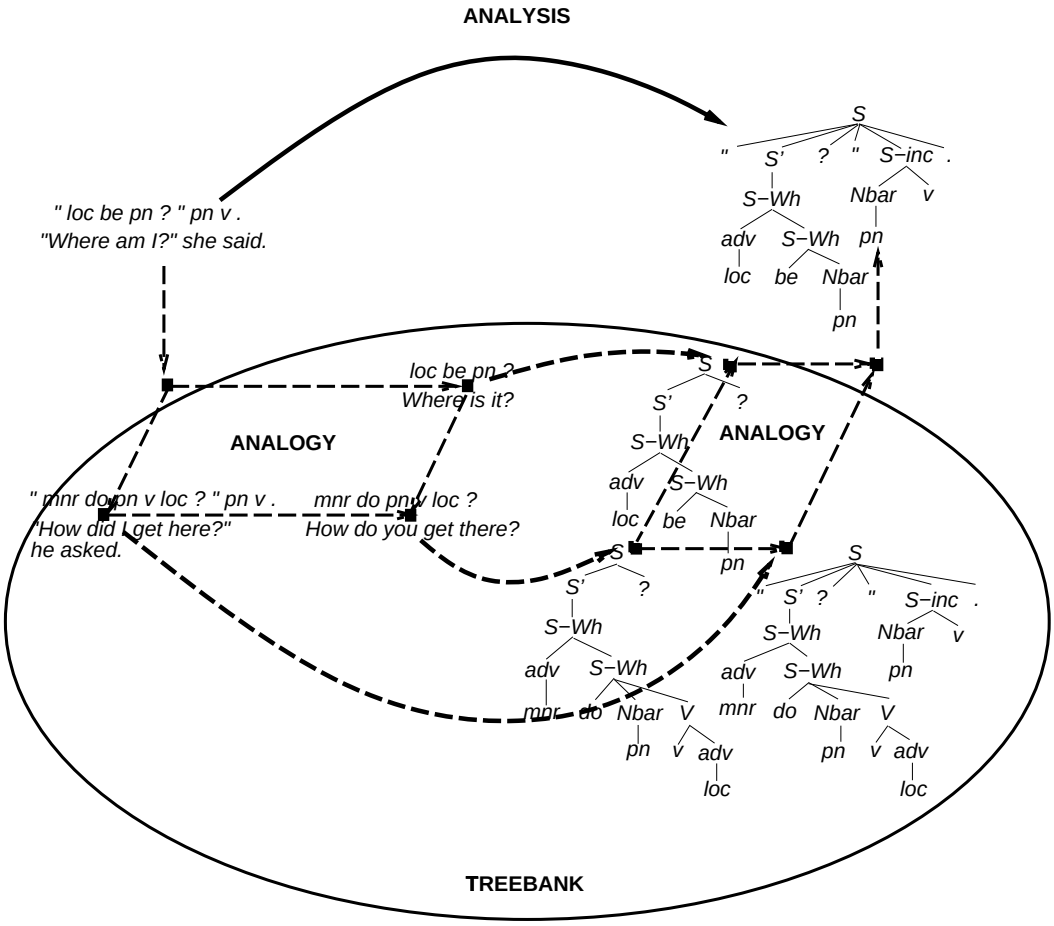


Figure 12.2: Analyse directe par analogie. Vue relative au corpus arboré. Le sens des flêches en pointillé montre l'écoulement du processus

la phrase nouvelle. Vérifier s'il s'agit bien de l'analyse de la phrase d'entrée est précisément l'objet des expériences présentées dans la suite de cet article. On comprend qu'il puisse exister plusieurs triplets de phrases formant une analogie avec une phrase nouvelle donnée. La technique présentée ici produit donc éventuellement plusieurs analyses possibles pour la même phrase. Et ces analyses peuvent être différentes ou égales. Autrement dit, la même analyse peut être produite de plusieurs façons possibles. Il semble raisonnable de penser que les analyses produites le plus grand nombre de fois pour la même phrase seront les meilleures. Nous retrouvons là une notion similaire à celle de fréquence, étudiée par Mańczak dans le fonctionnement diachronique de l'analogie (p. 68). En résumé, la méthode est schématisée dans la figure 12.2 (p. 296).

A priori, le coût de la recherche des triplets (A, B, C) dans $\mathcal{A}^3$ est cubique. Il pourrait être facilement divisé par deux en prenant en compte la permutation des moyens (p. 116). Alternativement, étant donné une phrase D, rechercher un triplet (A, B, C) de $\mathcal{A}^3$ tel que l'analogie $A : B \doteq C : D$ soit vraie, peut se réaliser de la façon équivalente suivante. À D donné, pour

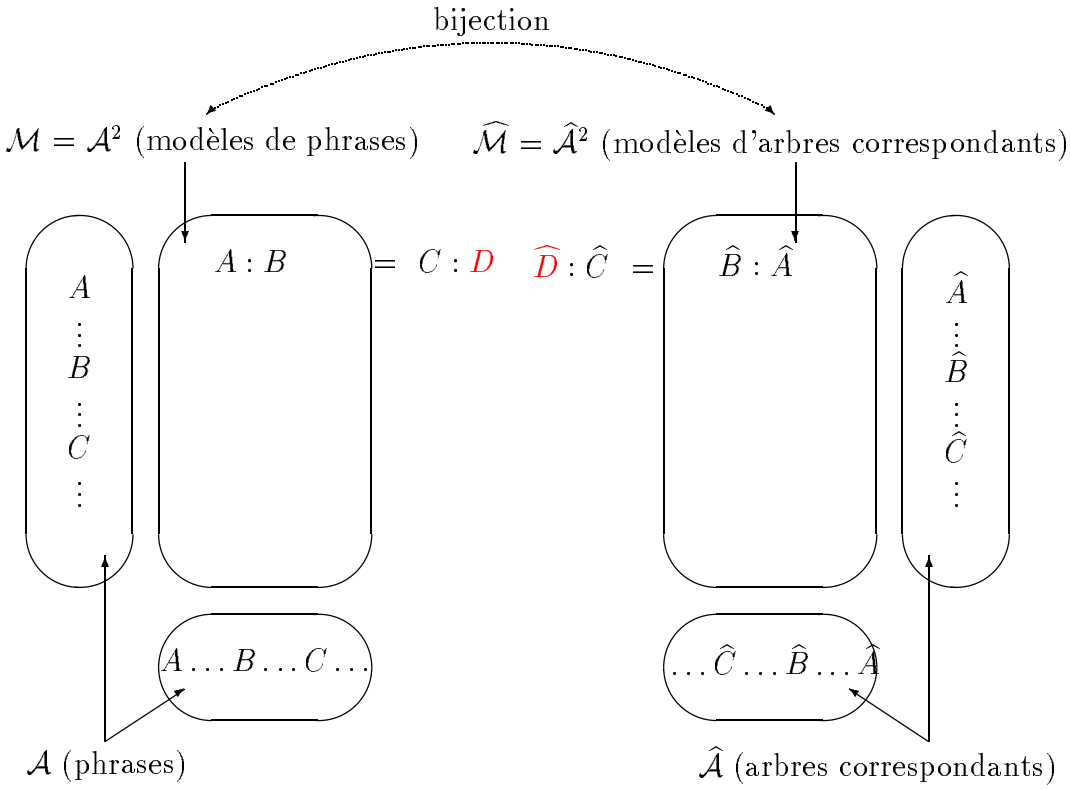


Figure 12.3: Analyse directe par analogie. Vue relative à l'homomorphisme. D est l'entrée, $\widehat{D}$ le résultat

chaque couple (A, B) de $\mathcal{A}^2$, on résoud l'équation analogique $B : A \doteq D : C$ équivalente par inversion des rapports (p. 116). Si une ou plusieurs solutions C existent, on n'a plus alors qu'à tester leur appartenance à $\mathcal{A}$. Bien sûr, nous n'indiquons pas cette façon de procéder par hasard. La raison en est que, théoriquement, l'étape de recherche apparaît alors comme la reconnaissance de l'appartenance de D à l'ensemble Λ_1 du langage de chaînes analogiques paresseux $\Lambda(\mathcal{A}, \mathcal{A}^2)$ (voir p. 174). Les couples (A, B) apparaissent comme des éléments de l'ensemble $\mathcal{M} = \mathcal{A}^2$ des modèles et l'on peut les noter $A \to B$. En plus, pratiquement, cette façon de procéder est très rentable. En effet, on passe d'un espace de recherche cubique à un espace de recherche quadratique. Le gain en temps est assuré parce que la vérification de l'appartenance de C à $\mathcal{A}$ peut se faire linéairement par des méthodes de hachage[7]. Et parce que la résolution des équations analogiques est en moyenne extrèmement rapide. Cela provient de ce que l'on a très souvent seulement à vérifier que l'équation analogique n'a pas de solution. Cela prend souvent un temps sous-linéaire en la taille des chaînes. Dans nos expériences, proportionnellement, pour un D

[7]Dans notre implémentation, elle est logarithmique car nous utilisons des arbres semi-équilibrés à la Adel'son-Velskiĭ et Landis.

donné, le nombre d'équations analogiques n'ayant pas de solutions est énorme. Elles représentent plus de 99,9% des équations analogiques considérées !

Le schéma de la figure 12.2 (p. 296) présente la méthode sous l'angle du corpus arboré. Il considère le corpus arboré comme un tout. Une présentation plus conforme à nos vues passe par la notion d'homomorphisme. Dans cette vue, le domaine des phrases, ou plus exactement des chaînes de classes morpho-syntaxiques, et le domaine des représentations structurales doivent être séparés. Entre eux existe une mise en correspondance. À chaque phrase correspond une représentation structurale. Il est important de souligner que, si les domaines sont considérés comme des ensembles, la correspondance n'est pas forcément bijective. Premièrement, elle n'est pas forcément injective, car un seul élément du domaine des chaînes de classes morpho-syntaxiques peut représenter plusieurs phrases et ces phrases peuvent s'être vues attribuer des descriptions structurales différentes. Par conséquent, un même élément du domaine des chaînes de classes morpho-syntaxiques peut correspondre à plusieurs élements du domaine des représentations structurales. Deuxièmement, elle n'est pas forcément surjective, car une même représentation structurale peut éventuellement correspondre à des phrases n'ayant pas la même représentation au niveau des classes morpho-syntaxiques. C'est typiquement le cas si les représentations structurales normalisent certains phénomènes de surface.

Nous baptisons la méthode présentée jusqu'à présent *analyse directe par analogie*, par opposition à la méthode étendue aux langages de chaînes analogiques que nous allons voir maintenant. En effet, dans ce que nous venons d'exposer, l'analyse d'une phrase nouvelle donnée doit s'effectuer du premier coup. Nous venons de voir que la recherche des triplets revenait à la reconnaissance de l'appartenance de D à l'ensemble Λ_1 du langage de chaînes analogiques paresseux $\Lambda(\mathcal{A}, \mathcal{A}^2)$ où $\mathcal{A}$ est l'ensemble des chaînes de classes morpho-syntaxiques. Nous pouvons donc à nouveau schématiser la méthode, mais cette fois-ci à la lumière des homomorphismes entre domaines (voir figure 12.3, p. 297). Nous avons presque déjà des homomorphismes entre langages de chaînes analogiques. La représentation du corpus arboré sous une forme compatible avec la notion d'homomorphisme en fait deux ensembles munis d'une correspondance entre eux. L'ensemble des chaînes de classes morpho-syntaxiques est noté $\mathcal{A}$. Celui des formes parenthésées des représentations structurales est noté $\widehat{\mathcal{A}}$.

Analyse récursive par analogie

Dans la méthode précédente, lorsque la phrase C obtenue par résolution de l'équation analogique $B : A \doteq D : C$ n'appartient pas au corpus arboré, il est naturel d'envisager une application récursive de la méthode pour essayer d'obtenir une analyse de cette phrase, et après retour en arrière, une analyse pour la phrase d'entrée D. Cette méthode est illustrée dans la figure 12.4 (p. 299). Nous l'appelons *analyse récursive par analogie*. Cela revient à tester l'appartenance de D à $\Lambda(\mathcal{A}, \mathcal{M})$ entier. Pratiquement, cette façon de faire est dangereuse, car le calcul peut éventuellement ne pas se terminer.

Le nombre d'applications récursives de la méthode peut être limité par un seuil donné r. Dans le cas de l'analyse directe par analogie, ce seuil est de 1. Avec un seuil de r, cela revient à tester l'appartenance de D à Λ_r. Comme nous l'avons dit plus haut, intuitivement, l'utilisation d'un seuil supérieur à 1 devrait accroître le nombre de phrases analysées.

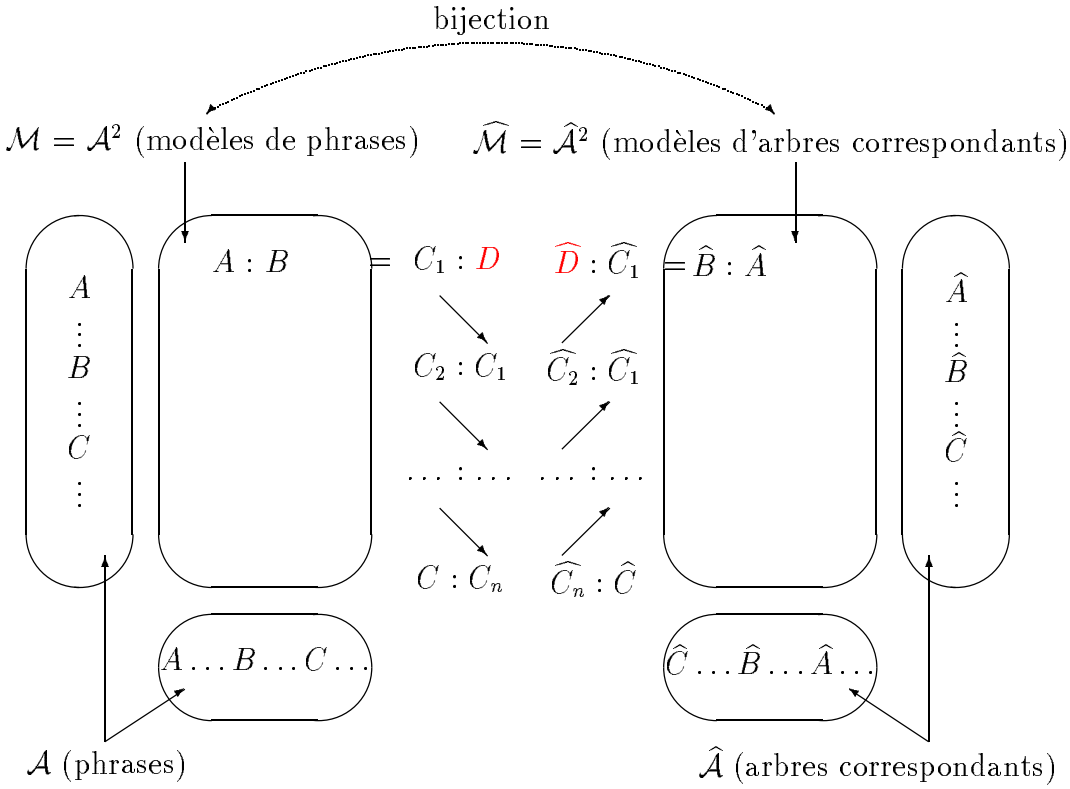


Figure 12.4: Analyse récursive par analogie. Vue relative aux langages de chaînes analogiques. D est l'entrée, $\widehat{D}$ le résultat

Convergence de la méthode

Une autre façon d'éviter que les calculs ne se terminent pas est la suivante. Bien qu'il n'y ait formellement aucune raison pour que les longueurs de $C_1, \ldots, C_n$ diminuent, cela aurait l'avantage d'impliquer la convergence de l'application récursive de la méthode. En effet, dans ce cas, C_n devrait tendre vers la chaîne vide. Cette propriété peut être obtenue en imposant la restriction que la longueur des B soit plus grande que celle des A pour tous les modèles $A \to B$ de $\mathcal{M}$. Le langage considéré est alors, par définition (p. 173), un langage de chaînes analogiques décroissant.

12.2.2 Protocole d'expérimentation

Pour vérifier la possibilité de réalisation et surtout la validité des différentes méthodes proposées, nous avons réalisé un certain nombre d'expériences dans lesquelles nous avons, autant que faire ce peut, utilisé toujours les mêmes données. Dans tous les cas, nous avons essayé d'analyser 1 553 nouvelles phrases avec un corpus arboré de 5 000 phrases. Nous l'avons fait dans deux langues différentes. Il s'agit d'expériences *ouvertes* parce que les *données de base* sont utilisées pour l'analyse de phrases nouvelles appartenant à un ensemble distinct, appelé *ensemble de test*. Cela s'oppose à des expériences *fermées* pour lesquelles les 1 553 phrases auraient été prises parmi les 5 000 phrases des données de base. Les termes de *données de base* et *ensemble de test* étant pratiques et usuels nous les emploierons dans la suite pour décrire nos expériences.

Deux langues, deux types de représentations linguistiques Nos données proviennent de deux corpus arborés, ATR-Lancaster[8] et ATR-NEC[9], différents sous plusieurs aspects :

- l'un est un corpus de phrases anglaises, l'autre est en japonais ;
- l'un consiste en des textes recopiés au hasard de l'internet, l'autre est limité à un domaine puisqu'il consiste en dialogues de réservation de chambres d'hôtel ;
- l'un utilise des représentations en constituants, l'autre des représentations en dépendance.

Données de base Lors des expériences, nous utilisons 5 000 phrases comme données de base. Leurs caractéristiques figurent dans la table 12.1 qui montre les tailles moyennes des phrases (en tant que séquences de classes morpho-syntaxiques) et des représentations linguistiques, ainsi que leur distribution.

Ensembles de test Nous avons tenté l'analyse de 1 553 phrases anglaises ou japonaises à l'aide des données de base décrites précédemment[10]. Rappelons de nouveau que les entrées de l'analyse sont des chaînes de classes morpho-syntaxiques. Nos expériences supposent une détermination automatique des

[8]Corpus anglais, rassemblé et annoté en collabration entre ATR et l'université de Lancaster. Voir BLACK *et al.*, *Beyond skeleton parsing: Producing a comprehensive large–scale general–english treebank with full grammatical analysis*, 1996.

[9]Corpus rassemblé à ATR et annot'é à NEC sous notre direction. Les structures linguistiques utilisées, fondées sur TESNIÈRE, *Eléments de syntaxe structurale*, 1959, s'appliquent particulièrement bien à la description du japonais. Voir LEPAGE *et al.*, *An annotated corpus in japanese using Tesnière's structural syntax*, 1998 et aussi LEPAGE, *Tesnière's structural syntax: notations for tree-banking using* `boardedit`, 1996.

[10]1 553 est un nombre « au hasard ». En fait, la taille du corpus arboré japonais à notre disposition était de 6 553 phrases, si bien que nous avons pris les 5 000 premières phrases comme données de base, et le reste comme ensemble de test.

Tableau 12.1: Caractéristiques des données de base.

tailles	min	max	moyenne ± écart type	nbre de classes ≠ ou d'étiq. de nœud ≠
phrases anglaises	1	99	12,19 ± 11,16	626
arbres anglais	4	299	33,19 ± 30,48	746
phrases japonaises	1	33	7,89 ± 3,61	201
arbres japonais	1	50	10,71 ± 5,77	208

Tableau 12.2: Caractéristiques des ensembles de test (chaînes de classes morpho-syntaxiques).

tailles	min	max	moyenne ± écart type	nbre de classes ≠
phrases anglaises	1	72	8,28 ± 8,94	386
phrases japonaises	1	25	8,58 ± 3,85	133

Tableau 12.3: Caractéristiques des ensembles de test (structures linguistiques).

tailles	min	max	moyenne ± écart type	nbre d'étiq. de nœud ≠
arbres anglais	4	199	22,85 ± 24,78	475
arbres japonais	1	39	11,81 ± 6,25	141

classes morpho-syntaxiques des mots des phrases, et pour le japonais un processus de segmentation. Cela n'est pas irréaliste, car on dispose maintenant d'outils relativement fiables pour réaliser ces opérations.

En fait, notre propos est d'évaluer les différents modes d'analyse par analogie. Or, pour chacune des 1 553 phrases de l'ensemble de test, nous connaissions déjà les chaînes de catégories morpho-syntaxiques et les structures linguistiques associées. Ces structures sont évidemment celles que nous souhaitons idéalement obtenir par analyse. C'est pourquoi nous les appelons *réponses exactes*. Les résultats obtenus lors des expériences ont donc été comparés à elles.

Évaluation La qualité des résultats a d'abord été examinée en comptant le nombre d'analyses produites par phrase, et plus particulièrement le nombre de réponses exactes obtenues. En traitement automatique des langues par méthodes statistiques, on évalue aussi souvent les résultats à l'aide d'une comparaison structurale entre les analyses produites et les réponses exactes (nombre de chevauchements) par le décompte des inclusions géométriques entre les deux structures, mais sans prendre en compte les différences d'étiquettes de nœud[11]. Nous avons utilisé ce critère habituel sur les résultats japonais. Nous avons aussi utilisé une méthode plus précise en calculant la distance d'édition entre les analyses produites et les réponses exactes[12]. Ici, la distance d'édition compte les insertions, les suppressions et les remplacements avec un même poids. Avec cette mesure, non seulement les étiquettes sont comparées, mais toute différence structurale est comptabilisée. Cette mesure caractérise mieux les analyses produites et permet en plus de donner la répartition des analyses par rapport aux réponses exactes. La qualité d'une analyse est inverse de sa distance à la structure exacte (une distance de 0 signifie que l'analyse est égale à la structure exacte).

12.2.3 Expériences d'analyse directe

Couverture Une première vue des résultats des expériences est donnée par la table 12.4 (p. 304). Elle donne la couverture, c'est-à-dire la proportion des phrases des ensembles de test pour lesquelles au moins une analyse a été obtenue: 49% pour l'anglais, 70% pour le japonais. Cela donne aussi évidemment la proportion de phrases pour lesquelles aucune analyse n'a été obtenue soit parce qu'aucune analogie n'a pu être trouvée au niveau des classes morpho-syntaxiques, soit parce qu'aucune équation analogique n'a pu être résolue entre structures linguistiques : 51% pour l'anglais, 30% seulement pour le japonais, En recherche documentaire, la couverture serait l'équivalent

[11]Voir Black *et al.*, *Beyond skeleton parsing: Producing a comprehensive large–scale general–english treebank with full grammatical analysis*, 1996.

[12]Il s'agit d'une extension de la distance d'édition entre chaînes aux arbres. Voir Selkow, *The tree-to-tree editing problem*, 1977. Formellement, les deux distances, entre chaînes et entre arbres, s'étendent aussi naturellement aux forêts. Voir Lepage, *Non-directionality and self-assessment in an example-based system using genetic algorithms*, 1994, p. 618.

du rappel, autrement dit le contraire du silence. Le silence est simplement le nombre de phrases qui n'ont pas été analysées par la méthode.

Cette table donne aussi le nombre de phrases pour lesquelles un réponse exacte a été obtenue au moins une fois : 32% pour l'anglais, et 58% pour le japonais. On peut considérer que ces résultats sont assez élevés, si l'on observe que la taille des ensemble de test n'est qu'un peu plus d'un tiers de celle des données de base.

Nous avons déjà montré que le processus d'analyse directe par analogie réalise en fait le test d'appartenance à Λ_1 des phrases à analyser. Intuitivement, un seuil supérieur à 1 devrait augmenter la couverture, c'est-à-dire le nombre de phrases pour lesquelles on devrait obtenir une analyse. Nous allons examiner cette hypothèse, hélas avec un succès mitigé, dans des expériences présentées plus bas.

Nombre d'analyses par phrases Rappelons qu'il peut y avoir plusieurs analyses pour une même phrase. La table 12.5 (p. 304). donne le nombre total d'analyses produites. Elle donne aussi le nombre de réponses exactes obtenues, avec, pour les données japonaises, le nombre d'analyses sans chevauchement structural avec la structure exacte. Ces résultats peuvent être comparés aux résultats de la première table pour obtenir le nombre d'analyses produites en moyenne par phrase. En moyenne donc, il y a $14\ 560\ 531/1\ 553 = 9\ 376 \approx 9 \times 10^3$ analyses (éventuellement égales) par phrase anglaise et $20\ 180\ 932/1\ 553 = 12\ 994 \approx 13 \times 10^3$ analyses par phrase japonaise. Pour que la méthode soit efficace, il est nécessaire que ce nombre très élevé soit contre-balancé par la fiabilité des analyses. Avant d'aborder ce point, observons que le fait que le nombre d'analyses produites sans chevauchement avec la structure exacte soit de 97% pour le japonais montre que les analyses produites ne sont pas réparties au hasard, mais qu'elles se concentrent heureusement autour des réponses exactes.

Fiabilité Il est très satisfaisant d'observer que les réponses exactes surpassent en nombre les autres analyses. Pour l'anglais, leur proportion est de 59% du total des analyses ; pour le japonais, cette proportion s'élève à 79% (voir la table 12.5, p. 304). Ces proportions sont l'équivalent de la précision en recherche documentaire, c'est-à-dire le contraire du bruit. Intuitivement, en effet, le bruit est le nombre d'analyses inexactes produites.

Le problème posé par toute technique produisant plusieurs résultats est le choix de la meilleure solution parmi les résultats proposés. Or, les graphes précédents peuvent être interprétés comme montrant que les analyses les plus fréquentes sont aussi les meilleures. Cela compenserait heureusement le nombre élevé d'analyses produites par phrase. Nous avons déjà émis cette hypothèse plus haut.

Nous avons aussi examiné globalement les meilleurs analyses, c'est-à-dire les analyses produites qui sont les plus proches des réponses exactes. Pour l'anglais, 85% des meilleures analyses sont des réponses exactes ; pour le japonais, 95%. On ne peut qu'en conclure que la méthode est assez fiable. La

Tableau 12.4: Répartition des phrases par analyse

	analyse produite pour la phrase ?				nombre
	exacte	sans chevauchement	oui	non	phrases total
anglais	495 (32%)	- -	767 (49%)	786 (51%)	1553 (100%)
japonais	903 (58%)	1051 (68%)	1089 (70%)	464 (30%)	1553 (100%)

Tableau 12.5: Répartition du nombre d'analyses produites

nombre d'analyses	exactes	sans chevauchement	total
anglais	8 528 412 (59%)	-	14 560 531
japonais	16 018 025 (79%)	19 579 758 (97%)	20 180 932

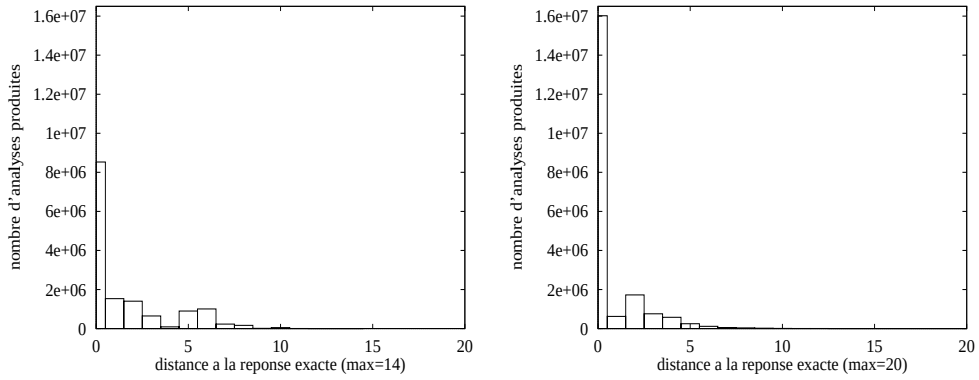


Figure 12.5: Distributions des analyses selon leurs distances aux réponses exactes (à gauche : anglais, à droite : japonais)

distribution globale des analyses selon leurs distances aux réponses exactes est donnée dans la figure 12.5 (p. 304).

Pour une phrase donnée, considérée individuellement, on souhaiterait que la meilleure analyse soit aussi la plus fréquente dans l'ensemble des analyses produites. Cette hypothèse est heureusement confirmée en première approximation : pour l'anglais, sur les 495 phrases pour lesquelles au moins une réponse exacte est obtenue, la meilleure analyse, c'est-à-dire la réponse exacte, est aussi la plus fréquente pour 468 phrases ; cela représente déjà $468/767 = 61\%$ de toutes les phrases analysées. C'est aussi le cas pour 616 des 767 phrases ayant une analyse (80%), lorsque les analyses à la même distance de la structure exacte sont comptabilisées ensemble[13]. Pour le japonais, les chiffres correspondants sont de 640 phrases exactement analysées, déjà $640/1089 = 59\%$ de toutes les phrases analysées, et de 744 sur 1089 phrases analysées, soit 68%. Retenir l'analyse la plus fréquente comme la meilleure semble donc judicieux pour l'analyse par analogie.

Il faut remarquer que la fiabilité des analyses les plus fréquentes est plus élevée pour les données anglaises, bien que le nombre de phrases analysées soit inférieur. Cela s'explique sans doute par le nombre plus élevé de classes morpho-syntaxiques utilisées.

Tout ce que nous venons de dire semble confirmer l'intuition que l'on a de la fréquence des analyses produites par une telle méthode : si une analyse est produite en grand nombre, c'est qu'elle reflète un plus grande regularité contenue dans le corpus. Elle ne saurait donc qu'en être meilleure.

Taille du vocabulaire et taille de l'ensemble de test Les deux expériences pécédentes, réalisées l'une avec un ensemble de classes morpho-syntaxiques relativement grand (environ 700) et des phrases longues et l'autre avec un ensemble plus petit (200 classes morpho-syntaxiques) et des phrases plus courtes, montrent que les résultats dépendent fortement de ces deux paramètres. Les données de base anglaises étaient une collection de phrases ouvertes, qui sont donc plus longues pour cette raison, alors que les phrases japonaises partageaient le même domaine, celui des réservations d'hôtel.

La taille des données de base joue certainement aussi un rôle. Sans doute, plus elle est importante, plus les chances de trouver des analogies sont élevées.

En conséquence, la technique est sans doute sensible à l'équilibre entre la taille des données de base, la taille du vocabulaire des classes morpho-syntaxiques, et la longueur des phrases et la taille des structures linguistiques dans les données de base. Des expériences restent à réaliser pour préciser l'influence de tous ces paramètres.

Phrases utiles Appelons *triplets utiles* les triplets fournissant des analyses. Comme le nombre de triplets examinés par notre algorithme est $5\,000 \times 5\,000/2 \approx 12 \times 10^6$ dans chaque expérience, la proportion de triplets

[13] Il peut y avoir plusieurs meilleures analyses lorsque l'on n'obtient pas la réponse exacte. Un calcul plus précis devrait les compter séparément.

utiles est approximativement de $9 \times 10^3/12 \times 10^6 = 0,075\%$ pour les données en anglais et de $13 \times 10^3/12 \times 10^6 \approx 0,1\%$ pour les données en japonais. Il peut être intéressant de considérer les phrases qui apparaissent dans les triplets utiles. On peut alors appeler de telles phrases des *phrases utiles* ou *représentatives*. Ces phrases sont pertinentes pour l'analyse des phrases de l'ensemble de test. À ce titre, elles peuvent constituer un point de départ pour construire une base regroupant les phrases dont on sait qu'elles sont utilisées dans l'analyse de données extérieures aux données de base. On peut alors songer à des applications dérivées. Par exemple, si l'ensemble de test est un texte en tant que tel, et si chaque phrase des données de base est étiquetée par un étiquette reflétant son genre, son domaine, etc., une application possible pourrait être de caractériser le genre ou le domaine du texte en examinant les étiquettes des phrases utiles qui servent à l'analyser.

12.2.4 Expériences de réduction de l'ensemble des modèles

Dans les expériences précédentes, certaines questions pratiques restent en suspens. Premièrement, bien que nous soyons passé à une recherche quadratique en la taille du corpus, est-il possible d'accélérer la méthode en réduisant encore l'espace de recherche ? Deuxièmement, l'utilisation de la récursivité permettrait-elle vraiment d'augmenter la couverture de l'analyse par analogie, c'est-à-dire le nombre de phrases pour lesquelles on obtient une analyse[14] ?

Nous allons d'abord nous intéresser à la réduction de l'ensemble des modèles, pour ensuite examiner l'influence de la récursivité

Dans les expériences précédentes, le modèle sous-jacent était celui des langages de chaînes analogiques paresseux (p. 174). L'ensemble $\mathcal{M}$ des modèles est le produit cartésien $\mathcal{A} \times \mathcal{A}$ de l'ensemble $\mathcal{A}$ des phrases. Son cardinal est donc $|\mathcal{A}|^2$. Dans les expériences précédentes, $|\mathcal{A}| = 5000$, et le produit cartésien a donc vingt-cinq millions d'éléments ! Dans des applications raisonnables, on peut envisager que l'ensemble $\mathcal{A}$ atteignent quelques centaines de milliers de phrases ou d'arbres attestés. La taille de l'ensemble des modèles $\mathcal{M}$ devient alors gigantesque, et l'on peut s'interroger sur la pertinence de considérer balourdement tous les couples de $\mathcal{A}$. La paresse n'est pas toujours économique. Il est souhaitable de réduire d'une manière ou d'une autre.

Mesures Les expériences rapportées maintenant ont été réalisées seulement sur le corpus japonais décrit plus haut. Comme précédemment, la qualité des résultats a été estimée par comptage du nombre d'analyses produites par phrase, et en particulier, par le nombre de fois où la réponse exacte est obtenue.

De même que précédemment nous avons aussi calculé la distance entre chaque analyse produite et la réponse exacte. Nous visualiserons ces résultats sous la forme de graphes où l'abscisse représentera toujours l'erreur relative

[14] Cette partie reprend les résultats présentés dans Taillefer & Lepage, *A series of experiments with recursive analysis by analogy*, 2000.

c'est-à-dire la distance à la réponse exacte divisée par la taille de la réponse exacte. Par exemple, une erreur de 0,5 signifie que l'indexeur du corpus arboré devrait modifier la moitié de l'analyse produite pour obtenir la réponse exacte[15]. En ordonnées, nous portons le nombre d'analyses produites ayant la même erreur relative. Par exemple, une valeur de 15 pour une erreur relative de 0,5 signifie que la méthode a produit 15 analyses contenant autant d'erreurs que la moitié de la taille de la réponse exacte. En plus, il faut noter que la barre située en zéro donne directement le nombre de réponses exactes obtenues.

Restriction aux langages décroissants

À la suite de notre proposition de faire converger les C_n vers la chaîne vide, nous imposons la restriction que tous les modèles $A \rightarrow B$ soient tels que $|A| < |B|$. Cette restriction implique la diminution en taille des C_n. Le langages sous-jacent est alors un langage de chaînes analogiques décroissant (voir p. 173). Dans la pratique, cette restriction divise le nombre de couples de $\mathcal{A}^2$ à considérer par deux. Une conséquence immédiate est donc une diminution du nombre de modèles qui réduira d'autant le temps d'exécution de la méthode. Nous allons montrer ci-dessous que, fort heureusement, cela n'affecte pas la qualité des résultats de façon notable.

Pour examiner l'influence de la restriction, nous avons tracé les graphes pour différentes valeurs de k, dans les deux cas suivants :

- sans la restriction (figure 12.6).
- avec la restriction (figure 12.7).

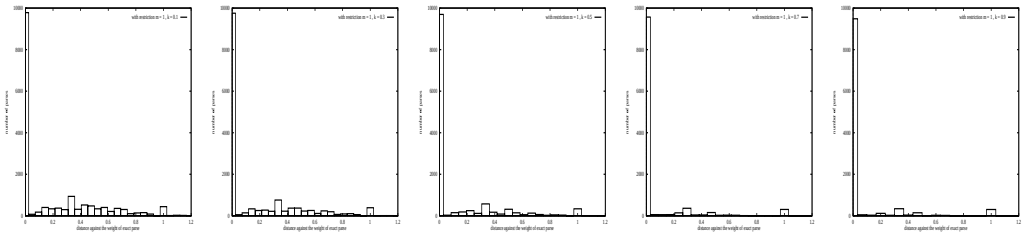


Figure 12.6: Nombre de modèles sans restriction. En abscisses, distance à la réponse exacte divisée par sa taille ; en ordonnées, nombre d'analyses produites.

Aucune différence majeure n'est visible entre les deux cas. Une analyse plus fine montre en fait qu'il y a moins de mauvaises analyses avec la restriction que sans, mais en nombre peu significatif. En résumé :

[15] C'est la situation que nous avions considérée avec un éditeur de textes et arbres comme dans Lepage, *Un éditeur pour la construction de banques d'arbres*, 1996 et dans nos expériences d'extension d'un corpus arboré rapportées dans Auclerc & Lepage, *Aides à l'analyse pour la construction de banque d'arbres : étude de l'effort*, 2001.

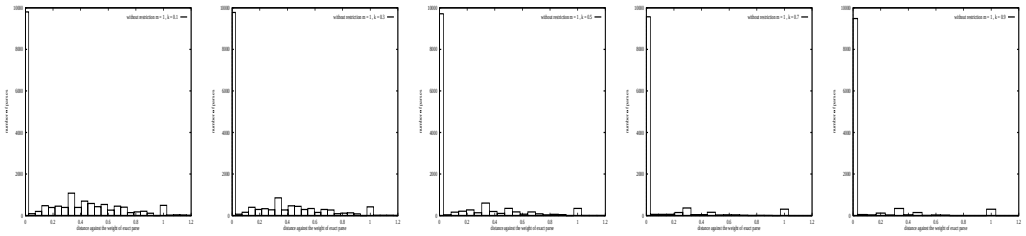


Figure 12.7: Nombre de modèles avec restriction. En abscisses, distance à la réponse exacte divisée par sa taille; en ordonnées, nombre d'analyses produites.

- dans les deux cas, il y a plus de bonnes analyses que de mauvaises;
- la restriction semble donc profitable:
  - elle ne change pas la nature des analyses;
  - elle réduit légèrement le nombre des mauvaises analyses;
  - elle augmente donc la fiabilité exprimée comme le rapport du nombre de réponses exactes au nombre total d'analyses produites.

Contrainte de proximité sur les modèles

Certaines propriétés de l'analogie nous autorisent à n'appliquer que les modèles qui aient la meilleure chance de fournir des solutions. Seuls ces modèles pourraient constituer l'ensemble $\mathcal{M}$ des modèles. Aussi, peut-on envisager l'introduction de contraintes pour sélectionner seulement les modèles utiles. Mais il y a danger que la réduction en temps de calcul obtenue se fasse au détriment de la qualité des résultats. Afin de mesurer l'influence de telles contraintes sur la qualité des résultats, nous aurons soin de proposer des contraintes paramétées. La contrainte peut correspondre à un trait particulier du modèle $A \rightarrow B$ utilisé dans l'analogie $A : B \doteq C : D$ où D est la phrase de départ, $A \rightarrow B$ un modèle de $\mathcal{M}$, et C la solution de l'équation analogique.

Plusieurs contraintes peuvent être considérées:

1. $|A| = k \times |B|$ avec $0 < k < 1$.
2. $|A| = k \times (\sigma(A,B) + \sigma(A,C))$ et où σ est la similitude. Cette contrainte provient de la propriété suivante de l'analogie: $A : B \doteq C : D \quad \Rightarrow$ $|A| \leq \sigma(A,B) + \sigma(A,C)$
3. $2 \times \sigma(A,B)/|A| + |B| \geq k$ avec $0 < k < 1$. Cette contrainte aussi provient de la même propriété.

Nous nous sommes intéressé seulement à la troisième contrainte. En effet, elle semble plus intuitive que la seconde et s'exprime seulement à l'aide d'un pourcentage donné par k. Aussi, ne jouant que sur $A \rightarrow B$, elle peut s'appliquer dans un prétraitement global sur l'ensemble $\mathcal{M}$ des modèles. Sous

une forme équivalente (voir p. 102), elle impose un seuil sur les distances entre A et B :

$$\delta(A,B) \leq (1-k) \times |A| + |B|$$

Elle exprime donc une propriété intuitive et globale de proximité entre les chaînes constitutives des modèles $A \rightarrow B$.

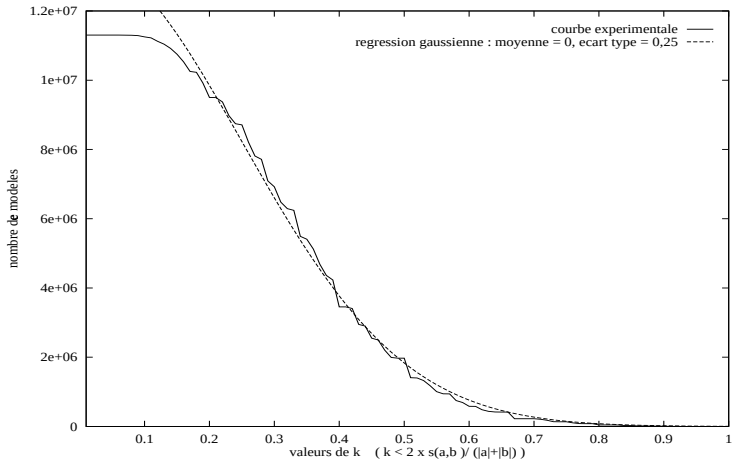


Figure 12.8: Variation de la taille de l'ensemble des modèles pour k variant de 0 à 1

Influence sur la taille de l'ensemble $\mathcal{M}$ des modèles Le graphe 12.8 montre l'évolution de la taille de l'ensemble des modèles pour la contrainte $\frac{2 \times \sigma(A,B)}{|A|+|B|} \geq k$ avec une centaine de valeurs de k comprises entre 0 et 1. Le plus grand nombre de modèles obtenu est d'à peu près 11 millions. Pour $k = \frac{1}{4}$, on en obtient environ 9 millions. Pour $k = \frac{1}{2}$, environ cent quarante mille. Le nombre maximal théorique de modèles est de 12 millions et demie. Cette valeur est atteinte lorsque l'on calcule l'ensemble des modèles sans la contrainte et sans compter les répétitions. On note que la courbe a le profil d'une courbe de Gauss dont l'équation générale est :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2} \times \frac{(x-m)^2}{\sigma^2}\right)$$

avec m la moyenne et σ l'écart type. Dans notre cas, la courbe est centrée, donc $m = 0$. Nous avons calculé les paramètres de la régression gaussienne, qui est montrée en pointillés sur le graphe.

En résumé, on peut dire que :

- la taille de l'ensemble $\mathcal{M}$ des modèles décroît quand k augmente. C'était bien là l'effet recherché.
- la courbe a l'air d'une gaussienne

– cela signifierait que la contrainte k ne changerait pas la distribution des données.
– cela confirmerait que la taille des phrases de notre corpus a une « bonne » distribution.

Influence sur la qualité des analyses

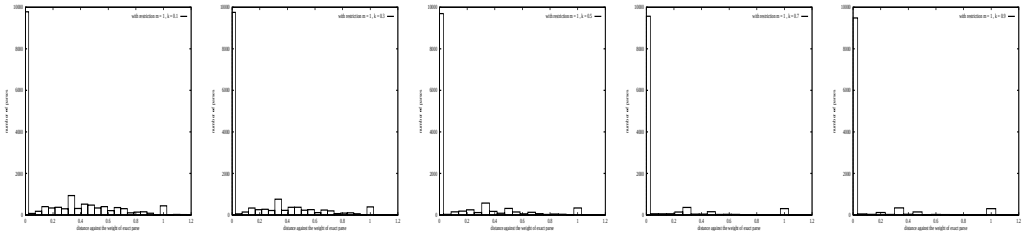


Figure 12.9: Influence de la contrainte. En abscisses, distance à la réponse exacte divisée par sa taille ; en ordonnées, nombre d'analyses produites.

Qualité pour k=0,1, 0,3, 0,5, 0,7 et 0,9. Les graphes de la figure 12.9 donnent les résultats obtenus pour différentes valeurs de k. On note que, comme nous l'attendions, le nombre de mauvaises analyses diminue quand k augmente. On observe aussi une diminution du nombre des bonnes réponses, mais cette diminution est peu significative. On peut résumer cela comme suit.

- généralement, plus k se rapproche de 0, plus les transformations linguistiques autorisées sont importantes ;
- par conséquent, plus k est proche de 0, plus il y a de chances d'avoir de mauvaises analyses, puisque les transformations autorisées sont plus lâches.

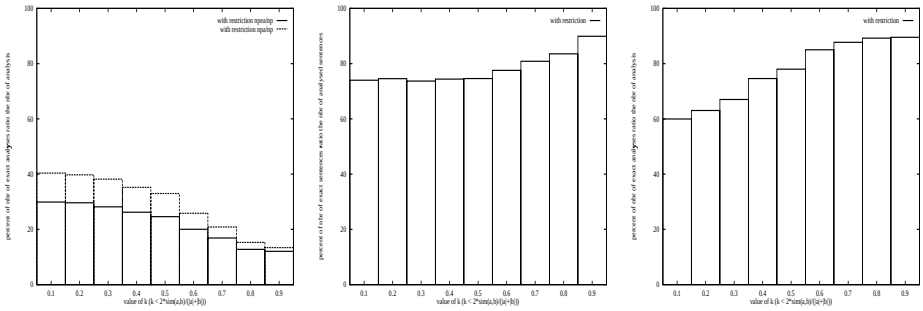


Figure 12.10: Justesse de la méthode. En abscisses, k=0,1, 0,3, 0,5, 0,7, 0,9 ; en ordonnées, nombre d'analyses produites.

Précision pour k=0,1, 0,3, 0,5, 0,7 et 0,9. Il est aussi intéressant d'étudier la précision de la méthode. Nous avons calculé différentes mesures sur les résultats précédents.

1. rapport du nombre absolu de phrases analysées et de phrases exactement analysées (c'est-à-dire nombre de phrases pour lesquelles au mons une réponse exacte est produite) : c'est un indicateur du rappel de la méthode, notion opposée à celle de bruit.

2. rapport du nombre de phrases exactement analysées au nombre de phrases analysées : c'est un indicateur de la précision de la méthode, notion opposée à celle de silence.

3. rapport du nombre de réponses exactes au nombre total de phrases : c'est un indicateur de la justesse de la méthode.

La figure 12.10 donne le rapport de phrases analysées aux phrases exactement analysées. On a l'impression que le nombre de phrases analysées converge avec k sur le nombre de phrases exactement analysées.

On peut donc dire que :

- lorsque k augmente,
  - les analyses deviennent meilleures ;
  - la justesse absolue et la précision par phrase augmentent ;
  - cependant, le rappel diminue.

12.2.5 Expériences d'analyse jusqu'à une certaine couche

L'extension de la méthode d'analyse directe par analogie à l'analyse récursive se fait en tirant bénéfice du modèle sous-jacent de langage de chaînes analogiques. Intuitivement on peut penser que l'application récursive de la méthode augmentera le nombre de phrases analysées. Hélas, les résultats ne montrent pas d'augmentation importante, et elle stagne après un seuil atteint rapidement.

Nous avons aussi calculé la qualité des analyses produites en fonction d'un seuil r. Nous nous attendions à une augmentation de cette qualité lorsque r augmente jusqu'à atteindre la qualité de la méthode directe. Nous verrons que, cependant, là encore, notre attente sera déçue.

Les expériences ont été menées avec une application simultanée de la contrainte de proximité des modèles. Afin de pouvoir comparer les résultats, toutes les mesures ont été effectuées avec deux mêmes valeurs de k : 0,8 et 0,9. Comme dans les graphes précédents, les graphes de la figure 12.11 portent en abscisse la distance à la réponse exacte divisée par sa taille et en ordonnée le nombre d'analyses produites.

Le résumé de l'analyse de ces résultats est le suivant.

- on n'observe aucun changement notable lorsque r augmente ;
- pour k=0,9, on obtient les mêmes résultats quelle que soit la valeur r=2, 3, 4 ;

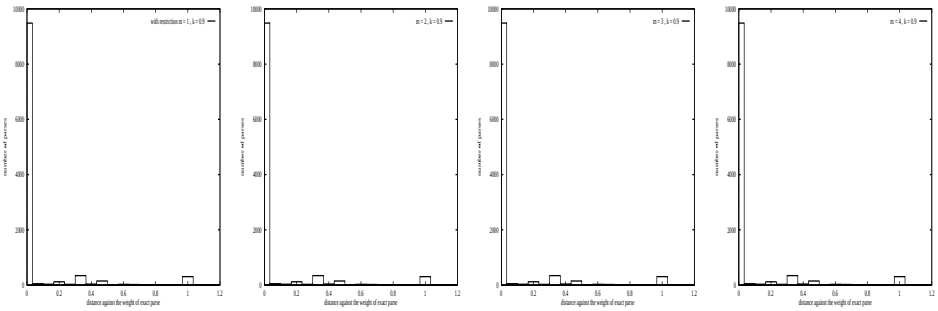


Figure 12.11: Récursivité. En abscisses, distance à la réponse exacte divisée par sa taille ; en ordonnées, nombre d'analyses produites.

- pour k=0,8, on obtient les mêmes résultats quelle que soit la valeur r=3, 4 ;

Par conséquent, il semble que le seuil de récursion optimal soit obtenu presque immédiatement. L'impression générale est donc assez décevante. Il reste cependant à la confirmer par des expériences supplémentaires avec d'autres valeurs de k.

12.2.6 Synthèse des résultats obtenus

L'analyse directe par analogie a permis d'obtenir une couverture de la moitié des phrases dans le cas d'un corpus arboré anglais avec des structures en constituants et un ensemble de catégories morpho-syntaxiques relativement important, et de deux tiers dans le cas d'un corpus arboré japonais avec des structures en dépendances sur moins de catégories morpho-syntaxiques. Un résultat particulièrement encourageant est que sa fiabilité est relativement élevée. Les réponses exactes représentent en effet 85% des meilleurs analyses produites dans le premier cas, et 95% dans le second.

Nous avons mesuré l'influence d'une restriction à un type de langages de chaînes analogiques particuliers, les langages de chaînes analogiques décroissants. Cette restriction n'a pas d'influence évidente sur les résultats, quoiqu'elle augmente légèrement la précision de la méthode. Afin de réduire l'ensemble des modèles, et donc de réduire l'espace de recherche, nous avons aussi mesuré l'influence d'une contrainte exercée sur la proximité des éléments dans les modèles utilisés. Cette contrainte ne semble pas accroître le nombre de réponses exactes mais elle contribue à diminuer le nombre de mauvaises analyses, en en éliminant. Le point négatif de ces expériences sur l'analyse structurale reste l'influence de la récursivité. Bien qu'une augmentation du nombre de réponses exactes ait été attendue quand le seuil augmente, aucune influence significative n'a pu être observée dès après une valeur relativement faible de ce seuil.

Nous concluons donc que, pour l'analyse structurale par analogie, la restriction aux langages de chaînes analogiques décroissants et la contrainte sur la

proximité des éléments des modéles semblent pouvoir s'utiliser sans dommage sur les résultats. Et il semble suffisant de se cantonner à l'analyse directe.

L'un des aspects positifs de la méthode générale proposée est qu'elle peut être appliquée immédiatement dès que l'on dispose d'un corpus arboré. Aucun prétraitement lourd, n'est requis à part une analyse morpho-syntaxique. Mais on peut considérer qu'une telle étape est de coût ridicule en l'état actuel des outils du traitement automatique des langues[16]. Cette immédiateté illustre les espoirs que nous avions placés dans l'analogie, et dans sa possible contribution aux approches à moindre effort en traitement automatique des langues (p. 94 et 252).

[16]Dans le cas d'un corpus arboré dont les représentations structurales sont en constituants, l'analyse morphosyntaxique peut se réduire à prendre la frontière des arbres, c'est-à-dire la séquence des feuilles des arbres, comme séquence de catégories morpho-syntaxiques.

12.3 Traduction automatique par analogie

Attaquons-nous maintenant à la traduction automatique. Le travail que nous allons rapporter n'est encore qu'embryonnaire, mais il constitue en fait le but ultime de nos réalisations. La montée vers ce but implique de gravir plusieurs marches. Nous avons déjà monté sur celle de l'analyse structurale. Nous venons de le voir. Il nous faut mettre pied sur celle qui représente l'aspect proprement dit de la traduction.

Or, avec tout ce que nous avons vu jusqu'à présent, l'intuition de base est facile. Pour introduire l'analyse par analogie à partir de la conjugaison, nous avions simplement observé qu'un changement des domaines intervenant dans les homomorphismes était suffisant. Il en est de même pour la traduction. On obtient immédiatement un système de traduction automatique en utilisant un homomorphisme entre deux domaines de phrases. Le premier domaine est l'ensemble des phrases en langue source, le second celui de l'ensemble des phrases en langue cible. En les mettant en correspondance, la méthode vue pour la conjugaison ou l'analyse s'applique à nouveau[17].

12.3.1 Maquette de traduction directe

Montrons un exemple simpliste d'application de cette idée[18]. Il s'agit d'une maquette de traduction automatique entre le français et le japonais. Nos données de base seront un sous-ensemble des phrases de Satou déjà vues plus haut (p. 277), légèrement modifiées et avec leur traduction française. Ces phrases sont donc en correspondance puisqu'alignées (voir figure 12.12, p. 315). Avec ces données, nous avons ébauché une répartition en grammaire et dictionnaire. La première partie des données est en effet constituée de phrases commutant selon les dimensions du genre, entre masculin et féminin, ou l'affirmation, avec des phrases affirmatives ou négatives, etc. La deuxième partie des données est constituée d'un petit lexique. La correspondance entre données du français et du japonais n'est pas bijective, comme nous l'avions mentionné pour l'analyse structurale. On notera donc que le mot *professeur* reçoit deux équivalents en japonais : 先生[19] ou 教授[20].

La maquette utilise le même moteur que celui utilisé pour la conjugaison des verbes français et pour l'analyse structurale. Son fonctionnement est donc identique. Il prend une phrase en entrée, par exemple *tu es infirmière*. Il essaie alors de trouver trois phrases françaises qui entretiennent avec elle une relation

[17]Cette technique fait l'objet d'un brevet au Japon Lepage, 用例機会翻訳装置, 1999.

[18]Cette partie reprend le contenu d'un exposé fait à Kyōto devant les membres de Kōin (光陰), groupe de chercheurs francophones du Kansaï: Lepage, *De l'analogie*, 2000, p. 10 à 11. Voir aussi Lepage & 白井諭 (Sirai Satosi), 言語学的比例類推空間相似, 2001, p. 91 et 92.

[19]先生 /sensei/ (maître, professeur). S'emploie pour n'importe quelle relation entre enseignant et élève, pas seulement dans l'enseignement scolaire, mais aussi dans les arts, le sport, etc.

[20]教授 /kyouzyu/ (professeur titulaire d'université). Titre universitaire au-dessus de celui de maître de conférence 助教授 /zyokyouzyu/.

tu es une fille.	あなたは少女だ。
tu es un garçon.	あなたは少年だ。
il est un garçon.	彼は少年だ。
elle est une fille.	彼女は少女だ。
elle est professeur.	彼女は先生だ。
il est professeur.	彼は先生だ。
elle est infirmière.	彼女は看護婦だ。
tu n'es pas infirmière.	あなたは看護婦ではない。
il n'est pas mon professeur.	彼は私の先生ではない。
professeur	先生
professeur	教授
marchand de vin	お酒屋さん
fille	少女
garçon	少年

Figure 12.12: Maquette de traduction français–japonais. Les données du système

tu es infirmière.	あなたは看護婦だ。	(× 21)
tu es une infirmière.	あなたは看護婦だ。	(× 1)
tu n'es pas mon infirmière.	あなたは私の看護婦ではない。	(× 3)
	あなたは看私の護婦ではない。	(× 2)
elle n'est pas mon infirmière.	彼女は私の看護婦ではない。	(× 22)
	彼女は看私の護婦ではない。	(× 18)
	彼女はの私看護婦ではない。	(× 1)
tu n'es pas professeur.	あなたは先生ではない。	(× 20)
	あなたは教授ではない。	(× 1)
tu n'es pas mon professeur.	あなたは私の先生ではない。	(× 21)
	あなたは私の教授ではない。	(× 1)
tu n'es pas marchand de vin.	あなたはお酒屋さんではない。	(× 2)

Figure 12.13: Maquette de traduction français-japonais. Exemples de traductions produites

d'analogie. Ici, il trouve: *elle est une fille*, *tu es une fille*, *elle est infirmière*. En effet, on a:

$$\textit{elle est une fille} : \textit{tu es une fille} \doteq \textit{elle est infirmière} : \textit{tu es infirmière}$$

En appliquant le principe de correspondances des analogies du français au japonais, on aura du côté japonais:

$$\text{彼女は少女だ} : \text{あなたは少女だ} \doteq \text{彼女は看護婦だ} : x \quad \Rightarrow \quad x = \text{あなたは看護婦だ}$$

On peut donc proposer la traduction あなたは看護婦だ comme traduction de la phrase de départ *tu es une infirmière*. Effectivement, ces deux phrases sont bien traduction l'une de l'autre.

De même que pour l'analyse, la recherche de triplets peut être effectuée en décomposant. On recherche d'abord des couples dans l'ensemble des modèles, on résoud ensuite l'équation analogique, et l'on vérifie enfin l'appartenance des solutions à l'ensemble de départ. On peut alors, comme pour l'analyse, appliquer récursivement la méthode. On retrouve alors le modèle d'homomorphisme entre langages de chaînes analogiques. La figure 12.13 (p. 315) donne un certain nombre de traductions produites par notre maquette avec ce modèle. Les chiffres figurant entre parenthèses indiquent la fréquence avec laquelle une traduction a été synthétisée. De même que dans les applications précédentes, il est naturel de penser que, pour une même entrée, les traductions les plus fréquentes sont aussi les meilleures.

Faisons deux remarques à propos des fréquences. Premièrement, parmi les traductions de la phrase *elle n'est pas mon infirmière*, des phrases ne faisant pas sens apparaissent. Dans un cas, le mot 看護婦 (infirmière) a été découpé de façon malheureuse et ces morceaux dispersés à droite et à gauche dans la phrase. Notre explication est que cette erreur, semblable à celles que nous avions observées dans l'expérience de production aveugle de phrases (p. 285), provient d'erreurs de segmentation de l'algorithme utilisé. Cependant, il est rassurant de constater que la traduction produite le plus fréquemment pour cette phrase, 22 fois en tout, est la bonne: 彼女は私の看護婦ではない。, les autres traductions défectueuses étant produites moins fréquemmment. Deuxièmement, toujours à propos de fréquence, on observe un comportement heureux dans la traduction de la phrase *tu n'es pas professeur*. Ici deux traductions, exactes, sont produites. Elles ne diffèrent que par l'usage du mot traduisant *professeur*. Celle qui utilise le terme 先生 est la plus fréquente, puisqu'elle est produite 20 fois. L'autre utilisant le terme plus spécifique de 教授 n'est produite qu'une fois. Or, dans notre corpus, le mot 先生 apparaissait quatre fois, l'autre mot une seule fois. La première phrase, utilisant des commutations avec le premier mot, a donc naturellement une fréquence plus élévée. La fréquence de production reflèterait donc une notion de fréquence d'usage. Un indice supplémentaire en est apporté par le fait que la phrase semblable *tu n'es pas mon professeur* est traduite de façon semblable par deux phrases montrant des répartitions en fréquence similaires.

12.3.2 Prototype de traduction directe

Bien que les possibilités d'une telle maquette ne soient quand même pas fameuses comparées à celles de n'importe quel système de traduction automatique un peu sérieux, du point de vue méthodologique, plusieurs points extrêmement positifs peuvent déjà être mis en avant. Nous les détaillerons plus bas. Nous préférons faire d'abord face à une critique naturelle à l'encontre de notre maquette. On peut en effet nous reprocher la taille ridicule des données précédemment utilisées.

Afin d'évaluer la couverture de la méthode, nous avons donc réappliqué la méthode à la traduction automatique entre le japonais et l'anglais sur des données brutes, mais bien plus importantes[21]. Il s'agit d'un ensemble de 150 000 phrases et 3 000 mots alignés dans ces deux langues. De nouveau, chaque langue constitue un domaine. Insistons sur le fait qu'il n'y a dans cette expérience aucune autre connaissance que la donnée brute des chaînes de symboles: ni catégories syntaxiques, ni codes sémantiques, ni résultat d'analyse morphologique, ni découpe en mots, rien. Pour 500 phrases nouvelles, avec la méthode directe, sans récursivité, des traductions sont produites dans 48% des cas. La figure 12.14 (p. 317) montre des traductions produites pour deux phrases, retranscrites sans modification aucune. On y observera des déplacements indus de caractères, le sceau de cette méthode. Les deux phrases japonaises signifient « *Une bière, s'il vous plaît !* » et « *Pas d'alcool.* ».

ビールを一杯お願いします。	*A glass of beer, please, please.* *A glass of been dressirg, please.* *I'd like a glass of beer, please.* *A glases of beer, please.*
アルコールはけっこうです。	*That's all ralcohoghl.* *No liquor, thanks.* *No space for liquor, thank you.* *No kind of liquor dessert, thanks.* *I'll pass kind of liquor on the dessert.*

Figure 12.14: Prototype de traduction japonais-anglais. Exemples de traductions produites

[21] Ces résultats préliminaires ont été présentés dans LEPAGE, *Formalisation de l'analogie entre chaînes de symboles*, 2001, p. 128.

12.3.3 Avantages méthodologiques

Les résultats précédents montrent bien que l'on ne peut prétendre résoudre le problème de la traduction automatique à l'aide de la seule analogie, mais la mise en œuvre de la maquette et du prototype permettent déjà de dégager certains avantages que nous allons décrire maintenant.

Moindre effort

Le premier avantage rejoint le souci que nous avions mentionné dans notre introduction sur les approches à moindre effort en traitement automatique des langues. Dans les deux applications précédentes, il faut noter que les données utilisées n'ont subi aucun traitement préliminaire. Les phrases et les mots sont seulement appariés d'une langue sur l'autre. Or, pratiquement, on dispose déjà de méthodes d'appariement automatique[22]. Ce prétraitement des données pourrait donc éventuellement se faire automatiquement.

À l'extrême, on pourrait prétendre que le type de système que nous proposons ne demanderait l'écriture d'aucune grammaire que ce soit, chose absolument nécessaire à la construction d'un système normal de traduction automatique. Une grammaire découpe explicitement à l'avance les données de langues et les classifient. Ici, c'est l'analogie qui se charge, implicitement, de faire ce découpage. Et ce sont les commutations qui font, implicitement, les classements (voir p. 47 à propos de la définition des catégories par Varron et p. 70 pour la syntaxe distributionnelle) . Et nous pensons que la faible couverture obtenue dans nos expériences provient d'un manque d'éléments susceptibles d'intervenir dans des analogies. Or, la méthode offre une manière simple d'augmenter les données afin d'augmenter la couverture. En effet, par l'ajout aux données de base des phrases non traduites, et par leur traduction dans la langue cible, on augmente à peu de frais le système. Cependant, il faut veiller à ce que la qualité des autres résultats ne pâtissent pas d'ajouts inconsidérés. Se pose donc le problème de la pertinence linguistique de ces ajouts. Toute une recherche reste à effectuer dans cette direction. Pour la faciliter, nous proposons d'introduire des traces d'exécution. Elles devraient permettre de comprendre les différences d'exécution avant et après l'ajout ou le retrait de données aux différents domaines. Il s'agira de visualiser les chemins empruntés lors de la constitution d'analogies et lors de la résolution d'équations analogiques, dans chacun des domaines. Comme ces chemins peuvent être très nombreux et mettent en jeu systématiquement à chaque étape au moins trois éléments à chaque fois, leur représentation est a priori difficile. Nous devrons sans doute recourir à des procédés assez élaborés de visualisation.

Distinctions ou coupures profitables

Se contenter d'ajouter au système des phrases non traduites avec leur traduction est évidemment peu satisfaisant théoriquement. Mais pratiquement aussi,

[22]Voir, par exemple, Debili & Sammouda, *Appariement de phrases de textes bilingues français-anglais et français-arabes*, 1992.

cela peut s'avérer peu rentable à moyen terme parce que peu économique au sens du nombre des modèles utiles (voir p. 305 et 308). On peut donc s'interroger sur le moyen de ne rajouter que les données vraiment utiles au système. Et même, dès le départ de ne construire que ces données utiles. Or, de ce point de vue, notre méthode offre encore quelques avantages que nous allons maintenant décrire. Chacun correspond en fait à une distinction ou coupure méthodologique qui est en quelque sorte sublimée paradoxalement dans l'avantage pratique qu'elle apporte. De plus chacune de ces coupures révèle un aspect théorique positif de la méthode du point de vue linguistique.

Pour exposer ces avantages, constatons que trois tâches différentes apparaissent dans la construction d'un système selon notre modèle :

- description monolingue de la langue source ;
- description monolingue de la langue cible ;
- appariement, c'est-à-dire établissement des correspondances par traduction.

Les deux premières tâches sont strictement monolingues. Seule la troisième est bilingue. L'opposition entre les deux premières tâches trace une coupure entre les deux langues mises en jeu dans la traduction. L'opposition entre les tâches monolingues d'une part et la troisième, bilingue, d'autre part, dessine une seconde coupure.

Coupures entre langues La sublimation de cette première distinction méthodologique est le fait que notre méthode construit des systèmes bidirectionnels. Et il s'agit là d'une bidirectionalité pure. En effet, plus que de bidirectionalité, c'est carrément d'indifférence aux données qu'il s'agit, puisque comme nous l'avons vu, le moteur est universel[23]. L'échange du premier et du second domaines est en effet immédiatement réalisable. Dans notre réalisation informatique, il suffit d'échanger les arguments d'appel au moteur général. Notre système se démarque donc des systèmes dont le formalisme est certes bidirectionnel, mais qui réclament deux compilations différentes pour produire, séparément, un moteur pour l'analyse et un moteur pour la génération.

Du point de vue de la mise au point, une retombée avantageuse immédiate de la bidirectionalité concerne la phase de test de chaque description monolingue d'un système. En effet, un test monolingue de la morphologie flexionnelle ou dérivationelle et de la grammaire pour une langue donnée est envisageable immédiatement. Il suffit de dupliquer le domaine représentant la langue à tester. On obtient alors un premier et un second domaine identiques. Apparier chaque élément avec lui-même dans ces deux domaines est trivial. En fait, dans notre réalisation informatique, il suffit de passer le même argument deux fois. Dès lors, faire jouer la méthode de traduction proposée revient en quelque sorte à faire de la traduction automatique d'une langue en cette même langue. Le moteur universel proposé permet bien ce test, immédiatement et

[23]La tentation est grande d'y voir un solveur général, à la mode d'un moteur Prolog.

sans aucune modification. Dans un système de traduction automatique de seconde génération, c'est-à-dire avec un transfert, il serait nécessaire, pour réaliser un test du même type, d'écrire un transfert « bidon ». Et cette écriture n'est pas toujours aisée. Ici, aucun travail n'est à faire.

Coupures en micro-descriptions À l'intérieur d'un même domaine, la présentation des données nécessaires à la réalisation d'un système de traduction automatique selon notre méthode n'impose aucune distinction de quelque ordre que ce soit. On a donc formellement la disparition de l'opposition classique entre dictionnaire et grammaire, opposition d'habitude présente dans la réalisation des systèmes de traduction automatique. Cette opposition n'est en rien essentielle à notre méthode. La disparition de cette opposition nous semble en fait souhaitable du point de vue épistémologique. En effet, personne ne sait vraiment définir ce qu'est une phrase[24], ni non plus ce qu'est un mot[25]. Le problème du degré de figement des expressions montre bien qu'il n'existe pas de frontière nette entre ce qui relève du vocabulaire et ce qui ressortit à la grammaire. Cela rejoint les éternels débats sur où finit la morphologie et où commence la syntaxe[26]. En définitive, il faut bien reconnaître que l'on sait seulement couper des morceaux de langue. Or c'est précisément cette possibilité qu'offre l'organisation que nous proposons. Le concepteur d'un système pourra donc librement ranger *jeune fille* et *pomme de terre* en opposition avec leur pluriel *jeunes filles* et *pommes de terre* dans la même partie de ces données que celle qui lui tient lieu de dictionnaire, s'il l'entend ainsi. Dans la maquette et le prototype que nous avons montrés plus haut, et comme nous l'avions noté au passage (p. 314), nous avions procédé en regroupant les données implicitement en deux parties : une partie de dictionnaire et une partie de phrases, cette seconde partie pouvant être considérée comme une partie de syntaxe. Il y a donc, librement, possibilité d'introduction de n'importe quelle coupure par blocs de description. Cette possibilité rejoint la proposition de Sergei Nirenburg de construire des systèmes de traduction automatique par assemblage de micro-théories décrivant chacune un problème particulier de syntaxe ou de morphologie. L'absence de coupure imposée dans la description des données est sublimée par la possibilité de décrire celles-ci de façon modulaire.

Coupures en descriptions monolingues et bilingues Pour continuer ce que nous mentionnions à propos du transfert, notre proposition supprime la conception d'une étape de transfert au sens habituel du terme. En ce sens, la

[24]Voir MOUNIN, *Dictionnaire de la linguistique*, 1974, p. 262. L'entrée *phrase*, rédigée par Joseph DONATO, donne cinq définitions possibles différentes de ce concept pourtant considéré comme intuitif, définitions qui ne sont pas équivalentes.

[25]L'article MOT du même dictionnaire se termine crânement par : *Le mot n'est pas une réalité de linguistique générale.* Mais l'on peut tout de même dire que la notion à bien une réalité pratique !

[26]Certains linguistes proposent quand même des critères assez formels. Voir, par exemple, MEL'ČUK, *Dependency syntax: Theory and practice*, 1988, p. 105 à 144, sur les types de dépendance syntagmatique.

méthode s'oppose aux systèmes de deuxième génération de traduction automatique caractérisés par les trois étapes successives d'analyse, de transfert et de génération. Si transfert il y a, il est décrit de façon purement statique par la mise en correspondance des données des deux langues. La conséquence en est qu'il ne saurait y avoir influence directe de la syntaxe de la langue source sur la syntaxe des traductions produites. Cela crée une différence d'avec les systèmes de première génération, cette fois. En effet dans ces systèmes, la phrase en langue source est en quelque sorte remplacée petit à petit par la traduction en langue cible. Le processus de traduction automatique de première génération peut être vu comme une réécriture sur une seule et même bande d'entrée et de sortie. Or, dans notre méthode, à la base, il y a deux domaines séparés. Les données du domaine d'entrée ne pénètrent jamais dans le domaine cible. Par conséquent, la syntaxe de la langue source ne saurait en aucun cas influencer la syntaxe de la langue cible. Par exemple, grosso modo, en japonais, le verbe est toujours en fin de phrase. On ne trouvera jamais de verbe en milieu de phrase japonaise comme résultat de traduction parce que, pour une raison ou pour une autre, le transfert du français au japonais aurait oublié de déplacer un groupe verbal. Si dans un résultat de traduction, un tel cas apparaissait jamais, la faute ne saurait être attribuée à un quelconque reste de syntaxe française oublié lors du traitement, mais bien à une commutation malheureuse due à l'analogie, en langue cible, le japonais, et en langue cible seulement.

À notre avis, la coupure en descriptions monolingues et bilingue permet au concepteur d'un système de traduction de se concentrer séparément sur les problèmes relevant de la traduction proprement dite. De nouveau, cela permet une approche modulaire. Ainsi, par exemple le travail sur l'expression de la triade temps-mode-aspect pourrait se faire comme suit. Dans une partie de description monolingue, on se concentrera, à l'intérieur d'un même système de langue, sur les oppositions entre formes verbales en contexte en donnant les commutations pertinentes. Les équivalences de traduction appartiennent à la partie bilingue. Décrire la traduction exacte d'une forme verbale relève de la confrontation de deux systèmes et n'implique pas forcément les mêmes exemples que ceux de la partie monolingue. Dans la partie bilingue de description, c'est précisément ceux-là qu'il faudrait donner, et ceux-là seulement. De cette façon on se concentrera sur les différences entre les systèmes, les régularités étant décrites par défaut par la traduction réciproque des exemples monolingues. On retrouve là l'opposition entre l'analogie et l'anomalie et le traitement des exceptions tel que nous l'avions déjà mentionné dans notre exemple de conjugaison des verbes français (p. 252).

Coupures en chemins indépendants Le dernier trait positif de notre méthode vient de la façon dont le calcul s'effectue. Cela l'oppose de nouveau aux systèmes de traduction automatique de deuxième génération. Dans ces systèmes les étapes d'analyse, de transfert et de génération se succèdent dans le temps. Ici, cela est encore vrai pour un seul chemin menant à une traduction, mais cela n'est plus vrai globalement. En effet, rien n'empêche de présenter les traductions dès qu'elles sont produites. On peut comprendre que

les traductions provenant de l'analyse de la phrase en langue source comme élément de Λ_1 et de la génération de la traduction comme élément de $\widehat{\Lambda_1}$ soient données d'abord. De façon globale, donc, il y a exécution simultanée des étapes d'analyse, de transfert[27] et de génération. Si l'on considère la résolution d'une équation analogique comme prenant un temps constant, on peut alors scander le rythme de production des traductions. Cette scansion est donnée dans le tableau 12.6. Dans chaque ligne, la somme $i+j$ des indices i et j portés par les sous-ensembles Λ_i et $\widehat{\Lambda_j}$ des langages de chaînes analogiques qui représentent les langues source et cible respectivement, est égale au temps donné par l'étape de scansion.

Tableau 12.6: Scansion de la traduction par analogie. Cas où les ensembles de phrases attestées $\mathcal{A}$ et $\widehat{\mathcal{A}}$ sont en correspondance totale

Temps de la scansion	Analyse : la phrase d'entrée appartient à	Génération : la traduction produite appartient à
0	$\Lambda_0 = \mathcal{A}$	$\widehat{\Lambda_0} = \widehat{\mathcal{A}}$
2	Λ_1	$\widehat{\Lambda_1}$
3	Λ_1 Λ_2	$\widehat{\Lambda_2}$ $\widehat{\Lambda_1}$
⋮	⋮	⋮

On remarquera que le temps de scansion 1 n'existe pas dans ce tableau. Cela provient du fait que, conformément à ce que nous avons implicitement admis jusqu'à présent, toute phrase de $\mathcal{A}$ est supposée avoir au moins une traduction dans $\widehat{\mathcal{A}}$ et que toute phrase de $\widehat{\mathcal{A}}$ est supposée être traduction d'au moins une phrase de $\mathcal{A}$. Or, pratiquement, cela n'est pas nécessaire. Le tableau de la scansion gagne en généralité si on admet que certaines phrases attestées de la langue source n'ont pas de traduction dans $\widehat{\mathcal{A}}$ et que certaines phrases de $\widehat{\mathcal{A}}$ ne sont pas des traductions de $\mathcal{A}$. Chaque temps de scansion y devient possible (voir le tableau 12.7, p. 323).

Le temps nécessaire à la production d'une traduction reflète sa difficulté. Plus court est le temps de production, moins l'analyse ou la génération requiert de calcul car plus proches des données de base sont la phrase à traduire et la traduction produite. Ainsi, à l'extrême, les traductions produites à l'étape 0 sont simplement les traductions existant déjà dans les données de base du système. Aucun calcul n'est nécessaire pour les obtenir.

À un temps donné de la scansion, des traductions déjà proposées à des temps antérieurs peuvent être à nouveau produites. Elles ont, comme nous l'avons déjà vu plus haut (p. 316), une fréquence donnée particulière à l'étape de scansion à laquelle elles viennent d'être produites. Cette fréquence s'ajoute

[27] On peut en fait négliger le transfert puisqu'il n'y en a pas réellement dans notre méthode.

Tableau 12.7: Scansion de la traduction par analogie. Cas où les ensembles de phrases attestées $\mathcal{A}$ et $\widehat{\mathcal{A}}$ ne correspondent pas totalement

Temps de la scansion	Analyse : la phrase d'entrée appartient à	Génération : la traduction produite appartient à
0	$\Lambda_0 = \mathcal{A}$	$\widehat{\Lambda_0} = \widehat{\mathcal{A}}$
1	$\Lambda_0 = \mathcal{A}$ Λ_1	$\widehat{\Lambda_1}$ $\widehat{\Lambda_0} = \widehat{\mathcal{A}}$
2	$\Lambda_0 = \mathcal{A}$ Λ_1 Λ_2	$\widehat{\Lambda_2}$ $\widehat{\Lambda_1}$ $\widehat{\Lambda_0} = \widehat{\mathcal{A}}$
3	$\Lambda_0 = \mathcal{A}$ Λ_1 Λ_2 Λ_3	$\widehat{\Lambda_3}$ $\widehat{\Lambda_2}$ $\widehat{\Lambda_1}$ $\widehat{\Lambda_0} = \widehat{\mathcal{A}}$
$\vdots$	$\vdots$	$\vdots$

à la fréquence déjà obtenue précédemment. De cette façon, la liste des traductions peut être réactualisée au fil du temps selon l'ordre de fréquences. On espère toujours, en effet, que les traductions les plus fréquentes seront les meilleures.

12.3.4 Projet de traduction automatique du groupe ℵ

Buts de ce projet Après avoir paré notre proposition de tous les avantages possibles tant du point de vue conceptuel, pratique que théorique, soulignons que nous gardons tout de même une vue réaliste de nos recherches. Nous l'avions mentionné en introduction : ce n'est pas parce que nous savons résoudre une équation analogique que nous avons résolu les problèmes du traitement automatique des langues, encore moins de la linguistique. Les résultats obtenus avec notre maquette ou avec notre prototype de traduction automatique le montrent, nous sommes tout de même loin du compte. Cependant, nous désirons savoir jusqu'où l'on peut aller à la seule force de l'analogie, *analogiae vi tantum*. Nous pensons qu'il s'agit là d'une expérience nécessaire pour mettre en lumière par la pratique les phénomènes hors de portée de l'analogie. Par exemple, déjà, nous savons, pour l'avoir éliminé en conséquence de notre formalisation, que nous ne saurions traiter les redoublements.

Pluralité de domaines Mais le prototype de système de traduction automatique que nous avons mentionné plus haut ne saurait être satisfaisant. Il montre des défauts auxquels il faut nécessairement remédier. Les remplacements erronés de caractères proviennent évidemment d'un manque de con-

trainte au niveau de la segmentation des mots. Des déplacements indus de groupes entiers proviennent évidemment d'un manque de contrainte à un niveau structural. Pour pallier ces problèmes nous proposons d'introduire plusieurs domaines qui sont appelés à se contraindre les uns les autres.

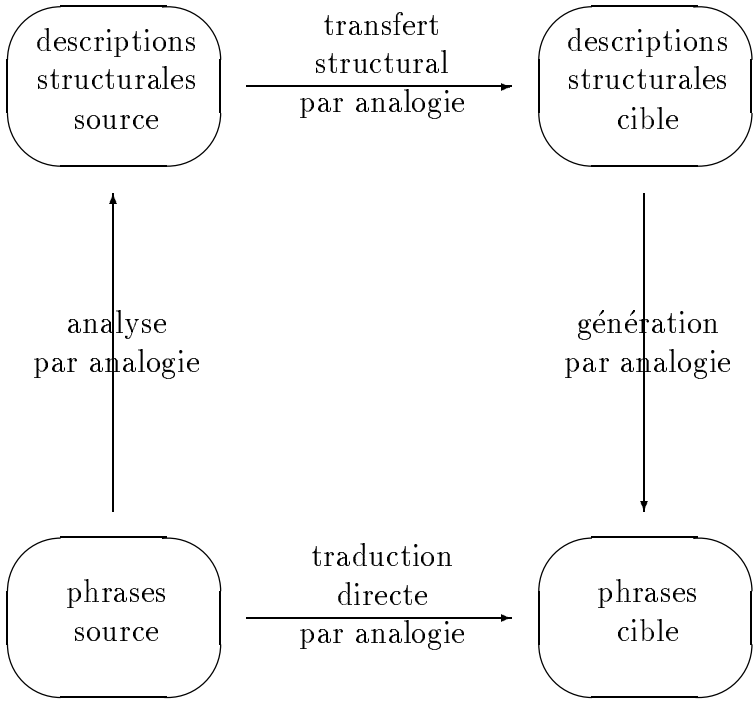


Figure 12.15: Assemblage des différents homomorphismes présentés plus haut dans un projet de traduction automatique par analogie

La vision d'une pluralité de domaines découle naturellement de tout ce que nous avons présenté jusqu'à présent. Nous venons de présenter une maquette de traduction automatique directe par analogie. Elle est matérialisée par la flêche horizontale du bas de la figure 12.15. Nous avons montré un peu plus haut comment réaliser l'analyse structurale par analogie. Cela correspond à la flêche ascendante à gauche dans la même figure. Nous avions mentionné que toutes nos applications étaient bidirectionnelles. En conséquence, rien ne nous empêche de faire de la génération en prenant comme premier domaine un ensemble de représentations structurales, et comme second domaine un ensemble de phrases de la langue cible. La flêche descendante à droite dans la même figure montre donc une étape de génération. En résumé nous venons d'obtenir dans la figure les trois côtés d'un carré. Il est alors loisible de fermer le carré. Cela correspond simplement à la réintroduction d'une véritable étape de transfert structural. En définitive, un système empruntant les flêches de gauche, puis du haut, puis de droite, suivrait simplement les trois étapes classiques de la traduction automatique de deuxième génération. Ici, chacune de

ces étapes peut être réalisée par un homomorphisme entre différents domaines. Mais, en plus, dans notre figure, apparaît une flêche en bas, correspondant à la traduction directe par analogie. Notre souhait n'est pas de faire s'exécuter l'un à la suite de l'autre chacun de ces processus, mais bien de les voir à l'œuvre dans une exécution simultanée. Et en plus, afin de contraindre éventuellement par une certaine découpe en mots, il serait souhaitable et naturel de rajouter dans un tel schéma deux domaines de catégories morpho-syntaxiques pour chaque langue. On obtiendrait alors la configuration de la figure 12.16.

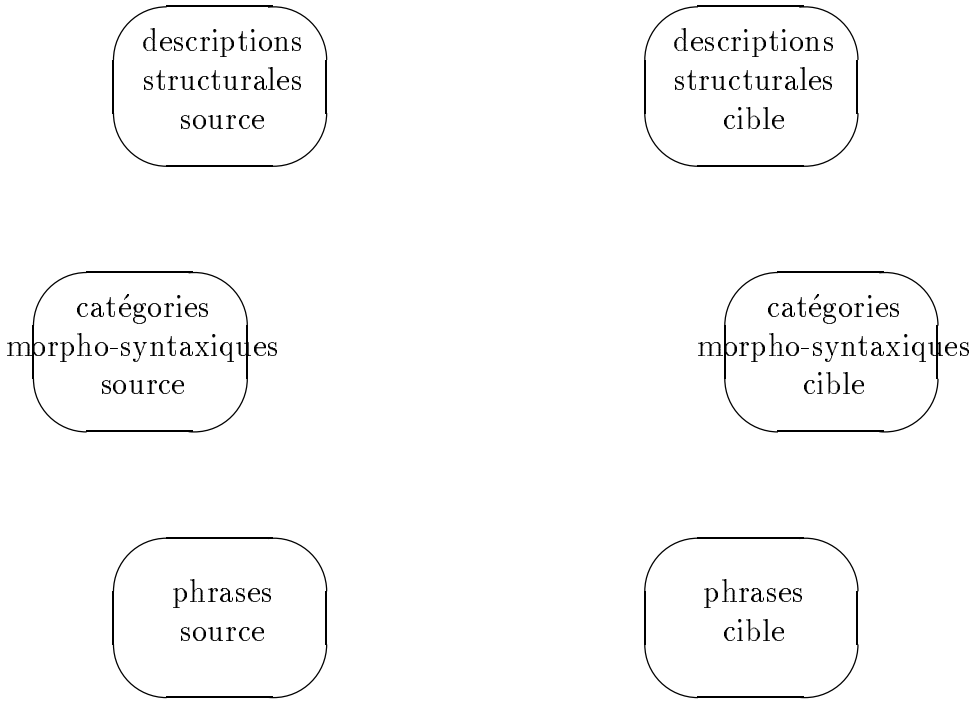


Figure 12.16: Domaines souhaitables dans un projet de traduction automatique par analogie

Tableau noir Les programmes que nous avons réalisés jusquà présent se limitent à un seul homomorphisme entre deux domaines seulement. Dans nos programmes, la résolution des équations analogiques dans le second domaine est intimement imbriquée dans la recherche des analogies dans le premier domaine (voir les algorithmes 8.6 et 8.7, p. 230 et 231). Notre projet nous force à une phase de reconception de nos programmes. Nous donnons ici très succintement le schéma de conception auquel nous sommes arrivé.

L'exécution dans chaque domaine doit devenir indépendante des autres. La correspondance entre domaines doit elle aussi devenir indépendante des programmes. Nous répondons à ces impératifs de la façon suivante. Premièrement, dans chaque domaine, nous nous contentons d'un algorithme de vérification

de l'appartenance au langage paresseux décroissant dont l'ensemble des éléments attestés est l'ensemble des données du domaine. Deuxièmement, les correspondances entre domaines sont données par l'intermédiaire de numéros d'éléments dans un grand tableau à la mode de celui montré dans la figure 12.8. Dès lors, tout résultat provenant d'un domaine est affiché, avec l'analogie qui a permis de le produire, dans le même tableau, qui sert de tableau noir à l'exécution général. Des numéros nouveaux pour des éléments nouvellement produits peuvent être attribués automatiquement dans chaque domaine par un simple processus de gestion de ces numéros. Dès lors, un nouveau numéro, avec l'équation analogique qui a permis de l'obtenir, mais sans contrepartie dans les autres domaines est interprété comme un ordre de création de données correspondantes dans chacun des autres domaines.

Tableau 12.8: Correspondances entre éléments des différents domaines

domaine 1	domaine 2	domaine 3	domaine 4	domaine 5
23	*45*	*24*	*124* *56*	*254* *24* : *3* ≐ *56* : *x* ⇒ *x* = *234*
	9 : *67* ≐ *5* : *x* ⇒ *x* = *142*	*56* : *45* ≐ *32* : *x* ⇒ *x* = *167*		

Le schéma détaillé est encore en cours de définition. Mais sa réalisation ne devrait pas poser de difficulté majeure, puisque nous disposons déjà de toutes les briques élémentaires. À moyen terme, cette architecture nous offrira une base générale de développement et d'expérimentation. Les données à notre disposition représentent 156 000 mots, expressions, phrases exemples et exemples d'usage en deux langues, japonais et anglais, avec leurs catégories morpho-syntaxiques et leur description structurale. Avec ces données nous espérons bien être à même d'effectuer des expériences de mesure de couverture et de qualité. Grâce à cela, nous espérons pouvoir épingler les phénomènes que l'analogie seule ne saurait traiter.

Conclusion : synthèse et spéculations

On peut comprendre que l'analogie entre chaînes de symboles ait été peu étudiée jusqu'à présent quand on sait que le point de vue génératiViste gouvernait l'étude des langages formels jusqu'à une date récente. Comme le courant génératiViste s'est précisément développé en réaction au courant structuraliste, et entendait l'éliminer sociologiquement, il rejeta l'analogie, avec l'induction, comme possible outil de test pour la grammaticalité, arguant de ce que, selon le bon sens, l'analogie en logique peut conduire au faux, chose connue depuis les Sophistes.

Mais le prétexte fallacieux qu'une application aveugle de l'analogie peut conduire au faux en logique, ou à l'agrammaticalité en syntaxe, est de même nature que prétendre qu'aucune grammaire au sens chomskyen du terme ne saurait être adéquate pour la description des langues parce que certaines grammaires formelles produiraient des structures jamais rencontrées dans aucune langue humaine. L'analogie ne constitue certainement pas l'horizon indépassable de la linguistique, mais on peut sans crainte affirmer qu'elle joue un rôle dans les langues.

Une étude historique de l'**analogie** nous a permis de dégager les deux articulations constitutives de l'analogie, la **conformité** et le **rapport**, ainsi que ses deux notions constitutives, la **similarité** et la **contiguïté**. Nous avons proposé une formalisation générale qui reprend ces articulations et ces notions. Cela nous a permis de donner certains résultats formels sur l'analogie de façon général, mais aussi dans une particularisation aux ensembles, aux multi-ensembles et aux chaînes de symboles. Pour ce qui est des chaînes de symboles, le versant de la similarité et ses implications semblent assez bien compris. En revanche, le versant de la contiguïté nous échappe encore.

Nous avons aussi introduit un procédé fondé sur l'analogie, qui permet de définir certains langages formels à partir de modèles, les **langages de chaînes analogiques**. Il est important de noter que les grammaires de chaînes analogiques ne font pas usage de non-terminaux. La grammaticalité, c'est-à-dire l'appartenance à un langage, est testée par comparaison avec certaines chaînes attestées de ce langage. Cette approche, par réduction à des formes attestées nous semble intuitivement proche de la façon dont les humains vérifient la grammaticalité des phrases nouvelles. Nous avons démontré que ce procédé possède une propriété remarquable, qui, à notre sens, constitue un argument de poids en faveur d'une modélisation de la langue par ce procédé. Il s'agit de la propriété de croissance constante des longueurs, une des propriétés proposées pour définir le modérément sous-contexte, qui, pour certains, caractériserait la puissance formelle apparente des langues humaines. Nous pressentons que la complexité d'analyse en temps des langages formels $\{a^n\}$, $\{a^n b^n\}$ et $\{a^n b^n c^n\}$, qui sont de classe différente dans la classification de Chomsky-Schützenberger, serait identique lorsqu'ils sont analysés comme langages de chaînes analogiques. Selon nous, si cela est vérifié, cela constitue un argument fort en faveur de l'analogie. Cela montrerait que l'analogie permettrait de présenter un autre point de vue que la hiérarchie de Chomsky-Schützenberger, ce que d'autres procédés, comme par exemple les grammaires contextuelles, faisaient aussi déjà.

Enfin, nous avons montré comment la notion d'**homomorphismes** permet de rendre compte d'un type restreint de métaphore. Notre proposition a été de transporter les analogies d'un domaine à un autre, de les conserver, et donc de sélectionner des structures dans un autre domaine. Pour ce qui est de nos applications à l'analyse structurale par analogie ou à la traduction automatique, le paysage peint par l'analogie est le suivant : entre des éléments de types distincts, nous établissons une correspondance de forme libre. Un ou plusieurs éléments de la source peuvent correspondre à un ou plusieurs éléments du but, certains éléments peuvent ne pas avoir de correspondants. Nous avons montré comment toutes ces notions théoriques pouvaient être utilisées pour réaliser l'analyse structurale ou la traduction automatique au stade de prototype, deux applications particulières du traitement automatique des langues.

Selon le dictionnaire universel de Pierre Larousse[28], n'en déplaise à Bachelard, l'analogie fonderait vraiment toute connaissance. Nous formulons le vœu que les vues que nous allons maintenant exposer réconcilient les tenants de l'analogie et ses adversaires. Nous sommes en effet d'accord avec ceux qui mettent en garde contre les dangers de l'analogie, en ce sens que l'analogie dont ils parlent est l'analogie qui chevauchent deux domaines, une analogie à la Nagao. Mais nous aimerions apporter notre soutien aux tenants de l'analogie comme fondement de la plupart de nos connaissances en mettant en avant notre modèle de la métaphore restreinte. Les analogies chevauchant deux domaines, obtenues dans ce cadre, et nous ne parlons que de celles-là, nous semblent, elles, beaucoup plus solides que n'importe quelle autre. Sous-tendues par deux analogies qui existent respectivement dans deux domaines différents, de telles analogies n'expriment pas que des correspondances de domaine à domaine, elles sont la conséquence de la conservation d'une structure d'un domaine à un autre. En ce sens, elles ne sont qu'une des faces d'une correspondance beaucoup plus massive entre les domaines. Et elles sont bien une face d'un parallélépidède difforme dans nos représentations. Elles émergent en fait de la confrontation entre deux structures données.

Nous allons maintenant insister sur le caractère donné de ces structures. Il découle du fait que la formalisation que nous avons proposée de l'analogie ne prend pas en compte d'autre niveau que celui des symboles et ne considère les symboles que sous l'angle de leur égalité. En ce sens, cette formalisation est donc aveugle à toute signification des symboles ou à tout autre relation entre symboles. En fait, plutôt que de caractère aveugle on pourrait parler de myopie. Illustrons ce point sur deux transcriptions différentes du japonais.

待ちます ：待つ ≒ 働きます ： x
machimasu : *matsu*≒*hatarakimasu*: x [29]

[28] LAROUSSE, *Grand dictionnaire universel*, 1865 a 1876, p. 312 et suivantes, articles ANALOGIE, ANALOGIQUE et ANALOGUE.

[29] 待ちます /matimasu/ (attendre) forme de politesse marquée. 待つ /matu/ (même sens), forme non marquée. 働きます /hatarakimasu/ (travailler) forme marquée. La seconde équation analogique utilise la transcription Hepburn qui est une transcription anglo-saxonne et bâtarde du japonais : les voyelles y ont la même valeur qu'en italien, mais les consonnes sont transcrites à l'anglaise. Pour quiconque connaît le japonais, ces deux équations analo-

Les deux équations analogiques précédentes ne peuvent être résolues dans le cadre de notre formalisation car celle-ci repose sur rien d'autre que l'égalité des symboles. En revanche, elle rend bien compte de l'analogie suivante.

$$matimasu : matu \doteq hatarakimasu : x \quad \Rightarrow \quad x = hataraku \ ^{30}$$

C'est seulement si l'on donnait une interprétation aux chaînes de symboles, c'est-à-dire si l'on ne se cantonnait plus au niveau des symboles, que l'on pourrait énoncer que notre formalisation dépendrait du codage. Mais il n'y a aucun codage au niveau auquel nous nous sommes placé.

La conséquence par contraposée de cette myopie ou de ce **caractère aveugle** est que la donnée seule d'un nombre de symboles suffit à définir entièrement la structure analogique sur l'ensemble des chaînes formées de ces symboles. En effet, il y a isomorphisme de structure analogique entre ensembles de chaînes de symboles construites sur des ensembles de symboles de même cardinal. Cet isomorphisme est basé sur l'isomorphisme trivial qui conserve la concaténation :

$$\mathrm{f}(A.A') = \mathrm{f}(A).\mathrm{f}(A')$$

Mais plus, la structure analogique d'un ensemble ${\mathcal{V}_n}^*$ de chaînes de symboles construites sur un ensemble $\mathcal{V}_n$ de n symboles est incluse dans la structure analogique de tout ensemble de chaînes de symboles construites sur un ensemble de $m > n$ symboles. En effet, il existe au moins C_m^n parties de ${\mathcal{V}_m}^*$ isomorphes à ${\mathcal{V}_n}^*$ par le type d'isomorphisme précédent[31]. Le **caractère universel** de l'analogie repose donc à la fois sur cette inclusion de structures et sur l'indifférence aux symboles par isomorphisme. En d'autres termes, la structure analogique est immanente, exactement de la manière que la structure de $\mathbb{N}$ est immanente. Et il n'y a ici place pour aucune liberté. Mais, de même que nous sommes incapables d'embrasser $\mathbb{N}$, de même nous sommes incapables d'embrasser, en un seul mouvement de pensée, la totalité de la structure analogique d'une ensemble de chaînes construites sur un ensemble fini de symboles. Il nous faut découvrir petit à petit cette structure, de la même façon que nous construisons $\mathbb{N}$ au coup par coup, au fur et à mesure de nos besoins, en additionnant un n à un m. Le **caractère créateur** de l'analogie dont nous parlions dans notre introduction est donc en réalité un faux-semblant. La création dont nous sommes capables là est juste du même ordre que celle de l'addition entre nombres entiers. En plus, de la même façon que tout entier peut apparaître comme une fonction de $\mathbb{N}$ dans $\mathbb{N}$, pour l'addition,

$$\begin{array}{rccc} m : & \mathbb{N} & \rightarrow & \mathbb{N} \\ & n & \mapsto & n + m \end{array}$$

de la même façon, tout couple de chaînes définit une application de l'ensemble des chaînes dans l'ensemble des parties de cet ensemble, pour l'analogie.

giques devraient produire les solutions 動く et *hataraku*.

[30] Cette transcription est donnée dans la recommandation officielle du gouvernement japonais, dite *kunrei*. Instruction, recommandation officielle, se dit en effet 訓令 /kunrei/.

[31] C'est simplement le nombre de possibilités de prendre n éléments parmi m, c'est-à-dire C_m^n.

$$
\begin{array}{rccl}
(A,B): & \mathcal{V}^* & \rightarrow & \wp(\mathcal{V}^*) \\
 & C & \mapsto & \alpha(A,B,C) \\
 & & & = (\ A:B\ )(C) \\
 & & & = \{D \in \mathcal{V}^* \ / \quad A:B \doteq C:D\ \ \}
\end{array}
$$

L'équation $A:B \doteq C:x \quad \Rightarrow \quad x=D$ est en effet seulement une autre notation pour $D=(A:B)(C)$, qui peut s'interpréter comme l'application de la fonction $(A:B)$ à C. Dans nos développements sur les langages de chaînes analogiques, nous avons noté de tels couples $A \rightarrow B$ et nous les avons appelés des modèles. De la même façon, encore, qu'aucune variable n'intervient dans notre connaissance immédiate de l'addition, de la même façon, aucune variable n'intervient dans notre connaissance immédiate de l'analogie. Logiquement donc, aucune variable n'apparaît dans les modèles et par conséquent, les langages de chaînes analogiques ne font pas usage de non-terminaux, contrairement aux usages classiques en théorie des langages. Un effet de cela est que les modèles généraux $A \rightarrow B$ se situent entre le remplacement d'un symbole par un autre, et sa conservation. Dans le cas général, par analogie, on transforme à moitié, on laisse inchangé à moitié. Un modèle du type $a \rightarrow b, a \neq b$, s'applique sur tout mot contenant le symbole a, et le modèle est simplement l'expression du remplacement d'une occurrence du symbole a par le symbole b. Par exemple, la résolution de l'analogie $a:b \doteq aa:x$ donne $x=ab$ ou $x=ba$. Mais par tout modèle du type $a \rightarrow a$, toute chaîne C sera laissée inchangée, car $a:a \doteq C:x \quad \Rightarrow \quad x=C.$ On a donc, dans le cas général, à l'aide d'une seule notation, la notation de deux phénomènes habituellement vus comme différents. Par exemple, pour l'équation analogique *je chante* : *tu chantes* $\doteq$ *je mange* : $x \quad \Rightarrow \quad x =$ *tu manges* l'échange de *je* en *tu* ressortit à l'échange, alors que la conservation du radical *mange*, ressortit à l'instanciation de *mange* à *chante*. En fait, quels symboles d'un modèle seront vraiment soumis à conservation n'est connu qu'en présence du troisième terme de l'équation analogique. Par conséquent, cette fusion de deux opérations comporte le danger inhérent à toute extension, celui d'en faire trop. Dans le cas particulier de l'analogie, c'est la source de ce qui apparaît comme une ambiguïté. Par exemple, l'équation analogique *je chante* : *tu chantes* $\doteq$ *je déjeune* : x admet deux solutions : $x =$ *tu déjeunes* et $x =$ *je détuunes*.

L'analyse d'une chaîne de symboles, c'est-à-dire la vérification de son appartenance à un langage de chaînes analogiques est en fait le parcours de la structure immanente de l'espace analogique. C'est une tâche qui ressortit à la pseudo-création que nous mentionnions plus haut. Il ne s'agit en fait que de l'exploration d'un donné. De cela résulte que la structure analogique universelle ne nous est normalement connue qu'à travers le filtre de notre langue, c'est-à-dire à travers un codage particulier par des symboles donnés. Cela, de la même façon que l'addition dans $\mathbb{N}$ ne nous est d'abord formulable qu'à travers un codage très particulier, celui des nombres de notre langue.

Et c'est ici qu'intervient le second volet de notre recherche, celui sur les homomorphismes. En effet, dans notre monde, les purs symboles n'existent

pas. Les symboles que nous manipulons sont en réalité souvent décomposables sous un angle quelconque, c'est-à-dire réinterprétables. C'est donc ici que se placerait toute une recherche qui reste à faire sur les homomorphismes entre structures analogiques induits par des homomorphismes tels que l'image d'un symbole élémentaire soit une chaîne d'autres symboles. En raison de l'unité fondamentale de structures, il s'agirait en fait de l'étude de sortes d'endomorphismes. Or c'est ici que se place le seul véritable caractère créateur de l'analogie, mais qui est lui-même encore un faux-semblant. En réalité, la seule liberté que nous ayons est celle de décider la façon dont s'effectue la réinterprétation des symboles, c'est-à-dire l'angle sous lequel on peut les décomposer. Cela revient dans notre vue de l'analogie à promulguer les correspondances initiales entre structures analogiques, celles dont on peut partir pour explorer les analogies conservées par homomorphisme. Mais ici aussi, une fois les correspondances initiales données, la structure de la correspondance est totalement définie par les structures analogiques sous-jacentes des deux espaces en correspondance. Sous l'angle des homomorphismes, on peut faire la distinction entre analogies au seul niveau des symboles et analogies systématiques au sens de la correspondance. Distinction qui rejoint l'articulation en anomalie et analogie. Au seul niveau des symboles, les analogies sont toutes, par définition, vraies. Mais si elles sont conservées dans le second espace analogique par la correspondance que l'on se donne, alors elles peuvent être déclarées systématiques, c'est-à-dire appartenant au système. Ainsi,

$$peindrai : peindre \doteq pourfendrai : pourfendre$$

est une analogie vraie au sens des symboles et au sens du français. C'est donc une analogie systématique en français. Tandis que

$$peindrai : peindre \doteq viendrai : viendre$$
$$peindrai : peindre \neq viendrai : venir$$

ne sont pas des analogies systématiques en français. L'une n'est pas vraie au sens de la conjugaison française, l'autre est fausse au sens des symboles. Dès lors se pose le problème de la réinterprétation des analogies entre symboles et de leur influence sur le système d'une langue donnée. Cela rejoint les travaux de Kuryłowicz. De façon général, le problème est donc de trouver une interprétation analogique à des structures locales imposées à nous de l'extérieur par un système global. Par exemple, comment faire en sorte d'imposer la conformité suivante ?

$$\square : \bigcirc \doteq \blacksquare : \bullet$$

Proposer un modèle de raisonnement qui permette cela, c'est montrer comment s'opère l'émergence de signes nouveaux qui fondent les symboles donnés. Et selon nous, c'est là, et seulement là que réside notre liberté. Une tentative de réponse pourrait poser un certain principe d'économie. On chercherait des

analogies minimales en un certain sens, pour les mettre en correspondance avec la conformité imposée. Par exemple, l'analogie minimale dont tous les termes sont différents est, au renommage de symboles et aux formes équivalentes de l'analogie près :

$$0 : 1 \doteq 01 : 11$$

L'établissement de la correspondance

$$0 = \square,\ 1 = \bigcirc,\ 01 = \blacksquare \text{ et } 11 = \bullet$$

permet un interprétation des symboles $\square$, $\bigcirc$, $\blacksquare$ et $\bullet$ dans laquelle les deux derniers termes de droite sont les premiers plus quelque chose. En l'occurrence, ce qui est ajouté ici serait la couleur noire. Une correspondance avec $00 : 01 \doteq 10 : 11$ aurait mené, elle, à une interprétation en forme géométrique et couleur. Mais ces analogies aussi jouent au seul niveau symbolique. Les informations dans chacune de ces catégories n'interviennent donc qu'en tant qu'elles s'opposent l'une à l'autre. Elles ne représentent pas immédiatement les couleurs blanche et noire, ou les formes de carré et de cercle, mais seulement l'opposition entre blanc et noir ou entre carré et cercle.

Ainsi donc lorsqu'une conformité de type analogique ne peut être établie directement, si une fonction de codage f peut être trouvée, pour laquelle une fonction inverse existe, le schéma de raisonnement suivant joue :

$$\mathrm{f}(A) : \mathrm{f}(B) \doteq \mathrm{f}(C) : D \qquad \Leftrightarrow \qquad A : B \equiv C : \mathrm{f}^{-1}(D)$$

Dans cette formulation, seule l'analogie de gauche est résoluble dans notre formalisation. De façon générale, cette formule explique comment certaines analogies au sens large peuvent être comprises comme presque systématiques. En effet, à un niveau plus élevé, les réinterprétations des symboles peuvent être faites dans un système connu. Ainsi la résolution de l'équation analogique suivante

$$abc : abd \doteq ijk : x \qquad \Rightarrow \quad x = ijl$$

peut être comprise par le passage par une numérotation des caractères de l'alphabet en binaire. Par exemple, avec Unicode :

$$\texttt{011000010110001001100011} : \texttt{011000010110001001100100} \doteq \texttt{011010010110101001101011} : x$$
$$\Rightarrow \quad x = \texttt{011010010110101001101100} = ijl$$

Et elles peuvent aussi passer par des descriptions formelles plus élaborées comme dans le cas

qui peut être comprise par la mise en correspondance avec le codage

$$\begin{array}{l} obj(gros)\& \\ obj{=}cercle \end{array} : \begin{array}{l} obj(petit){\subset}obj(gros) \\ \&obj{=}cercle \end{array} \doteq \begin{array}{l} obj(gros)\& \\ obj{=}carré \end{array} : x$$

ce qui permet bien d'obtenir la solution :

$$x = \begin{array}{l} obj(petit){\subset}obj(gros) \\ \&obj{=}carré \end{array}$$

Notre vue des choses s'incrit donc dans une vue saussurienne du signe. En effet, si, dans un homomorphisme entre domaines structurés par l'analogie, le premier domaine est un domaine de signifiants, c'est-à-dire de chaînes de symboles de la langue, et si le second domaine est un domaine de signifiés, c'est-à-dire de représentations du monde, alors, les correspondances établies sont des associations de signifiant à signifié, c'est-à-dire, en termes saussuriens, des signes. Ainsi donc, clairement, la formalisation de la métaphore que nous avons proposée montre comment de nouveaux signes peuvent être découverts par exploration en parallèle des structures analogiques des deux domaines. Le modèle dynamique que nous avons détaillé plus haut et qui peut s'appliquer dans les deux sens, du domaine du signifiant à celui du signifié, comme du signifié au signifiant, rend compte de la façon dont peut avoir lieu un travail constant de découvertes de nouvelles correspondances entre signifiants et signifiés. Du point de vue du découvreur, ces découvertes apparaissent comme des création de signes. Pour le moment, nous ne sommes en mesure que de proposer un tel modèle dans le cas où les signifiants comme les signifiés sont des chaînes de symboles. Si un jour nous sommes en mesure de vérifier des analogies et de résoudre des équations analogiques entre ondes sonores d'une part, grâce à une extension aux fonctions continues, et entre images d'autre part, grâce à une extension à deux dimensions, alors, notre modèle pourrait acquérir une certaine épaisseur.

Avec des linguistes comme Itkonen, nous croyons qu'il y a caché là un critère de scientificité positif qui est celui du caractère analogique des formalisations. Attention de nouveau, il ne s'agit pas ici de faire du Gentner, ce qui consisterait en l'application vague de quelque métaphore. Nous entendons par caractère analogique le fait que la formalisation (partielle) que nous avons proposée joue mécaniquement sur les symboles utilisés par une théorie proposée pour tirer, dans le cadre de la théorie proposée, de nouvelles propositions de la théorie. Car l'analogie permet la découverte de nouvelles propositions dans la théorie. Elle ne permet pas seule l'émergence de nouveaux signes, et encore moins de nouveaux symboles. Mais la liberté réside dans l'affirmation initiale de certains signes. En raison de l'immanence de la structure homomorphique entre deux espaces analogiques, nous devons alors nous contenter de découvrir les nouveaux signes impliqués par cette initialisation. Dans un tel cadre, le caractère aveugle de l'analogie n'est pas une calamité, c'est au contraire une bénédiction. Sans lui, la confrontation entre plusieurs espaces analogiques ne serait pas possible. Une analogie dans un seul domaine n'est en effet pas suffisante à l'établissement de la vérité, puisque, comme nous venons de le voir elle est déjà contenue dans une structure immanente. En revanche, la vérité peut sourdre de la confrontation de plusieurs espaces. Du point de vue

de l'explorateur-découvreur, la liberté à poser des correspondances initiales, c'est-à-dire des signes initiaux, permet de provoquer l'émergence de nouveaux signes par la confrontation de structures analogiques sur lesquelles il travaille. En résumé donc, la liberté de créer ou de remettre en question des correspondances entre espaces analogiques préside à la création de nouveaux signes d'une théorie. La seule résolution des analogies dans l'un seulement des espaces de symboles préside à l'émission de nouvelles propositions dans le cadre de la nouvelle théorie. La vérification de la conservation des analogies d'un espace à un ou plusieurs autres préside, en dernier lieu, à l'établissement de la vérité.

Dans un tel cadre, la découverte de lieux où la conservation des analogies ne joue plus permet la réfutation de la formalisation. On sait que le seul critère de scientificité reconnu est pour le moment celui proposé par Popper, la réfutabilité. Il n'y aurait pas de théorie vraie, il n'y aurait pas de vérifiable, il ne pourrait y avoir que du réfutable. C'est seulement la proposition d'expériences de réfutation qui ferait d'une théorie qu'elle est scientifique. Ce que nous venons de décrire pourrait donner lieu à un autre critère de scientificité. Son corset est l'immanence de la structure analogique, sa liberté est dans ses signes initiaux. Ainsi donc, si le critère de Popper est externe à une théorie et joue sur l'adéquation avec le monde, celui-là est en quelque sorte interne à la théorie et joue sur la partie de la modélisation. Les deux nous semblent nécessaires, en ce qu'il contraignent une théorie, l'un de l'intérieur, l'autre de l'extérieur, l'un du côté de la formalisation, l'autre du côté de l'adéquation avec le monde.

Annexes

Annexe A

Classification morphologique des phénomènes

Nous donnons ici des exemples d'analogies classifiés selon une vue traditionnelle des phénomènes en morphologie, c'est-à-dire par préfixation, suffixation ou infixation. En fait, cette classification apparaît complexe parce que, pensons-nous, elle essaie de rendre compte de l'analogie sous un angle erroné. Rappelons que la formalisation (partielle) que nous avons proposée de l'analogie entre chaînes de symboles ne se sert pas de ces notions. Nous avions insisté, lorsque nous avions critiqué les modèles de Nagao et d'Yvon, sur le fait que l'approche qui considère une transformation à la fois nous apparaissait comme peu satisfaisante pour deux raisons. Premièrement, parce qu'elle mettait en avant la notion dynamique de transformation, alors que l'analogie privilégie la notion statique de rapport. Deuxièmement, parce qu'elle se limitait à une seule transformation à la fois. Or, plusieurs transformations sont absolument nécessaires pour rendre compte des phénomènes de la langue. À cause de cette vue, on ne saurait délimiter clairement certains cas. Ainsi, l'analogie *wyszedłeś* : *wyszłaś* ≐ *poszedłeś* : *poszłaś* [1] de la conjugaison polonaise est difficile à classifier du point de vue purement formel. Doit-on y voir un échange de suffixes (*-edłeś* avec *-łaś*) ? S'agit-il d'un simple échange d'infixes (*-edłe* et *-ła* dans *wysz–ś* ou *posz–ś*) ? Ou encore d'un infixation multiple (*-ed-e* et *-a* dans *wysz–ł–ś* ou *posz–ł–ś*) ? La question est difficile à trancher du point de vue de la langue même, puisque, étymologiquement, on peut reconstruire un radical *'st* qui apparaît sous une forme mouillée *jść* dans les infinitifs *pójść* et *wyjść*, et prendrait la forme *-szd* au passé masculin seulement. Mais ce dernier radical n'apparaît nulle part ailleurs dans la conjugaison.

Voici donc maintenant, pêle-mêle, sous chaque rubrique un certain nombre d'exemples en différentes langues.

[1] *Wyszedłeś* (tu es sorti) et *poszedłeś* (tu es allé) sont les formes du masculin. *Wyszłaś* et *poszłaś* sont les formes correspondantes au féminin (tu es sortie) et (tu es allée).

A.1 Insertions ou suppressions d'un préfixe

Allemand : $lachen : \ddot{u}berlachen \doteq setzen : x \quad \Rightarrow \quad x = \ddot{u}bersetzen$ [2]
Français : $fini : infini \doteq exact : x \quad \Rightarrow \quad x = inexact$
Français : $faire : défaire \doteq visser : x \quad \Rightarrow \quad x = dévisser$
Français : $aventure : mésaventure \doteq alliance : x \quad \Rightarrow \quad x = mésalliance$
Polonais : $zwyczajny : nadzwyczajny \doteq mierny : x \quad \Rightarrow \quad x = nadmierny$ [3]

A.2 Remplacement de préfixes

Polonais : $odjechać : wyjechać \doteq odjeździsz : x \quad \Rightarrow \quad x = wyjeździsz$ [4]
Français : $antigouvernemental : progouvernemental \doteq antilibéral : x \quad \Rightarrow \quad x = prolibéral$

A.3 Insertion ou suppression d'un suffixe

Chinois : 科学 : 科学家 ≒ 政治 : $x \quad \Rightarrow \quad x =$ 政治家 [5]
Latin : $oratorem : orator \doteq honorem : x \quad \Rightarrow \quad x = honor$ [6]
Français : $répression : répressionnaire \doteq réaction : x \quad \Rightarrow \quad x = réactionnaire$

A.4 Remplacement de suffixes

Français : $mangerons : mangerais \doteq tremperons : x \quad \Rightarrow \quad x = tremperais$
Japonais : 食べる : 食べます ≒ 認める : $x \quad \Rightarrow \quad x =$ 認めます [7]

A.5 Insertions, suppressions ou remplacements d'un préfixe et d'un suffixe

Japonais : 自由 : 不自由な ≒ 用意 : $x \quad \Rightarrow \quad x =$ 不用意な [8]

[2] *Lachen* (rire), *überlachen* (rire exagérément), *setzten* (poser), *übersetzen* (traduire). Analogie purement formelle.

[3] *Zwyczajny* (habituel), *nadzwyczajny* (exceptionnel), *mierny* (mesuré), *nadmierny* (qui est outre mesure). Ces quatre adjectifs sont au nominatif masculin singulier.

[4] *Odjechać* (partir, s'en aller), *wyjechać* (sortir), *odjeździsz* (tu pars, tu t'en vas), *wyjeździsz* (tu sors).

[5] 科学 /kēxué/ (la science), 科学家 /kēxuéjiā/ (un scientifique), 政治 /zhèngzhì/ (la politique), 政治家 /zhèngzhìjiā/ (un homme politique).

[6] *Orator* (un orateur), nominatif, *oratorem* même sens, accusatif, *honor* (l'honneur), nominatif, *honorem* même sens, accusatif.

[7] 食べる /taberu/ (manger), forme de politesse neutre, 食べます /tabemasu/ même sens, forme marquée de politesse, 決める /kimeru/ (décider), forme de politesse neutre, 決めます /kimemasu/ même sens, forme marquée de politesse.

[8] 自由 /ziyuu/ (liberté), 不自由な /huziyuuna/ (géné, privé de, handicapé), 用意 /youi/ (préparation), 不用意な /huyouina/ (imprévoyant, imprudent).

Malais-indonésien : $tinggal : ketinggalan \doteq duduk : x \Rightarrow x = kedudukan$ [9]

Malais-indonésien : $kawan : mengawani \doteq keliling : x \Rightarrow x = mengelilingi$ [10]

Malais-indonésien : $keras : mengeraskan \doteq kena : x \Rightarrow x = mengenakan$ [11]

Polonais : $tłumaczymy : tłumaczone \doteq mijamy : x \Rightarrow x = mijane$ [12]

A.6 Insertion ou suppression d'un infixe

Japonais : 乗る : 乗せる ≒ 寄る : $x \Rightarrow x =$ 寄せる [13]

Akkadien : $ukaššad : uktanaššad \doteq ušakšad : x \Rightarrow x = uštanakšad$ [14]

A.7 Remplacement d'un infixe

Polonais : $wyszedłeś : wyjechałeś \doteq poszedłeś : x \Rightarrow x = pojechałeś$ [15]

A.8 Insertions, suppressions ou échanges de plusieurs infixes

Polonais : $wyszedłeś : wyszłaś \doteq poszedłeś : x \Rightarrow x = poszłaś$ [16]

Allemand : $lang : längste \doteq scharf : x \Rightarrow x = schärfste$ [17]

Allemand : $sprechen : du\ sprächest \doteq nehmen : x \Rightarrow x = du\ nähmest$ [18]

Polonais : $zgubiony : zgubieni \doteq zmartwiony : x \Rightarrow x = zmartwieni$ [19]

---

[9] *Tinggal* (demeurer, habiter), *ketinggalan* (rater, manquer), *duduk* (asseoir, installer), *kedudukan* (situation, grade).

[10] *Kawan* (un ami ; un troupeau (ou groupe d'animaux)), *mengawani* (accompagner), *keliling* (le pourtour, les environs), *mengelilingi* (entourer).

[11] *Keras* (raide, dur), *mengeraskan* (raidir, durcir), *kena* (toucher, être en contact, être attaché), *mengenakan* (mettre [un vêtement], attacher, infliger).

[12] *Tłumaczymy* (nous expliquons), *tłumaczone* (expliqué), participe passé neutre singulier, *mijamy* (nous passons, nous évitons), *mijane* (passé, évité), participe passé neutre singulier.

[13] 乗る /noru/ (monter [en voiture]), 乗せる /noseru/ (faire monter), 寄る /yoru/ (tirer [à soi]), 寄せる /yoseru/ (faire s'approcher, attirer).

[14] KURYŁOWICZ, *L'apophonie en sémitique*, 1961, p. 64.

[15] *Wyszedłeś* (tu es sorti [à pied]), masculin, *wyjechałeś* (tu es sortie [en véhicule]), féminin, *poszedłeś* (tu es allé [à pied]), masculin, *pojechałeś* (tu es allée [en véhicule]), féminin.

[16] *Wyszedłeś* (tu es sorti [à pied]), masculin, *wyszłaś* (tu es sortie), féminin, *poszedłeś* (tu es allé [à pied]), masculin, *pojechałeś* (tu es allée), féminin.

[17] *Lang* (long), *längste* (le plus long), *scharf* (aiguisé), *schärfste* (le plus aiguisé).

[18] *Sprechen* (parler), *du sprächest* (tu aurais parlé), *nehmen* (prendre), *du nähmest* (tu aurais pris).

[19] *Zgubiony* (perdu), masculin singulier, *zgubieni* (perdus), masculin personnel pluriel, *zmartwiony* (tracassé, perplexe), masculin singulier, *zmartwieni* (tracassés, perplexes), masculin personnel pluriel.

Allemand : $\textit{fliehen} : \textit{er floh} \doteq \textit{schließen} : x \quad \Rightarrow \quad x = \textit{er schloß}$ [20]
Arabe : $\textit{huzila} : \textit{huzāl} \doteq \textit{ṣudi‘a} : x \quad \Rightarrow \quad x = \textit{ṣudā‘}$ [21]
Arabe : $\textit{arsala} : \textit{mursilun} \doteq \textit{aslama} : x \quad \Rightarrow \quad x = \textit{muslimun}$ [22]
Arabe : $\textit{arsala} : \textit{mursil} \doteq \textit{aslama} : x \quad \Rightarrow \quad x = \textit{muslim}$ [23]
Arabe : $\textit{kataba} : \textit{kātib} \doteq \textit{sakana} : x \quad \Rightarrow \quad x = \textit{sākin}$ [24]
Arabe : $\textit{kataba} : \textit{maktab} \doteq \textit{sakana} : x \quad \Rightarrow \quad x = \textit{maskan}$ [25]
Hébreu : $\underline{\textit{ia}}\textit{ḥmo}\bar{\textit{d}} : \textit{maḥmå}\bar{\textit{d}} \doteq \underline{\textit{ia}}\textit{‘a}\bar{\textit{b}}\textit{o}\underline{\textit{r}} : x \quad \Rightarrow \quad x = \textit{ma‘a}\bar{\textit{b}}\textit{å}\underline{\textit{r}}$ [26]

A.9 Interversions

Proto-sémitique : $\underline{\textit{ia}}\textit{sriqu} : \textit{sariq} \doteq \underline{\textit{ia}}\textit{nqimu} : x \quad \Rightarrow \quad x = \textit{naqim}$ [27]

[20] *Fliehen* (voler), *er floh* (il vola, il volait), *schließen* (fermer), *er schloß* (il ferma, il fermait).

[21] Kuryłowicz, *L'apophonie en sémitique*, 1961, p. 114.

[22] *Arsala* (il envoya), *mursilun* (un envoyé), *aslama* (il se convertit (sous-entendu à l'Islam)), *muslimun* (un converti, c'est-à-dire un musulman). *Mursil* et *muslim* sont les formes marquées.

[23] *Arsala* (il envoya), *mursil* (un envoyé), *aslama* (il se convertit (sous-entendu à l'Islam)), *muslim* (un converti, c'est-à-dire un musulman). *Mursil* et *muslim* sont les formes vernaculaires. Voir Kouloughli, *Grammaire de l'arabe d'aujourd'hui*, 1994, p. 87, où il est dit que, phonétiquement, seules les flexions de l'état « d'annexion » sont obligatoirement réalisées, c'est-à-dire, pour notre exemple, *muslimu* au nominatif, *muslima* à l'accusatif et *muslimi* au génitif. Les formes de l'état indéterminé *muslimun* au nominatif, *musliman* à l'accusatif et *muslimin* au génitif, qui sont obtenues par ajout d'un *tanwīn* (ibidem, p. 79), peuvent être réalisées *muslim* et sont toujours élidées à la pause.

[24] *Kataba* (il écrivit), *kātib* (un écrivain), *sakana* (il habita), *sākin* (un habitant).

[25] *Kataba* (il écrivit), *maktab* (un lieu où l'on écrit), *sakana* (il habita), *maskan* (une habitation).

[26] Kuryłowicz, *L'apophonie en sémitique*, 1961, p. 127.

[27] Kuryłowicz, *L'apophonie en sémitique*, 1961, p. 89.

Annexe B

Quelques données morphologiques

B.1 Conjugaison latine

Nous donnons tout d'abord un exemple ou la régularité est parfaite. Les verbes latins ont une conjugaison relativement semblable dans l'ensemble. Cependant, on note des différences selon la voyelle finale du radical. Il existe trois sortes de voyelles finales longues : *a*, *e* et *i*. Les terminaisons sont des suffixes commençant par des voyelles brèves. Par conséquent, si le radical d'un verbe se termine par une consonne, alors son type exhibe une voyelle brève, à l'infinitif un *ĕ*. Chacun des ces quatre types majeurs est exemplifié par un verbe modèle : *amāre*, *delēre*, *finīre* et *capĕre*.

Voici une description de Quintilien :

> (7) La comparaison se fait de la même manière pour les verbes. Si quelqu'un, à l'imitation des anciens, prononçait brève la pénultième de *feruere*, il serait convaincu de mal parler, parce que tous les verbes qui ont l'indicatif terminé en *eo*, lorsque l'infinitif de ces verbes est en *ere*, ont toujours ce premier *e* long, *prandeo*, *pendeo*, *spondeo*, *prandere*, *pendere*, *spondere* ;
>
> (8) tandis que ceux qui n'ont qu'un *o* à l'indicatif, et qui ont aussi l'infinitif en *ere*, comme *lego*, *dico*, *curro*, ont cet *e* bref, *legere*, *icere*, *currere*...[1]

Le type le plus régulier, qui était évidemment le type productif (avec les verbes exprimant la transformation en -*īre*), est le type des verbes en -*āre*. Nous donnons ci-dessous quelques formes de cette conjugaison avec les temps présent, imparfait et futur.

$$\begin{array}{lcllcl} amare & : & amas & \doteq & ambulare & : & ambulas \\ amare & : & amas & \doteq & ambulare & : & ambulas \\ amare & : & amat & \doteq & ambulare & : & ambulat \end{array}$$

[1] QUINTILIEN, *L'institution oratoire*, 2000, chapitre 6, § 7 à 9.

$$
\begin{array}{lcllcl}
amare & : & amamus & \doteq & ambulare & : ambulamus \\
amare & : & amatis & \doteq & ambulare & : ambulatis \\
amare & : & amant & \doteq & ambulare & : ambulant
\end{array}
$$

$$
\begin{array}{lcllcl}
amare & : & amabam & \doteq & ambulare & : ambulabam \\
amare & : & amabas & \doteq & ambulare & : ambulabas \\
amare & : & amabat & \doteq & ambulare & : ambulabat \\
amare & : & amabamus & \doteq & ambulare & : ambulabamus \\
amare & : & amabatis & \doteq & ambulare & : ambulabatis \\
amare & : & amabant & \doteq & ambulare & : ambulabant
\end{array}
$$

$$
\begin{array}{lcllcl}
amare & : & amabo & \doteq & ambulare & : ambulabo \\
amare & : & amabis & \doteq & ambulare & : ambulabis \\
amare & : & amabit & \doteq & ambulare & : ambulabit \\
amare & : & amabimus & \doteq & ambulare & : ambulabimus \\
amare & : & amabitis & \doteq & ambulare & : ambulabitis \\
amare & : & amabunt & \doteq & ambulare & : ambulabunt
\end{array}
$$

B.2 Déclinaison polonaise

Nous donnons maintenant un exemple de déclinaison dans une langue slave. Il s'agit de la déclinaison des noms féminins en polonais. Grossièrement, cette déclinaison est régulière, mais elle exhibe tout de même deux lieux d'anomalie.

Singulier

$$
\begin{array}{lcllcl}
\text{Vocatif:} & kobieta & : kobieto & \doteq & perła & : perło \\
\text{Accusatif:} & kobieta & : kobietę & \doteq & perła & : perłę \\
\text{Génitif:} & kobieta & : kobiety & \doteq & perła & : perły \\
\text{Datif:} & kobieta & : kobiecie & \neq & perła & : perle \\
\text{Instrumental:} & kobieta & : kobietą & \doteq & perła & : perłą
\end{array}
$$

Pluriel

$$
\begin{array}{lcllcl}
\text{Nominatif:} & kobieta & : kobiety & \doteq & perła & : perły \\
\text{Vocatif:} & kobieta & : kobiety & \doteq & perła & : perły \\
\text{Accusatif:} & kobieta & : kobiety & \doteq & perła & : perły \\
\text{Génitif:} & kobieta & : kobiet & \neq & perła & : pereł \\
\text{Datif:} & kobieta & : kobietom & \doteq & perła & : perłom \\
\text{Instrumental:} & kobieta & : kobietami & \doteq & perła & : perłami
\end{array}
$$

Les irrégularités apparaissent au datif singulier et au génitif pluriel.

Au datif singulier, un phénomène de palatalisation a lieu sur les goupes consonantiques finals durs du radical. On peut en fait considérer que la désinence

du datif singulier est *-'e*, où l'apostrophe note la palatalisation. Avec cette interprétation, l'analogie est rétablie : *kobieta* : *kobiet'e* ≐ *perła* : *perł'e*. On compte trois, voire quatre, niveaux possibles de palatalisation en polonais, illustrés dans le tableau B.1[2]. Ce tableau montre que le *-ta* de *kobieta* doit donc passer à *-cie*, et que le *-ła* (consonne dure) de *perła* doit passer à *-le*.

Tableau B.1: Palatalisation au datif singulier féminin ou neutre

radical	datif	degrés possibles de palatalisation
-b-	*-bie*	*b* : *b'*
-p-	*-pie*	*p* : *p'*
-f-	*-fie*	*f* : *f'*
-w-	*-wie*	*w* : *w'*
-m-	*-mie*	*m* : *m'*
-n-	*-nie*	*n* : *n'*
-r-	*-rze*	*r* : *rz*
-ł-	*-le*	*ł* : *l*
-sł-	*-śle*	*sł* : *śl*
-t-	*-cie*	*t* : *ć* : *c* : *cz*
-st-	*-ście*	*st* : *ść* : *szcz*
-d-	*-dzie*	*d* : *dź* : *dz*
-zd-	*-ździe*	*z d* : *źdź* : *żdż*
-s-	*-sie*	*s* : *ś* : *sz*
-z-	*-zie*	*z* : *ź* : *ż*
-k-	*-ce*	*k* : *k'* : *cz* : *c*
-sk-	*-sce*	*sk* : *sk'* : *szcz* : *sc*
-g-	*-dze*	*g* : *g'* : *ż* : *dz*
-zg-	*-zdze*	*zg* : *zg'* : *żdż* : *zdz*
-ch	*sze*	*ch* : *ch'* : *sz* : *ź*

Au génitif pluriel, les féminins et les neutres perdent toute désinence. On peut dire aussi qu'ils prennent une désinence zéro *-∅*. L'accent du mot, à de rares exceptions près, toujours sur la pénultième en polonais, recule donc d'une voyelle dans ce cas. Mais si aucune voyelle n'est disponible, comme c'est le cas pour *perł*, et s'il y a possibilité d'insertion d'une voyelle entre deux consonnes dasn le groupe consonantique final, alors cette épenthèse a lieu. Ici, on a donc : *perł* → *perel*.

[2]Voir Jagodziński, *Gramatyka języka polskiego*, 2001, rubrique *Alternacje w polskiej morfologii* (Alternance dans la morphologie polonaise).

B.3 Morphologie dérivationnelle du malais

Nous donnons maintenant un exemple en morphonologie dérivationnelle dans une langue austronésienne, le malais-indonésien. L'exemple que nous avons choisi est la formation des factitifs ou des V_0 selon la nomenclature de Mel'čuk[3] qui présente quelques irrégularités formelles.

lanjut	:	*melanjut*	$\doteq$	*lompat*	:	*melompat*	[4]
lanjut	:	*melanjut*	$\doteq$	*rasa*	:	*merasa*	[5]
lanjut	:	*melanjut*	$\doteq$	*rokok*	:	*merokok*	[6]
lanjut	:	*melanjut*	$\doteq$	*masak*	:	*memasak*	[7]
lanjut	:	*melanjut*	$\doteq$	*nilai*	:	*menilai*	[8]
lanjut	:	*melanjut*	$\doteq$	*nyanyi*	:	*menyanyi*	[9]
buka	:	*membuka*	$\doteq$	*buka*	:	*membuka*	[10]
buka	:	*membuka*	$\neq$	*potong*	:	*memotong*	[11]
pakai	:	*memakai*	$\doteq$	*potong*	:	*memotong*	[12]
cuci	:	*mencuci*	$\doteq$	*dapat*	:	*mendapat*	[13]
cuci	:	*mencuci*	$\doteq$	*cantas*	:	*mencantas*	[14]
cuci	:	*mencuci*	$\neq$	*tulis*	:	*menulis*	[15]
ambil	:	*mengambil*	$\doteq$	*isi*	:	*mengisi*	[16]
ambil	:	*mengambil*	$\doteq$	*ukur*	:	*mengukur*	[17]
ambil	:	*mengambil*	$\doteq$	*epak*	:	*mengepak*	[18]
ambil	:	*mengambil*	$\doteq$	*olah*	:	*mengolah*	[19]
ambil	:	*mengambil*	$\doteq$	*gunting*	:	*menggunting*	[20]
ambil	:	*mengambil*	$\doteq$	*hitung*	:	*menghitung*	[21]
ambil	:	*mengambil*	$\neq$	*kirim*	:	*mengirim*	[22]

[3] Mel'čuk *et al.*, *Dictionnaire explicatif et combinatoire du français contemporain. Recherche lexico-sémantiques I*, 1984, p. 7.

[4] Exemples essentiellement tirés de 舟田 京子 (Hunada Kyouko), やさしい初歩のインドネシア語, 1997, p. 148 et 149. *Lanjut* (long, prolongé), *melanjut* (continuer). *Lompat* (sauter), *melompat* (sauter).

[5] *Rasa* (goûter, sentir), *merasa* (goûter, sentir, ressentir).

[6] *Rokok* (fumer), *merokok* (fumer [une cigarette]).

[7] *Masak* (cuire), *memasak* (faire cuire).

[8] *Nilai* (valoir), *menilai* (évaluer, estimer, juger).

[9] *Nyanyi* (chanter), *menyanyi* (interpréter [une œuvre]).

[10] *Buka, membuka* (ouvrir).

[11] *Pakai* (utiliser, porter [un vêtement]), *memakai* (utiliser, mettre [un vêtement]).

[12] *Potong* (une tranche [de pain]). *memotong* (couper).

[13] *Cuci* (nettoyer), *mencuci* (essuyer, laver). *Dapat* (obtenir, pouvoir). *mendapat* (acquérir, attraper, obtenir, gagner).

[14] *Cantas* (couper, tailler), *mencantas* (tailler [un arbre]).

[15] *Tulis* (écrire), *menulis* (écrire).

[16] *Ambil* (prendre), *mengambil* (déposséder, retirer ; prendre, adopter). *Isi* (contenu), *mengisi* (remplir, faire entrer, bourrer).

[17] *Ukur* (se mesurer), *mengukur* (mesurer).

[18] *Epak* (être empaqueté), *mengepak* (empaqueter).

[19] *Olah* (manière, façon), *mengolah* (construire, planifier).

[20] *Gunting* (ciseaux), *menggunting* (cisailler, couper).

[21] *Hitung* (compte), *menghitung* (faire le compte, calculer).

[22] *Kirim* (envoyer, poster) *mengirim* (envoyer).

simpan	:	*menyimpan*	$\doteq$	*sapu*	:	*menyapu* [23]
simpan	:	*menyimpan*	$\doteq$	*sara*	:	*menyara* [24]

Afin de forcer la rigueur du système, c'est-à-dire pour le rendre complètement analogique, nous avons déjà mentionné que les descriptions du malais postulent un archiphonème *N* qui se comporterait différemment selon la voyelle ou la consonne qu'il précède en composition. Cet archiphonème n'intervient pas que dans la formation des factitifs en *meN-*. Il est aussi présent dans le préfixe *peN-* qui sert à former les noms d'agents ou d'instruments. Par exemple, on a la série: *kirim* (envoyer), *mengirim* (poster), *pengirim* (envoyeur). Ou encore *sapu* (nettoyer, balayer), *menyapu* (balayer, enduire), *penyapu* (un balai). Dans le cas des voyelles, la réalisation de l'archiphonème est la consonne simple *ng*. Il s'agit bien de la nasale notée par γ en grec: ἄγγελος (envoyé, ange) ou ἀνάγκη (la nécessité, le destin), et par *ng* en allemand: *Menge* (un ensemble). Les règles de composition sont données dans le tableau B.2.

Tableau B.2: Réalisation de l'archiphonème *N*

-*N* + consonne	résultat de la composition
-*N* + *b*-	-*mb*-
-*N* + *p*-	-*m*-
-*N* + *c*-	-*nc*-
-*N* + *d*-	-*nd*-
-*N* + *j*-	-*nj*-
-*N* + *t*-	-*n*-
-*N* + *g*-	-*ngg*-
-*N* + *h*-	-*ngh*-
-*N* + *k*-	-*ng*-
	(disparition de *N*)
-*N* + *l*-	-*l*-
-*N* + *r*-	-*r*-
-*N* + *m*-	-*m*-
-*N* + *n*-	-*n*-
-*N* + *ny*-	-*ny*-
-*N* + *w*-	-*w*-
-*N* + *y*-	-*y*-

De ce tableau, il apparaît clairement que l'on a les transformations phonétiques habituelles pour une nasale, sauf dans le cas de consonnes sourdes simples, qui sont absorbées dans la réalisation de l'archiphonème *N*.

[23] *Simpan* (garder, mettre de côté), *menyimpan* (garder, conserver, avoir). *Sapu* (nettoyer, balayer), *menyapu* (balayer, enduire).

[24] *Sara* (traitement, pension), *menyara* (entretenir, financer).

Tableau B.3: Règles de réalisation de l'archiphonème *N*

-*N* + consonne	résultat de la composition
-N + labiale sonore ou composée	*-m + la consonne*
-N + dentale sonore ou composée	*-n + la consonne*
-N + vélaire sonore ou composée	*-ng + la consonne*
-N + labiale sourde simple (p)	*-m*
-N + dentale sourde simple (t)	*-n*
-N + vélaire sourde simple (k)	*-ng*

Annexe C

Distance d'édition entre chaînes

Dans cette section[1], nous introduisons les distances d'édition et nous montrons que ce sont de vraies distances au sens mathématique sur l'ensemble des chaînes si et seulement si elles sont définies à partir de distances sur les éléments du vocabulaire. On montre aussi quelques autres propriétés.

C.1 Opération d'édition

DÉFINITION **29 (Opération d'édition)** *Soit* $(a,b) \in (\mathcal{V} \cup \{\varepsilon\})^2$, (a,b) *est appelé une* opération d'édition *et est noté* $a \to b$ *si et seulement si* $a \neq b$. *On note par* $\mathcal{E}$ *l'ensemble des opérations d'édition.*

Une opération d'édition $a \to b$ est appelée une *insertion* si et seulement si $a = \varepsilon$; c'est une *suppression* si et seulement si $b = \varepsilon$ et c'est un *remplacement* sinon.

Une chaîne $u \in \mathcal{V}^*$ dérive directement vers la chaîne $v \in \mathcal{V}^*$ si et seulement si

$$\exists\, a \to b \in \mathcal{E},\ \exists\, (x,y) \in (\mathcal{V}^*)^2 \ /\ u = x.a.y \wedge v = x.b.y$$

LEMME **22 (Unique opération d'édition)** *Si* $u \in \mathcal{V}^*$ *dérive directement vers* $v \in \mathcal{V}^*$ *alors,* $\exists!\, a \to b \in \mathcal{E},\ \exists\, (x,y) \in (\mathcal{V}^*)^2 \ /\ u = x.a.y \wedge v = x.b.y$

Ce lemme signifie que l'opération d'édition qui intervient dans la dérivation directe est unique. Mais l'endroit où elle s'applique peut ne pas être unique. Par exemple, soit la dérivation directe de aaa vers aa par l'unique opération d'édition $a \to \varepsilon$. Cette opération peut s'appliquer sur chacub des trois a de la chaîne.

DÉMONSTRATION : Supoosons qu'il existe deux opérations d'éditiona $\to b$ et $a' \to b'$. Alors, il existe x, y, x', y' dans $\mathcal{V}^*$ tels que

$$\begin{cases} u &= x.a.y &= x'.a'.y' \\ v &= x.b.y &= x'.b'.y' \end{cases}$$

[1] Reprise de l'article LEPAGE, *Regular languages have regular neighbourhoods for the Wagner and Fischer distance*, 1994, p. 14–16.

Il y a quatre cas, qui peuvent être factorisés en deux cas, si l'on remarque que x est un préfixe de x', et soit y est un préfixe de y' soit y' est un préfixe de y. On a donc le système d'équations suivant :

$$\left\{\begin{array}{lcl} a.y'' & = & x''.a' \\ b.y'' & = & x''.b' \end{array}\right. \vee \left\{\begin{array}{lcl} a.y & = & x''.a'.y'' \\ b.y & = & x''.b'.y'' \end{array}\right.$$

qui n'est satisfait que lorsque $a = b$ and $a' = b'$ (ce qui est impossible par définition des opérations d'édition), ou lorsque $a = a'$ et $b = b'$. CQFD

LEMME **23** *Si $u \in \mathcal{V}^*$ dérive directement vers $v \in \mathcal{V}^*$ avec $b \to c$ comme opération d'édition, alors, pour tout $a \in \mathcal{V}$, $a.u$ dérive directement vers $a.v$ avec la même opération d'édition.*

DÉMONSTRATION : Soit $b' \to c'$ l'opération d'éditionimpliquée dans la dérivation directe de $a.u$ vers $a.v$.

$$\exists (x', y') \in (\mathcal{V}^*)^2 \ / \ \begin{array}{lcl} a.u & = & x'.b'.y' \\ a.v & = & x'.c'.y' \end{array}$$

$x' = \varepsilon$ est impossible car $b' \neq c'$ par définition des opérations d'édition. Donc, $x' = a.x''$ et

$$\left\{\begin{array}{lcl} a.u & = & a.x''.b'.y' \\ a.v & = & a.x''.c'.y' \end{array}\right.$$

ce qui implique que $u = x''.b'.y' \wedge v = x''.c'.y'$. Par unicité de l'opération d'éditiondans une dérivation directe, nécessairement, $b = b' \wedge c = c'$. CQFD

On propose maintenant une extension naturelle de la dérivation directe. Une chaîne $u \in \mathcal{V}^*$ dérive vers une chaîne $v \in \mathcal{V}^*$ si et seulement si

$$\exists\, (u_1, u_2, \ldots, u_n) \in (\mathcal{V}^*)^n / \ \begin{array}{l} u = u_1 \wedge \\ v = u_n \wedge \\ \forall i \,/\, 1 \leq i < n,\ u_i \text{ direct } u_{i+1} \end{array}$$

C.2 Distance de Wagner et Fischer

On peut affecter à chaque opération d'édition $a \to b$ une valeur $\delta(a, b)$. Par conséquent, toute dérivation directe peut aussi se voir affectée la valeur de son unique opération d'édition. Pour une dérivation de u_1 vers u_n, on peut donc calculer le minimum sur toutes les dérivations possibles. C'est ainsi qu'est définie la distance de Wagner et Fischer[2].

[2]Voir WAGNER & FISCHER, *The string-to-string correction problem*, 1974 pour une présentation différente.

DÉFINITION **30 (Distance de Wagner et Fischer)** *Soit $\mathcal{V}$ un vocabulaire,* dist *une fonction de $(\mathcal{V} \cup \{\varepsilon\})^2$ dans $\mathbb{R}^+$,* dist *est étendue à $(\mathcal{V}^*)^2$ de la façon suivante : $\forall (u,v) \in (\mathcal{V}^*)^2$,*

$$\begin{aligned} u = v &\Leftrightarrow \delta(u,v) = 0 \\ u \neq v &\Leftrightarrow \delta(u,v) = \min(\sum_{i=1}^{i=n-1} \delta(u_i, u_{i+1})) \end{aligned}$$

sur toutes les dérivations $(u_1, \ldots, u_n)$ de u vers v.

DÉMONSTRATION : Nous devons prouver l'existence de $\delta(u,v)$ pour tout $(u,v) \in (\mathcal{V}^*)^2$ / $u \neq v$. Soit $u = a_1 \ldots a_m$ et $v = b_1 \ldots b_n$ avec $a_i, b_i \in \mathcal{V}$. De façon évidente, la dérivation

$$(a_1 \ldots a_m,\ \ldots,\ a_m,\ \varepsilon,\ b_1,\ b_1.b_2,\ \ldots,\ b_1 \ldots b_n)$$

existe. CQFD

La distance de Wagner and Fischer peut être définie d'une autre façon. Nous ne prouverons pas l'équivalence entre les deux définitions.

DÉFINITION **31 (Distance de Wagner et Fischer)** *Soit $\mathcal{V}$ un vocabulaire,* dist *une distance sur $\mathcal{V} \cup \{\varepsilon\}$,* dist *peut être étendue à $\mathcal{V}^*$ de la façon suivante : $\forall (a,b) \in (\mathcal{V})^2$, $\forall (u,v) \in (\mathcal{V}^*)^2$,*

$$\begin{aligned} \delta(a.u, \varepsilon) &= \delta(a,\varepsilon) + \delta(u,\varepsilon) \\ \delta(a.u, b.v) &= min(\begin{array}{l} \delta(a,\varepsilon) + \delta(u, b.v), \\ \delta(a,b) + \delta(u,v), \\ \delta(\varepsilon,b) + \delta(a.u, v) \end{array}) \end{aligned}$$

dist *est la distance de Wagner et Fischer sur $\mathcal{V}^*$.*

Cette définition est celle qui est habituellement retenue, car directement applicable à l'implantation du calcul des distances entre chaînes de symboles par programme.

C.3 Distances au sens mathématique

Nous rappelons la définition d'une distance au sens mathématique.

DÉFINITION **32 (Distance)** *Soit $\mathcal{V}$ un ensemble,* dist *une fonction de $\mathcal{V} \times \mathcal{V}$ dans $\mathbb{R}^+$, l'ensemble des réels positifs,* dist *est une distance sur $\mathcal{V}$ si et seulement si*

- *(égalité)*
  $\forall (a,b) \in \mathcal{V}^2,\ \delta(a,b) = 0 \Leftrightarrow a = b$
- *(commutativité)*
  $\forall (a,b) \in \mathcal{V}^2,\ \delta(a,b) = \delta(b,a)$

- *(inégalité triangulaire)*
  $\forall (a,b,c) \in \mathcal{V}^3, \delta(a,c) \leq \delta(a,b) + \delta(b,c)$

La distance de Wagner et Fischer est intéressante parce qu'elle vérifie le théorème suivant.

THÉORÈME **25** *Soit $\mathcal{V}$ un vocabulaire et* dist *une distance (au sens mathématique) sur $\mathcal{V} \cup \{\varepsilon\}$, alors la distance de Wagner et Fischer sur $\mathcal{V}^*$ est une distance au sens mathématique.*

DÉMONSTRATION : Nous devons vérifier que la distance de Wagner et Fischer ne prend que des valeurs positives ou nulles, et qu'elle vérifie les axiomes d'égalité de commutativité et l'inégalité triangulaire.

Valeurs positives ou nulles Comme dist est une distance sur $\mathcal{V} \cup \{\varepsilon\}$, ses valeurs sont toutes positives ou nulles. Selon la première définition de la distance de Wagner et Fischer, ses valeurs sont obtenues uniquement par addition de valeurs de distance sur $\mathcal{V} \cup \{\varepsilon\}$, donc, les valeurs de la distance de Wagner et Fischer sont positives ou nulles.

Axiome d'égalité Supposons qu'il existe $u, v \in \mathcal{V}^*$ tels que $u \neq v \wedge \delta(u,v) = 0$, alors il existerait une dérivation de u vers v de coût zéro. Il existerait donc une opération d'édition$a \rightarrow b$ $(a \neq b)$ de coût zéro. Ceci est impossible car dist est une distance sur $\mathcal{V} \cup \{\varepsilon\}$.

Commutativité Evident par construction de la distance de Wagner et Fischer et par commutativité de dist sur $\mathcal{V} \cup \{\varepsilon\}$.

Inégalité triangulaire Considérons la dérivation $(v_1, v_2, \ldots, v_n)$ de u vers w, telle qu'il existe un k / $1 \leq k \leq n$ pour lequel $v_k = v$. Une telle dérivation est une dérivation de u vers w et aussi une dérivation de u vers v suivie d'une dérivation de v vers w. La distance de Wagner et Fischer étant le minimum sur toutes les dérivations possibles, l'inégalité triangulaire est vérifiée.

$$\delta(u,w) \leq \delta(u,v) + \delta(v,w)$$

CQFD

Dans la suite, nous supposons que $\mathcal{V}$ est fini et a au moins deux éléments.

C.4 Propriétés

LEMME **24** $\forall (a,b,c) \in (\mathcal{V} \cup \{\varepsilon\})^3$,

$$\delta(a.b, a.c) = \delta(b,c)$$

DÉMONSTRATION : Si $a = \varepsilon$, trivial. On suppose $a \neq \varepsilon$. Si $b = c$, l'égalité devient triviale : $\delta(a.b, a.c) = 0 = \delta(b,c)$. On suppose que $b \neq c$. Considérons la dérivation $(u_1 = a.b, u_2 = a.c)$ impliquant l'unique opération d'édition $b \rightarrow c$. Parce que la distance de Wagner et Fischer est le minimum sur toutes les dérivations possibles, on a :

$$\delta(a.b, a.c) \leq \delta(b,c)$$

Réciproquement, supposons que $(u_1 = a.b, u_2, \ldots, u_{n-1}, u_n = a.c)$ est la dérivation pour laquelle $\delta(a.b, a.c)$ est atteinte. Si $b \to c$ est impliquée alors

$$\delta(a.b, a.c) \geq \delta(b, c)$$

sinon il existe deux opérations d'édition $b \to e$ et $f \to c$ avec $e, f \in \mathcal{V} \cup \{\varepsilon\}$ qui rendent compte de la suppression ou du remplacement de b et de l'insertion ou du remplacement de c. Par conséquent,

$$\delta(a.b, a.c) \geq \delta(b, e) + \delta(f, c)$$

Nécessairement, on doit rendre compte de la suppression de e et de l'insertion de f aussi, le même raisonnement s'applique donc encore. Comme une dérivation est finie, il existe en fin de compte une dérivation de b vers c telle que

$$\begin{aligned} \delta(a.b, a.c) \quad \geq \quad & \delta(b, e_1) + \delta(e_1, e_2) + \\ & \cdots + \delta(f_2, f_1) + \delta(f_1, c) \end{aligned}$$

pour laquelle soit $f_n = e_n$ ou $\delta(e_n, f_n)$ est impliquée. L'inégalité trianguaire étant vérifiée pour dist sur $\mathcal{V} \cup \{\varepsilon\}$, on a

$$\begin{aligned} \delta(a.b, a.c) \quad \geq \quad & \delta(b, e_1) + \delta(e_1, e_2) + \\ & \cdots + \delta(f_2, f_1) + \delta(f_1, c) \\ \geq \quad & \delta(b, c) \end{aligned}$$

En combinant les deux inégalités, on a

$$\delta(a.b, a.c) = \delta(b, c)$$

CQFD

LEMME **25** $\forall a \in \mathcal{V}$, $\forall (v, w) \in (\mathcal{V}^*)^2$,

$$\delta(a.v, a.w) = \delta(v, w)$$

DÉMONSTRATION : Soit $(a.u = v_1, \ldots, v_i, \ldots, v_n = a.v)$ la dérivation pour laquelle $\delta(a.u, a.v)$ est atteinte. Considérons le premier v_i pour lequel a n'est pas préfixe. Necéssairement, la dérivation directe précédente était de $a.v_i$ vers v_i. Si on supprime cette dérivation directe, on obteint une dérivation de u vers $a.v$. Par définition de dist comme minimum sur toutes les dérivations, on a :

$$\delta(u, a.v) \leq \delta(a.u, a.v) - \delta(a, \varepsilon)$$

Maintenant, si on ajoute une dérivation directe de $a.v$ vers v à la fin de la dérivation précédemment obtenue, on obtient une dérivation de u vers v. Parce que dist est le minimum sur toutes les dérivations, on a :

$$\delta(u, v) \leq \delta(a.u, a.v) - \delta(a, \varepsilon) + \delta(\varepsilon, a)$$

c'est-à-dire $\delta(u,v) \leq \delta(a.u, a.v)$ Réciproquement, considérons une dérivation $(u = {v'}_1, \ldots, {v'}_n = v)$ pour laquelle $\delta(u,v)$ est atteinte. On peut construire la dérivation $(a.u = {v'}_1, \ldots, a.{v'}_n = a.v)$ de $a.u$ vers $a.v$. Par le lemme de l'unique opération d'édition, cette dérivation fait intervenir les mêmes opérations d'édition. Donc son coût est le mêmem que celui de la dérivation de u vers v, c'est-à-dire $\delta(u,v)$. Par définition de dist comme minimum sur toutes les dérivations, on a

$$\delta(a.u, a.v) \leq \delta(u,v)$$

Les deux inégalités donnent le résultat recherché. CQFD

LEMME **26 (Préfixe)** $\forall (u,v,w) \in (\mathcal{V}^*)^3$,

$$\delta(u.v, u.w) = \delta(v,w)$$

DÉMONSTRATION : Trivial étant donné le lemme précédent. Soit $u = a_1 \ldots a_n$. Alors,

$$\begin{aligned} \delta(v,w) &= \delta(a_n.v, a_n.w) \\ &= \cdots \\ &= \delta(a_1 \ldots a_n.v, a_1 \ldots a_n.w) \end{aligned}$$

De façon évidente, le lemme est aussi vraie si u est un suffixe au lieu d'être un préfixe, *i.e.* $\delta(v.u, w.u) = \delta(v,w)$. CQFD

LEMME **27 (Séparation)** $\forall (u,v,u',v') \in (\mathcal{V}^*)^4$,

$$\delta(u.v, u'.v') \leq \delta(u,u') + \delta(v,v')$$

DÉMONSTRATION : Par application du lemme du préfixe,

$$\begin{aligned} \delta(u,u') + \delta(v,v') &= \delta(u.v, u'.v) \\ &\quad + \delta(u'.v, u'.v') \end{aligned}$$

ce qui donne, par inégalité triangulaire :

$$\delta(u.v, u'.v') \leq \delta(u,u') + \delta(v,v')$$

CQFD

LEMME **28 (Séparation exacte)** $\forall (u,v,w) \in (\mathcal{V}^*)^3, \exists (u',v') \in (\mathcal{V}^*)^2 \ / \ w = u'.v' \wedge$

$$\delta(u.v, w) = \delta(u,u') + \delta(v,v')$$

DÉMONSTRATION : Par application du lemme de la séparation, on a

$$\delta(u.v, u'.v') \leq \delta(u, u') + \delta(v, v')$$

Considérons $(u.v = w_1, \ldots, w_n = w)$ une dérivation pour laquelle $\delta(u.v, w)$ est atteinte.

$$\delta(u, v) = \sum_{i=1}^{i=n-1} \delta(w_i, w_{i+1})$$

La dérivation (w_i, w_{i+1}) fait intervenir une unique opération d'édition $a_i \rightarrow b_i$, par conséquent,

$$w_i = w'_i.a_i.w''_i \wedge w_{i+1} = w'_i.b_i.w''_i \wedge$$

$$\begin{aligned}
&\delta(w_i, w_{i+1}) \\
&\quad = \quad \delta(a_i, b_i) \\
&\quad = \quad \delta(w'_i, w'_i) + \delta(a_i, b_i) + \delta(w''_i, w''_i) \\
&\quad = \quad \delta(w'_i.a_i, w'_i.b_i) + \delta(w''_i, w''_i)
\end{aligned}$$

Si on considère deux dérivations directes successives (w_{i-1}, w_i, w_{i+1}), soit $w'_{i-1}.b_{i-1}$ est un préfixe de $w'_i.a_i$ ou le contraire ou elles sont égales. Dans le premier cas,

$$\begin{aligned}
\delta(w_{i-1}, w_{i+1}) \quad = \quad & \delta(w'_{i-1}.a_{i-1}, w'_{i-1}.b_{i-1}) \\
& + \ \delta(w'''_i.a_i, w'''_i.b_i) \\
& + \ \delta(w''_i, w''_i)
\end{aligned}$$

avec $w_{i-1} = w'_{i-1}.a_{i-1}.w'''_i.a_i.w''_i$ et w_{i+1} obtenue en remplaçant les a par des b. Le second cas est semblable. Le troisième cas impique que $b_{i-1} = a_i$, ce qui signifie que le passage par a_i n'est pas nécessaire. Donc, pour toute la dérivation,

$$\begin{aligned}
&\delta(u.v, w) \quad = \\
&\quad (\sum_{j=1}^{N-1} \delta(w'_j.a_j, w'_j.b_j)) + \delta(w'_N, w'_N)
\end{aligned}$$

avec $a_j \rightarrow b_j$ des opérations d'édition de la dérivation (w_i), pas nécessairement dans le même ordre.

Par construction,

$$u.v = (\prod_{i=1}^{N-1} w'_i.a_i).w'_N$$

et w est obtenue en remplaçant les a par des b, c'est-à-dire en appliquant les opérations d'édition. Maintenant, pour un u donné préfixe de $u.v$, il est clair que nous pouvons construire u' et v' en appliquant les opérations d'édition aux places convenables. CQFD

Bibliographie

David ABERCROMBIE, « Extending the Roman alphabet: Some orthographic experiments of the past four centuries » (Extension de l'alphabet romain : les essais orthographiques des quatre derniers siècles), *in* ASHER & HENDERSON, *Towards a History of Phonetics*, 1981, p. 207–224.

Académie française, *Dictionnaire de l'Académie française*, Paris, Chez la veuve de Jean-Baptiste Coignard, 1694.
`http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos/ACADEMIE-/PREMIERE/premiere.fr.html`, page consultée le 26 septembre 2001.

Académie française, *Dictionnaire de l'Académie française*, Paris, Chez la veuve de Jean-Baptiste Coignard, 1835.
`http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos/ACADEMIE-/SIXIEME/sixieme.fr.html`, page consultée le 26 septembre 2001.

Académie française, *Dictionnaire de l'Académie française*, Paris, Chez J.J. Smits et C^{e}, an VI de la République (1798).
`http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos/ACADEMIE-/CINQUIEME/cinquieme.fr.html`, page consultée le 26 septembre 2001.

Pascal ACOT, *L'histoire des sciences*, coll. Que sais-je ? Paris, PUF, 1999.

Georgiĭ Maximovitch ADEL'SON-VELSKIĬ et Evgeniĭ Mikhaĭlovich LANDIS, « An algorithm for the organization of information » (Algorithme pour l'organisation de l'information), *Soviet Math.*, vol. 146, p. 263–266, 1962.

Hassan AÏT-KACI, *A lattice theoretic approach to computation based on a calculus of partially ordered type structures* (Approche théorique du calcul par les treillis, reposant sur le calcul de structures typées munies d'un ordre partiel), Thèse de doctorat, Université de Pennsylvanie, 1984.

Lloyd ALLISON et Trevor I. DIX, « A bit string longest common subsequence algorithm » (Un algorithme de calcul du plus long sous-mot commun entre chaînes binaires), *Information Processing Letter*, vol. 23, p. 305–310, 1986.

ARISTOTE, *Poétique*, Paris, Gallimard, collection *tel*, 1996, Trad. J. Hardy.

ARISTOTE, *Éthique à Nicomaque*, Paris, Librairie philosophique J. Vrin, 1997 [1er tirage 1990], Trad. J. Tricot.

Ronald E. ASHER et Eugénie J. A. HENDERSON, *Towards a History of Phonetics* (Vers une histoire de la phonétique), Edinburgh, Edinburgh University Press, 1981.

Nicolas AUCLERC et Yves LEPAGE, « Aides à l'analyse pour la construction de banque d'arbres : étude de l'effort », *in* Actes de TALN-2001, p. 53–62, Tours, 2001.
`http://www.slt.atr.co.jp/~lepage/ps/taln01-2.ps.gz`.

Gaston BACHELARD, *La formation de l'esprit scientifique*, Paris, Librairie philosophique J. Vrin, 1996, [1^e éd. 1938].

Bruno BACHIMONT, *Herméneutique matérielle et artéfacture : des machines qui pensent aux machines qui donnent à penser*, thèse d'épistémologie, École polytechnique, 1991.

Jean-Pierre BARTHÉLEMY et Alain GUÉNOCHE, *Les arbres et les représentations des proximités*, Paris, Masson, 1988.

M.-C. BARTHOLY, J.-P. DESPIN et G. GRANDPIERRE, *La science : epistémologie générale*, Paris, Éditions Magnard, 1978.

Kenneth R. BEESLEY, « Consonant Spreading in Arabic Stems » (Diffusion consonantique en arabe), *in* Proceedings of COLING-ACL'98, vol. I, p. 117–123, Montréal, 1998.

Henri BERGSON, *L'évolution créatrice*, Paris, Presses Universitaires de France, collection Quadrige, 1998, [1^e éd. 1941].

Henri BERGSON, *Le rire*, Paris, Presses Universitaires de France, collection Quadrige, 1999, [1^e éd. 1940].

Alan A. BLACK, « Finite State Machines from Feature Grammars » (Machines à nombre fini d'états obtenues à partir de grammaires de traits), *in* International Workshop on Parsing Technologies, Carnegie Mellon University, p. 277–285, Pittsburgh, 1989.

Ezra BLACK, Stephen EUBANK, KASHIOKA Hideki, David MAGERMAN, Roger GARSIDE et Geoffrey LEECH, « Beyond Skeleton Parsing: Producing a Comprehensive Large–Scale General–English Treebank with Full Grammatical Analysis » (Au-delà d'une analyse squelettique : production d'une banque d'arbres de grand échelle de l'anglais tout-venant avec analyse grammaticale complète), *in* Proceedings of COLING 96, p. 107–112, København, 1996.

Leonard BLOOMFIELD, *Language* (Le langage), New York, Holt, 1933.

Christian BOITET, *On the automatic transformation of a set of EBMT Constituent Boundary Patterns into a Context-Free Grammar, and associated bottom-up algorithms* (Sur la transformation d'un ensemble de patrons de marqueurs de constituants de EBMT en grammaire hors-contexte et sur les

algorithmes ascendants associés), Rapport technique ATR-IT-0071, ATR-ITL, 1994.

Jean-Élie BOLTANSKI, *La linguistique diachronique*, coll. Que sais-je ? Paris, PUF, 1995.

Antoine Augustin COURNOT, *Essais sur les fondements de nos connaissances et sur les caractères de la critique philosophique*, Paris, Hachette, 1851.

Christopher CULY, « The Complexity of the Vocabulary of Bambara » (La complexité du vocabulaire bambara), *Linguistics and Philosophy*, vol. 8, p. 345–351, 1985.

Robert I. DAMPER et John E.G. EASTMAN, « Pronouncing Text by Analogy » (Prononciation de textes par analogie), *in* Proceedings of COLING-96, p. 268–269, København, 1996.

Alain de LIBERA, *La philosophie médiévale*, coll. Que sais-je ? Paris, PUF, 1992.

Ferdinand de SAUSSURE, *Cours de linguistique générale*, Lausanne et Paris, Payot, 1995, [1^{e} éd. 1916].

Fathi DEBILI et Elyès SAMMOUDA, « Appariement de phrases de textes bilingues français-anglais et français-arabes », *in* Proceedings of COLING-92, vol. 2, p. 517–524, Nantes, 1992.

Gilles DELEUZE, *Anti-Œdipe et Mille plateaux*, 1974.
`http://www.deleuze.fr.st/TXT/140174.html`, page consultée le 27 novembre 2000, Transcription du cours donné à Vincennes le 14 janvier 1974.

Gilles DELEUZE, *Sur Leibniz*, 1980.
`http://www.deleuze.fr.st/TXT/150480.html`, page consultée le 27 novembre 2000, Transcription du cours donné à Vincennes le 15 avril 1980.

Denis DIDEROT et Jean LE ROND d'ALEMBERT, *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, par une Société de Gens de lettres, entre 1751 et 1774.
`http://encyclopedie.inalf.fr/`, page consultée le 12 octobre 2001.

Aelius DONATIUS, *De partibus orationis – Ars minor* (Les parties du discours – Résumé), 1994.
`http://ccat.sas.upenn.edu/jod/texts/donatus.4.html`, page consultée le 27 novembre 2000, Texte latin.

Françoise DOUAY, « La contre-analogie – Réflexion sur la récusation de certaines analogies pourtant bien formées cognitivement », *texto*, vol. 10, p. 155–170, 1985.
`http://www.revue-texto.net/nouveautes/Contre-Analogie-/Contre-analogie.html`, page consultée le 21 mars 2001.

David R. Dowty, Lauri Karttunen et Arnold M. Zwicky, *Natural language processing – Psychological, computational, and theoretical perspectives* (Traitement automatique des langues – perspectives psychologiques, calculatoires et théoriques), Cambridge, Cambridge University Press, 1985.

S. Duckett (sous la direction de), *Dictionnaire de la conversation et de la lecture – Dictionnaire raisonné des notions générales les plus indispensables à tous par une société de savants et de gens de lettres*, Paris, Librairie Firmin Didot, 1864.
`http://www.chass.utoronto.ca/epc/langueXIX/duckett/`, page consultée le 4 décembre 2001.

Oswald Ducrot et Jean-Marie Schaeffer, *Nouveau dictionnaire encyclopédique des sciences du langage*, Paris, Seuil, 1978.

Alick Elithorn et Ranan Banerji, *Artificial & Human Intelligence* (Intelligences artificielle et humaine), Elsevier Science, 1984.

Euclide, *Les quinze livres des éléments géométriques d'Euclide : plus le livre des donnez... trad. en françois*, Paris, I. Dédin, 1632, Trad. D. Henrion.
`http://gallica.bnf.fr/` ou `http://www-mathdoc.ujf-grenoble.fr-/NUMDAM/euclide.htm`, page consultée le 11 juillet 2001.

Euclide, *Euclidis Elementa* (Les éléments d'Euclide), Leipzig, B.G. Teubner, 1883, edidit et latine interpretatus est I.L. Heiberg, Dr. Phil.

Euclide, ユークリッド原論 (Les éléments d'Euclide), 東京, 共立出版株式会社, 1996, 訳・解説 中村幸四郎・寺坂英孝・伊東俊太郎・池田美恵.

J.-P. Guillaume D.E. Kouloughli G. Bohas, *The Arabic Linguistic Tradition* (La tradition linguistique arabe), London–New York, Routledge, 1990.

Dedre Gentner, « Structure Mapping: A Theoretical Model for Analogy » (Applications de structures : un modèle théorique de l'analogie), *Cognitive Science*, vol. 7, n° 2, p. 155–170, 1983.

Paul Ghils, « Langage et contradiction », *Bulletin interactif du Centre International de Rechreches et Études Transdisciplinaires (CIRET)*, vol. 13, p. 18, 1998.
`http://perso.club-internet.fr/nicol/ciret/bulletin/b13-/b13c18.htm`, page consultée le 11 octobre 2001.

Maurice Gross, « On the Failure of Generative Grammar » (L'échec de la grammaire générative), *Language*, vol. 55, n° 4, p. 859–885, 1979.

Rogers P. Hall, « Computational Approaches to Analogical Reasoning: A Comparative Analysis » (Approches calculatoires du raisonnement logique : analyse comparative), *Artificial Intelligence*, vol. 39, n° 1, p. 39–120, 1989.

Nabil HATHOUT, « Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes », *in* Actes de TALN-2001, p. 223–232, Tours, 2001.

D.S. HIRSCHBERG, « Algorithms for the longest common subsequence problem » (Algorithmes de résolution du problème de la plus longue sous-chaîne commune), *Journal of the ACM*, vol. 24, n° 4, p. 664–675, 1977.

Robert R. HOFFMAN, « Monster Analogies » (Étranges analogies), *AI Magazine*, vol. 11, p. 11–35, 1995.

Douglas HOFSTADTER et the Fluid Analogies Research Group, *Fluid Concepts and Creative Analogies* (Concepts fluides et analogies créatrices), New York, Basic Books, 1994.

舟田 京子 (HUNADA Kyouko), やさしい初歩のインドネシア語 (Premiers pas faciles en indonésien), 東京, 南霊堂, 1997.

井上 史雄 (INOUE Humio), « 言語の構造の変遷・東北方言音韻史を例として » (Évolution de la structure des langues : exemple de l'histoire des phonèmes du dialecte du Touhoku), *in* 柴田 武編 (SIBATA Takesi), 言語の構造, 1980, p. 159–175.

Institut des Langues de Pékin, *Manuel de chinois pratique*, vol. I, Pékin, La Presse Commerciale, 1995.

Institut National de la Langue Française, *Trésor de la langue française informatisé*, INaLF, C.N.R.S., 2000.
`http://zeus.inalf.fr/tlf.htm`, page consultée le 12 octobre 2001.

Esa ITKONEN, « Iconicity, analogy, and universal grammar » (Iconicité, analogie et grammaire universelle), *Journal of Pragmatics*, vol. 22, p. 37–53, 1994.

Esa ITKONEN et Jussi HAUKIOJA, « A rehabilitation of analogy in syntax (and elsewhere) » (Une réhabilitation de l'analogie en syntaxe (et ailleurs)), *in* KERTÉSZ, *Metalinguistik im Wandel: die kognitive Wende in Wissenschaftstheorie und Linguistik*, 1997, p. 131–177.

Grzegorz JAGODZIŃSKI, *Gramatyka języka polskiego* (Grammaire de la langue polonaise), 2001.
`http://www.republika.pl/grzegorj/isopl/gram1.html`, page consultée le 30 novembre 2001, Page internet.

Roman JAKOBSON, *Essais de linguistique générale*, Paris, Éditions de Minuit, 1963.

Roman JAKOBSON, *Question de poétique*, Paris, Seuil, 1973.

Aravind K. JOSHI, « Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural description? » (Grammaires d'arbres adjoints : quelle quantité de sous-contexte est nécessaire à l'obtention de descriptions structurales raisonnables ?), *in* DOWTY *et al.*, *Natural language processing – Psychological, computational, and theoretical perspectives*, 1985, p. 206–250.

Aravind K. JOSHI, K. VIJAY-SHANKER et David WEIR, « The Convergence of Mildly Context-Sensitive Grammar Formalisms » (Convergence des formalismes grammaticaux légèrement sous-contexte), *in* SELLS *et al.*, *Foundational issues in natural language processing*, 1991, p. 31–81.

Martin KAY, « Functional unification grammar: a formalism for machine translation » (Les grammaires fonctionnelles par unification : un formalisme pour la traduction automatique), *in* Proceedings of COLING-84, p. 75–78, Stanford, 1984.

András KERTÉSZ, *Metalinguistik im Wandel: die kognitive Wende in Wissenschaftstheorie und Linguistik* (Métalinguistique en marche : le tournant cognitif en théorie de la connaissance et en linguistique), Frankfurt a/M, Peter Lang, 1997.

Djamel KOULOUGHLI, *Grammaire de l'arabe d'aujourd'hui*, Paris, Pocket, collection Langues pour tous, 1994.

Joseph B. KRUSKAL et Mark LIBERMAN, « The symmetric time warping problem: from continuous to discrete » (Le problème symétrique de la distortion en temps : du continu au discret), *in* SANKOFF & KRUSKAL, *Time warps, string edits, and macromolecules*, 1999, p. 125–161.

九鬼 周造 (KUKI Syuuzou), 「いき」の構造 (La structure de l'*iki*), 東京, 岩波文庫, 1999, [1^e éd. 1930].

Jerzy KURYŁOWICZ, « La nature des procès dits « analogiques » », *Acta Linguistica*, vol. 5, p. 15–37, 1949.

Jerzy KURYŁOWICZ, *Esquisses linguistiques*, Wrocław–Warszawa–Kraków, Ossolineum, 1961.

Jerzy KURYŁOWICZ, *L'apophonie en sémitique*, Wrocław–Warszawa–Kraków, Ossolineum, 1961.

Pierre LAROUSSE, *Grand dictionnaire universel*, Paris, Librairie Larousse, 1865 à 1876.

Pierre LE GOFFIC, *Les formes conjugées du verbe français – oral et écrit*, Paris, Ophrys, 1997.

Wilhelm Gottfried LEIBNIZ, *Nova methodus pro maximis et minimis* (Nouvelle méthode pour le calcul des maximums et minimums), 1646–1716. `http://pro.wanadoo.fr/1011/LEIB.HTM`, page consultée le 4 décembre 2000, Page internet.

Yves LEPAGE, « Non-directionality and Self-assessment in an Example-based System Using Genetic Algorithms » (Non-directionalité et auto-évaluation d'un système par l'exemple utilisant des algorithmes génétiques), *in* Proceedings of COLING-94, vol. I, p. 616–621, Kyoto, 1994. `http://www.slt.atr.co.jp/~lepage/ps/coling94.ps.gz`.

Yves LEPAGE, « Regular languages have regular neighbourhoods for the Wagner and Fischer distance » (Pour la distance de Wagner et Fischer, les voisinages des langages réguliers sont réguliers), *in* Proceedings of the International Conference on Linguistic Applications, p. 13–19, Pulau Pinang, 1994. `http://www.slt.atr.co.jp/~lepage/pdf/icla94.fr.pdf.gz` (version française).

Yves LEPAGE, *Tesnière's structural syntax: notations for tree-banking using* `BoardEdit` (Syntax structurale de Tesnière: notations pour la banque d'arbres à l'aide de `BoardEdit`), Rapport technique ATR-IT-0176, ATR-ITL, 1996.

Yves LEPAGE, « Un éditeur pour la construction de banques d'arbres », *in* Actes de TALN-96, p. 104–111, Marseille, 1996. `http://www.slt.atr.co.jp/~lepage/ps/taln96.ps.gz`.

Yves LEPAGE, « Corpus Contraction by Sentence Extraction Using Analogy » (Contraction de textes par extraction de phrases analogiques), *in* Proceedings of NLPRS-97, p. 457–462, Phuket, 1997. `http://www.slt.atr.co.jp/~lepage/ps/nlprs97.ps.gz`.

Yves LEPAGE, « String Approximate Pattern-Matching » (Filtrage approximatif de chaînes), *in* 情報処理学会第 55 回全国大会, vol. 3, p. 139–140, 福岡工業大学, 1997. `http://www.slt.atr.co.jp/~lepage/ps/ipsj97.fr.ps.gz` (version française).

Yves LEPAGE, « Solving Analogies on Words: an Algorithm » (Un algorithme pour la résolution d'analogies entre mots), *in* Proceedings of COLING-ACL'98, vol. I, p. 728–735, Montréal, 1998. `http://www.slt.atr.co.jp/~lepage/pdf/coling98.fr.pdf.gz` (version française).

Yves LEPAGE, « 用例機会翻訳装置 » (Appareil de traduction automatique par l'exemple), 日本特許庁 (office japonais des brevets), n° du brevet: 特許第 2948151 号, 1999.

Yves LEPAGE, « Analogy + Tables = Conjugation » (Analogie + Tables = Conjugaison), *in* G. FRIEDL & H. MAYR (sous la direction de), Proceedings of NLDB'99, p. 197–201, Klagenfurt, 1999.
`http://www.slt.atr.co.jp/~lepage/pdf/nldb99.fr.pdf.gz` (version française).

Yves LEPAGE, « Apparatus and method for producing analogically similar word based on pseudo-distances between words » (Appareil et méthode de production de mots analogiquement similaires fondée sur le calcul de pseudo-distances entre mots), Office européen des brevets, n° de demande du brevet : 99115561.5, 1999.

Yves LEPAGE, « Open Set Experiments with Direct Analysis by Analogy » (Expériences d'analyse directe par analogie), *in* Proceedings of NLPRS-99, p. 363–368, Beijing, 1999.
`http://www.slt.atr.co.jp/~lepage/pdf/nlprs99.fr.pdf.gz` (version française).

Yves LEPAGE, *De l'analogie*, 2000, notes rédigées de l'exposé donné devant les membres du groupe Kōin (光陰).

Yves LEPAGE, « Analogy and formal languages » (analogie et langages formels), *in* Proceedings of FG/MOL 2001, p. 373–378, Helsinki, 2001.
`http://www.elsevier.nl/locate/entcs/volume47.html` et
`http://www.slt.atr.co.jp/~lepage/pdf/fgmol01.pdf.gz`.

Yves LEPAGE, « Formalisation de l'analogie entre chaînes de symboles », *in* Actes de CAp-2001, plateforme AFIA-2001, p. 117–131, Grenoble, 2001.
`http://www.slt.atr.co.jp/~lepage/pdf/afia01.pdf.gz`.

Yves LEPAGE et Shinichi ANDO, « Similarity search apparatus for searching unit string based on similarity » (Appareil de recherche par similarité pour la recherche de chaînes composées d'unités fondée sur la similarité), Brevet américain, n° du brevet : US 6,009,424, 1999.

Yves LEPAGE et Shin-Ichi ANDO, « Saussurian analogy: a theoretical account and its application » (Analogie saussurienne : un compte-rendu théorique et son application), *in* Proceedings of COLING-96, p. 717–722, København, 1996.
`http://www.slt.atr.co.jp/~lepage/ps/coling96.ps.gz`.

Yves LEPAGE et 安藤　真一 (ANDOU Sin-Iti), « 類似検索装置 » (Appareil de recherche approximative), 日本特許庁 (office japonais des brevets), n° du brevet : 特許第 2747443 号, 1998.

Yves LEPAGE et 安藤　真一 (ANDOU Sin-Iti), « 用例主導型言語構造解析装置 » (Appareil d'analyse structurale de la langue guidée par des exemples), 日本特許庁 (office japonais des brevets), n° du brevet : 特許第 3135221 号, 2000.

Yves LEPAGE, ANDO Shin-ichi et IIDA Hitoshi, « The snow–ball effect of analogy » (Effet boule de neige de l'analogie), *in* 言語処理学会第 3 回年次大会発表論文集, p. 289–292, 京都大学, 1997 年.
`http://www.slt.atr.co.jp/~lepage/pdf/nlpj98.pdf.gz`.

Yves LEPAGE, ANDO Shin-Ichi, AKAMINE Susumu et IIDA Hitoshi, « An annotated corpus in Japanese using Tesnière's structural syntax » (Un corpus japonais annoté avec la syntaxe structurale de Tesnière), *in* S. KAHANE & A. POLGUÈRE (sous la direction de), ACL-COLING Workshop on Processing of Dependency-Based Grammars, p. 109–115, Montréal, 1998.
`http://www.slt.atr.co.jp/~lepage/pdf/pdgb98.pdf.gz`.

Yves LEPAGE et 白井諭 (SIRAI Satosi), « 言語学的類推による生成文における非文法生の分析 » (Analyse de l'agrammaticalité des phrases produites par analogie linguistique), *in* 言語処理学会第 6 回年次大会発表論文集, p. 219–222, 北陸先端大学, 2000 年.
`http://www.slt.atr.co.jp/~lepage/ps/nlpj00.ps.gz`.

Yves LEPAGE et 白井諭 (SIRAI Satosi), « 言語学的比例類推空間相似 » (Homomorphismes d'espaces munis de l'analogie linguistique), *in* 言語処理学会第 7 回年次大会発表論文集, p. 90–92, 東京大学, 2001 年.
`http://www.slt.atr.co.jp/~lepage/pdf/nlpj01-1.pdf.gz`.

V.I. LEVENSHTEIN, « Binary codes capable of correcting deletions, insertions and reversals » (Codes binaires susceptibles de correction par suppressions, insertions et remplacements), *Soviet Physics-doklady*, vol. 10, n° 8, p. 707–710, 1966.

Witold MAŃCZAK, « Tendances générales des changements analogiques », *Lingua*, vol. VII, p. 298–325 et 387–420, 1958.

Witold MAŃCZAK, *Z zagadnień językoznawstwa ogólnego* (Problèmes de linguistique générale), Wrocław – Warszawa – Kraków, Ossolineum, 1970.

Benoît MANDELBROT, « Les constantes chiffrées du discours », *in* MARTINET, *Encyclopédie de la Pléïade : linguistique*, 1968, p. 46–56.

Yannick MARCHAND et Robert I. DAMPER, « A Multistrategy Approach to Improving Pronounciation by Analogy » (Une approche multistratégique pour l'amélioration de la prononciation par analogie), *Computational Linguistics*, vol. 26, n° 2, p. 195–219, 2000.

Solomon MARCUS, « Contextual Grammars » (Les grammaires contextuelles), *Revue roumaine de mathématiques théoriques et appliquées*, vol. 14, p. 1525–1534, 1969.

Solomon MARCUS, Carlos MARTÍN-VIDE et Gheorghe PĂUN, *Contextual Grammars versus Natural Languages* (Grammaires contextuelles et langue), Technical Report 44, Turku center for Computer Science, 1996.

J. MARTINET (sous la direction de), *Encyclopédie de la Pléïade : linguistique*, Paris, Gallimard, 1968.

Carlos MARTÍN-VIDE, *Mathematical and computational analysis of natural language* (Analyse mathématique et calculatoire de la langue), Amsterdam / Philadelphia, John Benjamins Publishing Co., 1998.

Igor MEL'ČUK, Nadia ARBATCHEWSKI-JUMARIE, Léo ELNITSKY, Ldija IORDANSKAJA et Adéle LESSARD, *Dictionnaire explicatif et combinatoire du français contemporain. Recherche lexico-sémantiques I*, Montréal, Presses de l'université de Montréal, 1984.

Igor A. MEL'ČUK, *Vers une linguistique Sens-Texte. Leçon inaugurale* (Analyse mathématique et calculatoire de la langue), Paris, Collège de France, Chaire internationale, 10 janvier 1997.

Igor A. MEL'ČUK, *Dependency Syntax: Theory and Practice* (Syntaxe en dépendance : théorie et pratique), New York, State University of New York Press, 1988.

Jens MICHAELIS et Marcus KRACHT, *Logical Aspects of Computational Linguistics* (Aspects logiques de la linguistique computationnelle), Number 1328 in LNCS/LNAI. Berlin, Springer Verlag, 1997.

Jean-Claude MILNER, *Introduction à une science du langage*, Paris, Seuil / Des Travaux, 1989.

Roger K. MOORE, « A Dynamic Programming Algorithm for the Distance Between Two Finite Areas » (Algorithme de programmation dynamique pour le calcul de la distance entre deux aires finies), *IEEE Transactions on pattern analysis and machine intelligence*, vol. PAMI-1, n° 1, p. 86–88, 1979.

Martine MORENON, *Roman Jakobson*, 1997.
`http://pro.wanadoo.fr/martine.morenon/1jakobso.htm`, page consultée le 6 décembre 2000, Page internet.

Georges MOUNIN, *Histoire de la linguistique – Des origines au XX^e siècle*, Paris, Quadrige / Presses Universitaires de France, 1967.

Georges MOUNIN, *Clefs pour la linguistique*, Paris, Bibliothèques 10/18, Seghers, 1968.

Georges MOUNIN, *Dictionnaire de la linguistique*, Paris, Quadrige / Presses Universitaires de France, 1974.

長尾 誠 (NAGAO Makoto), « A Framework of a Mechanical Translation between Japanese and English by Analogy Principle » (Un cadre pour la traduction automatique entre le japonais et l'anglais selon un principe d'analogie), *in* ELITHORN & BANERJI, *Artificial & human intelligence*, 1984, p. 173–180.

小川 芳男 (Ogawa Yosio), 林 大 (Hayasi Takesi) et 他編集 (et al.), 日本語教育辞典 (Dictionnaire de l'enseignement du japonais), 東京, 大修館書店, 1988.

Attribué à Olympiodore, *Prolégomènes à la philosophie de Platon*, Paris, Coll. Belles-lettres, 1990, Texte établi par L.G. Westerink, trad. J. Trouillard.

Blaise Pascal, *Œuvres complètes*, L'intégrale. Paris, Seuil, 1963.

Hermann Paul, *Prinzipien der Sprachgeschichte* (Principes de l'histoire des langues), Tübingen, Niemayer, 1920, 5^e éd., [1^e éd. 1880].
`http://www.gutenberg.aol.de/paulh/prinzip/paulvorr.htm`, page consultée le 21 mars 2001.

Vito Pirelli et Stefano Federici, « "Derivational" paradigms in morphonology » (Paradigmes « dédivationnels » en morphophonologie), *in* Proceedings of COLING-94, vol. I, p. 234–240, Kyōto, 1994.

Geoffrey K. Pullum, « Generative grammar » (Les grammaires génératives), *in* The MIT Encyclopedy of Cognitive Sciences, p. 340–343, Cambridge.
`http://ling.ucsc.edu/~pullum/locker/mitecs_gengram.html`, page consultée le 31 août 2001.

Geoffrey K. Pullum, « Model-Theoretical Syntax » (Théories des modèles en syntaxe), *in* Proceedings of FG/MOL 2001, p. ??–??, Helsinki, 2001.
`http://www.elsevier.nl/locate/entcs/volume47.html`.

Quintilien, *L'institution oratoire*, Université catholique de Louvain, Bibliotheca Classica Selecta, 2000, traduction française de *De institutione oratoria*.
`http://bcs.fltr.ucl.ac.be/Quint/quint1,6.html`, page consultée le 4 décembre 2001.

Victor Sadler et Ronald Vendelmans, « Pilot implementation of a bilingual knowledge bank » (Implémentation pilote d'une banque de connaissances bilingue), *in* COLING-90, vol. 3, p. 449–451, Helsinki, 1990.

Francisco José Zamora Salamanca, « La tradición histórica de la analogía lingüística » (La tradition historique de l'analogie linguistique), *Revista Española de Lingüística*, vol. 14, n° 2, p. 367–419, 1984.

Morris Salkoff, *Une grammaire en chaîne du français*, Paris, Dunod, 1973.

San Antonio, *Si, Signore !*, Paris, Fleuve Noir, 1974.

David Sankoff et Joseph Kruskal, *Time warps, string edits, and macromolecules* (Distortion en temps, édition de chaînes et macromolécules), Lausanne et Paris, The David Hume Series, CSLI Publications, 1999.

佐藤 理史 (SATOU Satosi), *Example-based Machine Translation* (Traduction automatique par l'exemple), Thèse de doctorat, Université de Kyōto, 1991.

Stanley M. SELKOW, « The Tree-to-Tree Editing Problem » (Le problème de l'édition d'un arbre en un autre), *Information Processing Letter*, vol. 6, n° 6, p. 184–186, 1977.

P. SELLS, S. SHIEBER & T. WASOW (sous la direction de), *Foundational Issues in natural language processing* (Problèmes fondamentaux en traitement automatique des langues), Cambridge, MIT Press, 1991.

Stuart M. SHIEBER, « Evidence against the Context-Freeness of Natural Language » (Preuve contre le caractère hors-contexte de la langue), *Linguistics and Philosophy*, vol. 8, p. 333–343, 1985.

柴田 武編 (SIBATA Takesi), 言語の構造 (La structure des langues), 東京, 大修館書店, 1980.

Royal SKOUSEN, *Analogical modeling of language* (Modélisation analogique de la langue), Dordrecht, Kluwer, 1989.

鈴木 信太郎 (監修) (sous la dir. de SUZUKI Sintarou), *Dictionnaire standard japonais-français* – スタンダード和仏辞典, 東京, 大修館, 1991, [1^{e} éd. 1970].

Edward STANKIEWICZ, *Baudouin de Courtenay i podstawy współczesnego językoznawstwa* (Baudouin de Courtenay et les bases de la linguistique contemporaine), Wrocław, Ossolineum, 1986.

Luc STEELS, « Origin of Syntax in Visually Grounded Robotic Agents » (Apparition de la syntaxe chez des robots équipés de vision), *Information and Control*, vol. 64, p. 100–118, 1985.

Luc STEELS, « Language Learning and Language Contact » (Apprentissage des langues et contact entre langues), *in* Workshop Notes of the ECML/MLnet Familiarization Workshop on Empirical Learning of Natural Language Processing Tasks, p. 11–24, Prague, 1997.
`http://www.csl.sony.fr/General/Publications-/ByTopic.php3?topic=language-dynamics`.

Eric STEINHART, « Truth Conditions for Metaphors » (Conditions de vérité par analogie pour les métaphores), *Metaphor and Symbolic Activity*, vol. 9, n° 3, p. 161–178, 1994.

Eddy TAILLEFER et Yves LEPAGE, *A series of experiments with recursive analysis by analogy* (Jeu d'expériences d'analyses récursives par analogie), Rapport technique ATR-S-0007, ATR-SLT, 2000.

Lucien TESNIÈRE, *Eléments de syntaxe structurale*, Paris, Klincksieck, 1959.

Esko UKKONEN, « Algorithms for Approximate String Matching » (Algorithmes de filtrage tolérant), *Information and Control*, vol. 64, p. 100–118, 1985.

Marcus Terentius VARRO, *De lingua latina* (La langue latine), Paris, Coll. Belles-lettres, 1954, Trad. J. COLLART.

Philippe VERHAEGEN, « Image, diagramme, métaphore », *Recherches en communication*, , n° 1, p. 19–47, 1994.

Karl VERNER, « Eine Ausnahme der ersten Lautverschiebung » (Une exception à la première mutation consonantique), *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Indogermanischen Sprachen*, vol. 23, n° 2, p. 97–130, 1875.
`http://www.utexas.edu/cola/depts/lrc/iedocctr/ie-docs/lehmann-/reader/Chapter11.html`, page consultée le 27 août 2001.

Robert A. WAGNER et Michael J. FISCHER, « The String-to-String Correction Problem » (Le problème de la correction d'une chaîne en une autre), *Journal for the Association of Computing Machinery*, vol. 21, n° 1, p. 168–173, 1974.

WIEM, *Wielka Encyklopedia Internetowa Multimedialna*, Kraków, Fogra, 1996–2001, édition électronique de *Popularna Encyklopedia Powszechna* (Encyclopédie populaire universelle).
`http://wiem.onet.pl/`, page consultée le 17 octobre 2001.

Sun WU et Udi MANBER, « Fast Text Searching Allowing Errors » (Recherche rapide de texte tolérante), *Communications of the ACM*, vol. 35, n° 10, p. 83–91, 1992.

Ю А Шрейдер (You. A. SCHREIDER), Равенство, сходство, порядок (Égalité, similarité, ressemblance), Издательство « Наука », 1975, Trad. angl. Martin GREENLINGER.

François YVON, « Paradigmatic Cascades: a Linguistically Sound Model of Pronunciation by Analogy » (Cascades de paradigmes : un modèle linguistiquement justifié de prononciation par analogie), *in* Proceedings of ACL-EACL-97, p. 428–435, Madrid, 1994.

Jean Marie ZEMB, *Vergleichende Grammatik Französisch-Deutsch – Comparaison de deux systèmes – Teil 1*, Dudenverlag, 1978.

Jean Marie ZEMB, *Vergleichende Grammatik Französisch-Deutsch – Comparaison de deux systèmes – Teil 2*, Dudenverlag, 1984.

George Kingsley ZIPF, *Human Behavior and the Principle of Least Effort* (Comportement humain et principe de moindre effort), Wesley, 1949.

Index

Index des noms cités

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

V

Z

Index des langues citées

A

B

C

D

E

F

G

H

I

J

L

M

N

P

S

T

Z

Index des notions

A

C

D

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V