Université de Montréal

Moranapho : Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques.

par
Jean-François Lavallée

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Août, 2010

Université de Montréal
Faculté des arts et des sciences

Ce mémoire intitulé:

Moranapho : Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques.

présenté par:

Jean-François Lavallée

a été évalué par un jury composé des personnes suivantes:

Guy Lapalme,	président-rapporteur
Philippe Langlais,	directeur de recherche
Nadia El-Mabrouk,	membre du jury

Mémoire accepté le: .....

RÉSUMÉ

Récemment, nous avons pu observer un intérêt grandissant pour l'application de *l'analogie formelle* à l'analyse morphologique. L'intérêt premier de ce concept repose sur ses parallèles avec le processus mental impliqué dans la création de nouveaux termes basée sur les relations morphologiques préexistantes de la langue. Toutefois, l'utilisation de ce concept reste tout de même marginale due notamment à son coût de calcul élevé. Dans ce document, nous présenterons le système à base de graphe *Moranapho* fondé sur l'analogie formelle. Nous démontrerons par notre participation au Morpho Challenge 2009 (Kurimo et al., 2009) et nos expériences subséquentes, que la qualité des analyses obtenues par ce système rivalise avec l'état de l'art. Nous analyserons aussi l'influence de certaines de ses composantes sur la qualité des analyses morphologiques produites. Nous appuierons les conclusions tirées de nos analyses sur des théories bien établies dans le domaine de la linguistique. Ceci nous permet donc de fournir certaines prédictions sur les succès et les échecs de notre système, lorsqu'appliqué à d'autres langues que celles testées au cours de nos expériences.

Mots clés: intelligence artificielle, apprentissage machine, analyse morphologique non supervisée, analogie formelle, approche à base de graphe.

ABSTRACT

Recently, we have witnessed a growing interest in applying the concept of *formal analogy* to unsupervised morphology acquisition. The attractiveness of this concept lies in its parallels with the mental process involved in the creation of new words based on morphological relations existing in the language. However, the use of *formal analogy* remain marginal partly due to their high computational cost. In this document, we present *Moranapho*, a graph-based system founded on the concept of formal analogy. Our participation in the 2009 Morpho Challenge (Kurimo et al., 2009) and our subsequent experiments demonstrate that the performance of *Moranapho* are favorably comparable to the state-of-the-art. We studied the influence of some of its components on the quality of the morphological analysis produced as well. Finally, we will discuss our findings based on well-established theories in the field of linguistics. This allows us to provide some predictions on the successes and failures of our system when applied to languages other than those tested in our experiments.

Keywords: artificial intelligence, machine learning, unsupervised learning of morphology, formal analogy, graph-based approach.

TABLE DES MATIÈRES

RÉSUMÉ . . . . . **iii**

ABSTRACT . . . . . **iv**

TABLE DES MATIÈRES . . . . . **v**

LISTE DES TABLEAUX . . . . . **viii**

LISTE DES FIGURES . . . . . **ix**

LISTE DES SIGLES . . . . . **xi**

REMERCIEMENTS . . . . . **xii**

CHAPITRE 1 :INTRODUCTION **1**

1.1 Tâche 1

1.1.1 Description 2

1.1.2 Motivations 4

1.2 Contributions 6

CHAPITRE 2 :CONCEPTS DE BASE EN MORPHOLOGIE **8**

2.1 Du mot au morphème 8

2.2 Les types de morphèmes 9

2.2.1 Morphème dérivationnel et flexionnel 9

2.2.2 Racine, lexème, base et *stem* 10

2.3 Phénomènes morphologiques 12

2.3.1 Affixation 12

2.3.2 Mutation du *stem* 13

2.3.3 Composition 14

2.3.4 Supplétion 14

2.3.5 Redoublement 14

2.3.6 Patron-racine 15

2.4 Typologie morphologique des langues 16

2.4.1 Analytique 17
2.4.2 Agglutinante 17
2.4.3 Fusionnelle 18
2.4.4 Polysynthétique 18
2.4.5 Templatique 19
2.5 Paradigme 19
2.6 Productivité 20
2.7 Analogie 21
2.8 Morphologie basée sur le mot 22

CHAPITRE 3 :ÉTAT DE L'ART 24
3.1 Approches courantes à l'acquisition morphologique 24
3.1.1 Calcul de l'incertitude 24
3.1.2 Minimisation de la taille de description 25
3.1.3 Utilisation du contexte 26
3.1.4 Paradigme 27
3.2 Approches marginales à l'acquisition morphologique 28
3.2.1 Approche à base de règle 28
3.2.2 Analogie 30
3.3 Le système *Morfessor* 31

CHAPITRE 4 :ANALOGIE FORMELLE 33
4.1 Identification des analogies 34
4.2 Système purement analogique 35

CHAPITRE 5 :LE SYSTÈME *MORANAPHO* 36
5.1 Règle morphologique 36
5.1.1 Extraction de règles par distance d'édition 38
5.1.2 Extraction de règles par analogie 38
5.1.3 Sélection des règles 40
5.2 Regroupement hiérarchique des mots du lexique 41
5.3 Décomposition morphologique des mots du lexique 44

CHAPITRE 6 :ÉVALUATION **47**
6.1 Données 47
6.2 Métrique d'évaluation 48

CHAPITRE 7 :RÉSULTATS ET ANALYSES **50**
7.1 Comparaison à l'état de l'art 50
7.1.1 Morpho Challenge 2009 50
7.1.2 Évaluation en condition optimale 53
7.2 Étude de l'impact des composantes du système sur les performances . . . 56
7.2.1 Impact des valeurs d'hyperparamètres 57
7.2.2 Impact du degré de l'analogie 58
7.2.3 Impact du processus d'acquisition de règles 59

CHAPITRE 8 :CONCLUSION **61**

BIBLIOGRAPHIE **63**

LISTE DES TABLEAUX

5.I Valeurs des métriques de qualité pour certaines règles extraites d'un lexique anglais. 42

6.I Principales caractéristiques de la référence `RefCelex`. 48
6.II Résultats obtenus en n'identifiant qu'un seul suffixe en anglais. 49

7.I Résultats officiels du Morpho Challenge 2009. 52
7.II Évaluation des analyses du Morpho Challenge sur la métrique EMMA. . . 53
7.III Évaluation des systèmes *Moranapho* et *Morfessor* sur les lexiques épurées de langues germaniques. 54
7.IV Statistiques sur la distribution des morphèmes des analyses de *Moranapho* et *Morfessor*. 56
7.V Impact des hyperparamètres ρ et τ sur les performances de *Moranapho*. . 58
7.VI Performance de *Moranapho* selon la méthode d'extraction de règles utilisée. 60

LISTE DES FIGURES

1.1 Extrait de la référence anglaise. 3
1.2 Alignement français-anglais des termes des segments de phrase *guerre nucléaire* et *centrale thermique*. 5
1.3 Alignement français-allemand des termes des segments de phrase *guerre nucléaire* et *centrale thermique* avant et après la décomposition morphologique 6

2.1 Exemples de dérivation. 10
2.2 Exemples de flexion. 11
2.3 Exemples d'affixation. 13
2.4 Exemples de redoublement. 15
2.5 Exemples de morphologie patron-racine. 16
2.6 Le principe d'agglutination démontré par quelques termes turcs. 18

4.1 Exemples d'analogies pour les principaux phénomènes morphologiques. . . 33
4.2 Factorisations de l'analogie [*cordially* : *cordial* = *lovely* : *love*]. . . . 34
4.3 Factorisations induites par analogie pour certains mots de différentes langues. 35

5.1 Alignement obtenu par distance d'édition entre les mots *unattractive* et *attraction*. 38
5.2 Deux factorisations de degré 2 de l'analogie anglaise [*addict* : *addictive* = *distinct* : *distinctive*]. 40
5.3 Exemple d'arbre hiérarchique avant le partitionnement. 42
5.4 Graphe des mots atteignables depuis le mot anglais *disabled* 43

7.1 Extraits des analyses produites par *Moranapho* et *Morfessor CatMAP* sur les lexiques de langues germaniques. 55
7.2 Exemples de partitionnements obtenus par notre système sur le lexique anglais de *DonPropre*. 57
7.3 Performance de *Moranapho* selon la limite imposée sur le degré des analogies considérées. 59

LISTE DES ALGORITHMES

5.1 Algorithme d'extraction de règles par analogie. 39
5.2 Algorithme de construction de l'arbre hiérarchique. 43
5.3 Algorithme de décomposition en morphème d'un mot m d'un ADM. . . . 46

LISTE DES SIGLES

Al.	*Alia*
All.	Allemand
ADM	Arbre de Dérivation de Mot
Ana.	Analogie
AMNS	Analyse Morphologique Non Supervisée
Ang.	Anglais
Arb.	Arabe
D.E.	Distance d'Édition
Dist.	Distinct
F1	F-mesure
Fin.	Finnois
Fr.	Fréquence
HMM	*Hidden Markov Model*
MBM	Morphologie Basée sur le Mot
MDL	*Minimum Description Length*
Mora.	*Moranapho*
Morf.	*Morfessor*
Morph.	Morphème
Nb.	Nombre
Ner.	Néerlandais
P. ex.	Par exemple
Pr.	Précision
Rp.	Rappel
Tur.	Turc
WFS	*Word Formation Strategy*

REMERCIEMENTS

Je voudrais remercier toute l'équipe du RALI pour son soutien exceptionnel ainsi que pour l'environnement de travail favorable qui ont rendu ces deux dernières années si agréables. Je tiens à remercier plus particulièrement mon directeur de recherche, M.Philippe Langlais, pour l'encadrement de premier ordre que j'ai reçu. J'aimerais remercier aussi M.Florian Boudin, M.Alexandre Patry et M.Pierre Paul Monty qui ont été très généreux de leur temps. Je voudrais aussi mentionner Mlle Geneviève Lavallée dont les commentaires et les conseils sur l'utilisation correcte des locutions conjonctives et prépositionnelles m'ont grandement aidé lors de la rédaction de ce document. Finalement, je remercie mes parents qui m'ont appuyé tout au long d'un parcours scolaire sinueux et qui ont rendu mon retour aux études possible.

CHAPITRE 1

INTRODUCTION

Traditionnellement, le mot est l'unité de base des modèles de langue. Bien qu'adapté aux langues occidentales les plus répandues, telles que l'anglais et le français, on observe une dégradation des performances de ces modèles sur les langues où il y a peu de données d'entraînement disponibles ou sur les langues dont la morphologie est plus complexe. Pour ces cas, la transition vers un modèle basé sur le *morphème*[1] semble tout appropriée, car elle permettrait de retirer le maximum d'information des données disponibles (section 1.1.2). Cependant, les lexiques morphologiques (*p. ex.* voir la figure 1.1) de qualité sont plutôt rares, ils ne sont habituellement disponibles que pour quelques langues très répandues. De plus, un effort considérable doit être fait pour tenir ces lexiques à jour dus à la création régulière de nouveaux termes. Ceci a donc motivé un bon nombre de chercheurs à s'intéresser à la possibilité de construire automatiquement le lexique morphologique d'une langue donnée. Cette tâche, détaillée à la section 1.1.1 et que nous appellerons analyse morphologique non supervisée (AMNS), est très complexe.Ceci est en partie dû au grand nombre de phénomènes morphologiques existants et de la grande variabilité de leur utilisation d'une langue à l'autre. Ces phénomènes morphologiques, ainsi que d'autres concepts linguistiques nécessaires à la compréhension de ce mémoire, sont expliqués au chapitre 2. Nous terminerons ce chapitre par une courte description de la contribution apportée par le travail présenté dans ce mémoire.

1.1 Tâche

Comme mentionné dans l'introduction et discuté à la section 1.1.2, il peut être avantageux d'effectuer un prétraitement qui normalise et affine l'unité de base des données linguistiques dans le but de maximiser l'information qui peut en être tirée. À cette fin, il est d'usage courant d'utiliser le *stemming* qui consiste à remplacer les mots contenus dans les ressources linguistiques par leurs *stem*s. Il est toutefois possible d'aller plus loin en remplaçant les mots non pas simplement par leur *stem*, mais par leur décomposition

[1] La plus petite unité significative du langage (voir section 2.1).

en morphèmes. Pour ce faire, il est nécessaire d'avoir accès à un lexique associant chacun des mots à sa composition morphologique. Malheureusement, de tels lexiques sont rarement disponibles et il est donc nécessaire d'utiliser un algorithme capable d'effectuer automatiquement l'analyse morphologique de la langue étudiée, tel que celui décrit dans cet ouvrage. Le concept d'analyse morphologique et le contenue du lexique que nous souhaiterions obtenir sont détaillés à la section 1.1.1.

1.1.1 Description

Bien que ne bénéficiant pas d'autant de visibilité que la recherche d'informations ou la traduction automatique, l'analyse morphologique est tout de même un domaine faisant l'objet d'un effort de recherche considérable. Comme pour tous les domaines rassemblant un nombre important de chercheurs, il y a des divergences sur la définition même du problème à résoudre. La différence de point de vue découle habituellement du type de morphologie jugée pertinente par le scientifique et est donc fortement dépendante de l'application postulée des analyses obtenues. Certains chercheurs ne s'intéressent qu'à la morphologie *flexionnelle*, réduisant ainsi l'analyse morphologique au *stemming*, alors que d'autres, plus ambitieux, tenteront d'émuler l'analyse telle que produite par un linguiste. Le système que nous décrivons dans cet ouvrage ne se limite à aucune application et à aucun type de langue en particulier. Par conséquent, la tâche à laquelle nous nous attaquons correspond plutôt au deuxième point de vue. C'est-à-dire que nous tentons de produire une analyse similaire à celle que produirait un linguiste en considérant tous les types de morphologie, et ce, de manière non supervisée. Notre système reçoit donc en entrée le lexique de la langue étudiée comprenant tous les mots à analyser (1^{re} colonne de la figure 1.1). Sans aucune information supplémentaire, le système tente de produire la décomposition en morphèmes (2^{e} colonne de la figure 1.1) de chacun des mots du lexique. Il est important de noter que la représentation textuelle des morphèmes dans le lexique est un choix arbitraire et que le lexique de la figure 1.1 serait tout aussi valide si on remplacerait par exemple l'étiquette de morphème `ist-s` par `M-1` ou toute autre chaîne de caractère n'étant pas utilisé pour identifier un autre morphème. En effet, ce qui importe est d'identifier la distribution du morphème dans la langue et non de trouver sa représentation graphémique la plus significative. Nous pouvons observer à la figure 1.1 que nous ne nous limitons pas aux cas où la frontière de morphème est bien définie (*p.*

ex. `antibiotic`). En effet, nous prenons aussi en compte les cas où une segmentation du mot ne suffit pas pour obtenir sa composition en morphème (*p. ex.* `bodies`). De plus, la tâche ne se limite pas à l'identification des *morphes*[2]. En effet, notre référence prend en compte qu'un même morphème peut être représenté par plus d'un morphe comme dans le cas du morphème `pluriel` des termes `botanists` et `fishes` qui est représenté respectivement par `s` et `es`. La relation inverse, c'est-à-dire lorsque 2 morphèmes peuvent être représentés par un même morphe, est aussi considérée. Par exemple, le pluriel et la 3^{e} personne du singulier sont tous les deux fréquemment représentés par un `s` (*p. ex.* `botanists` et `relives`). Toutefois, aucun système existant ne tente de couvrir l'entièreté des phénomènes couverts par notre lexique de référence, l'état de l'art actuel du domaine étant encore loin d'être capable de fournir une qualité d'analyse équivalente.

Mot	Morphèmes
antibiotic	anti-p biotic-s
body	body-N
bodies	body-N +PL
antibody	anti-p body-N
antibodies	anti-p body-N +PL
botanist	botany-N ist-s
botanists	botany-N ist-s +PL
botanists'	botany-N ist-s +PL +GEN
fishes	fish-N +PL
flies	fly-N +PL, fly-V +3SG
lock	lock-N
locksmith	lock-N smith-N
relives	re-p live-V +3SG

Figure 1.1 – Extrait de la référence anglaise. Les flexions présentes dans cet extrait sont le pluriel(+PL), le génitif(+GEN) et la 3^{e} personne du singulier (+3SG). Les symbole -N (Nom) et -V (Verbe) permettent de désambiguïser les racines partageant la même représentation graphémique (fly). Les morphèmes dérivationnels sont désambiguïsés par -p (préfixe) et -s (suffixe).

[2]La représentation du morphème dans le mot (voir section 2.1).

1.1.2 Motivations

Depuis quelques années, les systèmes basés sur l'apprentissage machine ont gagné en popularité dans le domaine du traitement des langues. Fréquemment, ces systèmes atteignent leurs objectifs en utilisant des statistiques sur la distribution des mots calculées dans un corpus d'apprentissage. Bien entendu, pour que les statistiques sur la distribution d'un mot soient significatives, ce mot doit avoir été observé plusieurs fois lors de l'apprentissage. Par conséquent, ces systèmes requièrent une grande quantité de données pour être performants. Malheureusement, il n'est pas toujours facile d'obtenir suffisament de données de bonne qualité. Il est donc important de pouvoir maximiser l'information extraite d'une quantité limitée de données.

La technique la plus commune et d'utiliser le *stemming* qui consiste à normaliser les données en remplaçant chacun des mots par son *stem*. Par exemple, ceci consisterait en français à remplacer tous les adjectifs par leurs formes au masculin singulier et à mettre tous les verbes à l'infinitif. Cependant, pour plusieurs langues le *stemming* ne semble pas être suffisant (Braschler et Ripplinger, 2004; Martin et al., 2003), car le *stem* est lui même très complexe. Ceci est particulièrement fâcheux, car ce sont ces mêmes langages qui souffrent le plus du manque de données d'apprentissage. En effet, le grand nombre de combinaisons de morphèmes permises par les langues morphologiquement complexes a pour conséquence que le nombre de mots distincts est aussi très élevé. Par conséquent, les corpus d'apprentissage doivent être très volumineux pour espérer couvrir convenablement la langue étudiée. De plus, même en respectant cette condition, il est fort probable que plusieurs des mots rencontrés lors de l'utilisation du système n'auront pas été observés en entraînement.

Pour ces cas, il est généralement admis que l'utilisation de la composition morphologique du mot permettrait de pallier ce problème. Ceci est appuyé par plusieurs études couvrant un large éventail de tâches classiques du traitement des langues tel que la recherche d'informations (Braschler et Ripplinger, 2004; Kurimo et al., 2009), la traduction (Toutanova et al., 2008; Cartoni, 2009) et la reconnaissance de la parole (Saraswathi et Geetha, 2007). Malheureusement, les lexiques morphologiques de qualités produits par des experts et couvrant de façon convenable la langue étudiée ne sont généralement disponibles que pour les langues où le potentiel de la décomposition morphologique est

minime. Ceci n'est pas surprenant, car plus la langue est morphologiquement complexe, plus il est impensable de produire manuellement une liste exhaustive de tous les mots de la langue accompagnés de leur composition morphologique. Par exemple, il est (presque) tout aussi futile de tenter de lister tous les mots de la langue inuktitut que de vouloir énumérer toutes les phrases (simples) possibles en français. En effet, il n'est pas rare en inuktitut qu'un seul mot contienne tous les éléments nécessaires (sujet-verbe-complément) à la formation d'une phrase grammaticalement correcte en français (section 2.4.4).

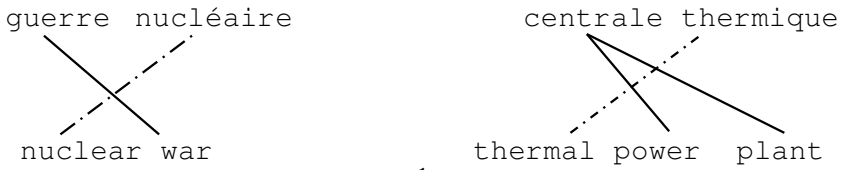

Figure 1.2 – Alignement français-anglais des termes des segments de phrase *guerre nucléaire* et *centrale thermique*.

L'intérêt de la morphologie en traitement des langues est en grande partie dû à sa capacité à gérer les mots qui n'ont pas été observés lors de l'entraînement. Pour mieux illustrer le potentiel de la décomposition morphologique, nous utiliserons un exemple tiré de la traduction automatique. Dans notre exemple, notre système construit son modèle de traduction sur 2 segments de phrases connus dans les 2 langues (*guerre nucléaire* et *centrale thermique*) dans le but d'en traduire un 3^{e} (*centrale nucléaire*) connu dans une seule langue. Lors d'une traduction du français vers l'anglais, il est possible d'obtenir la traduction du 3^{e} segment (*nuclear power plant*) à partir de nos 2 exemples de traduction si l'alignement correct (*guerre* - *war*, *nucléaire* - *nuclear*,*centrale* - *power plant*, *thermique* - *thermal*) entre les termes a été identifié (figure 1.2). Cependant, il en est tout autrement lors d'une traduction vers une langue morphologiquement plus complexe telle que l'allemand. En effet, un alignement par mot ne nous donne aucune information nous permettant de traduire le 3^{e} segment[3], car chacun des segments français est englobé par un seul terme allemand (haut de la figure 1.3). Par contre, si l'on observe attentivement les termes allemands, on remarque qu'il existe une correspondance en tout point semblable à celle présente en anglais entre les éléments sémantiques et graphémiques des segments. C'est-à-dire que dans les 2 langues, le segment est com-

[3]Le segment *centrale nucléaire*.

posé du même nombre d'unités et que chacune de ces unités a un équivalent partageant la même signification dans l'autre langue. La seule différence est que le découpage en unité sémantique en anglais est rendu explicite par des espaces alors qu'en allemand la frontière entre les éléments contenus dans le segment est implicite. Donc en remplaçant ces termes par leur décomposition en morphème (bas de la figure 1.3), on obtient un alignement aussi précis des unités significatives que lors de la traduction vers l'anglais (*guerre - krieg*, *nucléaire - atom*,*centrale - kraft werk*,*thermal - wärme*). Grâce à cet alignement, nous pouvons déterminer que la traduction du 3^{e} segment est *atomkraftwerk*.

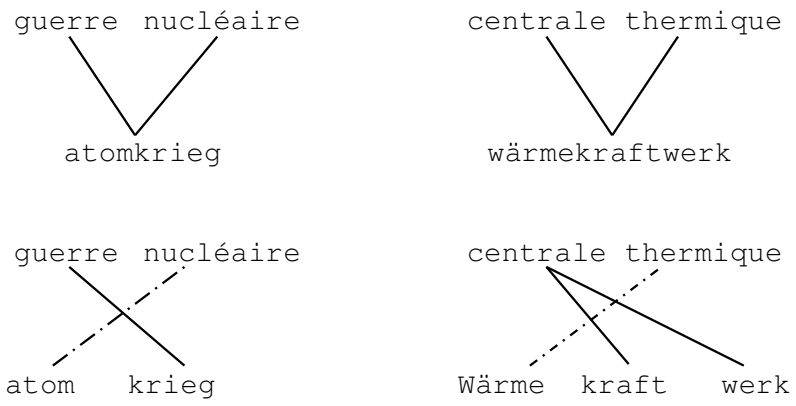

Figure 1.3 – Alignement français-allemand des termes des segments de phrase *guerre nucléaire* et *centrale thermique* avant (haut) et après (bas) la décomposition morphologique.

1.2 Contributions

Plusieurs études antérieures ont démontré qu'il existait un lien entre l'analogie formelle et la morphologie (section 3.2.2). Cependant, à cause des coûts de calcul prohibitifs de l'analogie, les systèmes analogiques antérieurs doivent imposer des restrictions sur la taille du lexique, la flexibilité de l'analogie ou la sélection des mots considérés. Le système *Moranapho* présenté dans ce mémoire contourne ces limitations en généralisant l'information capturée par analogie, permettant ainsi de réduire le nombre de calculs nécessaires. Par conséquent, le système *Moranapho* est à notre connaissance le seul système analogique capable d'analyser des lexiques de taille équivalente à ceux gérés par les autres approches tout en obtenant des résultats avantageusement comparables. Ceci est démontré par les

résultats que nous avons obtenus au Morpho Challenge 2009 (Kurimo et al., 2009).

L'information analogique est généralisée en extrayant de chaque analogie des règles de réécriture permettant d'identifier les autres paires de mots suivant le même procédé analogique. Ces règles de réécriture sont alors utilisées pour construire un graphe joignant entre eux les mots analogiquement liés. Le graphe est ensuite utilisé pour construire des communautés de mots partageant une racine commune. De ces communautés nous extrayons pour chacun des mots une analyse morphologique détaillée prenant en compte un large éventail de phénomènes morphologiques. Nous mettrons en évidence que les résultats obtenus par un tel système surpassent non seulement le système analogique de Langlais (2009), mais aussi le système à base de graphe de Bernhard (2010) et le système MDL *Morfessor* (Creutz et Lagus, 2005) sur plusieurs langues. Nous discuterons aussi de la polyvalence de notre système relativement aux autres approches. En plus de tout cela, nous examinerons en profondeur certaines composantes du système et comment elles influencent les analyses morphologiques produites.

Les résultats publiés dans ce document ont été précédemment présentés dans plusieurs articles différents (Lavallée et Langlais, 2009, 2010a,b), mais aucune analyse globale des résultats n'a encore été publiée. Nous tenterons donc de donner ici une analyse cohérente des différentes observations précédemment présentées.

CHAPITRE 2

CONCEPTS DE BASE EN MORPHOLOGIE

L'objectif de ce chapitre est d'établir un vocabulaire de termes morphologiques facilitant ainsi la discussion des motivations linguistiques justifiant nos décisions et les forces et faiblesses qui en résultent. Nous ne décrirons cependant que les concepts nécessaires à l'argumentation des choix effectués et à la délimitation du champ d'application possible du système développé. Nous ferons aussi abstraction de certaines difficultés du domaine pour simplifier la compréhension. Notez aussi qu'il n'y a pas unanimité sur tous les sujets abordés, il se peut qu'il y ait conflit entre nos définitions et celles d'un autre auteur. Le lecteur intéressé par les subtilités du domaine peut se référer aux ouvrages utilisés dans la rédaction de ce chapitre : (Katamba et Stonham, 2006; Spencer et Zwicky, 1998; Aronoff et Fudeman, 2005; Jensen, 1990; Fradin, 2003; Bubenik, 1999; Mel'čuk, 1993, 2006; Saussure, 1968; Anderson, 1992).

2.1 Du mot au morphème

Pour plusieurs, le mot peut sembler être la plus fine unité du langage. Ceci est principalement dû à sa délimitation relativement explicite dans de nombreuses langues. Cependant, il existe des unités du langage plus fines. La division en son des mots d'une langue, appelée **phonème**, en est la plus petite. Il existe une unité intermédiaire entre le mot et le phonème appelé **morphème**. Le morphème est une unité abstraite qui correspond à la plus petite unité significative du langage. Par exemple, le mot *inaliénable* est composé des trois morphèmes *NON*, *ALIÉNER* et *VERBE* → *ADJECTIF*. Chacun de ces morphèmes a une réalisation concrète composée d'un ou plusieurs phonèmes appelés **morphes**. Par exemple, la réalisation en morphe du morphème *NON* est *in*[1]. Plusieurs facteurs influencent la réalisation en morphe du morphème. Par exemple, le morphème *PLURIEL* n'est pas réalisé par le même morphe dans les mots *chats (s)* et *chevaux* (*aux*). Les réalisations diverses d'un même morphème sont appelées **allomorphes**. Le terme de **morphème homonyme** est utilisé pour identifier la relation inverse ; c'est-à-

[1]Pour simplifier, nous considérerons les morphes comme étant de nature graphémique plutôt que phonétique.

dire lorsque deux morphèmes différents ont un morphe en commun. Comme dans le cas du morphe *ment* qui ne représente pas le même morphème dans le mot *déménagement* (FAIT DE) que dans *calmement* (DE FAÇON). Pour démontrer que ce sont réellement deux morphèmes différents, il suffit d'observer que certaines racines peuvent être vues avec les deux morphèmes (*p. ex.* *tiédissement* et *tièdement*).

La frontière entre la morphologie et l'étymologie peut parfois sembler floue. Il est important de noter que pour être considérée comme un morphème, la signification du morphe doit pouvoir être reconnue sans trop de difficulté par un natif de la langue. Prenons pour exemple le terme *hélicoptère*. Il est possible de décomposer ce mot en 2 parties *héli*(*SPIRALE*) et *ptère*(*AILES*). Bien que ces composantes soient des morphèmes de la langue grec, il serait plus controversé de les considérer comme tels lors de l'analyse d'un mot français. L'analyse morphologique habituellement privilégiée dans ce cas sera donc *HÉLICOPTÈRE*.

2.2 Les types de morphèmes

Tous les morphèmes ne sont pas égaux. Leurs rôles dans la composition d'un mot et leurs caractéristiques varient. Une des divisions majeures est entre les morphèmes **libres** et **liés**. Les morphèmes libres sont ceux pouvant se retrouver seuls dans un texte sans être attaché à aucun autre morphème. Comme les mots *feu* ou *chat* qui sont tous deux composés d'un seul morphème libre. Les morphèmes qui ne sont pas libres sont liés. Comme le morphème *able*[2] qui doit être combiné à au moins un autre morphème (*p. ex.* *capable*, *serviable* et *vendable*).

Une autre dimension permettant de séparer les morphèmes est le rôle qu'ils ont dans le mot. Nous décrirons dans la section 2.2.2 les morphèmes formant la base sémantique du mot. Les morphèmes opérant des transformations sur ces morphèmes de base sont décrits à la section 2.2.1.

2.2.1 Morphème dérivationnel et flexionnel

Le phénomène de construction de mots le plus fréquent est la transformation d'un mot par flexion ou dérivation. Ceci est presque toujours effectué par l'ajout d'un morphème

[2]Pour faciliter la compréhension des exemples, nous appellerons les morphèmes tous simplement par leur morphe lorsque la distinction entre morphème et morphe n'est pas nécessaire.

Dérivation	Base	Dérivé	Sens
	cigare	*cigarette*	Petit cigare
ette	*fille*	*fillette*	Petite fille
	bûche	*bûchette*	Petite bûche
	parfait	*imparfait*	Non parfait
in	*probable*	*improbale*	Peu probable
	délibile	*indélibile*	Non délibile
	virus	*viral*	
al	*colosse*	*colossal*	Transforme le nom en adjectif
	ami	*amical*	
	poli	*politesse*	
esse	*gentil*	*gentillesse*	Transforme l'adjectif en nom
	petit	*petitesse*	

Figure 2.1 – Exemples de dérivation.

lié au mot de base.

La **dérivation** est une transformation qui change le sens ou la classe du mot. Les morphèmes *ette* et *al* en sont 2 exemples. Le premier ajoute le sens de petit et le second transforme le nom en adjectif (figure 2.1).

La **flexion** par contre ne change pas le sens du mot, elle existe pour des raisons grammaticales. Les exemples sont nombreux, comme le morphème indiquant le pluriel ou les temps de verbes en français. La figure 2.2 en répertorie quelques exemples.

2.2.2 Racine, lexème, base et *stem*

Dans la littérature sur l'AMNS, il est fréquent de voir les termes racine (*root*), lexème (*lemma*), base (*base*) et *stem* utilisés de façon interchangeable. Ceci malgré le fait que ces termes définissent des concepts différents, bien que liés.

Un **lexème** est un concept abstrait qui regroupe toutes les représentations d'un même élément du vocabulaire. Il est fréquemment représenté par sa forme lexicale en majuscule. Par exemple, le lexème *GRAND* peut prendre les représentations suivantes : *grandes*, *grand*, *grandeur*... etc. Un autre exemple serait le lexème *VOIR* qui comprend les représentations suivantes : *voir*, *vu*, *vue*... etc.

Pour sa part, la **racine** est le noyau irréductible du mot sans aucun morphème d'at-

Inflection	Base	Infléchis	Sens
ons	*partir* *pendre* *rougir*	*partirons* *prendrons* *rougirons*	Future de la 3ieme personne du pluriel
a	*partir* *pendre* *rougir*	*partira* *prendra* *rougira*	Future de la 3ieme personne du singulier
s	*voiture* *chien* *dernier*	*voitures* *chiens* *derniers*	Pluriel du nom

Figure 2.2 – Exemples de flexion.

taché. C'est donc le morphème que l'on retrouve dans toutes les représentations du lexème associé. Par exemple, le lexème *MARCHER* qui comprend entre autres les mots *marche*, *marcherons*, *marcher* et *marcherai* a pour racine *marche*. Les racines sont fréquemment des morphèmes libres (*p. ex.* *chat*, *marche*, *vol*..), mais peuvent aussi être des morphèmes liés. La racine *préd* (*p. ex.* *prédateur*, *prédatrice*, *prédation*...) est un cas de morphème lié.

Le terme ***stem*** indique la partie du mot avant toute flexion. Il y a fréquemment confusion entre le terme racine et *stem* pour la bonne raison que dans plusieurs cas, ils sont identiques. Par exemple, la racine et le *stem* du mot *chiens* sont dans les deux cas *chien*. Par contre, dans le cas du mot *voyageurs*, la racine est *voyage*, mais le *stem* est *voyageur*. Bien que *s* soit un morphème flexionnel, *eur* en est un dérivationnel et fait donc parti du *stem*. La distinction entre racine et *stem* est particulièrement évidente dans les cas de mots composés. Le *stem* du mot *portemanteau* est *portemanteau*, mais il contient les racines *porte* et *manteau*.

Finalement, **base** réfère à toutes séquences qui peuvent être fléchies ou dérivées. Par conséquent, une racine est une base. Le *stem* est aussi une base, mais seulement lorsqu'on parle de morphologie flexionnelle.

2.3 Phénomènes morphologiques

Il existe une multitude de phénomènes par lesquels les morphèmes peuvent être combinés. Bien que la plupart des langues utilisent plusieurs de ces techniques, leur prévalence varie beaucoup d'une langue à l'autre ce qui rend le développement d'un système réellement universel très difficile.

2.3.1 Affixation

L'affixation est de loin le phénomène le plus commun des langues européennes et la grande majorité des systèmes d'AMNS ne reconnaissent que ce procédé morphologique. Cette prédominance est problématique, car elle donne l'impression que l'AMNS est beaucoup plus simple qu'elle ne l'est réellement et que le *stemming* équivaut à l'analyse morphologique. Les sections suivantes tenteront de dissiper le mythe, mais ici nous nous concentrerons sur les trois cas d'**affixations** : préfixation, suffixation et infixation. Chacun de ces cas tient son nom à sa position relative à la racine dans le mot. La **préfixation** indique que l'affixe se retrouve devant la racine et la **suffixation** est toujours après la racine. Pour être en présence d'un cas d'**infixation**, l'affixe doit se retrouver **dans** la racine. Ce cas est très rare aussi bien en français qu'en anglais, mais il est fréquent dans d'autres langues telles que le tagalog ou l'ulwa. La figure 2.3 contient des exemples de ces trois types d'affixations. Notez que la distinction entre ces trois concepts dépend de sa position relativement à la racine à laquelle il est lié et non à sa position dans le mot. Par conséquent, un préfixe ne se retrouve pas nécessairement au début du mot. Par exemple, dans le mot *anti**néo**libéralisme*, *néo* n'est pas un cas d'infixation, mais bien un cas de préfixation. Un préfixe ou un suffixe peut se retrouver à n'importe qu'elle position dans un mot. Comme le suffixe *s* se retrouvant au milieu du mot composé anglais *sword**s**man*, car il est lié à la racine *sword*. Il y a une quatrième forme d'affixation appelée **circumfixe** qui décrit un morphe divisé en 2 parties, une avant et l'autre après la racine. Cependant, son existence est controversée, car certains décrivent plutôt ce phénomène comme un cas de préfixation et de suffixation simultané.

Phénomène	Affixe	Racine	Infléchis	Sens
Préfixation	re	*lire* *dire* *écrire*	*relire* *redire* *réécrire*	Recommencer
Suffixation	s	*chat* *chien* *travailleur*	*chats* *chiens* *travailleurs*	Pluriel
Infixation	ka	*asna "linge"* *arakbus "fusil"* *súlu "chien"*	*askana "son linge"* *arakkabus "son fusil"* *súkalu "son chien"*	Possessif (ulwa)

Figure 2.3 – Exemples d'affixation.

2.3.2 Mutation du *stem*

Un autre cas fréquent dans plusieurs langues est la **mutation du *stem*** (*stem mutation*). Ce phénomène n'opère pas en ajoutant des phonèmes au mot, mais en modifiant ceux existants[3]. Les cas les mieux connus sont les cas d'**apophonie**, principalement dus aux phénomènes d'**ablaut** et d'**umlaut** qu'on retrouve dans les langues germaniques telles que l'allemand (*p. ex.* *vater* ⟶ *väter*) et l'anglais (*p. ex.* *sing* ⟶ *sang, drink* ⟶ *drank*). L'apophonie ne fait référence qu'au cas de mutation de voyelles, le terme de **mutation consonante** est utilisé lorsque la mutation implique une consonne. Le celte et le breton sont particulièrement connus pour la mutation de la première consonne (*p. ex.* *breur "frère"* ⟶ *preur "ton frère"*, *bag "bateau"* ⟶ *pag "ton bateau"* (breton)). Il est à noter que nous ne sommes en présence d'une mutation du *stem* que si la modification est motivée morphologiquement. Par exemple, dans la forme pluriel des mots anglais se terminant par *y* (*p. ex.* *body*), ce dernier est remplacé par un *i* (*p. ex.* *bodies*). Pourtant, nous sommes en présence d'un cas de suffixation, car la motivation de cette transformation est phonologique. La modification permet de préserver le son *i*, car *yes* aurait une tout autre sonorité.

[3]Cependant, dans certaines langues tel le latin, il est fréquent que la mutation du *stem* soit combinés à l'affixation.

2.3.3 Composition

La **composition** consiste à former un nouveau mot en combinant deux racines. Les exemples sont fréquents dans plusieurs langues tels le français (*p. ex.* portemanteau) et l'anglais (*p. ex.* waterfront, fireman). L'allemand est célèbre pour l'utilisation fréquente de ce procédé et surtout pour les longs mots qui en résultent (*p. ex.* zusammengehoerigkeitsgefuehl).

2.3.4 Supplétion

La **supplétion**[4] décrit le cas où la dérivation ou la flexion du mot transforme complètement la forme de base sans respecter de règles établies. Les cas de supplétion se retrouvent dans plusieurs langues et sont caractérisés par l'incapacité à identifier la racine du mot sans en connaître la signification. Par exemple, sans connaître la signification du verbe *ira*, il est impossible d'identifier sa relation avec sa forme à l'infinitif *aller*. Ceci contraste avec les cas d'affixation (*p. ex.* *partira*) où il est beaucoup plus aisé de reconnaître la racine (*p. ex.* *partir*) en ne se basant que sur la forme. Le mot anglais *worst* est un autre cas de supplétion où il n'y a aucune lettre de partagée entre la forme et le mot de base *bad*. Notez que cette transformation ne partage aucune caractéristique commune avec les autres cas de dérivation par ce même morphème. Ceci est vrai autant pour la règle de dérivation standard par affixation de *er* (*p. ex.* *late* → *later*) ou pour un autre cas de supplétion (*p. ex.* *good* → *better*). Lorsque seulement une partie du mot est modifiée, nous sommes dans un cas de **supplétion partielle** (*p. ex.* *think* → *thought*). La distinction entre ce phénomène et la mutation du *stem* provient de la difficulté à reconnaître le mot d'origine, soit par l'absence de règle expliquant le changement ou par le taux élevé de mutation de la forme de base.

2.3.5 Redoublement

Le **redoublement** est un phénomène plutôt rare dans les langues occidentales, mais fréquent dans plusieurs autres. Il consiste en la répétition d'une partie ou de l'ensemble de

[4] Selon Mel'čuk (2006), la supplétion n'est pas un phénomène morphologique au même titre que les autres phénomènes décrits dans cette section. Cependant, comme elle remplit le même rôle, c'est-à-dire de modifier la forme de base pour indiquer une différence morphologique, nous l'avons tout de même incluse dans cette section.

Langue	Base		Redupliqué	
Ilokano	ulo	tête	ululo	têtes
	mula	plante	mulmula	plantes
Malay	orang	homme	orang-orang	gens
	pĕrĕmpuan	femme	pĕrĕmpuan-pĕrĕmpuan	femmes
Dakota	fič	mauvais	fikfič	mauvais (Pluriel)
	hàska	grand	hàskaska	grands
Chinois	ganjing	propre	ganganjingjing	très propre
	luosuo	bavard	luoluosuosuo	très bavard
Chichewa	mwamûna	homme	mwamúnámuna	homme macho
	m-kâzi	femme	mkázíkazi	belle femme
	mu-nthu	humain	munthumúnthu	altruiste

Figure 2.4 – Exemples de redoublement.

la racine pour marquer une différence morphologique. Le redoublement est fréquemment utilisé pour signifier le pluriel ou pour amplifier la signification du mot. La figure 2.4 illustre l'utilisation du redoublement dans plusieurs langues.

2.3.6 Patron-racine

La morphologie **patron-racine** [5] est un phénomène qui diffère de manière fondamentale des autres phénomènes illustrés précédemment. Ce type de morphologie caractérise les langues sémitiques du Moyen-Orient tels l'arabe et l'hébreu et se démarque par la discontinuité des morphes. Dans une telle langue, la construction du mot n'est pas faite en affixant la racine, mais en y appliquant un patron particulier. En arabe par exemple, la racine est composée de 3 consonnes entre lesquelles on ajoute une séquence particulière de voyelles selon un arrangement bien défini pour construire le mot (voir figure 2.5). La construction est similaire en hébreu où les trois consonnes de la racine (*p. ex.* `MLX`) sont entrelacées de voyelle suivant un patron pour donner les différentes flexions et dérivations de la racine (*p. ex.* `MaLaX "il règne"`, `MoLeX "celui qui règne"`, `MeLex "roi"`). Dans ces exemples, une des voyelles en isolation n'a aucune signification. Donc dans le cas de `MaLaX`, nous ne sommes pas en présence de 3 morphes (`MLX` + `a` + `a`),

[5] Anglicisme tiré de root-and-pattern morphology. Aucune traduction française satisfaisante n'a été trouvée dans la littérature.

Racine $C_1C_2C_3$		Singulier $C_1aaC_2iC_3$		Pluriel $C_1uC_2C_2aaC_3$	
KTB	"LIVRE"	KaaTiB	Écrivain	KuTTaaB	Écrivains
QR'	"LIRE"	QaaRi'	Lecteur	QuRRaa'	Lecteurs
TJR	"AFFAIRE"	TaaJiR	Marchand	TuJJaaR	Marchands

Figure 2.5 – Exemples du phénomène patron-racine de l'arabe. La description du patron a été faite pour correspondre à la latinisation utilisée par notre source. D'autres ouvrages peuvent présenter des latinisations différentes pour les mêmes mots. Pour cette raison, il est probable que la description du patron ne corresponde pas au patron tel qu'il est réellement en arabe.

mais bien de 2 (`MLX` + C_1aC_2aC_3). Ce type de phénomène ne peut pas être facilement modélisé par segmentation du mot (*p. ex.* `chien+s`). Les linguistes utilisent donc habituellement des patrons indiquant l'alternance consonne-voyelle (*p. ex.* `CvCvC`) d'où le nom de ce procédé.

2.4 Typologie morphologique des langues

Même si la morphologie varie énormément d'une langue à l'autre, il existe tout de même d'étonnantes similarités entre elles. Les linguistes ont donc créé une typologie regroupant les langues selon leurs caractéristiques morphologiques, bien qu'aucune langue ne corresponde parfaitement à sa catégorie. La typologie souligne donc plutôt les caractéristiques dominantes de la langue étudiée. Par exemple, le français est habituellement considéré comme étant fusionnel bien que les procédés morphologiques typiques des langues agglutinantes et analytiques y sont chose courante. La typologie est donc plutôt utilisée comme une échelle continue utilisée pour comparer les langues entre elles. Autrement dit, toutes les langues sont fusionnelles, mais certaines le sont plus que d'autres. Par exemple, pour décrire le français, on pourrait dire qu'il est plus fusionnel que l'anglais, mais moins que le latin, tout en étant moins analytique que l'anglais. Dans les prochaines sous-sections, nous décrirons chacune des 5 catégories typiquement utilisées en typologie morphologique.

2.4.1 Analytique

Les langues **analytiques**, aussi appelées **isolantes**, sont caractérisées par un faible taux de morphèmes par mot et la rareté des morphèmes liés. Dans ces langages, le mot est fréquemment une racine ou un composé de racine sans aucune inflexion ou dérivation. Les exemples classiques de ce type de langage sont le chinois et le vietnamien. L'anglais est aussi très analytique pour une langue européenne comme cela est démontré par le pluriel et le singulier du verbe. Ce dernier est indiqué seulement par le pronom `I` ou `we` sans modification à la réalisation du verbe (*p. ex.* `I walk`, `we walk`). Par contre, dans une langue moins analytique, tel le français, le pluriel est associé à une transformation de la forme (*p. ex.* `je marche`, `nous marchons`). L'anglais n'est toutefois pas totalement analytique, car le verbe est tout de même infléchi pour indiquer que l'action a été complétée (*p. ex.* `cook` → `cooked`), ce qui n'est pas caractéristique de ce type de langues. Pour ce cas, le chinois est plus analytique, car le morphème utilisé est libre au lieu de lié. Ce morphème est représenté par le mot `le` ajouté après le verbe (*p. ex.* `chão "cook"` → `chão le "cooked"`).

2.4.2 Agglutinante

Les langues **agglutinantes**, tout comme les langues analytiques, sont caractérisées par une relation 1 à 1 entre le morphe et le morphème, mais diffèrent par un nombre élevé de morphèmes par mot. Les exemples de la Figure 2.6 démontrent bien l'élément prédominant de ce type de langue. Même pour un non-initié à la langue turque, il est possible de reconnaître les morphes présents dans l'exemple précédent et d'inférer le morphème associé à chacun. Une autre caractéristique importante de ce type de langue est la constance des morphes d'un mot à l'autre. En effet, l'allomorphie est plutôt rare et les mots sont construits en agglutinant (d'où le nom) ensemble plusieurs morphes sans en modifier la forme. Par conséquent, une connaissance de quelques morphes de base permet d'analyser une multitude de mots. Cependant, les mots peuvent devenir très longs et la racine peut devenir difficile à identifier.

el	main	elimde	dans ma main
elim	ma main	ellerim	mes mains
eler	les mains	ellerimde	dans mes mains

Figure 2.6 – Le principe d'agglutination démontré par quelques termes turcs.

2.4.3 Fusionnelle

On appelle **fusionnelle** ou **flexionnelle** les langages dont les mots sont composés de plusieurs morphèmes, mais de peu de morphes. Ce qui implique qu'un morphe représente fréquemment plusieurs morphèmes dans un même mot. Cette catégorie caractérise la plupart des langues européennes, probablement dues à l'influence du latin qui est fortement flexionnel. La conjugaison des verbes en français est un bon exemple de ce type de langue. Par exemple, la fin de verbe `ai` (*p. ex.* `je pourrai`) indique la première personne, le singulier et le futur. La modification de seulement un de ces morphèmes transforme complètement le morphe commun. Si l'on remplace le morphème `SINGULIER` de l'exemple précédent par `PLURIEL`, le morphe `ai` est substitué par `ons` (*p. ex.* nous pourrons). Ceci indique qu'aucun des trois morphèmes n'a de réalisation propre dans ce mot. Pour mieux illustrer ce principe, on peut comparer cette dernière flexion à celle des adjectifs de cette même langue qui suivent un processus d'agglutination. En comparant les formes `grand` (Masculin singulier), `grande` (Féminin singulier) et `grandes` (Féminin pluriel), on peut voir que le féminin est représenté par le morphe `e` et le pluriel par `s`. Les mots sont obtenus en juxtaposant ces morphes et non par trois morphes totalement différents comme il serait d'usage dans une langue purement fusionnelle. Les mots des langues fusionnelles ont donc tendance à être courts malgré la richesse de l'information qu'ils contiennent. Par contre, cela rend les mots plus difficiles à analyser morphologiquement en raison du grand nombre de formes irrégulières et à l'absence de correspondance claire entre le morphe et le morphème.

2.4.4 Polysynthétique

Les langues **polysynthétiques** sont similaires aux langues agglutinantes en raison de la correspondance 1 à 1 entre le morphe et le morphème et par la constance de la réalisation des morphèmes. Ce qui les différencie est l'étendue de l'agglutination. En effet,

les mots d'une langue polysynthétique sont composés en général de plusieurs morphèmes racines. Il est même fréquent qu'un seul mot contienne toute l'information véhiculée par une phrase complète dans un autre langage, englobant ainsi le sujet, le verbe et le complément. Par exemple le mot groenlandais *tuttusivuq "il a vue un caribou"* est composé de 3 morphèmes : Le complément *tuttu "caribou"*, le verbe *si "vue"* et le pronom *vuq "il a"*. Ce type de morphologie est commun auprès des populations autochtones d'Amérique.

2.4.5 Templatique

Les langages **templatiques**[6] se distinguent des autres typologies par l'importance du positionnement des lettres (morphologie patron-racine, section 2.3.6). Contrairement aux autres catégories, les morphes disjoints sont fréquents et la composition graphémique n'est pas suffisante pour identifier le morphe. En effet, il faut aussi prendre en considération la façon dont le morphe est entrelacé avec la racine. L'analyse morphologique classique par segmentation est donc mal adaptée à ces langues. L'exemple habituel de ce type de langue est l'arabe et l'hébreu où la morphologie patron-racine est prédominante. L'arabe démontre aussi le fait que les typologies ne sont pas exclusives, car non seulement cette langue est l'exemple par excellence pour illustrer les langues templatiques, elle est aussi fortement flexionnelle.

2.5 Paradigme

Le **paradigme** est un regroupement de lexèmes partageant les mêmes règles de flexion. Donc, tous les lexèmes suivant un même paradigme partagent les mêmes morphèmes flexionnels et ces morphèmes sont réalisés par les mêmes morphes. Une des caractéristiques les plus importantes du paradigme est sa constance qui permet de mettre un peu d'ordre dans le système de flexion des langues fusionnelles. Malgré cela, un lexème peut diverger de façon mineure de son paradigme. Dans ce cas, on le dit **incomplet**. Les verbes français illustrent parfaitement le concept de paradigme. Tous les verbes français peuvent être observés avec le même ensemble de combinaison de morphèmes (temps, personne et nombre), mais la réalisation de ces morphèmes change d'un verbe à l'autre

[6] Anglicisme du terme *templatic*. Aucune traduction française décente n'a été trouvée pour ce terme.

(*p. ex.*`je reçois`, `je parle`). Le morphe dépend de l'appartenance du verbe à un des 3 paradigmes déterminés par la fin du verbe à l'infinitif (`-er`,`-oir` et `-re`).

2.6 Productivité

Un mythe répandu à propos du lexique est que les mots qui le composent sont prédéterminés et statiques. Contrairement aux phrases, le locuteur mémoriserait un ensemble limité de mots qu'il utiliserait au besoin. Peu d'experts adhèrent à cette vision et le lexique est plutôt considéré comme une entité en constante mutation à laquelle le locuteur crée de nouveaux mots selon ses besoins. Anshen et Aronoff (1988) ont avancé la théorie selon laquelle le locuteur a 3 processus à sa disposition pour trouver le mot juste. Il peut rechercher le mot dans son lexique mental ou il peut en créer un nouveau soit par règle ou par analogie. Le lexique n'est donc pas seulement composé des mots existants de la langue, mais aussi de tous les mots potentiels. Cependant, tous ces mots n'ont pas le même potentiel de réalisation, certains processus de création sont plus probables que d'autres. Les morphologues[7] utilisent le terme de **productivité** pour indiquer le potentiel qu'une certaine transformation morphologique soit utilisée dans la création d'un nouveau mot plutôt qu'une autre concurrente. L'exemple typique de processus concurrent est la suffixation de `ness` (*p. ex.* `cordialness`) et de `ity` (*p. ex.* `productivity`) en anglais qui indiquent tout les deux la mesure d'une qualité. La première est habituellement considérée plus productive que la deuxième. Cependant, il n'y a pas de formule bien établie pour déterminer la productivité d'un processus morphologique. À cause de cela, certains chercheurs préfèrent considérer la productivité comme étant une valeur stricte. Ce point de vue n'est pas partagé par la majorité et différentes formules ont été proposées avec un succès mitigé . L'estimation la plus simple est que la productivité dépend du nombre de formes construites de cette façon. Si l'on applique cette technique à notre exemple de `ness`/`ity`, la première est comme prévu plus productive. Cependant, cette technique est fortement critiquée, car elle ne tient pas compte des contraintes d'applications des processus de transformation. Un processus pouvant s'appliquer à un grand nombre de racines, mais de façon non systématique sera donc plus productif qu'un processus ap-

[7]Notez que morphologue n'est pas un terme existant. Il a été créé pour les besoins du texte courant. Il existait d'autres possibilités telles que morphologueur ou morphologiste. On pourrait donc dire que si cette préférence est généralisée, `-gue` est plus productive que `-giste`.

plicable systématiquement à un nombre limité de racines. Il a donc été suggéré que la productivité d'un processus soit calculée en tenant compte du nombre de formes auxquelles il peut s'appliquer. Dans ce cas, le taux de productivité dépendra des contraintes considérées et variera donc significativement d'un linguiste à l'autre. Par exemple, si la contrainte acceptée sur le calcul de la productivité de *ness*/*ity* est de considérer seulement les adjectifs, *ness* restera dominant. Par contre, en considérant seulement les termes scientifiques, *ity* devient plus productif. Un des travaux influents sur cette question est celui de (Baayen, 1992) qui suggère que la productivité doit être calculée non pas sur un lexique, mais sur un corpus. Selon lui, la productivité dépend des hapax, car ils ont une probabilité plus élevée d'avoir été créés par un processus productif. L'idée derrière cela est que les mots courants proviennent généralement du lexique mental de l'auteur contrairement aux hapax qui ont probablement été créés par le locuteur pour une utilisation particulière. Donc, la formule qu'il suggère pour calculer la productivité d'une transformation morphologique (*p. ex.* `X-able`) est $Prod = n_1/N$ où N est le nombre de mots du texte respectant cette transformation (*p. ex.* `vivable`, `aimable`) et n_1 est le nombre de ces N mots qui sont des hapax. Cette métrique a été critiquée pour les résultats absurdes qu'elle obtient sur certains cas. Par exemple, une transformation vue une seule fois aura une productivité de 1. Ce qui voudrait dire qu'elle est totalement productive bien que cela soit très peu probable.

2.7 Analogie

L'analogie est un concept clé dans la théorie de l'évolution du langage de Saussure (1968) permettant d'expliquer l'introduction de nouveaux mots et la transformation des mots existants. Une forme analogique est une forme créée à l'image d'une autre notée sous la forme [*x* : *y* = *z* : *t*] où la quatrième proportionnelle t est la nouvelle forme introduite au langage. Par exemple, l'analogie [*porter* : *portable* = *poster* : *t*] aurait comme résultat *t* = *postable*. L'analogie agit donc en faveur de la régularité et tend à unifier les procédés de dérivation et de flexion. Il est cependant évident que tout groupe de 3 mots du vocabulaire ne forme pas nécessairement une analogie. Pour cela, les mots doivent être liés dans le sens et dans la forme, un de ces éléments seuls n'est pas suffisant (Anderson, 1992). Si les mots ne sont liés que dans la forme, le mot résultant n'aura aucun sens propre

comme illustré par l'analogie suivante : [*faire* : *taire* = *foire* : *t*] dont le résultat est *t* = *toire*. Par contre, si l'on observe l'analogie [*livre* : *couverture* = *maison* : *t*], il est clair que la signification de *t* est l'enveloppe d'une maison, mais aucune base ne permet d'identifier la forme de *t*. Même si ces éléments sont réunis, l'analogie n'est pas nécessairement valide elle doit aussi être basée sur une relation régulière. Par exemple, l'analogie [*ear* : *hear* = *eye* : *t*] *t* = *heye* respecte nos 2 conditions de base, mais la relation formelle entre *ear* et *hear* est trop isolée pour former une analogie. Par isolé, nous voulons dire que la relation *X* → *hX* se retrouve dans peu de formes, voire aucune autre.

La morphologie et l'évolution du langage sont deux domaines différents. Par conséquent, l'importance du concept d'analogie dans le domaine de l'évolution de la langue n'implique en aucun cas que ce même concept est pertinent en morphologie. Cependant, il y a certainement des liens à faire entre les phénomènes morphologiques et l'analogie. Selon Saussure, la création analogique apparaît comme un chapitre particulier du phénomène de l'interprétation[8]. Le concept d'analogie est donc un processus créatif dans le sens qu'il crée de nouveaux mots, mais il est non créatif dans le sens que les unités de base (morphèmes) de cette création doivent préexister dans la langue. Saussure conçoit donc la création analogique de deux façons, par la reconstruction de la 4^e proportionnelle, telle qu'utilisée dans les exemples précédents, et par analyse. L'analyse consiste à décomposer le mot en ses unités pour expliquer la transformation. Par exemple, la formule [*aimer* : *aimable* = *entamer* : *x*] revient à dire qu'il est possible de remplacer *er* par *able*. Cette dernière formulation met en évidence le morphe *able* présent dans la forme *aimable* et *entamable*.

2.8 Morphologie basée sur le mot

Comme les sections précédentes de ce chapitre le démontrent, l'école de pensée dominante en morphologie considère le morphème comme étant l'élément central de leur théorie. La forme du mot ne serait que le résultat des morphèmes qui le composent et de règles phonétiques qui s'appliquent. Cette vision des choses n'est pas partagée par tous. Les partisans de l'approche de la morphologie basée sur le mot rejettent jus-

[8] Distinction des unités.

qu'à l'existence même du morphème. Selon eux, la morphologie ne devrait pas être étudiée comme l'agencement d'unités, mais plutôt comme une règle expliquant la relation sémantique entre 2 mots. Dans le livre *Pace Pānini* (Ford et al., 1997), les auteurs soutiennent que toute relation morphologique entre 2 mots du lexique peut s'exprimer par une stratégie de formation de mot ("Word Formation Strategy" ou WFS). Ces WFS sont des règles généralisant la relation morphologique entre 2 termes X et X' sous le format suivant $/X/_{\alpha} \leftrightarrow /X'/_{\beta}$. Les symboles α et β étant des catégories lexicales et ' représentant la différence formelle entre les deux formes. Par exemple, la relation entre les termes `gros` et `grosse` ou `bas` et `basse` serait $|X|_{Adj.M} \leftrightarrow |Xse|_{Adj.F}$ [9]. Des règles plus complexes peuvent aussi être admises pour capturer des relations plus complexes telles que des règles de formation patron-racine et des mutations du *stem*. Par exemple, la relation entre les termes espagnols `cuento` et `contamos` serait exprimée par la règle $|Xue(C(C))o| \leftrightarrow |Xo(C(C))amos|$[10].

[9] Le caractère | a été utilisé au lieu de / pour indiquer que la règle est représentée par des graphèmes au lieu de par des phonèmes pour des raisons de simplicité. La règle phonétique serait $/X/_{Adj.M} \leftrightarrow /Xs/_{Adj.F}$

[10] $/Xwe(C(C))o/ \leftrightarrow /Xo(C(C))àmos/$

CHAPITRE 3

ÉTAT DE L'ART

Depuis plus de 50 ans, de nombreux chercheurs se sont intéressés au domaine de l'AMNS[1]. Au cours de ces années, un grand nombre de solutions ont été proposées mais la plupart des travaux peuvent être divisés en 4 grandes familles d'approches qui seront décrites dans ce chapitre (section 3.1). Nous mentionnerons aussi certains travaux plus en marge du domaine (section 3.2) en plus de décrire en profondeur le système le plus influent du moment (section 3.3).

3.1 Approches courantes à l'acquisition morphologique

3.1.1 Calcul de l'incertitude

Une des approches les plus répandues à l'AMNS consiste à calculer des statistiques sur les séquences de lettres des mots du lexique. Ces statistiques mesurent le niveau d'incertitude de la prédiction de la lettre suivant une séquence de lettres donnée. L'idée derrière cela est qu'un morphème devrait être vu en combinaison avec plusieurs autres morphèmes. Par conséquent, les lettres situées aux frontières d'un morphème devraient être difficiles à prévoir. Ce type de système est loin d'être nouveau, Harris (1955) en a fait la description il y a plus de 50 ans. Il proposait de segmenter le mot aux endroits où le nombre de phonèmes pouvant possiblemement suivre dépasse un seuil donné. Par la suite, Hafer et Weiss (1974) ont approfondi les recherches de Harris en testant plusieurs métriques de segmentation, dont l'entropie. Ils démontrèrent aussi qu'il était possible de travailler sur les lettres au lieu des phonèmes. Ce qui, bien que théoriquement moins solide linguistiquement, est beaucoup plus pratique. Plus récemment, Bernhard (2008) a obtenu des résultats comparable à l'état de l'art en utilisant cette technique pour obtenir automatiquement un dictionnaire de préfixes et de suffixes d'une langue. De plus, *Morfessor CatMAP*, la variante la plus performante du populaire système *Morfessor* (Creutz et Lagus, 2005), utilise aussi une métrique dérivée de l'entropie pour classifier les morphèmes qu'il trouve. Cependant, ce type d'approche considère que les morphèmes ont

[1] Analyse Morphologique Non Supervisée.

nécessairement une frontière nette dans le mot. Ceci n'est pas toujours vrai et plusieurs phénomènes morphologiques sont tout simplement ignorés. De plus, même si une frontière non ambiguë existe, certains morphèmes ne seront pas identifiés, car la prémisse n'est pas toujours vraie. Il existe des morphèmes n'apparaissant que dans un nombre limité de contextes graphémiques. Comme le suffixe anglais `ive` (*p. ex.* `addictive`) qui est toujours précédé d'un `t` ou bien d'un `s` et dont l'entropie est par conséquent faible.

3.1.2 Minimisation de la taille de description

D'autres chercheurs abordent l'analyse morphologique comme un problème de compression. Ils décomposent les mots de façon à permettre l'encodage optimal du lexique en utilisant le concept du *Minimum Description Length* (MDL). L'idée fondatrice du MDL est que la meilleure façon de capturer une caractéristique régulière des données est de construire un modèle capable de les décrire de façon optimale (Rissanen, 2008). Les morphèmes étant en grande partie définis comme un phénomène régulier, cette approche semble donc appropriée. Cette idée est le fondement des deux systèmes les plus influents du domaine actuellement, *Morfessor* (Creutz et Lagus, 2005) et *Linguistica* (Goldsmith, 2001). *Linguistica* utilise le principe MDL pour regrouper les mots partageant la même *signature*[2]. Ce système impose une structure stricte racine-suffixes aux mots analysés. Ceci rend le système peu apte à gérer les langages morphologiquement complexes. *Morfessor* comble cette lacune en n'imposant aucune structure prédéfinie, permettant ainsi d'identifier les phénomènes morphologiques apparaissant à toute position dans le mot. Ce système a tendance à segmenter les séquences très fréquentes quelques soit sa position dans le mot. Par exemple, le mot `entrevue` serait découpé en deux morphèmes `ent+revue` car `ent` est un suffixe commun du français qui identifie la 3^{ieme} personne du pluriel. La variante *Morfessor CatMAP* (Creutz et Lagus, 2005) règle ce problème en utilisant un HMM qui conditionne la segmentation d'un morphe suggérée par l'algorithme de MDL sur le type de morphe le précédant (voir section 3.3). L'approche MDL a l'avantage d'avoir de solides fondements mathématiques. Par contre, les fondements linguistiques ne sont pas aussi solides. De plus, tout comme l'approche par incertitude (section 3.1.1), elle ne détecte que les cas où les morphèmes sont tout simplement juxtaposés les uns aux autres.

[2]Concept introduit par Goldsmith similaire aux paradigmes.

3.1.3 Utilisation du contexte

La plupart des approches considèrent la morphologie comme un phénomène interne au mot. Par conséquent, les mots sont analysés de façon détachée les uns des autres. Par contre, le contexte d'un mot dans un corpus peut contribuer grandement à en reconnaître certains morphèmes. Il est facile d'imaginer qu'un mot contenant le morphème *PLURIEL* aura dans son entourage des mots tels que *les*, *des* ou *mes*. Ce type d'approche se démarque en raison de son indépendance envers la représentation en morphe du morphème. Un terme comme *hiboux* aura un contexte similaire à *chats* et permettra donc d'identifier le lien morphologique unifiant les 2 formes. Ceci est très intéressant dans le cas de formes irrégulières comme démontré par Yarowsky et Wicentowski (2000) pour les verbes irréguliers anglais. Malheureusement, ce type de technique s'applique seulement à la morphologie inflectionnelle (pluriel, genre, temps... etc.). Le contexte seul n'est pas suffisant à une analyse morphologique poussée et est donc fréquemment combiné à d'autres approches. La façon la plus courante d'utiliser le contexte en AMNS est de tenter de regrouper les mots par catégorie syntaxique en se basant sur le contexte. La catégorie syntaxique étant une des contraintes les plus fréquentes sur la morphologie, l'ajout de cette information est certainement un atout important à tout système. Par exemple, en français, le morphe *able* peut être suffixé au verbe pour obtenir un adjectif (*p. ex.* *scier* → *sciable*). Ceci pourrait laisser croire que le mot *cartable* a été dérivé de cette façon du verbe *carter*. L'information syntaxique permet d'éviter cette erreur, car *cartable* n'étant pas un adjectif, il n'a pu être produit par cette règle. L'information syntaxique s'intégrant relativement bien à une multitude de techniques, elle fait partie de multiples systèmes. Par exemple, le contexte a été utilisé en combinaison au système *Linguistica* pour prédire les morphèmes pouvant être affixés à une racine, et ce, même si la forme affixée n'existe pas dans le lexique (Hu et al., 2005). Pour ce faire, les préfixes et les suffixes de chacun des mots du lexique sont obtenus par une approche MDL (*Linguistica*). Par la suite, les listes d'affixes similaires ayant des racines se retrouvant dans des contextes proches dans un corpus sont regroupées. Il est aussi mentionné dans cet article qu'il est possible d'utiliser le contexte pour reconnaître les *allomorphes* (voir section 2.1). Bordag (2008) montre qu'il est avantageux d'enrichir d'informations syntaxiques un modèle basé sur l'incertitude (Section 3.1.1). Au lieu de construire un modèle global à tous les

mots, il en fait un par groupe de mots partageant des contextes similaires. Dans la même veine, Can et Manandhar (2009) a employé un algorithme permettant d'identifier les *paradigmes* (voir section 2.5) sur les mots d'un même regroupement syntaxique obtenu par similitude de contexte. Le contexte peut aussi être utilisé pour détecter plus d'information que la catégorie syntaxique. Schone et Jurafsky (2001) utilisent le contexte pour catégoriser sémantiquement les mots. Cette information est utilisée pour restreindre les liens morphologiques potentiels obtenus par calcul de l'incertitude.

3.1.4 Paradigme

Le concept de paradigme (section 2.5) est un concept majeur en morphologie. L'idée générale est que les *lexèmes* (voir section 2.2.2) peuvent être regroupés selon les flexions qui leur sont appliquées. Plusieurs systèmes tirent avantage de ce concept avec succès. En particulier le système *Paramor* (Monson et al., 2007) qui obtient régulièrement les meilleurs résultats aux ateliers morpho challenge. *Paramor* recherche les suffixes partageant le même *stem* (voir section 2.2.2). Les *stem*s ayant été observés avec le même ensemble de suffixes sont ensuite regroupés en paradigme. Comme il est probable que plusieurs des représentations d'un lexème du paradigme ne soient pas présentes dans le lexique, les paradigmes similaires sont aussi fusionnés. Le modèle mathématique utilisé par *Paramor* est plutôt simpliste et ne permet pas d'attribuer une probabilité à l'appartenance d'un mot à un paradigme. Chan (2006) a tenté de développer un algorithme moins heuristique. Cet algorithme probabiliste basé sur Latent Dirichlet Allocation n'obtient malheureusement des résultats compétitifs que dans les cas supervisés ou semi-supervisés. Tous ces algorithmes, tout comme la plupart des approches décrites précédemment, ne sont adaptés qu'aux cas où le mot est construit par juxtaposition de morphes ; c'est-à-dire lorsque la frontière de morphèmes est non ambigüe. Cependant, elle a tout de même la capacité de gérer certaines irrégularités dans les morphes. Ceci grâce à la constance des paradigmes. Donc dans le cas où deux paradigmes potentiels divergent de peu dans leurs flexions, on peut identifier le morphème commun entre deux formes même s'il n'y a pas de graphèmes communs. Prenons comme exemple les lexèmes de la langue anglaise *SING* (*sing*, *sang*, *sings* et *singing*) et *LOOK* (*look*, *looked*, *looks* et *looking*). Sans utiliser le concept de paradigme, il serait difficile d'identifier la relation entre *sang* et *looked* due à l'irrégularité du lexème *SING*. Par contre, comme les flexions des lexèmes *SING*

et *LOOK* partagent plusieurs similitudes, on peut en conclure qu'ils font partie du même paradigme et que donc *sang* est fort probablement la flexion équivalente à *looked*. Un point important à noter est que bien que théoriquement le concept de paradigme ne s'applique qu'à la morphologie flexionnelle, en pratique cette technique peut aussi détecter un nombre limité de cas de morphologie dérivationnelle.

3.2 Approches marginales à l'acquisition morphologique

3.2.1 Approche à base de règle

La principale critique que l'on pourrait apporter aux approches les plus courantes en AMNS est qu'elles ne peuvent que capturer un nombre limité de processus morphologiques. En effet, la plupart des systèmes ne peuvent que décomposer le mot limitant ainsi l'analyse aux cas d'agglutination de morphes. De ce fait, seulement les cas *d'affixations* (section 2.3.1)[3] sont identifiés sans même permettre de mutation mineure au *stem*. Cela malgré que les gains potentiels de permettre ces mutations sont élevées pour certaines langues. Ceci est démontré par l'amélioration des résultats (Kurimo et al., 2009) obtenus pour la langue anglaise par le système *Allomorfessor* (Virpioja et Kohonen, 2009) comparé au système sur lequel il est fondé. L'amélioration est obtenue en ajoutant au système probabiliste *Morfessor* la probabilité d'une mutation du *stem* lié à l'affixation de la base. Le but de cela est de permettre des analyses plus cohérentes en regroupant sous une même dénomination plusieurs morphes en relation d'allomorphie. Par exemple, l'analyse du mot *bodies* sera *body+es* contrairement à *bod+ies* pour *Morfessor*. L'analyse d'*Allomorfessor* permet de faire le lien entre *bodies* et *body*. La mutation du *stem* est modélisée par une règle de réécriture associée à l'affixe tel que es(*y|i*) dans notre exemple.

Malgré cela, l'amélioration apportée par *Allomorfessor* n'est que superficielle et l'analyse reste limitée aux cas d'affixations. Ceci est dû aux fondements théoriques sur lesquels reposent ce système et plusieurs autres qui considèrent le morphème comme l'élément central de la morphologie, le mot n'étant qu'une juxtaposition de morphe. La perception de la tâche d'AMNS découlant intuitivement de cette vision de la morphologie

[3]Les systèmes se limitent fréquemment à la *suffixation* bien que la *préfixation* soit de plus en plus courante. La détection des cas *d'infixation* reste très marginale.

se résume à 2 étapes : identifier les morphes de la langue et ensuite trouver l'agencement de morphes ayant produit chacun des mots du lexique. La tendance envers cette conception de l'AMNS est mise en évidence par la pratique qui consiste à utiliser les systèmes d'analyse morphologique pour résoudre le problème de segmentation des mots. Dans cette dernière tâche, un texte sans espaces et ponctuations est fourni au système qui doit segmenter ce texte en mots. Cette conception découle d'une simplification de la pensée dominante dans le domaine de la morphologie. Il existe par contre, des théories alternatives qui permettraient de contourner ces limitations. Par exemple, un système basé sur la théorie de la morphologie basée sur le mot (MBM) (section 2.8) aurait potentiellement la capacité de modéliser un grand nombre de processus morphologiques. L'expérience a été tentée par Neuvel et Fulop (2002) qui ont développé un système qui prédit les formes potentielles de la langue en se basant sur une liste de mots associés à leur catégorie syntaxique. Ceci est fait en observant les ressemblances et les différences entre chacune des paires de mots de la liste fournie dans le but de trouver des règles de réécriture mettant en relation les formes morphologiquement liées. Les paires de mots ayant les mêmes divergences sont regroupées et une règle de réécriture est extraite de chacun des groupes. Le contexte d'application le plus strict applicable à tous les éléments du groupe est ajouté à la règle. Par exemple, la règle obtenue des paires *deception-deceive*, *reception-receive* et *perception-perceive* est */*##ception/*$_{Ns}$ $\leftrightarrow$ */*##ceive/*$_{V}$. Les symboles # et * indiquent des caractères quelconques restant identiques dans les 2 formes. La différence entre # et * est que # doit correspondre à un caractère alors que * peut être n'importe quelle chaîne de caractère. Cette approche est applicable à plusieurs langages, car elle peut identifier non seulement les affixations (*p. ex.* do-doable /##/$_{v}$ $\leftrightarrow$ /##able/$_{adj}$), mais aussi des cas qui seraient impossibles à analyser par segmentation (*p. ex.* KaaTTiB-KuTTaaB /#aa##i#/$_{Ns}$ $\leftrightarrow$ /#u##a#/$_{Np}$). Cependant, cette approche ne peut donner une analyse morphologique approfondie telle que celle obtenue par les autres systèmes. Ce système peut identifier qu'il existe une relation morphologique entre deux mots, mais ne peut pas reconnaître la nature de cette relation. Il y a aussi une ambiguïté évidente pour les langues *polysynthétiques* (section 2.4.4) où il y a plusieurs relations potentielles. Par exemple, dans le cas de *qimmiqarpuq "il a un chien"* il est difficile de dire si le mot est en relation avec le verbe *avoir* ou avec le nom commun *chien*.

D'autres chercheurs moins puristes utilisent le principe de règle morphologique tiré de la théorie MBM pour ensuite en extraire une décomposition en morphèmes qui ne se limite pas en une segmentation du mot. Notre système ainsi que le système *MorphoNet* (Bernhard, 2010) correspondent à cette description. Ces 2 systèmes ont des structures similaires. Ils tentent en premier lieu d'identifier les règles mettant en relation morphologique les mots du lexique. Ensuite, un graphe illustrant ces relations est construit à l'aide des règles trouvées et des communautés de mots sont extraites de ce graphe. Ces communautés sont finalement utilisées pour extraire une décomposition morphologique pour chacun des mots. Cette analyse est obtenue en se basant sur les règles liant le mot analysé aux autres mots de la communauté. Cependant, les 2 systèmes divergent de façon substantielle sur la façon d'accomplir chacune de ces étapes. *MorphoNet* se démarque surtout par son formalisme de règle flexible basé sur les expressions régulières. Par exemple, la règle obtenue pour la paire de mots *revisiter-visite* serait $\hat{}re(.*)r\$ \rightarrow \backslash 1$.

3.2.2 Analogie

L'analogie est un concept en théorie linguistique qui permet d'expliquer l'évolution de la langue soit par la création de mots ou la modification de mots existants (Section 2.7). Bien que l'évolution de la langue et la morphologie soient deux domaines distincts, il est clair que la création de *néologismes*[4] suit fréquemment un procédé morphologique. L'ambiguïté entre les deux domaines est tellement commune qu'elle a été répertoriée parmi les 3 erreurs courantes de la morphologie par Mel'čuk (2006). Par exemple, l'introduction à la langue anglaise du terme *Google* comme verbe d'usage commun a rapidement été suivie par l'introduction de ses formes infléchies *Googled* et *Googling*. Il est probable que l'insertion de ces termes se soit fait par analogie entre *Google* et d'autres paires de mots morphologiquement liées telle que [*search* : *searched* = *Google* : *x*] où *x* = *Googled*. Cette analogie capture le morphe *ed* qui est une des représentations possibles du morphème *PASSÉ* en anglais.

Bien que l'analogie soit bien implantée en linguistique, elle a eu un départ difficile en AMNS. En effet, Pirrelli (1993) rapporte une correspondance faible entre l'analogie et les morphèmes. Heureusement, ses travaux subséquents ont donné des résultats encourageants par leur rappel très élevé sur la tâche d'identification de la racine des mots

[4] Tout mot récemment introduit dans la langue.

d'un lexique (Pirrelli, 1999). Il n'en reste pas moins que jusqu'à récemment, l'analogie ne fut pas très populaire dans le domaine de l'AMNS. Ceci est dû en partie au manque de résultats concluants et au coût de calcul prohibitif de cette technique. L'analogie a principalement gagné en popularité par son application en traduction où sa capacité à identifier les relations entre les mots fut implicitement utilisée pour traduire les mots inconnus d'un modèle de traduction (Lepage et Denoual, 2005; Langlais et Patry, 2008; Denoual, 2007). Pour ce qui est de la morphologie, Stroppa et Yvon (2005) ont démontré qu'il était possible d'utiliser l'analogie pour identifier le lexème et certaines caractéristiques des mots en anglais, en néerlandais et en allemand. Hathout (2002, 2008) quant à lui, classe automatiquement les mots d'un lexique en familles morphologiques en combinant l'analogie à de l'information syntaxique obtenue d'autres sources. Moreau et al. (2007) ont utilisé l'analogie pour faire de l'AMNS, mais en imposant des restrictions très strictes sur les analogies acceptées, sacrifiant ainsi une grande partie du potentiel de l'analogie. Récemment, Langlais (2009) a obtenu des résultats encourageants sur une tâche de décomposition morphologique des mots de la base de données CELEX en utilisant une technique simple détaillée à la section 4.2. Finalement, Goldsmith (2007) a combiné le principe d'analogie à celui du MDL et a obtenu de bons résultats sur de petits lexiques anglais et swahili.

3.3 Le système *Morfessor*

Le nom *Morfessor* regroupe sous une même appellation une série de modèles développés par Creutz et Lagus du Helsinky Institute of Technology. Par sa facilité d'utilisation, sa rapidité et ses bons résultats, *Morfessor* est rapidement devenu la référence dans le domaine. Ce système est basé sur le principe du MDL, tous comme le populaire système *Linguistica*. Il se différencie de ce dernier par sa capacité à analyser les mots ayant une composition morphologique riche due à une plus grande flexibilité sur les schémas d'agglutinations de morphes permis (section 3.1.2). Ceci se traduit par un rappel plus élevé sur certaines langues comme le finnois (Creutz et Lagus, 2007). Bien qu'il y ait eu plusieurs variantes du modèle, deux sont fréquemment utilisées *Morfessor Baseline* et *Morfessor CatMAP*. *Morfessor Baseline* est une version purement MDL alors que *Morfessor CatMAP* utilise les résultats de *Morfessor Baseline* en ajoutant certains

traitements supplémentaires.

Le mécanisme de *Morfessor Baseline* consiste à identifier parmi toutes les segmentations en morphèmes possibles celle qui en plus d'avoir un lexique de morphème concis permet d'encoder le lexique de façon optimale. Pour ce faire, *Morfessor* utilise le principe du maximum a posteriori[5] pour sélectionner la décomposition optimale parmi celles suggérées. La recherche de la segmentation optimale se fait de façon gloutonne. Les mots sont analysés un à un en ordre aléatoire. Pour chacun des mots, toutes les segmentations possibles sont considérées, mais seulement celle qui améliore le plus le résultat du maximum a posteriori est conservée. Tout le lexique est analysé de cette façon jusqu'à ce qu'une itération sur le lexique complet n'apporte pas une amélioration jugée suffisante.

La flexibilité de *Morfessor Baseline* n'a pas que des avantages. Le manque de modélisation de la structure des morphes cause dans certains cas la segmentation d'affixes typiques de la langue à des positions non orthodoxes. Par exemple, le système à tendance à suggérer des suffixes au début des mots comme dans ces exemples de la langue anglaise : `ed+ward`, `s+urge+on`, `s+well` (Creutz et Lagus, 2005). Pour résoudre cela, *Morfessor CatMAP* ajoute quelques étapes supplémentaires au mécanisme de base dans le but d'éliminer ces segmentations peu pertinentes. Ce système innove, entre autres, en utilisant la perplexité pour classifier (suffixe, préfixe, *stem*, bruit) chacun des morphes identifiés par le mécanisme de base. Ces catégories sont utilisées pour construire un modèle HMM qui calcule les probabilités de transition d'une catégorie à l'autre. Le modèle appris est ensuite utilisé pour éliminer les segmentations peu probables ou impossibles. Par exemple, la segmentation `s+well` ne serait pas suggérée, car la séquence `suffixe+stem` serait considérée impossible[6]. Les résultats du Morpho Challenge (Kurimo et al., 2009) semblent indiquer que ce type de traitement donne de meilleurs résultats que le modèle MDL seul sur plusieurs[7] langues.

[5] Le maximum a posteriori est une technique basée sur le théorème de Bayes permettant d'estimer une distribution à partir de données empiriques.

[6] *Morfessor CatMAP* utilise en plus des statistiques apprises, quelques règles pour assurer l'ordre attendu des morphèmes.

[7] Seulement l'anglais semble faire exception.

CHAPITRE 4

ANALOGIE FORMELLE

Une *analogie (proportionnelle)* est une relation entre 4 éléments notés [x : y = z : t] ce lisant "x est à y ce que z est à t". L'*analogie formelle*[1] (section 2.7) est un cas spécifique d'analogie proportionnelle où la relation analogique entre les 4 éléments est graphémique (*p. ex.* [cordially : cordial = lovely : love]). Depuis quelques années, certains chercheurs se sont intéressés aux utilisations possibles de l'analogie formelle et ont démontré que l'apprentissage analogique peut être appliqué à plusieurs problèmes du traitement automatique des langues tels que la traduction automatique et l'AMNS (section 3.2.2). L'intérêt premier d'appliquer l'analogie à la tâche d'AMNS provient du grand nombre de phénomènes morphologiques qu'elle peut capturer. En effet, comme on peut observer à la figure 4.1, seulement 2 des 8 phénomènes répertoriés ne peuvent être reconnus par l'analogie, soit la supplétion et le redoublement.

Phénomène	Analogie	Langue
Préfixation	[bicorne : tricorne = bicycle : tricycle]	français
Suffixation	[capital : capitaliste = commun : communiste]	français
Infixation	[asna : askana = arakbus : arakkabus])	ulwa
Mutation du *stem*	[ceann : cinn = fear : fir])	celte
Composition	[talg : talglicht = tee : teelicht]	allemand
Patron-racine	[nafs : nufuus = bank : bunuuk]	arabe
Supplétion	[bad : worse = good : ?]	anglais
Redoublement	[ulo : ululo = mula : ?]	ilokano

Figure 4.1 – Exemples d'analogies pour les principaux phénomènes morphologiques.

L'utilisation de l'apprentissage analogique par un système informatique requiert l'élaboration d'un algorithme capable d'identifier les éléments en relation analogique. Plusieurs solutions ont été suggérées (Pirrelli, 1993; Yvon et al., 2004; Goldsmith, 2007), mais nous nous intéresserons ici seulement à la définition proposée par Yvon et al. (2004) décrite à la section 4.1. Récemment, Langlais (2009) a conçu un système d'AMNS, décrit

[1]Dans le but d'alléger le texte, nous utiliserons tout simplement le terme d'analogie au lieu d'analogie formelle.

à la section 4.2, entièrement basé sur cette définition de l'analogie.

4.1 Identification des analogies

Dans le cadre de ce mémoire, nous avons réutilisé le moteur analogique conçu par Langlais (2009) basé sur la définition d'Yvon et al. (2004) qui définit l'analogie en terme de *factorisation*. Soit $\mathtt{x}$ étant une chaîne de caractères sur un alphabet Σ, une *factorisation* de $\mathtt{x}$, noté $f_{\mathtt{x}}$, est une séquence de n *facteurs* $f_{\mathtt{x}} = (f_{\mathtt{x}}^1, \ldots, f_{\mathtt{x}}^n)$, de sorte que $\mathtt{x} = f_{\mathtt{x}}^1 \odot \ldots \odot f_{\mathtt{x}}^n$, où $\odot$ dénote l'opération de concaténation. D'après Yvon et al. (2004) nous définissons l'analogie formelle comme étant :

Definition 1 $\forall(\mathtt{x},\mathtt{y},\mathtt{z},\mathtt{t}) \in \Sigma^{\star^4}$, $[\mathtt{x} : \mathtt{y} = \mathtt{z} : \mathtt{t}]$ ***ssi*** *il existe une factorisation* $(f_{\mathtt{x}}, f_{\mathtt{y}}, f_{\mathtt{z}}, f_{\mathtt{t}}) \in (\Sigma^{\star^d})^4$ *de* $(\mathtt{x},\mathtt{y},\mathtt{z},\mathtt{t})$ *de sorte que,* $\forall i \in [1,d]$, $(f_{\mathtt{y}}^i, f_{\mathtt{z}}^i) \in \left\{(f_{\mathtt{x}}^i, f_{\mathtt{t}}^i), (f_{\mathtt{t}}^i, f_{\mathtt{x}}^i)\right\}$. *La plus petite valeur de d pour laquelle cette définition s'applique est le degré de l'analogie.*

Selon cette définition, les 4 mots `cordially`, `cordial`, `lovely` et `love` sont en relation analogique, noté [`cordially` : `cordial` = `lovely` : `love`], car il existe une factorisation de ces 4 chaînes qui mettent en jeu des *alternances*. La figure 4.2 illustre deux factorisations de ces chaînes. Celle de gauche met en œuvre 4 facteurs par factorisation, tandis que celle de droite n'en utilise que 2, illustrant les alternances `love`/`cordial` et `ly`/ϵ capturées par cette analogie. Ces alternances sont appelées *cofacteurs* (Langlais et Patry, 2007). Il est important de noter que ces cofacteurs ne sont pas dirigés, c'est-à-dire que la paire `ly`/ϵ est égale à la paire ϵ/`ly`. Le cofacteur `ly`/ϵ capture en anglais le phénomène productif où un adverbe est dérivé d'un substantif ou d'un adjectif en lui ajoutant le suffixe `ly`. Il permettrait par exemple de rendre compte de la forme `flabbergastedly` à partir de la forme `flabbergasted` (ou l'inverse) même si cette dernière n'est pas présente

$\mathrm{f}_{\texttt{cordially}}$	$\equiv$	`cordia`	`l`	`l`	`y`	$\mathrm{f}_{\texttt{cordially}}$	$\equiv$	`cordial`	`ly`
$\mathrm{f}_{\texttt{cordial}}$	$\equiv$	`cordia`	ϵ	`l`	ϵ	$\mathrm{f}_{\texttt{cordial}}$	$\equiv$	`cordial`	ϵ
$\mathrm{f}_{\texttt{lovely}}$	$\equiv$	`love`	`l`	ϵ	`y`	$\mathrm{f}_{\texttt{lovely}}$	$\equiv$	`love`	`ly`
$\mathrm{f}_{\texttt{love}}$	$\equiv$	`love`	ϵ	ϵ	ϵ	$\mathrm{f}_{\texttt{love}}$	$\equiv$	`love`	ϵ

Figure 4.2 – Deux factorisations de l'analogie [`cordially` : `cordial` = `lovely` : `love`]. Le nombre de facteurs de la plus petite factorisation est appelé le degré de l'analogie ; soit 2 dans cet exemple.

dans le lexique. Il permet malheureusement de lier à tort des formes comme *flyable* et *fable*.

4.2 Système purement analogique

Langlais (2009) a démontré qu'il est possible d'analyser morphologiquement les mots d'une langue en utilisant seulement le concept d'analogie. Ce système, particulièrement remarquable de par sa simplicité, postule que les factorisations résultantes de la méthode d'identification des analogies d'Yvon et al. (2004) (section 4.1) correspondent fréquemment aux frontières de morphèmes des mots en relation analogique. Tel qu'illustré à la figure 4.3, pour chaque mot à analyser, le système conserve les fréquences d'occurrence de chaque factorisation qui a permis d'établir au moins une des relations analogiques l'impliquant. La décomposition en facteurs la plus fréquente est identifiée par le système comme la segmentation en morphèmes la plus vraisemblable. Par exemple, l'analyse du mot anglais *abolishing* suggéré par le système sera *abolish + ing* car cette factorisation entre en jeu dans 12 des 21 analogies impliquant ce mot. Bien que ceci puisse sembler très simple, cet algorithme requiert l'identification de toutes les analogies du lexique $\mathcal{L}$. Ceci rend cette méthode difficilement applicable aux lexiques les plus volumineux dû au temps de calcul élevé requis pour traiter les $|\mathcal{L}|^4$ analogies potentielles, et ce, même en utilisant la stratégie du *tree-count* proposée par Langlais et Yvon (2008).

abolishing (ANG.)		*abberufen* (ALL.)		*abdallardan* (TUR.)	
abolish ing	12	*ab berufen*	12	*abd allardan*	17
ab olishing	4	*a b b erufen*	12	*abdallar dan*	10
abol ishing	2	*abberufe n*	10	*a b da llardan*	9
a bo lishing	1	*a b beruf en*	6	*ab dallardan*	6
abolis hing	1	*abb erufen*	5	*abdallar d an*	5
abolish in g	1	*abberuf en*	5	*a b da l lardan*	4
		ab beruf en	2	*ab dallar dan*	2
		abbe rufe n	1	*abda llardan*	2

Figure 4.3 – Factorisations induites par analogie pour certains mots de différentes langues. La fréquence de la factorisation est indiquée à droite de la factorisation.

CHAPITRE 5

LE SYSTÈME *MORANAPHO*

Notre système produit pour chacun des mots d'un lexique $\mathcal{L}$ la liste des morphèmes qui le composent (section 1.1). La seule information requise par le système est l'ensemble des mots de $\mathcal{L}$ sans aucune annotation particulière. Brièvement, notre système applique un algorithme glouton de partitionnement (*clustering*) qui regroupe les mots du lexique partageant la même racine en exploitant un graphe de mots relié par règles morphologiques. Dans notre approche, ces règles morphologiques sont obtenues par analogie calculée parmi les mots de $\mathcal{L}$. Les sections suivantes décrivent en détail les différentes composantes du système.

5.1 Règle morphologique

Le coeur de notre système est un ensemble de règles de réécriture notées $\langle \alpha \rightarrow \beta \rangle$, où α et β sont des séquences de symboles de l'alphabet de la langue étudiée étendus du symbole $\star$ (*p. ex.* $\langle \star\texttt{ly} \rightarrow \star\epsilon \rangle$[1]). Ce dernier représente n'importe qu'elle combinaison non vide de symboles de l'alphabet. Nous notons l'application d'une règle $\mathcal{R}$ à un mot $\texttt{x}$ $\mathcal{R}(\texttt{x})$. Par exemple, l'application de la règle $\langle \star\texttt{ly} \rightarrow \star\epsilon \rangle$ à l'adverbe `literally` résulterait en sa forme nominale `literal` et serait notée $\langle \star\texttt{ly} \rightarrow \star\epsilon \rangle(\texttt{literally}) = \texttt{literal}$. Par la même logique, $[\mathcal{R}_1, \ldots, \mathcal{R}_n](\texttt{x})$ dénote la forme[2] résultante de l'application de n règles : $\mathcal{R}_n(\ldots \mathcal{R}_2(\mathcal{R}_1(\texttt{x}))\ldots)$. Dans le cas où la règle ne s'applique pas, le mot ne subit aucune transformation et donc $\mathcal{R}(\texttt{x}) = \texttt{x}$.

Il existe certaines contraintes que les règles de réécriture doivent respecter pour être considérées comme valides par notre système. Premièrement, nous imposons que $|\alpha| \geq |\beta|$[3], 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 de α, nous permettant ainsi de différencier facilement la préfixation, la suffixation et l'infixation. Ceci implique donc que les transformations capturées par une règle ne

[1] ϵ indique une chaîne vide.

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

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

peuvent qu'être contiguës. Par conséquent, certaines relations morphologiques pourront difficilement être représentées par une seule règle (*p. ex.* $\langle$*suggère* $\rightarrow$ *suggérer*$\rangle$). Ceci cause peu de problèmes, et est même possiblement avantageux, pour les phénomènes morphologiques les plus courants. Toutefois, les morphes flexionnels et dérivationnels disjoints sont tout de même fréquents dans certaines langues, en particulier celles de type templatique tel que l'arabe. Pour ces langues, notre format de règle ne permet pas de faire le lien entre certaines formes reliées morphologiquement, et ce, même si la relation a été identifiée par nos analogies. Par exemple, aucune règle valide ne peut faire le lien entre les formes arabes *KaaTiB* et *KuTaaB*, et ce, même si la relation est "connue" du système grâce aux analogies identifiées (*p. ex.* [*KaaTiB* : *KuTaaB* = *QaaRi'* : *QuRaa'*]). Pour capturer ce morphe, 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 2$^{\text{e}}$ contrainte. Il serait aussi possible d'identifier la relation par l'application successive des règles $\langle \star u \star \rightarrow \star aa \star \rangle$ et $\langle \star aa \star \rightarrow \star i \star \rangle$ mais malheureusement $\langle \star u \star \rightarrow \star aa \star \rangle$ ne respecte pas la 1^{re} contrainte. D'autres formats de règle 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 leurs usages n'introduiraient pas trop de bruit dans nos analyses.

Nos règles de réécriture sont obtenues par un processus basé sur l'analogie que nous détaillons à la section 5.1.2. Cependant, comme le calcul des analogies est très coûteux en temps, l'emploi d'une autre technique serait favorable si elle ne dégrade pas de façon majeure la qualité des résultats. Dans le but de nous assurer que l'usage de l'analogie est justifiable, nous l'avons comparé à une autre technique d'extraction détaillée à la section 5.1.1. Finalement, quelque soit la façon utilisée pour extraire les règles, une bonne partie d'entre elles ne seront tout simplement pas valides (*p. ex.* $\langle ab\star \rightarrow p\star \rangle$ qui lie $\langle$*about* $\rightarrow$ *pout*$\rangle$ et $\langle$*abolishing* $\rightarrow$ *polishing*$\rangle$). Il est donc important de pouvoir établir quelles sont les règles pertinentes de la langue analysée. Nous décrivons à la section 5.1.3 les différentes métriques de la qualité d'une règle que nous avons testées. Toutefois, quel que soit la méthode d'extraction ou la métrique choisie, le grand nombre de règles rencontrées nous oblige à appliquer un filtre éliminant les règles les moins vraisemblables pour des considérations de temps de calcul. Toutes les règles ayant un poids inférieur à ρ (section 7.2.1) sont donc enlevées en espérant qu'elles ne jouent pas un rôle important dans la morphologie de la langue étudiée.

5.1.1 Extraction de règles par distance d'édition

Afin d'évaluer les bénéfices obtenus par l'acquisition des règles de réécriture basées sur l'analogie, nous avons aussi tenté d'acquérir les règles en remplaçant l'analogie par un alignement par distance d'édition. Notre processus est inspiré par (Bernhard, 2010) et consiste à obtenir pour chaque mot de taille supérieure à la moyenne, les 20 mots du lexique les plus similaires à celui-ci selon la distance d'édition. Pour chacune de ces paires, nous calculons l'alignement par distance d'édition optimale[4] et les sections non alignées sont transformées en règle de réécriture. Par exemple, l'alignement entre *unattractive* et *attraction* de la figure 5.1 génère les règles $\langle un\star \rightarrow \epsilon\star \rangle$ et $\langle \star ve \rightarrow \star on \rangle$. Cette façon de procéder est beaucoup plus rapide qu'identifier les analogies parmi tous les mots de $\mathcal{L}$.

u	*n*	**a**	**t**	**t**	**r**	**a**	**c**	**t**	**i**	*v*	*e*
ϵ	ϵ	**a**	**t**	**t**	**r**	**a**	**c**	**t**	**i**	*o*	*n*

Figure 5.1 – Alignement obtenu par distance d'édition entre les mots *unattractive* et *attraction*. La section alignée est en gras et les sections non alignées sont en italique.

5.1.2 Extraction de règles par analogie

Tout comme Langlais (2009) (section 4.2), la prémisse de notre système est que les mots en relation analogique sont aussi fréquemment liés morphologiquement. Cependant, dès qu'un lexique dépasse les quelques centaines de milliers de mots, il devient déraisonnable de tenter de calculer toutes les analogies comme le fait son système purement analogique. Nous suggérons donc une technique permettant de généraliser l'information capturée par l'analogie de sorte que le système soit capable d'analyser morphologiquement tous les mots d'un lexique même si seulement un sous-ensemble de toutes les analogies a été calculé. Cette technique se base sur les factorisations obtenues de la technique d'identification des analogies proposée par d'Yvon et al. (2004) (section 4.1). Comme illustrée par l'algorithme 5.1, une règle de réécriture est tout simplement un cofacteur contextualisé. Une règle $\langle \alpha \rightarrow \beta \rangle$ est obtenue du cofacteur α/β auquel nous ajoutons le symbole $\star$ selon la position relative du cofacteur original par rapport aux

[4] Il peut en exister plusieurs. Dans ce cas, elles sont toutes conservées.

autres cofacteurs de l'analogie.

Algorithme 5.1 Algorithme d'extraction de règles par analogie.

1. Pour toutes les analogies a de l'ensemble des analogies identifiées $\mathcal{A}$.
   (a) Pour i est entre 1 et degré(a).
      i. Extraire les facteurs α et β du cofacteur à la position i de l'analogie a.
      ii. Si $i > 1$.
         A. $\beta \leftarrow \star \odot \beta$.
         B. $\alpha \leftarrow \star \odot \alpha$.
      iii. Si $i <$ degré(a).
         A. $\beta \leftarrow \beta \odot \star$.
         B. $\alpha \leftarrow \alpha \odot \star$.
      iv. Si $(|\alpha| > |\beta|) \vee (|\alpha| = |\beta| \wedge \alpha > \beta)$
         A. Ajouter $\langle \alpha \rightarrow \beta \rangle$ à l'ensemble de règles $\mathcal{R}$
      v. Sinon
         A. Ajouter $\langle \beta \rightarrow \alpha \rangle$ à l'ensemble de règles $\mathcal{R}$

Dans notre exemple de l'analogie [*cordially* : *cordial* = *lovely* : *love*] (figure 4.2), nous sommes en présence d'une analogie impliquant les 2 cofacteurs *love*/*cordial* et ϵ/*ly*. Les deux règles obtenues à partir de ces cofacteurs seront donc $\langle$*love*$\star \rightarrow$ *cordial*$\star\rangle$ et $\langle \star$*ly* $\rightarrow \star\epsilon\rangle$. Une telle décomposition de l'analogie n'est pas sans rappeler le concept d'analyse de l'analogie introduit par Saussure (1968) (section 2.7). Un des avantages de ce procédé relativement à la technique par distance d'édition de la section 5.1.1 est que le contexte d'application de la règle est capturé sans aucun effort supplémentaire. Nous illustrerons ceci par l'analogie de la figure 5.2 qui peut être factorisée de multiples façons. De cette analogie, les règles $\langle \star$*ive* $\rightarrow \star\epsilon\rangle$ et $\langle \star$*tive* $\rightarrow \star$*t*$\rangle$ seront obtenues. La première capturant la frontière du morphe *ive* alors que la seconde permet d'identifier que le suffixe *ive* est habituellement précédé d'un *t*[5]. Cette information supplémentaire, si elle est bien utilisée, permettra d'établir la relation entre *secret* et *secretive* tout en évitant de faire l'erreur de lier *end* et *endive*. Nous appelons la relation entre les règles $\langle \star$*ive* $\rightarrow \star\epsilon\rangle$ et $\langle \star$*tive* $\rightarrow \star$*t*$\rangle$ une *équivalence*. Deux règles $\mathcal{R}_1$ et $\mathcal{R}_2$ sont équivalentes dans le contexte du mot *x* si $\mathcal{R}_1$(*x*) = $\mathcal{R}_2$(*x*).

[5]Le morphe *ive* est toujours précédé d'un *t* ou d'un *s*.

f_{addict}	$\equiv$	addict	ϵ	f_{addict}	$\equiv$	addic	t
$f_{addictive}$	$\equiv$	addict	ive	$f_{addictive}$	$\equiv$	addic	tive
$f_{distinct}$	$\equiv$	distinct	ϵ	$f_{distinct}$	$\equiv$	distinc	t
$f_{distinctive}$	$\equiv$	distinct	ive	$f_{distinctive}$	$\equiv$	distinc	tive

Figure 5.2 – Deux factorisations de degré 2 de l'analogie anglaise [addict : addictive = distinct : distinctive].

5.1.3 Sélection des règles

La théorie de l'analogie d'Anderson (1992) nous apprend que les mots en relation analogique doivent être non seulement liés graphémiquement, mais aussi sémantiquement (section 2.7). Malheureusement, notre moteur analogique, n'ayant aucune information permettant d'inférer le sens des mots, ne se base que sur la forme pour établir un lien analogique. Ceci implique donc que plusieurs analogies fortuites (*p. ex.* [pear : spear = care : scare]) se seront glissées dans notre ensemble d'analogies. Les règles extraites de ces analogies seront fort probablement non pertinentes pour la langue étudiée (*p. ex.* $\langle s\star \rightarrow \epsilon\star \rangle$). Selon Anderson (1992), il y a un troisième facteur à prendre en compte pour juger de la validité d'une analogie : la régularité de la relation. Intuitivement, cela porte à croire que les relations (règles) régulières seront plus pertinentes que les relations isolées. Nous pouvons estimer la régularité d'une règle par le nombre d'analogies desquelles la règle a été extraite. Cependant, la fréquence crée un biais envers les règles composées de facteurs courts, car elles ont une plus grande probabilité de se retrouver par hasard dans une analogie (*p. ex.* [stop : top = stick : tick], [tar : star = word : sword]). Par exemple, dans un de nos lexiques anglais, la règle $\langle anti\text{-}\star \rightarrow \epsilon\star \rangle$ est observée 2 472 fois, comparé à la règle fortuite $\langle ka\star \rightarrow \epsilon\star \rangle$ qui est vue 13 839 fois.

Pour résoudre ce problème, nous nous inspirons encore une fois du domaine de la linguistique et plus particulièrement du concept de *productivité* (section 2.6). En résumé, ce principe tente d'évaluer le degré d'utilisation d'un processus particulier dans la création d'un nouveau mot. Par exemple, en français, l'ajout d'un s au pluriel (*p. ex.* terres) est très productif, alors que l'ajout de aux (*p. ex.* chevaux) l'est probablement moins. Toutefois, le concept de productivité est plutôt abstrait et relève plus de l'intuition du linguiste que d'un processus bien défini[6]. L'utilisation de ce concept en informatique s'avère donc

[6]Toutefois, certains linguistes ont tenté de formaliser le concept de productivité (section 2.6).

plutôt difficile. Heureusement, il existe en intelligence artificielle une mesure nommée la *productivité de la règle*[7] qui correspond au ratio entre le nombre d'applications valides d'une règle sur le nombre d'applications possibles de cette même règle. Outre le nom, ces deux concepts ont en commun qu'ils tentent tous deux d'évaluer le taux d'utilisation d'un processus. Il est cependant important de noter que malgré cette similitude, ce sont tout de même des concepts très différents. Toutefois, nous croyons que la productivité d'une règle sera élevée si cette règle correspond à un processus productif linguistiquement. 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 (équation 5.1). Les règles $\langle \texttt{anti-}\star \rightarrow \epsilon\star \rangle$ et $\langle \texttt{ka}\star \rightarrow \epsilon\star \rangle$ de notre exemple précédent ont des productivités respectives de 0,9490 et 0,2472.

$$prod(\mathcal{R}) = \frac{|\{x \in \mathcal{L} : \mathcal{R}(x) \in \mathcal{L} \wedge \mathcal{R}(x) \neq x\}|}{|\{x \in \mathcal{L} : \mathcal{R}(x) \neq x\}|} \tag{5.1}$$

Une des faiblesses de la productivité est que l'information sur la fréquence est perdue. De plus, les règles peu fréquentes ont tendance à être très productives même lorsqu'elles ne sont pas pertinentes et doivent donc être éliminées. Nous avons donc testé une métrique supplémentaire (équation 5.2) qui combine le logarithme de la fréquence à la productivité de la règle[8].

$$\text{f-prod}(\mathcal{R}) = \text{prod}(\mathcal{R}) \times \frac{\log |\mathcal{R}|}{\log \sum_{\mathcal{R} \in \mathcal{S}} |\mathcal{R}|} \tag{5.2}$$

La table 5.I contient quelques exemples de règles extraites d'un lexique anglais utilisé dans nos expériences ainsi que leurs valeurs pour les trois métriques décrites précédemment.

5.2 Regroupement hiérarchique des mots du lexique

L'ensemble des règles collectées à l'étape précédente constitue la matière première de notre système. À l'aide de ces règles, nous construisons un arbre tel que celui de la figure 5.3. Les noeuds de cet arbre sont les mots du lexique. Un arc entre deux noeuds n_a et n_b, noté $n_a \rightarrow n_b$, est étiqueté d'une séquence de règles qui lorsque appliquées au mot n_a résultent en le mot n_b. Dans notre exemple, toutes les séquences ne contiennent

[7] Nous utiliserons par la suite tout simplement le terme de productivité.

[8] Le dénominateur assure que le résultat sera une valeur entre 0 et 1. Notre système impose cette contrainte.

Tableau 5.I – Fréquence, productivité et la fréquence combiné à la productivité de règles extraites d'un lexique anglais.

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

qu'une seule règle. Une fois construit, l'arbre est partitionné en éliminant les relations jugées peu probables. Le résultat de cette opération est une forêt d'arbres où tous les mots d'un même arbre ont été vraisemblablement dérivés de la même racine. Nous appelons un arbre de cette forêt un *Arbre de Dérivation de Mot* ou *ADM*.

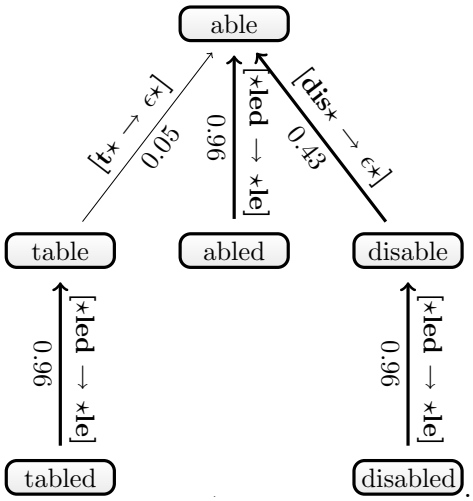

Figure 5.3 – Exemple d'arbre hiérarchique avant le partitionnement. L'arc fin sera vraisemblablement supprimé lors du partitionnement résultant en 2 ADMs.

L'arbre est construit par l'algorithme 5.2 inspiré de l'algorithme glouton de partitionnement hièrarchique par lien simple (*single-linkage*) (Newman, 2004). Contrairement à l'algorithme standard, nous n'avons pas besoin d'identifier le lien le plus fort du graphe à chaque itération, mais seulement le lien le plus fort impliquant un mot donné. Ceci a pour avantage que le graphe n'a pas à être construit en entier d'un seul coup.

Algorithme 5.2 Algorithme de construction de l'arbre hiérarchique.

1. Pour tous les mots m du lexique $\mathcal{L}$.
   (a) Construire le graphe $\mathcal{G}$ des mots pouvant être atteints en appliquant un nombre strictement positif de règles au mot m (figure 5.4).
   (b) Trouver le mot b de $\mathcal{L}$ qui maximise un score de similarité $\mathcal{S}$ (équation 5.3) selon $\mathcal{G}$
   (c) Ajouter l'arc $m \rightarrow b$ à l'arbre

En effet, comme le démontre l'étape 1a de l'algorithme 5.2, seulement le graphe du voisinage du mot à intégrer à l'arbre est nécessaire à l'algorithme. La figure 5.4 illustre le graphe $\mathcal{G}$ construit à cette étape pour le mot ***disabled***. Notez que les noeuds du graphe ne sont pas limités aux mots du lexique. Ceci nous permet de relier notamment le mot ***disable*** à ***able***, et ce, même si le mot ***disable*** était absent du lexique. Un autre point important illustré par cette figure est qu'il peut y avoir plusieurs arcs entre deux noeuds.

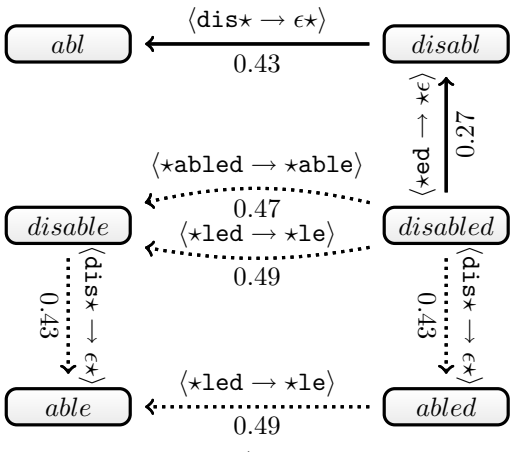

Figure 5.4 – Graphe des mots atteignables depuis le mot anglais ***disabled***. Les arcs en pointillés sont ceux considérés lors du calcul du score entre ***disabled*** et ***able***.

Il est important pour la compréhension de l'algorithme de bien faire la différence entre l'arbre hiérarchique et le graphe. L'arbre capture les relations de dérivations et de flexions entre les mots du lexique alors que le graphe n'est qu'un outil nécessaire au calcul d'un score permettant l'identification de ces relations (étape 1b de l'algorithme 5.2). Le score (équation 5.3) d'un lien potentiel entre deux mots m et w est calculé en sommant

le poids de chacun des chemins du graphe partant de m jusqu'à w.

$$score(m, w) = \sum_{[\mathcal{R}_1,\ldots,\mathcal{R}_n](m)\equiv w} poids([\mathcal{R}_1, \ldots, \mathcal{R}_n]) \tag{5.3}$$

Le poids d'un chemin entre m et w, représenté par une séquence de k règles $[\mathcal{R}_1, \ldots, \mathcal{R}_k]$, est calculé par le produit du poids attribué aux k règles du chemin selon une métrique donnée (équation 5.4).

$$poids([\mathcal{R}_1, \ldots, \mathcal{R}_k]) = \prod_{i=1}^{k} métrique(\mathcal{R}_i) \tag{5.4}$$

Un arc $m \rightarrow b$ entre le mot m et le mot lexique b qui maximise l'équation 5.3 ($b \equiv \text{argmax}_{w\in\mathcal{L}}\, score(m, w)$) est ajouté à l'ADM. Cet arc est étiqueté par la séquence de règles $[\mathcal{R}_1, \ldots, \mathcal{R}_k]$ du chemin entre m et b ayant le plus haut score. Dans notre exemple, la forme `disable` totalisant un score de $0,96$ est celle qui est sélectionnée et la séquence $[\langle \star\texttt{led} \rightarrow \star\texttt{le}\rangle]$ étiquette dans l'ADM l'arc `disabled` $\rightarrow$ `disable`.

Une fois que l'arbre est construit, il est partitionné en éliminant les relations dont les poids sont inférieurs à un certain seuil τ (section 7.2.1). Par exemple, le partitionnement de l'arbre de la figure 5.3[9] résulterait en la suppression du lien entre `table` et `able` et nous donnerait donc deux ADMs dont les racines respectives seraient `able` et `table`.

5.3 Décomposition morphologique des mots du lexique

Jusqu'à maintenant, nous avons identifié la racine de chacun des mots du lexique. Toutefois, nous désirons obtenir une analyse complète de la composition en morphème de ces mots (section 1.1). La liste des morphèmes d'un mot du lexique est obtenue en se guidant sur les règles de réécriture entre le mot analysé et la racine de son ADM. La base de notre algorithme d'extraction des morphèmes est plutôt simple. Chaque noeud de l'ADM contient la segmentation en morphèmes (potentiels) du mot qui lui est associé. Dans le cas de la racine, le seul morphème est le mot lui-même. Pour tout autre noeud n, l'ensemble des morphèmes est constitué en regroupant les morphèmes du noeud parent p et de ceux des règles étiquetant l'arc $n \rightarrow p$. Par exemple, dans le cas d'un ADM de

[9] $\tau = 0.25$

deux mots qui contient l'arc *disarmed* → *arm* étiqueté par [⟨*dis*⋆ → ϵ⋆⟩,⟨⋆*ed* → ⋆ϵ⟩], les morphèmes de *disarmed* seraient [*arm*,*dis*,*ed*] ; *dis* et *ed* provenant de la partie gauche de leur règles respectives.

Malheureusement, ce procédé s'est avéré impraticable tel quel et a nécessité quelques ajustements pour 2 raisons. Premièrement, la règle qui étiquette un arc de l'ADM correspond rarement à un morphe valide dû au contexte d'application contenu dans nos règles. Ceci est illustré par le lien *tabled* → *table* de la figure 5.3 qui est étiqueté par la règle ⟨⋆*led* → ⋆*le*⟩ au lieu de ⟨⋆*ed* → ⋆*e*⟩. Deuxièmement, dans certains cas, un mot n'est pas directement attaché à sa base, mais à un autre mot dérivé de cette même base. Par conséquent, l'analyse morphologique du mot contiendra des morphèmes superflus provenant des mots qui s'interposent entre lui et sa base la plus proche. Par exemple, le mot *disabling* pourrait être lié à *disabled* au lieu de *disable* par la règle ⟨⋆*ing* → ⋆*ed*⟩. Dans ce cas, il faudrait enlever de la liste des morphèmes de *disabling* résultante [*dis*,*able*,*ed*,*ing*] le morphème superflu *ed* provenant de *disabled*. Heureusement, la partie droite de la règle de réécriture (*p. ex.* ⟨⋆*ing* → ⋆*ed*⟩) permet d'identifier les morphèmes en trop provenant d'une relation intermédiaire. Bien que, encore une fois, le contexte d'application de la règle complique les choses. Ainsi, si dans notre exemple précédent le noeud *disabled* avait contenu le morphème *led* au lieu de *ed*, se baser sur la partie droite de la règle ⟨⋆*ing* → ⋆*ed*⟩ ne suffirait pas à identifier le morphème superflus *led*.

Pour toutes ces raisons, nous avons raffiné l'implémentation de notre algorithme de base en substituant chacune des règles qui étiquettent un arc de l'ADM à une règle équivalente (section 5.1.2) choisie selon certains critères, tel qu'illustré par l'algorithme 5.3[10]. La règle que nous choisissons est la règle ayant la fréquence la plus élevée et qui est cohérente avec le reste de l'ADM. Par cohérente, nous voulons dire qu'il n'y a pas de conflit entre les frontières de morphème tacite à la règle. Par exemple, dans le cas du mot *disabled*, la règle ⟨⋆*ing* → ⋆*ed*⟩ et ⟨⋆*ed* → ⋆*e*⟩ sont cohérentes car la frontière de morphème implicite au deux règles se situe après le *l* (*disabl+ed*). Par contre, la règle ⟨⋆*led* → ⋆*le*⟩ n'est pas cohérente avec les deux règles précédentes car la frontière de morphème implicite se situe après le *b* (*disab+led*).

Bien que ces améliorations à l'algorithme de base nous aient permis de résoudre les

[10]Dans le but de simplifier, nous omettons le cas où plus d'une règle étiquette un lien.

Algorithme 5.3 Algorithme de décomposition en morphème d'un mot m d'un ADM.

1. Si le mot à analyser m est la racine de l'ADM.
   (a) Retourner $[m]$ et terminer l'exécution.
2. Trouver les morphèmes $\mathcal{P}$ du père p de m.
3. Trouver toutes les règles équivalentes $\mathcal{E}$ à la règle étiquetant $m \rightarrow p$.
4. Si $\exists\ \langle \alpha \rightarrow \beta \rangle \in \mathcal{E}$, $\beta \in \mathcal{P}$
   (a) Retourner $\mathcal{P} \cup [\alpha] - [\beta]$
5. Sinon
   (a) Trouver la règle $\langle \alpha \rightarrow \beta \rangle$ ayant la fréquence la plus élevée dans $\mathcal{E}$.
   (b) Retourner $\mathcal{P} \cup [\alpha]$

problèmes que nous avons identifiés précédement, notre système souffre tout de même d'un manque de cohérence flagrant dans le choix des étiquettes de morphème. En effet, comme la sélection de l'étiquette identifiant un morphème est un choix local à un ADM, il est possible qu'une même règle produise deux, voir plusieurs, étiquettes différentes d'un ADM à l'autre. Par exemple, la règle $\langle \star\texttt{ed} \rightarrow \star\texttt{e} \rangle$ pourrait produire l'analyse [*cable*,*d*] pour la forme *cabled* alors que cette même règle résulterait en l'analyse [*dance*,*ed*] pour la forme *danced*. Dans le premier cas, l'étiquette provient de la règle équivalente la plus fréquente $\langle \star\texttt{d} \rightarrow \star\epsilon \rangle$ alors que dans le deuxième cas, le choix a été fait pour préserver la cohérence avec la règle $\langle \star\texttt{ing} \rightarrow \star\texttt{ed} \rangle$ liant *dancing* à *danced*. Ceci pourrait être résolu simplement en attribuant une étiquette à chaque groupe de règles équivalentes qui serait systématiquement utilisé pour tout arc étiqueté par une règle de ce groupe. Cependant, il serait plus intéressant de fusionner les arbres similaires assurant ainsi une certaine cohésion des étiquettes tout en intégrant le concept de paradigme à notre système. Ceci ne serait pas sans rappeler (Goldsmith, 2007) qui unifiait par MDL des automates à états finis capturant les affixations possibles des mots d'un même paradigme.

CHAPITRE 6

ÉVALUATION

Afin d'évaluer le rendement de notre système, nous utilisons les scripts[1] fournis par les organisateurs du Morpho Challenge. Ces scripts mesurent la qualité des analyses produites en terme de précision, rappel et f-mesure en les comparant à un lexique morphologique produit manuellement. En bref, la méthode d'évaluation utilisée repose sur l'intuition que deux mots choisis au hasard devraient partager le même nombre de morphèmes dans les analyses suggérées par le système que dans la référence. La section 6.2 contient plus de détail sur le processus d'évaluation et sur les raisons qui nous ont poussés à choisir cette métrique plutôt qu'une autre. Les jeux de données et les références utilisées pour nos expériences sont décrits à la section 6.1.

6.1 Données

Nos expériences ont été effectuées sur 2 jeux de données distincts. Le premier, que nous appellerons *DonPropre*, est composé de 3 lexiques de langue germanique soit l'anglais (72 628 mots), l'allemand (311 000 mots) et le néerlandais (321 926 mots). Ces lexiques contiennent des noms communs, des adjectifs et des verbes sans aucune erreur typographique. L'attrait principal d'utiliser ces langues est que le niveau de complexité morphologique varie considérablement d'une langue à l'autre tout en partageant certaines caractéristiques communes. Ceci facilite l'analyse des résultats en permettant de mieux délimiter les caractéristiques de la langue influençant la qualité des résultats. Comme deuxième ensemble de données d'entrée, nous avons utilisé les données officielles du Morpho Challenge 2009[2] qui couvrent un plus grand éventail de langues. Ces 5 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). Nous utiliserons l'appellation *DonBruitée* pour identifier ce jeu de données.

[1] http ://www.cis.hut.fi/morphochallenge2009/evaluation.shtml
[2] http ://www.cis.hut.fi/morphochallenge2009/datasets.shtml

Les évaluations internes ont été effectuées sur une référence extraite de Celex (Baayen et al., 1995) nommée *RefCelex*. La table 6.I donne les principales caractéristiques de cette référence. Cette référence ne contient que 3 des langues étudiées. La deuxième référence est celle utilisée par les organisateurs de Morpho Challenge pour évaluer les analyses des systèmes en compétition. Seuls les organisateurs ont accès à cette référence que nous nommerons par la suite *RefMorpho*.

Tableau 6.I – Principales caractéristiques de la référence *RefCelex*. De gauche à droite : la taille en nombre de mots du lexique, le nombre de morphèmes distincts, la moyenne du nombre d'analyses possibles pour un mot, le nombre de morphèmes moyens pour une analyse, le nombre moyen de mots avec lesquels un mot partage au moins 1 morphème.

	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

6.2 Métrique d'évaluation

Pour évaluer le niveau de concordance entre les résultats obtenus par AMNS et une référence linguistique produite manuellement, nous utilisons la métrique proposée par Kurimo et al. (2007a). 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 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 des avantages de cette métrique est qu'elle est indépendante de l'étiquette attribuée au morphème, ce qui est essentiel à l'évaluation de toute approche non supervisée. De plus, elle tient compte de tous les types de procédés morphologiques (section 2.3) et peut donc être appliquée à n'importe qu'elle langue. Qui plus est, cette métrique gère aussi l'allomorphie (section 2.1), les morphèmes homonymes (section 2.1) et l'ambiguïté d'ana-

lyse[3]. Cependant, Spiegler et Monson (2010) ont tout de même critiqué cette métrique en raison de la facilité avec laquelle elle peut être abusée en ajoutant un même morphème à toutes les analyses ou en proposant plusieurs solutions alternatives. Nous avons aussi observé une très grande disparité entre les gains obtenus par l'identification de certains morphèmes. Comme on pouvait s'y attendre, l'impact de l'identification d'un morphème sur le score dépend fortement de sa fréquence dans la référence. Ce qui est surprenant est l'importance de ce phénomène. Comme le démontre la table 6.II, un système capable d'identifier seulement la terminaison *s* comme étant un morphe de l'anglais obtiendrait un pointage relativement élevé alors que l'identification du morphe *man* n'apporterait qu'un gain mineur. Ceci a pour effet de mettre l'accent sur la morphologie flexionnelle (beaucoup plus fréquente) alors que les cas de morphologie dérivationnelle ou de composition sont peu récompensés.

Tableau 6.II – Précision (Pr.), Rappel (Rp.) et F-mesure (F1) obtenus en n'identifiant qu'un seul suffixe en anglais. Fr.(m) indique la fréquence analogique en millier.

Morphe	Fr.(m)	Pr.	Rp.	F1
⋆*s*	6 940	93,43	20,47	33,44
⋆*ed*	568	95,40	1,48	2,90
⋆*man*	7	100,00	0,00	0,14

Malgré cela, nous utiliserons tout de même cette métrique pour plusieurs raisons. Premièrement, elle est bien établie dans le domaine dû à sont utilisation à Morpho Challenge depuis les 3 dernières années. Deuxièmement, les scripts d'évaluations sont librement accessibles sur le web[4]. Finalement, les autres métriques populaires sont plutôt limitées quant aux types de morphologie et aux procédés morphologiques qu'elles peuvent évaluer les rendant peu adaptées à l'évaluation d'un système se voulant multilingue. Le lecteur doutant de notre bonne volonté dans le choix de la métrique d'évaluation peut se référer aux résultats obtenus par nos analyses du Morpho Challenge 2009 sur la métrique EMMA lors de l'évaluation effectuée par Spiegler (2010) (section 7.1.1).

[3]L'ambiguïté d'analyse sert à désigner le cas où il y a plusieurs analyses morphologiques valides pour un même mot.

[4]http ://www.cis.hut.fi/morphochallenge2009/evaluation.shtml

CHAPITRE 7

RÉSULTATS ET ANALYSES

Ce chapitre contient la description, les résultats et l'analyse des nombreuses expériences impliquant le système *Moranapho*. Ces expériences ont été divisées en deux groupes selon l'objectif visé. La section 7.1 décrit les expériences réalisées dans le but d'évaluer les performances du système relativement aux autres systèmes d'AMNS. Le deuxième groupe décrit à la section 7.2, rassemble les expériences effectuées pour valider certaines de nos hypothèses.

7.1 Comparaison à l'état de l'art

Nos premières expériences ont été effectuées dans le cadre de notre participation à la compétition d'AMNS Morpho Challenge 2009[1] (section 7.1.1). Malgré un handicap causé par le départ tardif du projet combiné à des données d'apprentissage peu avantageuses pour notre type d'approche, notre système s'est tout de même bien classé. Nous avons néanmoins effectué une deuxième série de tests (section 7.1.2) dans des conditions plus favorables.

7.1.1 Morpho Challenge 2009

L'édition 2009 du Morpho Challenge a connu un record de participation avec 10 participants et 15 systèmes en compétition. À ces 15 systèmes s'ajoutent 2 références fortes et 1 naïve provenant des organisateurs. Les 2 références fortes sont les systèmes *Morfessor Baseline* et *Morfessor CatMAP* (section 3.3) alors que la référence naïve, appelée *Letters*, consiste à retourner comme liste de morphèmes la décomposition en lettres du mot analysé (*p. ex.* `fire` [`f + i + r + e`]).

Nous avons soumis aux organisateurs, en plus des analyses produites par *Moranapho*[2], celles provenant du système de Langlais (2009) (section 4.2) sous le nom de *Rali-Ana*. Les analyses de ces 2 systèmes ont été produites sur un même ensemble d'analogies obtenu

[1] La tâche et la métrique d'évaluation de cette compétition sont les mêmes que celles décrites respectivement aux chapitres 1 et 6.

[2] Appelé *Rali-Cof* à l'époque.

en exécutant notre moteur analogique pendant une semaine sur les lexiques du jeu de données *DonBruitée* fournis par les organisateurs. Nous avons ainsi recueilli un grand nombre d'analogies[3] qui malgré leur nombre, ne constitue qu'une petite partie de toutes les analogies existantes au sein de ces lexiques. Le tableau 7.I donne les résultats officiels (Kurimo et al., 2009) de l'évaluation pour nos 2 soumissions ainsi que pour 2 variantes du système *Morfessor* (section 3.3) et du système *MorphoNet* (section 3.2.1). Nous avons ajouté aux résultats un rang global mesurant la performance de chacune des approches sur l'ensemble des langues. Le rang est obtenu en attribuant à chacun des systèmes un pointage calculé en additionnant la position dans le classement pour chacune des langues. Les systèmes sont ensuite classés en ordre croissant de pointage. Seulement la variante la plus performante de chacun des systèmes est conservée[4]. L'arabe a été omis du calcul en raison du taux relativement faible de participation et aux résultats décevants obtenus par l'ensemble des systèmes. En effet, le système ayant la f-mesure la plus élevée pour cette langue est la référence naïve *Letters*. Notre opinion est que le problème provient en grande partie de la référence utilisée[5] qui contenait des morphèmes redondants faussant ainsi le calcul des résultats. Par exemple, le mot *AbEv*[6] est associé aux morphèmes [*bEv + fEl + bEv + Verb + Imperative + 2P*] dans la référence. Notez que cette liste contient 2 fois le morphème *bEv* et que le morphème *+Verb* est une généralisation du morphème *+Imperative*. Toutefois, nous croyons qu'aucun des systèmes en compétition, sauf exception de *MorphoNet*, n'est adapté au type de morphologie patron-racine que l'on retrouve en arabe et que donc, les résultats auraient été faibles même avec une référence adéquate. Pour ces mêmes raisons, nous ferons abstraction de l'arabe lors de la discussion des résultats.

Notre première observation est que *Moranapho* surpasse non seulement le système purement analogique *Rali-Ana*, mais aussi le système à base de graphe *MorphoNet* sur toutes les langues. Comme *Moranapho* et *MorphoNet* se rejoignent sur plusieurs points, ceci laisse présager que l'analogie apporte un gain réel (section 7.2.3). Nous pouvons aussi observer que nous surpassons *Morfessor Baseline* sur toutes les langues considérées sauf l'anglais. Selon notre méthode de classement, nous terminons au troisième rang de la

[3] De 11 (arabe) à 52 (turc) millions.
[4] Exception faite du système *Morfessor* où les deux variantes ont été conservées.
[5] L'arabe ne fait plus partie de la compétition Morpho Challenge en 2010.
[6] L'arabe est latinisé par translittération Buckwalter.

Tableau 7.I – F-mesure et le rang global des systèmes analogiques, des systèmes *Morfessor* et du système *MorphoNet* sur le jeu de données `DonBruitée` évalués sur la référence `RefMorpho` dans le cadre de l'atelier Morpho Challenge 2009.

	Ang.	Fin.	Tur.	All.	Arb.	Rang global /12
Moranapho	55,30	38,81	**46,40**	45,57	4,18	3
Rali-Ana	44,10	17,63	21,69	24,55	8,41	11
Morfessor Baseline	**59,84**	26,75	29,67	35,87	**12,03**	6
Morfessor CatMAP	50,50	**44,61**	45,49	**50,30**	*ND*	2
MorphoNet	55,13	33,34	41,19	41,71	9,39	5

compétition, très proche de la deuxième place occupée par *Morfessor CatMAP* que nous surclassons tout de même sur 2 des 4 langues. Le seul système obtenant systématiquement de meilleurs résultats que *Moranapho* est le compétiteur de longue date `Paramor` (section 3.1.4).

Récemment, une deuxième évaluation a été effectuée sur les analyses de la majorité des systèmes en compétition[7] en plus de certains des meilleurs systèmes ayant participé aux Morpho Challenge en 2007 ou en 2008[8] (Kurimo et Varjokallio, 2008; Kurimo et al., 2007b). Cette évaluation a été produite par Spiegler et Monson (2010) pour tester leur métrique d'évaluation `EMMA` (section 6.2). Le tableau 7.II contient une partie des résultats de cette évaluation tirée du rapport technique (Spiegler, 2010). Notez que sur cette métrique, le système *Morfessor CatMAP* domine avec une première position sur toutes les langues considérées à l'exception du finnois où notre système se classe premier. Globalement, nous conservons notre 3e 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 le Morpho Challenge depuis quelques années, se glisse en 4e place sur cette métrique.

En conclusion, les résultats obtenus à cette compétition sont très encourageants. D'autant plus qu'il n'y avait que 3 mois entre le début du développement de *Moranapho* et la compétition. Nous n'avons donc pas eu le temps de trouver les valeurs d'hyperparamètres

[7] Les systèmes ne se retrouvant pas dans la seconde évaluation sont *Morfessor Baseline* et la référence naïve *Letters*.

[8] Ces systèmes sont des variantes du système *Paramor* et le système par calcul d'incertitude de Bernhard (2008) (section 3.1.1).

Tableau 7.II – F-mesure et le rang global des systèmes analogiques, du système *Morfessor CatMAP* et du système *MorphoNet* sur le jeu de données `DonBruitée` évalués par la métrique `EMMA`.

	Ang.	Fin.	Tur.	All.	Arb.	Rang global /11
Moranapho	72,84	**45,09**	41,36	58,74	29,31	3
Rali-Ana	65,23	36,06	37,34	51,37	**31,54**	8
Morfessor CatMAP	**73,13**	40,12	**45,40**	**63,14**	26,97	1
MorphoNet	72,22	37,62	39,00	59,74	31,32	5

maximisant les performances de nos systèmes.

7.1.2 Évaluation en condition optimale

Malgré les résultats compétitifs obtenus à Morpho Challenge, plusieurs facteurs ont pu nuire à la qualité des analyses morphologiques que nous avons produites. Premièrement, les expériences effectuées par la suite démontrent que les valeurs d'hyperparamètres utilisées dans le cadre du Morpho Challenge 2009 ne sont pas optimales (section 7.2.1). De plus, il est fort probable que l'utilisation d'un lexique bruité favorise les modèles MDL (*Morfessor*) au détriment des systèmes à base de règles (*Moranapho*, *MorphoNet*). Nous croyons que le principe même du MDL offre une certaine protection aux bruits en imposant que tous les morphèmes se retrouvant dans une analyse soient aussi fréquemment identifiés dans les analyses des autres mots du lexique. En ce qui concerne les modèles à base de règles, leur flexibilité les rend particulièrement vulnérables à l'insertion de formes fautives dans le lexique.

Nous avons donc effectué une autre série de tests utilisant les lexiques propres du jeu de données `DonPropre` que nous avons évaluée sur la référence `RefCelex` par la même métrique que celle utilisée pour le Morpho Challenge. Nous avons aussi profité de l'occasion pour tester, en plus de la productivité, la f-prod qui combine la productivité à la fréquence (section 5.1.3). Les résultats obtenus sur cette tâche par les 2 variantes de *Moranapho* et les 2 variantes du système *Morfessor* publiquement disponibles[9] se retrouvent dans le tableau 7.III.

[9]http ://www.cis.hut.fi/projects/morpho/

Tableau 7.III – Précision (Pr.), Rappel (Rp.) et F-mesure (F1) des systèmes *Moranapho* (`Mora.`) et *Morfessor* (`Morf.`) sur le jeu de données `DonPropre` évalués sur la référence `RefCelex` par la métrique du Morpho Challenge.

	Anglais			Néerlandais			Allemand		
	Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
Mora. prod	76,26	57,58	65,50	69,37	43,50	53,32	54,50	64,40	58,97
Mora. f-prod	81,06	56,40	**66,41**	67,21	47,10	**55,25**	58,31	61,35	**59,63**
Morf. Base	71,65	61,82	66,24	79,52	38,10	51,37	80,54	27,73	41,13
Morf. Cat	64,71	63,13	63,76	70,42	45,51	55,11	68,98	42,53	52,50

Une première observation est que tout comme sur les lexiques de `DonBruitée`, *Morfessor CatMAP* surpasse *Morfessor Baseline* sur toutes les langues sauf l'anglais. Cela est probablement dû à la simplicité de la morphologie anglaise pour laquelle des procédures plus simples d'AMNS sont peut-être mieux adaptées à la tâche. Deuxièmement, nous sommes heureux de noter que la version de *Moranapho* utilisant la 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 respectivement bons en anglais et en néerlandais, notre système performe bien sur les deux langues. Ceci démontre donc la nécessité d'évaluer les systèmes d'AMNS sur plusieurs langues et que la stabilité d'une approche sur plusieurs langues est une mesure qui est intéressante à prendre en compte. Pour ce qui est de l'allemand, notre système surpasse clairement les systèmes *Morfessor* avec 7 points de plus de f-mesure comparé à *Morfessor CatMAP* dû à une avance de plus de 20 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.

Afin d'expliquer les disparités entre les résultats de *Moranapho* et de *Morfessor CatMAP*, nous avons inclus des extraits d'analyses produites par ces 2 systèmes (figure 7.1). Nous pouvons observer à partir de ces exemples que les 2 systèmes sont aptes à cap-

		Moranapho	*Morfessor CatMAP*
Eng.	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
Dut.	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
Ger	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 7.1 – Extraits des analyses produites par *Moranapho* et *Morfessor CatMAP* sur le jeu de données `DonPropre`.

turer les cas d'affixation de morphes flexionnels et 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 anglais par contre, *Morfessor CatMAP* a tendance à surestimer le nombre d'affixes alors que *Moranapho* semble s'être mieux adapté à la simplicité morphologique de cette langue telle qu'illustrée par les analyses des mots `redecorating` et `skateboarding`. Les 2 approches détectent aussi plutôt bien les cas de morphologie compositionnelle (*p. ex.* `skateboarding`). Cependant, *Moranapho* traite la composition de la même manière que l'affixation, malgré que la fréquence d'occurrence d'un affixe est beaucoup plus élevée que celle d'un composant. Ceci a pour conséquence que seulement les composants fréquemment utilisés sont détectés, comme le `board` de `skateboard` ou le `national` de `nationalliberalem`, alors que les cas plus rares lui échappent. Notre système se démarque en allemand par sa capacité à détecter les apophonies, tel que pour le terme `länge` (`longueur`) dérivé de `lang` (`long`)[10]. Ceci peut en partie expliquer l'écart important entre *Moranapho* et *Morfessor* sur cette langue. En effet, l'avantage principal de notre approche sur *Morfessor CatMAP* est que les cas de mutation du *stem* ou d'affixation modifiant le *stem* sont gérés de façon élégante. Par exemple, notre approche établit le lien entre les mots néerlandais `officium` et `officia` alors que *Morfessor CatMAP* ne peut gérer adéquatement ce cas, car il n'y

[10]Néanmoins, notre succès est amoindri par l'étiquette peu générique choisie par le système pour représenter cette apophonie.

a pas de frontière de morphème bien définie. Des cas similaires sont aussi observables en anglais (*p. ex.* `bacteriologists`, `redecorating`).

Tableau 7.IV – Le nombre de morphèmes distincts (Morph. Dist.) et le nombre de morphèmes moyens pour une analyse (Morph./Mot) des analyses de *Moranapho* et *Morfessor* sur le jeu de données `DonPropre`.

	Anglais		Néerlandais		Allemand	
	Morph. Dist.	$\frac{\text{Morph.}}{\text{Mot}}$	Morph. Dist.	$\frac{\text{Morph.}}{\text{Mot}}$	Morph. Dist.	$\frac{\text{Morph.}}{\text{Mot}}$
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

Un autre point de comparaison intéressant entre 2 systèmes est de regarder la distribution des morphèmes. Comme indiqué par le tableau 7.IV, le 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 plutôt qu'il y a un manque de cohérence dans les étiquettes de morphèmes attribués. Comme expliqué à la section 5.3, le problème vient du fait que le choix de l'étiquette est local à un ADM. Ce problème est amplifié par la tendance de *Moranapho* à sursegmenter[11] le graphe. Par exemple, dans le meilleur des cas, les partitions `marry` et `marriage` de la figure 7.2 devraient n'en former qu'une seule. Néanmoins, la moyenne de morphèmes par mot de *Moranapho* est plus proche de la réalité que celui des systèmes *Morfessor* bien que nous sous-estimons constamment cette valeur pour toutes les langues.

7.2 Étude de l'impact des composantes du système sur les performances

Lors de la soumission de nos analyses à Morpho Challenge 2009, peu de temps s'était écoulé depuis le début du projet et beaucoup de questions étaient sans réponses. Les expériences effectuées par la suite nous ont permis d'élucider certaines de ces interrogations telles que :

[11] Il est même fréquent qu'une partition contienne un seul mot créant ainsi un morphème hapax.

Racines	Mots
abandon	abandon abandoned abandoning abandonment abandons
marriage	intermarriage intermarriages marriage marriageable marriages remarriage remarriages
marry	intermarried intermarries intermarry intermarrying married marries marry marrying remarried remarries remarry remarrying unmarried

Figure 7.2 – Exemples de partitionnements obtenus par notre système sur le lexique anglais de *DonPropre*.

1. Quel est l'impact des valeurs d'hyperparamètres choisies sur la qualité des analyses ? (section 7.2.1)
2. Quel niveau de flexibilité doit-on laisser au système lors de l'identification des analogies ?(section 7.2.2)
3. Est-ce que l'utilisation de l'analogie formelle améliore les performances du système ? (section 7.2.3)

7.2.1 Impact des valeurs d'hyperparamètres

La table 7.V montre les résultats obtenus par notre système sur le jeu de données du Morpho Challenge *DonBruitée* évalués sur la référence *RefCelex* (anglais et allemand) selon les différentes valeurs d'hyperparamètres qui le contrôlent : ρ[12] et τ[13]. On peut premièrement constater que les résultats obtenus sur la référence *RefCelex* sont similaires à ceux obtenus sur la référence *RefMorpho* avec les mêmes valeurs d'hyperparamètres (tableau 7.I). Ceci n'est pas surprenant, car elles sont toutes deux tirées 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 *RefCelex* 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 par le fait qu'en moyenne, le nombre de mots en relation est

[12]Le poids minimal d'une règle.
[13]Le score minimum qu'un chemin entre deux mots doit avoir pour que ces mots soient regroupés dans un même ADM.

plus élevé pour l'allemand que pour l'anglais (tableau 6.I). Donc en réduisant le seuil, plus de relations sont créées ce qui avantage les langues morphologiquement plus complexes. Pour ce qui est du filtre ρ, si la valeur de τ est bien choisie, on gagne à lui attribuer la valeur la plus petite possible. Cependant, ceci augmente les temps de calcul et les gains obtenus par la réduction de ρ deviennent minimes passé un certain point.

Tableau 7.V – 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 *DonBruitée* évalué sur la référence *RefCelex*. Les analyses soumises à Morpho Challenge 2009 sont soulignées et les meilleurs résultats sont en gras.

		$\rho = 0{,}20$			$\rho = 0{,}25$			$\rho = 0{,}30$		
		Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
$\tau = 0{,}20$	Ang.	51,98	52,13	51,81	59,55	47,94	52,92	66,43	44,80	53,32
	All.	60,01	42,59	49,39	66,99	37,71	47,81	70,72	32,11	43,71
$\tau = 0{,}25$	Ang.	61,02	48,85	54,04	60,59	47,90	53,31	65,67	44,76	53,05
	All.	**68,89**	**41,66**	**52,24**	67,07	40,15	49,61	70,56	32,11	43,70
$\tau = 0{,}30$	Ang.	**65,97**	**46,75**	**54,52**	66,20	45,29	53,59	66,97	44,72	53,46
	All.	68,89	41,66	51,38	69,87	36,24	47,02	71,03	32,11	43,82
$\tau = 0{,}35$	Ang.	71,46	43,09	53,54	71,27	40,81	51,65	72,45	39,62	50,94
	All.	70,66	33,42	44,75	72,14	30,01	41,60	74,05	25,99	37,89
$\tau = 0{,}40$	Ang.	74,22	41,15	52,76	76,10	38,88	51,26	76,72	37,61	50,25
	All.	71,48	32,43	44,00	74,99	28,98	41,09	75,99	24,38	36,30

7.2.2 Impact du degré de l'analogie

Comme mentionné à la section 4.1, le degré d'une analogie indique le nombre minimal de cofacteurs impliqués dans l'analogie. Par exemple, l'analogie [*reader* : *unreadable* = *dreamer* : *undreamable*] implique 3 cofacteurs *un*/ϵ, *dream*/*read* et *able*/ϵ et est donc de degré 3. Puisque les mots anglais contiennent rarement plus de 3 morphèmes, il est peut-être avantageux de retirer les analogies de degré élevé pour éviter qu'elles introduisent du bruit dans les analyses. Pour valider ceci, nous avons retiré de notre ensemble d'analogies celles ayant un degré supérieur à un seuil donné. Les résultats de cette expérience se retrouvent à la figure 7.3.

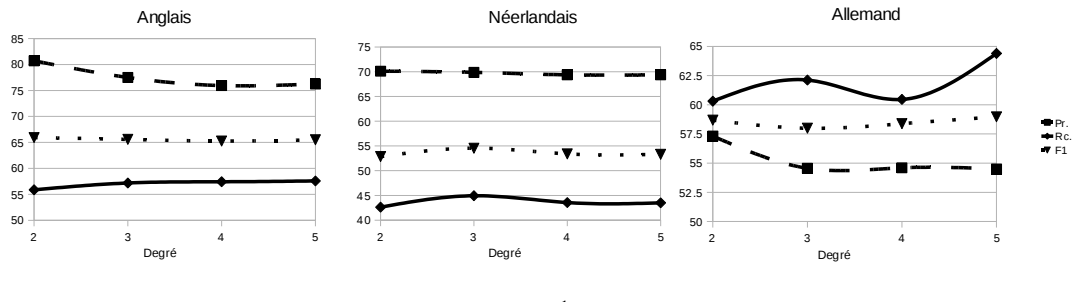

Figure 7.3 – Précision (Tiret), Rappel (Pleine) et F-mesure (Pointillée) de *Moranapho* selon la limite imposée sur le degré des analogies considérées. Les analyses ont été produites sur le jeu de données `DonPropre` et ont été évaluées sur la référence `RefCelex` par la métrique d'évaluation de Morpho Challenge.

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 dû à 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 (tableau 6.I). La courbe pour l'allemand est par contre un peu plus inusitée. 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. Aussi, l'allemand est la seule langue étudiée où les règles qui ne peuvent être extraites par une analogie de degré 2 sont communes. Un cas typique de ceci est l'apophonie qui peut par exemple lier un verbe au présent (*p. ex.* `fallen`) à sa forme au passé (*p. ex.* `fällen`). De plus, autoriser les analogies de degré supérieur à 2 peut être utile dans les cas de composition impliquant plusieurs composants tels que [`talg`*s* : `talg`*lichts* = `tee` : `tee`*licht*] (degré 3) ou [`atomkraftwerk`*en* : `atomkrieg`*en* = `kraftwerks` : `kriegs`*s*] (degré 4). La perte en rappel entre le degré 3 et 4 en l'allemand est due à l'introduction d'une règle causant une inconsistance dans l'appellation d'un morphème commun de cette langue.

7.2.3 Impact du processus d'acquisition de règles

Les résultats que nous avons présentés à la section 7.1 démontrent que des performances rivalisant avec l'état de l'art peuvent être obtenues par un système basé sur l'analogie formelle. Toutefois, ceci ne démontre 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 basée sur la distance d'édition détaillée à la section 5.1.1. En un mot, nous alignons par distance d'édition des paires de mots proches du lexique et les sections non alignées sont transformées en règles de réécriture. Puisque le calcul de la productivité n'utilise aucune information particulière à l'analogie, nous utilisons cette métrique pour noter les règles des 2 systèmes. Comme le démontre la table 7.VI, l'utilisation de la distance d'édition réduit la qualité des analyses pour toutes les langues bien que la différence soit minime en anglais. La similarité entre les résultats anglais des 2 variantes tend à indiquer que l'analogie n'apporte pas de bénéfice sur les langues morphologiquement peu complexes sur lesquelles une technique plus simple est tout aussi adaptée. Néanmoins, le système à base d'analogies performe clairement mieux que celui basé sur la distance d'édition pour le néerlandais. Nous avons été surpris d'observer que le gain prodigué par l'analogie en allemand est inférieur à celui observé en néerlandais. Ceci va à l'encontre de notre première intuition qui était que l'écart entre les 2 variantes augmenterait avec la complexité de la langue à l'avantage de l'analogie due au plus grand nombre de phénomènes qu'elle peut capturer. Toutefois, l'incrément majeur en rappel tend à indiquer qu'il y a bien eu plus de règles d'identifiées, mais à un coût élevé en précision. Ceci pourrait vouloir dire qu'une partie des règles supplémentaires trouvées par analogie ne sont pas valides. Une autre explication serait que le format de règle utilisé ne restreint pas assez l'application des règles et que par conséquent, des règles valides sont appliquées à des mots auxquels elles ne devraient pas l'être.

Tableau 7.VI – 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.). Les analyses ont été produites sur le jeu de données `DonPropre` et ont été évaluées sur la référence `RefCelex` par la métrique d'évaluation de Morpho Challenge.

	Anglais			Néerlandais			Allemand		
	Pr.	Rp.	F1	Pr.	Rp.	F1	Pr.	Rp.	F1
Moranapho Ana.	76,26	**57,58**	**65,50**	69,37	**43,50**	**53,32**	54,50	**64,40**	**58,97**
Moranapho D.E.	**78,74**	55,18	64,79	**74,86**	33,24	45,86	**69,52**	45,91	55,22

CHAPITRE 8

CONCLUSION

Nous avons présenté dans ce mémoire le système à base d'analogie *Moranapho* dont les performances rivalisent avec l'état de l'art. Ce système se différencie du système par analogie de Langlais (2009) par sa capacité à généraliser l'information capturée passivement par une analogie. Un des avantages de cette abstraction est que seulement un sous-ensemble des analogies impliquant les mots d'un lexique est nécessaire pour produire les analyses de l'ensemble des mots du lexique. Par conséquent, ce système est applicable à des lexiques beaucoup plus volumineux que ceux analysables par le système de segmentation suggéré par Langlais (2009).

De plus, notre système surpasse les systèmes *Morfessor* sur les lexiques propres et obtient des résultats supérieurs à *Morfessor Baseline* et comparables à *Morfessor Cat-MAP* lorsque les lexiques sont bruités. Nos expériences ont aussi démontré que les règles de réécriture extraites par analogie formelle sont de meilleure qualité que celles obtenues par un processus basé sur la distance d'édition. Ces résultats appuient donc notre hypothèse que l'analogie formelle peut être utilisée efficacement pour réaliser de manière non supervisée l'analyse morphologique d'une langue.

Néanmoins, plusieurs pistes de recherche n'ont pas été explorées. Premièrement, bien que la généralisation des analogies permet de réduire le nombre d'analogies nécessaires au bon fonctionnement du système, nous ne savons pas quelle est la proportion d'analogies qui doivent être calculées. Nos expériences préliminaires sur le sujet semblent indiquer que de calculer les analogies sur seulement 10% des mots d'un lexique anglais en comptant 300 000 serait suffisant pour trouver tous les principaux cas de flexions et de dérivations.

Deuxièmement, bien que notre format de règle capture un niveau intéressant d'information contextuelle, d'autres alternatives seraient à considérer. Une possibilité serait d'utiliser des expressions régulières comme celles employées par le système *MorphoNet*. Un tel format de règle nous permettrait de gérer de façon élégante la morphologie patron-racine, mais risquerait néanmoins d'introduire du bruit dans nos analyses.

Troisièmement, un des problèmes majeurs de *Moranapho* que nous avons identifié est le manque de cohérence dans le choix des étiquettes de morphème utilisées dans nos

analyses (section 5.3). En effet, comme cette décision est prise localement pour chaque ADM, l'étiquette identifiant un même morphème peut changer d'un ADM à l'autre. Pour corriger ceci, plusieurs options s'offrent à nous. Certaines sont très simples, comme d'imposer une étiquette particulière à un regroupement de règles équivalentes. D'autres par contre sont plus complexes, tels que de fusionner les arbres similaires avant de prendre la décision.

Finalement, nous avons observé une tendance à la sursegmentation des ADM que nous produisons. Il existe plusieurs algorithmes de partitionnement autre que celui que nous avons implanté qui pourraient potentiellement donner de meilleurs résultats. Cependant, rien ne prouve qu'améliorer la qualité de la partition du graphe aurait un impact majeur sur la qualité des analyses morphologiques qui en découle. Une première étape serait d'évaluer les gains en performance obtenus par un partitionnement parfait du graphe produit à partir de la référence exacte.

BIBLIOGRAPHIE

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

Frank Anshen et Mark Aronoff. Producing morphologically complex words. *Linguistics*, 26:641–655, 1988.

Mark Aronoff et Kirsten Fudeman. *What is Morphology ?* Blackwell Publishing, Malden, 2005.

Harald Baayen. Quantitative aspects of morphological productivity. *Yearbook of Morphology 1991*, pages 109–149, 1992.

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

Delphine Bernhard. Simple morpheme labelling in unsupervised morpheme analysis. Dans *Lecture Notes in Computer Science : Cross Language Evaluation Forum 2007 (CLEF'07)*, pages 873–880, Berlin, 2008. Springer.

Delphine Bernhard. Morphonet : Exploring the use of community structure for unsupervised morpheme analysis. Dans *Lecture Notes in Computer Science : Cross Language Evaluation Forum 2009 (CLEF'09)*, Corfu, 2010. Springer.

Stefan Bordag. Unsupervised and knowledge-free morpheme segmentation and analysis. Dans *Lecture Notes in Computer Science : Cross Language Evaluation Forum 2007 (CLEF'07)*, pages 881–891, Berlin, 2008. Springer.

Martin Braschler et Bärbel Ripplinger. How effective is stemming and decompounding for german text retrieval ? *Information Retrieval*, 7(3-4):291–316, 2004.

Vit Bubenik. *An Introduction to the Study of Morphology*. Lincom Europa, Muenchen, 1999.

Burcu Can et Suresh Manandhar. Unsupervised learning of morphology by using syntactic categories. Dans *Notes de travail du Cross Language Evaluation Forum 2009 (CLEF'09)*, Corfu, 2009.

Bruno Cartoni. Lexical morphology in machine translation : a feasibility study. Dans *European Chapter of the Association for Computational Linguistics 2009 (EACL'09)*, pages 130–138, Athens, 2009. Association for Computational Linguistics.

Erwin Chan. Learning probabilistic paradigms for morphology in a latent class model. Dans *Special Interest Group on Computational Phonology 2006 (SIGPHON'06)*, pages 69–78, Morristown, 2006. Association for Computational Linguistics.

Mathias Creutz et Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. Dans *Adaptive Knowledge Representation and Reasoning 2005 (AKRR'05)*, volume 5, pages 106–113, Espoo, 2005.

Mathias Creutz et Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1): 1–34, 2007.

Étienne Denoual. Analogical translation of unknown words in a statistical machine translation framework. Dans *Machine Translation Summit 2007 (MT Summit'07)*, pages 8–14, Copenhagen, 2007.

Alan Ford, Rajendra Singh et Gita Martohardjono. *Pace Pānini*. Peter Lang, New-York, 1997.

Bernard Fradin. *Nouvelles approches en morphologie*. Press Universitaire de France, Paris, 2003.

John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27:153–198, 2001.

John Goldsmith. Morphological analogy : Only a beginning, 2007.

Margaret A. Hafer et Stephen F. Weiss. Word segmentation by letter success varieties. *Information Storage and Retrieval*, 10:371–385, 1974.

Zellig S. Harris. From phoneme to morpheme. *Language*, 31(2):190–222, 1955.

Nabil Hathout. From wordnet to celex : acquiring morphological links from dictionaries of synonyms. Dans *Language Resources and Evaluation 2002 (LREC'02)*, pages 1478–1484, Las Palmas de Gran Canaria, 2002.

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

Yu Hu, Irina Matveeva, John Goldsmith et Colin Sprague. Using morphology and syntax together in unsupervised learning. Dans *Psychocomputational Models of Human Language Acquisition 2005 (PsychoCompLA'05)*, pages 20–27, Ann Arbor, 2005. Association for Computational Linguistics.

Jonh Thayer Jensen. *Morphology : Word Structure in Generative Grammar*. Jonh Benjamins Publishing Company, Amsterdam, 1990.

Francis Katamba et Jonh Stonham. *Morphology*. Palgrave macmillan, New-York, 2e édition, 2006.

M. Kurimo, M. Creutz et M. Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard — morpho challenge 2007. Dans *Notes de travail du Cross Language Evaluation Forum 2007 (CLEF'07)*, Budapest, 2007a.

Mikko Kurimo, Mathias Creutz et Ville Turunen. Overview of morpho challenge in clef 2007. Dans *Notes de travail du Cross Language Evaluation Forum 2007 (CLEF'07)*, Budapest, 2007b.

Mikko Kurimo et Matti Varjokallio. Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – morpho challenge 2008. Dans *Notes de travail du Cross Language Evaluation Forum 2008 (CLEF'08)*, Aarhus, 2008.

Mikko Kurimo, Sami Virpioja, Ville Turunen, Graeme Blackwood et William Byrne. Overview and results of morpho challenge 2009. Dans *Notes de travail du Cross Language Evaluation Forum 2009 (CLEF'09)*, Corfu, 2009.

Philippe Langlais. Étude quantitative de liens entre l'analogie formelle et la morphologie constructionnelle. Dans *Traitement Automatique des Langues Naturelles 2009 (TALN'09)*, Senlis, 2009.

Philippe Langlais et Alexandre Patry. Translating unknown words by analogical learning.

Dans *Empirical Methods in Natural Language Processing - Computational Natural Language Learning 2007(EMNLP-CoNLL'07)*, pages 877–886, Prague, 2007.

Philippe Langlais et Alexandre Patry. Enrichissement d'un lexique bilingue par apprentissage analogique. *Traitement Automatique des Langues*, 49 (varia):13–40, 2008.

Philippe Langlais et François Yvon. Scaling up analogical learning. Rapport technique 2008D014, Paritech, INFRES, Paris, 2008.

Jean François Lavallée et Philippe Langlais. Unsupervised morphological analysis by formal analogy. Dans *Notes de travail du Cross Language Evaluation Forum 2009 (CLEF'09)*, Corfu, 2009.

Jean François Lavallée et Philippe Langlais. Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques. Dans *Traitement Automatique des Langues Naturelles 2010 (TALN'10)*, Montréal, 2010a.

Jean François Lavallée et Philippe Langlais. Unsupervised morphological analysis by formal analogy. Dans *Lecture Notes in Computer Science : Cross Language Evaluation Forum 2009 (CLEF'09)*, pages 618–625, Corfu, 2010b. Springer.

Yves Lepage et Étienne Denoual. Purest ever example-based machine translation : Detailed presentation and assessment. *Machine Translation*, 29:251–282, 2005.

Joel Martin, Howard Johnson, Benoit Farley et Anna Maclachlan. Aligning and using an english-inuktitut parallel corpus. Dans *Proceedings de HLT-NAACL 2003*, pages 115–118, Edmonton, 2003. Association for Computational Linguistics.

Igor A. Mel'čuk. *Cours de morphologie générale Vol. II.* Les presses de l'Université de Montréal, Montréal, 1993.

Igor A. Mel'čuk. *Aspects of the Theory of Morphology.* Mounton de Gruyter, Berlin, 2006.

Christian Monson, Jaime Carbonell, Alon Lavie et Lori Levin. Paramor : Minimally supervised induction of paradigm structure and morphological analysis. Dans *Special Interest Group on Computational Morphology and Phonology (SIGMORPHON'07)*, pages 117–125, Prague, 2007. ACL.

Fabienne Moreau, Vincent Claveau et Pascale Sébillot. Automatic morphological query expansion using analogy-based machine learning. Dans *European Conference on Information Retrieval 2007 (ECIR'07)*, pages 222–233, 2007.

Sylvain Neuvel et Sean A. Fulop. Unsupervised learning of morphology without morphemes. Dans *Special Interest Group on Computational Morphology and Phonology (SIGMORPHON'02)*, pages 31–40. ACL Publications, 2002.

Mark E. J. Newman. Detecting community structure in network. *The European Physical Journal B*, 38:321–330, 2004.

François Pirrelli, Vito et Yvon. Analogy in the lexicon : a probe into analogy-based machine learning of language. 1999.

Stefano Pirrelli, Vito et Federici. An analogical way to language modelling : Morpheus. *Acta Linguistica Hungarica*, 41:235–263, 1993.

J. Rissanen. Introduction an introduction to the mdl principle, 2008.

Selvarajan Saraswathi et Thekkumpurath Geetha. Comparison of performance of enhanced morpheme-based language model with different word-based language models for improving the performance of tamil speech recognition system. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(3):9, 2007.

Ferdinand Saussure. *Cours de linguistique générale*, volume 1. Otto Harrassowitz, Wiesbaden, rudolf engler édition, 1968.

Patrick Schone et Daniel Jurafsky. Knowledge-free induction of inflectional morphologies. Dans *North American Chapter of the Association for Computational Linguistics 2001 (NAACL'01)*, pages 183–191, Pittsburgh, 2001.

Andrew Spencer et Arnold Zwicky. *The Handbook of Morphology*. Blackwell Publishing, Oxford, 1998.

Sebastian Spiegler. Emma : A novel evaluation metric for morphological analysis - experimental results in detail. Rapport technique CSTR-10-004, University of Bristol, Bristol, 2010.

Sebastian Spiegler et Christian Monson. Emma : A novel evaluation metric for morphological analysis. Dans *Conference on Computational Linguistics 2010 (CoLing'10)*, Beijing, 2010.

Nicolas Stroppa et François Yvon. An analogical learner for morphological analysis. Dans *Computational Natural Language Learning 2005 (CoNLL'05)*, pages 120–127, Ann Arbor, 2005.

Kristina Toutanova, Hisami Suzuki et Achim Ruopp. Applying morphology generation models to machine translation. Dans *ACL 2008*, pages 514–522, Colombus, 2008.

Sami Virpioja et Oskar Kohonen. Unsupervised morpheme discovery with allomorfessor. Dans *Notes de travail du Cross Language Evaluation Forum 2009 (CLEF'09)*, Corfu, 2009.

David Yarowsky et Richard Wicentowski. Minimally supervised morphological analysis by multimodal alignment. Dans *Association for Computational Linguistics 2010 (ACL'10)*, pages 207–216, Morristown, 2000. Association for Computational Linguistics.

François Yvon, Nicolas Stroppa, Arnaud Delhay et Laurent Miclet. Solving analogical equations on words. Rapport technique D005, École Nationale Supérieure des Télécommunications, Paris, 2004.