Une méthode non-supervisée pour la segmentation morphologique et l'apprentissage de morphotactique à l'aide de processus de Pitman-Yor

Kevin Löser Alexandre Allauzen
Université Paris-Sud & LIMSI-CNRS, 91403 Orsay, France
loser@limsi.fr, allauzen@limsi.fr

RÉSUMÉ

Cet article présente un modèle bayésien non-paramétrique pour la segmentation morphologique non supervisée. Ce modèle semi-markovien s'appuie sur des classes latentes de morphèmes afin de modéliser les caractéristiques morphotactiques du lexique, et son caractère non-paramétrique lui permet de s'adapter aux données sans avoir à spécifier à l'avance l'inventaire des morphèmes ainsi que leurs classes. Un processus de Pitman-Yor est utilisé comme *a priori* sur les paramètres afin d'éviter une convergence vers des solutions dégénérées et inadaptées au traitemement automatique des langues. Les résultats expérimentaux montrent la pertinence des segmentations obtenues pour le turc et l'anglais. Une étude qualitative montre également que le modèle infère une morphotactique linguistiquement pertinente, sans le recours à des connaissances expertes quant à la structure morphologique des formes de mots.

ABSTRACT

An unsupervised method for joint morphological segmentation and morphotactics learning using Pitman-Yor processes

We describe a non-parametric bayesian model for unsupervised morphological segmentation. This semi-Markov model uses latent morph classes in order to model the morphotactics of the lexicon, and its non-parametric framework allows to automatically adapt the number of classes and the morph inventory to the complexity of the data. A Pitman-Yor process is used as a prior on the model parameters in order to avoid convergence towards degenerate solutions that would be unsuited to natural language processing tasks. Experimental results show the quality of the segmentations produced by our method on Turkish and English. Moreover, a qualitative study reveals that our model infers linguistically sound morphotactics, without relying on expert knowledge regarding the word structure.

MOTS-CLÉS : Morphologie, Apprentissage non-supervisé, Modèles bayésiens non-paramétriques.

KEYWORDS: Morphology, Unsupervised learning, Nonparametric bayesian models.

1 Introduction

L'analyse morphologique est une tâche importante dans de nombreux domaines du traitement automatique des langues comme la reconnaissance automatique de la parole (Vergyri *et al.*, 2004; Xiang *et al.*, 2006), la traduction automatique (Lee, 2004; Goldwater & McClosky, 2005) ou la recherche d'informations (Claveau, 2012). Cette analyse revêt souvent la forme d'une segmentation morphologique, qui consiste à décomposer un mot en une suite de morphèmes. Ce type de décomposition permet alors de réduire la variété lexicale, en particulier pour les langues considérées comme morphologiquement riches, où la taille du vocabulaire pose des problèmes de couverture et de généralisation aux approches usuelles en traitement automatique des langues [1].

Le manque de ressources décrivant la morphologie de nombreuses langues et domaines a favorisé le développement de modèles d'apprentissage non-supervisé permettant l'acquisition automatique d'analyses morphologiques à partir de textes. En l'absence de données étiquetées, l'apprentissage non-supervisé doit inclure dans la modélisation de la tâche un biais permettant d'infléchir la recherche d'une solution dans une direction appropriée. Cette inflexion peut s'induire de manière efficace grâce à des heuristiques (Bernhard, 2006; Keshava & Pitler, 2006) ou par la sélection de morphèmes candidats sur des critères comme l'entropie entre les n-grammes adjacents (Harris, 1955). Par ailleurs, le principe de longueur de description minimale (ou MDL pour *Minimum Description Length*) propose un cadre d'apprentissage pour orienter la sélection d'un modèle vers une représentation compacte du lexique (Brent *et al.*, 1995; Goldsmith, 2001; Creutz & Lagus, 2007). Ce cadre s'applique également aux modèles log-linéaires (Poon *et al.*, 2009; Narasimhan *et al.*, 2015) dans lesquels les connaissances expertes sont exprimées sous forme de caractéristiques.

Une autre manière d'envisager le principe de longueur de description minimale est de se placer dans un cadre bayésien non-paramétrique. La distribution *a priori* sur les paramètres permet alors de contrôler la complexité du modèle en favorisant des distributions parcimonieuses. Dans (Goldwater *et al.*, 2006), les auteurs proposent d'utiliser un processus de Pitman-Yor en tant qu'*a priori*. Le caractère non-paramétrique permet alors de contourner la difficile définition d'un inventaire heuristique et fini des morphèmes. Ces travaux ont été par la suite étendus à un modèle morphologique basé sur les grammaires hors-contexte (Sirts & Goldwater, 2013). Néanmoins, ces travaux s'appuient sur une définition préalable de la morphotactique de la langue et donc des types de morphèmes (*e.g.* racine et suffixe). Par contre, (Snyder & Barzilay, 2008) introduit la notion de classe abstraite de morphèmes, mais dans un cadre multilingue. L'objectif dans ces travaux est d'apprendre conjointement dans plusieurs langues le processus de segmentation morphologique, les classes permettent l'appariement de morphèmes dans différentes langues.

Dans cet article, nous proposons un modèle bayésien non-paramétrique pour la segmentation morphologique. La particularité de ce modèle décrit à la section 2 réside dans l'introduction de classes latentes de morphèmes dans un modèle semi-markovien qui étend les travaux récents de (Uchiumi *et al.*, 2015). Le cadre non-paramétrique permet au modèle d'inférer par lui même l'inventaire des morphèmes, mais aussi les classes de morphèmes qui lui permettent de modéliser les caractéristiques morphotactiques du lexique. Le caractère parcimonieux des distributions rencontrées dans le traitement automatique des langues est contrôlé grâce au processus de Pitman-Yor, utilisé comme *a priori* sur les paramètres du modèle semi-markovien permettant l'induction de la segmentation en morphèmes et de classes de morphèmes. L'inférence des analyses morphologiques est décrite à la

1. Le caractère parcimonieux des données textuelles, allié à une forte production morphologique implique que la plupart des mots ont des occurences faibles dans des corpus qui peuvent-être de taille trop réduite.

section 3. Elle s'effectue en marginalisant les paramètres grâce à un échantillonnage de Gibbs. Enfin, les expériences relatées à la section 4 utilisent le cadre d'évaluation du *Morpho Challenge* sur les langues turque et anglaise. Les résultats montrent la capacité de ce modèle à inférer des segmentations morphologiques pertinentes.

2 Apprentissage bayésien et le processus de Pitman-Yor

L'apprentissage non-supervisé s'appuie sur la définition d'un modèle génératif, lorsqu'on se place dans un cadre probabiliste. Partant de l'hypothèse d'un ensemble de variables inobservées, ou latentes, le modèle définit la probabilité jointe des observations (les données) et des variables latentes. Parmi les exemples connus, citons les modèles d'alignement mot-à-mot (Brown *et al.*, 1990) où les variables latentes sont les liens d'alignement entre les mots d'une paire de phrases parallèles, et les modèles de thèmes (*topic models*) (Hofmann, 2001; Blei *et al.*, 2003) qui cherchent à expliquer une collection de documents avec un ensemble de thèmes qu'il s'agit de découvrir. Dans le cadre de cet article, l'objectif est de pouvoir analyser un lexique à partir d'un ensemble de morphèmes et de proposer une segmentation des mots de ce lexique grâce à ces unités induites des données.

Considérons dans un premier temps un modèle de segmentation simple pour lequel un mot m est segmenté en une séquence de K morphèmes $\boldsymbol{\mu} = \mu_1, ..., \mu_K$. Chaque morphème appartient à un ensemble fini, le vocabulaire $\mathcal{V}$. Une paramétrisation possible est alors d'envisager un modèle n$-gramme$ sur les morphèmes et l'algorithme EM (*Expectation-Maximization*) peut être utilisé pour optimiser la vraisemblance. Néanmoins, trois difficultés se posent. La première est qu'il faut définir au préalable l'ensemble des morphèmes $\mathcal{V}$ et que celui-ci doit-être fini. De plus, le critère du maximum de vraisemblance ne prend pas en compte la complexité des modèles. Par conséquent, la solution privilégiée sera une solution dégénérée typique du sur-apprentissage : chaque mot donne lieu à un morphème ; cette solution revient alors à un apprentissage par cœur des exemples d'apprentissage, incapable de se généraliser à de nouvelles formes de mots. Ce comportement a déjà été documenté et le lecteur peut se reporter par exemple à l'article (Goldwater *et al.*, 2009) pour plus de détails. Enfin, cette modélisation ne tient pas compte des caractéristiques morphotactiques, à savoir que les séquences de morphèmes peuvent suivre des règles morphologiques propres à la langue, comme par exemple le fait que certains mots puissent se décomposer selon le schéma *préfixe* $\rightarrow$ *racine* $\rightarrow$ *suffixe*. Ce schéma fréquent correspond à une séquence de catégories de morphèmes qu'il est important de prendre en compte afin de donner au modèle la possibilité d'analyser les observations de manière compacte et ainsi augmenter sa capacité de généralisation.

Dans cet article, nous proposons d'envisager l'analyse morphologique grâce à un modèle bayésien non-paramétrique et hiérarchique afin de résoudre ces trois limitations. Dans un cadre bayésien, la distribution *a priori* sur les paramètres alliée à leur marginalisation permet d'orienter l'apprentissage vers des distributions parcimonieuses appropriées au traitement des langues, tout en laissant au modèle la possibilité de s'adapter à la compléxité inhérente de la tâche. Le modèle que nous proposons utilise des classes de morphèmes afin de rendre compte des propriétés morphotactiques de la langue. Enfin, le caractère non-paramétrique se justife par la liberté qu'il procure au modèle d'adapter le nombre de classes et de morphèmes à la complexité des données.

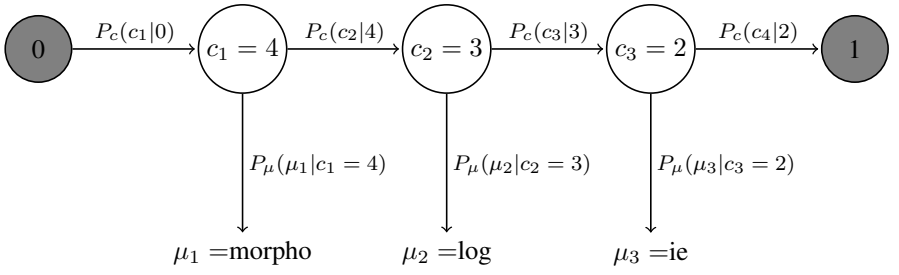

FIGURE 1 – Exemple du processus génératif d'une analyse morphologique. Chaque distribution impliquée (P_c et P_μ) est issue d'un processus de Pitman-Yor hiérarchique. Dans le cas de la figure, P_c est une distribution bi-gramme.

2.1 Histoire générative

Une manière de définir notre modèle est de relater son histoire générative, c'est-à-dire la manière dont le modèle explique la formation des mots observés.

1. En premier lieu, une séquence de classes morphologiques est engendrée de la manière suivante : partant de la classe initiale $c_0 = (0)$, chaque classe c_i est tirée selon la distribution d'un modèle n-gramme $P_c(c_i|c_{i-n+1}^{i-1})$, où c_{i-n+1}^{i-1} désigne le contexte constitué des $n-1$ classes précédentes $c_{i-n+1} \dots c_{i-1}$. La génération de la séquence $\boldsymbol{c} = c_0, c_1, ..., c_K$ se termine lors de l'émission de la classe $c_K = (1)$.

2. Chaque classe c_i de cette séquence engendre indépendemment une chaîne de caractères (un morphème) μ_i selon la loi $P_\mu(\mu_i|c_i)$.

Ce processus génératif est représenté à la figure 1. Il s'appuie sur les deux distributions conditionnelles P_c and P_μ afin d'engendrer ce que nous appellerons par la suite une *analyse* morphologique, soit le couple $(\boldsymbol{c} = c_0 \dots c_K, \boldsymbol{\mu} = \mu_0 \dots \mu_K)$. La probabilité d'une analyse peut alors se calculer selon :

$$P_a(\mathbf{c}, \boldsymbol{\mu}) = \prod_{i=0}^{K} P_c(c_i|c_{i-n+1}^{i-1}) \cdot \prod_{i=1}^{K-1} P_\mu(\mu_i|c_i).$$

Deux raisons justifient le choix de cette histoire générative. Tout d'abord, l'introduction des classes permet de modéliser les caractéristiques morphotactiques de la langue. Alors que nous pourrions imposer des patrons morphotactiques comme *préfixe* → *racine* → *suffixe*, il nous parait préférable de laisser au modèle toute latitude afin de découvrir ce type de règles de successions et sans figer *a priori* le nombre de classes. Afin d'illustrer ce choix, en langue allemande, les règles de combinaisons peuvent être très variables, avec par exemple l'enchaînement de plusieurs préfixes ($\texttt{ab}_{pre} - \texttt{ge}_{pre} - \texttt{lehn}_{stem} - \texttt{t}_{suf}$), ou la présence de racines multiples ($\texttt{um}_{pre} - \texttt{welt}_{stem} - \texttt{ver}_{pre} - \texttt{schmutz}_{stem} - \texttt{ung}_{suf}$). Nous faisons l'hypothèse que chaque classe engendre un morphème indépendemment de son contexte car d'une part, cela rend les calculs plus efficaces et d'autre part, cette hypothèse reflète la définition usuelle du morphème en linguistique : une unité minimale et indépendante (Harris, 1955).

2.2 Le processus de Pitman-Yor hiérarchique

L'histoire générative définit un modèle semi-markovien (Levinson, 1986) faisant intervenir deux distributions : la chaîne de Markov modélisant la séquence de classes P_c, et la distribution P_μ permettant la génération d'un morphème connaissant sa classe. Dans un cadre bayésien, ces distributions et les paramètres associés sont considérés comme des variables aléatoires issues de distributions dites *a priori*. Afin de pouvoir imposer une forme particulière respectant les spécificités de la langue, nous utilisons comme distribution *a priori* un processus de Pitman-Yor (Pitman & Yor, 1997), dont la caractéristique est de favoriser l'émergence de solutions parcimonieuses, *i.e.* des distributions expliquant les données avec peu de morphèmes et de classes. Nous retrouvons ainsi le comportement souhaitable du principe de longueur de description minimale déjà utilisé en analyse morphologique. En d'autres termes, ce type d'*a priori* contraint l'apprentissage à trouver un compromis entre deux situations extrêmes : le cas du sur-apprentissage, où chaque mot est considéré comme un morphème ; ou au contraire le cas du sous-apprentissage dans lequel chaque caractère est un morphème.

Considérons la distribution $P_c(c_i|c_{i-n+1}^{i-1})$ qui pour un contexte c_{i-n+1}^{i-1} donné est une distribution multinomiale sur l'ensemble des classes possibles. Le processus de Pitman-Yor (PYP) (Teh, 2006) est utilisé comme *a priori* sur les paramètres :

$$P_c(c_i|c_{i-n+1}^{i-1}) \sim PYP(\theta_c, d_c, G),$$

où les hyperparamètres θ_c et d_c représentent respectivement le terme de concentration et de décompte. G est une distribution dite de base permettant la création d'une nouvelle classe. Le tirage d'un PYP est un ensemble, potentiellement infini mais dénombrable, de réels positifs qui somment à un, donc une distribution de probabilité. Il y a donc un modèle de Pitman-Yor par distribution n-gramme, soit par contexte c_{i-n+1}^{i-1} possible. Les hyperparamètres peuvent être propres à chacun des contextes, mais nous avons choisi de tous les fixer à la même valeur.

Une propriété interessante du PYP est la possibilité d'inférer la probabilité d'une classe c_i dans son contexte à partir de l'observation préalable d'un ensemble d'analyses notée a_{-k}. Cette estimation utilise la représentation dite du « processus du restaurant chinois » ou CRP. La probabilité $P_c(c_i|c_{i-n+1}^{i-1}, a_{-k})$ se calcule ainsi :

$$\frac{n_{a_{-k}}\left(c_i \leftarrow c_{i-n+1}^{i-1}\right) - d_c \cdot n_{a_{-k}}\left(c_i \overset{b}{\leftarrow} c_{i-n+1}^{i-1}\right) + \left(\theta_c + n_{a_{-k}}\left(c_i \overset{b}{\leftarrow} c_{i-n+1}^{i-1}\right) \cdot d_c\right) G(c_i)}{\sum_{c \in \mathcal{C}} n_{a_{-k}}\left(c \leftarrow c_{i-n+1}^{i-1}\right) + \theta_c}. \quad (1)$$

Afin d'expliciter cette formule, nous allons détailler le processus de génération conditionnellement à un ensemble d'analyses préalables a_{-k} de la classe c_i à la suite des classes c_{i-1} et c_{i-2}, soit dans le cas d'un modèle trigramme (dans cette explication, le numérateur est omis) :

– Si la classe c_i a déjà été observée, elle peut être engendrée avec une probabilité proportionnelle à

$$n_{a_{-k}}(c_i \leftarrow c_{i-1}c_{i-2}) - d_c \cdot n_{a_{-k}}(c_i \overset{b}{\leftarrow} c_{i-1}c_{i-2}),$$

où $n_{a_{-k}}(c_i \leftarrow c_{i-1}c_{i-2})$ désigne le nombre de fois que c_i apparaît à la suite de c_{i-1}, c_{i-2} dans les analyses a_{-k}, et $n_{a_{-k}}(c_i \overset{b}{\leftarrow} c_{i-1}c_{i-2})$ le nombre de fois qu'elle est apparue en faisant appel à la distribution de base.

– Une nouvelle classe c_i peut également être engendrée par la distribution de base G. Dans ce cas, sa probabilité est proportionnelle à

$$\theta_c + d_c \cdot n_{a_{-k}}(c_i \overset{b}{\leftarrow} c_{i-1}c_{i-2})G(c_i).$$

Modèle de classe morphologique	$P_c(c_i\|c_{i-n+1}^{i-1})$	$\sim PYP(\theta_c, d_c, P_c(c_i\|c_{i-n+2}^{i-1}))$	pour $n \geq 2$
	$P_c(c_i\|\emptyset)$	$\sim PYP(\theta_c, d_c, \delta(c^*))$	pour $n = 1$
Modèle de génération de morphème	$P_\mu(\mu_i\|c_i)$	$\sim PYP(\theta_\mu, d_\mu, P_\mu(\mu_i))$	
	$P_\mu(\mu_i)$	$\sim PYP(\theta_p, d_p, P^{car}(\mu_i))$	

TABLE 1 – Définition du modèle d'analyse morphologique, où δ représente la distribution de Dirac et $c*$ l'identifiant d'une nouvelle classe. Les deux modèles sont issus d'un processus de Pitman-Yor hiérarchique.

Si la classe dans son contexte est observée dans a_{-k}, il existe deux manière de l'engendrer. Le choix est fait par un tirage aléatoire. De plus, dans la formule 1 les paramètres de la distribution n-gramme n'apparaissent pas. En effet cette estimation résulte de la marginalisation des paramètres [2], ce qui implique qu'ils n'apparaissent pas explicitement dans les calculs.

Dans le cas d'un modèle n-gramme, un choix possible pour la distribution de base G est d'utiliser la distribution d'ordre inférieur. Dans ce cas d'un modèle trigramme, G peut être la distribution bigramme $P_c(c_i|c_{i-1})$. Il est alors possible de récursivement reproduire le processus : $P_c(c_i|c_{i-1})$ est à son tour considéré comme un tirage dans un PYP avec une distribution de base $P_c(c_i|\emptyset)$. Nous obtenons ainsi un *processus de Pitman-Yor hiérarchique*. Pour $P_c(c_i|\emptyset)$, nous avons fait le choix d'une distribution où toute la masse de probabilité est allouée à l'identifiant d'une nouvelle classe (une distribution de Dirac). Ce choix permet de favoriser la création d'une nouvelle classe lorsque cette distribution est échantillonnée.

Les distributions $P_\mu(\mu_i|c_i)$ sont également engendrées par un PYP d'hyperparamètres θ_μ et d_μ et de distribution de base $P_\mu(\mu_i)$. Cette distribution de base est elle-même issue d'un PYP d'hyperparamètres θ_p et d_p et de base distribution P^{car} où P^{car} est un modèle n-gramme sur les caractères (définissant une distribution de probabilités sur les chaînes de caractères favorisant celles qui ressemblent à un morphème de la langue traitée). Dans le cadre de nos expériences, nous avons choisi pour P^{car} un simple modèle trigramme de caractères appris sur le corpus d'apprentissage. Les hyperparamètres θ_p et d_p ainsi que la distribution P^{car} sont partagés par toutes les classes.

3 Inférence bayésienne

Dans l'inférence bayésienne, les paramètres des modèles sont considérés comme des variables aléatoires. Ils ne sont présents qu'implicitement car l'objectif est de prédire les analyses morphologiques du lexique. Pour cela, la distribution que l'on cherche à estimer est $\mathbb{P}(a_1, \ldots, a_{|L|}|w_1, \ldots, w_{|L|})$, soit la distribution jointe des analyses $a_1, \ldots, a_{|L|}$ connaissant le lexique $L = \{w_1, \ldots, w_{|L|}\}$. Cela revient à intégrer sur l'espace des paramètres ce qui explique leur absence. Cette distribution n'est pas analytiquement accessible, mais des méthodes d'inférence approchée existent. Dans cet article nous effectuons un échantillonnage de Gibbs, où chaque analyse est tour à tour rééchantillonnée conditionnellement aux autres. En d'autres termes, pour chaque $k = 1, \ldots, |L|$, il est possible de produire des échantillons de la distribution $\mathbb{P}(a_k|a_{-k}, w_1, \ldots, w_{|L|}) = \mathbb{P}(a_k|w_k, a_{-k})$. Sous certaines conditions, cette procédure d'échantillonnage produit des échantillons sous la distribution d'intérêt.

2. En d'autres termes, la prédiction d'une classe dans son contexte se fait en intégrant sur l'ensemble des paramètres la probabilité jointe de la classe et des paramètres. Pour plus de détail, le lecteur peut se reporter à (MacKay, 2002) ou (Teh, 2006)

3.1 Conditionnement des modèles de classes et de morphèmes par rapport aux autres analyses

Tout d'abord, exprimons le modèle de séquence de classes et le modèle d'émission de morphèmes conditionnellement aux autres analyses a_{-k}. Pour alléger les notations, le conditionnement par rapport à a_{-k} est omis dans le reste de cette section. En reprenant les notations de la section 2.2, pour $n \geq 2$, la probabilité $P_c(c_i|c_{i-n+1}^{i-1})$ s'estime de la manière suivante :

$$\frac{n_{a_{-k}}\left(c_i \leftarrow c_{i-n+1}^{i-1}\right) - d_c \cdot n_{a_{-k}}\left(c_i \overset{b}{\leftarrow} c_{i-n+1}^{i-1}\right) + \left(\theta_c + n_{a_{-k}}\left(c_i \overset{b}{\leftarrow} c_{i-n+1}^{i-1}\right) \cdot d_c\right) P_c(c_i|c_{i-n+2}^{i-1})}{\sum_{c \in \mathcal{C}} n_{a_{-k}}\left(c \leftarrow c_{i-n+1}^{i-1}\right) + \theta_c}$$

Nous rappelons que la distribution de base de $P_c(c_i|c_{i-n+1}^{i-1})$ est $P_c(c_i|c_{i-n+2}^{i-1})$ pour $n \geq 2$. Pour $n = 1$ la distribution de base est une distribution de Dirac qui alloue toute sa masse de probabilité à l'identifiant d'une nouvelle classe, afin que le nombre de classes engendrées s'adapte aux données. Ainsi pour $n = 1$, nous avons l'expression suivante :

$$P_c(c_i|\emptyset) = \frac{n_{a_{-k}}\left(c_i \leftarrow \emptyset\right) - d_c \cdot n_{a_{-k}}\left(c_i \overset{b}{\leftarrow} \emptyset\right)}{\sum_{c \in \mathcal{C}} n_{a_{-k}}\left(c \leftarrow \emptyset\right) + \theta_c}$$

si c_i est une classe observée parmi les autres analyses, ou :

$$P_c(c_i|\emptyset) = \frac{\theta_c + n_{a_{-k}}\left(c_i \overset{b}{\leftarrow} \emptyset\right) \cdot d_c}{\sum_{c \in \mathcal{C}} n_{a_{-k}}\left(c \leftarrow \emptyset\right) + \theta_c}$$

si c_i est une classe jamais rencontrée auparavant. Cette probabilité résiduelle permet la création de nouvelles classes. De même, notons $n_{a_{-k}}\left(c_i \rightarrow \mu_i\right)$ le nombre de fois que le morphème μ_i a été engendré sous la classe c_i parmi les analyses a_k, et $n_{a_{-k}}\left(c_i \overset{b}{\rightarrow} \mu_i\right)$ le nombre de fois que cette génération a eu lieu en faisant appel à la distribution de base. Nous rappelons que la distribution de base de $P_\mu(\mu_i|c_i)$ est une distribution $P_\mu(\mu_i)$ partagée par toutes les classes, qui est elle-même engendrée par un processus de Pitman-Yor ayant pour distribution de base un modèle trigramme de caractères P^{car}. Notons $n_{a_{-k}}\left(\emptyset \rightarrow \mu_i\right)$ le nombre de fois que $P_\mu(\mu_i)$ a été échantillonné parmi les analyses a_{-k} (c'est-à-dire : $n_{a_{-k}}\left(\emptyset \rightarrow \mu_i\right) = \sum_{c \in \text{class IDs}} n_{a_{-k}}\left(c \overset{b}{\rightarrow} \mu_i\right)$), et notons $n_{a_{-k}}\left(\emptyset \overset{b}{\rightarrow} \mu_i\right)$ le nombre de fois qu'il a été échantillonné en faisant appel à la distribution de base P^{car}. La probabilité d'engendrer un morphème μ_i sous une classe c_i est donnée par une formule du CRP analogue à ci-dessus :

$$P_\mu(\mu_i|c_i) = \frac{n_{a_{-k}}\left(c_i \rightarrow \mu_i\right) - d_\mu \cdot n_{a_{-k}}\left(c_i \overset{b}{\rightarrow} \mu_i\right) + \left(\theta_\mu + d_\mu \cdot n_{a_{-k}}\left(c_i \overset{b}{\rightarrow} \mu_i\right)\right) P_\mu(\mu_i)}{\sum_{\mu \in \mathcal{V}} n_{a_{-k}}\left(c_i \rightarrow \mu\right) + \theta_\mu}$$

$$P_\mu(c_i) = \frac{n_{a_{-k}}\left(\emptyset \rightarrow \mu_i\right) - d_p \cdot n_{a_{-k}}\left(\emptyset \overset{b}{\rightarrow} \mu_i\right) + \left(\theta_p + d_p \cdot n_{a_{-k}}\left(\emptyset \overset{b}{\rightarrow} \mu_i\right)\right) P^{\text{car}}(\mu_i)}{\sum_{\mu \in \mathcal{V}} n_{a_{-k}}\left(\emptyset \rightarrow \mu\right) + \theta_p}$$

3.2 Échantillonnage des analyses

Nous décrivons à présent comment échantillonner une analyse étant donné une forme de mot w, et conditionnellement à toutes les analyses précédentes a_{-k}. Le nombre d'analyses pour une forme

est exponentiellement grand : si une analyse éclate le mot en σ segments, il y a $n_{classes}^{\sigma}$ différentes manières d'assigner une classe à chaque segment. Donc le nombre total d'analyses possibles est : $\sum_{\sigma=1}^{|w|} \binom{|w|-1}{\sigma-1} \cdot n_{classes}^{\sigma} = n_{classes} \cdot (1 + n_{classes})^{|w|-1}$. Afin d'éviter que l'échantillonnage soit de complexité exponentielle, nous utilisons donc une méthode de programmation dynamique similaire à l'algorithme forward-backward utilisé pour les modèles de Markov cachés (Rabiner, 1989). Pour faciliter la présentation, nous présenterons uniquement le cas où le modèle de classes est d'ordre 1, mais il est facile de généraliser au cas de modèles d'ordres supérieurs.

Tout d'abord, fixons quelques notations : $w_j^{j'}$ représente la sous-chaîne de caractères de w délimitée par les positions j et j' (telle que $w = w_0^{|w|}$), et $\mathcal{C}$ représente l'ensemble des classes (y compris un identifiant pour une nouvelle classe vide). Dans ce qui suit, toutes les distributions de probabilités seront conditionnées aux analyses précédentes, selon les formules données à la section 3.1.

Tout d'abord, calculons la probabilité d'engendrer un suffixe $w_j^{|w|}$ de w suivant la classe c_i. Nous noterons $\beta(j, c_i)$ cette probabilité. Pour la calculer, il nous faut sommer sur toutes les classes c_{i+1} suivant c_i, et sur toutes les positions j' possibles pour la prochaine limite de morphème :

$$\beta(j, c_i) = \mathbb{P}(w_j^{|w|} | c_i, a_{-k}) = \sum_{c_{i+1} \in \mathcal{C}} \sum_{j'=j+1}^{|w|} P_c(c_{i+1}|c_i) \cdot P_\mu(w_j^{j'}|c_{i+1}) \cdot \beta(j', c_{i+1})$$

avec le cas limite suivant :

$$\beta(|w|, c_i) = P_c\left((1)|c_i\right)$$

Ces deux formules montrent que β peut efficacement se calculer avec une complexité $\mathcal{O}\left(|w|^2 \cdot |\mathcal{C}|^2\right)$ à l'aide d'une méthode de programmation dynamique, en remplissant une matrice bidimensionnelle indexée par $c \in \mathcal{C}$ et $j = |w|, |w| - 1, \ldots, 0$.

Afin d'échantillonner une analyse du mot w, nous engendrons tout d'abord la classe initiale $c_0 = (0)$. Puis à chaque itération, nous utilisons la procédure d'échantillonnage suivante. Supposons que l'analyse partiellement générée est donnée par la suite de classes $c_0, \ldots, c_{i-1}$ et les positions de limites de morphèmes $j_0 = 0, j_1, j_2, \ldots, j_{i-1}$. Autrement dit, $w_0^{j_{i-1}}$ a déjà été segmenté en un segment $w_{j_0}^{j_1}$ assigné à la classe c_1, suivi d'un segment $w_{j_1}^{j_2}$ assigné à la classe c_2, et ainsi de suite jusqu'au dernier segment engendré $w_{j_{i-2}}^{j_{i-1}}$ assigné à la classe c_{i-1}. Alors la probabilité de placer la prochaine limite de morphème à la position j_i et d'assigner le segment résultant $w_{j_{i-1}}^{j_i}$ à la classe c_i est donnée par :

$$\mathbb{P}(c_i, j_i | c_{i-1}, j_{i-1}) = \frac{1}{Z} \cdot P_c(c_i|c_{i-1}) \cdot P_\mu(w_{j_{i-1}}^{j_i}|c_i) \cdot \beta(j_i, c_i)$$

où Z est une constante de normalisation obtenue en sommant ces probabilités sur $c_i \in \mathcal{C}$ et $j_i \in \{j_{i-1} + 1, \ldots, |w|\}$. Par conséquent, nous répétons l'échantillonnage de cette distribution afin d'engendrer la position de la prochaine limite de morphème et la prochaine classe de l'analyse, jusqu'à atteindre la position $|w|$ (fin du mot), auquel point la classe terminale (1) sera engendrée.

Notons qu'en échantillonnant de la sorte, l'algorithme doit conserver un compte des niveaux de la hiérarchie de Pitman-Yor auxquels les échantillons ont été émis. Par exemple, si une classe c_i a été engendrée après c_{i-1}, elle a pu être engendrée soit par la distribution de niveau supérieur (d'ordre 1), soit en faisant appel à la distribution de niveau inférieur (d'ordre 0). De même, si une classe c_i

engendre un segment $w_j^{j'}$, ceci peut être expliqué soit comme un échantillon direct de la distribution de morphèmes conditionnelle aux classes $P_\mu(w_j^{j'}|c_i)$, soit comme un échantillon qui a été produit en faisant appel à la distribution de base (partagée par toutes les classes) $P_\mu(w_j^{j'})$. Une telle comptabilité est nécessaire afin d'évaluer les termes $n\left({}_{c_i}\overset{b}{\leftarrow}{}_{c_{i-m+1}^{i-1}}\right)$ et $n\left({}_{c_i}\overset{b}{\rightarrow}{}_{\mu_i}\right)$ (voir section 3.1).

Au début du processus d'échantillonnage, le modèle ne dispose d'aucune analyse. Plus précisément, étant donnée notre liste de formes de mots $w_1, \ldots, w_L$, nous avons échantillonné une première analyse a_1 pour w_1, puis une analyse a_2 pour w_2 conditionnellement à l'analyse précédente a_1, puis une analyse a_3 pour w_3 conditionnellement à toutes les analyses précédentes a_1 et a_2, et ainsi de suite. Puis une fois que les analyses $a_1 \ldots a_L$ ont été échantillonnées pour tous les mots, nous avons itéré de nouveau sur l'ensemble du corpus : c'est-à-dire rééchantillonné a_1 conditionnellement à $a_2, \ldots, a_L$, puis rééchantillonné a_2 conditionnellement à $a_1, a_3, \ldots, a_L$, et ainsi de suite.

4 Expériences

Nous avons évalué notre méthode sur les données de la compétition *Morpho Challenge* 2005. Les données sont une longue liste de formes de mots non annotées afin d'entraîner le modèle, ainsi qu'un ensemble plus restreint de formes de mots annotées avec des segmentations produites par des annotateurs. Comme métrique d'évaluation, nous calculons la F-mesure sur toutes les limites de morphèmes prédites par le modèle comparées à celles fournies par les annotateurs. Chaque fois qu'une forme a été estimée décomposable de plusieurs manières différentes par les annotateurs, plusieurs segmentations ont été fournies, et seule la segmentation pour laquelle le modèle a la F-mesure la plus élevée est retenue.

4.1 Données

Nous avons évalué sur deux langues : le turc et l'anglais. Le turc est une langue agglutinante et morphologiquement riche qui est particulièrement adéquate au traitement par des modèles de morphologie concaténative. La liste de mots turcs est extraite de textes de prose et de publications collectées du Web, de journaux d'information et et de sports. Elle contient 582923 formes dans le corpus d'entraînement et 774 formes dans le corpus de test. L'anglais a une morphologie bien plus simple, mais nous avons choisi cette langue afin d'évaluer qualitativement la pertinence de la morphotactique apprise par notre modèle, puisqu'il n'y avait pas de locuteurs turcs parmi les auteurs. La liste de mots anglais est extraite de publications et de romans du projet Gutenberg, d'un échantillon du corpus Gigaword anglais, ainsi que de l'ensemble du corpus Brown. Elle contient 167377 formes dans le corpus d'entraînement et 532 formes dans le corpus de test.

4.2 Résultats quantitatifs

Nous avons consigné nos résultats dans le tableau 2, et nous avons également inclus les résultats de l'algorithme de référence utilisé dans la compétition (*Morfessor Categories-MAP* (Creutz & Lagus, 2005)) ainsi que les trois compétiteurs ayant obtenu les meilleures F-mesures pour chaque langue. Afin de choisir les hyperparamètres, nous avons fixé d_c et d_μ à $1/6$, et nous avons fait varier θ_c et θ_μ

TABLE 2 – Résultats sur la tâche de segmentation non supervisée du *Morpho Challenge*

Corpus turc		
Méthode	F-mesure	Précision
Morfessor Categories-MAP	70.7	77.5
Bernhard (b)	65.3	65.4
Bordag "Comb"	57.0	79.9
Choudri & Dang "Summaa"	55.4	58.8
PYP-SHMM ($\theta_c = \theta_\mu = 1, d_c, d_\mu = 1/6$)	58.9	72.7
PYP-SHMM ($\theta_c = \theta_\mu = 10, d_c, d_\mu = 1/6$)	57.7	56.1
Corpus anglais		
Méthode	F-mesure	Précision
Morfessor Categories-MAP	66.2	85.1
Pitler & Keshava "RePortS"	76.8	76.2
Bernhard (a)	66.6	67.7
Bernhard (b)	62.4	55.2
PYP-SHMM ($\theta_c = 1, \theta_\mu = 5, d_c, d_\mu = 1/6$)	64.0	67.8
PYP-SHMM ($\theta_c = \theta_\mu = 10, d_c, d_\mu = 1/6$)	63.2	64.2

dans l'ensemble $\{1, 5, 10\}$, en conservant les deux réglages de (θ_c, θ_μ) ayant obtenu les meilleurs résultats sur l'ensemble de test.

Résumons les méthodes employés par les compétiteurs. La méthode *Bordag "Comb"* consiste à induire dans un premier temps des clusters de mots partageant des profils de coocurrence similaire (et nécessite par conséquent un corpus de texte et pas juste une liste de mots). Une fois ces clusters induits (l'idée étant que les mots au sein d'un cluster partageront des propriétés morphologiques similaires), une heuristique de "letter successor varieties" (Hafer & Weiss, 1974) est appliquée afin de conjecturer des limites de morphèmes. Ensuite, sur le corpus segmenté, un classifieur est appris afin de pouvoir généraliser et effectuer des segmentations sur des nouvelles formes de mots.

Les méthodes *Bernhard* (Bernhard, 2006) sont des méthodes heuristiques se basant sur une morphotactique comprenant quatre catégories de morphèmes : *préfixes, racines, suffixes* et *éléments de liaison*. Une segmentation préliminaire est obtenue en examinant les variations des probabilités de transition entre les parties du mots (idée similaire à celle des "letter successor varieties"). Puis dans chaque segmentation, un segment est identifié comme racine selon des règles heuristiques, et les segments précédant et suivant ce mot sont ajoutés respectivement à une liste de préfixes et de suffixes. Ensuite, un ensemble de racines est acquis en examinant toutes les manières possibles de soustraire des préfixes ou des suffixes aux mots de la liste, et en soumettant ces racines potentielles à des critères heuristiques de validation. De même la méthode *Pitler & Keshava* (Keshava & Pitler, 2006) utilise des heuristiques à base de "letter successor varieties" afin d'induire des listes de préfixes et de suffixes.

Bien que notre méthode soit loin de surpasser *Morfessor*, nous voyons que pour chacune des langues, elle est compétitive avec les meilleurs algorithmes qui ont été soumis. De plus, un argument en faveur de notre méthode est son cadre mathématique unifié : en effet, plutôt que d'utiliser une chaîne d'heuristiques, et elle se base seulement sur l'échantillonage d'un modèle génératif conditionnellement à des formes observées. De plus, notre méthode n'incorpore aucun a priori morphotactique, et ne présuppose pas que les mots observés devraient être formés par exemple par l'enchaînement d'un préfixe, d'une racine et d'un suffixe.

TABLE 3 – Illustration des classes morphologiques induites pour les données anglaises. La concentration est mesurée comme le nombre de morphèmes qui, classés du plus fréquent au moins fréquent, représentent 80% de la masse de probabilité au sein de la classe.

Classe	Concentration	Exemple de morphèmes générés selon $P(\mu\|c)$
« suffixes »	40	`s e ed ing es 's a y us er o ers ly ia ation i um an en ic on is ian s' ate ie e's os ius ies in man ness ted t`
« racines »	13256	`ver rezh chang sag vers fritz tan va mutt ba deduc sha flor omo ding be domestic b kay e voic greg def physi commander sancti drown clar psorias condemn figur fist cowhid whiten orchestr`
« suffixes »	634	`or's oo ed ers ling ing land icaut en um in air ors i ia ots ius zen eck atic ich ita man o idas onne st er bil t n al es fold`
« préfixes »	133	`un re de s be in a dis con over pro ma l' inter pre k co d' ca out mis la bi ka pa di en sub sa trans car under mc du`
	57	`freckl inflict herme drill gould ocular soli possessive liff awkward scissor keill adiabenian china buridan mann`

4.3 Résultats qualitatifs

A présent, nous examinons la morphologie apprise par notre modèle sur le corpus anglais afin de juger de sa pertinence linguistique. Cinq classes morphologiques ont été induites (en plus des classes initiale et terminale). Parmi toutes les analyses effectuées, les séquences de classes les plus fréquentes sont (en fréquence décroissante) `3-4`, `3-2`, `3-4-2`, `5-3-4`, `5-3-2`, `3`, `3,4,4`, `5-3`, et à elles seules, ces huit séquences représentent 94% des analyses. Cela laisse supposer que sans qu'on lui ait fourni d'a priori, notre modèle a bien inféré une morphotactique de type *préfixe* (classe `5`) → *racine* (classe `3`) → *suffixes* (classes `2` et `4`). Parmi les séquences de classes moins fréquentes, on trouve par exemple des doubles racines (par exemple, parmi les analyses inférées : $\texttt{un}_5\texttt{philo}_3\texttt{soph}_3\texttt{ically}_4$, $\texttt{house}_3\texttt{warm}_3\texttt{ing}_2$, $\texttt{photo}_3\texttt{electr}_3\texttt{onic}_4$). La classe `6` n'apparaît que dans des séquences `6` et `6-2`, et la plupart des analyses 6-2 sont un nom propre suivi du `'s` de possession (parmi les analyses inférées : $\texttt{liffe}_3\texttt{'s}_2$, $\texttt{tirana}_6\texttt{'s}_6$).

L'analyse des classes de morphèmes induites (tableau 3) confirme bien notre constatation ci-dessus. Nous avons consigné dans ce tableau quelques exemples de morphèmes pour chaque classe induite, en échantillonnant ces exemples selon la distribution de probabilité $P(\mu|c)$. On observe en effet de nombreux suffixes dans les classes `2` et `4` (la classe `2` étant plus restreinte et contenant surtout des suffixes de nature grammaticale comme les marques du pluriel, du passé, du gérondif...), et de nombreux préfixes dans la classe 5. De plus, le modèle a appris à distinguer entre des classes morphologiques fermées (les classes de préfixes et suffixes, relativement restreintes) et une classe morphologique ouverte : la classe `3`, contenant de nombreuses racines. La classe `6` est plus difficile à interpréter, mais on remarque qu'elle contient de nombreux noms propres.

5 Conclusion

Nous avons présenté un modèle génératif modélisant la formation des mots d'un lexique à l'aide d'un modèle semi-Markov caché (SHMM). Le modèle semi-Markov caché est formé d'une hiérarchie à deux niveaux combinant un modèle de séquences de classes morphologiques à un modèle de génération de morphèmes conditionnellement aux classes. Cette structure permet d'analyser un lexique de manière non-supervisée, et à trois niveaux. Premièrement, les transitions correspondent à des limites entre morphèmes, et permettent d'obtenir des segmentations morphologiques. Deuxièmement, la génération de morphèmes étant conditionnée par les classes, notre modèle à la capacité de regrouper les morphèmes induits en classes morphologiques, comme par exemple en préfixes, suffixes et racines. Troisièmement, le modèle markovien de séquences de classes constitue un modèle de la morphotactique sous-jacente au lexique. De plus, nous avons soumis notre SHMM à un *a priori* de Pitman-Yor, ce qui procure deux avantages. D'une part, le processus de Pitman-Yor impose une contrainte de parcimonie aux distributions de probabilité de notre modèle, ce qui garantit que l'inférence ne convergera pas vers des explications triviales du lexique, où chaque mot ou chaque caractère serait considéré comme un morphème. D'autre part, il permet de conserver un nombre potentiellement infini de classes, et d'adapter le nombre de classes effectivement induites à la structure des données. Nous avons ensuite montré en quoi l'utilisation d'une méthode simple de programmation dynamique permet de surmonter des problèmes d'intractabilité dans l'échantillonnage du modèle. Après avoir évalué notre méthode sur les données en anglais et en turc du *Morpho Challenge* 2005, nous avons constaté que celle-ci est compétitive avec les meilleurs méthodes soumises. De plus, une analyse qualitative de notre méthode sur l'anglais montre que la morphotactique inférée est linguistiquement pertinente, et que sans qu'aucun *a priori* quant à la structure de la langue n'ait été fourni, notre modèle apprend par lui-même à distinguer l'existence de préfixes, de racines et de suffixes, ainsi qu'à faire la distinction entre des classes morphologiques fermées et des classes ouvertes.

Nous envisageons dans des travaux ultérieurs d'analyser la performance de notre modèle lorsque ses hyperparamètres ne seront pas fixés, mais inférés automatiquement. En particulier, il serait intéressant de voir si le modèle parvient à adapter automatiquement ses hyperparamètres de concentration en fonction du "degré d'ouverture" d'une classe, en inférant par exemple une concentration élevée pour les classes de préfixes et de suffixes, et une concentration faible pour les classes de racines. Une autre piste de recherche serait de mesurer l'impact de l'initialisation sur notre modèle, et de vérifier si par exemple en l'initialisant avec des segmentations produites par une autre méthode déjà efficace, il serait capable d'améliorer encore davantage ces segmentations. Enfin, une dernière piste de recherche serait de modifier la structure du modèle de génération de morphèmes afin de prendre en compte des phénomènes de morphologie non-concaténative, comme par exemple l'entrelacement entre schèmes vocaliques et racines consonantiques dans les langues sémitiques.

Références

BERNHARD D. (2006). Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of 2nd Pascal Challenges Workshop*, p. 19–24.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

BRENT M. R., MURTHY S. K. & LUNDBERG A. (1995). Discovering morphemic suffixes a case study in MDL induction. In *In Fifth International Workshop on AI and Statistics,Ft*, p. 264–271.

BROWN P. F., COCKE J., PIETRA S. D., PIETRA V. J. D., JELINEK F., LAFFERTY J. D., MERCER R. L. & ROOSSIN P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, **16**(2), 79–85.

CLAVEAU V. (2012). Unsupervised and semi-supervised morphological analysis for information retrieval in the biomedical domain. In *Proceedings of COLING 2012*, p. 629–646, Mumbai, India : The COLING 2012 Organizing Committee.

CREUTZ M. & LAGUS K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In *In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*.

CREUTZ M. & LAGUS K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, **4**(1), 3 :1–3 :34.

GOLDSMITH J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, **27**(2), 153–198.

GOLDWATER S., GRIFFITHS T. L. & JOHNSON M. (2006). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18*.

GOLDWATER S., GRIFFITHS T. L. & JOHNSON M. (2009). A Bayesian framework for word segmentation : Exploring the effects of context. *Cognition*, **112**(1), 21–54.

GOLDWATER S. & MCCLOSKY D. (2005). Improving statistical MT through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 676–683, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

HAFER M. A. & WEISS S. F. (1974). Word segmentation by letter successor varieties. *Information storage and retrieval*, **10**(11), 371–385.

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

HOFMANN T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, **42**(1-2), 177–196.

KESHAVA S. & PITLER E. (2006). A simpler, intuitive approach to morpheme induction. In *Proceedings of 2nd Pascal Challenges Workshop*, p. 31–35.

LEE Y.-S. (2004). Morphological analysis for statistical machine translation. In D. M. SUSAN DUMAIS & S. ROUKOS, Eds., *HLT-NAACL 2004 : Short Papers*, p. 57–60, Boston, Massachusetts, USA : Association for Computational Linguistics.

LEVINSON S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, **1**(1), 29 – 45.

MACKAY D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. New York, NY, USA : Cambridge University Press.

NARASIMHAN K., BARZILAY R. & JAAKKOLA T. (2015). An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, **3**, 157–167.

PITMAN J. & YOR M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **25**(2), 855–900.

POON H., CHERRY C. & TOUTANOVA K. (2009). Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 209–217.

RABINER L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, p. 257–286.

SIRTS K. & GOLDWATER S. (2013). Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, **1**, 255–266.

SNYDER B. & BARZILAY R. (2008). Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08 : HLT*, p. 737–745, Columbus, Ohio : Association for Computational Linguistics.

TEH Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 985–992.

UCHIUMI K., TSUKAHARA H. & MOCHIHASHI D. (2015). Inducing word and part-of-speech with Pitman-Yor hidden semi-Markov models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1774–1782, Beijing, China : Association for Computational Linguistics.

VERGYRI D., KIRCHHOFF K., DUH K. & STOLCKE A. (2004). Morphology-based language modeling for Arabic speech recognition. In *In Proc. of ICSLP*, p. 2245–2248.

XIANG B., NGUYEN K., NGUYEN L., SCHWARTZ R. & MAKHOUL J. (2006). Morphological decomposition for Arabic broadcast news transcription. In *In Proc. of ICASSP*, volume 1, p. I–I.