Apprentissage non supervisé de familles morphologiques : comparaison de méthodes et aspects multilingues

Delphine Bernhard

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LIMSI-CNRS
B.P. 133
F-91403 Orsay CEDEX
Delphine.Bernhard@limsi.fr

RÉSUMÉ. *Cet article décrit MorphoClust et MorphoNet, deux méthodes pour l'apprentissage non supervisé de familles morphologiques. MorphoClust forme des familles par groupements successifs, de manière similaire aux méthodes de classification ascendante hiérarchique. La méthode MorphoNet est quant à elle fondée sur la détection de communautés dans des réseaux lexicaux. Les nœuds de ces réseaux représentent des mots et les liens des règles de transformation morphologique acquises automatiquement à partir de mots graphiquement similaires. Nous appliquons ces deux méthodes à un lexique bilingue anglais-allemand, de manière isolée et sous forme combinée, et évaluons les résultats obtenus en utilisant la base de données lexicales CELEX.*

ABSTRACT. *This article describes MorphoClust and MorphoNet, two methods for the unsupervised acquisition of morphological families. MorphoClust builds families by iterative conflations, similarly to hierchical clustering methods. The MorphoNet method relies on community detection in lexical networks. The nodes of these networks stand for words while edges represent morphological transformation rules which are automatically acquired based on graphical similarities between words. The two methods are applied to a German-English bilingual lexicon, both in isolation and in combination. We evaluate the results using the CELEX lexical database.*

MOTS-CLÉS : *morphologie, apprentissage non supervisé, combinaison de systèmes.*

KEYWORDS: *morphology, unsupervised learning, system combination.*

1. Introduction

L'analyse morphologique est une tâche primordiale dans divers domaines du traitement automatique des langues comme la reconnaissance de la parole, la communication alternative et augmentée, la traduction automatique ou la recherche d'information. Dans ce dernier cas, l'utilité des connaissances morphologiques se justifie par la proximité sémantique des variantes flexionnelles ou dérivationnelles appartenant à une même famille morphologique.

Les ressources morphologiques ne sont toutefois pas disponibles à l'heure actuelle pour toutes les langues et tous les domaines, ou sont difficilement applicables à certaines données, notamment celles qui comportent des fautes d'orthographe ou des néologismes. De nombreux travaux récents ont, de fait, cherché à acquérir automatiquement des connaissances morphologiques à l'aide de méthodes non supervisées ou semi-supervisées, se caractérisant par une approche multilingue, peu ou pas dépendante de la langue cible et nécessitant un minimum de ressources langagières. L'objectif affiché est non seulement de produire des ressources morphologiques, mais également de mieux comprendre les phénomènes sous-jacents à la formation de mots dans diverses langues, parfois très éloignées (Kurimo *et al.*, 2006).

Les méthodes sur lesquelles reposent ces systèmes sont par ailleurs variées : comparaison de graphies (Zweigenbaum et Grabar, 2000), recherche d'analogies (Lepage, 1998 ; Stroppa et Yvon, 2006), modèles probabilistes (Creutz et Lagus, 2005) ou segmentation par optimisation (Goldsmith, 2001 ; Creutz et Lagus, 2002). Les systèmes proposés dans la littérature se distinguent également par le type de résultats obtenus : mots découpés en segments, analyse morphémique complète ou liens morphologiques entre mots.

Les travaux présentés dans cet article relèvent du dernier type car ils consistent en l'acquisition de familles morphologiques, c'est-à-dire des groupes de mots liés deux à deux par un lien morphologique d'affixation (préfixation ou suffixation) ou de composition. Nous décrivons et comparons deux méthodes : MorphoClust et MorphoNet. MorphoClust forme des familles par groupements successifs, d'une manière similaire aux méthodes de classification ascendante hiérarchique. MorphoNet représente les relations morphologiques sous forme de réseau lexical, afin d'utiliser des algorithmes fondés sur les graphes pour détecter des communautés dans le réseau obtenu.

Les deux méthodes prennent pour point de départ une liste de mots et les groupent en familles. Elles ont en outre la particularité d'être non supervisées : elles ne sont donc pas *a priori* liées à une langue ou a un domaine précis. MorphoClust a pour l'heure été appliqué au français, à l'anglais et à l'allemand (Bernhard, 2007), tandis que MorphoNet a été utilisé pour l'anglais, l'allemand, le finnois, le turc et l'arabe (Bernhard, 2010).

Les contributions de ce travail portent sur divers points. Notre première contribution est de proposer deux méthodes efficaces et non dépendantes de la langue cible pour identifier des familles morphologiques. Notre seconde contribution consiste en la

combinaison translingue de ces méthodes. Nous montrons que la combinaison permet d'améliorer les résultats en allemand. Enfin, nous évaluons les méthodes et analysons les résultats obtenus en allemand et en anglais.

Nous détaillons dans un premier temps l'état de l'art des méthodes d'analyse morphologique automatique en TAL (section 2). Puis nous décrivons les méthodes MorphoClust (section 3) et MorphoNet (section 4) avant de présenter et d'analyser les résultats obtenus en allemand et en anglais (section 5), ainsi que les résultats de la combinaison translingue des méthodes (section 6). Enfin, nous examinons les limites des systèmes pour proposer des améliorations futures (section 7).

2. Analyse morphologique en traitement automatique des langues (TAL)

Nous décrivons ici les diverses approches d'analyse morphologique adoptées en TAL, en détaillant tour à tour les approches à base de règles, celles par apprentissage supervisé, et enfin les méthodes d'apprentissage non supervisé. Dès que cela est pertinent, nous détaillons également les aspects multilingues de ces méthodes.

2.1. *Analyse morphologique à base de règles*

Il existe divers types de méthodes d'analyse morphologique à base de règles : (i) les méthodes simples, basées sur des règles heuristiques, telles que les algorithmes de désuffixation, (ii) les transducteurs, utilisés essentiellement pour l'analyse et la génération en morphologie flexionnelle et (iii) les analyseurs morphosémantiques, qui visent à la fois une analyse morphologique et sémantique des mots.

2.1.1. *Désuffixation et racinisation*

Les méthodes de désuffixation, également connues sous le nom de racinisation (*stemming* en anglais) sont surtout utilisées en recherche d'information, en particulier au moment de l'indexation. La racinisation est effectuée par l'application de règles permettant de réduire les mots à des radicaux qu'ils partagent avec d'autres mots. Une classe d'équivalence regroupe alors l'ensemble des mots partageant le même radical. L'algorithme de désuffixation le plus connu est celui de Porter (1980), originellement développé pour l'anglais. Son principe est applicable à diverses langues, mais les heuristiques et leur ordre d'application doivent alors être adaptés. Il existe actuellement des versions de l'algorithme de Porter pour les langues suivantes : français, espagnol, portugais, italien, roumain, allemand, néerlandais, suédois, norvégien, danois, russe, finnois, hongrois et turc. [1]

1. Disponibles à l'adresse suivante : `http://snowball.tartarus.org/` [Visitée le 31.05.2010].

2.1.2. *Transducteurs et morphologie à deux niveaux*

Contrairement à la racinisation, la morphologie à deux niveaux (Koskenniemi, 1984) est une méthode non directionnelle, c'est-à-dire qu'elle peut être utilisée à la fois pour l'analyse et la génération. Les règles sont représentées sous la forme de transducteurs mettant en correspondance le niveau de surface et le niveau lexical de l'analyse. La morphologie à deux niveaux est particulièrement bien adaptée aux langues fortement suffixées comme le finnois ou le turc (ten Hacken et Lüdeling, 2002).

2.1.3. *Analyse morphosémantique*

Les approches morphosémantiques donnent un poids important aux informations sémantiques et ne se limitent donc pas à une analyse purement morphosyntaxique. Des systèmes de ce type ont été proposés pour le vocabulaire médical en anglais (Pratt et Pacak, 1969), en allemand (Hahn *et al.*, 2003) et en français (Lovis *et al.*, 1995). Le système DériF (Namer, 2009), initialement développé pour le français, a, quant à lui, été étendu avec succès au vocabulaire médical anglais (Deléger *et al.*, 2009). Enfin, Cartoni (2009) présente une approche morphosémantique multilingue d'analyse de néologismes construits dans une langue source et de génération de leur traduction dans une langue cible appliquée aux mots préfixés du français et de l'italien.

2.2. *Méthodes d'apprentissage supervisé*

Les méthodes d'apprentissage supervisé nécessitent des données d'apprentissage annotées, associant les données d'entrée aux résultats désirés. Les capacités du système ne dépendent donc pas uniquement de ses propriétés intrinsèques mais avant tout du contenu, de la qualité et de la taille du corpus d'apprentissage. van den Bosch et Daelemans (1999) présentent un système d'apprentissage fondé sur le stockage en mémoire des instances de la base d'apprentissage et appliqué aux analyses morphologiques pour le néerlandais tirées de la base lexicale CELEX (Baayen *et al.*, 1995). Stroppa et Yvon (2006) présentent, quant à eux, une méthode d'apprentissage reposant sur le principe d'analogie et appliquée à l'apprentissage de la morphologie de l'anglais, de l'allemand et du néerlandais à partir des annotations contenues dans CELEX.

2.3. *Méthodes d'apprentissage non supervisé et faiblement supervisé*

Les méthodes que nous venons de présenter nécessitent soit des règles et des lexiques spécifiques à la langue, soit des données annotées permettant l'apprentissage supervisé. Or ces ressources ne sont pas toujours disponibles et leur construction est une tâche longue et complexe. Les méthodes d'analyse morphologique non supervisée ne nécessitent pas de ressources spécifiques et procèdent généralement à partir d'une simple liste de mots de la langue cible, voire d'un corpus. Ces méthodes visent essentiellement le découpage des mots en segments morphémiques, tandis que les

approches par classification sont plus rares. Les avancées dans ce domaine ont été récemment stimulées par la compétition internationale Morpho Challenge [2] qui vise à comparer les systèmes pour des langues (anglais, allemand, finnois, turc, arabe) et des applications diverses (recherche d'information, reconnaissance de la parole, traduction automatique). Les systèmes adoptent généralement les méthodes et heuristiques détaillées dans la suite.

2.3.1. *Comparaison de graphies*

Une des manières les plus immédiates d'acquérir des informations sur la morphologie d'une langue est de comparer la graphie des mots, par diverses méthodes : (i) distances orthographiques (Baroni *et al.*, 2002) ; (ii) repérage du plus long préfixe ou suffixe commun (Jacquemin, 1997 ; Gaussier, 1999 ; Zweigenbaum et Grabar, 2000) ; (iii) inclusion d'autres mots (Keshava et Pitler, 2006 ; Demberg, 2007 ; Bernhard, 2007) ; (iv) analogies (Lepage, 1998 ; Hathout, 2005 ; Moreau et Claveau, 2006).

2.3.2. *Transitions entre sous-chaînes de caractères*

Selon une idée formulée en 1955 par Harris (1955), il est possible d'identifier des segments morphémiques en découpant les représentations phonémiques des mots suivant le nombre de phonèmes différents qui peuvent suivre une séquence initiale de phonèmes (*successor variety*). Un nombre élevé de possibilités indique une frontière morphémique vraisemblable. Cette heuristique a depuis été appliquée aux caractères orthographiques et utilisée dans divers systèmes (Keshava et Pitler, 2006 ; Bernhard, 2006 ; Spiegler *et al.*, 2009).

2.3.3. *Segmentation par compression*

Les méthodes de segmentation par compression reposent sur le principe de longueur minimale de description (*MDL* : *Minimum Description Length*). L'intuition sous-jacente est que la morphologie, grâce à ses régularités, permet une représentation compacte des mots de la langue. L'objectif de la segmentation est alors de trouver un dictionnaire de segments morphémiques et un encodage des mots du corpus à l'aide de ces segments qui soient les plus courts possibles. Ce principe est mis en œuvre par de nombreux auteurs, parmi lesquels Goldsmith (2001) et Creutz et Lagus (2002).

2.3.4. *Modèles probabilistes*

De nombreuses méthodes récentes se placent dans un cadre probabiliste et utilisent l'inférence bayésienne (Creutz et Lagus, 2005), les modèles bayésiens hiérarchiques (Snyder et Barzilay, 2008) ou encore les modèles probabilistes génératifs (Spiegler *et al.*, 2009). Ces approches modélisent des hypothèses généralistes sur la morphologie comme des règles morphotactiques sur les séquences de morphèmes acceptables.

2. `http://www.cis.hut.fi/morphochallenge2009/`

2.3.5. *Approches faiblement supervisées*

Les approches non supervisées peuvent également être combinées à des approches à base de règles, notamment afin de corriger leur résultat. Tepper et Xia (2010) définissent ainsi des règles de réécriture contextuelle appliquées aux résultats d'une analyse non supervisée afin de prendre en compte les cas d'allomorphie en anglais et en turc. L'intervention humaine reste toutefois faible car le temps nécessaire à l'écriture de telles règles est très restreint.

2.4. *Apprentissage de familles morphologiques*

Les deux approches présentées dans cet article visent l'acquisition de familles de mots morphologiquement liés et sont ainsi à rapprocher des méthodes de désuffixation fondées sur des heuristiques dépendantes de la langue cible.

Quelques méthodes de classification morphologique non supervisée ont été proposées. Adamson et Boreham (1974) calculent un coefficient de similarité des mots à partir des bigrammes de lettres puis utilisent un algorithme de classification ascendante hiérarchique reposant sur ces coefficients. Gaussier (1999) utilise également un algorithme de classification ascendante hiérarchique fondé sur une mesure de similarité calculée à partir de la productivité des paires de suffixes liant deux mots. Jacquemin (1997) définit quant à lui une mesure de distance entre suffixes permettant de calculer une distance entre classes.

D'autres méthodes passent par une phase intermédiaire de segmentation ou d'analyse pour identifier les familles morphologiques et n'ont pas recours à un algorithme de classification. Schone et Jurafsky (2001) identifient des mots liés par une paire de préfixes, suffixes ou circumfixes avant de procéder au groupement de ces variantes morphologiques en utilisant le contexte d'occurrence des mots. Moon *et al.* (2009) adoptent une démarche similaire, qui consiste à d'abord identifier des bases et des affixes flexionnels candidats, avant de procéder à leur regroupement dans des familles morphologiques. Ce regroupement est contraint par l'utilisation d'informations sur les frontières entre documents du corpus.

Les méthodes de classification décrites précédemment ont été appliquées essentiellement à des langues fréquemment étudiées en TAL telles que l'anglais, l'allemand ou le néerlandais, mais également à des langues rares telles que l'uspanteko, une langue maya (Moon *et al.*, 2009).

Nous détaillons maintenant nos deux systèmes, MorphoClust et MorphoNet.

3. MorphoClust

La méthode MorphoClust groupe les mots en familles d'une manière similaire aux méthodes de classification ascendante hiérarchique utilisées en analyse de données. Elle diffère toutefois des approches présentées dans l'état de l'art qui reposent

sur une mesure unique de similarité ou de distance morphologique. En effet, MorphoClust utilise divers indices, à diverses étapes de l'algorithme, et inclut une phase de bootstrapping. De plus, elle fait l'hypothèse que les mots peuvent être préfixés ou suffixés, mais ne fait pas de différence explicite entre affixes dérivationnels et flexionnels. Enfin, elle accorde un rôle important au procédé de préfixation, ce qui la rend particulièrement adaptée au traitement des mots construits, issus du vocabulaire de spécialité.

MorphoClust prend pour point de départ les données suivantes :

– une liste des mots d'un corpus M ;

– une liste de préfixes P ;

– une liste de signatures (ou paires de suffixes) S.

Les deux dernières listes sont obtenues à partir de la première à l'aide du module d'apprentissage d'affixes décrit dans (Bernhard, 2006). Celui-ci utilise les probabilités transitionnelles entre sous-chaînes pour repérer les zones de faible probabilité et ainsi découper les mots en radical et affixes. Nous avons adapté ce module pour qu'il produise non seulement une liste de préfixes et de suffixes mais également une liste de paires de suffixes qui apparaissent avec la même base et qui sont donc mutuellement substituables. Par exemple, les suffixes de la paire (*s*,*ique*) peuvent se combiner à la base *climat* pour former les mots *climat**s*** et *climat**ique***. La même signature se retrouve dans les paires de mots *volcan**s*** – *volcan**ique*** et *océan**s*** – *océan**ique***. La notion de signature est présente dans de nombreux travaux en acquisition automatique de connaissances morphologiques, parfois sous des dénominations différentes : *paires de suffixes* (Gaussier, 1999), *règles morphologiques* (Grabar et Zweigenbaum, 1999) ou *schémas de suffixation* (Hathout, 2005).

Nous allons maintenant détailler l'ensemble des étapes menant à l'acquisition des familles morphologiques.

3.1. *Familles initiales*

Avant apprentissage, il y a autant de familles que de mots dans la liste donnée en entrée : chaque mot constitue sa propre famille. Les familles formées au cours du processus d'apprentissage sont représentées par un radical R. De plus, chaque famille comprend deux sous-familles, sauf si elle correspond à une feuille dans la hiérarchie : dans ce cas, elle contient un mot unique et n'a pas de sous-famille.

3.2. *Étape 1 : regroupement de familles à partir de l'inclusion de mots*

Le premier critère de regroupement des familles est l'inclusion de mots : il s'agit de repérer les mots formés par préfixation à partir d'un autre mot de la liste, selon une procédure détaillée ci-après.

Soient :

– $m_1, m_2, \ldots, m_i$ et m_j des mots de longueur minimale égale à 4 ;
– $F_1, F_2, \ldots, F_i$ des familles telles que $F_1 = [m_1], F_2 = [m_2], \ldots, F_i = [m_i]$;
– F_j une famille telle que $F_j = [m_j]$.

Les familles $F_1, F_2, \ldots, F_i$ et F_j sont regroupées pour former une nouvelle famille F_k si $m_1 = E_1 + m_j, m_2 = E_2 + m_j, \ldots, m_i = E_i + m_j$
où E_1, E_2, …, E_i représentent une suite maximale d'un ou plusieurs préfixes de la liste P, éventuellement séparés par des tirets, tels que chaque préfixe ait une longueur minimale de 3.
Le radical de la nouvelle famille F_k est m_j.

Par exemple, si F_1 = [sub-océaniques], F_2 = [océaniques] et F_3 = [intra-océaniques] alors il est possible de former une nouvelle famille F_4 telle que $F_4 = F_1 \cup F_2 \cup F_3$ = [sub-océaniques, océaniques, intra-océaniques]. En effet, les mots *sub-océaniques* et *intra-océaniques* contiennent tous le mot *océaniques*. De plus, ils débutent par les préfixes *sub+* et *intra+*. Le radical de la nouvelle famille est *océaniques*.

3.3. *Étape 2 : regroupement de familles à partir des préfixes*

Après avoir procédé à un premier regroupement des mots en fonction des mots inclus, nous utilisons d'autres critères de regroupement, fondés sur la comparaison des graphies des radicaux des familles existantes et des préfixes auxquels ils peuvent être associés. En effet, lorsque deux mots partagent un même préfixe et que leurs bases sont graphiquement similaires, alors il y a de fortes chances pour qu'ils soient également morphologiquement liés. Prenons l'exemple des mots suivants : *neuro-oncologist* et *neuro-oncology*. Ces deux mots débutent tous deux par le préfixe *neuro-* suivi d'une même chaîne de caractères de longueur 7 : *oncolog*. La combinaison de deux indices, à savoir le partage d'un préfixe, suivi d'une chaîne commune, est un indice suffisant dans la plupart des cas pour conclure que les mots sont morphologiquement liés.

Nous appliquons ces remarques de la manière suivante.

Soient :

– F_1 et F_2 deux familles ;
– R_1 le radical représentant F_1 ;
– R_2 le radical représentant F_2.

Les deux familles F_1 et F_2 sont regroupées dans une nouvelle famille F_3 ssi :

1) $R_1 = \alpha + s_1$ et $R_2 = \alpha + s_2$, où α est une chaîne de caractères de longueur minimale égale à 4 et s_1 et s_2 sont des chaînes de caractères différant au moins par leur premier caractère.

2) Il existe au moins un mot $m_1 \in F_1$ et un mot $m_2 \in F_2$ tels que m_1 et m_2 incluent le même préfixe.

Le radical R_3 de la nouvelle famille F_3 est le mot le plus court parmi R_1 et R_2.

Par exemple, si :

– F_1 = [océanique, intra-océanique] avec R_1 = *océanique*,

– F_2 = [océaniques, sub-océaniques, intra-océaniques] avec R_2 = *océaniques*,

alors il est possible de former une nouvelle famille :

$F_3 = F_1 \cup F_2 =$ [océanique, intra-océanique, océaniques, sub-océaniques, intra-océaniques].

En effet, R_1 et R_2 partagent une chaîne initiale commune de longueur 9, *océanique*, et les mots *intra-océanique* de F_1 et *intra-océaniques* de F_2 ont en commun le préfixe *intra*. Le radical de F_3 est le radical le plus court, à savoir *océanique*.

3.4. *Étape 3 : regroupement de familles à partir des signatures*

La dernière étape de la classification consiste à utiliser la liste de signatures S donnée en entrée et à découvrir de nouvelles signatures à partir des regroupements opérés lors des étapes précédentes. Ces signatures vont permettre à leur tour d'effectuer de nouveaux regroupements, selon le principe du *bootstrapping*. Le processus se termine lorsqu'il n'est plus possible de découvrir de nouvelles signatures.

3.4.1. *Découverte de nouvelles signatures*

La découverte de nouvelles signatures se fait à partir des familles déjà constituées au cours des étapes précédentes. Les mots non préfixés de chaque famille sont comparés deux à deux afin d'obtenir une liste de signatures, selon la méthode suivante.

Soient m_1 et m_2 deux mots non préfixés appartenant à la famille F
tels que $m_1 = \alpha + s_1$ et $m_2 = \alpha + s_2$
avec $|\alpha| \geq 4$ et s_1 et s_2 des chaînes de caractères différant au moins par leur premier caractère.
Nous appellerons signature la paire de suffixes (s_1,s_2) et $sig(F, F)$ l'ensemble des signatures formées à partir d'une famille F, c'est-à-dire par comparaison bijective des mots non préfixés de F. Toutes ces signatures sont ajoutées à la liste des signatures S.

Prenons l'exemple de la famille suivante, formée lors des étapes 1 et 2 :

[trachyandésite, andésite, trachy-andésite, andésites, trachy-andésites, trachyandésites, andésitique, trachy-andésitique, trachyandésitique, trachy-andésitiques, trachyandésitiques, andésitiques].

La comparaison des graphies des mots non préfixés de cette famille conduit à l'identification des paires de suffixes suivantes : (ϵ, s), ($e, ique$), ($e, iques$), ($es, ique$) et ($es, iques$) (voir figure 1).

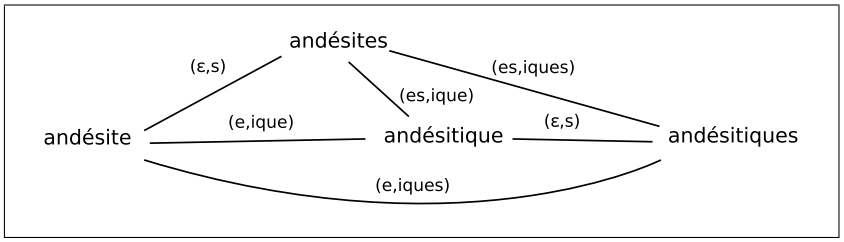

Figure 1. *Identification de signatures*

3.4.2. *Fusion de familles à l'aide des signatures*

Les signatures ainsi acquises sont utilisées pour fusionner des familles. Le critère d'agglomération repose sur un indice p qui mesure la proportion de signatures valides partagées entre deux familles que l'on cherche à fusionner.

Soient :

– F_1 et F_2 deux familles ;

– l_1 le nombre de mots non préfixés de F_1 ;

– l_2 le nombre de mots non préfixés de F_2 ;

– S la liste de signatures fournies en entrée et découvertes à partir des familles déjà constituées.

$$p = \frac{|sig(F_1, F_2) \cap S|}{l_1 \cdot l_2}$$

Dans les expériences relatées dans la suite de cet article, nous avons fusionné deux familles lorsque $p \geq 0,5$.

Prenons l'exemple des familles représentées sur la figure 2. Les signatures connues sont représentées par un arc plein tandis que les signatures inconnues sont représentées en pointillés. Ces deux familles sont fusionnées car le rapport du nombre de signatures connues sur le nombre total de signatures possibles est égal à 0,5.

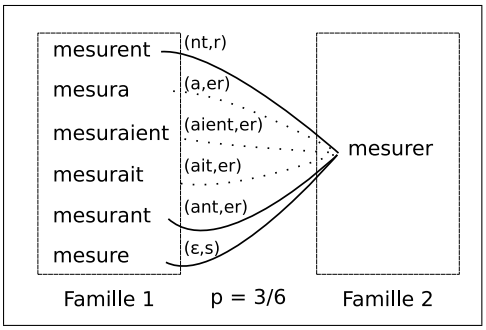

Figure 2. *Fusion de familles*

Le dendrogramme de la figure 3 illustre l'intérêt des regroupements effectués aux diverses étapes de la méthode. La seule famille formée à l'issue des deux premières étapes est [satellites, microsatellites, microsatellite, sous-satellite, satellite, mini-satellite]. L'étape de fusion de familles à partir des signatures partagées permet le regroupement de mots comme [satellitaire, satellitaires] ou [satellisation, satellisait].

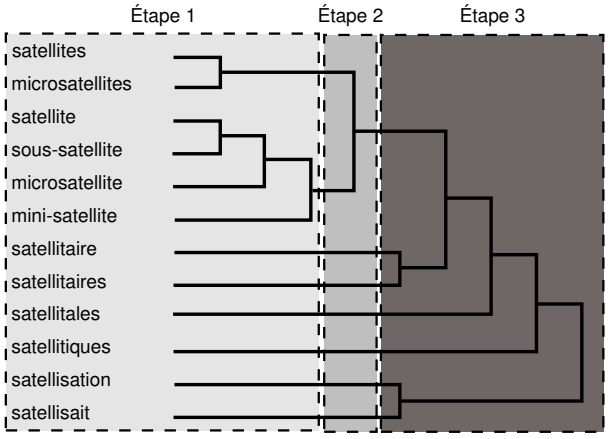

Figure 3. *Familles obtenues par classification à l'issue des trois étapes*

L'étape 3 d'agglomération à partir des signatures partagées est répétée tant que de nouvelles signatures sont acquises à partir des regroupements effectués et tant que ces

signatures permettent de regrouper des familles. Le nombre de signatures différentes augmente fortement au cours des premières itérations, puis se stabilise. Le processus d'acquisition de nouvelles signatures, et par conséquent d'apprentissage de familles, s'achève au bout de 10 à 15 itérations.

4. MorphoNet

La méthode MorphoNet repose sur la construction de réseaux lexicaux qui encodent les relations morphologiques entre mots. Des réseaux lexicaux similaires ont également été décrits par Hathout (2002). Notre approche diffère cependant de celle de Hathout sur les aspects suivants : (i) elle est totalement non supervisée et utilise uniquement une liste de mots en entrée, tandis que la méthode de Hathout repose sur WordNet, et (ii) elle ne se limite pas à l'opération de suffixation car les réseaux lexicaux sont construits à partir de règles de transformation morphologique qui ne font aucune hypothèse sur la structure interne des mots. De plus, MorphoNet utilise un algorithme de détection de communautés dans les graphes qui n'a jamais, à notre connaissance, été appliqué à l'analyse morphologique auparavant.

4.1. *Acquisition de règles de transformation morphologique*

La première étape de la méthode MorphoNet consiste à acquérir un ensemble de *règles de transformation morphologique*. Ces règles définissent des opérations de substitution permettant de générer des variantes morphologiques d'un mot. Nous utilisons la notation suivante pour une règle r : `patron` $\rightarrow$ `rempl`, où le `patron` est une expression régulière et `rempl` encode les opérations de substitution comportant des références au contenu des groupes identifiés dans le patron. Par exemple, la règle `^(.+)ly$` $\rightarrow$ `\1` peut être appliquée au mot *totally* pour produire le mot *total*.

Les règles de transformation sont une généralisation de la notion de *paire d'affixes* que l'on retrouve dans de nombreuses méthodes d'apprentissage de la morphologie, sous des dénominations différentes : modèle (*pattern*) (Hathout, 2002), règles (*rules*) (Schone et Jurafsky, 2000), ou transformations (*transforms*) (Freitag, 2005).

L'avantage des règles de transformation ainsi définies est qu'elles ne se limitent pas en principe aux procédés de construction de mots par concaténation, ce qui peut s'avérer nécessaire pour des langues gabaritiques comme l'arabe, dans le cas de paires de mots telles que *kataba* (il a écrit) et *kutiba* (il a été écrit). Ainsi, ces règles de transformation visent à remédier à certaines limites bien connues des systèmes d'analyse morphologique non supervisée qui font l'hypothèse que les mots sont formés par concaténation de morphèmes.

L'acquisition de ces règles se fait à partir d'un sous-ensemble L de la liste de mots M fournie en entrée. Dans nos expériences, nous avons utilisé les 10 000 mots les plus fréquents dont la longueur est supérieure à la longueur de mots moyenne. De plus, nous avons conservé toutes les règles apprises à partir d'au moins deux paires de

mots différentes. L'algorithme d'extraction des règles de transformation est décrit en détail dans l'algorithme 1.

Algorithme 1 Algorithme d'acquisition de règles de transformation morphologique, à partir d'une liste de mots L.

1: $règles \leftarrow \emptyset$
2: $n \leftarrow \mathsf{longueur}(L)$
3: **pour** $i = 1$ à n **faire**
4: $m \leftarrow L[i]$
5: $mots\_similaires \leftarrow \mathsf{trouve\_mots\_similaires}(m, L[i+1:n], seuil)$
6: **pour** m_2 dans $mots\_similaires$ **faire**
7: $r \leftarrow \mathsf{trouve\_règle}(m, m_2)$
8: ajouter r à $règles$
9: **fin pour**
10: **fin pour**
11: **retourner** $règles$

Pour chaque mot m de la liste L, la méthode recherche les mots graphiquement similaires pour un seuil de similarité donné (ligne 5, trouve_mots_similaires) grâce à l'algorithme de reconnaissance structurelle Ratcliff-Obershelp [3].

Par exemple, pour le mot *democratic*, les mots similaires suivants sont obtenus : *undemocratic*, *democratically*, *democrats*, *democrat's*, *anti-democratic*. L'acquisition de règles (ligne 7, trouve_règle) s'effectue par comparaison du mot cible avec tous les mots similaires pour identifier les sous-chaînes identiques [4] ; par exemple, pour la paire de mots similaires *democratic* et *undemocratic*, la règle suivante est extraite : `^un(.+)$` → `\1`.

4.2. *Construction d'un réseau lexical*

Les règles de transformation morphologique acquises lors de l'étape précédente sont utilisées pour construire un réseau lexical. Les nœuds du graphe représentent les mots de la liste d'entrée M. Deux mots m_1 et m_2 sont connectés par un lien s'il existe une règle de transformation r telle que $r(m_1) = m_2$. Le graphe ainsi obtenu est orienté en fonction de la direction des règles appliquées. La figure 4 donne un exemple de réseau lexical.

Certaines règles sont toutefois plus fiables que d'autres. Nous mesurons la fiabilité d'une règle de transformation r par la mesure de *productivité* P qui correspond au ratio $|E|/|G|$ de mots existants E de M générés par rapport à l'ensemble des mots

3. Nous avons utilisé l'implémentation disponible dans le module Python *difflib*, avec une valeur du paramètre *seuil* spécifiant la similarité minimale fixée à 0,8.
4. Les sous-chaînes identiques sont identifiées à l'aide de la fonction Python get_matching_blocks.

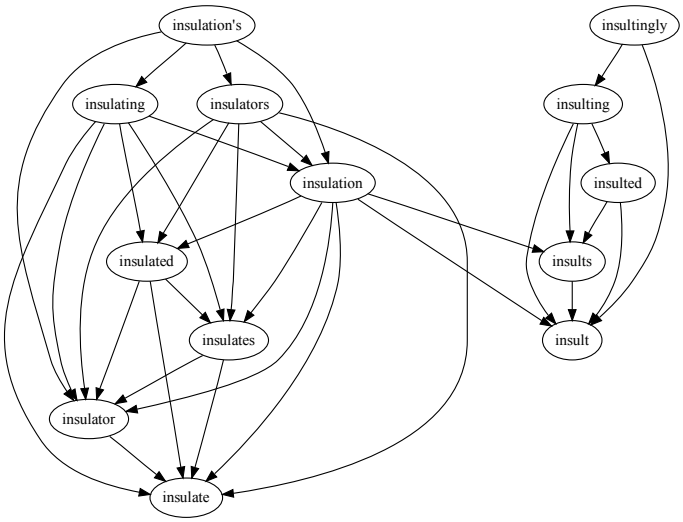

Figure 4. *Exemple de réseau lexical*

générés G par la règle. La productivité a précédemment été utilisée dans un contexte similaire pour pondérer des règles de réécriture (Lavallée et Langlais, 2009). La mesure de productivité permet d'éliminer les règles les moins fiables qui se trouvent sous une valeur seuil, ce qui permet d'obtenir un graphe moins dense.

4.3. *Acquisition de familles de mots*

Les réseaux lexicaux obtenus comprennent généralement une grande composante connexe, avec un ensemble de composantes connexes de plus faible taille. Il n'est donc pas possible d'extraire des familles de mots en identifiant simplement les composantes connexes. Par exemple, le réseau lexical représenté sur la figure 4 comprend une seule composante connexe, mais se décompose clairement en deux familles morphologiques différentes. L'acquisition de familles de mots peut être formulée comme un problème classique de détection de communautés dans les graphes, qui peut être résolu par l'utilisation d'algorithmes de classification spécifiques.

Les communautés correspondent à des groupes de nœuds densément interconnectés, qui se caractérisent par une grande densité de liens intragroupe et une plus faible densité intergroupe (Newman, 2004). Il existe diverses méthodes pour détecter les communautés dans des graphes. Par exemple, l'algorithme *Markov Clustering* (MCL) de van Dongen (2000) consiste à partitionner un graphe en simulant un processus de marche aléatoire ; MCL a été utilisé pour détecter des communautés dans un graphe de noms par Dorow *et al.* (2005). La méthode décrite par Newman

(Newman, 2004 ; Clauset *et al.*, 2004 ; Newman, 2006) repose, quant à elle, sur la notion de modularité Q qui mesure la qualité d'une division des nœuds en communautés. La modularité est importante quand il y a beaucoup de liens au sein des communautés et peu de liens entre elles. Cette méthode a été appliquée à des données langagières par Matsuo *et al.* (2006) pour des graphes construits à partir de mesures de similarité entre mots.

Cette dernière méthode présente l'avantage de détecter automatiquement le nombre optimal de communautés. Il n'est donc pas nécessaire de définir *a priori* le nombre de communautés souhaitées et donc de procéder à des ajustements de ce paramètre. La modularité compare le nombre de liens au sein d'une communauté au nombre de liens attendus :

$$Q = \sum_i (e_{ii} - (\sum_j e_{ij})^2) \quad [1]$$

où e_{ii} est la fraction de liens dans le graphe qui connectent des nœuds appartenant à la communauté i, e_{ij} est la moitié de la fraction de liens du graphe qui connectent des nœuds de la communauté i à ceux de la communauté j et $\sum_j e_{ij}$ est la fraction de liens connectés à des nœuds de la communauté i.

Pour une bonne division du graphe en communautés, le nombre de liens au sein des communautés excède ce qui serait attendu par pur hasard, ce qui correspond à une valeur de modularité Q positive. La modularité est importante quand il y a beaucoup de liens dans des communautés et peu de liens entre communautés. La figure 5 illustre les résultats de la méthode de Newman pour le réseau lexical de la figure 4 : dans ce cas précis, deux communautés ont été détectées.

La plus grande difficulté de mise en œuvre de cet algorithme réside dans la découverte de la division qui donne la meilleure valeur pour Q. Il est bien sûr impossible de tester chaque division possible du graphe. Newman (2004) a donc proposé une méthode de classification ascendante hiérarchique, qui prend pour point de départ des communautés constituées par un nœud unique. Les communautés sont groupées de manière itérative, en choisissant le regroupement qui correspond à la plus grande augmentation (ou la plus petite diminution) de la valeur Q. La meilleure partition du graphe est celle pour laquelle la modularité Q est maximale.

Nos précédentes expériences avec la méthode de Newman ont toutefois démontré qu'elle tend à détecter des communautés trop grandes, conduisant à une baisse de la précision (Bernhard, 2010). Nous avons donc ajouté une contrainte supplémentaire lors de la fusion de deux familles qui consiste à mesurer la densité des liens entre communautés.

La densité des liens entre deux communautés A et B est définie de la manière suivante :

$$D_{AB} = \frac{\mathsf{nombre\_de\_liens}(A, B)}{|A| \times |B|} \quad [2]$$

où $\mathsf{nombre\_de\_liens}(A, B)$ est le nombre de liens reliant des nœuds de la communauté A à des liens de la communauté B, et $|A|$ et $|B|$ sont le nombre de nœuds

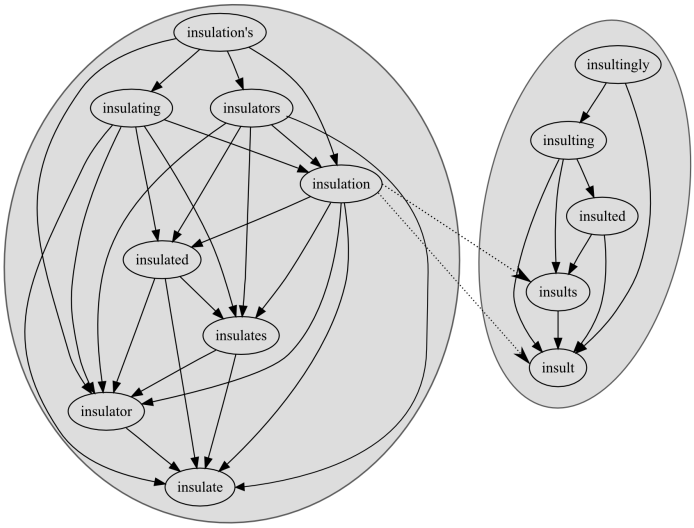

Figure 5. *Illustration du Newman Clustering sur un réseau lexical : deux communautés ont été détectées*

dans les communautés A et B, respectivement. La densité minimale est fixée par le paramètre d, compris entre 0 et 1.

5. Évaluation

Nous avons évalué les deux méthodes en les appliquant à l'anglais et à l'allemand. Bien que l'allemand et l'anglais appartiennent tous deux à la famille des langues germaniques, ces deux langues présentent des caractéristiques et différences intéressantes pour l'évaluation d'un système dans un contexte multilingue :

– l'allemand se caractérise par un système flexionnel bien plus complexe que l'anglais et possède notamment un système de déclinaison avec quatre cas (nominatif, accusatif, datif et génitif), trois genres grammaticaux (féminin, masculin et neutre) et deux nombres (singulier et pluriel) donnant lieu à des flexions rendant compte de propriétés grammaticales très diversifiées ;

– les deux langues ont recours au procédé de composition. Toutefois, les composés en allemand forment généralement un seul mot graphique, tandis que les composés anglais sont souvent séparés par des espaces ;

– l'allemand se caractérise par l'utilisation de deux types particuliers d'affixes que l'on ne retrouve pas en anglais : (i) les circumfixes en conjugaison, pour former le participe passé, et (ii), les interfixes (ou éléments de liaison, en allemand *Fugenelemente*) pour la formation des composés ;

– l'allemand et l'anglais présentent des cas de flexion par apophonie, conduisant à des changements de voyelle dans le radical, tels que *drink, drank, drunk* en anglais ou *trinken, trank, getrunken* en allemand.

Les systèmes ont également été appliqués à d'autres langues, lors d'études antérieures. Les résultats de MorphoClust en français sont comparables aux niveaux de performance obtenus en anglais (Bernhard, 2007). Par ailleurs, MorphoNet a obtenu des résultats encourageants en finnois et en turc lors de la compétition MorphoChallenge 2009, surpassant le système de référence *Morfessor* (Creutz et Lagus, 2002).

5.1. *Données d'évaluation*

Nous avons extrait des familles morphologiques de référence pour l'anglais et l'allemand à partir des segmentations contenues dans la base CELEX (Baayen *et al.*, 1995). CELEX fournit deux types d'informations morphologiques essentielles en trois langues (allemand, anglais et néerlandais) : les mots fléchis sont liés à leur lemme et les lemmes sont segmentés, de manière à pouvoir identifier la base minimale. La figure 6 décrit les liens morphologiques présents dans CELEX pour les mots anglais associés à la base *concern*.

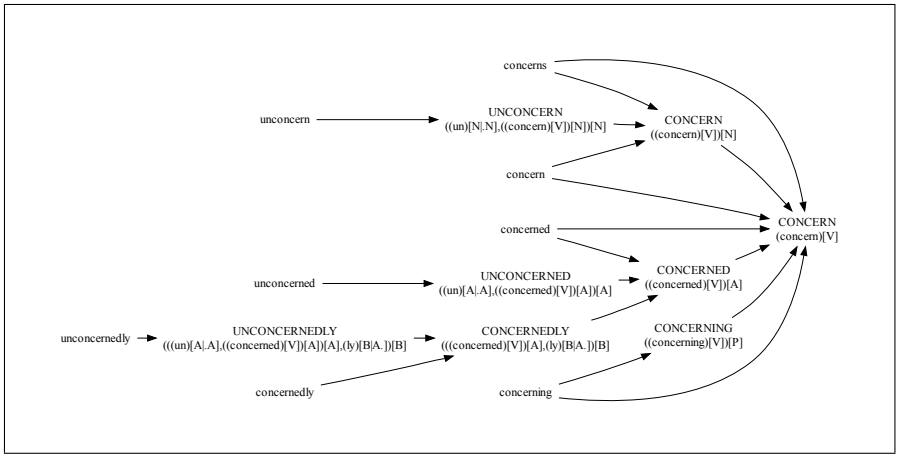

Figure 6. *Liens morphologiques dans CELEX. Les mots sont représentés en caractères minuscules tandis que les lemmes apparaissent en majuscules. Les décompositions proposées pour les lemmes sont également représentées*

Nous avons obtenus des familles morphologiques de référence en prenant en compte les opérations de flexion, de dérivation, de composition et de conversion. Nous obtenons ainsi 15 410 familles de référence pour l'anglais et 11 899 familles pour l'allemand. Nous avons également extrait les familles sans prendre en compte l'opération de composition, très fréquente en allemand, ce qui donne 21 686 familles de référence pour l'anglais et 29 368 pour l'allemand.

5.2. *Méthode d'évaluation*

Nous avons évalué les familles induites par rapport aux familles de référence en utilisant les mesures proposées par Schone et Jurafsky (2000) et Schone et Jurafsky (2001). La méthode d'évaluation consiste à faire la somme des nombres de mots corrects (vrais positifs C), insérés (faux positifs I) et supprimés (faux négatifs D) dans les familles morphologiques de tous les mots de la liste d'évaluation. Si X_w est l'ensemble des mots appartenant à la famille morphologique du mot w selon le système à évaluer et Y_w est l'ensemble des mots appartenant à la famille morphologique de w selon CELEX ou toute autre base de référence, alors :

$$C = \sum_{\forall w} \frac{|X_w \cap Y_w|}{|Y_w|} \quad [3]$$

$$D = \sum_{\forall w} \frac{|Y_w - (X_w \cap Y_w)|}{|Y_w|} \quad [4]$$

$$I = \sum_{\forall w} \frac{|X_w - (X_w \cap Y_w)|}{|Y_w|} \quad [5]$$

Lors du calcul de ces valeurs, seule l'intersection des mots de la base de référence et de la liste de mots analysés par le système est utilisée.

Par exemple, supposons que le système propose la famille suivante et que l'on cherche à évaluer la famille proposée pour le mot *concern* :

[*concern* ; *concerningly* ; *concerns* ; *concerted* ; *concert* ; *concerning* ; *concerned* ; *concern*].

Si *concerningly* n'appartient pas à la base de référence, alors :

X_w = [*concern* ; *concerns* ; *concerted* ; *concert* ; *concerning* ; *concerned* ; *concern*]

De plus, la famille de référence pour *concern* est la suivante : Y_w = [*concerned* ; *concerns* ; *concerning* ; *concern*]

Donc, pour le mot *concern* : $C_w = \frac{4}{4} = 1,0$, $D_w = \frac{0}{4} = 0$ et $I_w = \frac{3}{4} = 0,75$. On procède de même pour l'ensemble des mots et on calcule la somme de l'ensemble des valeurs de C_w, D_w et I_w pour obtenir C, D et I.

À partir de ces valeurs, il est également possible de calculer la précision, le rappel (et par conséquent la F-mesure) du système. La précision est égale à $C/(C + I)$ et le rappel à $C/(C + D)$.

Nous avons également utilisé la mesure de *pureté* qui est définie dans le domaine de la classification non supervisée (*clustering*) pour évaluer les clusters obtenus. La pureté est définie de la manière suivante (Manning *et al.*, 2008). Soient $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ l'ensemble des familles de mots obtenues et $\mathbb{C} = \{c_1, c_2, \ldots, c_j\}$ l'ensemble des familles attendues. La pureté d'une classification se calcule alors comme suit :

$$\text{pureté}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j| \qquad [6]$$

5.3. *Méthodes de référence*

Nous avons comparé les résultats obtenus par notre système à ceux de deux méthodes *baseline* : une première méthode basique qui consiste à considérer que chaque mot forme sa propre famille morphologique (*mots*) et l'algorithme de racinisation de Porter [5]. L'algorithme de racinisation utilise des règles définies manuellement et différentes pour chaque langue afin d'éliminer les suffixes à la fin des mots et identifier une racine commune à plusieurs mots qui constituent ainsi une famille morphologique.

5.4. *Résultats obtenus*

Les résultats obtenus par MorphoClust et par MorphoNet ainsi que par les méthodes de référence sont détaillés dans les tableaux 1 (avec composition) et 2 (sans composition). Les listes de mots sont tirées d'un lexique bilingue allemand-anglais, provenant du site FreeDict [6]. Le lexique comprend 61 655 mots anglais et 67 601 mots allemands. Les résultats de MorphoClust ont été obtenus pour une valeur du paramètre N du module d'apprentissage des affixes égale à 5 (pour une description de ce paramètre, voir (Bernhard, 2006)). N est un paramètre permettant de contrôler le processus d'apprentissage des affixes. Plus N est grand, plus le nombre de préfixes, de suffixes et, par conséquent, de signatures, est important. Les résultats indiqués pour MorphoNet correspondent à une valeur seuil de 0,1 pour la productivité et 0,1 pour la densité. Nous avons choisi de conserver les même valeurs de paramètres dans les deux langues afin d'étudier les performances de MorphoClust et MorphoNet dans le contexte d'une application directe à de nouvelles langues, sans réglage préalable des paramètres.

MorphoClust obtient généralement de meilleurs résultats que MorphoNet en terme de F-mesure dans les deux langues, compte tenu de son meilleur rappel. De plus, MorphoClust est globalement plus performant que la racinisation à base de règles, qui a une très bonne précision mais un plus faible rappel. MorphoNet est plus performant que MorphoClust uniquement en allemand, lorsque l'évaluation ne tient pas compte du procédé de composition. MorphoNet est également plus précis, avec une précision d'environ 80 à 90 %, proche de celle de la méthode de racinisation, et une meilleure pureté. Le rappel des deux méthodes est bien plus faible en allemand, notamment lorsque l'évaluation prend en compte les mots composés.

5. Nous avons utilisé le *wrapper* Python de Snowball, disponible sur le site suivant : `http://snowball.tartarus.org/download.php`

6. `http://www.freedict.org`.

		Précision	Rappel	F-mesure	Pureté
Allemand	Mots	100,00	12,42	22,09	1,00
	Racinisation	**96,73**	20,20	33,42	0,99
	MorphoClust	75,83	**31,53**	**44,54**	0,94
	MorphoNet	92,85	26,36	41,07	0,98
Anglais	Mots	100,00	22,41	36.61	1,00
	Racinisation	**95,62**	57,15	71,54	0,99
	MorphoClust	80,02	**67,53**	**73,25**	0,94
	MorphoNet	87,12	54,58	67,11	0,96

Tableau 1. *Résultats obtenus en allemand et en anglais (avec prise en compte des relations de composition).*

		Précision	Rappel	F-mesure	Pureté
Allemand	Mots	100,00	31,24	47,61	1,00
	Racinisation	**97,48**	46,36	62,83	0,99
	MorphoClust	50,79	**60,01**	55,02	0,88
	MorphoNet	82,18	55,15	**66,00**	0,95
Anglais	Mots	100,00	27,96	43,71	1,00
	Racinisation	**96,00**	66,55	78,60	0,99
	MorphoClust	80,72	**77,89**	**79,28**	0,94
	MorphoNet	86,58	64,02	73,61	0,96

Tableau 2. *Résultats obtenus en allemand et en anglais (sans prise en compte des relations de composition).*

Le tableau 3 liste quelques familles de mots identifiées en utilisant MorphoClust et MorphoNet. L'examen plus approfondi des résultats montre que différents types de variantes sont groupés par l'une et/ou l'autre méthode :

– variantes graphiques et orthographiques comme *tumor* (variante américaine) et *tumour* (variante britannique) ;

– variantes flexionnelles comme *friend* et *friends* ;

– variantes dérivationnelles suffixées comme *absorb* et *absorbency* et préfixées comme *countermeasure* et *measure* ;

– composés comme *Naturschutzgebiet* et *Gebiet* en allemand.

MorphoClust et MorphoNet commettent deux types d'erreur : sur-regroupement, c'est-à-dire le groupement de mots qui n'appartiennent pas tous à la même famille morphologique et sous-regroupement, c'est-à-dire l'absence de regroupement pour

MorphoClust	MorphoNet
absorber ; *absorbing* ; *absorb* ; absorbencies ; absorbency ; *absorbed* ; absorbable ; absorbabilities ; *absorbs* ; absorbent ; absorbability ; *absorbingly*	*absorb* ; *absorbed* ; *absorber* ; *absorbing* ; *absorbingly* ; *absorbs*
document ; documentaries ; *documenting* ; *documented* ; *documentation* ; documentary ; documents ; *documental* ; *documentable*	*document* ; *documentable* ; *documental* ; *documentation* ; *documented* ; *documenting* ; undocumented
friendless ; friendship ; friendlier ; friends ; friendliness ; friendliest ; friendships ; friendlessness ; *friendly* ; *friend*	*friend* ; *friendly* ; unfriendly
migrated ; migratory ; migrate ; *migrations* ; migrant ; migratorily ; *migration* ; migrates ; *migrational* ; migrating	*migration* ; immigration ; immigrations ; *migrational* ; *migrations* ; remigration
sparkle ; *sparkled* ; sparks ; *sparkler* ; *sparkles* ; *sparklers* ; sparked ; sparking ; spark ; *sparkling*	*sparkle* ; *sparkled* ; *sparkler* ; *sparklers* ; *sparkles* ; *sparkling*

Tableau 3. *Exemples de familles de mots obtenues en anglais. Les mots en italique sont présents dans les deux familles*

des mots appartenant à la même famille. La première erreur a pour conséquence de faire baisser la précision du système, tandis que la seconde conduit à une baisse du rappel. Par exemple, MorphoClust est incapable d'identifier le lien entre *undocumented* et *documented* compte tenu des contraintes imposées sur la longueur des préfixes ; celui-ci est toutefois correctement identifié par MorphoNet, qui est capable d'identifier des préfixes courts. MorphoNet est quant à lui limité par les règles de transformation acquises lors de la première étape et l'élimination des règles peu productives, ce qui explique son moindre rappel. Ainsi, les mots qui comportent plusieurs suffixes, comme *absorbabilities* ou *friendlessness*, ne sont pas correctement analysés par MorphoNet. Ce problème pourrait être résolu en adoptant une approche similaire à MorphoClust, qui consiste à acquérir itérativement de nouvelles règles morphologiques à partir des regroupements effectués.

6. Combinaison de MorphoClust et MorphoNet

Comme nous l'avons montré, MorphoClust et MorphoNet produisent des résultats différents en fonction de la langue et du type de procédé morphologique considéré. D'une manière générale, il est rare qu'une méthode d'analyse morphologique non supervisée ait des niveaux de performance équivalents dans toutes les langues. Par exemple, lors de la compétition MorphoChallenge 2009 (Kurimo *et al.*, 2010), le sys-

tème qui a obtenu les meilleurs résultats pour la langue arabe (Spiegler *et al.*, 2009), surpassant de loin les autres systèmes, s'est classé parmi les moins bons en anglais. Dans ce cas, il est généralement intéressant de combiner plusieurs méthodes, afin de profiter des avantages de chaque système et de gommer leurs inconvénients. Cette approche a été adoptée par Atwell et Roberts (2006) et Spiegler *et al.* (2009) pour la segmentation morphologique non supervisée et consiste en un mécanisme de vote entre systèmes pour choisir les positions de segmentation d'un mot.

Il a par ailleurs été montré que la combinaison translingue permet d'améliorer sensiblement les résultats de l'analyse morphologique non supervisée : la performance du système de segmentation morphologique proposé par Snyder et Barzilay (2008) augmente lorsque l'apprentissage est effectué de manière translingue, sur des paires de langues.

Nous avons donc cherché à combiner à la fois nos deux méthodes et les résultats obtenus en anglais et en allemand afin d'étudier l'effet de la combinaison sur les résultats.

6.1. *Méthode de combinaison*

La méthode de combinaison consiste à construire un réseau lexical à partir des familles morphologiques produites par MorphoNet et MorphoClust. Un lien est créé entre deux mots s'ils sont classés dans la même famille par au moins deux méthodes. La combinaison translingue utilise les traductions présentes dans le lexique bilingue Freedict d'où sont tirées les listes de mots traitées et y associe une contrainte de similarité graphique [7]. Par exemple, si *rückberechnen* et *berechnen* font partie de la même famille morphologique en allemand, et que l'on dispose des traductions suivantes en anglais pour *rückberechnen* [*recalculate*] et *berechnen* [*charge* ; *compute* ; *calculate* ; *bill*], alors le lien suivant est ajouté dans le réseau lexical combiné en anglais : *recalculate – calculate*. Les paires de traduction *recalculate – charge*, *recalculate – compute* et *recalculate – bill* ne sont pas validées car leur proximité graphique est insuffisante.

La combinaison translingue présente l'avantage d'ajouter des liens morphologiques dont la complexité est variable entre langues : par exemple, la flexion par umlaut en allemand que l'on retrouve dans la paire *Umstand – Umstände* est difficile à identifier de manière non supervisée car elle modifie la graphie de la base. Toutefois, l'équivalent traduit de cette paire de mots en anglais correspond à une forme de flexion très fréquente en anglais : *circumstance – circumstances*.

7. La proximité graphique est mesurée à l'aide de la méthode *quick_ratio* du module Python *difflib*.

6.2. *Résultats*

Les résultats de la combinaison translingue des méthodes sont détaillés dans les tableaux 4 et 5.

	Précision	Rappel	F-mesure	Pureté
Allemand	93,80	26,95	41,87	0,99
Anglais	94,69	53,34	68,24	0,98

Tableau 4. *Résultats de la combinaison translingue des méthodes (avec prise en compte de la composition)*

	Précision	Rappel	F-mesure	Pureté
Allemand	86,43	55,82	**67,83**	0,96
Anglais	94,77	62,71	75,48	0,98

Tableau 5. *Résultats de la combinaison translingue des méthodes (sans prise en compte de la composition)*

Les résultats de la combinaison sont mitigés : elle n'apporte aucune amélioration dans la plupart des cas, mais permet d'améliorer la F-mesure par rapport aux meilleurs résultats de MorphoNet en allemand lorsque l'on ignore les relations de composition. Dans ce cas, la combinaison semble avant tout utile pour identifier de manière fiable les relations de flexion et de dérivation. La combinaison profite donc à la langue la plus complexe, pour laquelle la marge de progression est la plus large : en effet, les résultats monolingues pour l'anglais sont déjà très bons. Il faut toutefois noter que la combinaison conduit à une nette amélioration des niveaux de précision et de pureté, y compris pour l'anglais, qui se rapprochent de ceux du raciniseur à base de règles. La plupart des applications, comme la recherche d'information, nécessitent une bonne précision. Dans ce cas, la combinaison de systèmes pourrait s'avérer nécessaire pour ne pas introduire un bruit trop important dans le système, conduisant à une baisse de la qualité des résultats obtenus.

7. Discussion des résultats

L'analyse détaillée des résultats obtenus par les deux systèmes, ainsi que leur combinaison, révèle un certain nombre de limites inhérentes aux hypothèse sous-jacentes à la conception de MorphoClust et MorphoNet. Dans cette section, nous détaillons et examinons ces limites afin de proposer des améliorations futures à apporter aux systèmes.

7.1. *Traitement différencié de la flexion, de la dérivation et de la composition*

Comme la grande majorité des systèmes d'analyse morphologique non supervisée, MorphoClust et MorphoNet n'opèrent pas de distinction explicite entre variantes graphémiques, flexionnelles, dérivationnelles ou compositionelles, qui sont toutes regroupées dans une même famille. Cette absence de traitement différencié peut s'avérer néfaste dans certains cas, dans la mesure où la proximité sémantique entre variantes flexionnelle et variantes dérivationnelles n'est pas la même. Si l'on reprend les exemples du tableau 3, il est bien évident que *document* est plus similaire à *documents* qu'à *documentary*, à la fois morphologiquement et sémantiquement. Nos méthodes rendent toutefois disponibles des informations statistiques permettant de pondérer la proximité morphologique et *a fortiori* sémantique entre mots : fréquence et productivité des préfixes, signatures et règles de transformation morphologique, ainsi que position d'un mot dans l'arbre de classification pour la méthode MorphoClust ou dans le réseau lexical dans le réseau MorphoNet. Ces indices constituent autant de possibilités à explorer pour obtenir des indices de similarité morphologique utilisables dans des applications concrètes.

7.2. *Cas particuliers de l'allomorphie, de l'apophonie et de la supplétion*

Une des difficultés majeures rencontrées par les systèmes d'analyse morphologique non supervisée concerne les cas d'allomorphie, d'apophonie et de supplétion. L'allomorphie correspond aux cas où plusieurs formes de surface correspondent au même morphème : ainsi, le préfixe anglais *in+* peut prendre les formes *im-*, *in-* ou *ir-*. L'apophonie est particulièrement fréquente en anglais et en allemand et fait référence aux phénomènes de changement de voyelle au sein d'un radical. Enfin, dans les cas de supplétion, les formes prises par un lexème ne sont pas prévisibles morphologiquement, par exemple « aller » et « vais » en français. Ces divers phénomènes nécessitent un traitement spécifique et sont difficiles à prendre en compte dans des systèmes reposant essentiellement sur la comparaison graphique des mots. MorphoClust et MorphoNet n'échappent pas à ce problème, qui conduit ainsi à la constitution de familles séparées pour les exemples suivants :

– cas de doublement de consonne en fin de radical en anglais : [*swim, swims*], [*swimmer, swimming*] ;

– ablaut en anglais : [*swim, swims*], [*swam*] [*swum*], et en allemand : [*sinkt ;sink ;sinkend ;sinken ;sinkende ;versinkt ;versinkend*], [*versank ;sanken ;sank*], [*gesunken*] ;

– umlaut en allemand : [*backt ;backe ;backen ;backte ;backend*], [*Bäckerei ;Bäckereien ;Bäcker*].

Il paraît difficile de prendre en compte ce type de phénomènes de manière totalement non supervisée, même s'il existe des tentatives dans ce sens, par exemple pour traiter l'apophonie (Demberg, 2007). Une solution intermédiaire, proposée par Tepper

et Xia (2010), consiste en l'utilisation de règles manuelles, appliquées aux analyses obtenues de manière non supervisée et dépendantes de la langue cible. Il s'agit de règles de réécriture orthographique, qui représentent les variations orthographiques régulières, conditionnées par un contexte particulier. L'utilisation de telles règles semble constituer une piste d'amélioration prometteuse, permettant d'intégrer à faible coût des informations morphologiquement motivées par la langue cible à des systèmes non supervisés. En particulier, de telles règles pourraient reposer sur les modèles d'espaces thématiques existant entre autre pour le français (Bonami et Boyé, 2003 ; Bonami et Boyé, 2005) ou l'espagnol (Boyé et Hofherr, 2006).

8. Conclusion et perspectives

Nous avons présenté deux méthodes non supervisées d'acquisition de familles morphologiques. La première, MorphoClust, procède par classification ascendante hiérarchique, tandis que la seconde, MorphoNet, repose sur la détection de communautés dans des réseaux lexicaux. Malgré la simplicité des méthodes, qui ne comportent pas de règles prédéfinies, les résultats obtenus en anglais et en allemand sont très bons : MorphoClust dépasse dans la majorité des cas la F-mesure obtenue par un raciniseur à base de règles heuristiques. L'analyse des résultats montre que les liens morphologiques découverts sont variés : flexion, dérivation et composition. De plus, l'apprentissage est effectué uniquement à partir d'une liste de mots et n'utilise aucune ressource externe. Les approches sont donc applicables à diverses langues.

Les perspectives d'amélioration des systèmes sont multiples. En effet, il est pour l'heure impossible pour un mot d'appartenir à deux voire à plusieurs familles différentes. Ceci est souhaitable pour les mots composés, qui font partie de plusieurs familles morphologiques. Il faudrait donc explicitement traiter des cas de composition. Les informations contextuelles, disponibles dans les corpus, pourraient également permettre d'améliorer encore les résultats, notamment en terme de précision en validant le fusionnement de deux familles en fonction de la similarité de leurs contextes d'occurrence. Enfin, nous n'avons pas testé la combinaison translingue des systèmes que pour l'anglais et l'allemand. Il serait intéressant d'évaluer les performances de cette combinaison avec un plus grand nombre de langues, voire des corpus de domaines différents. Une autre amélioration possible de la combinaison consisterait en l'utilisation d'algorithmes d'agrégation de clusters développés dans d'autres domaines, pour d'autres types de données.

Les systèmes de classification proposés peuvent être utilisés comme module préalable à une analyse morphologique plus poussée comme la segmentation. Nous avons, lors d'une étude précédente, proposé une méthode simple de génération d'analyses morphologiques à partir des regroupements opérés par MorphoNet (Bernhard, 2010).

Les perspectives applicatives directement envisageables concernent la recherche d'information et la classification de documents. En recherche d'information, les familles morphologiques peuvent être utilisées pour l'expansion de requêtes (Moreau

et Claveau, 2006). Pour la catégorisation de documents, les familles peuvent servir de descripteurs des documents à classer (Witschel et Biemann, 2006). Enfin, la combinaison translingue des méthodes permet d'identifier les règles de transformation morphologique équivalentes dans les langues cibles, qui pourraient servir à l'extension de lexiques multilingues utilisables en traduction automatique, similairement à Langlais et Patry (2007).

9. Bibliographie

Adamson G. W., Boreham J., « The use of an association measure based on character structure to identify semantically related pairs of words and document titles. », *Information Storage and Retrieval*, vol. 10, p. 253-260, 1974.

Atwell E., Roberts A., « Combinatory Hybrid Elementary Analysis of Text », *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, Venice, Italy, p. 41-45, April, 2006.

Baayen R. H., Piepenbrock R., Gulikers L., *The Celex Lexical Database (Release 2) [CD-ROM]*, Linguistic Data Consortium, Philadelphia, PA, 1995.

Baroni M., Matiasek J., Trost H., « Unsupervised discovery of morphologically related words based on orthographic and semantic similarity », *Proceedings of the ACL Workshop on Morphological and Phonological Learning 2002*, p. 48-57, 2002.

Bernhard D., « Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment », *in* M. Kurimo, M. Creutz, K. Lagus (eds), *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, Venice, Italy, p. 19-23, April, 2006.

Bernhard D., « Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique », *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles – TALN 2007 (communications orales)*, Toulouse, France, p. 367-376, 5-8 juin, 2007.

Bernhard D., « MorphoNet : Exploring the Use of Community Structure for Unsupervised Morpheme Analysis », *Multilingual Information Access Evaluation Vol. I, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Revised Selected Papers*, vol. 6241/2010 of *Lecture Notes in Computer Science*, Springer, Corfu, Greece, p. 598-608, September 30-October 2, 2010.

Bonami O., Boyé G., « Supplétion et classes flexionnelles », *Langages*, vol. 37, n° 152, p. 102-126, 2003.

Bonami O., Boyé G., « Construire le paradigme d'un adjectif », *Recherches linguistiques de Vincennes*, vol. 34, p. 77-98, 2005.

Boyé G., Hofherr P. C., « The structure of allomorphy in Spanish verbal inflection », *Cuadernos de Lingüística del Instituto Universitario Ortega y Gasset*, vol. 13, p. 9-24, 2006.

Cartoni B., « Lexical Morphology in Machine Translation : A Feasibility Study », *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece, p. 130-138, March, 2009.

Clauset A., Newman M. E. J., Moore C., « Finding community structure in very large networks », *Physical Review E*, 2004.

Creutz M., Lagus K., « Unsupervised Discovery of Morphemes », *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, p. 21-30, 2002.

Creutz M., Lagus K., « Inducing the Morphological Lexicon of a Natural Language from Unannotated Text », *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland, 15-17 June, 2005.

Deléger L., Namer F., Zweigenbaum P., « Morphosemantic parsing of medical compound words : Transferring a French analyzer to English », *International Journal of Medical Informatics*, vol. 78, p. 48-55, 2009.

Demberg V., « A Language-Independent Unsupervised Model for Morphological Segmentation », *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, p. 920-927, June, 2007.

Dorow B., Widdows D., Ling K., Eckmann J.-P., Sergi D., Moses E., « Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination », *MEANING-2005, 2nd Workshop organized by the MEANING Project*, 2005.

Freitag D., « Morphology Induction from Term Clusters », *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, Michigan, p. 128-135, June, 2005.

Gaussier E., « Unsupervised learning of derivational morphology from inflectional lexicons », *Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing*, University of Maryland, 1999.

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

Grabar N., Zweigenbaum P., « Acquisition automatique de connaissances morphologiques sur le vocabulaire médical », *in* P. Amsili (ed.), *Actes de TALN 1999*, Cargèse, p. 175-184, 12–17 juillet, 1999.

Hahn U., Honeck M., Shulz S., « Subword-Based Text Retrieval », *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, Big Island, Hawaii, January 06 - 09, 2003.

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

Hathout N., « From WordNet to CELEX : acquiring morphological links from dictionaries of synonyms », *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, p. 1478-1484, 2002.

Hathout N., « Exploiter la structure analogique du lexique construit : une approche computationnelle », *Cahiers de Lexicologie*, 2005.

Jacquemin C., « Guessing morphology from terms and corpora », *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 156 - 165, 1997.

Keshava S., Pitler E., « A Simpler, Intuitive Approach to Morpheme Induction », *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, Venice, Italy, p. 31-35, April, 2006.

Koskenniemi K., « A general computational model for word-form recognition and production », *Proceedings of the 22nd annual meeting of the Association for Computational Linguistics*, p. 178-181, 1984.

Kurimo M., Creutz M., Varjokallio M., Arisoy E., Saraclar M., « Unsupervised segmentation of words into morphemes – Challenge 2005 : An Introduction and Evaluation Report », *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, Venice, Italy, p. 1-11, 12 April, 2006.

Kurimo M., Virpioja S., Turunen V. T., Blackwood G. W., Byrne W., « Overview and Results of Morpho Challenge 2009. », *Multilingual Information Access Evaluation Vol. I, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Revised Selected Papers*, Lecture Notes in Computer Science, Corfu, Greece, September 30 - October 2, 2010.

Langlais P., Patry A., « Translating Unknown Words by Analogical Learning », *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 877-886, 2007.

Lavallée J.-F., Langlais P., « Morphological Acquisition by Formal Analogy », *Working Notes for the CLEF 2009 Workshop*, 2009.

Lepage Y., « Solving analogies on words : an algorithm », *Proceedings of the 17th international conference on Computational Linguistics*, p. 728-734, 1998.

Lovis C., Michel P.-A., Baud R., Scherrer J.-R., « Word Segmentation Processing : A Way to Exponentially Extend Medical Dictionaries », *in* R. A. Greenes, H. E. Peterson, D. J. Protti (eds), *Proceedings of the 8th World Congress on Medical Informatics*, p. 28-32, 1995.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

Matsuo Y., Sakaki T., Uchiyama K., Ishizuka M., « Graph-based Word Clustering using a Web Search Engine », *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, p. 542-550, July, 2006.

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

Moreau F., Claveau V., « Extension de requêtes par relations morphologiques acquises automatiquement », *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, p. 181-192, 2006.

Namer F., *Morphologie, Lexique et TAL : l'analyseur DériF*, TIC et Sciences cognitives, London : Hermes Sciences Publishing, 2009.

Newman M. E. J., « Fast algorithm for detecting community structure in networks », *Phys. Rev. E*, 2004.

Newman M. E. J., « Modularity and community structure in networks », *PNAS*, vol. 103, n° 23, p. 8577-8582, 2006.

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

Pratt A. W., Pacak M. G., « Automated processing of medical English », *Proceedings of the 1969 conference on Computational linguistics*, p. 1-23, 1969.

Schone P., Jurafsky D., « Knowledge-Free Induction of Morphology Using Latent Semantic Analysis », *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, Lisbon, Portugal, September, 2000.

Schone P., Jurafsky D., « Knowledge-Free Induction of Inflectional Morphologies », *Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics*, p. 1-9, 2001.

Snyder B., Barzilay R., « Unsupervised Multilingual Learning for Morphological Segmentation », *Proceedings of ACL-08*, Columbus, Ohio, p. 737-745, June, 2008.

Spiegler S., Golenia B., Flach P., « PROMODES : A Probabilistic Generative Model for Word Decomposition », *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

Stroppa N., Yvon F., « Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie », *Traitement Automatique des Langues*, vol. 47, n° 1, p. 33-59, 2006.

ten Hacken P., Lüdeling A., « Word Formation in Computational Linguistics », *Proceedings of TALN 2002*, vol. 2, p. 61-87, 2002.

Tepper M., Xia F., « Inducing Morphemes Using Light Knowledge », *Journal of ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 9, n° 3, p. 1-38, 2010.

van den Bosch A., Daelemans W., « Memory-based morphological analysis », *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, p. 285-292, 1999.

van Dongen S., Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht, May, 2000.

Witschel H. F., Biemann C., « Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization », *in* S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, vol. 1, Joensuu, Finland, p. 197-204, 2006.

Zweigenbaum P., Grabar N., « Liens morphologiques et structuration de terminologie », *Actes de IC 2000 : Ingénierie des Connaissances*, p. 325-334, 2000.