DÉFENSE ET ILLUSTRATION DE L'ANALOGIE

Yves LEPAGE
ATR – Laboratoires télécommunication langue parlée
Hikaridai 2-2-2, Seika-tyō, Sōraku-gun, 619-0288 Kyōto, Japon
yves.lepage@slt.atr.co.jp

Résumé – Abstract

L'argumentation générativiste contre l'analogie tenait en trois points: l'hypothèse de l'inné, celle du hors-contexte et la surproduction. Des résultats théoriques et expérimentaux reposant sur une formulation calculatoire nouvelle de l'analogie contribuent de façon constructive à la réfutation de ces points.

The generativists had three arguments against analogy: the innate hypothesis and the context-free hypothesis, and overgeneration. Theoretical and experimental results, based on a recently obtained computational expression of analogy, contribute to the refutation of these three points in a constructive way.

Mots-clés

Formalisation de l'analogie, langages de mots analogiques, mise en correspondance d'espaces analogiques.

Formalisation of analogy, languages of analogical strings, mapping of analogical spaces.

1 INTRODUCTION

Une analogie est une égalité de rapports énoncée par « A est à B ce que C est à D » et notée par $A : B = C : D$ [1]. Au seul niveau des chaînes de caractères, des exemples d'analogies sont : *elle est infirmière* : *tu es infirmière* = *elle est professeur* : *tu es professeur* ou : *venir* : *il vient* = *prévenir* : *il prévient*. Une équation analogique consiste en la donnée des trois premiers termes seulement : *venir* : *il vient* = *devenir* : `x`.

La tradition grammaticale utilise de façon implicite l'analogie. Les grands hommes qui ont présidé à la naissance de la science linguistique (Paul 20, chap. V et XII), Baudouin de Courtenay (Stankiewicz 86, p. 26–27), (Saussure 16, partie III, chap. 4) la mentionnent tous explicitement, et s'accordent tous à en reconnaître les limites. Seul le courant

[1] La phrase « l'atome est un système solaire » n'est donc pas une analogie ; c'est une métaphore. En revanche, la prase « les électrons sont au noyau atomique ce que les planètes sont au soleil » est bien une analogie.

générativiste l'a récusé formellement (voir (Itkonen 94)) en trois points que nous examinons maintenant l'un après l'autre.

2 L'HYPOTHÈSE DE L'INNÉ

Le premier argument contre l'analogie veut qu'aucune procédure de découverte ne prenne place dans l'acquisition des langues qui serait une faculté innée : c'est parce que les enfants seraient soumis à une quantité trop faible de données, qu'il ne saurait y avoir de procédure de découverte.

Puisque notre propos est une défense constructive de l'analogie, nous nous demandons s'il est possible, à l'aide de la seule analogie et d'un nombre réduit de données, de produire des phrases nouvelles, et ce en quelle quantité. Cette question achoppait jusqu'à présent sur le fait qu'il n'existait pas d'algorithme de résolution d'analogies, c'est-à-dire de procédure de calcul capable de produire, quand c'est possible, un quatrième terme à partir de trois termes données. Après examen de nombreux exemples et axiomatisation des observations, nous avons proposé un algorithme qui repose essentiellement sur les conditions suivantes caractérisant les analogies entre chaînes de symboles (Lepage 00).

THÉORÈME 1 *Soit $\mathcal{V}$ un alphabet. Soit* dist *la distance d'édition canonique (insertion et suppression seulement).* $\forall (A,B,C,D) \in (\mathcal{V}^*)^4$,

$$A : B = C : D \quad \Rightarrow \quad \begin{cases} \delta(A,B) = \delta(C,D) \\ \delta(A,C) = \delta(B,D) \\ |A| + |D| = |B| + |C| \\ \gamma(A,B,C,D) = \frac{1}{2} \times (|B| + |C| - \delta(A,B) - \delta(A,C)) \end{cases}$$

où $\gamma(A,B,C,D)$ est le nombre de symboles en commun, dans le même ordre, à la fois à A, B, C et D.

Comme expérience, nous avons extrait d'une banque de phrases japonaises les 153 phrases contenant le même symbole « 持 » /mot-/ (avoir, tenir) et résolu toutes les équations analogiques possibles. Nous avons obtenu 1 248 phrases différentes, ce qui représente 8 fois le nombre de phrases de base. Le nombre de phrases naturelles produites s'élève à 453. Ici donc, pour un nombre de phrases de base relativement petit, le nombre de phrases correctes produites par application aveugle de l'analogie est triple du nombre de phrases de base. L'affirmation selon laquelle les données seraient trop faibles pour que l'analogie puisse jouer un rôle quelconque apparaît donc douteuse.

3 L'HYPOTHÈSE DU HORS-CONTEXTE

Le second argument contre l'analogie prolonge le premier : c'est qu'il existerait un modèle universel des langues sous-jacent qui ne serait que paramétré lors de l'apprentissage d'une langue donnée. Les tenants du générativisme ont mis en avant les langages formels hors-contexte comme modèle sous-jacent de syntaxe universelle. Dans la classification de Chomsky, ils se situent entre le régulier et le sous-contexte.

dénomination	réguliers $\subset$	hors-contexte $\subset$	sous-contexte
types de règles	$\begin{cases} A \to Ba \\ A \to a \end{cases}$	$\begin{cases} A \to BC \\ A \to a \end{cases}$	$w \to w' \ / \ \lvert w\rvert \leq \lvert w'\rvert$
exemples classiques	$\{a^n\}$	$\{a^n b^n\}$	$\{a^n b^n c^n\}$ $\{a_1^n a_2^n \ldots a_m^n\}$ $\{a^n b^m c^n d^m\}$

Cette classification repose sur la complexité croissante des chaînes apparaissant dans les dérivations. Or, l'existence des symboles non-terminaux est un présupposé très fort. Cette classification n'est pas non plus très intuitive, car les trois exemples classiques pour chacune des trois grandes familles sont à première vue semblables. Leur séparation dans trois familles différentes va à l'encontre de l'intuition.

On sait que l'hypothèse du hors-contexte a été détruite par la découverte de deux structures similaires, l'une dans la morphologie du bambara (Culy 85), l'autre dans la syntaxe de la variante zurichoise du suisse-allemand (Shieber 85). Dans ces deux cas, le cœur de la démonstration repose sur l'apparition du schéma sous-contexte $a^n b^m c^n d^m$.

$$\mathcal{L} \cap a^* b^* c^* d^* = \{ a^n b^m c^n d^m \} \quad \text{est sous-contexte} \ \Rightarrow \ \mathcal{L} \text{ est au moins sous-contexte.}$$

Afin de rendre compte de ces faits, certains ont pensé que la famille des langages formels permettant la formalisation des langues devait être élargie vers le sous-contexte, mais pas trop. (Joshi & al. 85) a proposé d'appeler cette famille la « famille des langages légèrement sous-contexte », et il la caractérise par quatre propositions. Parmi celles-ci, (Marcus & al. 96) soutiennent que le point clé serait, avec l'analyse en temps polynomial, la propriété de croissance constante des longueurs :

Définition 1 (Croissance constante des longueurs) *Un langage $\mathcal{L}$ de cardinal infini vérifie la propriété de croissance constante des longueurs si (et seulement si)*

$$\exists k \in \mathbb{N} \quad / \quad \forall w \in \mathcal{L}, \ \ (|w| < k \quad \vee \quad \exists w' \in \mathcal{L} \ / \ 0 < |w| - |w'| \leq k)$$

Notre propos est de montrer que l'analogie répond de façon constructive aux problèmes posés par la classification de Chomsky. À partir de la notion de dérivation analogique,

Définition 2 (Dérivation analogique) *Soit $\mathcal{V}$ un alphabet. Soit $\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$ dont les éléments (v, v') sont notés $v \to v'$. La dérivation analogique modulo $\mathcal{M}$, notée $\vdash_{\mathcal{M}}$, est définie de la façon suivante.*

$$\forall (w, w') \in \mathcal{V}^* \times \mathcal{V}^*, \quad w \vdash_{\mathcal{M}} w' \quad \Leftrightarrow \quad \exists v \to v' \in \mathcal{M} \ / \quad v : v' = w : \mathtt{x} \ \Rightarrow \ \mathtt{x} = w'$$

on peut définir, de façon standard, la famille des « langages de mots analogiques » :

Définition 3 (Langages de mots analogiques) *Soit $\mathcal{V}$ un alphabet. Soit $\mathcal{A} \subset \mathcal{V}^*$ et $\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$, tous deux finis. Soit $\vdash_{\mathcal{M}}^{+}$ la fermeture transitive de la dérivation analogique $\vdash_{\mathcal{M}}$, alors, $\Lambda(\mathcal{A}, \mathcal{M})$ est le langage de mots analogiques défini de la façon suivante*

$$\Lambda(\mathcal{A}, \mathcal{M}) = \mathcal{A} \cup \{ \ w \in \mathcal{V}^* \ / \ \exists w' \in \mathcal{A}, \ w' \vdash_{\mathcal{M}}^{+} w \ \}$$

Comme d'habitude, on utilise l'induction structurale pour engendrer tous les éléments d'un langage de mots analogiques: en partant des éléments de $\mathcal{A}$, on applique toutes les analogies possibles avec les éléments de $\mathcal{M}$ comme modèles. Dans l'expérience de la section 2, $\mathcal{A}$ était l'ensemble des 153 phrases de base, et $\mathcal{M}$ était $\mathcal{A} \times \mathcal{A}$. Mais nous n'avions appliqué les dérivations qu'une seule fois. Si nous avions appliqué récursivement et indéfiniment les dérivations, afin d'engendrer tout le langage $\Lambda(\mathcal{A}, \mathcal{M})$, le nombre de phrases produites aurait été certainement bien supérieur, voire infini.

On remarquera que la définition des langages de mots analogiques ne fait pas usage de non-terminaux. La grammaticalité est testée simplement, *in fine*, par comparaison avec les chaînes de $\mathcal{A}$, dites chaînes attestées, aprés réduction par analogie selon les modèles de $\mathcal{M}$.

On montre aisément, par induction et par utilisation d'une hypothèse sur la concaténation d'analogies (Lepage 00), que les trois exemples classiques illustrant les trois grandes classes de la classification de Chomsky sont, exprimés sous forme de langages de mots analogiques, très simples (une seule chaîne attestée, un seul modèle) et tout à fait semblables.

THÉORÈME 2

$$\begin{aligned} \{a^n \ / n \geq 1\} &= \Lambda(\{a\}, \{a \rightarrow aa\}) \\ \{a^n b^n \ / n \geq 1\} &= \Lambda(\{ab\}, \{ab \rightarrow aabb\}) \\ \{a^n b^n c^n \ / n \geq 1\} &= \Lambda(\{abc\}, \{abc \rightarrow aabbcc\}) \end{aligned}$$

Une démonstration identique permet d'obtenir que le langage qui sert de base aux contre-exemples contre l'hypothèse du hors-contexte est aussi un langages de mots analogiques.

THÉORÈME 3 $\{a^m b^n c^m d^n \ / \ n, m \geq 1\} = \Lambda(\{abcd\}, \{abcd \rightarrow abbcdd, abcd \rightarrow aabccd\})$

Enfin, le lien avec le légèrement sous-contexte est donné par le résultat remarquable suivant. Rappelons que la croissance constante des longueurs est l'une des quatre propositions plus ou moins informelles caractérisant le légèrement sous-contexte.

THÉORÈME 4 *Tout langage de mots analogiques vérifie la propriété de croissance constante des longueurs.*

4 LA SURPRODUCTION

Le dernier point de l'argumentation des générativistes repose sur le fait que l'analogie, trop lâche, ne saurait constituer un critère de grammaticalité. Nous avions déjà observé dans notre expérience de la section 2 que les phrases grammaticalement incorrectes étaient majoritaires parmi les phrases produites par analogie. De plus, Chomsky argue qu'une analogie valable grammaticalement peut ne pas l'être en sens :

$$\textit{Max peint le mur en rouge} : \frac{\textit{Max peint le}}{\textit{mur rouge}} = \textit{Max voit le mur en rouge} : \frac{\textit{Max voit le}}{\textit{mur rouge}}$$

La réponse à cet argument est de refuser de voir l'analogie à l'œuvre dans un seul espace de symboles. Ainsi, (Itkonen & Haukioja 97, p. 150-156) montrent comment faire fonctionner l'analogie à la fois sur les structures syntaxiques et sur les structures sémantiques, et

(Paul 20, p. 190) insistait déjà sur le fait que l'analogie joue aussi bien au niveau des formes sonores, que du sens. Nous proposons nous aussi de faire jouer l'analogie (au seul niveau des symboles) dans plusieurs espaces à la fois, et de faire résulter la grammaticalité de la correspondance des analogies.

Bien que cela s'écarte (en apparence seulement) du problème de la grammaticalité, illustrons cette idée par une maquette de traduction automatique. Partons de l'alignement de deux ensembles de segments, l'un en français, l'autre en japonais, c'est-à-dire de deux espaces analogiques en correspondance.

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
fille	少女
garçon	少年

Si l'on propose au système une phrase nouvelle, par exemple, « *tu es infirmière* », le système recherche trois phrases en relation d'analogie avec la phrase d'entrée. Ici, il trouve : « *elle est une fille* », « *tu es une fille* » et « *elle est infirmière* ». L'analogie est la suivante :

$$\textit{elle est une fille} : \textit{tu es une fille} = \textit{elle est infirmière} : \textit{tu es infirmière}$$

Le système transpose alors l'analogie précédente aux phrases japonaises, en remplaçant simplement par les phrases correspondantes :

$$\text{彼女は少女だ} : \text{あなたは少女だ} = \text{彼女は看護婦だ} : \mathtt{x}$$

La résolution de l'équation analogique produit la phrase « あなたは看護婦だ », proposée comme traduction de la phrase d'entrée. De cette façon, la traduction d'une phrase est obtenue grâce aux analogies possibles à la fois dans les deux espaces, le français et le japonais. Nous avons obtenu ici une solution en un seul coup, mais une application récursive de la méthode fait de chacun des espaces un langage de mots analogiques, tel que défini dans la section 3. Voici quelques traductions obtenues par notre maquette.

tu es une infirmière.	あなたは看護婦だ。
elle n'est pas mon infirmière.	彼女は私の看護婦ではない。
tu n'es pas mon professeur.	あなたは私の先生ではない。 あなたは私の教授ではない。

5 CONCLUSION

Loin de nous l'idée de faire de l'analogie la seule et unique opération à l'œuvre dans la langue. Mais nous considérons qu'il y a place pour l'analogie en linguistique formelle. Le but premier de cet article n'était pas vraiment de « défendre » l'analogie en réfutant les arguments génératives, puisque cela a déjà été fait par ailleurs. Il s'agissait plutôt pour nous de l'« illustrer » en montrant qu'une formalisation adéquate de l'analogie répond particulièrement bien, et point par point, à toutes ces critiques. Nous avons montré, premièrement, que l'application aveugle de l'analogie sur un ensemble petit de données produit un nombre non négligeable de nouveaux énoncés ; deuxièmement, que la formalisation des langages de mots analogiques permet de se libérer des non-terminaux, laisse conforme l'intuition quant à certains exemples classiques de langages formels et s'approche du légèrement sous-contexte ; troisièmement, que la surproduction de l'analogie peut être considérée comme bénéfique au sens où elle autorise une formalisation de la langue par mise en correspondances de plusieurs espaces analogiques.

Références

Christopher Culy, (1985), The Complexity of the Vocabulary of Bambara (La complexité du vocabulaire bambara), *Linguistics and Philosophy*, vol. 8, pp. 345-351.

Esa Itkonen & Jussi Haukioja, (1997), A rehabilitation of analogy in syntax (and elsewhere) (Une réhabilitation de l'analogie en syntaxe (et ailleurs)), in András Kertész (ed.) *Metalinguistik im Wandel: die kognitive Wende in Wissenschaftstheorie und Linguistik* Frankfurt a/M, Peter Lang, pp. 131-177.

Esa Itkonen, (1994), Iconicity, analogy, and universal grammar (Iconicité, analogie et grammaire universelle), *Journal of Pragmatics*, vol. 22, pp. 37-53.

Aravind K. Joshi, K. Vijay-Shanker, & David Weir, (1991), The Convergence of Mildly Context-Sensitive Grammar Formalisms (Convergence des formalismes grammaticaux légèrement sous-contexte), in P. Sells, S. Shieber & T. Wasow (eds), *Foundational Issues in natural language processing* (Problèmes fondamentaux du traitement de la langue), MIT Press, Cambridge, pp. 31–81.

Yves Lepage, (2000), Languages of Analogical Strings (Langages de mots analogiques), *actes de COLING 2000*, vol 1, Saarbrücken, pp. 488–494.

Solomon Marcus, Carlos Martin-Vide, Gheorghe Păun, (1987), *Contextual Grammars versus Natural Languages* (Grammaires contextuelles et langue), Turku center for Computer Science, TUCS technical Report No 44.

Hermann Paul, (1920) [1$^{\text{ère}}$éd. 1880], *Prinzipien der Sprachgeschichte* (Principes d'histoire des langues), Niemayer, Tübingen, 5$^{\text{è}}$ éd.

Ferdinand de Saussure, (1995) [1$^{\text{ère}}$éd. 1916], *Cours de linguistique générale*, publié par Charles Bally et Albert Séchehaye, Payot, Lausanne et Paris.

Stuart M. Shieber, (1985), Evidence against the Context-Freeness of Natural Language (Preuve contre le caractère hors-contexte de la langue), *Linguistics and Philosophy*, vol. 8, pp. 333-343.

Edward Stankiewicz, (1986), *Baudouin de Courtenay i podstawy współczesnego językoznawstwa* (Baudouin de Courtenay et les bases de la linguistique contemporaine), Ossolineum, Wrocław.