Moranapho : un système multilingue d'analyse morphologique fondé sur l'analogie formelle

Jean-François Lavallée* — **Philippe Langlais****

* *Centre de Recherche en Informatique de Montréal*
405, avenue Ogilvy, bureau 101
Montréal (Québec) H3N 1M3

jean-francois.lavallee@crim.ca

** *RALI / DIRO Université de Montréal*
C.P. 6128 Suc. Centre-Ville
H3C3J7 Montréal, Canada

felipe@iro.umontreal.ca

RÉSUMÉ. *Dans le cadre de Morpho Challenge 2009, nous avons mis au point un système d'analyse morphologique non supervisée qui tire profit d'analogies de formes entre mots. À l'issue de notre participation nous avons pris le temps de corriger certaines lacunes de notre système et d'en analyser les caractéristiques principales. Dans cet article, nous mettons l'accent non pas sur les performances de notre système (au demeurant plus qu'encourageantes) à cette compétition, mais plutôt sur les acquis de ces expériences postcompétition. Nous espérons que ce travail contribuera à démontrer que l'analogie formelle est non seulement une alternative viable mais également compétitive à l'analyse morphologique non supervisée.*

ABSTRACT. *Within the framework of Morpho Challenge 2009, we developed an unsupervised morphological analysis system based on formal analogies between words. After the evaluation campaign, we took time to correct some problems in our system and to analyze its main characteristics. This paper focusses on the knowledge we acquired during this post-mortem exercise. We hope this work will contribute to show that formal analogy is not only a viable solution, but is also a competitive one.*

MOTS-CLÉS : *analyse morphologique non supervisée, analogie formelle, approche à base de graphes.*

KEYWORDS: *unsupervised morphology acquisition, formal analogy, graph-based approach.*

1. Introduction

L'acquisition non supervisée d'informations morphologiques tente d'extraire à partir d'un corpus textuel non annoté des connaissances morphologiques d'une langue. Ces connaissances peuvent être exprimées ou évaluées de différentes façons. Beaucoup de travaux, incluant ceux que nous présentons ici, s'intéressent à la décomposition des mots d'une langue en leurs morphèmes[1] ; voir par exemple (Spiegler et Flach, 2010) pour une réalisation récente d'un modèle génératif de segmentation. D'autres tentent plutôt de regrouper ensemble des mots reliés morphologiquement (Hathout, 2002 ; Schone et Jurafsky, 2000). Enfin, certains s'intéressent davantage à l'utilisation de ces connaissances afin, par exemple, de lisser des modèles de langue (Ghaoui *et al.*, 2005 ; Sarikaya *et al.*, 2008) ou encore de traiter des mots hors vocabulaire dans un système de traduction (Langlais et Patry, 2007).

Dans le cadre de Morpho Challenge 2009, nous avons mis au point un système d'analyse morphologique non supervisée qui tire profit d'analogies de formes identifiées entre des quadruplets d'entrées du lexique (p. ex. `calmement` est à `calme` ce que `rapidement` est à `rapide`). À l'issue de notre participation nous avons pris le temps de corriger certaines lacunes de notre système et d'en analyser les paramètres principaux. Dans cet article, nous mettons l'accent non pas sur les performances de notre système (au demeurant plus qu'encourageantes) à cette compétition, mais plutôt sur les acquis de ces expériences postcompétition. Nous espérons démontrer que l'analogie formelle est une alternative viable aux approches proposées dans la littérature, qui permet de traiter de manière uniforme différents phénomènes morphologiques.

Le reste de l'article est organisé comme suit. Nous brossons, en section 2, quelques caractéristiques distinctives d'approches dédiées à l'acquisition non supervisée de la morphologie d'une langue. Nous rappelons, en section 3, les notions élémentaires sur l'analogie formelle qui sont nécessaires à la reproductibilité de notre approche. Nous détaillons, en section 4, le système d'analyse morphologique fondé sur l'analogie formelle que nous avons développé dans le cadre de l'édition 2009 de la campagne Morpho Challenge. Nous décrivons, en section 5, le protocole expérimental mis en place dans cette campagne et analysons les performances de *Moranapho* en comparaison à d'autres systèmes de référence. Nous disséquons ensuite, en section 6, les différents composants de notre système et tentons de caractériser ses propriétés. À la lumière de cette étude, nous discutons, en section 7, les phénomènes morphologiques pour lesquels notre système est bien adapté et ceux qui posent problème et concluons cette étude en décrivant des pistes de recherche que ce travail suggère.

2. Travaux reliés

Il existe de nombreux travaux en acquisition non supervisée de la morphologie d'une langue. Nous tentons dans cette section de caractériser quelques traits distinc-

1. La plus petite unité significative du langage.

tifs de ces travaux, sachant que les familles d'approches que nous décrivons sont loin d'être exclusives. Les travaux diffèrent tout d'abord quant au matériel utilisé en amont des systèmes. De nombreuses études, dont la nôtre, utilisent une simple liste de formes. D'autres tirent profit du contexte des mots et traitent donc des textes (Yarowsky et Wicentowski, 2000 ; Goldwater *et al.*, 2009). De manière plus marginale, Moon *et al.* (2009) montrent que l'usage de multiples documents permet d'affiner l'analyse morphologique, les termes candidats à l'appariement morphologique apparaissant naturellement dans un même document.

Le type d'approche déployé distingue également les différents travaux. Une approche répandue, que nous qualifions d'entropique, repose sur l'hypothèse que la prédiction du caractère qui suit ou précède un mot est moins fiable que celle d'un caractère interne à un mot. Harris (1955) a décrit un système fondé sur cette hypothèse. De nombreuses variantes ont depuis été évaluées (Hafer et Weiss, 1974 ; Bernhard, 2006 ; Bordag, 2006).

Une autre famille d'approches que nous appelons génératives, tente de prédire la segmentation d'un ensemble de formes à l'aide d'un lexique de segments[2] et d'un modèle calculant la probabilité d'une segmentation (étant donné l'ensemble de segments). Dans le cas de l'approche MDL (*Minimum Description Length*), la généralisation offerte par le modèle est contrôlée en pénalisant les lexiques de morphèmes sur la base de leur taille. Les systèmes populaires *Linguistica* (Goldsmith, 2001) et *Morfessor* (Creutz et Lagus, 2005b) utilisent cette approche. L'usage d'une distribution *a priori* sur les segments permet d'atteindre le même effet dans un cadre d'inférence bayésienne (Goldwater, 2006).

Une autre méthodologie consiste à regrouper les mots d'un lexique en paradigmes, puis d'identifier les affixes communs à ces groupes. Le système *ParaMor* (Monson *et al.*, 2007) appartient à cette famille et obtient régulièrement de bons résultats aux ateliers Morpho Challenge.

L'apprentissage analogique (Stroppa et Yvon, 2005) – que nous mettons à contribution dans ce travail – constitue une autre approche qui a reçu récemment l'attention de plusieurs chercheurs. Hathout montre qu'il est possible de regrouper automatiquement les mots d'un lexique en familles morphologiques en recoupant l'analogie à des données sémantiques données *a priori* (Hathout, 2002 ; Hathout, 2008). Langlais (2009) décrit un système de segmentation de mots fondé sur le seul concept d'analogie formelle ; ce système sera décrit en détail plus loin et servira de point de comparaison au système *Moranapho* que nous proposons. Goldsmith (2009) décrit également l'analogie de formes comme un moyen pertinent de réduire la redondance dans un système de type MDL.

De tous ces travaux, le système *MorphoNet* (Bernhard, 2010) est certainement celui qui se rapproche le plus du nôtre. Un ensemble de règles de transformation est extrait de paires de mots graphémiquement proches extraites elles-mêmes d'un lexique.

2. Ces segments peuvent provenir d'une quelconque des approches que nous discutons.

Ces règles servent à construire un réseau lexical dont les nœuds sont les mots du lexique, les arcs étant étiquetés par les règles permettant de transformer un mot en un autre. Un algorithme de regroupement est ensuite appliqué afin d'extraire des familles morphologiques, à partir desquelles l'analyse des mots est produite.

Notre système diffère de *MorphoNet* sur plus d'un point. Premièrement, nous utilisons l'analogie formelle pour extraire nos règles tandis que Bernhard (2010) fait usage de similarités graphémiques. Nous montrons dans cet article que l'usage des analogies, bien que beaucoup plus lourd en pratique, amène à de meilleurs résultats. Deuxièmement, les règles de *MorphoNet* ne sont pas pondérées et l'algorithme de regroupement ne tire donc aucun profit de cette information. Nous montrons que le score donné à une règle influence les performances de notre système. Enfin, nos algorithmes de regroupement et d'analyse morphologique diffèrent, bien qu'il soit difficile de les comparer. Les résultats supérieurs de *Moranapho* à Morpho Challenge 2009 confirment globalement la validité des choix que nous avons faits.

3. Analogie formelle

Une analogie (proportionnelle) est une relation entre quatre éléments notée $[x : y = z : t]$ se lisant "x est à y ce que z est à t". L'analogie formelle est un cas particulier d'analogie où la relation entre les quatre éléments est graphémique (p. ex. $[calmement : calme = rapidement : rapide]$). Plusieurs définitions de l'analogie formelle[3] ont été proposées dans la littérature. Hathout (2002) s'intéresse aux analogies mettant en œuvre des paires de formes qui partagent chacune le même préfixe (p. ex. $[marcher : marchons = parler : parlons]$). Moreau *et al.* (2007) autorisent des analogies mettant en œuvre à la fois une opération de préfixation et de suffixation comme dans $[déchargera : rechargerions = déclassera : reclasserions]$. Une contrainte sur la taille des séquences communes à chaque paire de formes est également introduite. Ces deux définitions sont des cas particuliers discutés dans (Pirrelli et Yvon, 1999).

Des définitions plus générales sont également disponibles. Lepage (1998) propose un algorithme capable de rendre compte d'opérations morphologiques plus complexes comme les cas de morphologie patron-racine présents notamment en arabe (p. ex. $[arsala : mursilun = aslama : muslimun]$[4]). Dans la présente étude, nous nous appuyons sur une définition proposée par Yvon *et al.* (2004) et reprise dans (Stroppa et Yvon, 2005). Cette définition qui s'appuie sur la notion de *factorisation* est une généralisation de la définition proposée par Lepage (1998).

Définition 1 – On appelle *n-factorisation* d'une forme x définie sur un alphabet Σ, une séquence de n *facteurs* $f_x \equiv (f_x^1, \ldots, f_x^n)$, avec $\forall i, f_x^i \in \Sigma^*$, telle que $f_x^1 \odot f_x^2 \odot \cdots \odot f_x^n = x$, où $\odot$ dénote l'opérateur de concaténation.

3. Nous utiliserons le terme d'analogie par la suite.

4. Cette analogie provient d'un tutoriel donné par Yves Lepage à TALN 2006.

Selon cette définition, (*calm*, *e*, *m*, *ent*) est une 4-factorisation du mot *calmement*. Il est important de noter que la notion de factorisation est un concept défini de manière formelle et que rien ne contraint les facteurs à correspondre à des morphes (représentations graphémiques d'un morphème) et donc *a fortiori* encore moins à des morphèmes. On peut alors définir une analogie formelle comme suit.

Définition 2 – Un quadruplet de formes (*x*, *y*, *z*, *t*) est une analogie formelle **ssi** il existe un quadruplet de n-factorisations (f_x, f_y, f_z, f_t) de *x*, *y*, *z* et *t* respectivement vérifiant : $\forall i \in [1, n] : (f^i_y, f^i_z) \in \{(f^i_x, f^i_t), (f^i_t, f^i_x)\}$. La plus petite valeur de n pour laquelle cette définition s'applique est nommée le *degré* de l'analogie ; les factorisations associées sont qualifiées de *minimales*.

Par exemple, [*calmement* : *calme* = *rapidement* : *rapide*] est une analogie formelle car le quadruplet de 4-factorisations indiqué en deuxième colonne de la figure 1 vérifie la définition 2. Il existe également une 2-factorisation de ces formes (troisième colonne) vérifiant la définition ; ainsi, l'analogie est de degré 2.

		f^1	f^2	f^3	f^4		f^1	f^2
$f_{x=calmement}$	$\equiv$	*calm*	*e*	*m*	*ent*	$\equiv$	*calme*	*ment*
$f_{y=calme}$	$\equiv$	*calm*	ϵ	*e*	ϵ	$\equiv$	*calme*	ϵ
$f_{z=rapidement}$	$\equiv$	*rapid*	*e*	*m*	*ent*	$\equiv$	*rapide*	*ment*
$f_{t=rapide}$	$\equiv$	*rapid*	ϵ	*e*	ϵ	$\equiv$	*rapide*	ϵ

Figure 1. *Deux factorisations de l'analogie* [*calmement* : *calme* = *rapidement* : *rapide*]

Définition 3 – Soit d le degré d'une analogie, et (f_x, f_y, f_z,f_t) un quadruplet de d-factorisations associé, nous appelons *alternances* les d paires de facteurs définies par, $\forall i \in [1, d]$:

$$\alpha_i/\beta_i = \begin{cases} f^i_z/f^i_x & \text{si } f^i_x \equiv f^i_y \\ f^i_z/f^i_y & \text{sinon} \end{cases}$$

Dans l'exemple de la figure 1, la factorisation minimale de la troisième colonne met en lumière les alternances : *rapide*/*calme* et *ment*/ϵ. L'alternance *ment*/ϵ capture le phénomène productif en français où un adverbe est dérivé d'un substantif ou d'un adjectif en lui ajoutant le suffixe *ment*.

4. Le système *Moranapho*

Notre système s'appuie sur la même hypothèse fondatrice que le système décrit par Langlais (2009), à savoir qu'une analogie lie entre elles des formes qui sont morphologiquement apparentées. Ainsi, [*calmement* : *calme* = *rapidement* : *rapide*] lie *calme* à *calmement* et *rapide* à *rapidement*. Nous observons cependant que

l'information capturée passivement par une analogie (ici le passage d'une forme substantive à une forme adverbiale) est fortement lexicalisée. L'analogie précédente ne nous donne, par exemple, aucune information sur la relation entre *rapide* et *rapides* ou entre *vive* et *vivement*. Ce constat nous a amené à fonder notre système sur un ensemble de *règles de réécriture* qui généralisent l'information capturée par une analogie. Ces règles sont collectées à partir des analogies identifiées en corpus par une procédure décrite dans la section suivante.

4.1. *Règles de réécriture*

4.1.1. *Format des règles*

Nous considérons les règles $\langle \alpha \rightarrow \beta \rangle$ où α et β sont les deux facteurs d'une alternance (voir la section 3). Nous imposons de plus que $|\alpha| \geq |\beta|$[5], de sorte que l'application d'une règle à un mot produise toujours un mot de taille inférieure ou égale. Nous ajoutons le symbole $\star$ à gauche et/ou à droite de ces facteurs pour indiquer l'existence d'un facteur non vide à gauche et/ou à droite de l'alternance considérée. Ceci est fait dans le but de distinguer les trois types d'affixations communes à de nombreuses langues[6].

Dans l'exemple de la figure 1, les règles $\langle \star\texttt{ment} \rightarrow \star\epsilon \rangle$ et $\langle \texttt{rapide}\star \rightarrow \texttt{calme}\star \rangle$ sont produites. L'absence du symbole $\star$ à droite de *ment* dans la première règle interdit, par exemple, de lier les formes *pigmenté* et *pigé* alors que la forme *activement* peut être liée à la forme *active*.

Nous notons $\mathcal{R}(x)$, l'application d'une règle $\mathcal{R}$ à une forme x. Par extension directe, $[\mathcal{R}_1, \ldots, \mathcal{R}_n](x)$ dénote la forme[7] résultant de l'application de n règles : $\mathcal{R}_n(\ldots \mathcal{R}_2(\mathcal{R}_1(x))\ldots)$. Dans le cas où la règle ne s'applique pas, on a $\mathcal{R}(x) = x$. Ainsi :

$$\begin{array}{lcl} \langle \star\texttt{ation} \rightarrow \star\texttt{er} \rangle(\texttt{marchandisation}) & \equiv & \texttt{marchandiser} \\ [\langle \star\texttt{ation} \rightarrow \star\texttt{er} \rangle, \langle \star\texttt{is}\star \rightarrow \star\epsilon\star \rangle](\texttt{marchandisation}) & \equiv & \texttt{marchander} \\ \langle \star\texttt{ment} \rightarrow \star\epsilon \rangle(\texttt{pigmenté}) & \equiv & \texttt{pigmenté} \end{array}$$

Obtenir les règles requiert d'identifier les analogies mettant en œuvre les mots du lexique. Dans la mesure où les règles opèrent une généralisation de l'information capturée par analogie, il n'est pas nécessaire de les identifier toutes. En pratique, toutefois, nous en avons calculé un grand nombre par lexique (voir la section 5.3.1).

5. Si les deux facteurs sont de la même taille, l'ordre alphabétique est utilisé.

6. Soit la préfixation (p. ex. *cycle* → *bi*cycle), la suffixation (p. ex. *commun* → *communiste*) et l'infixation (p. ex. *asna* → *askana*, en langue Ulwa).

7. Afin de simplifier l'exposé, nous omettons le cas où l'application d'une règle peut générer plusieurs formes.

4.1.2. *Sélection des règles*

Anderson (1992) suggère que les mots en relation analogique doivent être non seulement liés graphémiquement, mais doivent également partager des propriétés sémantiques. Notre moteur analogique n'ayant aucune information permettant d'inférer le sens des mots, plusieurs analogies fortuites (p. ex. [*age* : *sage* = *table* : *stable*]) sont identifiées, à partir desquelles des règles non pertinentes sont extraites (p. ex. $\langle s\star \rightarrow \epsilon\star \rangle$). Anderson (1992) recommande de prendre en considération un troisième facteur afin de juger de la validité d'une analogie : la régularité de la relation. Intuitivement, cela porte à croire que les règles de réécriture régulières seront plus pertinentes que les autres.

Nous pouvons estimer la régularité d'une règle par le nombre d'analogies à partir desquelles la règle a été extraite. Cependant, la fréquence crée un biais envers les règles composées de facteurs courts qui ont plus de chance d'être impliqués par hasard dans une analogie (p. ex. [*savant* : *avant* = *seau* : *eau*], [*avoir* : *savoir* = *oie* : *soie*]). Pour résoudre ce problème, nous calculons pour chaque règle sa *productivité* qui correspond au ratio entre le nombre d'applications valides d'une règle sur le nombre d'applications possibles de cette même règle. Nous estimons la productivité d'une règle $\mathcal{R}$, notée prod($\mathcal{R}$), par le ratio du nombre de fois où l'application de cette règle mène à une forme valide du lexique $\mathcal{L}$ sur le nombre de formes de $\mathcal{L}$ auxquelles elle peut être appliquée :

$$\texttt{prod}(\mathcal{R}) = \frac{|\{x \in \mathcal{L} : \mathcal{R}(x) \in \mathcal{L} \wedge \mathcal{R}(x) \neq x\}|}{|\{x \in \mathcal{L} : \mathcal{R}(x) \neq x\}|} \quad [1]$$

Un score similaire a, par exemple, été utilisé par Yvon (1997) dans son modèle génératif de prononciation par analogie. Ce score a tendance à favoriser les règles peu fréquentes. Nous avons à cet effet utilisé une troisième mesure, nommée `f-prod`, qui combine les deux mesures précédentes[8] :

$$\texttt{f-prod}(\mathcal{R}) = \texttt{prod}(\mathcal{R}) \times \frac{\log |\mathcal{R}|}{\log \sum_{\mathcal{R}} |\mathcal{R}|} \quad [2]$$

Le tableau 1 contient quelques règles extraites d'un lexique anglais utilisé dans nos expériences ainsi que les valeurs des trois métriques décrites précédemment. On observe que la règle $\langle \star a\star \rightarrow \star\epsilon\star \rangle$ (p. ex.[*caries* : *cries* = *pain* : *pin*]), malgré le fait qu'elle soit fréquente, possède une productivité faible. À l'inverse, la règle $\langle \star$*able's* $\rightarrow \star$*able*$\rangle$ (p. ex.. [*cable* : *cable's* = *vegetable* : *vegetable's*]) est peu fréquente, mais sa productivité est maximale avec `prod` et modérement importante avec `f-prod`.

Dans notre système, nous appliquons de plus un filtre qui élimine les règles ayant un poids (fréquence, `prod` ou `f-prod`) inférieur à un seuil ρ (dont l'influence est étudiée en section 5).

8. Nous avons testé différentes façons de combiner ces deux mesures, la formule indiquée étant celle pour laquelle nous avons observé de manière informelle les meilleurs résultats.

$\mathcal{R}$	$\lvert\mathcal{R}\rvert$	prod($\mathcal{R}$)	f-prod($\mathcal{R}$)
$\langle\star$'s $\rightarrow \star\epsilon\rangle$	2 225 258	0,93	0,700
$\langle\star$a$\star \rightarrow \star\epsilon\star\rangle$	288 743	0,04	0,003
$\langle\star$ing $\rightarrow \star$ed$\rangle$	4 669	0,54	0,235
$\langle\star$-based $\rightarrow \star\epsilon\rangle$	3 226	0,98	0,408
$\langle\star$co $\rightarrow \star$as$\rangle$	68	0,10	0,002
$\langle\star$able's $\rightarrow \star$able$\rangle$	65	1,00	0,215

Tableau 1. *Fréquence,* prod *et* f-prod *de règles extraites d'un lexique anglais*

4.2. *Partionnement des mots du lexique*

À l'aide des règles collectées, nous construisons une forêt d'arbres, ou ADM pour *arbre de dérivation de mot*, tel que celui représenté à gauche dans la figure 2. Les nœuds des ADMs sont les mots du lexique. Un arc entre deux nœuds n_a et n_b, noté $n_a \rightarrow n_b$, est étiqueté d'une séquence de règles qui lorsqu'elles sont appliquées au mot n_a résultent en le mot n_b (dans notre exemple, les séquences contiennent une seule règle). Chaque ADM regroupe les mots dérivant vraisemblablement de la même racine et capture donc les relations de dérivation et de flexion entre les mots du lexique. Par construction, un mot du lexique $\mathcal{L}$ n'appartient qu'à un seul ADM.

La forêt est construite à l'aide de l'algorithme 1 inspiré de l'algorithme glouton de partitionnement hiérarchique par lien simple (*single-linkage*) (Newman, 2004). Contrairement à l'algorithme standard, nous n'avons pas besoin de construire le graphe de voisinage de tous les mots de $\mathcal{L}$, mais seulement le voisinage du mot m que l'on souhaite placer dans un ADM à un moment donné (étape a). Ce voisinage est exploré en construisant pour m un graphe dont les nœuds sont des séquences de caractères atteignables par application sur m de règles de réécriture, les arcs dénotent la règle permettant de passer d'une séquence à l'autre. Ce calcul du graphe des dérivations correspond à l'étape de recherche dans l'approche de Yvon (1997). Dans notre cas, le processus est simplifié par notre hypothèse que la taille de la partie gauche d'une règle est supérieure ou égale à la taille de la partie droite (voir la section 4.1.1).

Le graphe est ensuite parcouru (étape b) afin d'identifier le mot qui possède la plus grande affinité morphologique avec m. Pour le mot $b \equiv \mathrm{argmax}_{w \in \mathcal{L} \cap \mathcal{G}}\, score(m, w)$, un arc $m \rightarrow b$ est ajouté à la forêt (étape c) et est étiqueté par la séquence de règles $[\mathcal{R}_1, \ldots, \mathcal{R}_k]$ du chemin (c) entre m et b ayant le plus haut score :

$$c \equiv \underset{[\mathcal{R}_1,\ldots,\mathcal{R}_k]:[\mathcal{R}_1,\ldots,\mathcal{R}_k](m)\equiv b}{\mathrm{argmax}} score(m, b)$$

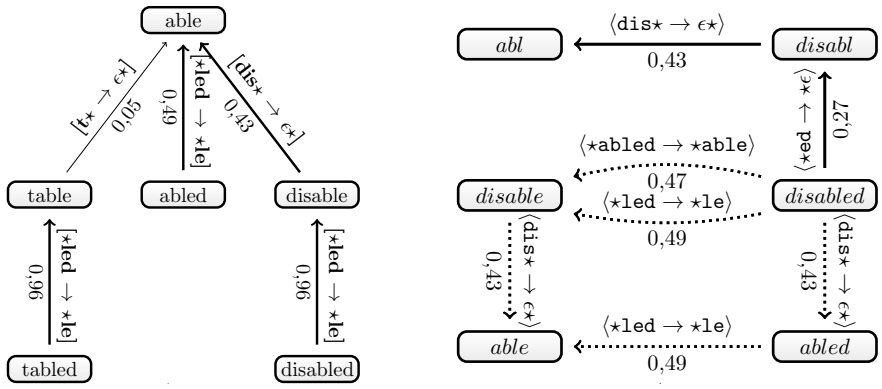

Figure 2. *À gauche : exemple d'arbre avant l'élagage (l'arc représenté en trait fin sera vraisemblablement supprimé, résultant en 2* ADMs*). À droite : graphe des séquences de caractères produites à partir du mot anglais* `disabled`*. Les arcs en pointillé sont ceux considérés lors du calcul du score entre* `disabled` *et* `able`

Algorithme 1 Algorithme de construction de la forêt $\mathcal{F}$ d'ADMs

$\mathcal{F} \leftarrow \phi$

for all mots m du lexique $\mathcal{L}$ **do**

a) Construire le graphe $\mathcal{G}$ des séquences de caractères pouvant être produites en appliquant un nombre strictement positif de règles au mot m (partie droite de la figure 2).

b) Trouver le mot b de $\mathcal{G} \cap \mathcal{L}$ qui maximise un score d'affinité morphologique au mot m (équation 3).

c) Ajouter à $\mathcal{F}$ l'arc $m \rightarrow b$.

Éliminer de $\mathcal{F}$ les arcs dont le score est inférieur à un seuil τ.

Le score d'affinité morphologique entre deux mots w_1 et w_2 de $\mathcal{G}$ est calculé en sommant le score de chacun des chemins du graphe allant de w_1 à w_2 :

$$score(w_1, w_2) = \sum_{[\mathcal{R}_1,\ldots,\mathcal{R}_n](w_1)\equiv w_2} sc([\mathcal{R}_1, \ldots, \mathcal{R}_n]) \qquad [3]$$

où le score d'un chemin entre w_1 et w_2, représenté par une séquence de k règles $[\mathcal{R}_1, \ldots, \mathcal{R}_k]$, est calculé par le produit de la valeur de la métrique (`prod` ou `f-prod`) de chacune des règles impliquées :

$$sc([\mathcal{R}_1, \ldots, \mathcal{R}_k]) = \prod_{i=1}^{k} \text{prod}(\mathcal{R}_i)$$

La partie droite de la figure 2 illustre le graphe $\mathcal{G}$ construit lors du traitement du mot *disabled*. Notons que les nœuds du graphe ne sont pas limités aux mots du lexique ; ceci permettrait par exemple de relier le mot *disabled* à *able* même si *disable* était absent du lexique. Dans notre exemple, la forme *disable* totalisant un score de $0,96$ $(0,47 + 0,49)$ est celle qui est sélectionnée à l'issue de l'étape b et la séquence $[\langle \star led \rightarrow \star le \rangle]$ étiquette dans l'ADM concerné l'arc *disabled* $\rightarrow$ *disable*.

Lorsque tous les mots de $\mathcal{L}$ ont été traités (boucle principale de l'algorithme 1), les arcs des ADMs dont les poids sont inférieurs à un seuil τ (voir la section 6.1) sont éliminés. Avec un seuil de $\tau = 0,25$, le lien entre *table* et *able* dans l'ADM de la figure 2 serait supprimé ; il en résulterait deux ADMs de nœuds racine respectifs *able* et *table*.

4.3. *Décomposition morphologique des mots du lexique*

La dernière étape de *Moranapho* consiste à dériver une analyse en morphes[9] des mots de chaque ADM. Cette opération correspond à parcourir en *profondeur d'abord* chaque ADM en calculant pour chaque nœud (mot du lexique $\mathcal{L}$) sa décomposition. Dans le cas de la racine d'un ADM, le seul morphe est le mot lui-même. Pour tout autre nœud n, l'ensemble des morphes est constitué en regroupant les morphes du nœud parent p et de ceux des règles étiquetant l'arc $n \rightarrow p$. Par exemple, dans le cas d'un ADM de deux mots qui contient l'arc *disarmed* $\rightarrow$ *arm* étiqueté par $[\langle dis\star \rightarrow \epsilon\star \rangle, \langle \star ed \rightarrow \star\epsilon \rangle]$, les morphes identifiés de *disarmed* sont [*arm*,*dis*,*ed*] ; *dis* et *ed* provenant de la partie gauche de leurs règles respectives.

Ce procédé, simple en apparence, cache un certain nombre de situations qui en pratique rendent cette étape beaucoup plus complexe qu'il n'y paraît. En particulier, la règle qui étiquette un arc de l'ADM ne correspond pas nécessairement à un morphe valide, en raison du contexte d'application contenu dans nos règles. Ceci est illustré par le lien *tabled* $\rightarrow$ *table* de la figure 2 qui est étiqueté par la règle $\langle \star led \rightarrow \star le \rangle$ au lieu de $\langle \star ed \rightarrow \star e \rangle$.

Notre solution, détaillée dans (Lavallée, 2010), fait intervenir la notion d'*équivalence de règles* selon une forme m. Deux règles $\mathcal{R}_1$ et $\mathcal{R}_2$ sont dites m-équivalentes pour une forme m donnée si l'application à m de l'une ou l'autre de ces règles amène à la même forme : $\mathcal{R}_1(m) \equiv \mathcal{R}_2(m)$. Les morphes d'un nœud m sont extraits à partir de la règle m-équivalente la plus fréquente (intuitivement, $\langle \star ed \rightarrow \star e \rangle$ est plus fréquente que $\langle \star led \rightarrow \star le \rangle$) qui est cohérente avec le reste de l'ADM. Par cohérente, nous entendons que les règles d'une même branche d'un ADM peuvent s'appliquer de manière non concurrentielle. Par exemple, dans le cas de l'ADM liant le mot *disabling* à la racine *able*, les règles $\langle \star ing \rightarrow \star ed \rangle$ et $\langle dis\star \rightarrow \epsilon\star \rangle$ sont

9. Techniquement, les segments identifiés par notre système ne sont pas nécessairement des morphes, mais nous utiliserons ce terme, plutôt que *candidat morphe*, afin de simplifier l'exposé.

cohérentes car leur application n'interfère pas, de même que les règles $\langle \star ing \rightarrow \star ed \rangle$ et $\langle \star ed \rightarrow \star e \rangle$ car elles partagent le facteur $\star ed$.

Notre procédure de décomposition souffre malgré tout du fait que le morphe associé à une règle est un choix local à chaque ADM ; il est donc possible qu'une même règle soit associée à plusieurs morphes différents d'un ADM à l'autre. Par exemple, la règle $\langle \star ed \rightarrow \star e \rangle$ pourrait produire l'analyse [cable,d] pour la forme cabled alors que cette même règle pourrait amener l'analyse [dance,ed] pour la forme danced.

Ceci pourrait être résolu en associant un morphe à chaque groupe de règles équivalentes qui serait systématiquement utilisé pour tout arc étiqueté par une règle de ce groupe. Il serait encore plus intéressant de fusionner les arbres similaires, assurant ainsi une certaine cohésion des analyses de notre système tout en intégrant le concept de paradigme à notre système. Ceci ne serait pas sans rappeler (Goldsmith, 2009) qui unifie des automates à états finis capturant les affixations possibles des mots d'un même paradigme.

5. Expériences

5.1. *Tâche et métriques d'évaluation*

La tâche que nous considérons ici est la tâche principale des compétitions Morpho Challenge. Étant donné un lexique d'une langue constitué d'une liste de mots, nous devons pour chaque mot de ce lexique produire sa décomposition morphologique.

Pour évaluer le niveau de concordance entre les résultats obtenus par un système et une référence linguistique (produite manuellement) nous calculons la f-mesure proposée par Kurimo *et al.* (2007) et qui tient lieu de métrique officielle dans les campagnes Morpho Challenge. Cette métrique, nommée MC dans la suite, repose sur l'intuition que deux mots choisis au hasard devraient partager le même nombre de morphèmes[10] communs dans les analyses produites par le système et par la référence[11]. L'évaluation en précision est faite en formant de façon aléatoire des paires de mots partageant au moins un morphème commun dans les analyses proposées. L'existence de ces paires est ensuite vérifiée dans la référence et le nombre de paires correctement identifiées est normalisé par le nombre total de paires identifiées. Le rappel est obtenu de la même façon, mais en inversant les rôles de la référence et de l'analyse suggérée. Cette métrique prend en considération le fait qu'une paire de mots peut partager plus d'un morphème et qu'il peut y avoir plus d'une analyse possible pour un mot. Un script[12] est disponible qui permet de calculer cette métrique simplement.

10. Bien qu'un système ne produise pas nécessairement des morphèmes, nous utilisons ce terme afin de simplifier l'exposé.
11. Les systèmes étant non supervisés, les étiquettes de morphèmes utilisées par un système ne correspondent *a priori* pas aux étiquettes de la référence.
12. http://www.cis.hut.fi/morphochallenge2009/evaluation.shtml

Récemment, Spiegler et Monson (2010) montraient que la métrique officielle des campagnes Morpho Challenge possède plusieurs faiblesses. Notamment, elle favorise les systèmes retournant plusieurs analyses d'un même mot et surévalue un système qui ajouterait systématiquement un morphème (factice) à chaque analyse fournie. Les auteurs proposent une métrique, EMMA, qui pallie en grande partie ces faiblesses. L'idée d'EMMA est de rechercher la meilleure correspondance entre les analyses produites automatiquement et celles de la référence, dans le but de réétiqueter les analyses candidates. L'analyse automatique et la référence deviennent alors directement comparables et il est ainsi facile de calculer les taux de précision et de rappel. Il est à noter que bien qu'EMMA ait été calculée par les organisateurs de la campagne Morpho Challenge 2010, elle n'est pas la métrique officielle qu'ils ont retenue (Kurimo *et al.*, 2010). Nous l'utiliserons cependant dans le but de vérifier que nos observations ne sont pas liées à une métrique particulière.

	Nb. mots	Nb. morphèmes distincts	$\frac{\text{Analyses}}{\text{Mot}}$	$\frac{\text{Morphèmes}}{\text{Analyse}}$	$\frac{\text{Paires}}{\text{Mot}}$
ANG.	72 628	16 388	1,07	2,15	21,93
NER.	321 926	30 620	1,12	2,78	116,25
ALL.	311 000	13 102	1,23	3,35	327,75

Tableau 2. *Principales caractéristiques de la référence* CELEX. *De gauche à droite : la taille en nombre de mots du lexique, le nombre de morphèmes distincts, la moyenne du nombre d'analyses par mot, le nombre de morphèmes moyens pour une analyse, le nombre moyen de mots avec lesquels un mot partage au moins un morphème*

5.2. *Données*

Nos expériences ont été effectuées sur deux jeux de données que nous décrivons brièvement.

5.2.1. CELEX

Ce corpus est composé de trois lexiques de langue germanique soit l'anglais (72 628 mots), l'allemand (311 000 mots) et le néerlandais (321 926 mots). Ces lexiques sont ceux de la base CELEX (Baayen *et al.*, 1995) ; ils contiennent des noms communs, des adjectifs et des verbes sans aucune erreur typographique. La décomposition morphologique de référence est obtenue en combinant les morphèmes flexionnels du mot à la décomposition morphologique du lemme associé. Nous retirons par la suite des décompositions, tous les morphèmes flexionnels qui n'ont pas de représentations graphémiques[13]. Le tableau 2 donne les principales caractéristiques de cette référence.

13. Par exemple, les morphèmes `infinitif` et `singulier` sont retirés des analyses anglaises.

5.2.2. MORPHO

Le second ensemble est constitué des données officielles de Morpho Challenge 2009[14] qui couvrent un plus grand éventail de langues. Cinq lexiques ont été extraits du Web sans aucun filtrage sur les mots et contiennent donc, en plus des mots valides de la langue, des noms propres et des erreurs typographiques. Les langues représentées dans ce jeu de données sont l'anglais (384 903 mots), l'allemand (1 266 159 mots), le finnois (2 206 719 mots), le turc (617 298 mots) et l'arabe (19 243 mots). Les analyses morphologiques de référence sont connues seulement des organisateurs.

5.3. *Résultats*

5.3.1. *Morpho Challenge 2009*

Nos premières expériences ont été effectuées dans le cadre de notre participation à Morpho Challenge 2009 qui comptait dix équipes participantes et quinze systèmes en compétition auxquels s'ajoutaient deux systèmes de référence lancés par les organisateurs, *Morfessor Baseline* (Creutz et Lagus, 2005b) et *Morfessor CatMAP* (Creutz et Lagus, 2005a), ainsi qu'un système de base, *Letters*, consistant à retourner pour décomposition les lettres du mot analysé (p. ex. [`f + i + r + e`] pour `fire`).

Nous avons soumis aux organisateurs, en plus des analyses produites par *Moranapho*[15], celles provenant du système décrit dans (Langlais, 2009) sous le nom de *Rali-Ana*. Ce système met à profit les analogies identifiées en lexique afin de collecter les factorisations minimales associées à chacun des mots impliqués par une analogie. Par exemple, la forme anglaise `abolishing` (voir la figure 3) est impliquée dans vingt et une analogies qui induisent un total de six segmentations (ou factorisations) différentes. Le système retourne alors pour un mot donné la segmentation la plus fréquente (`abolish + ing`). Comme nous l'avons déjà souligné, rien dans la définition de la factorisation que nous avons donnée en section 3 ne force un facteur à correspondre à un morphe ou un morphème. Il convient également de noter que ce système ne permet pas de traiter les mots n'apparaissant dans aucune analogie. Nous verrons qu'il s'agit d'une limitation importante de ce système.

Les analyses de ces deux systèmes ont été produites sur un même ensemble d'analogies obtenu en exécutant notre moteur analogique pendant une semaine sur les lexiques du jeu de données MORPHO fournis par les organisateurs. Nous avons ainsi recueilli un grand nombre d'analogies[16] qui, malgré leur nombre, ne constituent qu'une petite partie de toutes les analogies existantes au sein de ces lexiques.

Le tableau 3 donne les résultats officiels de l'évaluation pour nos deux soumissions ainsi que pour deux variantes du système *Morfessor* et du système *Morpho-*

14. http://www.cis.hut.fi/morphochallenge2009/datasets.shtml
15. Appelé *Rali-Cof* au moment de la soumission.
16. De 11 (arabe) à 52 (turc) millions.

abolishing (ANG.)	
abolish ing	12
ab olishing	4
abol ishing	2
a bo lishing	1
abolis hing	1
abolish in g	1

Figure 3. *Factorisations induites par analogie pour le mot anglais* `abolishing`*. La fréquence de la factorisation est indiquée à droite*

Net[17]. Le rang global a été calculé par nos soins en attribuant à chacun des systèmes un pointage calculé en additionnant son rang dans le classement pour chacune des langues, exception faite de l'arabe, en raison des résultats décevants obtenus par l'ensemble des systèmes pour cette langue. En effet, le système ayant la f-mesure la plus élevée pour cette langue est la référence naïve *Letters*. Nous renvoyons le lecteur à (Lavallée, 2010) pour une explication de ce constat d'échec.

	ANG.	FIN.	TUR.	ALL.	ARB.	Rang global /12
Moranapho	55,3	38,8	**46,4**	45,6	4,2	3
Rali-Ana	44,1	17,6	21,7	24,6	8,4	11
Morfessor Baseline	**59,8**	26,8	29,7	35,9	**12,0**	6
Morfessor CatMAP	50,5	**44,6**	45,5	**50,3**	*ND*	2
MorphoNet	55,1	33,3	41,2	41,7	9,4	5

Tableau 3. *F-mesure (métrique* MC*) et rang global des systèmes analogiques, des systèmes Morfessor et du système MorphoNet sur le jeu de données* MORPHO *évalués sur la référence* MORPHO *dans le cadre de l'atelier Morpho Challenge 2009*

Notre première observation est que *Moranapho* surpasse non seulement le système purement analogique *Rali-Ana*, mais aussi le système à base de graphes *MorphoNet* sur toutes les langues. Comme *Moranapho* et *MorphoNet* se rejoignent sur plusieurs points, ceci laisse présager que l'analogie apporte un gain réel (nous y reviendrons en section 6.4). Nous observons également que nous surpassons *Morfessor Baseline* sur toutes les langues considérées sauf l'anglais. La simplicité de la morphologie anglaise pour laquelle des procédures plus simples sont peut-être mieux adaptées à la tâche peut expliquer ce constat. Selon notre méthode de classement, nous terminons au troisième rang de la compétition, très proche de la deuxième place occupée par *Morfessor*

17. L'ensemble des résultats de la compétition sont disponibles dans (Kurimo *et al.*, 2009).

CatMAP que nous surclassons néanmoins sur deux des quatre langues. Le seul système obtenant systématiquement de meilleurs résultats que *Moranapho* est le système *ParaMor*. Ces résultats sont très encourageants et militent en faveur du potentiel de l'analogie formelle pour la tâche d'analyse formelle non supervisée.

5.3.2. *Évaluation sur* CELEX

Nous avons effectué (hors compétition) une série de tests utilisant les lexiques du jeu de données CELEX. Nous en avons profité pour tester, en plus de la productivité, la `f-prod` qui combine la productivité à la fréquence (section 4.1.2). Les résultats obtenus sur cette tâche par les deux variantes de *Moranapho* et les deux variantes du système *Morfessor* publiquement disponibles[18] se retrouvent dans le tableau 4.

	Anglais			Néerlandais			Allemand		
	Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
`prod`	76,3	57,6	65,5	69,4	43,5	53,3	54,5	**64,4**	59,0
`f-prod`	**81,1**	56,4	**66,4**	67,2	**47,1**	**55,3**	58,3	61,4	**59,6**
base	71,7	61,8	66,2	**79,5**	38,1	51,4	**80,5**	27,7	41,1
cat	64,7	**63,1**	63,7	70,4	45,5	55,1	69,0	42,5	52,5

Tableau 4. *Précision (Pr.), Rappel (Rp.) et F-mesure (F1) de la métrique* MC *pour les systèmes Moranapho (`prod` et `f-prod`) et Morfessor (`base` et `cat`) sur le jeu de données* CELEX

Une première observation est que tout comme sur les lexiques de MORPHO, *Morfessor CatMAP* surpasse *Morfessor Baseline* sur toutes les langues sauf l'anglais. Deuxièmement, nous observons que la version de *Moranapho* utilisant la mesure `f-prod` surpasse tous les autres systèmes pour toutes les langues étudiées en terme de f-mesure. Cette variante obtient un gain mineur, mais constant relativement à la version utilisant la productivité seule. Toutefois, même si les gains en termes de f-mesure en anglais et en néerlandais peuvent sembler faibles relativement à *Morfessor*, la régularité du système *Moranapho* est frappante. En effet, alors que *Morfessor Baseline* et *Morfessor CatMAP* sont bons en anglais et en néerlandais, notre système se comporte de manière stable pour ces deux langues. Pour ce qui est de l'allemand, notre système surpasse clairement les systèmes *Morfessor* avec sept points de plus de f-mesure comparé à *Morfessor CatMAP*, ceci en raison d'une avance de plus de vingt points en rappel. La différence avec *Morfessor Baseline* est encore plus grande. Ceci est particulièrement intéressant, car les bénéfices potentiels de l'analyse morphologique augmentent avec la complexité morphologique de la langue.

	ANG.	FIN.	TUR.	ALL.	ARB.	Rang global /11
Moranapho	72,8	**45,1**	41,4	58,7	29,3	3
Rali-Ana	65,2	36,1	37,3	51,4	**31,5**	8
Morfessor CatMAP	**73,1**	40,1	**45,4**	**63,1**	27,0	1
MorphoNet	72,2	37,6	39,0	59,7	31,3	5

Tableau 5. *Performance des systèmes analogiques, du système Morfessor CatMAP et du système MorphoNet évaluée par la métrique* EMMA *sur le jeu de données* MORPHO

5.3.3. *Évaluation à l'aide de la métrique* EMMA

Nous montrons au tableau 5 des extraits d'une évaluation menée par Spiegler (2010) à l'aide d'EMMA sur les analyses de la majorité des systèmes en compétition à Morpho Challenge 2009[19] ainsi que certains des meilleurs systèmes ayant participé aux campagnes précédentes[20]. Le système *Morfessor CatMAP* domine sur toutes les langues considérées à l'exception du finnois où notre système se classe premier. Globalement, nous conservons notre troisième rang derrière *Morfessor CatMAP* et le système par calcul d'incertitude de (Bernhard, 2008) qui n'était pas en compétition en 2009. Le système *ParaMor*, qui domine Morpho Challenge depuis plusieurs années, se classe en quatrième position selon cette métrique.

	Anglais			Néerlandais			Allemand		
	Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
base	80,4	78,7	79,2	**79,9**	64,5	71,3	**75,0**	47,3	57,9
cat	76,3	**79,1**	77,0	68,9	64,0	66,5	69,4	57,2	62,5
`prod`	84,7	76,6	80,4	77,5	67,2	72,0	71,4	**60,1**	**65,1**
`f-prod`	**87,7**	76,1	**81,5**	77,0	**68,6**	**72,5**	73,3	54,1	62,0

Tableau 6. *Précision (Pr.), Rappel (Rp.) et F-mesure (F1) des systèmes Moranapho (`prod` et `f-prod`) et Morfessor (`base` et `cat`) sur le jeu de données* CELEX*, selon la métrique* EMMA

Nous avons appliqué à notre tour EMMA sur le jeu de données CELEX pour les variantes de notre système évaluées au tableau 4 ainsi que pour les systèmes *Morfessor Baseline* et *Morfessor CatMAP*. Comme l'ont observé les organisateurs de Morpho Challenge 2010, la complexité calculatoire d'EMMA ne permet pas de l'appli-

18. http://www.cis.hut.fi/projects/morpho/

19. Le système *Morfessor Baseline* est absent de cette évaluation.

20. Ces systèmes sont des variantes du système *ParaMor* et le système par calcul d'incertitude de Bernhard (2008).

quer à de grandes quantités de mots. Nous avons donc employé la procédure qu'ils suggèrent et qui consiste à échantillonner aléatoirement dix sous-ensembles de 1 000 mots chacun sur lesquels EMMA est calculée. Les scores reportés dans le tableau 6 sont obtenus en moyennant ces dix mesures. Notons par soucis de transparence, que notre méthode d'échantillonnage sélectionne aléatoirement des morphèmes dans la référence, puis des mots qui les contiennent. Ceci nous assure une meilleure représentation des différents phénomènes morphologiques à traiter qu'en échantillonnant au hasard 1 000 mots, comme le suggèrent Kurimo *et al.* (2010).

Les écarts entre les systèmes sont globalement moins marqués que ceux observés aux sections précédentes. La variante `f-prod` n'est notamment plus celle qui se comporte le mieux pour toutes les langues ; *a contrario*, la variante obtient avec EMMA une meilleure f-mesure que *Morfessor CatMAP* et *Morfessor Baseline* sur les trois langues étudiées.

5.3.4. *Analyse qualitative*

Observer la qualité d'un système à travers des taux de f-mesure ne donne qu'une vue étroite du potentiel d'un système. Afin de mieux appréhender les disparités entre les résultats de *Moranapho* et de *Morfessor CatMAP*, nous rapportons en figure 4 des extraits d'analyses produites par ces deux systèmes.

		Moranapho	*Morfessor CatMAP*
Ang.	redecorating	decorate+re+ing	re/P+de/P+cor/R+ating/S
	bacteriologists	bacteriology+s+ist	bacteri/R+ologist/S+s/S
	skateboarding	skate+ing+board	skate/R+boar/R+d/S+ing/S
Néer.	officium	officia+um	offic/R+i/S+um/S
	stukoffers	offer+s+stuk	stuk/R+offers/S
	langzaamheid	langzaam+heid	lang/P+zaam/R+heid/S
All.	schulhoefen	hoefe+n+schul	schul/R+hoefe/S+n/S
	länge	lang+¨nge	länge
	nationalliberalem	liberal+m+e+national	national/P+liberal/R+em/S

Figure 4. *Extraits des analyses produites par Moranapho et Morfessor CatMAP sur le jeu de données* CELEX

Nous observons à partir de ces exemples que les deux systèmes sont aptes à capturer les cas de préfixation et de suffixation, qu'ils soient flexionnels ou dérivationnels. Toutefois, *Moranapho* semble prendre plus de risques en allemand et en néerlandais, parfois pour le mieux (p. ex. *stukoffers*), mais pas toujours (p. ex. *nationalliberalem*). En revanche, *Morfessor CatMAP* a tendance en anglais à surestimer le nombre d'affixes alors que *Moranapho* semble mieux s'adapter à la simplicité morphologique de cette langue, telle qu'illustrée par les analyses des mots *redecorating* et *skateboarding*.

Les deux approches semblent détecter correctement les cas de morphologie compositionnelle (p. ex. *skateboarding*). Cependant, *Moranapho* traite la composition de la même manière que l'affixation, même si la fréquence d'occurrences d'un affixe est beaucoup plus élevée que celle des racines formant un mot composé. Ceci a pour conséquence que seulement les racines fréquemment combinées à d'autres racines sont détectées, comme *board* dans *skateboard* ou *national* dans *nationalliberalem*, alors que les cas plus rares lui échappent.

Notre système se démarque en allemand par sa capacité à détecter les mutations du *stem*[21], comme dans le cas de *länge* (*longueur*) dérivé de *lang* (*long*)[22]. Ceci peut en partie expliquer l'écart important entre *Moranapho* et *Morfessor* sur cette langue selon la métrique MC. Les cas de mutation du *stem* (p. ex. *länge*) ou d'affixation modifiant la racine (p. ex. *bacteriologists*) sont en effet pris en compte par notre système qui, par exemple, peut établir le lien entre les mots néerlandais *officium* et *officia* contrairement à *Morfessor CatMAP*. En effet, *Morfessor CatMAP* segmente le mot aux frontières de morphèmes, ce qui le rend mal adapté pour les cas où ces dernières sont mal définies.

Un problème partagé par les deux systèmes, ainsi que par la grande majorité des méthodes non supervisées, est que les analyses produites sont des décompositions en morphes plutôt qu'en morphèmes. En effet, les deux systèmes sont, par exemple, capables d'identifier le morphe *s* de *bacteriologists* ou *en* de *oxen*, mais ne peuvent trouver le morphème *pluriel* qui est représenté par ces morphes. Ceci les rend donc vulnérables à l'allomorphie et aux morphèmes homonymes.

	Anglais		Néerlandais		Allemand	
	Morph. $\neq$	$\frac{\text{Morph.}}{\text{Mot}}$	Morph. $\neq$	$\frac{\text{Morph.}}{\text{Mot}}$	Morph. $\neq$	$\frac{\text{Morph.}}{\text{Mot}}$
Moranapho	22 487	1,93	66 006	2,69	32 788	3,20
Morfessor Baseline	23 660	1,85	68 691	1,91	58 484	1,90
Morfessor CatMAP	14 531	2,40	29 586	2,70	22 254	2,98
Référence	16 388	2,15	30 620	2,78	13 102	3,35

Tableau 7. *Le nombre de morphèmes distincts (Morph. Dist.) et le nombre moyen de morphèmes par analyse (Morph./Mot) pour les sorties des systèmes Moranapho et Morfessor sur le jeu de données* CELEX

Il est également instructif de comparer globalement la distribution des morphèmes produits par un système à la distribution de référence. Le tableau 7 montre que le

21. Ce terme tiré de l'anglais *stem mutation* englobe les phénomènes modifiant le *stem* (p. ex. l'apophonie). Le terme *stem* indique la partie du mot commune à toutes ses flexions.

22. Néanmoins, notre succès est amoindri par la représentation peu générique choisie par le système pour représenter ce morphème.

nombre de morphèmes distincts identifiés par *Morfessor CatMAP* est beaucoup plus proche de la réalité que les estimations produites par *Moranapho*. Ceci n'indique pas nécessairement que les analyses sont fausses, mais trahit plutôt un manque de cohérence de notre système lors de l'attribution d'une étiquette identifiant un morphe. Comme expliqué à la section 4.3, le problème vient du fait que ce choix est local à un ADM. Néanmoins, la moyenne de morphèmes par mot de *Moranapho* est plus proche de la réalité que celle des systèmes *Morfessor* bien que nous sous-estimons constamment cette valeur pour toutes les langues.

6. *Moranapho* à la loupe

Les analyses soumises à Morpho Challenge 2009 résultent d'un effort de développement rapide pendant lequel nous avons dû faire des choix (d'implémentation, de valeurs d'hyperparamètres) que nous n'avons pas eu l'occasion de tester. Après la compétition, nous avons entrepris de mieux comprendre les limites de notre système que nous avons en partie réécrit, soit pour corriger certains problèmes (inévitables dans un logiciel de plus de 10 000 lignes de code C++ écrit à la hâte), soit pour ajuster certains procédés.

Dans cette section, nous mesurons tout d'abord (section 6.1) l'impact des hyperparamètres sur notre système. Nous nous interrogeons ensuite (section 6.2) sur la flexibilité que l'on doit laisser au système lors de l'identification des analogies. Nous étudions ensuite (section 6.3) l'aptitude de *Moranapho* à gérer les umlaut en allemand. Enfin, nous vérifions (section 6.4) que l'analogie est bien utile à l'acquisition des règles que nous utilisons dans notre système.

6.1. *Les hyperparamètres*

Le tableau 8 montre les résultats obtenus par notre système sur le jeu de données du Morpho Challenge MORPHO évalués sur la référence CELEX (anglais et allemand) selon les différentes valeurs d'hyperparamètres qui le contrôlent: ρ, le poids minimal d'une règle, et τ, le score minimal qu'un chemin entre deux mots doit avoir pour que ces mots soient regroupés dans un même ADM.

On peut premièrement constater que les résultats obtenus sur la référence CELEX sont similaires à ceux mesurés sur la référence MORPHO avec les mêmes valeurs d'hyperparamètres (voir le tableau 3). Ceci n'est pas surprenant, car les deux références sont toutes les deux extraites de CELEX. Par conséquent, il est fort probable que les valeurs d'hyperparamètres utilisées pour Morpho Challenge 2009 ne sont pas optimales, car ces mêmes valeurs sur la référence CELEX sont loin de donner les meilleurs résultats.

Comme on pouvait s'y attendre, augmenter le seuil τ ou la productivité minimale ρ améliore la précision au détriment du rappel. La valeur optimale du seuil τ pour l'anglais est de 0,3 contrairement à 0,25 pour l'allemand. Cette différence s'explique

		$\rho = 0{,}20$			$\rho = 0{,}25$			$\rho = 0{,}30$		
		Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
$\tau = 0{,}20$	ANG.	52,0	52,1	51,8	59,6	47,9	52,9	66,4	44,8	53,3
	ALL.	60,0	42,6	49,4	67,0	37,7	47,8	70,7	32,1	43,7
$\tau = 0{,}25$	ANG.	61,0	48,9	54,0	60,6	47,9	53,3	65,7	44,8	53,1
	ALL.	**69,0**	**41,7**	**52,2**	67,1	40,2	49,6	70,6	32,1	43,7
$\tau = 0{,}30$	ANG.	**66,0**	**46,8**	**54,5**	66,2	45,3	53,6	67,0	44,7	53,5
	ALL.	69,0	41,7	51,4	69,9	36,2	47,0	71,0	32,1	43,8
$\tau = 0{,}35$	ANG.	71,5	43,1	53,5	71,3	40,8	51,7	72,5	39,6	50,9
	ALL.	70,7	33,4	44,8	72,1	30,0	41,6	74,1	26,0	37,9
$\tau = 0{,}40$	ANG.	74,2	41,2	52,8	76,1	38,9	51,3	76,7	37,6	50,3
	ALL.	71,5	32,4	44,0	75,0	29,0	41,1	76,0	24,4	36,3

Tableau 8. *Impact des hyperparamètres ρ et τ sur la Précision (Pr.), le Rappel (Rp.) et la F-Mesure (F1) de notre système sur le jeu de données* MORPHO *évalué sur la référence* CELEX*, métrique* MC*. La configuration testée à Morpho Challenge 2009 est soulignée et les meilleurs résultats sont en gras*

par le fait qu'en moyenne, le nombre de mots en relation est plus élevé pour l'allemand que pour l'anglais (voir le tableau 2). Donc en diminuant τ, plus de relations sont créées, ce qui avantage les langues morphologiquement plus complexes. Pour ce qui est du seuil ρ, si la valeur de τ est bien choisie, on gagne à lui attribuer la valeur la plus petite possible. Ceci augmente cependant les temps de calcul, et les gains obtenus par la diminution de ρ deviennent minimes au-delà d'un certain point.

6.2. *Impact du degré de l'analogie*

Comme mentionné à la section 3, le degré d'une analogie indique le nombre minimal de cofacteurs impliqués dans l'analogie. Par exemple, l'analogie [*capital* : *anticapitalisme* = *commun* : *anticommunisme*] implique trois cofacteurs *anti*/ϵ, *commun*/*capital* et *isme*/ϵ et est donc de degré 3.

Puisque les mots anglais contiennent rarement plus de trois morphèmes, il est peut-être avantageux pour cette langue de retirer les analogies de degré élevé ; d'une part pour accélérer le processus d'acquisition des analogies et d'autre part, pour éviter d'introduire du bruit dans les analyses. Afin de vérifier ceci, nous avons retiré de notre ensemble d'analogies celles ayant un degré supérieur à un seuil donné ; le reste du système demeurant, quant à lui, inchangé. Les résultats de cette expérience sont résumés à la figure 5.

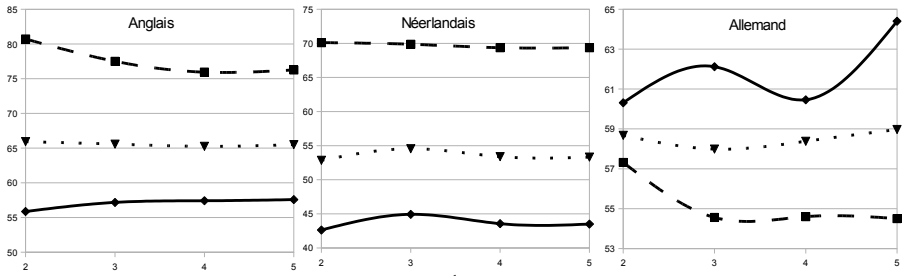

Figure 5. *Précision (tiret), Rappel (pleine) et F-mesure (pointillé) de Moranapho en fonction du degré maximal toléré pour une analogie. Les analyses ont été produites sur le jeu de données* CELEX *et ont été évaluées sur la référence* CELEX

Pour l'anglais et le néerlandais, le degré n'a pas un impact important sur la f-mesure. La courbe pour le néerlandais atteint son apogée au degré 3 alors que l'anglais l'atteint au degré 2 en raison d'une augmentation importante de la précision. Il est intéressant de noter que ces valeurs correspondent à la moyenne de morphèmes par mot de leurs langues respectives (voir le tableau 2). La courbe pour l'allemand est en revanche plus surprenante. C'est la seule langue étudiée pour laquelle la f-mesure augmente si l'on permet des analogies de plus haut degré. On pouvait s'y attendre, car les mots allemands contiennent plus de morphèmes en moyenne. L'allemand est également la seule langue étudiée où les règles qui ne peuvent être extraites par une analogie de degré 2 sont communes. Un exemple typique de ceci est la mutation du *stem* capturée par l'analogie de degré 4 $[l\underline{a}ng : l\ddot{\underline{a}}nge = st\underline{a}rk : st\ddot{\underline{a}}rke]$. De plus, autoriser les analogies de degrés supérieurs à 2 peut être utile dans les cas de composition impliquant plusieurs racines tels que $[\underline{talg}s : \underline{talg}lichts = \underline{tee} : \underline{tee}licht]$ (degré 3) ou $[atom\underline{kraft}werken : atom\underline{krieg}en = \underline{kraft}werks : \underline{krieg}s]$ (degré 4). La perte en rappel entre les degrés 3 et 4 en allemand est due à l'introduction d'une règle causant une inconsistance dans l'étiquetage d'un morphe (voir la section 4.3 pour une explication du problème).

6.3. *Étude d'un phénomène non concaténatif*

Nous avons souligné que notre système traitait de manière homogène différents phénomènes morphologiques. Nous souhaitons, dans cette section, vérifier plus précisément son aptitude à capturer des phénomènes de mutation du *stem* en allemand en mesurant à l'aide de la métrique EMMA l'aptitude de *Moranapho* à identifier la racine d'un mot dans les cas d'umlaut.

Nous avons pour cela extrait de la référence de CELEX les mots contenant une voyelle accentuée d'un tréma (p. ex. ä, ë...) et pour lesquels l'analyse morphologique de référence contenait deux morphèmes (afin de concentrer l'évaluation sur la racine

	Pr.	Rp.	F1
Morfessor CatMAP	58,3	58,1	58,2
`prod`	**73,2**	**73,1**	**73,2**
`f-prod`	65,5	65,3	65,4

Tableau 9. *Précision (Pr.), rappel (Rp.) et F-mesure (F1) selon la métrique* EMMA *des systèmes Moranapho (`prod` et `f-prod`) et du système Morfessor CatMAP sur la tâche d'identification de la racine*

du mot) et aucun tréma (p. ex. *`körbe`* : *`korb`* `+PL`). Seule la racine a été conservée dans les décompositions et un seul mot de la même famille a été retenu afin de ne pas comptabiliser des liens qui ne proviendraient pas de la bonne identification de la racine. La racine a finalement été ajoutée dans la liste des analyses, de sorte que des formes et leurs décompositions, à gauche de la figure 6, on ne retienne que celles de droite :

CELEX				Référence		
kommissär	:	*kommissar* `+ACC`				
kommissäre	:	*kommissar* `+PL`	$\Longrightarrow$	*kommissäre*	:	kommissar
kommissären	:	*kommissar* `+PL`		*kommissar*	:	kommissar
kommissäres	:	*kommissar* `+GEN`				
		...				

Figure 6. *Construction de la référence pour les cas d'umlaut en allemand*

Les sorties des systèmes *Moranapho* et *Morfessor CatMAP* (faisant usage de tout le lexique CELEX) sont évaluées sur les 974 termes de la référence ainsi construite. Les performances de ces systèmes sont résumées au tableau 9. On observe que les scores sont relativement élevés quelle que soit la méthode, pour la simple raison que ne pas décomposer un mot est payant dans la moitié des cas (car la racine est ajoutée à la référence avec elle-même comme décomposition). En fait, l'approche *Morfessor CatMAP* obtient une performance à peine supérieure à un système qui ne réaliserait aucune décomposition et qui obtiendrait 50 % de précision et de rappel. On observe que le meilleur système est *Moranapho* faisant usage de la métrique `prod`. L'autre variante a tendance à décomposer la racine ; elle est donc pénalisée par la métrique EMMA.

6.4. *Impact du processus d'acquisition de règles*

Les résultats que nous avons présentés démontrent que des performances rivalisant avec l'état de l'art peuvent être obtenues par un système fondé sur l'analogie

formelle. Toutefois, ceci ne prouve pas que d'autres façons d'acquérir les règles de réécriture ne donneraient pas de meilleurs résultats. Pour cette raison, nous avons développé une variante de *Moranapho* qui remplace l'analogie formelle par une méthode d'extraction de règles fondée sur la distance de Levensthein où le coût des opérations d'insertion, de suppression et de substitution est unitaire. Notre processus est inspiré par (Bernhard, 2010) et consiste à obtenir pour chaque mot de taille supérieure à la moyenne, les vingt mots du lexique les plus similaires à celui-ci selon la distance d'édition. Pour chacune de ces paires, nous calculons l'alignement de distance d'édition minimale[23] et les sections non alignées sont transformées en règles de réécriture.

Par exemple, l'alignement entre `inconcevable` et `concevoir` de la figure 7 génère les règles $\langle in\star \rightarrow \epsilon\star \rangle$ et $\langle \star\texttt{able} \rightarrow \star\texttt{oir} \rangle$. Cette façon de procéder est beaucoup plus rapide qu'identifier les analogies parmi tous les mots de $\mathcal{L}$.

i	*n*	**c**	**o**	**n**	**c**	**e**	**v**	*a*	*b*	*l*	*e*
ϵ	ϵ	**c**	**o**	**n**	**c**	**e**	**v**	*o*	*i*	*r*	

Figure 7. *Alignement obtenu par distance d'édition entre les mots* `inconcevable` *et* `concevoir`

Puisque le calcul de la productivité n'utilise aucune information particulière à l'analogie, nous l'utilisons pour noter les règles ainsi obtenues, tout comme nous le faisons dans la version analogique. Le tableau 10 montre que l'utilisation de la distance d'édition réduit la qualité des analyses pour toutes les langues bien que la différence soit minime en anglais. L'analogie ne semble donc pas apporter de bénéfice marquant sur les langues morphologiquement peu complexes, mais s'avère nettement préférable pour des langues plus complexes. Ce résultat n'exclut pas la possibilité qu'une distance d'édition ajustée puisse servir de premier filtre au calcul plus coûteux des analogies. Notre système sacrifierait alors sa simplicité conceptuelle au profit de sa rapidité de traitement.

	Anglais			Néerlandais			Allemand		
	Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
Moranapho ANA.	76,3	**57,6**	**65,5**	69,4	**43,5**	**53,3**	54,5	**64,4**	**59,0**
Moranapho D.E.	**78,7**	55,2	64,8	**74,9**	33,2	45,9	**69,5**	45,9	55,2

Tableau 10. *Précision (Pr.), rappel (Rp.) et F-mesure (F1) de Moranapho à base d'analogie (Ana) et de sa variante utilisant la distance d'édition (D.E.) sur le jeu de données* CELEX*, métrique* MC

23. Il peut en exister plusieurs, auquel cas tous sont conservés.

7. Discussion

Nous décrivons dans cet article un système fondé sur l'identification en corpus d'analogies formelles entre mots. Ces analogies sont analysées afin d'en extraire un ensemble de règles de réécriture qui sont mises à profit pour regrouper entre eux les mots du lexique qui sont reliés morphologiquement. Ce système a montré de bonnes performances à la campagne d'évaluation Morpho Challenge 2009 et s'est avéré meilleur que le système analogique de Langlais (2009). Des expériences réalisées après cette campagne nous ont permis de mieux comprendre les forces et les faiblesses de notre système. Parmi les résultats notables de ces expériences, nous constatons que la variante de notre système faisant usage du score `f-prod` surpasse les systèmes à l'état de l'art *Morfessor Baseline* et *Morfessor CatMAP*, et ce, pour les trois langues disponibles dans CELEX. La différence de performances mesurée en f-mesure est plus importante pour l'allemand, langue où des phénomènes de composition et de mutation du *stem* rendent la tâche plus difficile. Nous avons également montré que le choix des deux hyperparamètres qui contrôlent *Moranapho* gagnerait à être ajusté pour une langue donnée. Enfin, nous avons révélé que l'usage de la distance d'édition amenait à des règles moins fiables que celles obtenues par analogie formelle.

Contrairement à de nombreux autres systèmes, *Moranapho* peut produire des analyses qui ne sont pas de simples segmentations du mot d'origine ; les cas de mutation du *stem* et les affixations modifiant la racine sont, par exemple, pris en compte par notre système, comme nous l'avons illustré en figure 4. *Moranapho* gère également les phénomènes de composition, sans qu'un traitement particulier ne leur soit dédié. Cette simplicité conceptuelle contraste grandement avec les deux seuls systèmes non supervisés qui ont participé à Morpho Challenge 2010 et qui sont essentiellement des approches concaténatives. Lignos (2010) apporte à l'approche décrite dans (Lignos *et al.*, 2009) différents traitements dédiés permettant d'augmenter la couverture de leur analyseur en gérant notamment le cas des mots composés. Certaines variantes de ce système réussissent à surpasser le système *Morfessor CatMAP* sur l'anglais et le finnois, mais aucune variante ne parvient à le surpasser pour l'allemand et le turc. Nicolas *et al.* (2010) décrivent une approche concaténative, contenant de nombreuses heuristiques, et qui ne dépasse l'approche *Morfessor CatMAP* sur aucune langue selon la métrique EMMA.

Notre format de règles ne permet cependant pas de faire le lien entre certaines formes reliées morphologiquement, et ce, même si la relation a été identifiée par nos analogies. Il existe en effet certaines contraintes que les règles de réécriture doivent respecter pour être considérées comme valides par notre système. Premièrement, nous imposons que $|\alpha| \geq |\beta|$, de sorte que l'application d'une règle à un mot produise toujours un mot de taille inférieure ou égale. Deuxièmement, le symbole $\star$ ne peut se retrouver qu'aux extrémités de β et/ou de α. Ceci nous permet de différencier la préfixation de la suffixation et de l'infixation, mais contraint les transformations capturées à être contiguës. Par conséquent, certaines relations morphologiques pourront difficilement être représentées par une seule règle (p. ex. $\langle \texttt{suggère} \rightarrow \texttt{suggérer} \rangle$). Si ce choix peut s'avérer bénéfique pour les phénomènes morphologiques les plus

courants, il convient mal aux cas de morphologie patron-racine qui caractérisent des langues templatiques telles que l'arabe.

Ainsi, aucune règle valide selon notre format ne peut faire le lien entre les formes arabes *KaaTiB* et *KuTaaB*, et ce, même si la relation est « connue » du système grâce à des analogies comme $[\mathit{KaaTiB} : \mathit{KuTaaB} = \mathit{QaaRi'} : \mathit{QuRaa'}]$. Pour établir le lien entre ces deux mots, il nous aurait fallu obtenir la règle $\langle \star u \star aa \star \rightarrow \star aa \star i \star \rangle$ qui ne respecte pas la deuxième contrainte que nous venons de rappeler. Il serait également possible d'identifier la relation par l'application successive des règles $\langle \star u \star \rightarrow \star aa \star \rangle$ et $\langle \star aa \star \rightarrow \star i \star \rangle$; malheureusement $\langle \star u \star \rightarrow \star aa \star \rangle$ ne respecte pas la première contrainte.

D'autres formats de règles capables de gérer les cas de morphologie patron-racine ont été suggérés dans la littérature (Bernhard, 2010 ; Neuvel et Fulop, 2002), mais il nous reste toutefois à vérifier que leur usage n'introduirait pas trop de bruit dans nos analyses. Ceci n'est pas improbable compte tenu du fait que les résultats de *Moranapho* à Morpho Challenge 2009 étaient supérieurs à ceux de *MorphoNet*.

Nous avons par ailleurs relevé un ensemble de faiblesses de certains processus de notre système. En particulier, nous avons mentionné que l'analyse morphologique d'un mot (voir la section 4.3) dépend du choix de l'étiquette associée à un morphe et que ce choix est local à chaque ADM. Les analyses produites par *Moranapho* manquent donc de cohérence. Nous avons de plus observé que les regroupements de mots opérés par notre système (voir la section 4.2) présentent une tendance à la sur-segmentation. Ce sont deux problèmes que nous devons étudier. Aussi, nous n'avons pas réalisé l'étude du nombre d'analogies nécessaires aux bonnes performances de *Moranapho* : les expériences réalisées ici font toute usage d'un ensemble d'analogies calculé sur une période d'une semaine. Nous souhaitons donc mesurer l'influence du nombre d'analogies sur les performances de notre système.

Les expériences que nous présentons en section 6.1 montrent qu'il est possible (voire préférable) d'ajuster les hyperparamètres de *Moranapho* à une langue particulière. Une continuation naturelle de notre participation à Morpho Challenge 2009 consisterait à comparer notre système non supervisé à une variante ajustée à l'aide d'un corpus de développement. En fait, quatre des quinze systèmes ayant concouru à la campagne Morpho Challenge 2010 sont des systèmes semi-supervisés qui font usage d'un corpus de développement de 1 000 mots dont l'analyse morphologique était fournie par les organisateurs. Kohonen *et al.* (2010) décrivent notamment une extension semi-supervisée de *Morfessor Baseline*.

Enfin, si, comme nous espérons l'avoir démontré, *Moranapho* possède des caractéristiques intéressantes et, somme toute, atypiques dans le paysage des analyseurs morphologiques, il n'en reste pas moins que le processus guidant l'analyse produite par notre système reste pour le moment entièrement procédural. Notre objectif à moyen terme est d'intégrer les bénéfices de l'analogie formelle dans un modèle génératif. Les travaux de Goldsmith (2009) offrent une première piste possible d'intégration de l'information analogique dans un modèle MDL.

Remerciements

Ce travail a été permis grâce à une subvention MITACS. Nous remercions les relecteurs de la première version de cet article qui par leurs commentaires pertinents ont grandement contribué à l'amélioration de cet article.

8. Bibliographie

Anderson S. R., *A-Morphous Morphology*, Cambridge University Press, Cambridge, 1992.

Baayen R. H., Piepenbrock R., Gulikers L., « The CELEX lexical database (release 2) », , CD-ROM, Linguistic Data Consortium, Univ. of Pennsylvania, USA, 1995.

Bernhard D., « Unsupervised morphological segmentation based on segment predictability and word segments alignment », *2nd Pascal Challenges Workshop*, Venice, Italy, p. 19-24, 2006.

Bernhard D., « Simple Morpheme Labelling in Unsupervised Morpheme Analysis », *Lecture Notes in Computer Science : Cross Language Evaluation Forum 2007 (CLEF'07)*, Springer, Berlin, p. 873-880, 2008.

Bernhard D., « MorphoNet: Exploring the Use of Community Structure for Unsupervised Morpheme Analysis », *Lecture Notes in Computer Science : Cross Language Evaluation Forum 2009 (CLEF'09)*, Springer, Corfu, 2010.

Bordag S., « Two-step approach to unsupervised morpheme segmentation », *2nd Pascal Challenges Workshop*, Venice, Italy, p. 25-29, 2006.

Creutz M., Lagus K., « Inducing the Morphological Lexicon of a Natural Language from Unannotated Text », *Adaptive Knowledge Representation and Reasoning 2005 (AKRR'05)*, vol. 5, Espoo, p. 106-113, 2005a.

Creutz M., Lagus K., « Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. », *Publications in Computer and Information Science, Report A81*, Helsinki University of Technology, 2005b.

Ghaoui A., Yvon F., Mokbel C., Chollet G., « On the Use of Morphological Constraints in N-gram Statistical Language Model », *Interspeech*, Lisbon, Portugal, p. 1281-1284, Sept., 2005.

Goldsmith J., « Unsupervised Learning of the Morphology of a Natural Language », *Computational Linguistics*, vol. 27, p. 153-198, 2001.

Goldsmith J., « Morphological Analogy: Only a beginning », *Analogy in Grammar*, vol. 28, p. 137-164, 2009.

Goldwater S., Nonparametric Bayesian models of lexical acquisition, PhD thesis, Brown University, 2006.

Goldwater S., Griffiths T. L., Johnson M., « A Bayesian framework for word segmentation: Exploring the effects of context », *Cognition*, vol. 112, n° 1, p. 21-54, 2009.

Hafer M., Weiss S., « Word Segmentation by Letter Successor Varieties », *Information Storage and Retrieval*, vol. 10, p. 371-385, 1974.

Harris Z. S., « From Phoneme to Morpheme », *Language*, vol. 31, n° 2, p. 190-222, 1955.

Hathout N., « From wordnet to celex: acquiring morphological links from dictionaries of synonyms », *Language Resources and Evaluation 2002 (LREC'02)*, Las Palmas de Gran Canaria, p. 1478-1484, 2002.

Hathout N., « Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy », *Graph-based Methods for Natural Language Processing 2008 (Textgraphs'08)*, Manchester, p. 1-8, 2008.

Kohonen O., Virpioja S., Leppänen L., Lagus K., « Semi-supervised extensions to Morfessor Baseline », *in* M. Kurimo, S. Virpioja, V. T. Turunen (eds), *Proceedings of the Morpho Challenge 2010 Workshop*, TKK Reports in Information and Computer Science, TKK-ICS-R37, p. 30-34, 2010.

Kurimo M., Creutz M., Varjokallio M., « Unsupervised morpheme analysis evaluation by a comparison to a linguistic Gold Standard — Morpho Challenge 2007 », *Notes de travail du Cross Language Evaluation Forum 2007 (CLEF'07)*, Budapest, 2007.

Kurimo M., Virpioja S., Turunen V., Blackwood G., Byrne W., « Overview and Results of Morpho Challenge 2009. », *Notes de travail du Cross Language Evaluation Forum 2009 (CLEF'09)*, Corfu, 2009.

Kurimo M., Virpioja S., Turunen V. T., « Overview and Results of Morpho Challenge 2010 », *in* M. Kurimo, S. Virpioja, V. T. Turunen (eds), *Proceedings of the Morpho Challenge 2010 Workshop*, TKK Reports in Information and Computer Science, TKK-ICS-R37, p. 7-24, 2010.

Langlais P., « Étude quantitative de liens entre l'analogie formelle et la morphologie constructionnelle », *Traitement automatique des langues naturelles 2009 (TALN'09)*, Senlis, 2009.

Langlais P., Patry A., « Translating Unknown Words by Analogical Learning », *Empirical Methods in Natural Language Processing - Computational Natural Language Learning 2007(EMNLP-CoNLL'07)*, Prague, p. 877-886, 2007.

Lavallée J.-F.*, « Moranapho: Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques »*, Master's thesis, Université de Montréal, 2010.

Lepage Y., « Solving Analogies on Words: an Algorithm », *COLING-ACL*, Montreal, Canada, p. 728-734, 1998.

Lignos C., « Learning from Unseen Data », *in* M. Kurimo, S. Virpioja, V. T. Turunen (eds), *Proceedings of the Morpho Challenge 2010 Workshop*, TKK Reports in Information and Computer Science, TKK-ICS-R37, p. 35-38, 2010.

Lignos C., Chanz E., Marcus M. P., Charles Y., « A Rule-Based Unsupervised Morphology Learning Framework », *Lecture Notes in Computer Science : Cross Language Evaluation Forum 2009 (CLEF'09)*, Springer, Corfu, p. 618-625, 2009.

Monson C., Carbonell J., Lavie A., Levin L., « ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis », *Special Interest Group on Computational Morphology and Phonology (SIGMORPHON'07)*, ACL, Prague, p. 117-125, 2007.

Moon T., Erk K., Baldridge J., « Unsupervised morphological segmentation and clustering with document boundaries », *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, p. 668-677, August, 2009.

Moreau F., Claveau V., Sébillot P., « Automatic Morphological Query Expansion Using Analogy-Based Machine Learning », *29th European Conference on Information Retrieval (ECIR 2007)*, Roma, Italy, 2007.

Neuvel S., Fulop S. A., « Unsupervised learning of morphology without morphemes », *Special Interest Group on Computational Morphology and Phonology (SIGMORPHON'02)*, ACL Publications, p. 31-40, 2002.

Newman M. E. J., « Detecting community structure in network », *The European Physical Journal B*, vol. 38, p. 321-330, 2004.

Nicolas L., Farré J., Molinero M. A., « Unsupervised learning of concatenative morphology based on frequency-related from occurrence », *in* M. Kurimo, S. Virpioja, V. T. Turunen (eds), *Proceedings of the Morpho Challenge 2010 Workshop*, TKK Reports in Information and Computer Science, TKK-ICS-R37, p. 39-43, 2010.

Pirrelli V., Yvon F., « Analogy in the lexicon: a probe into analogy-based machine learning of language », 1999.

Sarikaya R., Afify M., Deng Y., Erdogan H., Gao Y., « Joint Morphological-Lexical Language Modeling for Processing Morphologically Rich Languages With Application to Dialectal Arabic », *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, n° 7, p. 1330 -1339, sept., 2008.

Schone P., Jurafsky D., « Knowledge-free induction of morphology using latent semantic analysis », *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*, ConLL '00, Association for Computational Linguistics, p. 67-72, 2000.

Spiegler S., EMMA: A Novel Evaluation Metric for Morphological Analysis - Experimental Results in Detail, Technical Report n° CSTR-10-004, University of Bristol, Bristol, 2010.

Spiegler S., Flach P., « Enhanced word decomposition by calibrating the decision threshold of probabilistic models and using a model ensemble », *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, p. 375-383, 2010.

Spiegler S., Monson C., « EMMA: A Novel Evaluation Metric for Morphological Analysis », *Conference on Computational Linguistics 2010 (COLING'10)*, Beijing, 2010.

Stroppa N., Yvon F., « An analogical learner for morphological analysis », *Computational Natural Language Learning 2005 (CoNLL'05)*, Ann Arbor, p. 120-127, 2005.

Yarowsky D., Wicentowski R., « Minimally supervised morphological analysis by multimodal alignment », *Association for Computational Linguistics 2010 (ACL'10)*, Association for Computational Linguistics, Morristown, p. 207-216, 2000.

Yvon F., « Paradigmatic Cascades: a linguistically Sound Model of Pronunciation by Analogy », *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, p. 428-435, 1997.

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