TALN 2009, Senlis, 24–26 juin 2009

Étude quantitative de liens entre l'analogie formelle et la morphologie constructionnelle

Philippe Langlais
Département d'Informatique et de Recherche Opérationnelle
Université de montreal
C.P. 6128 Suc. Centre-Ville
Montréal, H3C3J7, Qc, Canada
felipe@iro.umontreal.ca

Résumé. Plusieurs travaux ont récemment étudié l'apport de l'apprentissage analogique dans des applications du traitement automatique des langues comme la traduction automatique, ou la recherche d'information. Il est souvent admis que les relations analogiques de forme entre les mots capturent des informations de nature morphologique. Le but de cette étude est de présenter une analyse des points de rencontre entre l'analyse morphologique et les analogies de forme. C'est à notre connaissance la première étude de ce type portant sur des corpus de grande taille et sur plusieurs langues. Bien que notre étude ne soit pas dédiée à une tâche particulière du traitement des langues, nous montrons cependant que le principe d'analogie permet de segmenter des mots en morphèmes avec une bonne précision.

Abstract. Several studies recently showed the interest of analogical learning for Natural Language processing tasks such as Machine Translation and Information Retrieval. It is often admitted that formal analogies between words capture morphological information. The purpose of this study os to quantify the correlations between morphological analysis and formal analogies. This is to our knowledge the first attempt to conduct such a quantitative analysis on large datasets and for several languages. Although this paper was not geared toward tackling a specific natural language processing task, we show that segmenting a word token into morphemes can be accomplished with a good precision by a simple strategy relying solely on formal analogy.

Mots-clés : Apprentissage analogique, analogie formelle, analyse morphologique.

Keywords: Analogical Learning, Formal Analogies, Morphological Analysis.

1 Introduction

Une *proportion analogique* ou analogie, est une relation entre quatre items notée [x:y::z:t] qui se lit x est à y ce que z est à t. Dans cette étude, nous nous intéressons aux seules *analogies formelles*, c'est-à-dire des proportions repérées sur la seule base de relations graphiques entre les formes [1] en présence. Par exemple, l'analogie [*marche*:*démarchage*::*friche*:*défrichage*] capture que le mécanisme constructionnel liant les deux premières formes est identique à celui liant les deux dernières. Nous fournirons plus loin une définition opérationnelle de l'analogie formelle.

Plusieurs travaux récents ont montré que l'apprentissage analogique basé sur l'analogie formelle (ou analogie de forme) offre un cadre théorique adapté à différents problèmes canoniques du traitement automatique des langues. Nombreux de ces travaux s'intéressent à la traduction automatique, inspirés pour une large part des travaux de Lepage et Denoual (2005) qui montrent qu'il est possible de traduire des phrases d'un domaine limité [2] à l'aide du seul concept d'analogie formelle. Ils illustrent le potentiel de l'approche pour cinq directions de traduction (japonais, chinois, coréen, arabe vers l'anglais et anglais vers chinois). Langlais & Patry (2008) montrent à leur tour que ce concept permet de traduire des mots inconnus pour les paires de langues français, espagnol et allemand de et vers l'anglais ; une idée développée en parallèle par Denoual (2007) pour la paire de langues japonais/anglais. Langlais *et al.* (2009) montrent qu'il est également possible de traduire à l'aide du même principe des multi-termes du domaine médical et ce, pour dix directions de traduction (français, suéduois, finnois, espagnol et russe vers et depuis l'anglais).

Des travaux ont aussi été menés en recherche d'information où Moreau *et al.* (2007) ont montré que la prise en compte d'analogies formelles simples mettant en œuvre des opérations de préfixation et de suffixation permettaient d'enrichir avec succès les requêtes soumises à un système de recherche d'information, et ce dans six langues (allemand, anglais, espagnol, français, italien et portugais).

Plusieurs études portent plus spécifiquement sur la morphologie. Stroppa & Yvon (2005) ont notamment montré qu'il était possible de prédire le lemme d'un mot ainsi qu'un ensemble de traits (genre, nombre, etc.) pour trois langues (anglais, néerlandais et allemand). Les auteurs appliquent également l'apprentissage analogique à la prédiction de la structure morphologique d'un mot inconnu et rapportent des résultats comparables à une approche à l'état de l'art. N. Hathout à également présenté plusieurs travaux visant à extraire automatiquement d'un lexique et de ressources sémantiques des informations morphologiques régissant le lexique. Dans (Hathout, 2002), l'auteur croise des analogies de formes capturant des opérations de suffixation avec différentes relations (dont la synonymie) répertoriées dans la ressource Wordnet (Fellbaum, 1999) dans le but de réduire les analogies formelles à un sous-ensemble pertinent. On retrouve cette même idée dans (Hathout, 2008) où l'information sémantique est obtenue par marche aléatoire dans un graphe construit à partir de définitions des mots du lexique.

Malgré cet intérêt grandissant pour les analogies formelles, aucune étude n'a véritablement tenté d'analyser de manière quantitative les points de rencontre entre l'analogie formelle et la morphologie constructionnelle. Cette étude à pour objectif de combler ce manque.

Cet article se divise comme suit. Un rappel des définitions nécessaires à la reproductibilité de

1. Nous utilisons de manière interchangeable *forme* et *mot* dans cette étude.
2. La tâche étudiée consistait à traduire des phrases du corpus BTEC: des phrases de structures syntaxiques simples destinées à des touristes devant s'exprimer dans la langue locale.

cette étude est donné en section 2. Notre méthodologie expérimentale est ensuite décrite en section 3. Les points de rencontre entre l'analogie formelle et la morphologie sont analysés en section 4. La section 5 permet enfin de souligner les points saillants dégagés par cette étude.

2 Rappels sur l'anologie formelle

Plusieurs définitions de l'analogie formelle ont été proposées dans la littérature. Hathout (2002) s'intéresse par exemple aux analogies mettant en œuvre des paires de formes qui partagent le même préfixe (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 [*republishing*:*unpublished*::*rediscovering*:*undiscovered*]. 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 & Yvon (1999).

Des définitions plus générales sont également disponibles. Lepage (1998) propose un algorithme capable de rendre compte d'opérations morphologiques plus complexes comme la double infixation dans l'analogie [*arsala*:*mursilun*::*aslama*:*muslimun*]. Dans la présente étude, nous nous appuyons sur une définition proposée par Yvon *et al.* (2004) et reprise dans (Stroppa & Yvon, 2005). Cette définition qui s'appuie sur la notion de *factorisation* est une généralisation de la définition proposée par Y. Lepage.

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.

Par exemple, $(cor, dial, \epsilon, ly)$ est une 4-factorisation du mot anglais *cordially*. 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 et f_t) de *x*, *y*, *z* et *t* respectivement vérifiant: $\forall i \in [1, n]$: $(f_y^i, f_z^i) \in \{(f_x^i, f_t^i), (f_t^i, f_x^i)\}$. 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, [*cordially*:*cordial*::*appreciatively*:*appreciative*] est une analogie formelle car le quadruplet de 4-factorisations de la première colonne vérifie la définition 2:

$f_{cordially}$	$\equiv$	*cordia*	*l*	*l*	*y*	$f_{cordially}$	$\equiv$	*cordial*	*ly*
$f_{cordial}$	$\equiv$	*cordia*	ϵ	*l*	ϵ	$f_{cordial}$	$\equiv$	*cordial*	ϵ
$f_{appreciatively}$	$\equiv$	*appreciative*	*l*	ϵ	*y*	$f_{appreciatively}$	$\equiv$	*appreciative*	*ly*
$f_{appreciative}$	$\equiv$	*appreciative*	ϵ	ϵ	ϵ	$f_{appreciative}$	$\equiv$	*appreciative*	ϵ

Il existe également une 2-factorisation de ces formes (seconde colonne ci-dessus) vérifiant la définition, aussi l'analogie est de degré 2 et les factorisations minimales correspondantes mettent en lumière les *alternances* en jeu dans cette analogie: *cordial*/*appreciative* et ϵ/*ly*. 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 morphèmes.

3 Protocole expérimental

3.1 Démarche

Notre démarche consiste à extraire les relations analogiques liant les entrées d'un lexique afin de caractériser les facteurs impliqués dans les factorisations minimales calculées pour chaque analogie identifiée. Comme nous le verrons dans la section suivante, de très nombreuses analogies peuvent être identifiées en corpus, aussi fonctionnons-nous comme suit. Nous découpons notre lexique en deux parties disjointes, l'une appelée *mémoire* notée $\mathcal{L}_m$ et l'autre appelée *lexique de test* notée $\mathcal{L}_t$. Nous nous intéressons alors aux seules analogies qui mettent en œuvre une forme de $\mathcal{L}_t$ et trois formes de $\mathcal{L}_m$. Plus précisément, nous calculons:

$$\mathcal{A}(\mathcal{L}_t) = \{\ [x : y :: z : t] \mid t \in \mathcal{L}_t, \langle x, y, z\rangle \in \mathcal{L}_m^3\ \}$$

Cette stratégie en apparence simple cache des problèmes techniques non triviaux. Identifier les triplets de formes $\langle x, y, z\rangle$ dans $\mathcal{L}_m^3$ qui définissent avec une forme t de $\mathcal{L}_t$ une relation analogique est une opération à priori cubique en le nombre d'entrées dans $\mathcal{L}_m$, ce qui n'est pas faisable pour les lexiques considérés ici qui abritent plusieurs dizaines de milliers de formes (voir la section 3.2). Plusieurs solutions ont été proposées dans la littérature pour contourner partiellement ce problème, la plupart s'appuyant sur un échantillonnage des formes de la mémoire. Nous utilisons ici la méthode exacte proposée par Langlais & Yvon (2008) qui permet d'identifier efficacement **toutes** les analogies formelles dans un corpus. Cette stratégie tire profit d'une propriété sur le compte des symboles (caractères) des formes impliquées dans une analogie (Lepage, 1998):

$$[x : y :: z : t] \Rightarrow |x|_c + |t|_c = |y|_c + |z|_c \quad \forall c \in \Sigma$$

où $|x|_c$ dénote le nombre de symboles c dans la forme x. Soit $\mathcal{C}_\mathcal{L}\langle x, t\rangle$ l'ensemble des paires $\langle z, t\rangle$ d'entrées dans le lexique $\mathcal{L}$ qui vérifient avec la paire $\langle x, t\rangle$ cette propriété sur les comptes:

$$\mathcal{C}_\mathcal{L}\langle x, t\rangle = \{\langle y, z\rangle \in \mathcal{L}^2 \mid \langle y, z\rangle \neq \langle x, t\rangle \text{ et } |x|_c + |t|_c = |y|_c + |z|_c \quad \forall c \in \Sigma\}$$

La stratégie consiste à considérer tour à tour chaque entrée x de $\mathcal{L}_m$; ce qui introduit avec l'entrée t de $\mathcal{L}_t$, une contrainte sur le compte de symboles que les formes y et z doivent vérifier pour que le quadruplet définisse une analogie. Le pré-calcul d'une structure dédiée, tel que décrit dans (Langlais & Yvon, 2008) permet d'identifier efficacement les éléments de $\mathcal{C}_\mathcal{L}\langle x, t\rangle$. La propriété sur les comptes étant nécessaire mais non suffisante, une vérification des quadruplets qui forment véritablement une analogie est ensuite nécessaire. Formellement, nous construisons l'ensemble:

$$\mathcal{A}(\mathcal{L}_t) = \{\ [x : y :: z : t] \mid t \in \mathcal{L}_t, x \in \mathcal{L}_m, \langle y, z\rangle \in \mathcal{C}_{\mathcal{L}_m}\langle x, t\rangle\ \}$$

Stroppa (2005) décrit un algorithme qui vérifie que quatre formes sont en relation proportionnelle selon la définition 2. Nous avons modifié cet algorithme de manière à ce qu'en cas de succès, la factorisation minimale de chacune des formes soit retournée.

3.2 Corpus

Nous avons considéré dans cette étude trois lexiques extraits de la base de données lexicale CELEX (Baayen *et al.*, 1995). Nous nous sommes plus précisément concentrés sur les fichiers

de lemmes disponibles pour l'anglais, l'allemand et le néerlandais. La 23è ligne dans le fichier `DML` de la base est illustrée ci-après:

```
23\aalbessestruik\2\C\1\Y\Y\Y\aalbes+e+struik\NxN\N\N
\(((aal)[N],(bes)[N])[N],(e)[N|N.N],(struik)[N])[N]\N\N\N
```

Nous avons extrait pour chacune de ces lignes le lemme (*aalbessestruik* dans l'exemple), le statut de l'analyse morphologique (`C` pour composition) ainsi que les segmentations morphologiques simples et structurées disponibles [3]. Dans le cas de la seconde décomposition, la structure morphologique a été éliminée en réalisant une lecture en pré-ordre des morphèmes de l'arbre [4]. La structure `(((aal)[N],(bes)[N])[N],(e)[N|N.N],(struik)[N])[N]` est par exemple transformée en la séquence `aal+bes+e+struik`. Le résultat de ce traitement pour la ligne précédente donne [5]:

```
aalbessestruik C aalbes+e+struik aal+bes+e+struik
```

Afin de simplifier notre étude, nous avons éliminé du vocabulaire toute forme contenant un espace ou un tiret [6]. Les lemmes pour lesquels aucune décomposition morphologique structurée n'est disponible dans CELEX ont également été éliminés. Des trois tables ainsi construites, nous avons extrait des *lexiques de test* et des *lexiques mémoire*. Pour ce faire, nous avons sélectionné aléatoirement [7] 5 000 lignes de chaque table ; les différents lemmes associés constituant nos lexiques de test ; les lemmes des tables privées de ces lignes constituant nos lexiques mémoire. Le décompte du nombre de formes différentes dans chaque lexique est indiqué dans le tableau 1. Puisqu'un lemme peut intervenir plusieurs fois dans une table de CELEX (*above* apparaît par exemple quatre fois dans la table des lemmes anglais de CELEX), les lexiques de test contiennent moins de 5 000 formes différentes. On observe le nombre élevé de lemmes dans le lexique néerlandais, ce qui est dû entre autre à la nature fortement compositionnelle des mots dans cette langue. Par exemple, CELEX recense cinq lemmes construits à partir du lemme *psycholoog* (psychologue): *arbeidspsycholoog* (p. du travail), *bedrijfspsycholoog* (p. d'entreprise), *gedragspsycholoog* (p. du comportement), *ontwikkelingspsycholoog* (p. du développement), *persoonlijkheidspsycholoog* (p. de la personalité).

	$\lvert\mathcal{L}_m\rvert$	$\lvert\mathcal{L}_t\rvert$	$nb.$	$moy.$	$max.$
EN	32 142	4 892	15 781	2.4	7
DE	45 866	4 991	13 030	2.0	9
NL	114 385	4 991	29 153	2.6	9

TABLE 1 – Caractéristiques principales des tables extraites de CELEX. $\lvert\mathcal{L}_m\rvert$ et $\lvert\mathcal{L}_t\rvert$ dénotent le nombre de formes (lemmes) différentes dans les lexiques mémoire et test respectivement ; $nb.$ indique le nombre de morphèmes différents répertoriés dans le champ *décomposition structurée* des tables ; $moy.$ et $max.$ indiquent le nombre moyen et maximum de morphèmes par forme.

3. Pour les lemmes possédant plusieurs analyses morphologiques, nous avons considéré seulement la première.
4. Nous avons utilisé le script `stripcls.awk` disponible dans la distribution pour cela.
5. La décomposition simple (3è champ) n'est pas utilisée dans cette étude.
6. L'analogie formelle n'a aucune difficulté à traiter ces caractères.
7. À l'aide de la commande `sort -random-sort`.

4 Analyse

4.1 Quantités d'analogies

Les caractéristiques générales des analogies identifiées dans chaque langue sont fournies dans le tableau 2. Entre six millions (allemand) et 17 millions (néerlandais) d'analogies sont identifiées par lexique, avec plus de 1300 analogies en moyenne par forme de $\mathcal{L}_t$ (plus du double pour le néerlandais). Très peu de formes ne trouvent pas d'analogies (de l'ordre de 1% pour l'allemand et moins de 0.5% pour les deux autres langues). Ce sont souvent des mots d'emprunt ou des noms propres, comme *weltanschauung* (un mot d'origine allemande) et *chihuahua* (nom d'une ville mexicaine) qui apparaissent dans la liste des lemmes de la langue anglaise, ou *bodybuilding* qui apparaît dans le lexique allemand. Le nombre plus élevé d'analogies identifiées dans le cas du néerlandais s'explique par le plus grand nombre d'entrées dans le lexique mémoire.

	$\lvert\mathcal{A}(\mathcal{L}_t)\rvert$	moy	max	ϕ
EN	9 085 031	1857	4 692	18
DE	6 529 501	1308	40 264	57
NL	17 572 012	3520	129 698	10

TABLE 2 – Caractéristiques des analogies identifiées en corpus. $\lvert\mathcal{A}(\mathcal{L}_t)\rvert$ est le nombre total d'analogies ; moy et max désignent le nombre moyen et maximum d'analogies par forme dans $\mathcal{L}_t$; ϕ indique le nombre de formes de $\mathcal{L}_t$ pour lesquelles aucune analogie n'est identifiée.

4.2 Correspondance entre morphèmes et facteurs

Nous souhaitons savoir si les facteurs pris en compte dans les analogies correspondent aux morphèmes listés dans CELEX. Dans l'affirmative, nous pourrions par exemple utiliser le mécanisme analogique comme un générateur de morphèmes à la manière d'approches non supervisées développées par exemple dans le cadre du *Morpho Challenge* (Kurimo & Varjokallio, 2008).

À cet effet, nous calculons pour chaque analogie identifiée en corpus la factorisation minimale des quatre formes impliquées et maintenons le compte des facteurs impliqués dans ces factorisations. Nous obtenons donc une liste des facteurs, triée par ordre décroissant de fréquence. Les cinq facteurs les plus fréquemment impliqués dans les analogies de degré 2 [8] pour l'anglais et le néerlandais sont listés dans le tableau 3 ainsi que les cinq morphèmes les plus fréquents selon les décompositions structurées de $\mathcal{L}_t$ proposées par CELEX (colonne 4 dans nos tables lexicales). On remarque une intersection commune à ces deux listes.

Le taux de correspondance entre les deux listes peut être mesuré à l'aide des taux de précision et de rappel au rang k où la tâche associée consiste à identifier par analogie, les seuls morphèmes de CELEX. La précision au rang k est alors la proportion des k-premiers facteurs qui sont présents dans CELEX ; le rappel au rang k indique la proportion des morphèmes de CELEX listés dans les k-premiers facteurs. La figure 1 montre pour les trois langues les courbes de précision et de rappel obtenues en faisant varier k. Les mêmes tendances se dégagent pour

8. Nous évaluons toutes les analogies dans la suite.

m	CELEX (EN)	f	factorisation	m	CELEX (NL)	f	factorisation
ly	*word+less+**ly***	*ly*	*abject+**ly***	*s*	*aal+**s**+kruik*	*en*	*ahorn+**en***
ness	*yellow+**ness***	*ness*	*abrupt+**ness***	*ing*	*aan+berm+**ing***	*ing*	*jeuk+**ing***
er	*fish+**er**+man*	*y*	*bump+**y***	*er*	*biets+**er***	*heid*	*alert+**heid***
y	*sleep+**y**+head*	*er*	*box+**er***	*en*	*abeel+**en**+laan*	*er*	*arbeid+**er***
un	***un**+zip*	*s*	*eat+**s***	*heid*	*gelijk+**heid***	*af*	***af**+druk*

TABLE 3 – Les morphèmes (m) les plus fréquents selon les décompositions structurées de CELEX, et facteurs (f) les plus fréquents dans les analogies de degré 2 identifiées en corpus.

toutes les langues. Les facteurs les plus fréquents correspondent davantage à des morphèmes répertoriés par CELEX: de l'ordre de 50% (EN) à 70% (DE) de précision est mesurée sur les 100 facteurs les plus fréquemment rencontrés dans les analogies de degré 3 ou moins (figure de gauche). Cette précision décroît au fur et à mesure que k augmente. Les courbes de rappel suivent quant à elles une tendance inverse. La figure de droite est obtenue en considérant toutes les analogies (quelque soit leur degré). On observe une légère baisse de la précision dans ce cas.

Si globalement, les facteurs les plus fréquents impliqués dans les analogies identifiées en corpus correspondent souvent à des morphèmes, ces courbes montrent qu'il n'existe pas de correspondance biunivoque entre les deux types d'unités. Nous reviendrons sur ce résultat dans la prochaine section.

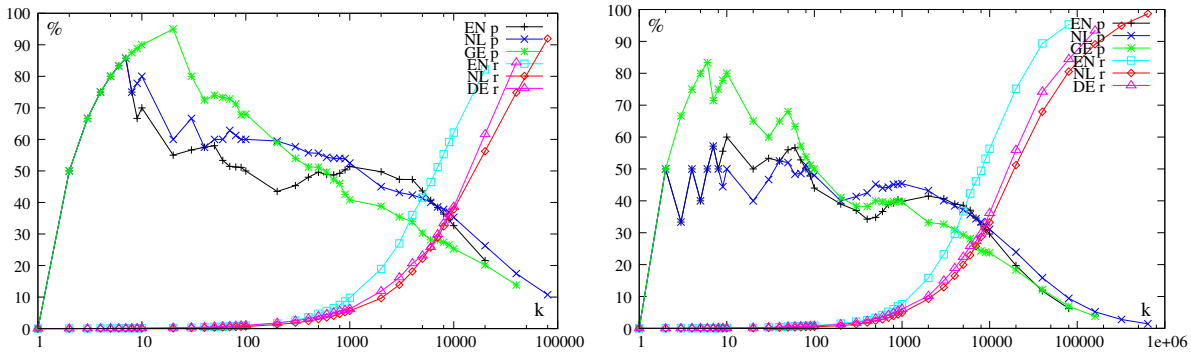

FIGURE 1 – Précision (p) et rappel (r) au rang k des facteurs impliqués dans les analogies identifiées en corpus. Dans la figure de gauche, seules les analogies de degré 3 ou moins sont considérées ; la figure de droite concerne l'ensemble des analogies.

4.3 Correspondance des segmentations induites par analogie

Nous nous intéressons maintenant à comparer la segmentation produite par CELEX à celle induite par analogie. Pour se faire, nous retenons pour chaque mot impliqué dans au moins une analogie l'ensemble des factorisations minimales de ce mot. Par exemple, la forme néerlandaise *prozabewerking* (prose + adaptation) est impliquée dans 118 analogies qui induisent un total de 26 segmentations différentes dont les cinq plus fréquentes sont indiquées dans le tableau 4. La décomposition dans CELEX de cette forme est induite par 16 analogies, dont [*prozabewerking*:*prozawerk*::*betekening*:*teken*] et [*prozawerk*:*invloed*::*prozabewerking*:*beinvloeding*]:

$f_{prozabewerking}$	$\equiv$	*proza*	*be*	*werk*	*ing*	$f_{prozawerk}$	$\equiv$	*proza*	*be*	*werk*	*ing*
$f_{prozawerk}$	$\equiv$	*proza*	ϵ	*werk*	ϵ	$f_{invloed}$	$\equiv$	*proza*	ϵ	*werk*	ϵ
$f_{betekening}$	$\equiv$	ϵ	*be*	*teken*	*ing*	$f_{prozabewerking}$	$\equiv$	ϵ	*be*	*teken*	*ing*
f_{teken}	$\equiv$	ϵ	ϵ	*teken*	ϵ	$f_{beinvloeding}$	$\equiv$	ϵ	ϵ	*teken*	ϵ

EN (16)		DE (26)		NL (26)	
$freq$	*factorisation*	$freq$	*factorisation*	$freq$	*factorisation*
18	**in+dent+ation**	92	unerbittlich+keit	18	p+r+ozabewerking
11	indent+ation	26	une+r+bittlichkeit	16	**proza+be+werk+ing**
7	ind+entation	14	**un+er+bitt+lich+keit**	14	prozab+e+werking
7	inden+tation	12	un+e+rbittlichkeit	12	pr+o+zabewerking
4	in+den+tation	12	unerbitt+lichkeit	10	proz+a+bewerking

TABLE 4 – Les cinq factorisations les plus fréquentes induites par analogie et leur fréquence associée pour un lemme de chaque langue étudiée. Le nombre total de factorisations trouvées est indiqué entre parenthèses. Les décompositions en gras sont celles fournies par CELEX.

Le tableau 5 évalue la segmentation produite par analogie. Le rang moyen r de la factorisation correspondant à la segmentation structurée de CELEX est inférieur à 3 pour l'anglais et l'allemand et est de 5 pour le néerlandais. On observe que dans 72% (allemand) à 85% (anglais) des mots testés [9], il existe au moins une analogie permettant d'induire une factorisation identique à la décomposition structurée de CELEX. En moyenne, l'analogie permet de produire de 9 (anglais) à 30 (néerlandais) factorisations par forme de $\mathcal{L}_t$ (nbf).

EN				DE				NL			
nb	%	nbf	r	nb	%	nbf	r	nb	%	nbf	r
11 257	85.1	9.3	2.2	22 471	72.0	22.7	2.3	60 586	80.9	29.9	4.9

TABLE 5 – Nombre moyen de segmentations induites par analogie (nbf) et rang (r) moyen de la segmentation CELEX associée. Lire le texte pour les détails.

Ces résultats indiquent qu'il est possible de segmenter morphologiquement un mot à l'aide des analogies l'impliquant ; ce qui peut sembler contradictoire avec les expériences précédentes où nous montrions qu'il n'existe pas une correspondance biunivoque entre les morphèmes et les facteurs. La différence entre les deux expériences réside essentiellement dans la prise en compte de la fréquence des factorisations dans la tâche de segmentation. Pour confirmer l'importance de cette information, nous mesurons sur la tâche consistant à identifier les morphèmes de CELEX, la précision et le rappel au rang k des facteurs identifiés par la factorisation la plus fréquente associée à chaque forme intervenant dans au moins une analogie. Les résultats sont consignés dans le tableau 6. Nous observons (entre parenthèse) des gains absolus substantiels de précision pour les trois langues par rapport aux expériences réalisées en section 4.2.

9. Nous ne considérons ici que les lemmes de CELEX dont la décomposition structurelle (4è champ dans nos lexiques) correspond au lemme ; $nb.$ indique leur nombre dans le tableau 5. La forme *abdication* n'est par exemple pas retenue car sa décomposition dans CELEX est *((abdicate)[V],(ion)[N|V.])[N]*, soit *abdicate+ion* dans notre format.

	EN				DE				NL			
k	précision		rappel		précision		rappel		précision		rappel	
100	84.0	(+31)	1.6	(+ϵ)	79.0	(+11)	1.2	(+ϵ)	72.0	(+12)	0.8	(+ϵ)
1 000	69.4	(+17)	13.2	(+3)	62.4	(+22)	9.5	(+3)	75.7	(+23)	8.1	(+3)
5 000	49.7	(+5)	47.3	(+5)	35.9	(+6)	27.4	(+4)	52.2	(+11)	28.0	(+6)

TABLE 6 – Identification des morphèmes de CELEX en considérant la factorisation la plus fréquente des formes de CELEX. Lire le texte pour les détails.

5 Discussion

Nous décrivons dans cet article une méthode qui permet d'identifier efficacement en corpus toutes les analogies formelles d'un lexique. Nous mettons à profit cette approche de manière à étudier à grande échelle les points de rendez-vous entre l'analogie formelle et l'organisation morphologique de lexiques anglais, allemand et néerlandais. Nous montrons que si les facteurs impliqués dans les analogies ne sont pas tous des morphèmes, le seul principe d'analogie formelle permet néanmoins de segmenter avec une bonne précision les mots d'un lexique en morphèmes.

Cette étude servira de base à des travaux que nous comptons mener sur l'acquisition non supervisée d'information morphologique à partir d'un lexique, dans la lignée des travaux menés dans le cadre de *Morpho Challenge* (Kurimo & Varjokallio, 2008). Plus précisément, nous nous concentrerons sur les deux tâches liées que nous avons abordées ici avec un esprit exploratoire, à savoir, l'identification non supervisée de morphèmes d'une langue et la segmentation d'une chaîne en morphèmes.

Nous pensons qu'il est possible de faire émerger davantage de "bons" morphèmes, en considérant non plus des lexiques de lemmes mais des lexiques de mots en nombre beaucoup plus importants. Nous pensons également mettre à profit des indices plus classiques dans ce type d'approches qui sont en grande partie inspirés des travaux de Harris (1955).

Remerciements

L'étude des points de rencontre entre l'analogie formelle et la morphologie m'a été initialement suggérée par les arbitres anonymes de l'article (Langlais & Patry, 2008) ; je les en remercie. Cette étude a bénéficié de discussions constructives que j'ai eu avec Pierre Zweigenbaum et François Yvon lors de mon dernier séjour au LIMSI. Je remercie également les trois relecteurs anonymes de cet article pour la justesse de leurs remarques. Cette étude a été partiellement financée par le Conseil de Recherches en Sciences Naturelles et en Génie du Canada.

Références

BAAYEN R. H., PIEPENBROCK R. & GULIKERS L. (1995). The CELEX lexical database (release 2). CD-ROM, Linguistic Data Consortium, Univ. of Pennsylvania, USA.

DENOUAL E. (2007). Analogical translation of unknown words in a statistical machine translation framework. In *Machine Translation Summit, XI*, Copenhagen.

C. FELLBAUM, Ed. (1999). *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge.

HARRIS Z. S. (1955). From phoneme to morpheme. *Language*, **31**(2), 190–222.

HATHOUT N. (2002). From wordnet to celex: acquiring morphological links from dictionaries of synonyms. In *Third International Conference on Language Resources and Evaluation*, p. 1478–1484, Las Palmas de Gran Canaria.

HATHOUT N. (2008). Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing*, p. 1–8, Manchester, United Kingdom.

KURIMO M. & VARJOKALLIO M. (2008). Unsupervised morpheme analysis evaluation by a comparison to a linguistic Gold Standard — Morpho Challenge 2008. Working Notes for the CLEF 2008 Workshop.

LANGLAIS P. & PATRY A. (2008). Enrichissement d'un lexique bilingue par apprentissage analogique. *Traitement Automatique des Langues (TAL)*, **49 (varia)**, 13–40.

LANGLAIS P. & YVON F. (2008). Scaling up analogical learning. In *22nd International Conference on Computational Linguistics (COLING 2008)*, p. 51–54, Manchester, United Kingdom.

LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2009). Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Á paraître dans 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, p. 9 pages, Athens, Greece.

LEPAGE Y. (1998). Solving analogies on words: an algorithm. In *COLING-ACL*, p. 728–734, Montreal, Canada.

LEPAGE Y. & DENOUAL E. (2005). Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, **29**, 251–282.

MOREAU F., CLAVEAU V. & SÉBILLOT P. (2007). Automatic morphological query expansion using analogy-based machine learning. In *29th European Conference on Information Retrieval (ECIR 2007)*, Roma, Italy.

PIRRELLI V. & YVON F. (1999). The hidden dimension: a paradigmatic view of data-driven NLP. *Journal of Experimental & Theroretical Artifical Intelligence*, **11**, 391–408.

STROPPA N. (2005). *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*. PhD thesis, ENST, ParisTech, Télécom, Paris, France.

STROPPA N. & YVON F. (2005). An analogical learner for morphological analysis. In *9th Conf. on Computational Natural Language Learning (CoNLL)*, p. 120–127, Ann Arbor, MI.

YVON F., STROPPA N., DELHAY A. & MICLET L. (2004). *Solving analogical equations on words*. Rapport interne D005, École Nationale Supérieure des Télécommunications, Paris, France.