REPRESENTATIVITE DE TEXTES PAR ANALOGIE

Yves LEPAGE
ATR-ITL Laboratoires communication et interprétation,
Hikaridai 2-2, Seika-tyō, Sōraku-gun, 619-02 Kyōto, Japon
lepage@itl.atr.co.jp

Résumé

Dans un travail précédent, nous avons proposé une explication calculatoire de l'analogie, ce phénomène linguistique par lequel un mot (voire une phrase) peut être produit à partir de trois autres. Pour un corpus donné, le nombre d'analogies par lesquelles une nouvelle phrase peut être obtenue croît, du moins initialement, comme le cube de la taille du corpus. Nous étudions la possibilité de réduire cette explosion : localement, en examinant la mémoire immédiate nécessaire pour une nouvelle phrase du corpus ; globalement, en construisant un corpus réduit représentatif du corpus de départ.

INTRODUCTION

L'analogie, un phénomène de grande importance dans la langue, n'a pas jusqu'à présent reçu beaucoup d'attention de la part de la communauté du traitement automatique de la langue (à part quelques notables exceptions [Yvon 89]), du fait du manque d'explication calculatoire. Comme une telle explication est maintenant disponible, l'étude de l'analogie sur de grands corpus est désormais possible.

1 L'ANALOGIE

1.1 Le phénomène linguistique

L'analogie a été décrite par les Junggrammatiker ainsi que par Saussure [Saussure 16]. C'est un phénomène qui apparaît en synchronie. Il explique la création de mots nouveaux et compréhensibles, mais non répertoriés. On en trouve de fréquents exemples dans la bouche des enfants :

$$prendrai : prendre = viendrai : \mathtt{x}$$
$$\mathtt{x} = viendre$$

L'analogie peut être considérée comme le phénomène général à l'œuvre en morphologie parce qu'il rend compte de la régularité des conjugaisons, déclinaisons ou dérivations. Mais il ne rend pas, bien sûr, compte des différences de paradigmes ni des exceptions. Pour toutes ces opérations morphologiques, l'analogie repose sur des opérations sur les chaînes de caractères comme la préfixation, la suffixation ou même l'infixation :

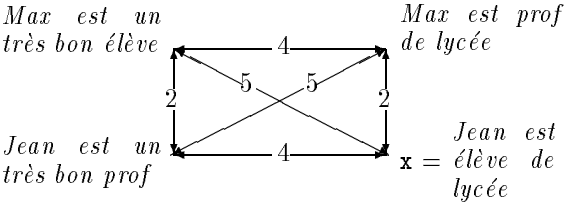
$$professeur : professoral = docteur : \mathtt{x}$$
$$\mathtt{x} = doctoral$$

Hermann Paul lui-même, l'un des Junggrammatiker, et aussi Bloomfield, avaient eu l'idée que, peut-être, l'analogie était aussi à l'œuvre en syntaxe, c'est-à-dire que c'est aussi un moyen de créer des phrases nouvelles à partir de trois autres phrases.

$$\begin{array}{l} Max\ est\ un \\ très\ bon\ élève \end{array} : \begin{array}{l} Jean\ est\ un \\ très\ bon\ prof \end{array} = \begin{array}{l} Max\ est\ prof \\ de\ lycée \end{array} : \mathtt{x}$$
$$\mathtt{x} = Jean\ est\ élève\ de\ lycée$$

1.2 Une explication calculatoire

Les quatre termes intervenant dans la relation d'analogie peuvent être placés aux sommets d'un rectangle dont les diagonales et les côtés opposés (qui représentent les distances entre les phrases) sont respectivement égaux.



Selon cette interprétation, l'analogie est équivalente à un système de trois équations [Lepage 96] (voir aussi [Lepage & Ando 96]).

Definition 1 (Analogie)

$$u : v = w : \mathtt{x} \overset{\Delta}{\Longleftrightarrow} \left\{ \begin{array}{rcl} \mathrm{dist}(u,v) & = & \mathrm{dist}(w,\mathtt{x}) \\ \mathrm{dist}(u,w) & = & \mathrm{dist}(v,\mathtt{x}) \\ \mathrm{dist}(v,w) & = & \mathrm{dist}(u,\mathtt{x}) \end{array} \right.$$

Les distances peuvent être données par les classiques distances d'édition ([Levenshtein 65],

[Wagner & Fischer 74]) définies à partir des opérations de base que sont l'insertion, la suppression et le remplacement :

- insertion (*elle mange* → *elle en mange*)
- suppression (*une grosse lampe* → *une lampe*)
- remplacement (*un beau livre* → *un petit livre*)

La distance d'édition donne le nombre minimal d'opérations d'édition nécessaires pour transformer une phrase en une autre.

1.3 Croissance du nombre d'analogies

D'après la définition 1, dans l'analogie interviennent quatre phrases, dont l'ordre est indifférent. Donc, étant donné un ensemble de n phrases, le nombre de relations d'analogie à vérifier pour connaître le nombre d'analogies dans lesquelles une nouvelle phrase intervient est le nombre d'arrangements de 3 objets dans cet ensemble :

$$A_n^3 = n \times (n-1) \times (n-2)$$

C'est un comportement asymptotique en $O(n^3)$!

Parmi ce nombre total d'arrangements, seulement un certain nombre verifiera la relation d'analogies, mais l'on peut s'attendre à ce que ce nombre croisse particulièrement vite.

Nous allons étudier ce nombre pour plusieurs raisons :

- premièrement, *pour lui-même*: quelle est la courbe exacte de sa croissance ? Croît-elle à l'infini ? Marque-t-elle un arrêt ?
- deuxièmement, à des fins de *limitations* :
  - localement : est-il possible de ne retenir qu'un certain nombre de phrases précédemment lues afin d'assurer quand même un nombre fixé d'analogies en moyenne pour la phrase suivante à lire ?
  - globalement : est-il possible de construire un corpus plus petit qui représente, modulo l'analogie, l'ensemble des phrases d'un grand corpus ?

1.4 Données d'expérience

Nos expérience ont été conduites sur un collection de textes de la banque d'arbres ATR–Lancaster [Black et al. 96]. Ces textes proviennent de diverses sources sur le réseau Internet, allant de pages d'accueil à des rapports médicaux ou économiques en passant par des descriptions géographiques, des postes à des groupes de discussions, *etc.* Pour nos expériences, l'ensemble des textes représente 10 000 phrases (à titre de comparaison, cet article compte un peu plus de 100 phrases).

Ces textes sont disponibles sous deux formes :

- *textes bruts* : le nombre de mots différents s'élève à 27 000;
- *étiquettes* : chaque mot se voit remplacé par une étiquette qui représente ses propiétés grammaticales et sémantiques. L'ensemble d'étiquettes est celui de l'équipe ATR–Lancaster. Dans nos textes apparaissent 1 500 étiquettes différentes.

L'analogie peut être calculée sous sa forme générale, sans contrainte sur les valeurs des distances impliquées. Nous avons déjà rendu compte de telles expériences [Lepage et al. 97]. Ici, nous imposerons à l'analogie que la distance entre deux phrases soit toujours inférieure strictement à la taille *minimale* des deux phrases. Cette contrainte est assez forte.

Pour chaque phrase (sous chacune des deux formes), nous calculons le nombre d'analogies dans lesquelles n'interviennent, avec la phrase considérée, que des phrases d'un rang inférieur dans le corpus. Nous appelons ce nombre le nombre d'analogies passées.

2 NOMBRE D'ANALOGIES PASSEES

2.1 Résultats

Le nombre d'analogies passées, pour les deux formes du corpus, est donné par la figure 1 pour chaque phrase du corpus. Les courbes ont été lissées en prenant la moyenne des 100 valeurs précédentes.

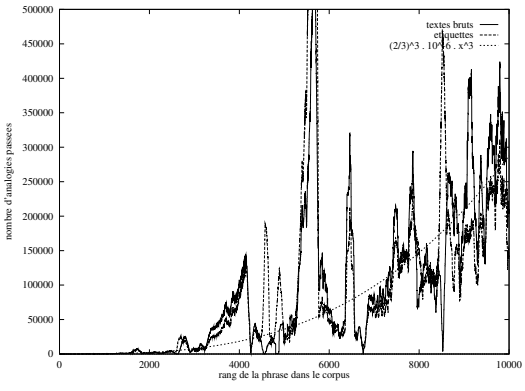


Figure 1: Nombre d'analogies passées (courbes lissées).

La forme générale des courbes est à peu près la même pour les deux formes du corpus. Cependant, nous constatons en particulier une nette disparité autour des rangs 4 500 et 8 500. Aux alentours de ces rangs, apparaissent des listes de

chiffres. Un chiffre isolé étant codé par la même étiquette, la proportion d'analogies sur de telles listes augmente nécessairement pour le texte sous forme d'étiquettes, alors qu'elle baisse pour le texte brute. D'autres disparités constatées sont expliquées par le fait que les étiquettes sont peu sensibles aux majuscules ou minuscules. Les deux courbes ont donc des allures différentes sur les morceaux de texte écrits entièrement en majuscules.

2.2 Croissance cubique

Dans les deux cas, la tendance générale des courbes n'est pas trop mal approchée par une cubique :

$$y = (K \times x)^3$$

Cela semble tout à fait naturel puisque quatre objets interviennent dans l'analogie, l'un deux étant fixé dans notre expérience. Une valeur de K a été obtenue de façon expérimentale : $2/3 \times 10^{-2}$.

Il ne semble pas que le nombre d'analogies passées marque un arrêt après un certain seuil, tout du moins pour ce qui est de notre corpus. De nouvelles expériences, extrêmement coûteuses en temps de calcul, sont donc encore nécessaires sur des corpus de plusieurs dizaines de milliers de phrases.

3 MEMOIRE IMMEDIATE

Puisque le nombre d'analogies dans lesquelles une phrase intervient est si élevé, et puisque l'on peut penser que toute application informatique de l'analogie se contentera d'un nombre réduit de ces analogies, il est intéressant de mesurer jusqu'où l'on peut oublier les phrases lues précédemment tout en garantissant un certain nombre fixé d'analogies pour la phrase suivante. Autrement dit, quelle est la mémoire immédiate, en nombre de phrases, suffisante pour garantir en moyenne un nombre fixé d'analogies pour la phrase suivante du corpus ? En d'autres termes encore, nous mesurons le plus petit $m(r,n)$ tel que la phrase de rang $r+1$ puisse être obtenue n fois par analogie à partir de triplets de phrases de rang compris entre $r - m(r,n)$ et r.

Cette mesure est d'importance pratique, puisque, dans toute application, la taille de cette mémoire sera déterminante pour la rapidité d'exécution.

3.1 Forme globale

Le graphe de la figure 2 montre les mesures (lissées) obtenues pour les textes bruts et étiquettes pour la valeur de $n = 1$.

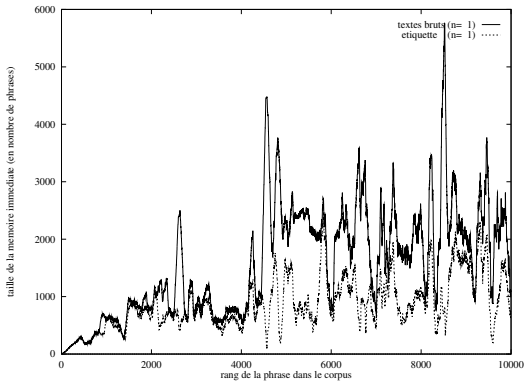


Figure 2: Taille de la mémoire immédiate garantissant un nombre fixé d'analogies.

Une propriété remarquable de ces courbes est leur forme relativement identique. Les courbes (que nous ne donnons pas ici) pour les valeurs $n = 10, 50, 100$ confirment cette impression, comme si elles étaient obtenues les unes à partir des autres par simple dilatation le long de l'axe vertical.

Les courbes pour les deux formes du corpus vérifient aussi cette propriété. Le fait de remplacer les mots par des étiquettes de propriétés grammaticales et sémantiques a pour effet d'abaisser la courbe (taille de la mémoire, de 1 600 en moyenne pour les textes bruts à 860 en moyenne pour les étiquettes), sans en changer la forme générale. On peut donc dire que les 1 500 étiquettes constituent une représentation économique et fidèle des mots (au nombre de 27 000), tout au moins en ce qui concerne l'analogie.

3.2 Hauteur absolue et mémoire moyenne

La taille de la mémoire immédiate (à lire en ordonnée) reflète la cohérence du texte : plus les phrases sont analogues dans un contexte réduit, plus la mémoire immédiate est petite ; plus les phrases du texte sont diverses, plus cette mémoire est importante.

Pour nos données, une mémoire immédiate de 2 500 phrases garantit en moyenne 100 analogies pour la phrase suivante. Bien sûr, des mémoires moindres, de 2 300, 2 000 et 1 600 phrases respectivement, sont suffisantes pour garantir 50, 10 analogies ou 1 seulement en moyenne.

3.3 Minimums locaux

On observe de nombreux minimums locaux. Si l'on pense que l'analogie rend compte d'une certaine similarité entre phrases, alors, il est raisonnable de penser que des phrases partageant le même style dans un texte devrait avoir la même mémoire. On devrait donc observer des croissances brutales

lors d'un changement de style, et une décroissance lorsque le style reste le même. En d'autres termes, les minimums locaux devraient correspondre a certains frontières entre textes (deux textes consécutifs peuvent partager le même style).

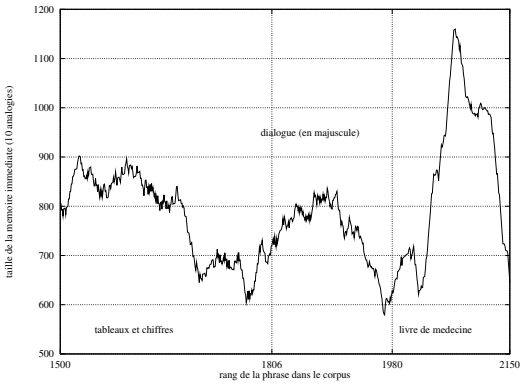


Figure 3: Discrimination de style par la mémoire immédiate.

Cette hypothèse est vérifiée et illustrée sur le graphe de la figure 3 sur une portion du corpus, pour laquelle les frontières de textes visualisées par les lignes verticales sont les plus proches des minimums locaux. Les trois morceaux de textes peuvent être assez bien caractérisés :

- de 1 500 à 1 739 : listes de faits (données géographiques et économiques sur des pays africains, liste des significations des codes maritimes, *etc.*) ;
- de 1 826 à 1 980 : transcription de dialogue écrit en majuscules ;
- de 1 990 à 2 150 : livre de médecine traitant du cancer.

4 REPRESENTATIVITE

Dans la section précédente, nous avons étudié la mémoire immédiate. Plutôt que de retenir les dernières phrases lues, on peut se demander si l'on ne peut pas ne retenir que certaines phrases, et ce depuis le début du corpus, de façon à garantir l'obtention de la phrase suivante par analogie. Ce problème est très proche de celui de la représentativité.

4.1 Définitions

La représentativité, est une notion beaucoup utilisée en linguistique mais qui n'a pas à notre connaissance de définition rigoureuse. Or, l'analogie permet d'en proposer une. Nous nous appuyons pour cela sur des notions simples d'algèbre moderne[1].

4.1.1 Générateurs, libres et bases

Pour un ensemble de phrases, ou corpus, $\mathcal{C}$, on veut obtenir un ensemble plus petit de phrases, $\mathcal{G}$, tel que toutes phrase de $\mathcal{C}$ puisse être obtenue par analogie à partir de phrases de $\mathcal{G}$. C'est la notion classique de générateur[2].

Definition 2 (Générateur) *Un ensemble $\mathcal{G}$ est un* générateur *de $\mathcal{C}$ si et seulement si*

$$\forall p \in \mathcal{C},\ \exists (p_1, p_2, p_3) \in \mathcal{G}^3 \ /\ p_1 : p_2 = p_3 : p$$

Evidemment, l'information contenue dans $\mathcal{B}$ ne doit pas être redondante relativement à l'analogie : aucune phrase de cet ensemble ne doit pouvoir être obtenue à partir de phrases du même ensemble par analogie. C'est la notion classique de libre.

Definition 3 (Libre) *Un ensemble $\mathcal{L}$ est* libre *si et seulement si $\mid \mathcal{L} \mid < 4$ ou*

$$\forall (p_1, p_2, p_3, p_4) \in \mathcal{L}^4,\ p_1 : p_2 = p_3 : p_4 \textit{ est faux}$$
$$(\textit{avec } p_1, p_2, p_3 \textit{ et } p_4 \textit{ tous différents})$$

Pour un corpus $\mathcal{C}$, on veut précisément obtenir un ensemble qui vérifie les deux définitions précédentes à la fois. Classiquement, un tel ensemble est appelé une base.

Definition 4 (Base) *Un ensemble $\mathcal{B}$ est une* base *de $\mathcal{C}$ si et seulement si c'est un générateur libre de $\mathcal{C}$.*

Nous pouvons alors proposer, relativement à l'analogie, une définition de la représentativité, ce vieux serpent de mer de la linguistique :

Definition 5 (Représentativité par analogie) *Un ensemble de phrases $\mathcal{B}$ est représentatif d'un ensemble de phrases $\mathcal{C}$ si et seulement si c'est une base pour $\mathcal{C}$.*

4.1.2 Analogie et structure de matroïde

Les définitions précédentes ne préjugent en rien de l'intersection avec l'ensemble de départ $\mathcal{C}$. Tous ces ensembles peuvent très bien être construits à partir d'éléments (des phrases, pour nous ici) n'appartenant pas à $\mathcal{C}$. Dans la suite, nous nous restreindrons au contraire au cas où ces ensembles sont construits uniquement à partir d'éléments de

[1] En fait, des notions classiques des espaces vectoriels, ou plus précisément, de théorie des matroïdes.

[2] Les définitions données ne font pas intervenir le fait que les éléments sont, pour nous ici, des phrases. Les distances d'édition pouvant s'appliquer aux mots (chaînes de caractères), aux phrases (chaînes de mots) ou aux arbres indifféremment, la definition 1 de l'analogie est valable pour tous ces objets. Par conséquent, les définitions suivantes sont aussi valables pour tous ces objets.

$\mathcal{C}$, sans aller plus bas que le niveau des phrases : nous ne construirons pas (pas encore !) de phrases nouvelles à partir de morceaux de phrases de $\mathcal{C}$, par exemple. Avec cette restriction, on rappelle les définitions suivantes :

Definition 6 (Générateur minimal) *Un générateur $\mathcal{G}$ de $\mathcal{C}$ est un* générateur minimal *si et seulement si*

$$\forall p \in \mathcal{G},\ \ \mathcal{G} \setminus \{p\} \textit{ n'est pas un générateur de } \mathcal{C}$$

Definition 7 (Libre maximal) *Un libre $\mathcal{L}$ dans $\mathcal{C}$ est un* libre maximal *si et seulement si*

$$\forall p \in \mathcal{C} \setminus \mathcal{F},\ \ \mathcal{F} \cup \{p\} \textit{ n'est pas libre}$$

Pour une structure de matroïde, les bases, les libres maximaux et les générateurs minimaux sont les mêmes ensembles, et ils possèdent la remarquable propriété qu'ils ont tous même cardinal. Malheureusement, la définition de l'analogie donnée en Déf. 1 n'implique pas une structure de matroïde dans le cas général. Lorsque nous construirons des bases de façon expérimentale, et puisqu'en fait nous sommes intéressés à obtenir le plus petit ensemble représentatif, il nous faudra malheureusement examiner ce problème de cardinalité.

4.2 Algorithme de construction

L'algorithme suivant construit une base $\mathcal{B}$ pour un corpus $\mathcal{C}$ donné.

$\mathcal{B}$:= n'importe quel triplet (p, p', p'') de $\mathcal{C}^3$
$\mathcal{C}' := \mathcal{C} \setminus \{p, p', p''\}$
tant que $\mathcal{C}' \neq \emptyset$
  a := n'importe quel élément de $\mathcal{C}'$
  $\mathcal{C}' := \mathcal{C}' \setminus \{a\}$
  si $\exists (b, c, d) \in \mathcal{B}^3$ / $a : b = c : d$ ①
    (b, c et d différents deux à deux)
    alors rien
    sinon $\mathcal{B} := \mathcal{B} \cup \{a\}$
  fin si
fin tant que

Il est facile de vérifier que $\mathcal{B}$ est bien une base pour $\mathcal{C}$. En effet, la condition de la ligne ① garantit que l'ensemble construit est un libre et un générateur de $\mathcal{C}$.

Cet algorithme est en fait un algorithme glouton, et il n'est pas trop surprenant qu'il marche lorsque l'on sait les relations privilégiées que ce type d'algorithme entretient avec les matroïdes.

Cependant, la présente définition de l'analogie n'implique pas une structure de matroïde. Si l'ensemble délivré par cet algorithme est bien une base, il n'est pas assuré qu'il soit libre maximal ou générateur minimal. Sa taille peut donc varier avec l'ordre dans lequel les éléments de $\mathcal{C}$ sont choisis au cours de l'exécution.

4.3 Expériences et discussion

4.3.1 Taille comparée pour les mots et les étiquettes

L'algorithme ci-dessus a été appliqué à notre corpus d'expérience. Ce qui nous intéresse est le cardinal de la base obtenue. C'est donc ce qui est tracé dans la figure 4, pour les textes sous leurs deux formes.

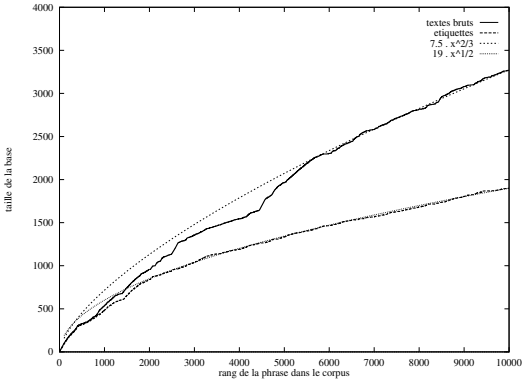


Figure 4: Taille de la base.

Les deux courbes, pour les textes bruts et étiquettes sont assez différentes l'une de l'autre. Elles peuvent être approchées respectivement par les fonctions :

$$y = 7.5 \times \sqrt[3]{x^2} \quad \text{et} \quad y = 19 \times \sqrt{x}$$

Ces courbes sont moins encourageantes que lorsqu'aucune contrainte n'est imposée au rectangle de l'analogie, comme dans [Lepage et al. 97], où les courbes sont nettement logarithmiques. Ces différences de résultats sont expliquées par le fait que, dans les expériences présentes, l'analogie à vérifier, devant être de bien meilleure «qualité», est plus rare.

4.3.2 Taille dans plusieurs expériences

Comme il a été dit plus haut, notre algorithme ne peut garantir que la base soit un générateur minimal. Sa taille peut donc varier en fonction de l'ordre d'apparition des phrases dans le corpus. Nous avons donc appliqué l'algorithme précédent pour plusieurs permutations, choisies au hasard, des 5 000 phrases. Certes, le nombre total de permutations possibles est 5 000!, et il est difficile de prétendre que ces expériences sont représentatives, mais on constate, dans toutes ces expériences, que, premièrement, les courbes sont assez identiques, et que surtout, la taille de la base représentative de 10 000 phrases de textes bruts est toujours comprise dans une fourchette de 3 000 à 3 500 phrases.

Du point de vue pratique, la taille d'un corpus représentatif de 10 000 phrases est réduit à moins

d'un tiers du corpus initial (3 200 phrases), voire un cinquième pour les étiquettes (2 000 phrases).

4.3.3 Génération à partir d'une base

Ce nombre de 3 200 est à rapprocher de la taille de la mémoire immédiate nécessaire pour assurer en moyenne 100 analogies: 2 500. Mais ce n'est qu'un rapprochement, car, dans un cas, il s'agit d'un nombre de phrases calculé sur l'ensemble du corpus, dans le second cas, il s'agit d'un contexte immédiat.

Cependant, le nombre de fois qu'une phrase du corpus de départ peut être obtenue par analogie avec des phrases du corpus représentatif, constitue, en quelque sorte, une mesure de la qualité du corpus représentatif. Ce nombre que nous avons calculé par souci de vérification est égal à 1 pour chaque phrase du corpus. Le corpus représentatif est donc «minimal» dans le sens où aucune phrase du corpus de départ n'est obtenue plus d'une fois.

Mais ce corpus représentatif n'est certainement pas minimal au sens où les phrases pouvant être engendrées à partir du corpus représentatif constitueraient exactement le corpus de départ. Ce dernier point est d'autant moins sûr quand aucune contrainte n'est imposée au rectangle de l'analogie comme dans [Lepage et al. 97], ou la taille du corpus representatif ne dépasse pas les 120 phrases pour 5 000 phrases.

CONCLUSION

Nous avons étudié le comportement de l'analogie (avec contrainte), telle que définie de façon calculatoire, sur des données réelles: quelques milliers de phrases de la banque d'arbres ATR–Lancaster.

Tout d'abord, le nombre d'analogies par lesquelles une phrase nouvelle peut être obtenue à partir de phrases passées a été mesuré. Nous avons mis en évidence une progression cubique, sans pouvoir déterminer si ce nombre cesse de croître après un certain seuil.

Ensuite, nous avons mesuré la mémoire immédiate nécessaire qui permet de garantir un certain nombre de relations d'analogies pour une nouvelle phrase. Les brusques chutes et montées observées marquent bien, semble-t-il, les changements de styles entre textes.

Enfin, à partir d'un grand corpus, nous avons montré comment construire un sous–corpus représentatif, tel que toute phrase du grand corpus puisse être obtenue par analogie à partir de phrases du sous–corpus. L'existence et la taille réduite de sous–corpus représentatifs est de bonne augure pour des applications de l'analogie en traitement automatique des langues.

BIBLIOGRAPHIE

[Black et al. 96] Ezra Black, Stephen Eubank, Kashioka Hideki, David Magerman, Roger Garside and Geoffrey Leech
Beyond Skeleton Parsing: Producing a Comprehensive Large–Scale General–English Treebank with Full Grammatical Analysis
Proceedings of COLING-96, Copenhagen, August 1996, pp. 107-112.

[Lepage 96] Yves Lepage
Ambiguities in analysis by analogy
Proceedings of MIDDIM-96, post–COLING seminar on interactive desambiguation, Christian Boitet ed., August 1996, pp. 93–100.

[Lepage & Ando 96] Yves Lepage & Ando Shin-Ichi
Un éditeur pour la construction de banques d'arbres
Actes de TALN-96, Marseille, mai 1996, pp. 104-111.

[Lepage et al. 97] Yves Lepage, Ando Shin-ichi, Iida Hitoshi
The snow–ball effect of analogy
Third annual meeting of the Japanese Association for Natural Language Processing, Kyoto, March 1997, pp. 289-292.

[Levenshtein 65] V.I. Levenshtein
Binary codes capable of correcting deletions, insertions and reversals
Dokl. Akad. Nauk SSSR, vol. 163, No. 4, August 1965, pp. 845-848.
English translation in *Soviet Physics-doklady* vol. 10, No. 8, February 1966, pp. 707-710.

[Saussure 16] Ferdinand de Saussure
Cours de linguistique générale
publié par Charles Bally et Albert Sechehaye, Payot, Lausanne et Paris, 1916.

[Wagner & Fischer 74] Robert A. Wagner and Michael J. Fischer
The String-to-String Correction Problem
Journal for the Association of Computing Machinery, Vol. 21, No. 1, January 1974, pp. 168-173.

[Yvon 89] François Yvon
Prononcer par analogie: motivation, formalisation et évaluation
Thèse, ENST, mai 1996.