Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie

Nicolas Stroppa* — **François Yvon****

* *National Centre for Language Technology, Dublin City University*
Dublin 9, Ireland
nstroppa@computing.dcu.ie

** *GET/ENST et LTCI, UMR 5141*
46, rue Barrault, F-75013 Paris
yvon@enst.fr

RÉSUMÉ. *Nous présentons un modèle d'apprentissage par analogie qui exploite la notion de proportions analogiques formelles ; cette approche présuppose de savoir donner un sens à ces proportions et de pouvoir implanter efficacement leur calcul. Nous proposons une définition algébrique de cette notion, valable pour les structures utilisées couramment pour les représentations linguistiques : mots sur un alphabet fini, structures attribut-valeur, arbres étiquetés. Nous présentons ensuite une application à une tâche concrète, consistant à apprendre à analyser morphologiquement des formes orthographiques inconnues. Des résultats expérimentaux sur plusieurs lexiques permettent d'apprécier la validité de notre démarche.*

ABSTRACT. *We present an analogical learning model which exploits the notion of formal analogical proportions; this approach requires the ability to define these proportions and to implement them efficiently. We propose an algebraic definition of this notion, which applies to a large range of structured objects: words, attribute-value structures, labeled trees. We test this inference model by applying it to the task of learning morphological analyses of unknown forms. We can assess the validity of the approach thanks to experimental results that are given for several lexicons.*

MOTS-CLÉS : *Apprentissage automatique, Raisonnement par analogie, Morphologie*

KEYWORDS: *Machine learning, Analogical reasoning, Morphology*

1. Introduction

Les outils du traitement automatique des langues ont été profondément renouvelés, au cours des dernières années, par la rapide dissémination de techniques d'apprentissage automatique (Mitchell, 1997), qui permettent d'acquérir automatiquement des connaissances et des règles de décision (exemplairement de désambiguïsation) à partir de corpus ou de ressources annotées (Manning et Schütze, 1999). Plusieurs raisons ont concouru à accélérer cette évolution, en particulier, la multiplication des sources de données disponibles et la démonstration répétée de l'efficacité de ces méthodes.

Si les approches à base de données s'opposent globalement aux approches à base de connaissances, cette notion recouvre de multiples sous-classes de méthodes, le choix d'une méthode particulière pour accomplir une tâche donnée dépendant largement des options théoriques et des buts poursuivis : cherche-t-on à construire un apprenti cognitivement réaliste ? un apprenti qualitativement performant ? un apprenti dont on saura analyser et exploiter les connaissances ? auquel on pourra incorporer des connaissances linguistiques ? qui sera capable de traiter des représentations variées ? etc. Nous nous intéressons, dans ce travail, à un mécanisme d'apprentissage particulier, fondé sur la recherche de relations d'analogies au sein d'un échantillon de données étiquetées. Ce mécanisme présente deux spécificités qui le distinguent des méthodes d'apprentissage les plus utilisées, notamment des méthodes statistiques, et qui justifient son intérêt : sa capacité à intégrer des formes d'*a priori* linguistiques et à s'accommoder simplement de représentations symboliques riches et variées.

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 (une relation d'analogie) entre une nouvelle instance d'un problème et des instances mémorisées du même problème dont la solution est connue. Une fois cet appariement établi, une solution à la nouvelle instance du problème peut être élaborée, à partir d'une ou plusieurs solutions dites analogues. La mise en œuvre de ce type d'apprentissage présuppose donc la capacité à rechercher et à exploiter de tels appariements, c'est-à-dire à pouvoir d'une part donner un sens à la notion de relation analogique, et, d'autre part, à savoir réaliser efficacement le calcul de ces relations.

L'adaptation directe d'une telle démarche en traitement automatique des langues est loin d'être évidente : la taille des bases de données disponibles, qui peuvent contenir des centaines de milliers d'instances, rend prohibitive la recherche d'appariements structurels complexes. Notre hypothèse est que l'exploitation d'appariements simples, c'est-à-dire ici fondés sur des relations de forme, entre représentations linguistiques permet, dans certaines situations, de détecter des relations plus profondes entre ces entités, et qu'à ce titre, les analogies formelles peuvent servir de fondement à des mécanismes de généralisation à partir de données.

Les données linguistiques présentent en effet une forme d'organisation *paradigmatique*, qui se manifeste par l'existence de systèmes plus ou moins réguliers d'oppositions et d'alternances formelles, ces oppositions pouvant marquer des distinctions plus profondes (par exemple sémantiques). Pour illustrer ce phénomène par un exemple

simple, l'alternance (répétée) entre les terminaisons '*-ons*' et '*-ez*' dans la conjugaison verbale est la marque d'une opposition entre 1ère et 2ème personnes. L'observation que ces systèmes d'oppositions donnaient lieu à des *analogies de forme*, prenant la forme de *rapports de proportions*[1], a suscité plusieurs tentatives visant à utiliser les concepts de l'apprentissage par analogie pour généraliser à partir de données linguistiques, que ce soit dans le cadre de l'apprentissage de la conversion orthographique-phonétique, (Yvon, 1999), de l'analyse morphologique (Lepage, 1999a; Pirrelli et Yvon, 1999b) ou encore de l'analyse syntaxique (Lepage, 1999b). Si ces travaux ont pu montrer le bénéfice que l'on pouvait tirer de ce biais d'apprentissage particulier, ils se sont principalement focalisés sur les oppositions entre séquences finies de symboles, en définissant de manière parfois *ad-hoc* la notion de proportionnalité. Ainsi, (Pirrelli et Yvon, 1999b) ne considère que des alternances de préfixe et de suffixe, quand (Lepage, 1998) part d'une définition algorithmique des proportions entre séquences.

Le traitement de descriptions linguistiques requiert la capacité de manipuler d'autres types de données, en particulier les structures attribut-valeur, ou encore les arbres étiquetés. La première contribution de ce travail est une définition générale des rapports de proportion, valant pour l'ensemble de ces types de données, ainsi que la présentation d'algorithmes efficaces pour les calculer. Nous montrons en particulier que, pour ce qui concerne les mots et les arbres, notre définition généralise celles proposées dans la littérature (Lepage, 1998; Lepage, 1999b). La seconde contribution de ce travail est de montrer comment ces proportions formelles peuvent servir de fondement à un mécanisme d'inférence *générique*, capable de s'adapter à diverses situations d'apprentissage et de s'appliquer à de grandes collections de données. Nous nous inspirons, pour ce faire, des propositions de (Pirrelli et Yvon, 1999a).

La troisième contribution de ce travail est d'illustrer la flexibilité et la pertinence de ce mécanisme d'inférence pour deux tâches d'analyse morphologique. Pour mettre en évidence la versatilité de ce mécanisme et son indépendance vis-à-vis des représentations des données (ici morphologiques), nous nous intéressons à deux situations très différentes : l'une qui exploite des représentations proches de celles que postulent les théories contemporaines de la morphologie lexématique (voir p. ex. (Fradin, 2003)) ; l'autre qui manipule des arbres d'analyse hiérarchique en constituants morphématiques, qui sont des représentions plus conformes aux visions traditionnelles de l'analyse morphologique. Pour apprécier les performances obtenues, nous conduisons une évaluation quantitative, en comparant la qualité des prédictions obtenues à celles d'un autre apprenti de l'état de l'art. D'un point de vue pratique, ces expériences permettent d'apprécier la validité de l'approximation effectuée en ne considérant que des analogies de forme, et d'en cerner les limitations. Incidemment, ce travail sur l'analyse morphologique apporte également des arguments en faveur de l'hypothèse d'une organisation globalement paradigmatique des lexiques considérés.

---

1. La vision de l'analogie que nous présentons ici, qui se focalise sur les *analogies de forme*, se rapproche de celles d'Aristote et de Saussure, pour qui elle est synonyme de *rapport de proportion* (ou plus précisément d'*égalité de rapports*) ; dans la suite, nous nous en tiendrons à cette seconde dénomination, qui nous semble nettement moins ambiguë que la première.

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 des rapports de proportion qui s'applique 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 le résultat d'expériences conduites sur des tâches d'analyse morphologique de formes orthographiques. Enfin, la section 5 est 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 et calcul de proportions

Dans cette section, nous introduisons notre modèle d'inférence analogique, s'inspirant principalement des travaux présentés dans (Pirrelli et Yvon, 1999a) ; ce modèle est ensuite contrasté avec d'autres modèles de l'apprentissage par analogie (section 2.2).

2.1. *Le mécanisme d'inférence*

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 attributs d'objets nouveaux, qui ne sont que partiellement informés. L'ensemble des attributs connus constitue *l'espace d'entrée* de la tâche d'apprentissage, les attributs à inférer en 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 en traitement automatique des langues, en particulier *l'étiquetage de données séquentielles*. Cette situation correspond en effet au cas où chaque objet est décrit par deux attributs (l'entrée et la sortie) prenant des valeurs séquentielles, l'apprenti devant pouvoir généraliser cette correspondance à de nouvelles séquences d'entrée.

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 *paresseux* (Mitchell, 1997; Aha, 1997). L'étape d'inférence s'effectue par identification de relations formelles de proportionnalité existant dans l'espace d'entrée, puis par reconstruction, toujours par proportionnalité, des attributs inconnus.

Formellement, nous supposons donnée une base d'apprentissage $\mathcal{A}$ d'objets représentés par un ensemble d'attributs. Ces attributs correspondent à des variables de types divers : des symboles dans un ensemble fini, des nombres, des chaînes de caractères, des arbres, etc. Par exemple, des entrées d'un lexique seront décrites par leur représentation orthographique (une séquence de lettres), leur représentation phonétique (une séquence de phonème), des propriétés morpho-syntaxiques (des variables catégorielles), un arbre d'analyse morphologique, leur fréquence en corpus, etc.

Nous supposons également donnée la possibilité de construire des rapports de proportion entre les objets de $\mathcal{A}$. Ces rapports 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, notée $A : B :: C : D$ et signifiant « A est à B ce que C est à D ». Lorsque les objets ne sont que partiellement informés, nous notons A^+ la partie connue de A et A^- sa partie inconnue. Soit alors X^+ la partie connue d'un objet X absent de la base d'apprentissage ; l'inférence analogique se formalise comme suit :

1) *recherche* : cette étape consiste en l'exploration de l'ensemble $\mathcal{T}$ des triplets de $\mathcal{A} \times \mathcal{A} \times \mathcal{A}$ défini par :

$$\mathcal{T}(X) = \{(A, B, C) \in \mathcal{A}^3,\ A^+ : B^+ :: C^+ : X^+\};$$

2) *transfert* : chaque triplet de $\mathcal{T}(X)$ donne lieu à une ou plusieurs hypothèses $\hat{X}^-$ concernant les attributs inconnus de X par calcul du quatrième de proportion de l'équation : $\hat{X}^- = A^- : B^- :: C^- :?$

Pour illustrer ce mécanisme, supposons que l'inférence analogique porte sur des entrées lexicales et que l'apprentissage exploite un ensemble d'instances représentées par leur forme graphique (séquence alphabétique) et leur catégorie syntaxique. Pour chaque entrée lexicale, A^+ se réduit donc à la forme graphique et A^- à la catégorie.

Soit alors la forme inconnue '*linguisiticiel*' : la découverte, durant l'étape de recherche, de la proportion '$logique : logiciel :: linguistique : linguisticiel$' dans l'espace des formes graphiques permet de suggérer, pendant l'étape de transfert, que la catégorie syntaxique de '*linguisticiel*' s'obtient par calcul du quatrième terme de la proportion : $Adj : Nom :: Adj :?$, soit ici 'Nom'.

2.1.1. *Remarques et compléments*

La procédure décrite dans les lignes précédentes est générale, dans la mesure où sa mise en œuvre repose presque uniquement sur la capacité à donner un sens et à effectuer deux d'opérations élémentaires :

– vérifier des rapports de proportion entre des attributs ou ensembles d'attributs, pendant l'étape de recherche ;

– calculer le quatrième terme d'une proportion pour un attribut ou ensemble d'attributs, durant l'étape de transfert.

La complexité globale de cette procédure dépend en pratique de la complexité de ces deux opérations, qui dépend des types de données qui sont manipulées. Ces questions sont discutées en détail dans les sections qui suivent. Notons toutefois qu'indépendamment de la complexité de ces calculs élémentaires, l'étape de recherche (1) demande d'explorer l'ensemble $\mathcal{A}^3$. Une exploration exhaustive naïve conduit à vérifier de l'ordre de $|\mathcal{A}|^3$ rapports de proportion, ce qui n'est que rarement réalisable. Pour limiter la complexité de cette étape, les expériences décrites en section 4 reposent sur une exploration partielle de l'espace de recherche, effectuée en sélectionnant aléatoirement, nous verrons comment, des triplets (A, B, C) « prometteurs » dans $\mathcal{A}^3$.

Ce mécanisme d'inférence, pour être complètement spécifié, demande également de définir la stratégie d'agrégation des hypothèses construites durant l'étape de transfert (2). Lorsque plusieurs hypothèses sont en effet proposées, une procédure de classement est nécessaire pour les départager, tout au moins pour les ordonner de la plus à la moins plausible. Dans ce travail, nous avons adopté une stratégie d'agrégation très simple : les hypothèses sont ordonnées de la plus à la moins fréquente. D'autres stratégies d'ordonnancement et d'agrégation sont naturellement envisageables, que nous discuterons à la section 5.

2.2. *Proportions et analogies*

L'apprenti présenté à la section précédente est un apprenti « paresseux » (Mitchell, 1997; Aha, 1997), dont la phase d'apprentissage se réduit à une mémorisation par cœur des données disponibles. Contrairement à la plupart des mécanismes d'apprentissage, notamment statistiques, il ne s'appuie pas sur la recherche d'un *modèle* pour les données (ou leurs frontières), que l'apprenti sélectionnerait par optimisation d'un critère numérique sur un ensemble de modèles possibles. L'approche que nous proposons présente donc une grande proximité avec d'autres apprentis paresseux tels que les méthodes de types k plus proches voisins ou apparentés (voir, par exemple, (Skousen, 1989; Jones, 1996; Daelemans et van den Bosch, 2005)), avec lesquelles nos algorithmes seront comparés à la section 4. Elle s'en rapproche également par son utilisation des fréquences pour agréger des hypothèses concurrentes.

Insistons toutefois sur une différence essentielle : les k plus proches voisins s'appuient sur l'hypothèse que plus deux objets se ressemblent (plus ils ont d'attributs en commun), plus ils sont susceptibles de partager d'autres attributs. Par contraste, l'hypothèse principale que nous formulons, et sur laquelle repose le succès de l'inférence, est que si deux *couples* d'objets sont en relation de proportionnalité pour certains attributs, alors il est plausible qu'ils soient également dans un rapport de proportionnalité pour d'autres attributs. C'est donc non pas la similarité (notion graduelle) entre objets qui justifie le transfert ou l'inférence des valeurs d'attributs inconnus, mais le parallélisme (notion booléenne) des relations de formes entre des couples d'objets. On trouvera dans (Pirrelli et Yvon, 1999a) divers arguments d'inspiration linguistique visant à étayer cette hypothèse.

Cette hypothèse n'est pas entièrement originale : elle est notamment centrale dans les travaux d'Y. Lepage (voir, par exemple (1999b; 1999a; 2003)), qui propose un mécanisme d'inférence identique au nôtre pour des applications d'analyse morphologique ou de traduction à base d'exemples. Les définitions qu'il utilise pour les rapports de proportion sont toutefois légèrement différentes de celles que nous proposons, comme il sera mentionné dans la section suivante. Elles sont également suggérées par (Delhay et Miclet, 2005), qui en envisage une généralisation, fondée sur une notion graduelle des rapports de proportionnalité.

La définition d'un apprenti fondé sur des rapports de proportionnalité peut finalement être vue comme la mise en œuvre simplifiée d'une forme *de raisonnement par analogie*, thème sur lequel il existe une longue tradition en Intelligence Artificielle. La capacité à identifier des relations analogiques entre des situations apparemment distinctes et d'utiliser ces relations pour résoudre des problèmes est, en effet, souvent présentée comme une capacité cognitive centrale (voir p. ex. (Gentner et al., 2001)). Cette observation a suscité un grand nombre de travaux visant à modéliser cette capacité, aussi bien par des modèles symboliques (p. ex. (Falkenhainer et al., 1989; Thagard et al., 1990; Hofstadter et the Fluid Analogies Research Group, 1995)) que subsymboliques (p. ex. (Plate, 2000)). Comme dans notre modèle, la généralisation s'établit en deux étapes : recherche et élaboration d'une relation analogique, puis transfert de la solution connue vers le nouveau problème. Ces travaux mettent toutefois l'accent sur le processus d'élaboration d'un *appariement structurel* entre une situation nouvelle et une situation mémorisée. L'appariement structurel vise à rapprocher des situations, qui, quoiqu'en apparence très dissemblables, 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* ce que *la terre* est au *soleil* ». Le calcul de proportions est un cas limite de ce type approche, dans lequel les relations considérées sont des relations formelles.

En résumé, le modèle d'inférence que nous proposons se situe à mi-chemin entre deux approches : des modèles du raisonnement par analogie, nous retenons l'idée de recherches de ressemblances qui dépassent les similarités de surface, que nous avons traduite en donnant un rôle central aux relations de proportionnalité. Des travaux conduits 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. Proportions sur des structures algébriques

Dans cette section, nous définissons formellement la notion de rapport de proportion, en débutant par une définition générale (à la section 3.1), qui est ensuite spécialisée pour diverses structures algébriques : treillis, monoïdes et ensembles d'arbres. Cette proposition détaille et étend (Stroppa et Yvon, 2005).

Ces définitions sont implantées dans un outil générique de recherche et de résolution de proportions analogiques, ALANIS (A Learning-by-Analogy Inferencer for Structured data), s'appuyant sur la bibliothèque de manipulation d'automates VAUCANSON (Lombardy et al., 2004), qui utilise massivement la programmation générique et le polymorphisme statique (Régis-Gianas et Poss, 2003).

3.1. *Les bases*

Comment définir en toute généralité la notion de rapport de proportion ? Débutons par un exemple qu'il nous semble intéressant de chercher à modéliser, impliquant de nouveau les quatre formes graphiques ('*logique*', '*logiciel*', '*linguistique*', '*linguisticiel*'), entre lesquelles on aimerait établir un rapport de proportion. On observe alors que prises deux à deux, ces forment partagent soit un préfixe (*log-* d'un côté, *linguist-* de l'autre), soit un suffixe (*-ique*, *-iciel*) ; et que chacune se segmente intégralement en préfixe+suffixe.

La notion de proportion entre quatre termes semble donc établie par la possibilité de *décomposer* chacun des termes en des termes plus petits (dans l'exemple précédent en deux termes) qui sont deux à deux échangeables, autrement dit qui *alternent*.

Cette intuition se formalise de la manière suivante. Soit U un ensemble muni d'une loi de composition interne associative $\oplus$. Pour exprimer la notion de décomposition, nous introduisons la notion de *factorisation* d'un élément u de U, définie comme suit.

Définition 1 *Une* factorisation *d'un élément* $u \in (U, \oplus)$ *est une séquence* $f_u = (u_1, \ldots, u_n) \in U^n$ *telle que* $u_1 \oplus \ldots \oplus u_n = u$. *On notera :* $f_u(i) = u_i$ *et* $|f_u| = n$.

En intégrant la contrainte d'alternance entre les termes d'une décomposition, nous aboutissons à la définition générale suivante pour les rapports de proportion.

Définition 2 (Proportion analogique) *Pour* $(x, y, z, t) \in U^4$, *on a* $x : y :: z : t$ *si et seulement s'il existe des factorisations* $(f_x, f_y, f_z, f_t) \in (U^d)^4$ *telles que*

$$\forall i \in [\![1, d]\!], (f_y(i), f_z(i)) \in \{(f_x(i), f_t(i)), (f_t(i), f_x(i))\}.$$

Cette définition est très générale, au sens où elle s'applique dès que U est munie d'une loi de composition interne associative. Dans le cas où la structure algébrique $(U, \oplus)$ 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 la proportionnalité : voir (Stroppa, 2005), qui démontre, en particulier, que la définition introduite ci-dessus est une *généralisation* du rapport de proportionnalité comme on le définit, par exemple, entre des éléments de $(\mathbb{N}, \times)$ ou encore entre les sommets d'un parallélogramme dans un espace vectoriel.

Cette définition vaut également pour de nombreuses autres structures algébriques classiques, telles que les espaces vectoriels, ainsi que des structures usuellement utilisées en représentation des connaissances. Nous nous intéressons, dans la suite de cette section, à certaines de ces structures, qui seront utilisées dans les expériences de la section 4, en débutant par les ensembles et les structures de traits.

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 et satisfaisant la loi d'absorption. L'ensemble des structures de traits, muni 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, des manipulations simples utilisant la commutativité permettent de simplifier la définition 2. (Stroppa, 2005) démontre en particulier le résultat suivant.

Proposition 1 (Proportion analogique (cas des treillis)) *Pour un quadruplet d'éléments dans un treillis $(x, y, z, t) \in (L; \wedge, \vee)^4$, on a $x : y :: z : t$ 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 (1) 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, 2005).

Un cas particulier important de la proposition 1 concerne les structures attribut-valeur, c'est-à-dire les structures de traits non-récursives. En effet, la mise en œuvre de la procédure d'inférence présentée à la section 2 requiert que l'on sache établir des proportions entre des ensembles de propriétés représentés par des structures attribut-valeur. On tire de la proposition 1 le résultat suivant, qui établit que de telles structures forment un rapport de proportion s'il est possible d'établir un rapport de proportion entre les valeurs de chacun des attributs.

Proposition 2 *Soit (x, y, z, t) un quadruplet de structures attribut-valeur, contenant chacune les attributs $f_1, \ldots, f_p$. On note $a.f_i$ la valeur de l'attribut i de la structure a. On a $x : y :: z : t$ si et seulement si :*

$$\forall i \in \{1, \ldots, p\},\ x.f_i : y.f_i :: z.f_i : t.f_i.$$

La Figure 1 illustre cette notion pour des structures attribut-valeur simples.

$$\begin{bmatrix}\text{graphie : } \textit{acteur}\\ \text{cat : Nom}\\ \text{genre : masc}\\ \text{nombre : sing}\\ \text{pers : -}\end{bmatrix} : \begin{bmatrix}\text{graphie : } \textit{penseur}\\ \text{cat : Nom}\\ \text{genre : masc}\\ \text{nombre : sing}\\ \text{pers : -}\end{bmatrix} :: \begin{bmatrix}\text{graphie : } \textit{actent}\\ \text{cat : Verbe}\\ \text{genre : -}\\ \text{nombre : plur}\\ \text{pers : 3}\end{bmatrix} : \begin{bmatrix}\text{graphie : } \textit{pensent}\\ \text{cat : Verbe}\\ \text{genre : -}\\ \text{nombre : plur}\\ \text{pers : 3}\end{bmatrix}$$

Figure 1. *Rapport de proportion entre structures attribut-valeur*

3.3. *Mots sur un alphabet fini*

Cette section est consacrée à l'étude d'une structure algébrique particulière, le monoïde libre, qui est un modèle classique en traitement des langues pour représenter des séquences. Partant de la définition 2, qui s'applique directement à des mots, nous étudions la question de la vérification et de la résolution des proportions. Nous montrons en particulier qu'il est possible de réaliser ces calculs à l'aide de transducteurs finis[2].

3.3.1. *Proportions entre mots*

Soit Σ un alphabet fini. On note $\Sigma^\star$ l'ensemble des séquences finies d'éléments de Σ, appellées *mots* sur Σ. L'ensemble $\Sigma^\star$, muni de l'opération de concaténation . est un monoïde libre, dont l'élément neutre est noté ε. Pour $w \in \Sigma^\star$, $w(i)$ désigne le i$^{\text{ème}}$ symbole de w. Dans ce cadre, la définition 2 donne lieu à la proposition suivante.

Proposition 3 (Proportion analogique entre mots) *Pour $(x, y, z, t) \in \Sigma^{\star 4}$, on a $x : y :: z : t$ ssi il existe des factorisations $(f_x, f_y, f_z, f_t) \in ((\Sigma^\star)^d)^4$ telles que :*

$$\forall i \in [\![1, d]\!], (f_y(i), f_z(i)) \in \{(f_x(i), f_t(i)), (f_t(i), f_x(i))\}.$$

Le plus petit d pour lequel de telles factorisations peuvent être trouvées est appelé le degré *de la proportion.*

Cette définition modélise effectivement la proportion (de degré 2) entre '*logique*', '*logiciel*', '*linguistique*' et '*linguisticiel*' : il suffit de prendre par exemple $x_1 = y_1 =$'*log*', $x_2 = z_2 =$'*ique*', $z_1 = t_1 =$'*linguist*', $y_2 = t_2 =$'*iciel*'.

On peut montrer (Yvon, 2003) qu'elle généralise la définition (algorithmique) de l'analogie entre mots proposée par (Lepage, 1998), avec laquelle elle partage les deux traits suivants. Étant donné trois termes, il est à la fois possible que le calcul du quatrième de proportion n'aboutisse à aucune solution, ou bien au contraire qu'il conduise à plusieurs solutions. (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 les solutions d'une équation analogique sont toutes de même longueur.

2. Ces définitions seront utilisées pour réaliser les expériences décrites dans la Section 4 : les mots (formels) de ces définitions seront alors des mots (de la langue) de l'anglais, de l'allemand et du néerlandais.

3.3.2. *Un solveur à états finis*

La proposition 3 donne lieu à des procédures efficaces pour vérifier un rapport de proportion ou résoudre le calcul du quatrième de proportion, qui s'appuie sur le formalisme des transducteurs finis. Nous esquissons ici les grandes lignes de ces procédures, et renvoyons à (Yvon, 2003; Yvon et al., 2004) pour la démonstration des principaux résultats qui sous-tendent cette construction, et à (Stroppa, 2005) pour une discussion des questions d'implantation des procédures. Pour débuter, nous introduisons les notions de *sous-mot complémentaire* et de *produit de mélange*.

3.3.2.1. Complémentarité

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

La relation de complémentarité de deux mots par rapport à x est une relation rationnelle. Le calcul des complémentaires de v par rapport à x s'effectue en prenant l'image de v par le transducteur fini T_x, illustré sur la Figure 2.

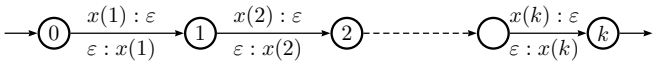


Figure 2. *Transducteur calculant la relation de complémentarité par rapport à x*

3.3.2.2. Mélange

Le *mélange* $u \bullet v$ de deux mots de $\Sigma^\star$ se définit comme suit (Sakarovitch, 2003) :

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

Le mélange de u et v contient tous les mots formés avec les lettres de u et de v et tels que si la lettre a précède la lettre b dans u ou v, alors cet ordre est respecté dans $u \bullet v$. Ainsi, si l'on prend $u = abc$ et $v = def$, alors les mots suivants : $abcdef$, $abdefc$, $adbecf\dots$ sont des éléments de $u \bullet v$; ce n'est pas le cas de $abefcd$, dans lequel d suit e, alors qu'il devrait le précéder. Cette opération se généralise à des langages : pour la suite de cette courte présentation du mélange, nous considérons ce cas plus général et identifions donc $u \bullet v$ avec $\{u\} \bullet \{v\}$.

Le produit de mélange de deux langages rationnels est rationnel ; l'automate le reconnaissant est construit en formant le produit des automates reconnaissant ces deux langages. 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-mot et de mélange sont reliées par la relation suivante :

$$\forall x, u, v \in \Sigma^\star, x \in u \bullet v \Leftrightarrow u \in x \backslash v.$$

3.3.2.3. Résolution

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

Proposition 4

$$\forall x, y, z, t \in \Sigma^\star, x : y :: z : t \Leftrightarrow x \bullet t \cap y \bullet z \neq \emptyset.$$

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

Proposition 5

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

Ce résultat énonce que l'ensemble des valeurs possibles pour le quatrième de proportion dans $x : y :: z :?$ est un ensemble rationnel, calculable par un transducteur fini.

Les constructions discutées dans cette section permette d'établir la complexité des procédures de vérification d'un rapport de proportion et de calcul du quatrième terme d'une proportion. Pour la vérification, il est nécessaire de construire les mélanges de x et y d'une part, de y et z d'autre part, conduisant à deux automates de taille respective $|x + 1| \times |y + 1|$ et $|z + 1| \times |t + 1|$, qu'il s'agit ensuite d'intersecter. Au total la complexité de la vérification est donc $O(|x| \times |y| \times |z| \times |t|)$. Le problème de la résolution est similaire : cette procédure demande d'intersecter l'automate représentant x avec le transducteur calculant le complémentaire de $y \bullet z$. Il existe donc un automate non-déterministe représentant l'ensemble des solutions possibles et possédant de l'ordre de $O(|x| \times |y| \times |z|)$ états. Une étude de l'implantation efficace de procédure exploitant ce formalisme, décrivant l'optimisation du calcul d'un quatrième terme *de degré minimum* est présentée dans (Stroppa, 2005).

Divers résultats complémentaires sont établis dans (Yvon, 2003; Yvon et al., 2004). En particulier :

– nous explorons diverses manières d'introduire une notion de gradualité dans les analogies (à l'aide du degré, mais également en considérant d'autres manières de valuer les rapports de proportions) ;

– nous montrons également comment généraliser ce résultat à des mots sur un alphabet Σ, qui serait lui-même muni d'une structure algébrique. Cette généralisation permet de définir des proportions entre, par exemple, des séquences de structures de traits.

3.4. *Proportions entre arbres étiquetés*

Le cas des arbres est plus problématique, dans la mesure où l'ensemble des arbres n'est pas « naturellement » muni d'une loi de composition interne associative. 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. Comme il est habituel en traitement des langues, les arbres considérés dans cette section sont des arbres étiquetés[3]. L'ensemble des arbres étiquetés par un alphabet L est noté $\mathcal{T}(L)$; on distinguera un sous-ensemble de variables $V \subseteq L$.

Définition 3 (Substitution) *Une* substitution *(unitaire) est une paire $(v \leftarrow t')$, où $v \in V$ est une variable et $t' \in \mathcal{T}(L)$ un arbre. L'application de la substitution $(v \leftarrow t')$ à un arbre t s'opère en remplaçant chaque feuille de t étiquetée par v par l'arbre t'. Le résultat de cette opération est noté $t \lhd_v t'$, ce qui définit un opérateur binaire $\lhd_v$.*

La notion de factorisation en résulte directement.

Définition 4 (Factorisation) *Une* factorisation *d'un arbre $t \in \mathcal{T}(L)$ est définie par une séquence de variables $(v_1, \ldots, v_{n-1})$ et une séquence de sous-arbres $(t_1, \ldots, t_n)$ telles que $t_1 \lhd_{v_1} \ldots \lhd_{v_{n-1}} t_n = t$.*

La définition de la proportion analogique entre arbres s'en déduit immédiatement.

Définition 5 (Proportion analogique (entre arbres)) *Pour $(x, y, z, t) \in \mathcal{T}(L)^4$, on a $x : y :: z : t$ ssi il existe des factorisations d'arbres $(f_x, f_y, f_z, f_t) \in \mathcal{F}(\mathcal{T}(L))^4$ telles que :*

$$\forall i \in [\![1, d]\!], (f_y(i), f_z(i)) \in \{(f_x(i), f_t(i)), (f_t(i), f_x(i))\},\ \textit{c.-à-d.}$$

$$\exists (x_i)_{i\in[\![1,d]\!]}, (y_i)_{i\in[\![1,d]\!]}, (z_i)_{i\in[\![1,d]\!]}, (t_i)_{i\in[\![1,d]\!]}\ \textit{t.q.}$$

$$\begin{array}{ll} x_1 \lhd_{v_1} \ldots \lhd_{v_{n-1}} x_n = x, & y_1 \lhd_{v_1} \ldots \lhd_{v_{n-1}} y_n = y, \\ z_1 \lhd_{v_1} \ldots \lhd_{v_{n-1}} z_n = z, & t_1 \lhd_{v_1} \ldots \lhd_{v_{n-1}} t_n = t, \end{array}$$

$$\textit{et } \forall i \in [\![1, d]\!], (y_i, z_i) \in \{(x_i, t_i), (t_i, x_i)\}.$$

La Figure 3 fournit un exemple d'une telle proportion entre arbres[4].

3. Ces développements seront utiles pour les expériences décrites au paragraphe 4.3.2, dans lesquelles les arbres de dérivation morphologique donnant l'analyse morphématique des mots sont des arbres étiquetés.

4. Notons que bien que les arbres puissent êtres considérés comme des structures de traits, la définition associée à la proposition 1 n'est pas adaptée aux phénomènes que l'on cherche à modéliser sur les arbres. En particulier, on peut vérifier que l'exemple de la Figure 3 n'est pas couvert par cette définition, justifiant le recours à la notion de substitution.

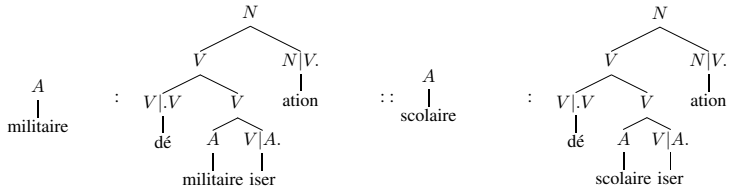


Figure 3. *militaire : démilitarisation :: scolaire : déscolarisation*

3.4.1. *Approximation par linéarisation*

Notre définition de la proportion analogique entre arbres est cohérente avec le cadre général présenté dans les sections 3.1, 3.2 et 3.3. En revanche, nous ne disposons actuellement pas de procédure efficace de calcul pour les opérations qui découlent de cette définition. De manière à pouvoir tout de même appliquer notre modèle d'inférence lorsque les objets considérés sont des arbres, notre approche consiste à appliquer une fonction de *linéarisation* qui associe de manière biunivoque un arbre et sa représentation linéarisée. Ainsi, après conversion des arbres en chaînes, il est possible de résoudre l'équation entre les représentations linéaires obtenues (cf. section 3.3.2), pour enfin construire une représentation arborée (si la chaîne obtenue correspond bien à une linéarisation d'arbre[5]). Cette approche est similaire à celle de (Lepage, 1999b) ; elle est de plus 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, 2005).

Proposition 6 *Pour* $(x, y, z, t) \in \mathcal{T}(L)^4$, *si* $x : y :: z : t$, *alors* $l(x) : l(y) :: l(z) : l(t)$ *pour* l *une « linéarisation appropriée ».*

Ainsi, pour l'exemple illustré sur la Figure 3, il est possible de résoudre l'équation constituée des trois premiers arbres en considérant l'équation entre chaînes suivantes : *(A(militaire))* : *(N(V(V\.V(dé))(.V(.A(militaire))(V\A.(iser))))(N\V.(ation)))* :: *(A(scolaire))* : *?*, dont la solution est : *(N(V(V\.V(dé))(.V(.A(scolaire))(V\A.(iser))))(N\V.(ation)))*.

3.4.1.1. Remarques concernant la linéarisation

Les résultats présentés ci-après légitiment l'utilisation du solveur d'équations disponibles pour le cas des chaînes, dont on sait qu'il implique uniquement des opérations rationnelles, ce qui permet de conclure sur la complexité de la procédure. Toutefois, puisque l'implication n'est vérifiée que dans un seul sens, cette utilisation correspond à une approximation par filtrage. En effet, le résultat d'implication nous permet de filtrer les configurations qui ne correspondent pas à une proportion, les autres étant conservées par défaut. Dans les cas pratiques que nous avons étudiés, cela ne semble

5. Il est en effet possible que la chaîne obtenue ne corresponde pas à un arbre ; c'est le cas par exemple de l'expression (mal) parenthésée *((as(d(*.

pas poser un réel problème, comme le suggère l'exemple discuté ci-dessus. Notons enfin qu'il est possible de se rapprocher d'un calcul exact en effectuant les calculs simultanément sur plusieurs linéarisations.

4. Expériences

Dans cette section, nous présentons deux tâches qui permettent d'illustrer l'exploitation de proportions analogiques dans le cadre de l'apprentissage de la morphologie. La première tâche consiste à déterminer les traits morpho-syntaxiques et le lemme associés à une forme inconnue. Dans la deuxième, il s'agit d'analyser une forme en lui associant une décomposition hiérarchique correspondant à sa structure interne.

Nous commençons par introduire les motivations d'un apprentissage pour ce type de tâches ; nous soulignons également les bénéfices d'une approche à base d'analogie dans ce contexte. Les lexiques utilisés (anglais, allemand et néerlandais) ainsi que le protocole expérimental sont ensuite décrits. Enfin, nous présentons et discutons les résultats de ces expériences, qui compare notre méthode implantée dans le logiciel ALANIS à un classifieur de l'état de l'art, TIMBL (Daelemans et al., 2004). Ces résultats complètent (Stroppa et Yvon, 2005).

4.1. *Analyser 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 à large couverture, 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 relativement réguliers (flexion, dérivation ou composition).

La caractérisation (l'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 et 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 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 existant, mais également à développer des sys-

tèmes 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 fines de la morphologie : les phénomènes flexionnels sont bien décrits, ce n'est pas encore le cas des autres phénomènes, même pour des langues les mieux étudiées, justifiant le recours à des méthodes d'apprentissage.

4.2. *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 et 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 et Wicentowski, 2000; Goldsmith, 2001; Schone et 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. D'une part, l'exploitation de proportions entre structures attribut-valeur dans lesquelles les formes sont non analysées (c.-à-d. structure interne ignorée) nous rapproche des modèles de la *morphologie lexématique*, à l'instar des travaux présentés notamment dans (Pirrelli et Yvon, 1999b; Lepage, 1999a). D'autre part, il est possible d'identifier des proportions analogiques aussi bien entre chaînes qu'entre arbres : on peut donc proposer une *analyse morphématique* (décomposition hiérarchique) à partir d'une forme inconnue. Ces deux expériences sont présentées dans les sections qui suivent.

4.3. *Description des tâches et données*

4.3.1. *Analyse flexionnelle*

Cette tâche consiste à analyser une forme graphique hors-vocabulaire, en la rattachant à un lexème, et en renseignant un certain nombre de traits morpho-syntaxiques : catégorie grammaticale, genre, nombre, cas, temps, etc. Un exemple de paire d'entrée/sortie pour cette tâche est[6] :

6. D'autres protocoles sont envisageables, consistant à informer progressivement les attributs définissant l'espace de sortie : calculer d'abord la catégorie, puis celle-ci étant connue, calculer

$$\text{entrée=}\textit{marcherons}\ ;\ \text{sortie=}\begin{bmatrix} Lexème : marcher \\ Catégorie : verbe \\ Personne : 1^{re} \\ Nombre : pluriel \\ Temps : futur\ de\ l'indicatif \end{bmatrix}.$$

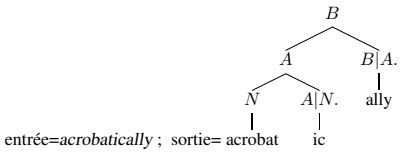
Les données utilisées pour cette tâche sont issues des lexiques CELEX (Burnage, 1990) pour l'anglais, l'allemand et le néerlandais ; leurs propriétés sont détaillées dans le tableau 1.

Corpus	Langue	Nombre d'entrées	Nombre d'entrées ambiguës	Pourcentage d'entrées ambiguës	Nombre moyen d'analyses
CELEX	Anglais	89 412	17 320	19,37%	1,80
	Néerlandais	324 249	39 332	12,13%	1,72
	Allemand	342 452	8 897	2,60%	1,05

Tableau 1. *Propriétés des données, tâche d'analyse flexionnelle*

4.3.2. *Analyse dérivationnelle*

L'objectif de cette tâche est d'effectuer une analyse dérivationnelle d'un lexème représenté par une forme graphique canonique. Cette analyse correspond à une structure hiérarchique qui donne la trace du processus de construction du lexème. Un exemple de paire d'entrée/sortie pour cette tâche est :



Dans la structure hiérarchique de sortie, les feuilles constituent les morphèmes composant le lexème. Les nœuds sont étiquetés par des catégories grammaticales. Un morphème étiqueté par une catégorie de la forme $A|N.$ est un morphème de type « affixe » ; une catégorie telle que $A|N.$ signifie que le morphème en question est suffixé (.) à un nom (N) pour former un adjectif (A).

Les données utilisées pour cette tâche sont également issues des lexiques CELEX ; leurs propriétés sont détaillées dans le tableau 2. Un exemple d'entrée ambiguë est illustré sur la figure 4. Cet exemple montre également un cas d'allomorphie, dans lequel *ab+negate+ion* donne *abnegation.*

---

le lemme, etc. Notons également qu'en inversant le rôle des entrées et des sorties, l'analyseur se mue en générateur (bidirectionalité).

Corpus	Langue	Nombre d'entrées	Nombre d'entrées ambiguës	Pourcentage d'entrées ambiguës	Nombre moyen d'analyses
CELEX	Anglais	46 129	7 309	15,84%	1,18
	Néerlandais	120 967	8 651	7,15%	1,06
	Allemand	49 936	1 415	2,83%	1,02

Tableau 2. *Propriétés des données, tâche d'analyse dérivationnelle*

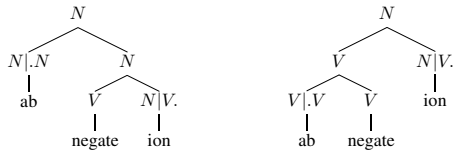


Figure 4. *abnegation : une entrée ambiguë pour l'analyse dérivationnelle*

4.4. *Protocole expérimental*

4.4.1. *Mesures des performances*

Soit x_i une entrée à analyser. Nous notons y_i l'ensemble des analyses possibles de x_i ; le fait de considérer un ensemble de solutions et non une solution unique permet de prendre en compte l'ambiguïté inhérente aux données. Pour un système d'apprentissage f, l'ensemble des solutions proposées pour une entrée x_i est noté $f(x_i)$. La précision p_i (resp. le rappel r_i) associée à un couple $(y_i, f(x_i))$ correspond au nombre de sorties correctes proposées rapporté au nombre total de sorties proposées (resp. attendues), soit :

$$p_i = \frac{|\, f(x_i) \cap y_i \,|}{|\, f(x_i) \,|}, \;\; r_i = \frac{|\, f(x_i) \cap y_i \,|}{|\, y_i \,|}.$$

Pour un ensemble d'entrées $\{x_1, x_2, \ldots, x_n\}$, la *micro-précision* globale p_m s'obtient en moyennant la précision sur l'ensemble des exemples. Il est également possible de pondérer les exemples par le nombre de sorties attendues ; on parlera alors de *macro-précision* (p_M). Les mêmes grandeurs sont définies pour le rappel.

$$p_m = \frac{1}{n} \sum_{i=1}^{n} p_i, \;\; p_M = \frac{\sum_{i=1}^{n} |\, y_i \,| p_i}{\sum_{i=1}^{n} |\, y_i \,|}.$$

Alors que les macro-mesures donnent plus de poids aux exemples ayant des sorties ambiguës, les micro-mesures traitent tous les exemples de manière identique. La différence entre les micro- et les macro-mesures permet de quantifier le comportement d'un algorithme vis-à-vis des entrées effectivement ambiguës.

L'approche adoptée pour évaluer les capacités de généralisation du modèle implique m étapes. À chaque étape, n exemples sont tirés aléatoirement et retirés de la base de données pour former la base de tests ; l'algorithme est entraîné sur les exemples restants (la base d'apprentissage) puis testé sur ces n exemples. Les mesures de rappel et de précision sont calculées pour cette étape ; ces mesures sont moyennées sur les m étapes.

Dans toutes les expériences effectuées, nous avons posé $m = 10$ et $n = 1000$. D'une part, le fait que le nombre de données disponibles est élevé (bases de centaines de milliers d'exemples) assure que la suppression de 1000 exemples ne perturbe pas l'apprentissage ; d'autre part, un échantillon de taille 1000 est suffisant pour estimer précisément les capacités de généralisation d'un système d'apprentissage.

4.4.2. *Exploration et ordonnancement*

Comme souligné dans la section 2.1, l'étape d'inférence nécessite l'exploration de l'ensemble $\mathcal{A}^3$; pour éviter une exploration exhaustive, nous procédons de la manière suivante. Tout d'abord nous regroupons les entrées qui partagent un critère commun donné. Dans la première tâche, les formes sont regroupées selon leur famille flexionnelle, dans la deuxième, deux formes sont regroupées si elles ont une racine commune ; on opère donc un regroupement en *paradigmes*. Nous imposons ensuite que durant la recherche de triplets, les deux premiers éléments font partie d'un même paradigme, ce qui paraît naturel d'un point de vue linguistique. Cette contrainte permet de fortement limiter la taille de l'espace de recherche (Stroppa, 2005). La deuxième étape exploite la redondance naturelle des données linguistiques et leur organisation en paradigmes, l'idée principale étant que les paradigmes sont très redondants et souvent interchangeables. L'approche suivie consiste à sélectionner aléatoirement un nombre n de paradigmes ; seules les entrées appartenant aux paradigmes sélectionnés sont examinées (l'algorithme est donc linéaire en fonction de n). Ces deux approximations permettent d'accélérer très significativement la recherche de triplets : sur une machine de bureau, l'ordre de grandeur de cette recherche pour une entrée est la seconde.

Pour une entrée donnée, l'étape de recherche peut retourner plusieurs triplets, et pour chaque triplet, le calcul du quatrième de proportion peut fournir plusieurs solutions : de manière générale, plusieurs analyses sont donc proposées par le système. Ce comportement est bienvenu puisqu'il permet de rendre compte de l'ambiguïté des données linguistiques. Toutefois, il est souhaitable de disposer d'un mécanisme d'ordonnancement de ces analyses, de façon à mettre en évidence les plus prometteuses et d'éliminer celles qui sont les moins plausibles. Le critère pris en compte de manière à ordonner les solutions est le suivant : puisqu'une même solution peut être proposée plusieurs fois par le système, nous les classons en fonction de leur fréquences d'apparitions dans la liste des solutions[7]. Plusieurs techniques de filtrage, fondées sur divers seuils ou sur la notion d'entropie sont exposées dans (Stroppa, 2005).

7. Si les objets manipulés sont des chaînes, il est également possible de pondérer chaque solution par une quantité liée au degré de l'équation ayant conduit à la solution en question ; voir (Stroppa, 2005) pour davantage de détail.

4.4.3. *Comparaison avec un classifieur*

Les expériences décrites ci-dessous comparent notre système à un classifieur de l'état de l'art, TIMBL (Daelemans et al., 2004). Celui-ci permet de résoudre des problèmes de catégorisation supervisée, c'est-à-dire des problèmes dans lesquels les entrées sont des vecteurs (numériques ou symboliques) de taille fixe et les sorties des étiquettes appartenant à un ensemble fini[8]. Étant donné l'aspect structuré (chaînes, arbres, etc.) des objets manipulés dans les tâches décrites ci-dessus, il n'est pas possible d'appliquer directement un algorithme de classification. Cette limitation peut toutefois être contournée, au prix d'une reformulation (parfois peu naturelle) du problème[9]. Cette comparaison permet de mettre en évidence la souplesse de notre modèle, capable de gérer directement (c.-à-d. sans reformulation) des objets structurés dès lors que les opérations de calcul des proportions analogiques sont définies.

4.5. *Résultats*

4.5.1. *Analyse flexionnelle*

Les résultats obtenus sur la tâche d'analyse flexionnelle sont illustrés sur le tableau 3, ventilés en fonction de la catégorie syntaxique des entrées. Pour ces expériences, le paramètre d (degré maximal) est fixé à 3 et n (nombre de familles flexionnelles tirées aléatoirement) à 150[10].

Ces résultats montrent que la tâche, du point de vue de la classification, n'est pas triviale. De très bons scores peuvent être atteints par TIMBL (adjectifs allemands, noms néerlandais), mais sont très dépendants de la langue et de la catégorie. Par exemple, alors que les adjectifs allemands sont plutôt reconnus correctement (89% de rappel et de précision), les verbes le sont beaucoup moins (57% de rappel et de précision). En outre, il existe une certaine homogénéité entre le rappel et la précision. Cela est cohérent avec la non-gestion de l'ambiguïté : puisqu'une et une seule solution est proposée par entrée, une entrée ambiguë conduit à une diminution à la fois de la précision et du rappel.

En ce qui concerne ALANIS, on remarque qu'il est d'autant plus performant qu'il s'appuie sur des paradigmes plus riches. En particulier, la pauvreté de la flexion des adjectifs anglais entraîne un faible taux de rappel ; par contraste, les résultats sont plus élevés, aussi bien en terme de rappel qu'en terme de précision, pour les verbes,

8. Voir aussi (Bayoudh, 2006), qui traite d'un classifieur fondé sur les proportions analogiques.
9. Les détails concernant ces reformulations ne sont pas exposés ici et peuvent être trouvés dans (Stroppa, 2005).
10. Plus n est est élevé, meilleures sont les performances (Stroppa, 2005). Ce paramètre est donc un compromis temps/performance et n'est pas à optimiser ; il sera déterminé par les besoins de l'application et le temps de calcul disponible. En ce qui concerne d, il faut le choisir au moins égal à 3 pour couvrir l'ensemble des phénomènes susceptibles de nous intéresser (p. ex. l'infixation) ; en revanche, le choisir plus grand augmente le coût computationnel sans améliorer significativement les résultats.

			micro rappel	macro rappel	micro précision	macro précision
Anglais	Noms	ALANIS	75,26	**76,33**	**95,37**	85,19
		TIMBL	**76,06**	65,29	79,67	**91,34**
	Verbes	ALANIS	**94,79**	**93,46**	**97,37**	**94,49**
		TIMBL	33,78	25,65	42,36	76,99
	Adjectifs	ALANIS	27,89	36,91	**87,67**	71,15
		TIMBL	**57,61**	**48,98**	62,86	**81,64**
Néerlandais	Noms	ALANIS	54,59	55,25	74,75	67,77
		TIMBL	**85,39**	**82,65**	**86,24**	**88,50**
	Verbes	ALANIS	**93,26**	**93,59**	**94,36**	**91,20**
		TIMBL	45,82	43,97	50,79	61,74
	Adjectifs	ALANIS	**90,02**	**89,16**	**95,33**	**86,83**
		TIMBL	76,75	73,69	78,29	81,22
Allemand	Noms	ALANIS	77,32	73,30	**81,70**	78,17
		TIMBL	**80,95**	**80,06**	81,45	**82,01**
	Verbes	ALANIS	**90,50**	**88,62**	**90,63**	**87,78**
		TIMBL	56,49	55,40	57,21	58,33
	Adjectifs	ALANIS	**99,01**	**98,90**	**99,15**	89,31
		TIMBL	89,31	88,71	89,57	**89,84**

Tableau 3. *Tâche d'analyse flexionnelle : résultats*

dont la conjugaison implique un plus grand nombre de formes. On observe en outre un comportement assez conservateur du mécanisme de généralisation ; quel que soit le rappel, la précision est toujours assez élevée (plus mauvais score de 67% pour les noms en néerlandais), signifiant que le système préfère se taire que se tromper. À l'inverse, TIMBL propose toujours une solution, au détriment parfois de la précision (42% seulement pour les formes verbales de l'anglais).

Le tableau 4 présente des exemples d'analyses correctes et incorrectes pour l'anglais. La forme *acclaimed* (1), bien que fortement ambiguë, est parfaitement analysée :

	Formes	Analyses
(a)	(1) acclaimed	V+Pa, V+P+Pe+1, V+P+Pe+2, V+P+Pe+3 V+S+Pe+1, V+S+Pe+2, V+S+Pe+3
	(2) advanced levels	N+P
	(3) smoky	A+Po
(b)	(4) understandable	attendue : A+Po proposée : *(aucune)*
	(5) dowser	attendue : N+S proposées : N+S, V+S+Pr+3

Codes : V=Verbe, A=Adjectif, P=Pluriel, Po=Positif, Pr= Présent de l'indicatif, I = Infinitif, Pa= Participe passé, Pe=Prétérit, 1,2,3= 1re , 2e , 3e personnes

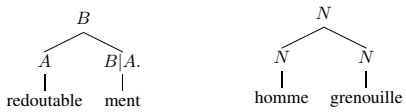
Tableau 4. *Analyse flexionnelle : exemples d'analyses correctes (a) et incorrectes (b)*

toutes les analyses possibles sont retrouvées. La forme composée *advanced levels* est correctement reconnue ; c'est également le cas de l'adjectif *smoky* (3), qui fait partie des adjectifs anglais sujets à flexion. La forme *understandable* (4), comme beaucoup d'autres adjectifs, n'est pas analysée. La forme *dowser* (5) donne lieu à deux analyses, dont une incorrecte, provenant de proportions telles que :

sleep (V+I) : *sleeps* (V+S+Pr+3) :: *dower* (V+I) : *dowser* (V+S+Pr+3).

4.5.2. *Analyse dérivationnelle*

Le tableau 5 représente les résultats obtenus respectivement par ALANIS et TIMBL sur la tâche d'analyse dérivationnelle. Les entrées sont distinguées selon qu'elles font intervenir de l'affixation ou de la composition. Une affixation correspond à l'ajout d'une feuille de type affixe, comme dans *redoutablement*, alors que la composition accole deux structures a priori autonomes, comme dans *homme-grenouille* :



Notons que dans cette tâche, nous cherchons à apprendre des structures entières ; cette mesure est en effet la plus « sévère » possible, et des sous-arbres corrects ne suffisent pas à faire augmenter le rappel et la précision. On peut donc considérer que cette tâche est nettement plus difficile que la précédente.

			micro rappel	macro rappel	micro précision	macro précision
Anglais	Composition	ALANIS	19,64	19,53	**88,33**	**91,49**
		TIMBL	**34,66**	**33,84**	59,43	61,52
	Affixation	ALANIS	**56,09**	**56,67**	**81,21**	**84,05**
		TIMBL	6,48	6,17	59,65	64,09
	Composition+Affixation	ALANIS	**17,21**	**16,26**	**84,68**	**89,99**
		TIMBL	10,99	9,86	74,86	82,24
Néerlandais	Composition	ALANIS	39,02	39,54	**88,58**	**90,39**
		TIMBL	**44,97**	**45,04**	77,93	80,39
	Affixation	ALANIS	**54,31**	**55,82**	**88,88**	**89,56**
		TIMBL	8,80	8,84	73,94	76,15
	Composition+Affixation	ALANIS	**30,73**	**31,78**	**84,64**	**88,13**
		TIMBL	25,47	23,46	80,14	84,55
Allemand	Composition	ALANIS	19,28	20,56	**94,11**	**95,16**
		TIMBL	**24,95**	**24,73**	68,53	69,23
	Affixation	ALANIS	**36,55**	**36,82**	**88,25**	**88,59**
		TIMBL	12,95	13,02	71,91	72,20
	Composition+Affixation	ALANIS	8,41	8,53	**92,24**	**95,60**
		TIMBL	**12,95**	**13,02**	71,91	72,21

Tableau 5. *Tâche d'analyse dérivationnelle : résultats*

Il n'est donc pas surprenant que les performances des deux systèmes soient bien moindres que sur la première tâche, et qualitativement assez comparables, même si la précision globale d'ALANIS est pratiquement systématiquement plus élevée que celle atteinte par TIMBL. Cela s'opère parfois au détriment du rappel ; ce qui illustre de nouveau le caractère conservateur du système, qui ne se prononce qu'avec une certaine assurance. En comparant les différents types de formes complexes, on note que la précision et le rappel se dégradent lorsque le procédé de construction s'éloigne du cas favorable que constitue l'affixation simple. Ces considérations s'appliquent surtout à ALANIS, qui s'appuie, pour analyser des formes affixées, sur l'existence de familles dérivationnelles telles que : *véritable* : *véritablement* :: *redoutable* : *redoutablement*. Par comparaison, les cas de composition, qui donnent lieu à des formes plus « isolées » donnent le plus souvent lieu à un échec de l'analyse.

Le tableau 6 présente des exemples d'analyses correctes et incorrectes pour l'anglais. La forme *algebraic* (1) est correctement analysée : suffixation de *ic* au nom *algebra* pour donner l'adjectif *algebraic*. La forme *historically* est ambiguë puisque deux analyses sont possibles ; ALANIS parvient à retrouver ces deux analyses. L'analyse est également correcte pour *acclimatise*, qui fait intervenir simultanément une préfixation et une suffixation. Pour la forme *acidulate* (4), difficile à analyser, ALANIS ne fournit aucune sortie. La forme *balanced* (5) donne lieu à une analyse non-attendue (donc comptabilisée comme incorrecte), mais elle pourrait en réalité être acceptée : le verbe *balance* donne le nom *balance*, auquel *ed* est suffixé pour former un adjectif. Dans le cas de *combustion* (6), une partie de l'analyse est reconstruite, mais la racine de l'arbre (étiquette N) est supprimée pour être remplacée par son fils gauche (étiquette V).

	Formes	Analyses
(a)	(1) algebraic	(A(N(algebra))(A|N.(ic)))
	(2) historically	(B(A(N(history))(A|N.(ic)))(B|A.(ally))) (B(A(N(history))(A|N.(ical)))(B|A.(ly)))
	(3) acclimatise	(V(V|.Nx(ac))(N(climate))(V|xN.(ize)))
(b)	(4) acidulate	attendue : (V(A(acid))(V|A.(ate))) proposée : *(aucune)*
	(5) balanced	attendue : (A(V(balanced))) proposée : (A(N(V(balance)))(A|N.(ed)))
	(6) combustion	attendue : (N(V(combust))(N|V.(ion))) proposée : (V(combust)(N|V.(ion)))

Tableau 6. *Analyse dérivationnelle : exemples d'analyses correctes (a) et incorrectes (b)*

5. Conclusion et perspectives

Dans cet article, nous avons présenté un apprenti exploitant un mécanisme de généralisation fondé sur le calcul de rapports de proportions formels. Après une présentation du processus d'inférence, nous avons, à partir d'une définition générale des rapports de proportion, détaillé des instanciations de cette définition à diverses structures algébriques classiquement utilisées pour modéliser des données linguistiques (structures de traits, ensembles, mots sur un alphabet fini et arbres) et discuté leur implantation. Ce modèle a été implanté dans une bibliothèque générique.

Nous avons ensuite, en prenant pour domaine d'application l'apprentissage de la morphologie, montré que cet apprenti était capable de s'adapter à des situations d'apprentissage variées : la situation d'analyse de formes inconnues, dans laquelle il s'agit de prédire simultanément un ensemble d'attributs catégoriels à partir d'une représentation graphique ; une situation de décomposition hiérarchique de lexèmes construits, dans laquelle il s'agit de prédire un arbre d'analyse, de nouveau à partir d'une représentation graphique.

Pour ces deux tâches, les performances quantitatives obtenues se comparent favorablement avec celles d'un apprenti de l'état de l'art, tout en présentant un double avantage : l'apprenti ne requiert pas de pré-traitement des données et il est *bidirectionnel* (entrée et sortie sont échangeables). Pour mieux apprécier les performances opérationnelles de cette approche de la « divination morpho-syntaxique », il reste toutefois 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 figées (p. ex. 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. Les résultats obtenus sur la tâche de calcul d'une structure hiérarchique, bien qu'encourageants, restent en deçà des attentes, surtout en terme de rappel.

L'augmentation de la précision requiert, en premier lieu, la mise en place d'un algorithme exact de calcul des proportions entre structures arborées. Pour ce qui concerne l'amélioration des taux de rappel, la piste la plus prometteuse semble l'utilisation, pour le calcul de proportions entre mots (qui sont celles qui échouent le plus souvent), de versions graduelles de l'analogie. Une première proposition en ce sens est formulée dans (Yvon et al., 2004) qui remarque que les proportions entre mots se déduisent d'une conception rigide des proportions entre lettres, selon laquelle les seules proportions possibles sont $a : a :: b : b$ et $a : b :: a : b$. En autorisant d'autres types de proportions « atomiques », éventuellement pondérées, il devient possible d'augmenter, tout en le valuant, l'ensemble des proportions possibles. L'étape suivante consistera à envisager des méthodes pour inférer automatiquement ces pondérations à partir de données.

D'un point de vue linguistique, ces expériences ont permis de confirmer l'intérêt de la démarche consistant à exploiter, par des mécanismes d'apprentissage idoines, l'organisation fortement paradigmatique des bases de données lexicales pour

les trois langues européennes étudiées, confirmant les résultats présentés p. ex. dans (Lepage, 1999a; Pirrelli et Yvon, 1999b), ou encore, pour le français, dans (Stroppa et Yvon, 2004). Pour poursuivre ce travail et généraliser ces conclusions, nous envisageons, d'une part, d'étendre les tests à d'autres types de langue, présentant d'autres types de régularités morphologiques ; d'autre part, de rendre le dispositif d'apprentissage plus « naturel », en considérant, par exemple, que les formes sont présentées dans un ordre donné, ont des fréquences variables, ou encore qu'elles ne peuvent être toutes mémorisées. Un objectif à plus long terme reste enfin d'appliquer ce modèle afin d'extraire des régularités syntaxiques.

Remerciements

Le premier auteur remercie Science Foundation Ireland (Principal Investigator Award 05/IN/1732, `http://www.sfi.ie`) pour avoir aidé partiellement cette recherche. Les auteurs remercient également les relecteurs anonymes, dont les remarques ont contribué à l'amélioration de ce texte.

6. Bibliographie

Aha D. W., « Editorial », *Artificial Intelligence Review*, vol. 11, n° 1-5, p. 7-10, 1997. Special Issue on Lazy Learning.

Bayoudh S., « Learning by analogy : a classification rule for binary and nominal data », *Actes de la Conférence d'Apprentissage (CAp 2006)*, Trégastel, France, 2006.

Brill E., « Some advances in transformation-based part of speech tagging », *Proceedings of the 12thNational Conference on Artificial Intelligence*, Seattle, WA, p. 722-727, 1994.

Burnage G., CELEX : A Guide for Users, Technical report, University of Nijmegen, Center for Lexical Information, 1990.

Daelemans W., van den Bosch A., *Memory-Based Language Processing*, Studies in Natural Language Processing, Cambridge University Press, 2005.

Daelemans W., Zavrel J., van der Sloot K., van den Bosch A., TiMBL : Tilburg Memory Based Learner, version 5.1, Reference Guide, Technical Report n° 04-02, ILK, 2004.

Dal G., Namer F., « Génération et analyse automatique de ressources lexicales construites utilisables en recherche d'information », *TAL*, vol. 47, n° 2, p. 423-445, 2000.

de Marcken C., Unsupervised Language Acquisition, PhD thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1996.

Delhay A., Miclet L., « Analogie entre séquences. Définition, calcul et utilisation en apprentissage supervisé », *Revue d'Intelligence Artificielle*, vol. 19, n° 4-5, p. 683-712, 2005.

Falkenhainer B., Forbus K. D., Gentner D., « The Structure-Mapping Engine : Algorithm and Examples », *Artificial Intelligence*, vol. 41, p. 1-63, 1989.

Fradin B., *Nouvelles approches en morphologie*, Presses Universitaires de France, Paris, France, 2003.

Gaussier E., « Unsupervised Learning of Derivational Morphology from Inflectional Lexicons », *Proceedings of the ACL Workshop on Unsupervised Methods in Natural Language Processing*, College Park, MD, p. 24-30, 1999.

Gentner D., Holyoak K. J., Kokinov B. (eds), *The Analogical Mind : Perspectives from Cognitive Science*, MIT Press, Cambridge, MA, 2001.

Goldsmith J., « Unsupervised Learning of the Morphology of Natural Languages », *Computational Linguistics*, vol. 27, n° 2, p. 153-198, 2001.

Hofstadter D. R., the Fluid Analogies Research Group (eds), *Fluid Concepts and Creative Analogies : Computer Models of the Fundamental Mechanisms of Thought*, Basic Books, New York, NY, 1995.

Jones D., *Analogical Natural Language Processing*, UCL Press, London, 1996.

Krovetz B., « Viewing Morphology as an Inference Process », *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, p. 191-202, 1993.

Lepage Y., « Solving Analogies on Words : an Algorithm », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, vol. 1, Montréal, Canada, p. 728-735, 1998.

Lepage Y., « Analogy + Tables = Conjugation », *Proceedings of the International Conference on Applications of Natural Language to Data Bases (NLDB 1999)*, Klagenfurt, Germany, p. 197-201, 1999a.

Lepage Y., « Open Set Experiments with Direct Analysis by Analogy », *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS 1999)*, Beijing, China, p. 363-368, 1999b.

Lepage Y., « Analogy and Formal Languages », *Proceedings of the 6th Conference on Formal Grammar and the 7th Meeting on Mathematics of Language (FG-MOL 2001)*, Helsinki, Finland, p. 373-378, 2001.

Lepage Y., « De l'analogie rendant compte de la commutation en linguistique », , Habilitation à diriger les recherches, Grenoble, France, 2003.

Lombardy S., Régis-Gianas Y., Sakarovitch J., « Introducing Vaucanson », *Theoretical Computer Science*, vol. 328, p. 77-96, 2004.

Manning C. D., Schütze H., *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA, 1999.

Mikheev A., « Automatic Rule Induction for Unknown Word Guessing », *Computational Linguistics*, vol. 23, n° 3, p. 405-423, 1997.

Mitchell T. M., *Machine Learning*, McGraw-Hill, 1997.

Namer F., « Flemm : Un analyseur Flexionnel du Français à base de règles », *Traitement Automatique des Langues*, vol. 41, n° 2, p. 523-547, 2000.

Pirrelli V., Yvon F., « Analogy in the Lexicon : a Probe into Analogy-based Machine Learning of Language », *Proceedings of the 6th International Symposium on Human Communication*, Santiago de Cuba, Cuba, 1999a.

Pirrelli V., Yvon F., « 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*, vol. 11, p. 391-408, 1999b.

Plate T. A., « Analogy retrieval and processing with distributed vector representations », *Expert systems*, vol. 17, n° 1, p. 29-40, 2000.

Porter M., « An algorithm for suffix stripping », *Program*, vol. 14, n° 3, p. 130-137, 1980.

Régis-Gianas Y., Poss R., « On orthogonal specialization in C++ : Dealing with efficiency and algebraic abstraction in Vaucanson », *Proceedings of the Parallel/High-performance Object-Oriented Scientific Computing (POOSC 2003)*, Darmstadt, Germany, p. 71-82, 2003.

Sakarovitch J., *Éléments de théorie des automates*, Vuibert, Paris, 2003.

Schone P., Jurafsky D., « Knowledge-Free Induction of Inflectional Morphologies », *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL 2001)*, Pittsburgh, PA, 2001.

Skousen R., *Analogical Modeling of Language*, Kluwer Academic Publishers, Dordrecht, 1989.

Stroppa N., Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles, PhD thesis, École Nationale Supérieure des Télécommunications, Paris, 2005.

Stroppa N., Yvon F., « Analogies dans les séquences : un solveur à états finis », *Actes de la 11e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2004)*, Fès, Maroc, 2004.

Stroppa N., Yvon F., « An Analogical Learner for Morphological Analysis », *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, Ann Arbor, MI, p. 120-127, 2005.

Sylberztein M., *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, Paris, 1993.

Thagard P., Holyoak K. J., Nelson G., Gochfeld D., « Analog retrieval by constraint satisfaction », *Artificial Intelligence*, vol. 46, n° 3, p. 259-310, 1990.

van den Bosch A., Daelemans W., « Memory-Based Morphological Analysis », *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, Maryland, USA, p. 285-292, 1999.

Yarowsky D., Wicentowski R., « Minimally supervised morphological analysis by multimodal alignment », *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, p. 207-216, 2000.

Yvon F., « Grapheme-to-phoneme conversion using multiple unbounded overlapping chunks », *Proceedings of the conference on New Methods in Natural Language Processing (NeMLaP II)*, Ankara, Turkey, p. 218-228, 1996.

Yvon F., « Pronouncing unknown words using multi-dimensional analogies », *Proceedings of the European Conference on Speech Application and Technology (Eurospeech)*, vol. 1, Budapest, Hungary, p. 199-202, 1999.

Yvon F., Finite-state machines solving analogies on words, Technical Report n° D008, École Nationale Supérieure des Télécommunications, Paris, France, 2003.

Yvon F., Stroppa N., Delhay A., Miclet L., Solving analogies on words, Technical Report n° D005, École Nationale Supérieure des Télécommunications, Paris, France, 2004.