

# Apprentissage par analogie et proportions formelles: contributions méthodologiques et expérimentales

Nicolas Stroppa, François Yvon

GET/ENST et LTCI, UMR 5141  
46, rue Barrault, 75013 Paris  
{stroppa,yvon}@enst.fr

**Résumé** : L'apprentissage par analogie exploite un mécanisme inductif en deux étapes : (i) construction d'un *appariement structurel* (d'une relation d'analogie) entre une nouvelle situation et des situations déjà connues ; (ii) transfert partiel des propriétés de la situation analogue vers la situation nouvelle. Cette approche pré-suppose la capacité à rechercher et à exploiter de tels appariements, ce qui implique de donner un sens à la notion de relation analogique et d'implanter efficacement leur calcul.

Dans cet article, nous proposons une définition de la notion de proportion analogique valant pour des structures algébriques quelconques. Nous montrons également comment cette définition s'instancie pour un certain nombre de structures algébriques classiques : structures attribut-valeur, ensembles, mots sur un alphabet fini, arbres étiquetés. Pour chacune de ces structures, nous discutons la complexité du calcul des relations de proportionnalité.

Nous présentons ensuite une application de ces résultats sur une tâche de traitement automatique des langues, consistant à apprendre à caractériser linguistiquement des formes orthographiques inconnues. Des résultats expérimentaux sur des lexiques anglais et français permettent d'apprécier la validité de notre démarche.

**Mots-clés** : Apprentissage par analogie, Traitement automatique des langues, Morphologie, Langages formels

## 1 Introduction

L'apprentissage par analogie (Gentner *et al.*, 2001) se fonde sur un mécanisme inductif en deux étapes : le premier temps consiste à construire un appariement structurel (d'une relation d'analogie) entre une nouvelle instance d'un problème avec des instances déjà résolues du même problème ; une fois cet appariement établi, une nouvelle solution peut être élaborée à partir d'une ou plusieurs solutions analogues. La mise en œuvre de ce type d'apprentissage présuppose donc la capacité à rechercher et à exploiter de tels appariements, soit d'une part de donner un sens à la notion de relation analogique, et, d'autre part, d'implanter efficacement le calcul de ces relations.

Dans certains domaines d'application, en particulier le traitement des langues, la taille des bases de données disponibles, qui contiennent typiquement des centaines de milliers d'instance, rend prohibitrice la recherche d'appariements structurels complexes. En revanche, ces domaines se caractérisent par des descriptions qui intègrent des attributs prenant des valeurs dans des ensembles de séquences (phonétiques, orthographiques, syntaxiques...), d'arbres ou de structures de traits : l'exploitation d'analogies purement formelles entre ces séquences permet dans certaines situations de détecter des analogies plus profondes entre entités linguistiques. Cette observation a donné lieu à de nombreuses tentatives pour exploiter les analogies de forme, en particulier pour ce qui concerne des tâches de prononciation automatique (Yvon, 1999), d'analyse morphologique (Lepage, 1999a; Pirrelli & Yvon, 1999a), ou syntaxique (Lepage, 1999b). Ces travaux se sont principalement focalisés sur les séquences finies de symboles, en définissant de manière restrictive et parfois ad-hoc la notion d'analogie formelle. L'objectif premier de cet article est de proposer une définition générale de l'analogie formelle sur des structures algébriques communément utilisées en traitement automatique des langues : structures attribut-valeur, mots sur un alphabet fini et arbres étiquetés, et, pour chacune de ces structures, d'implanter des algorithmes à même de traiter efficacement les larges volumes de données disponibles. Le second objectif est de confirmer expérimentalement, ici pour une tâche d'analyse morphologique, la validité de l'approximation effectuée en ne considérant que des analogies de forme.

La contribution de ce travail est donc double :

- nous proposons une définition unifiée de la notion de rapport de proportionnalité formel, définition qui s'instancie sur un grand nombre de structures algébriques communes et nous montrons que cette définition donne lieu à un calcul efficace de ces proportions. Pour ce qui concerne les mots et les arbres, notre définition généralise celle donnée dans (Lepage, 1998) ;
- l'application à l'apprentissage de relations morphologiques donne lieu à un modèle qui est compatible avec les théories contemporaines de la morphologie lexématique (i.e une morphologie sans morphème), voir e.g. (Matthews, 1974; Fradin, 2003). Ce modèle est également compatible avec une approche morphématique, comme le montrent nos expériences sur le calcul de la structure interne (hiérarchique) des formes construites, alors que la majorité des approches existantes se limite au calcul des frontières de morphèmes.

Cet article est organisé comme suit. La Section 2 est consacrée à une exposition de notre interprétation de l'apprentissage par analogie, qui est contrastée avec d'autres modèles du raisonnement par analogie ou de l'apprentissage à base d'instances. Nous introduisons ensuite une définition de l'analogie valant pour des structures algébriques quelconques (Section 3), en détaillant son instanciation dans un certain nombre de cas particuliers : structures attribut-valeur, ensembles, mots sur un alphabet fini, arbres étiquetés. La Section 4 présente une expérimentation conduite sur une tâche d'analyse morphologique de formes orthographiques inconnues. Des résultats expérimentaux obtenus sur des données de l'anglais et du français permettent d'apprécier la validité de notre démarche. La section 5 est enfin consacrée à un retour critique sur les résultats obtenus, permettant de mieux cerner les limitations de notre modèle et d'ouvrir les perspectives pour des travaux futurs.

## 2 Apprentissage par analogie

Dans cette section, nous introduisons notre modèle d'inférence analogique, qui s'inspire, pour une large part, des travaux présentés dans (Pirrelli & Yvon, 1999b); nous contrastons ensuite ce modèle avec d'autres modèles du raisonnement et de l'apprentissage par analogie (Section 2.2).

### 2.1 Un modèle d'inférence analogique

Nous considérons une tâche générique d'apprentissage automatique supervisé, qui consiste, à partir d'une base d'apprentissage décrivant des objets connus, à inférer des propriétés d'objets nouveaux, qui ne sont que partiellement informés. L'ensemble des propriétés connues constitue l'*espace d'entrée* de l'apprentissage, les propriétés à inférer constituant l'*espace de sortie*. Cette situation recouvre, en particulier, le cas de la catégorisation automatique, dans laquelle la sortie se réduit à une étiquette de classe, mais également de nombreuses autres configurations intéressantes.

Dans notre modèle, l'étape d'apprentissage se réduit à une mémorisation par cœur des objets connus : il s'agit donc d'un apprentissage à base d'instances ou apprentissage *paraséux* (Mitchell, 1997). L'étape d'inférence s'effectue par identification de relations formelles de proportionnalité existant dans l'espace d'entrée, puis par reconstruction analogique des attributs inconnus.

Formellement, nous supposons donnée une base d'apprentissage  $\mathcal{A}$  d'objets représentés par un ensemble fini de descripteurs pouvant prendre des formes très variables. Nous supposons également donnée la possibilité de construire des rapports de proportion formels entre les objets de  $\mathcal{A}$ . Ces rapports de proportion font l'objet de la Section 3. Pour l'heure, définissons simplement un rapport de proportion comme une relation impliquant quatre objets  $A, B, C$  et  $D$  et qui sera dénotée  $A : B :: C : D$ , signifiant *A est à B comme C est à D*. Lorsque les objets ne sont que partiellement informés, nous notons  $A^+$  la partie connue de  $A$  et  $A^-$  la partie inconnue. Soit alors  $X^+$  un objet partiellement décrit et absent de la base d'apprentissage, l'inférence analogique se formalise comme suit :

- recherche de l'ensemble  $\mathcal{T}$  des triplets de  $\mathcal{A} \times \mathcal{A} \times \mathcal{A}$  défini par :

$$\mathcal{T}(X) = \{(A, B, C) \in \mathcal{T}, A^+ : B^+ :: C^+ : X^+\}$$

- chaque triplet de  $\mathcal{T}(X)$  donne lieu à une ou plusieurs hypothèses  $\hat{X}^-$  sur les propriétés inconnues de  $X$  ; par résolution de l'*équation analogique* :  $\hat{X}^- = A^- : B^- :: C^- : ?$

Pour achever la description de cet algorithme, plusieurs aspects doivent être précisés :

- la stratégie d'exploration de  $\mathcal{A} \times \mathcal{A} \times \mathcal{A}$  : une stratégie d'exploration exhaustive conduit à évaluer  $|\mathcal{A}|^3$  triplets, ce qui n'est pas toujours réalisable. Par ailleurs,  $\mathcal{A}$  peut contenir des objets inégalement pertinents au regard de la tâche, suggérant de pondérer différemment les instances disponibles. Dans les expériences décrites à Section 4, nous avons procédé par échantillonnage sous contrainte, en considérant toutes les instances comme équiprobables.
- une stratégie pour résoudre les ambiguïtés : lorsque plusieurs hypothèses sont proposées, un classement des hypothèses doit être proposé. Dans le cadre de ce travail, les hypothèses sont ordonnées par une procédure de vote majoritaire.

## 2.2 Raisonner et apprendre par analogie

La capacité d'identifier des relations analogiques entre des situations apparemment distinctes et d'utiliser ces relations pour résoudre des problèmes, est souvent présentée comme une capacité cognitive centrale (voir e.g. (Gentner *et al.*, 2001)). Cette observation a suscité un grand nombre de travaux visant à modéliser cette capacité, aussi bien par des modèles symboliques (e.g. (Falkenheimer & Gentner, 1986; Thagard *et al.*, 1990; Hofstadter & Mitchell, 1995)) que subsymboliques (e.g. (Plate, 2000)). L'aspect central de ces modélisations concerne le processus de fabrication dynamique d'un *appariement structurel* entre une situation nouvelle et une situation mémorisée. L'appariement structurel vise à rapprocher des situations, qui, bien qu'en apparence (en surface) fort différentes, mettent en jeu des ensembles de relations qu'il est possible de reconnaître comme identiques : le système solaire se distingue en apparence de l'atome par sa taille ; il s'en rapproche par le fait que des parties du système sont en rotation autour d'un centre, suggérant une ressemblance entre les raisons qui causent cette rotation. La construction d'un appariement structurel entre deux situations mobilise donc plusieurs termes de la description de ces situations et les relations qu'ils entretiennent entre eux, permettant de construire des énoncés tels que : *l'électron est au noyau atomique comme la terre est au soleil*. Un cas limite de raisonnement par analogie est le raisonnement par cas, dans lequel toutes les situations envisagées se rapportent à un même problème, facilitant la construction des appariements et la construction de solutions.

Si l'analogie semble jouer un rôle central dans le raisonnement, elle est également invoquée pour expliquer des comportements humains n'impliquant pas de raisonnement conscient, en particulier pour des tâches liées à la production et à la perception du langage : accès lexical, prononciation de mots inconnus, analyse de construits morphologiques, etc. Dans ce contexte, production analogique s'oppose à production régulière. De nombreux modèles exploitant des mécanismes d'apprentissage automatique ont ainsi été proposés comme substituts possibles aux modèles à base de règles, allant des réseaux de neurones aux arbres de décision en passant par des méthodes de type *k*-plus proche voisins ou s'en rapprochant (voir, par exemple, (Skousen, 1989; Jones, 1996; Daelemans *et al.*, 1999)). Dans ces travaux, l'accent est mis sur la qualité des généralisations produites en présence de larges quantités de données, ces généralisations s'appuyant le plus souvent sur des similarités de surface.

Le modèle que nous proposons se situe à mi-chemin entre ces deux approches : du raisonnement par analogie, nous retenons l'idée de recherches de ressemblances qui dépassent les similarités de surface, ce qui se traduit dans notre procédure par le rôle central donné aux relations de proportionnalité. Des travaux sur les « analogies spontanées », en particulier celles qui se placent dans le paradigme de l'apprentissage paresseux, nous retenons le principe d'un apprentissage par cœur et l'utilisation de critères statistiques pour départager des hypothèses en compétition.

## 3 Relations d'analogies sur des structures algébriques

Dans cette section, nous considérons les relations de proportionnalité formelles, en débutant par une définition générale (Section 3.1), qui est ensuite instanciée pour di-

verses structures algébriques : treillis et monoïdes. Ces définitions sont implantées dans un outil générique de résolution d'analogies, s'appuyant sur la bibliothèque de manipulation d'automates Vaucanson, qui utilise massivement la programmation générique à base de templates (Lombardy *et al.*, 2003).

### 3.1 Les bases

La notion d'analogie implique deux notions clés : la *décomposition* de chaque objet sous la forme d'une combinaison de termes plus petits qui observent des *contraintes d'alternance*. Notons  $U$  un ensemble quelconque muni d'une loi de composition interne associative, notée  $\oplus$ . Pour exprimer formellement la notion de décomposition, nous introduisons la notion de *factorisation* d'un élément  $u$  de  $U$ , définie par :

#### Définition 1

Une factorisation de  $u \in U$  est une séquence  $u_1 \dots u_n$ , avec  $\forall i, u_i \in U$  et telle que :  $u_1 \oplus \dots \oplus u_n = u$ .

En intégrant la contrainte d'alternance entre les termes d'une décomposition, nous aboutissons à la définition générale suivante pour les proportions analogiques :

#### Définition 2 (Proportion analogique)

$(x, y, z, t) \in U$  constituent une proportion analogique, notée  $x : y :: z : t$  si et seulement si il existe des factorisations  $x_1 \oplus \dots \oplus x_n = x$ ,  $y_1 \oplus \dots \oplus y_n = y$ ,  $z_1 \oplus \dots \oplus z_n = z$ ,  $t_1 \oplus \dots \oplus t_n = t$  telles que  $\forall i, (y_i, z_i) \in \{(x_i, t_i), (t_i, x_i)\}$ . Le plus petit entier  $n$  pour lequel de telles factorisations existent est le degré de l'analogie.

Cette définition s'applique directement à de nombreuses structures algébriques classiques, telles que les espaces vectoriels, les groupes et les monoïdes libres. Ceci inclut en particulier des structures classiquement utilisées pour représenter des connaissances : les ensembles munis de l'opération d'union ensembliste, les structures de traits munies de l'opération de généralisation, qui sont étudiées à la section 3.2.

Dans le cas où la structure sous-jacente possède des propriétés supplémentaires (commutativité, existence d'un élément neutre, existence d'un inverse unique pour  $\oplus$ ), cette définition se simplifie et permet de retrouver des définitions bien connues de l'analogie (Stroppa & Yvon, 2005). Ainsi, dans le cas où  $(U, \oplus)$  est un groupe, (2) se simplifie en :

#### Définition 3 (Proportion analogique (dans un groupe))

$(x, y, z, t) \in U$  constituent une proportion analogique, notée  $x : y :: z : t$  si et seulement si :  $x \oplus t = y \oplus z$ .

On retrouve là une intuition classique, qui correspond au rapport de proportionnalité dans  $(\mathbb{N}, \times)$  et à la relation entre les sommets d'un parallélogramme dans un espace vectoriel.

## 3.2 Structures de traits, ensembles

L'ensemble des structures de traits et l'ensemble des parties d'un ensemble sont deux cas particuliers de *treillis*. Un treillis est une algèbre non-vide dont les deux opérations internes binaires (notées  $\vee$  et  $\wedge$ ) sont idempotentes, commutatives, associatives, mutuellement distributives et satisfaisant la loi d'absorption. L'ensemble des structures de traits, muni de des opérations d'unification et de généralisation est un treillis. C'est également le cas de l'ensemble des parties d'un ensemble, muni de l'union et de l'intersection. Si  $(U, \vee, \wedge)$  est un treillis, (2) se ramène à :

### Définition 4 (Proportion analogique (dans un treillis))

$(x, y, z, t) \in U$  constituent une proportion analogique, notée  $x : y :: z : t$  si et seulement si :

$$\begin{aligned}x &= (x \wedge y) \vee (x \wedge z) \\y &= (x \wedge y) \vee (t \wedge y) \\z &= (t \wedge z) \vee (x \wedge z) \\t &= (t \wedge z) \vee (t \wedge y)\end{aligned}$$

Cette définition s'applique directement au cas des structures de traits (avec ou sans réentrance) et des ensembles. Dans ce dernier cas, elle généralise le modèle proposé par (Lepage, 2001). En outre, la disparition des quantificateurs dans (4) réduit la vérification d'une relation analogique au calcul de 8 opérations atomiques : 4 unifications et 4 généralisations pour les structures de traits, 4 unions et 4 intersections dans le cas des ensembles. Une procédure de calcul efficace s'en déduit immédiatement. Le cas des multi-ensembles donne lieu à une définition similaire (Stroppa & Yvon, 2005).

## 3.3 Mots sur un alphabet fini

### 3.3.1 Relations analogiques sur les mots

Soit  $A$  un alphabet fini, on note  $\Sigma^*$  l'ensemble des séquences finies d'éléments de  $A$ , qu'on appellera *mots* sur  $A$ .  $\Sigma^*$ , muni de l'opération de concaténation  $.$  est un monoïde libre, dont l'élément neutre est le mot de longueur nulle, noté  $\varepsilon$ . Pour  $w \in \Sigma^*$ ,  $w(i)$  désigne le  $i$ ème symbole de  $w$ . Dans ce cadre, la définition (2) se réexprime par :

### Définition 5 (Proportion analogique entre mots)

$(x, y, z, t) \in \Sigma^*$  constituent une proportion analogique, notée  $x : y :: z : t$  si et seulement s'il existe des factorisations  $x_1 \dots x_n, y_1 \dots y_n, z_1 \dots z_n, t_1 \dots t_n$  respectivement de  $x, y, z$  et  $t$  telles que  $\forall i, (y_i, z_i) \in \{(x_i, t_i), (t_i, x_i)\}$ . Le plus petit entier  $n$  pour lequel de telles factorisations existent est le degré de l'analogie.

Un exemple d'analogie sur les mots est : *sers : resers :: donne : redonne*, avec  $x_1 = re, x_2 = sers...$

Cette définition généralise la définition de l'analogie entre mots proposée par (Lepage, 2001). Elle ne garantit ni qu'une équation analogique possède toujours une solution, ni inversement l'unicité d'une solution. (Lepage, 2001) donne également une série de conditions nécessaires pour qu'une équation ait au moins une solution, conditions qui s'appliquent également ici. En particulier, si  $t$  est solution de  $x : y :: z : ?$ , alors  $t$  contient tous les symboles de  $y$  et de  $z$  qui ne sont pas dans  $x$ , dans un ordre inchangé.

Un corollaire est que toutes les solutions d'une équation analogique ont la même longueur.

### 3.3.2 Un solveur à états finis

La définition (5) donne lieu à une procédure efficace de résolution d'une équation analogique qui s'appuie sur le formalisme des transducteurs à états finis. Nous esquissons ici les grandes lignes de cette procédure, et renvoyons à (Yvon, 2003; Stroppa & Yvon, 2005) pour la démonstration des principaux résultats qui sous-tendent cette construction. Pour débiter, nous introduisons les notions de *sous-mot complémentaire* et de *produit de mélange*.

#### Complémentarité

Si  $v$  est un sous-mot de  $x$ , on appelle langage complémentaire de  $v$  par rapport à  $x$ , noté  $v \setminus u$ , l'ensemble des sous-mot de  $v$  formés en supprimant les symboles présents dans  $x$ . Ainsi, par exemple,  $eeai$  est un sous-mot complémentaire de  $xmplr$  par rapport à  $xmplr$ . Lorsque  $u$  n'est pas un sous-mot de  $v$ ,  $v \setminus u$  est vide. Cette notion s'étend à des langages rationnels quelconques.

La relation de complémentarité de deux mots par rapport à  $w$  est une relation rationnelle ; le calcul des complémentaires de  $u$  par rapport à  $v$  s'effectue en prenant l'image de  $u$  par le transducteur à état fini  $T_w$ , dont la construction est illustrée sur la Figure 1.

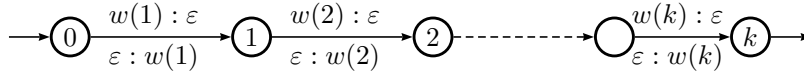


FIG. 1 – Automate calculant la relation de complémentarité par rapport  $w$

#### Mélange

Le *mélange*  $u \bullet v$  de deux mots de  $\Sigma^*$  est le langage défini, par exemple dans (Sakarovitch, 2003), comme suit :

$$u \bullet v = \{u_1 v_1 u_2 v_2 \dots u_n v_n, \text{ avec } u_i, v_i \in \Sigma^*, u = u_1 \dots u_n, w = v_1 \dots v_n\}$$

Le mélange de  $u$  et  $v$  contient tous les mots formés des symboles de  $u$  et de  $v$ , avec la contrainte que si  $a$  précède  $b$  dans  $u$  ou  $v$ , alors cet ordre est respecté dans  $u \bullet v$ . Ainsi, par exemple, si l'on prend  $u = abc$  et  $v = def$ , alors les mots suivants :  $abcdef$ ,  $abdefc$ ,  $adbecf$  ... sont dans  $u \bullet v$  ; ce n'est pas le cas de  $abefcd$ , dans lequel  $d$  suit  $e$ , alors qu'il devrait le précéder.

Le produit de mélange est également une opération rationnelle ; le mélange de deux mots est calculé en formant le produit des automates reconnaissant ces deux mots. Formellement, si  $K$  et  $L$  sont deux langages rationnels reconnus respectivement par  $A_K = (\Sigma, Q_K, q_K^0, F_K, \delta_K)$  et  $A_L = (\Sigma, Q_L, q_L^0, F_L, \delta_L)$ , avec  $A_K$  et  $A_L$  déterministes, l'automate  $A$  calculant  $K \bullet L$  se construit par :  $A = (\Sigma, Q_K \times Q_L, (q_K^0, q_L^0), F_K \times$

$F_L, \delta$ ), avec  $\delta$  définie par :  $\delta((q_K, q_L), a) = (r_K, r_L)$  si et seulement si soit  $\delta_K(q_K, a) = r_K$  et  $q_L = r_L$  soit  $\delta_L(q_L, a) = r_L$  et  $q_K = r_K$ .

Les notions de sous-mot et de mélange sont reliées par la relation suivante :

$$x \in u \bullet v \Leftrightarrow u \in x \setminus v$$

### Résolution

Ces notions étant posées, il est possible de réexprimer la notion de proportion analogique. Le résultat principal est énoncé par la proposition suivante (Yvon, 2003) :

#### Proposition 1

$$x : y :: z : t \Leftrightarrow x \bullet t \cap y \bullet z \neq \emptyset$$

L'intuition de cette proposition est que, pour que l'analogie soit établie, il faut non seulement que les symboles de  $x$  et  $t$  soient les mêmes que ceux de  $y$  et  $z$ , mais également que les symboles de  $x$  (et de  $t$ ) qui apparaissent dans  $y$  (et dans  $z$ ) conservent leur ordonnancement. Le corollaire suivant s'en déduit immédiatement :

#### Proposition 2

$$t \text{ est une solution de } x : y :: z : ? \Leftrightarrow t \in y \bullet z \setminus x$$

Ce résultat énonce que l'ensemble des solutions d'une équation analogique  $x : y :: z : ?$  est un ensemble rationnel, qui peut être calculé par un transducteur fini  $T$ . Divers résultats complémentaires sont établis dans (Yvon, 2003; Yvon *et al.*, 2004). En particulier, nous montrons que ce solveur analogique généralise l'approche fondée sur des distances d'édition proposée par (Lepage, 1998) et explorons diverses manières d'introduire une notion de gradualité dans les analogies (à l'aide du degré, mais également en considérant diverses autres valuation des proportions); nous montrons également comment généraliser ce résultat à mots sur un alphabet  $A$  qui est lui-même muni d'une structure algébrique, donnant lieu, par exemple, à des analogies entre séquences d'ensembles ou de structures de traits. Une étude de l'implantation efficace de solveurs analogiques exploitant ce formalisme, décrivant l'optimisation du calcul de la proportion analogique de degré minimum entre quatre termes, est présentée dans (Stroppa & Yvon, 2005).

## 3.4 Arbres

Le cas des arbres est plus problématique, l'ensemble des arbres n'étant pas naturellement muni d'une loi de composition interne. Pour pouvoir conserver l'idée générale de décomposition d'un objet sous forme d'une factorisation, nous opérons un détour par la notion de substitution. Les arbres considérés sont étiquetés.

#### Définition 6 (Substitution)

Une substitution est un ensemble fini de couples (variable, arbre). Appliquer la substitution  $\theta = \{v_1 \leftarrow t_1, \dots, v_n \leftarrow t_n\}$  à un arbre  $t$  consiste à remplacer chaque feuille de  $t$  étiquetée par  $v_i$  par l'arbre  $t_i$ . Le résultat de cette application est noté  $t\theta$ .

La composition de substitutions étant associative, l'ensemble des substitutions est un monoïde pour la composition. Il est alors possible de considérer des factorisations de substitutions, conduisant à la définition suivante de l'analogie entre arbres :

**Définition 7 (Proportion analogique (entre arbres))**

$(x, y, z, t) \in U$  constituent une proportion analogique, notée  $x : y :: z : t$  si et seulement s'il existe des factorisations de substitutions  $\theta_{x_1} \dots \theta_{x_n}$ ,  $\theta_{y_1} \dots \theta_{y_n}$ ,  $\theta_{z_1} \dots \theta_{z_n}$ ,  $\theta_{t_1} \dots \theta_{t_n}$  respectivement de  $\theta_x$ ,  $\theta_y$ ,  $\theta_z$  et  $\theta_t$  telles que  $\forall i, (\theta_{y_i}, \theta_{z_i}) \in \{(\theta_{x_i}, \theta_{t_i}), (\theta_{t_i}, \theta_{x_i})\}$  avec  $x = a\theta_x$ ,  $y = a\theta_y$ ,  $z = a\theta_z$ , et  $t = a\theta_t$  (où  $a$  est un arbre quelconque). Le plus petit entier  $n$  pour lequel de telles factorisations existent est le degré de l'analogie.

### 3.4.1 Approximation par linéarisation

La définition de l'analogie entre arbres proposée est cohérente avec le schéma général de proportion analogique, mais ne donne pas lieu directement à une procédure de calcul. Notre procédure de résolution d'équations analogiques entre arbres pré-suppose de disposer d'une fonction de *linéarisation* qui associe de manière bi-univoque un arbre et sa représentation linéarisée. On effectue alors une conversion des arbres en chaînes, avant de résoudre l'équation entre les représentations linéaires (cf. Section 3.3.2), pour construire enfin une représentation arborée. Cette approche est similaire à celle de (Lepage, 1999b) ; elle est justifiée par le résultat suivant qui vaut pour les linéarisations les plus usuelles, en particulier pour celle qui consiste à représenter un arbre sous forme d'une expression parenthésée (Stroppa & Yvon, 2005) :

**Proposition 3**

Si  $x : y :: z : t$ , alors  $l(x) : l(y) :: l(z) : l(t)$  pour  $l$  une linéarisation appropriée.

Le calcul d'une proportion analogique entre arbres est ainsi ramené à celui d'une proportion entre chaînes, dont on sait qu'il implique uniquement des opérations rationnelles, ce qui permet de conclure sur la complexité de la procédure. Toutefois, l'implication inverse dans la proposition n'étant pas vérifiée, cette méthode correspond uniquement à une approximation.

## 4 Quelques expérimentations

Dans cette section, nous présentons une application qui permet d'illustrer l'utilisation de proportions analogiques dans un cadre d'apprentissage. Cette application consiste à construire l'analyse morphologique d'une forme inconnue à partir d'une base lexicale. Après avoir introduit cette application et les bénéfices d'une approche à base d'analogie, nous décrivons les lexiques français et anglais utilisés et le protocole expérimental. Les résultats de ces expérimentations sont ensuite présentés et discutés.

## 4.1 Analyser des formes inconnues

### 4.1.1 Des formes inconnues

Les formes graphiques inconnues constituent une réalité incontournable pour qui s'intéresse au traitement automatique des langues : en dépit de la disponibilité de dictionnaires large couverture (e.g. Multext (Ide & Véronis, 1994)), ces formes inconnues continuent de représenter une majorité des types rencontrés dans les corpus journalistiques ou collectés sur la toile. Même en faisant abstraction des noms propres, dates et montants, qui fournissent les gros bataillons de formes inconnues, il subsiste un volant significatif de formes qui relèvent des catégories lexicales ouvertes, principalement noms, verbes, adjectifs et adverbes, et qui pour une large part sont construites par des procédés morphologiques réguliers (flexion, dérivation ou composition).

La caractérisation (analyse) de ces inconnus est pourtant de première importance pour de nombreux outils et applications pratiques, cette caractérisation pouvant, suivant les circonstances, prendre des modalités variables :

- regroupement de formes d'un même lexème pour des tâches d'indexation ou de fouille de textes (Porter, 1980; Gaussier, 1999; Dal & Namer, 2000) ;
- assignation d'une ou de plusieurs catégories et descriptions morpho-syntaxiques, pour des tâches d'étiquetage morpho-syntaxique (Brill, 1994; Mikheev, 1997) ;
- détection de la structure interne et des marques flexionnelles, par exemple, pour des tâches de prononciation automatique en synthèse ou en reconnaissance vocale (Yvon, 1996).

De nombreuses stratégies existent pour faire face à ce problème, consistant en premier lieu à étendre les dictionnaires existants, mais également à développer des systèmes à base de règles pour capturer les constructions les plus régulières : pour le français, citons en particulier le système INTEX (Sylberztein, 1993) et l'analyseur Flemm (Namer, 2000)). Cette approche se heurte toutefois à la disponibilité de descriptions suffisamment larges et précises de la morphologie : les phénomènes flexionnels sont bien décrits, ce n'est pas encore le cas des autres phénomènes, même pour des langues bien étudiées comme le français et l'anglais, justifiant le recours à des techniques d'apprentissage.

### 4.1.2 Apprendre la morphologie

Compte-tenu des besoins applicatifs évoqués ci-dessus, l'apprentissage automatique de régularités morphologiques, visant à analyser automatiquement des formes inconnues, a fait l'objet de multiples études. En plus des travaux cités précédemment, mentionnons, par exemple, (Krovetz, 1993; van den Bosch & Daelemans, 1999). En parallèle, s'est progressivement accumulée une importante littérature sur l'apprentissage non-supervisé de connaissances morphologiques à partir de corpus, avec des ambitions à la fois théoriques et applicatives, voir, par exemple, (de Marcken, 1996; Yarowsky & Wicentowski, 2000; Goldsmith, 2001; Schone & Jurafsky, 2001). Ces approches, pour l'essentiel, partagent un modèle théorique commun, dans lequel les formes sont construites par concaténation d'unités minimales morphématiques : l'apprentissage vise alors à inférer des collections de morphèmes et des procédures de segmentation.

Le modèle d'apprentissage proposé ici se démarque de ces approches et est, d'une

certaine manière, agnostique vis-à-vis de la théorie morphologique sous-jacente : la construction d'analogies sur des représentations lexicales intégrant une décomposition linéaire (mot) ou hiérarchique (arbre), il nous est possible de proposer des analyses morphématiques de formes inconnues. La construction d'analogie sur des structures attribuées dans lesquelles les formes sont non-analysées (sans structure interne) nous rapproche des modèles de la morphologie lexématique, à l'instar des travaux présentés notamment dans (Pirrelli & Yvon, 1999a; Lepage, 1999a). Ces deux types d'expérimentations sont présentées dans les sections qui suivent.

## 4.2 Données et protocole expérimental

Le protocole d'apprentissage utilisée est commun aux deux types d'expérimentations, la différence se situant dans la définition des espaces d'entrée et de sortie de la tâche. Chaque entrée des lexiques utilisés est constituée d'une forme graphique, d'un lemme et d'une analyse. Le lexique utilisé pour le français est issu du projet Multext (Ide & Véronis, 1994) ; pour l'anglais, il s'agit de CELEX (Burnage, 1990).

Le seul traitement opéré lors de l'apprentissage consiste à regrouper les lexèmes connus par famille flexionnelle ou dérivationnelle, afin de garantir qu'au moins deux termes entrant dans une proportion appartiennent à la même famille. L'analyse  $t^-$  d'un lexème inconnu  $t$  est inférée comme suit :  $n$  familles  $F_i$  sont tirées au hasard<sup>1</sup> ; pour chaque tirage, on construit toutes les analogies formelles  $x^+ : y^+ :: z^+ : t^+$  impliquant  $x$  et  $y$  dans  $F_i$  et  $z$  dans le lexique. Chaque triplet  $(x, y, z)$  ainsi construit donne lieu à une analyse possible pour  $t$ , par résolution de  $x^- : y^- :: z^- : t^-$ . Le candidat retenu est celui qui est majoritairement proposé.

Pour estimer le rappel et la précision de la méthode, 10 bases de tests sont constituées en tirant aléatoirement 1000 entrées du lexique. Une sortie est jugée correcte si elle correspond complètement à l'entrée du lexique<sup>2</sup>. Les résultats donnés sont les moyennes sur ces 10 tests.

### 4.2.1 Analyse lexématique

Chaque entrée du lexique est représentée sous la forme d'un vecteur multi-dimensionnel constitué de la forme orthographique, de son lemme et d'un ensemble de traits<sup>3</sup> morpho-syntaxiques spécifiant la catégorie, le genre, le nombre, la personne, le temps, etc. Quand un trait n'est pas pertinent pour une catégorie (eg. le temps pour les noms), il prend conventionnellement la valeur '-'. Ainsi, pour le lexique anglais, l'entrée :

```
replying for:reply for:V-pp--
```

exprime que `replying for` est une forme verbale (V) au participe (p) présent (p) dont le lemme est `reply for`. L'espace d'entrée de la tâche contient la seule forme

<sup>1</sup>Les expérimentations ont été effectuées avec  $n = 50, 100, 200$ .

<sup>2</sup>Cette mesure est la plus « sévère » possible : d'une part il existe des cas d'homographie, qui donnent lieu à des erreurs certaines ; pour les arbres, une mesure plus réaliste serait de comptabiliser le nombre de constituants bien identifiés.

<sup>3</sup>19 pour les entrées du lexique français, 6 pour le lexique anglais.

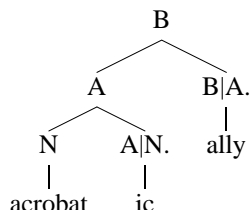
orthographique, les autres attributs définissant l'espace de sortie<sup>4</sup>. Les lexiques utilisés contiennent respectivement 288 000 formes (Multext) et 160 000 formes (CELEX).

### 4.2.2 Structure morphématique

Chaque entrée du lexique dérivationnel<sup>5</sup> de CELEX (43 582 entrées) contient trois attributs : le lemme, la racine, et une décomposition hiérarchique. Cette dernière exprime sous forme arborescente la structure interne de la forme, décomposée en unités morphématiques. Ainsi, le lemme *acrobatically* est représenté par :

```
acrobatically:acrobat:(((acrobat)[N],(ic)[A|N.])[A],(ally)[B|A.])[B]
```

Sa décomposition hiérarchique correspond à l'arbre :



## 4.3 Résultats

### 4.3.1 Analyse lexématique

Les tables 1 et 2 donnent les performances obtenues, en termes de rappel<sup>6</sup> et de précision respectivement pour les lexiques français et anglais. En première analyse, le rappel varie avec le nombre d'éléments dans les familles flexionnelles : très bon pour les verbes français, dont la conjugaison implique des dizaines de formes, faible pour les adjectifs anglais, qui sont invariables en nombre et pour lesquels les rares « flexions » correspondent à des comparatifs et des superlatifs. Pour pallier ce problème, nous envisageons de reproduire ces tests en regroupant les formes par famille *dérivationnelle*. Des résultats préliminaires montrent que ceci améliore nettement le rappel.

Dans presque tous les cas, la précision approche ou dépasse les 95%, s'approchant des 100% pour plusieurs catégories. Le différentiel de précision entre verbes et noms résulte de l'utilisation d'un principe du vote majoritaire, qui favorise les analyses prédisant un lemme verbal, du fait de la surreprésentation des formes verbales dans les deux lexiques.

La figure 2 représente les performances (résumées par le *F*-score) en faisant varier les paramètres utilisés pendant la phase de recherche : degré maximal des analogies considérées (entre 2 et 4) et nombre de tirages aléatoires :  $n \in \{50, 100, 200\}$ . Dans les deux cas, on note une amélioration sensible des performances. S'il est probable que

<sup>4</sup>D'autres protocoles étaient envisageable, consistant à informer progressivement les attributs définissant l'espace de sortie : calculer d'abord la catégorie, celle-ci étant connue, calculer le lemme... Notons également qu'en inversant le rôle des entrées et des sorties, l'analyseur se mue en générateur.

<sup>5</sup>Faute de données pour le français, cette expérimentation ne concerne que l'anglais.

<sup>6</sup>Rappelons que l'inférence analogique ne garantit pas un rappel de 100% : il arrive qu'aucune proportion formelle ne puisse être construite.

	Lemmes		Traits	
	Rappel	Précision	Rappel	Précision
Noms	0.87	0.94	0.90	0.97
Verbes	0.99	0.99	0.99	0.99
Adjectifs	0.97	0.98	0.98	0.99

TAB. 1 – Analyse flexionnelle du français

	Lemmes		Traits	
	Rappel	Précision	Rappel	Précision
Noms	0.77	0.95	0.77	0.95
Verbes	0.98	0.99	0.98	0.99
Adjectifs	0.25	0.70	0.25	0.70

TAB. 2 – Analyse flexionnelle de l'anglais

cette amélioration se confirme pour des valeurs encore supérieures de  $n$ , augmenter davantage le degré semble inutile.

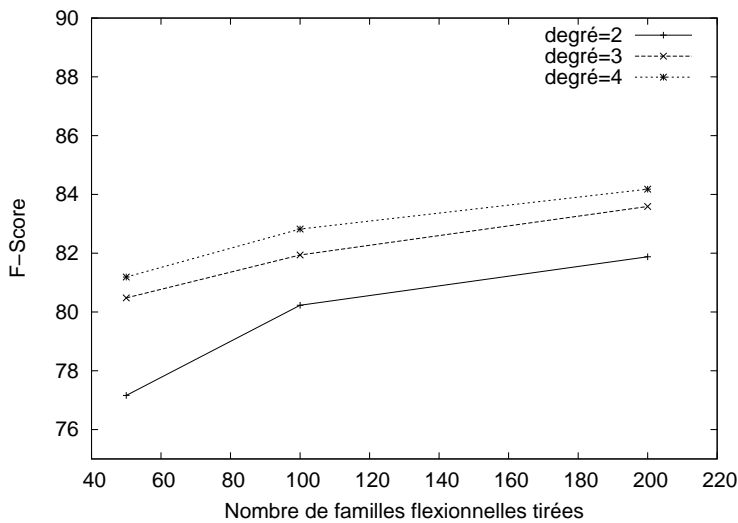


FIG. 2 – Influence des paramètres d'étape de recherche : lexique anglais

#### 4.3.2 Analyse hiérarchique

Cette seconde expérience porte sur la reconstruction des structures hiérarchiques de lemmes et donne un rappel de 30% pour une précision de 71%. Si le rappel est relativement faible, il convient de rappeler que tous les lemmes considérés ne sont pas analysables morphologiquement et qu'il est donc tout à fait fondé de ne pas propo-

ser, dans ces cas, d'analyse. La table 3 permet d'apprécier cet effet, en contrastant les performances pour trois types de lemmes : lorsque les formes sont effectivement des dérivés morphologiques, le rappel approche 60% et la précision 80%. En revanche, pour les composés, le système est presque toujours silencieux ; c'est encore le cas, dans une mesure moindre, quand le lemme résulte à la fois d'une dérivation et d'une composition.

	Décomposition hiérarchique	
	Rappel	Précision
Dérivé	0.58	0.78
Composé	0.01	0.45
Dérivé et Composé	0.14	0.69

TAB. 3 – Précision et rappel par type de dérivé

Une autre raison de la faiblesse du rappel est la très forte contrainte utilisée pendant l'étape de recherche, qui ne considère que des analogies de degré au plus trois. Des expérimentations complémentaires sont en cours, utilisant une stratégie de recherche élargie. En ce qui concerne la précision, elle atteint des niveaux corrects, d'autant plus que dans de nombreux cas d'erreurs, la solution proposée est partiellement correcte : ceci reste toutefois à quantifier précisément.

## 5 Discussion et Conclusion

### 5.1 Discussion

Ces expériences ont permis de confirmer la pertinence d'une approche à base d'analogie pour capturer les régularités présentes dans la morphologie de langues telles que le français et l'anglais, confirmant les résultats présentés eg. dans (Lepage, 1999a; Pirrelli & Yvon, 1999a). Pour mieux apprécier les performances de cette méthode, il reste à conduire des expériences sur des formes réellement inconnues. Travailler sur des lexiques biaise doublement les résultats : d'un côté, la tâche de l'analyseur est compliquée par l'analyse de formes complètement idiosyncrasiques (ie. les formes du verbe *être*) ; à l'inverse, cela garantit que pour de nombreuses formes testées, une forme morphologiquement apparentée existe dans l'ensemble d'apprentissage, ce qui augmente le rappel. De telles expériences sont actuellement en cours sur des corpus journalistiques.

Les résultats obtenus sur la tâche de calcul de la structure d'une forme inconnue, bien que prometteurs, restent en deçà des attentes, aussi bien en terme de rappel que de précision. Parmi les voies d'amélioration de nos algorithmes, nous envisageons en premier lieu de travailler sur l'étape de recherche. Le calcul exhaustif de l'ensemble des analogies dérivables d'un lexique reste computationnellement hors de portée : une piste consiste à utiliser, en plus du degré, d'autres types de valuations des analogies, puis à exploiter ces valuations pour construire de nouvelles heuristiques de recherche.

## 5.2 Conclusion

Dans cet article, nous avons présenté un modèle général d'inférence par analogie, fondé sur l'exploitation de relations formelles de proportionnalité entre objets décrits par des attributs prenant des formes variées. Nous avons également montré comment calculer ces proportions, en étudiant les cas des structures de traits, des ensembles, des mots sur un alphabet fini et des arbres. Ce modèle a été implanté dans une bibliothèque générique d'inférence analogique, et testé sur une tâche d'analyse morphologique de formes inconnues. Les résultats obtenus ont mis en évidence la bonne adéquation de ce modèle à la tâche considérée, ainsi que la nécessité d'améliorer le taux de rappel.

## Références

- BRILL E. (1994). Some advances in rule based part of speech tagging. In *Proceedings of AAAI*.
- BURNAGE G. (1990). *CELEX : A Guide for Users*. Rapport interne, University of Nijmegen, Center for Lexical Information, Nijmegen.
- DAELEMANS W., BOSCH A. V. D. & ZAVREL J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning*, **34**, 11.
- DAL G. & NAMER F. (2000). Génération et analyse automatique de ressources lexicales construites utilisables en recherche d'information. *Traitement Automatique des Langues*, **47**(2), 423–445.
- DE MARCKEN C. (1996). *Unsupervised Language Acquisition*. PhD thesis, Dpt of Computer Science, MIT.
- FALKENHEIMER B. & GENTNER D. (1986). The structure-mapping engine. In *Proceedings of the meeting of the American Association for Artificial Intelligence (AAAI)*, p. 272–277.
- FRADIN B. (2003). *Nouvelles approches en morphologie*. Paris, France : Presses Universitaires de France.
- GAUSSIER E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the ACL Workshop on Unsupervised Methods in Natural Language Processing*, College Park, MD.
- D. GENTNER, K. J. HOLYOAK & B. N. KONIKOV, Eds. (2001). *The analogical mind*. Cambridge, MA : The MIT Press.
- GOLDSMITH J. (2001). Unsupervised learning of the morphology of natural languages. *Computational Linguistics*, **27**(2), 153–198.
- HOFSTADTER D. & MITCHELL M. (1995). *The copycat project : A model of mental fluidity and analogy-making*, In D. HOFSTADTER & THE FLUID ANALOGIES RESEARCH GROUP, Eds., *Fluid Concepts and Creative Analogies*, chapter 5, p. 205–267. Basic Books.
- IDE N. & VÉRONIS J. (1994). MULTTEXT (Multilingual Tools and Corpora). In *Proceedings of the 14th COLING*, p. 588–592, Kyoto, Japan.
- JONES D. (1996). *Analogical Natural Language Processing*. London : UCL Press.
- KROVETZ B. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM-SIGIR*, p. 191–202.
- LEPAGE Y. (1998). Solving analogies on words : An algorithm. In *Proceedings of COLING-ACL '98*, volume 2, p. 728–735, Montréal, Canada.

- LEPAGE Y. (1999a). Analogy+tables=conjugation. In G. FRIEDL & H. MAYR, Eds., *Proceedings of NLDB'99*, p. 197–201, Klagenfurt, Germany.
- LEPAGE Y. (1999b). Open set experiments with direct analysis by analogy. In *Proceedings of NLPRS '99*, volume 2, p. 363–368, Beijing, China.
- LEPAGE Y. (2001). Analogy and formal language. *Electronic Notes in Theoretical Computer Science*, **47**, 1–12.
- LOMBARDY S., POSS R., RÉGIS-GIANAS Y. & SAKAROVITCH J. (2003). Introducing Vaucanson. In *Implementation and Application of Automata (CIAA 2003)*, p. 96–107.
- MATTHEWS P. (1974). *Morphology*. Cambridge : Cambridge University Press.
- MIKHEEV A. (1997). Automatic rule induction for unknown word guessing. *Computational Linguistics*, **23**(3), 405–423.
- MITCHELL T. M. (1997). *Machine Learning*. McGraw-Hill.
- NAMER F. (2000). Flemm : Un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, **41**(2), 523–547.
- PIRRELLI V. & YVON F. (1999a). Analogy in the lexicon : a probe into analogy-based machine learning of language. In *Proceedings of the 6th International Symposium on Human Communication*, Santiago de Cuba, Cuba.
- PIRRELLI V. & YVON F. (1999b). The hidden dimension : paradigmatic approaches to data-driven natural language processing. *Journal of Experimental and Theoretical Artificial Intelligence, Special Issue on Memory-Based Language Processing*, **11**, 391–408.
- PLATE T. A. (2000). Analogy retrieval and processing with distributed vector representations. *Expert systems*, **17**(1), 29–40.
- PORTER M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.
- SAKAROVITCH J. (2003). *Éléments de théorie des automates*. Vuibert, Paris.
- SCHONE P. & JURAFSKY D. (2001). Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*, Pittsburgh, PA.
- SKOUSEN R. (1989). *Analogical Modeling of Language*. Dordrecht : Kluwer.
- STROPPA N. & YVON F. (2005). *Formal models of analogical relationships*. Rapport interne, à paraître, ENST, Paris, France.
- SYLBERZTEIN M. (1993). *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson, Paris.
- THAGARD P., HOLOYAK K. J., NELSON G. & GOCHFELD D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, **46**(3), 259–310.
- VAN DEN BOSCH A. & DAELEMANS W. (1999). Memory-based morphological processing. In *Proceedings of ACL*, p. 285–292, Maryland.
- YAROWSKY D. & WICENTOWSKI R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL*, p. 207–216, Hong Kong.
- YVON F. (1996). *Prononcer par analogie : motivations, formalisations et évaluations*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris.
- YVON F. (1999). Pronouncing unknown words using multi-dimensional analogies. In *Proceedings of Eurospeech*, volume 1, p. 199–202, Budapest, Hungary.
- YVON F. (2003). *Finite-state machines solving analogies on words*. Rapport interne, ENST.
- YVON F., STROPPA N., DELHAY A. & MICLET L. (2004). *Solving analogies on words*. Rapport interne D005, Ecole Nationale Supérieure des Télécommunications, Paris, France.