TALN 2004, Session Poster, Fès, 19–21 avril 2004

Analogies dans les séquences : un solveur à états finis

Nicolas Stroppa et François Yvon
GET/ENST et LTCI, CNRS UMR 5141
46 rue Barrault - F-75 013 Paris
{stroppa,yvon}@enst.fr

Résumé - Abstract

L'apprentissage par analogie se fonde sur un principe inférentiel potentiellement pertinent pour le traitement des langues naturelles. L'utilisation de ce principe pour des tâches d'analyse linguistique présuppose toutefois une définition formelle de l'analogie entre séquences. Dans cet article, nous proposons une telle définition et montrons qu'elle donne lieu à l'implantation efficace d'un solveur d'équations analogiques sous la forme d'un transducteur fini. Munis de ces résultats, nous caractérisons empiriquement l'extension analogique de divers langages finis, correspondant à des dictionnaires de quatre langues.

Analogical reasoning provides us with an inferential mecanism of potential interest for NLP applications. An effective use of this process requires a formal definition of the notion of an analogy between strings of symbols. In this paper, we propose such a definition, from which we derive the implementation of a finite-state transducer solving analogical equations on sequences. We finally present the results of an empirical study of the analogical extension of several finite languages, corresponding to dictionnaries of four European languages.

Mots-clefs – Keywords

Apprentissage par Analogie, Automates finis
Analogical Learning, Finite-State Automaton

1 Introduction

1.1 Analogie et TAL

L'apprentissage par analogie est une méthode d'inférence qui offre une alternative à la fois aux méthodes à base de connaissance et aux modélisations probabilistes. Dans cette méthode, la connaissance linguistique n'est pas abstraite sous forme de règles ou de modèle stochastique ; à l'inverse, elle demeure *implicitement représentée* dans le corpus (Daelemans, 1996), la généralisation à de nouvelles données s'effectuant toujours dans un contexte défini (Lepage, 1999). Dans sa version "canonique", l'analyse par analogie d'un objet x lui associe une représentation $f(x)$ en deux temps :

– identification d'un rapport de proportionnalité impliquant x et 3 objets déjà connus, b, c, et d ; on note cette relation $x : b :: c : d$, signifiant que x *est à* b *ce que* c *est à* d ;
– construction d'une analyse $f(x)$ de x par *transfert analogique*, consistant à trouver $f(x)$ tel que : $f(x) : f(b) :: f(c) : f(d)$.

Ce principe est intéressant à plus d'un titre : en premier lieu, il est régulièrement utilisé pour décrire un certain nombre de phénomènes cognitifs ; dans le cadre de la linguistique traditionnelle, il est invoqué pour rendre compte de régularités linguistiques, en particulier morphologiques (Matthews, 1972) ; son utilisation en traitement automatique des langues a donné des résultats encourageants pour un certain nombre de tâches telles que : l'analyse et la génération morphologique (Lepage, 1999; Pirrelli & Yvon, 1999a; Tanguy & Hathout, 2002) ; la conversion orthographique-phonétique (Yvon, 1997; Yvon, 1999), ou encore la désambiguïsation sémantique (Federici *et al.*, 1997). Ces expériences ont également démontré que l'induction analogique permet de manipuler directement des données linguistiques sous une forme plus "naturelle" (Pirrelli & Yvon, 1999b; Hathout, 2002). Enfin, elle se révèle fournir un cadre permettant d'intégrer des connaissances hétérogènes issues de sources différentes.

1.2 Analogie dans les séquences

Utiliser l'analogie en TAL présuppose toutefois de définir cette notion pour des séquences symboliques, puisque telle est la forme sous laquelle se présentent le plus directement les objets linguistiques. Pourtant, l'analogie entre séquences n'a pas fait l'objet d'une définition formelle (voir toutefois (Lepage, 1998; Lepage, 2001) pour l'étude de diverses formalisations et de leurs propriétés). La mise en œuvre de stratégies d'inférence analogique pose également plusieurs problèmes théoriques et pratiques :

– la construction de tous les triplets d'objets analogues à x : une procédure naïve examine tous les quadruplets (x, b, c, d), soit une complexité cubique qui est excessive quand on manipule des bases d'exemples réelles, contenant des milliers d'éléments ;
– lorsqu'une équation analogique $a : b :: c :?$ a plusieurs solutions, il est nécessaire de disposer de critères exprimant des préférences sur les ensembles de solutions ;
– inversement, la recherche de triplets analogues peut échouer, entraînant l'échec du calcul de $f(x)$. Pour faire face à ces situations, il est nécessaire de disposer de notions plus affaiblies de l'analogie, qui permettent de garantir la construction d'au moins une analyse de x.

Dans cet article, nous nous proposons de fournir des éléments de réponse aux deux premières de ces questions. Dans un premier temps (Section 2), nous proposons une définition de l'analogie sur les séquences, et nous montrons (Section 3) que cette définition permet (i) de construire des transducteurs finis qui résolvent les analogies ; (ii) de calculer efficacement tous les quadru-

plets analogiques impliquant un élément particulier ; (iii) de définir et calculer efficacement une mesure quantitative, appelée degré, de la qualité d'une analogie.

Dans un second temps (Section 4), nous présentons les résultats d'une série d'expérimentations conduites sur des lexiques de plusieurs langues. Ces expériences visent d'une part à caractériser empiriquement l'extension analogique d'un langage ; d'autre part à valider l'hypothèse que nos algorithmes sont potentiellement utiles dans un contexte d'apprentissage automatique.

2 Analogies dans les séquences

2.1 Notations

Soit Σ un vocabulaire fini, $\Sigma^\star$ est l'ensemble des séquences finies de symboles de Σ (des *mots*) et ϵ le mot vide. Un mot x, de longueur $|\,x\,|$, se décompose de façon unique selon $x = x_1^{|x|}$ ou encore $x = x_1 \dots x_{|x|}$, avec $x_i \in \Sigma$. Classiquement, xy dénote la concaténation des mots x et y. Si $x = uzv$, on appelle u un *préfixe* de x, v un *suffixe* de x, et z un *facteur* de x. Un *sous-mot* de $x = x_1 \dots x_n$ est un mot v tel que $\exists I_v = \{i_1 \dots i_k\}, 1 \leq i_1 < i_2 \dots i_k \leq n$ et $v = x_{i_1} \dots x_{i_k}$. Réciproquement, tout ensemble ordonné I d'indices de $\{1 \dots \ |\,x\,|\ \}$ définit un sous-mot $x(I)$ de x. Si v est un sous-mot de x, nous noterons : $v \in x$.

Un automate fini A est un 5-uplet $(\Sigma, Q, q^0, F, \delta)$, avec Σ le vocabulaire, Q un ensemble fini d'états, $q^0 \in Q$ l'état initial, $F \subset Q$ l'ensemble des états finals et δ la fonction de transition associant des couples de $\Sigma \times Q$ avec des parties de Q. Le langage accepté par A est noté $L(A)$; si $\delta^\star$ dénote la fermeture transitive de δ il vient $L(A) = \{w, \delta^\star(q^0, w) \cap F \neq \emptyset\}$. Un transducteur fini est un automate fini comportant deux rubans : un pour l'entrée et un pour la sortie. En conséquence, ses transitions sont étiquetées par des paires $a : b$, avec a dans l'alphabet d'entrée Σ_1 et b dans l'alphabet de sortie Σ_2. Une présentation détaillée des transducteurs finis et des principaux algorithmes s'y rapportant est donnée dans (Roche & Schabes, 1997).

2.2 Analogies et équations analogiques

Dans cette section, nous donnons une définition de la notion d'analogie entre mots, de laquelle nous déduisons diverses propriétés classiques.

Définition 1 $(x, y, z, t) \in \Sigma^+$ *constituent une proportion analogique, notée* $x : y :: z : t$ *ssi* $\exists n > 0, \alpha_i, i = 1 \dots n, \beta_i, i = 1 \dots n \in \Sigma^\star$ *tels que :*

$$\begin{array}{llll} \textit{soit} & x = \alpha_1 \dots \alpha_n & t = \beta_1 \dots \beta_n & y = \alpha_1\beta_2\alpha_3 \dots \quad z = \beta_1\alpha_2\beta_3 \dots \\ \textit{soit} & x = \alpha_1 \dots \alpha_n & t = \beta_1 \dots \beta_n & y = \beta_1\alpha_2\beta_3 \dots \quad z = \alpha_1\beta_2\alpha_3 \dots \end{array}$$

avec $\forall i, \alpha_i\beta_i \neq \epsilon$. *Le plus petit* n *pour lequel une telle relation existe est le* degré *de l'analogie.*

Un exemple de proportion vérifiant cette définition est :

$$subjectif : subversif :: injection : inversion$$

avec $n = 3$ et les facteurs suivants : $\alpha_1 = sub$, $\alpha_2 = ject$, $\alpha_3 = if$, $\beta_1 = in$, $\beta_2 = vers$, $\beta_3 = ion$[1]. Le degré de cette proportion est donc 3. D'une manière générale, le degré mesure la qualité intrinsèque d'une proportion : plus il est faible, meilleure est l'analogie.

De cette définition se déduisent aisément les propriétés classiques de l'analogie (voir e.g. (Lepage, 2001)) :

$$\forall x \in \Sigma^+, x : x :: x : x \quad (1)$$

$$\forall x, y \in \Sigma^+ : x : x :: y : y \quad (2)$$

$$\forall x, y, z, t \in \Sigma^+ : x : y :: z : t \Rightarrow z : t :: x : y \quad (3)$$

$$\forall x, y, z, t \in \Sigma^+ : x : y :: z : t \Rightarrow x : z :: y : t \quad (4)$$

Une équation analogique est une proportion analogique pour laquelle seuls trois termes sur quatre sont connus ; le calcul du terme manquant correspondant à la résolution de l'équation.

Définition 2 *t est une solution de l'équation analogique $x : y :: z :?$ ssi $x : y :: z : t$.*

La définition (1) ne permet de garantir ni qu'une équation a toujours une solution, ni inversement l'unicité d'une solution. Ainsi, par exemple, $abc : def :: ijk :?$ ne peut être résolue ; alors que $c : ac :: bc :?$ a deux solutions : abc et bac. (Lepage, 2001) donne 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 Un solveur à états finis

3.1 Mot complémentaire et mélange

Mots complémentaires Nous commençons par introduire la notion de mots complémentaires. Si v est un sous-mot de x, on appelle complémentaire de v par rapport à x le sous-mot formé en conservant uniquement les symboles de x qui ne sont pas dans v. Ainsi, par exemple $eeaie$ est le complémentaire (vocalique) de $xmplr$ par rapport à $exemplaire$. Formellement :

Définition 3 *Si v est un sous-mot de x, correspondant à l'ensemble d'indices I_v, l'ensemble des sous-mot complémentaires de v par rapport à x est défini par $x \backslash v = \{y, y \in x, v = x(\{0 \dots | x| \} \backslash I_y)\}$. Si v n'est pas un sous-mot de x, $x \backslash v$ est vide.*

À chaque mot de $\Sigma^\star$ on associe alors une relation binaire sur $\Sigma^\star \times \Sigma^\star$, notée $\backslash_w$, et définie par :

Définition 4 *$u \backslash_w v$ si et seulement si $u \in w \backslash v$.*

Pour tout mot x de $\Sigma^\star$, il est aisé de construire un automate fini A_x qui reconnaît uniquement x en établissant une bijection entre les états de A_x et les préfixes de x. En ajoutant à chaque

[1] On peut tout aussi bien prendre $\alpha_2 = jecti$, $\alpha_3 = f$...

transition de A_x une transition spontanée, on dérive un automate S_x qui reconnaît exactement les sous-mots de x. En transformant S_x en un transducteur T_x tel que chaque transition de S_x étiquetée par x_i produit en sortie ϵ et chaque transition ϵ produit x_i, on obtient finalement une machine calculant la relation de complémentation par rapport à x. La relation de complémentarité est donc une relation rationnelle. Ces constructions sont illustrées sur la Figure 1.

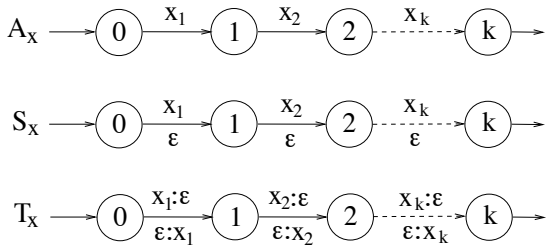


FIG. 1 – Automates et transducteurs pour x, ses sous-mots et la relation de complémentarité

On notera pour finir que les notions de sous-mots complémentaires et de relation de complémentarité s'étendent sans difficulté à des langages rationnels quelconques.

Mélange Le produit de mélange est défini par exemple dans (Sakarovitch, 2003) comme suit. Si u et v sont deux mots de $\Sigma^\star$, leur mélange $u \bullet v$ est le langage défini par :

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

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.

Cette opération s'étend naturellement aux langages, selon, pour les langages K et L :

$$K \bullet L = \bigcup_{x \in K, y \in L} x \bullet y$$

Il est connu que cette opération est une opération rationnelle, et que le mélange de deux mots est calculé en effectuant 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 δ 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-mots et de mélange sont reliées par la relation suivante :

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

qui énonce que u et v sont en relation complémentaire par rapport à x si et seulement si x appartient au mélange de ces deux mots.

3.2 Un solveur analogique

Ces notions préliminaires étant posées, il devient possible de ré-exprimer la notion de proportion analogique. Le résultat principal est énoncé par la proposition suivante :

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 aussi que les symboles de x (et aussi de t) apparaissant dans y et z apparaissent dans le même ordre. Ce résultat est établi formellement dans (Yvon, 2003), ainsi que le corollaire suivant :

Proposition 2

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

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. T est obtenu en construisant tout d'abord l'automate qui reconnnaît le mélange de y et z, duquel on dérive ensuite l'automate des complémentaires, par une méthode analogue à celle détaillée ci-dessus.

Deux résultats complémentaires sont également établis dans (Yvon, 2003) :
– le solveur analogique implanté sous la forme d'un transducteur fini généralise strictement l'approche fondée sur des distances d'édition proposée par (Lepage, 1998) ;
– le calcul du degré d'une analogie $x : y :: z : t$ est réalisé en "comptant" dans t le nombre de transitions entre fragments de y et de z ; pour réaliser directement ce calcul, il suffit de composer T avec un transducteur fini qui effectue un tel décompte.

Ces résultats théoriques permettent d'implanter efficacement le calcul de *l'extension analogique* d'un langage fini L, notée L^3, qui contient tous les mots entrant dans une relation de proportion analogique avec 3 mots de L. Dans la pratique, plutôt que calculer explicitement le transducteur $(L \bullet L) \backslash L$, qui contient un nombre d'états quadratique par rapport à l'automate A_L représentant L, il est plus efficace de le simuler, à partir d'une représentation de A_L.

4 Expérimentations

Dans cette section, nous présentons les résultats d'une première étude empirique de L^3, qui vise à comprendre le type de généralisation effectué par le principe analogique et à contrôler la validité de ce principe pour des tâches d'inférence. Nous cherchons principalement à répondre à deux questions : L^3 contient-il bien les mots "valides" absents de L ? L^3 ne surgénéralise-t-il pas abusivement L ? Outre leur intérêt théorique, ces questions sont également importantes dans l'optique d'utiliser notre modèle pour faire de l'apprentissage automatique. Avant en effet de chercher à transférer des relations analogiques existant dans un langage L vers un autre espace de représentation L', il convient de s'assurer au préalable que des séquences qui sont absentes de L, mais qui sont toutefois des séquences possibles, sont effectivement trouvées dans L^3.

En revanche, il est important de réaliser que ces expériences ne visent pas à tester un système d'analyse morphologique fondé sur l'apprentissage par analogie. La mise en place d'un tel système excèdant de loin le cadre de notre travail.

4.1 Estimation du rappel

L'inférence analogique est fondée si elle permet de généraliser correctement à de nouveaux exemples. Pour tester cette hypothèse pour des tâches d'analyse lexicale, nous allons montrer que l'analogie peut être utilisée comme une procédure effective de reconstruction de lexiques. Pour ce faire, nous utilisons différents lexiques et répétons l'expérience suivante : partitionner aléatoirement L en n sous-lexiques P_i de même taille et calculer le pourcentage moyen d'éléments de P_i qui appartiennent à l'extension analogique de $(L \setminus P_i)$, et qui sont donc recalculables à partir d'éléments de $(L \setminus P_i)$. La mesure obtenue, que nous appelons *densité analogique* s'apparente donc à une mesure de rappel, estimée ici par validation croisée à n blocs. La densité analogique est mesurée pour 11 lexiques, correspondant aux racines, aux lemmes et aux formes graphiques de quatre langues, le français, l'anglais, l'allemand et le néerlandais, avec $n = 10$. Les lexiques anglais, allemand et néerlandais proviennent du corpus Celex (Burnage, 1990), les lexiques français proviennent du corpus Multext (Ide & Véronis, 1994) (qui ne contient que lemmes et formes). Par ailleurs, pour l'ensemble des expérimentations, nous avons uniquement considéré les analogies ayant un degré inférieur ou égal à 2, ceci nous permettant d'obtenir plus rapidement une borne inférieure à la densité recherchée. Les résultats obtenus sont donnés dans le tableau 1. Le calcul de la densité d'un lexique contenant 300 000 entrées demande en moyenne 3 heures.

TAB. 1 – Densités analogiques de 4 langues (en %)

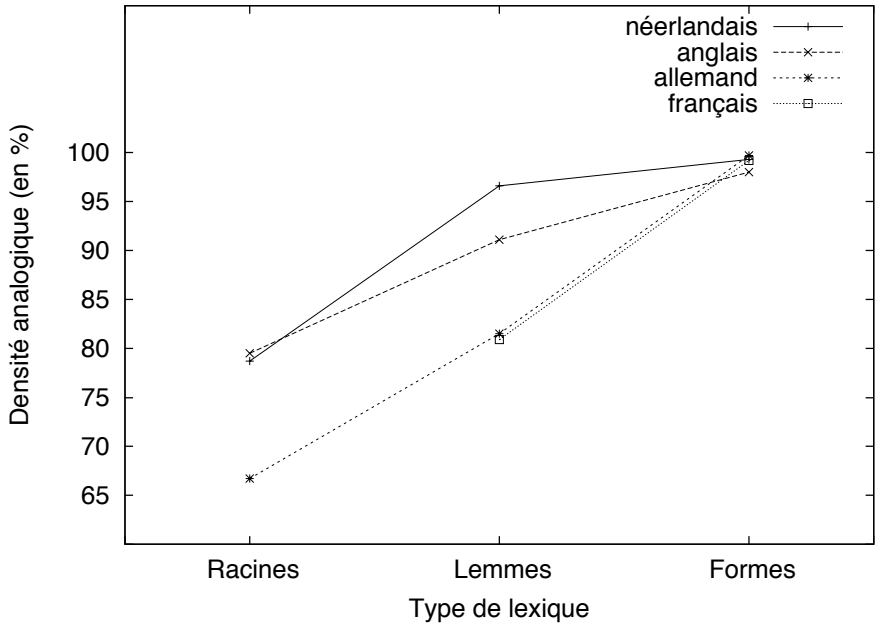
	Français	Anglais	Allemand	Néerlandais
Racines	-	79.5±3.1	66.7±4.7	78.7±3.5
Lemmes	80.9±3.9	91.1±3.4	81.5±3.4	96.6±1.4
Formes	99.2±0.6	98±1.4	99.7±0.5	99.3±0.8

Ces résultats permettent de faire deux constats. Premièrement, l'hypothèse selon laquelle le langage est fortement marqué par des régularités analogiques est vérifiée sur les quatre langues étudiées, comme en attestent des densités analogiques sur les lexiques de formes supérieures à 98%. Deuxièmement, on peut remarquer que les formes sont plus facilement reconstructibles par analogie que les lemmes, eux-mêmes plus facilement reconstructibles que les racines (Figure 2), ce qui met en évidence la capacité d'une modélisation analogique pour décrire des phénomènes de dérivation et de flexion morphologiques.

4.2 Estimation de la précision

Une seconde expérience vise à montrer que L^3 n'est pas surgénératif. En effet, il s'agit de vérifier non seulement que $(L \setminus P)^3$ contient une large partie de P, mais également qu'il ne contient que des éléments voisins de ceux de L. Le protocole expérimental consiste à considérer différents processus aléatoires de génération d'entrées lexicales modélisant de mieux en mieux L, pour vérifier que mieux on modélise L, plus on génère des mots de L^3. Un premier modèle, noté "Uniforme", génère un mot en (i) tirant une longueur l suivant un processus de Poisson ; (ii) tirant l lettres avec une probabilité uniforme. Deux modèles n-grammes sont également appris sur L, pour des valeurs de n valant 1 et 3. Chacun de ces deux modèles est utilisé pour générer 100 mots et l'on mesure la proportion de ces mots qui figurent dans L^3. La ligne "Réel" rappelle enfin les valeurs de densité calculées dans la section précédente. Les résultat obtenus sont rapportés dans le tableau 2.

FIG. 2 – Densités analogiques de 4 langues



Ces résultats vont tous dans le même sens : mieux le modèle de génération modélise L, plus il a une grande intersection avec L^3. Cela confirme notre hypothèse que L^3 n'est pas n'est pas composé d'éléments aléatoires, mais d'éléments "proches" de L. L'extension analogique d'un langage fournit donc une généralisation finie de L, qui inclut intégralement L ainsi que de nombreux mots "proches" de L.

5 Conclusions

Dans cet article, nous avons proposé une définition formelle de l'analogie entre séquences de symbole, de laquelle nous avons pu déduire une implantation d'un solveur d'équations analogiques sous la forme d'un transducteur fini. Ces résultats théoriques sont complétés par une étude empirique de l'extension analogique d'un langage, qui illustre, sur des données dictionnairiques, la validité de l'approche à base d'analogie.

Ce travail offre de nombreuses perspectives. Du point de vue théorique, nous envisageons principalement d'étendre ces modèles dans trois directions :

– définir, en utilisant les notions de mélange et de complémentation, des versions plus "faibles" de l'analogie entre séquences, permettant d'obtenir un meilleur rappel pour des tâches d'analyse de phrases ;
– définir des mesures plus riches que le degré de la qualité d'une analogie, intégrant en particulier des informations statistiques ou encore des contraintes linguistiques supplémentaires ;
– définir formellement l'analogie pour d'autres représentations classiquement utilisées en TAL, en particulier entre arbres, entre langages finis et entre graphes acycliques orientés.

Du point de vue empirique, nous envisageons dans un premier temps d'implanter un système complet et générique d'inférence analogique, qui sera évalué sur diverses tâches d'analyse lin-

TAB. 2 – Caractérisation de L^3 - densités en %

		Français	Anglais	Allemand	Néerlandais
Racines	Uniforme	-	1.0±0	7.7±0.9	1.9±0.3
	1-gramme	-	37.9±0.3	36.9±0.8	40.2±0.7
	3-gramme	-	66.6±1.0	63.4±0.8	69.7±0.8
	Réel	-	79.5±3.1	66.7±4.7	78.7±3.5
Lemmes	Uniforme	0±0	1.7±0.5	0±0	0.7±0.5
	1-gramme	34±0	38.2±1.0	18.6±0.7	25.9±0.9
	3-gramme	58.7±1.4	50.2±1.0	33.5±1.0	41.2±0
	Réel	80.9±3.9	91.1±3.4	81.5±3.3	96.6±1.3
Formes	Uniforme	1±0	0.0±0.0	0.9±0.3	1.0±0.0
	1-gramme	26.6±0.8	33.9±0.3	18.0±0.1	44.2±0.8
	3-gramme	61.5±0	53.0±0.6	57.7±1.3	65.3±2.6
	Réel	99.2±0.6	98.0±1.4	99.7±0.5	99.3±0.8

guistique.

Remerciements

Une partie de ce travail a été conduit dans le cadre du programme TCAN du CNRS. En particulier, il a bénéficié d'échanges fructueux avec L. Miclet, A. Delhaye de l'IRISA (Lannion) et de A. Cornuéjols du LRI (Orsay).

Références

BURNAGE G. (1990). *CELEX : A Guide for Users*. Rapport interne, University of Nijmegen, Center for Lexical Information, Nijmegen.

DAELEMANS W. (1996). Abstraction considered harmful : lazy learning of language processing. In H. J. V. DEN HERIK & A. WEIJTERS, Eds., *Proceedings of the sixth Belgian-Dutch Conference on Machine Learning*, p. 3–12, Maastricht, The Nederlands.

FEDERICI S., MONTEMAGNI S. & PIRRELLI V. (1997). Inferring semantic similarity from distributional evidence : an analogy-based approach to word sense disambiguation. In *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

HATHOUT N. (2002). From wordnet to celex : acquiring morphological links from dictionaries of synonyms. In *Proceedings of the Third Conference on Language, Resources and Evaluation, LREC'03*, Las Palmas de Gran Canaria, Espagne.

IDE N. & VÉRONIS J. (1994). MULTEXT (Multilingual Tools and Corpora). In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, p. 588–592, Kyoto, Japan.

LEPAGE Y. (1998). Solving analogies on words : An algorithm. In *Proceedings of COLING-ACL '98*, volume 2, p. 728–735, Montréal, Canada.

LEPAGE Y. (1999). Analogy+tables=conjugation. In G. FRIEDL & H. MAYR, Eds., *Proceedings of NLDB'99*, p. 197–201, Klagenfurt, Germany.

LEPAGE Y. (2001). Analogy and formal language. *Electronic Notes in Theoretical Computer Science*, **47**, 1–12.

MATTHEWS P. (1972). *Inflectional Morphology. A Theoretical Study Based on Aspects of Latin Verb Conjugation.* Cambridge : Cambridge University Press.

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.

ROCHE E. & SCHABES Y. (1997). Introduction to finite-state devices in natural language processing. In E. ROCHE & Y. SCHABES, Eds., *Finite State Natural Language Processing*, Cambridge, MA : The MIT Press.

SAKAROVITCH J. (2003). *Eléments de théorie des automates.* Vuibert, Paris.

TANGUY L. & HATHOUT N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In *Actes de la 9eme Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN'2002)*, Nancy, France.

YVON F. (1997). Paradigmatic cascades : a linguistically sound model of pronunciation by analogy. In *Proceedings of the 35th annual meeting of the ACL*, Madrid, Spain.

YVON F. (1999). Pronouncing unknown words using multi-dimensional analogies. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, p. 199–202, Budapest, Hungary.

YVON F. (2003). *Finite-state machines solving analogies on words.* Rapport interne, ENST, Paris.