Identification de l'Origine Linguistique : le Cas des Noms Propres

Erika Valencia et **François Yvon**
École Nationale Supérieure des Télécommunications
Département Informatique - 46 rue Barrault, 75 013 Paris
{valencia,yvon}@inf.enst.fr

Résumé

Nous nous intéressons dans cet article au problème de la détermination de l'origine linguistique d'une chaîne alphabétique *isolée*, dans le cadre particulier de la transcription orthographique--phonétique des noms propres. Les noms propres contribuent en effet pour une large part aux erreurs de prononciation commises par les systèmes de synthèse vocale, principalement du fait de l'abondance de noms propres empruntés à d'autres langues, pour lesquels les règles mises en œuvre dans le module de transcription conduisent souvent à des prononciations erronées (Hochberg *et al.* 1990; Métayer & Vincent 91).
Réduire le volume de ces erreurs nécessite de disposer d'un mécanisme d'identification de l'origine linguistique des noms propres. De tels mécanismes ont été proposés, reposant soit sur des règles (Carlson, Granström, & Lindström 1989; Belhoula 1993), soit sur des techniques à base d'analogie (Golding 1991), soit encore sur des modèles statistiques (Vitale 1991; Gustafson 1996).
Dans la lignée de ces derniers auteurs, nous explorons différents modèles statistiques permettant de discrimer des langues, en nous inspirant en particulier des techniques utilisées pour les tâches d'identification et de vérification du locuteur. Nous montrons, à partir d'expériences comparatives portant sur des noms propres de villes et de personnes que ces techniques permettent d'améliorer sensiblement la qualité des prédictions par rapport au modèle de Vitale.

1. Introduction

1.1. La synthèse des noms propres

Les noms propres contribuent pour une large part aux erreurs commises par les systèmes de transcription, l'une des raisons clairement identifiée de ces erreurs (Hochberg *et al.* 1990; Métayer & Vincent 91) étant l'abondance de noms propres empruntés à d'autres langues, dont la prononciation ne respecte pas les règles mises en œuvre dans le système de transcription. Ainsi, un nom d'origine italienne tel que *Benedetti* risque, s'il n'est pas préalablement reconnu comme étranger, d'être lu /bənədetti/ par un synthétiseur du français. Ces erreurs sont particulièrement pénalisantes pour l'acceptabilité d'un certain nombre de services de synthèse, qui requièrent, par exemple, que le nom de la personne utilisant le service soit (correctement) synthétisé.
Pour contourner cette difficulté, diverses techniques ont été proposées, faisant appel soit à des modèles statistiques (Vitale 1991; Gustafson 1996), soit à des techniques à base d'analogie (Golding 1991), ou encore à base de règles s'appuyant principalement sur des listes de préfixes/suffixes caractéristiques (Carlson, Granström, & Lindström 1989; Belhoula 1993). Ces techniques sont mises en œuvre, en amont de l'étape de transcription, pour assigner à chaque nom propre une origine. Suivant leur origine, les noms propres sont ensuite aiguillés vers des jeux de règles spécifiques, qui prennent en compte les particularités de chaque langue, de manière à parvenir à une prononciation qui, quoique différant de celle du nom propre dans sa langue d'origine (par exemple à cause de différences entre les inventaires segmentaux), intègre tout ou partie des conventions tacites qui s'appliquent aux emprunts (telle qu'en français, l'absence de nasalisation des voyelles dans les noms identifiés comme étant d'origine anglaise).
Les travaux préliminaires que nous présentons dans cet article visent principalement à tester des modèles statistiques éprouvés, de manière à étalonner en quelque sorte ce problème. Nous comptons ultérieurement mettre en œuvre d'autres techniques de discrimation de leangues, par exemple des techniques analogiques telles que celles présentées dans (Yvon 1996).

1.2. Description formelle du problème

Si l'on se place dans un cadre probabiliste, et que l'on dispose d'un modèle de chacune des langues à identifier, ce problème consiste à trouver, dans un ensemble de langues $\{L_i\}, i = 1 \ldots N$, celle qui est la plus probable conditionnellement à l'observation du nom propre x, en d'autres termes celle qui maximise $P(L_i|x)$. Ce programme se ramène, par application de la formule de Bayes, à trouver L^* rendant maximum $P(x|L^*)P(L^*)$. Ce problème de choix dans un ensemble fini de langues est analogue à celui de l'identification vocale de l'identité d'un locuteur dans un ensemble clos de locuteurs potentiels.
Dans la mesure où, de plus, le nom propre est susceptible d'appartenir à une langue qui n'appartient pas à l'ensemble de celles que l'on veut identifier, cette règle de décision doit être additionnellement complétée par un mécanisme de rejet, qui permet de refuser de classer un nom dans une des langues à reconnaître. Il existe différentes manières, plus ou moins *ad-hoc*, de mettre en œuvre un tel mécanisme :

- une première manière de procéder est de fixer directment un seuil pour $max_{L_i}(P(L_i|x))$, en-dessous duquel on préfère ne pas affecter le nom à une des langues à reconnaître. Nous parlerons dans ce cas de *rejet par seuillage*;
- une seconde manière de procéder, qui requiert l'utilisation d' exemples négatifs (i.e. qui n'appartiennent à aucune des langues à reconnaître), consiste à apprendre un modèle de ces exemples négatifs. Un mot

sera alors rejeté si son appartenance à l'ensemble des exemples négatifs est plus probable que son appartenance à une des langues à identifier. Nous appelerons ce procédé *rejet par caractérisation négative*.

– enfin, il est possible de mettre en œuvre, langue par langue, des tests du rapport de vraisemblance : $\frac{P(x|L_i)}{P(x|\overline{L_i})}$, où $\overline{L_i}$ représente le complémentaire de L_i dans l'ensemble des langues possibles. Ceci permet d'accepter ou de rejeter, langue par langue, et avec des niveaux de confiance prédéfinis, l'hypothèse que le nom identifié appartient ou non à la langue. Cette technique de rejet sera par la suite dénommée *rejet par test LR*. Dans un tel cadre, deux questions supplémentaires se posent. Tout d'abord, celle de la modélisation de $\overline{L_i}$:s'il est relativement aisé de modéliser une langue à partir d'exemples positifs, quels exemples négatifs doit on utiliser pour modéliser le mieux possible son complémentaire ? Ensuite, celle de la procédure de décision à associer aux tests : si le test du rapport de vraisemblance ne permet pas de conserver une et une seule langue, comment parvenir malgré tout à la meilleure solution possible ?

Il est enfin naturellement possible de combiner de différentes manières ces stratégies de rejet, par exemple en se donnant une caractérisation négative, et en mettant en œuvre des tests du rapport de vraisemblance, etc.

Dans ce qui suit, un des éléments importants de notre contribution est d'apporter des éléments de réponse sur les conséquences de l'adpotion de l'une ou l'autre de ces stratégies de rejet, en présentant les résultats d'études comparatives de performances de reconnaissance.

1.3. Les données d'apprentissage

Les données sur lesquelles nous nous sommes appuyés pour apprendre et tester nos modèles sont issues du projet LRE Onomastica (Schmidt *et al.* 1993; Jack *et al.* 1995). Nous disposons ainsi pour dix langues européennes (allemand, anglais, danois, espagnol, français, italien, néerlandais, norvégien, portugais, et suédois) de corpus allant de quelques milliers à quelques dizaines de milliers de noms propres (toponymes, prénoms, et patronymes). Pour les expériences décrites dans les sections qui suivent, nous n'avons pas tiré profit des informations, pourtant présentes dans ces corpus, concernant les représentations phonétiques des noms propres. Il est également à noter que les corpus d'Onomastica ne contiennent pas de données fréquentielles, dont nos modèles auraient pu tirer grand profit. On conservera néanmoins à l'esprit que de telles données pourraient, à peu de frais, être utilisées pour enrichir les modèles statistiques que nous décrivons.

2. Les différents modèles

2.1. Le modèle de Vitale

Le modèle qui servira de base de référence est le modèle proposé par Vitale 1991. Ce modèle se fonde sur la caractérisation d'une langue par l'ensemble des n-grammes apparaissant dans un échantillon de la langue. Les données fréquentielles relevées lors de l'apprentissage permettent de dériver directement, pour chaque n-gramme s du dictionnaire, une estimation de la probabilité de la langue conditionnellement à l'observation de la séquence, selon (on note $c_L(s)$ la fréquence d'apparition de s dans la langue s, et d_L le nombre de n-grammes différents attestés dans L) :

$$P(L|s) = \frac{\frac{c_L(s)}{d_L}}{\sum_{i=1}^{i=N} \frac{c_{L_i}(s)}{d_{L_i}}} \qquad (1)$$

Étant donné un mot inconnu $x = x_1 \ldots x_l$ à identifier, ces probabilités permettent de dériver, pour chaque langue L_i, un ensemble de $l - n + 1$ valeurs, correspondant à :

$$\{P(L_i|x_j \ldots x_{j+n-1}), \text{pour } j = 1 \ldots l - n + 1\}$$

Vitale en dérive un score global $S_x(L_i)$ en réalisant tout simplement la somme de ces valeurs. La langue qui est retenue est alors L^* telle que :

$$L^* = \underset{L_i}{\text{argmax}}\ S_x(L_i) \qquad (2)$$

Dans les expériences décrites ci-dessous, nous avons utilisé une version légèrement modifiée de cette procédure. En effet, se pose pour les langues européennes un problème que Vitale n'avait pas rencontré sur ces données américaines, à savoir que les différentes langues à distinguer ont des alphabets graphémiques de tailles inégales, suivant le nombre plus ou moins grand de caractères diacrisés qu'elles utilisent. À nombre d'exemples constants, un n-gramme est donc proportionnellement plus fréquent dans une langue qui utilise un alphabet plus restreint. Nous avons donc remplacé dans (1) d_L par $c_L = \sum_{\{s\}} c_L(s)$, ce qui permet de moins faire dépendre les probabilités $P(L|s)$ de la taille des alphabets.

Enfin, dans les expériences concernant les patronymes (voir la section , pour lesquelles il était nécessaire de faire appel à un mécanisme de rejet, nous avons uniquement utilisé le rejet par caractérisation négative, en apprenant un modèle de tous les exemples négatifs disponibles.

2.2. Les modèles de langage

Les autres mécanismes d'identification considérés reposent sur des modèles de langage, qui permettent de calculer, pour chaque chaîne graphique $x = x_1..x_l$, sa probabilité d'occurrence dans la langue L : $P(x|L)$. Cette probabilité est calculée en utilisant des modèles de n-grammes (Jelinek 1990) selon [1] :

$$P(x) = P(x_1)P(x_2|x_1) \ldots \prod_{i=n}^{i=l} P(x_i|x_{i-1} \ldots x_{i-n+1})$$

Cette famille de modèle se heurte, dès que n dépasse 2 ou 3, au problème de l'estimation correcte des paramètres correspondants aux probabilités d'occurence de n-grammes qui

1. Pour alléger les notations, nous omettons dans les formules qui suivent le conditionnement par rapport à la langue, et écrivons $P(x)$ pour $P(x|L)$.

n'apparaîssent pas dans le corpus d'apprentissage. Utiliser l'estimateur du maximum de vraisemblance pour ces séquences conduit à leur affecter une probabilité d'occurrence nulle, ce qui annule également la probabilité de tout mot les contenant. Dans la mesure où de tels n-grammes sont pourtant susceptibles d'apparaître (le corpus d'apprentissage ne contient jamais toutes les séquences possibles), cette manière de procéder conduit, dans notre contexte, à des taux de non-reconnaissance bien trop importants.

De nombreuses méthodes ont été proposées pour contourner ce problème, qui consistent soit à interpoler systématiquement les probabilités n-gramme par des combinaisons linéaires des probabilités des séquences de taille plus faible (modèles d'interpolation, voir (Federico *et al.* 1995) pour une revue), soit à remplacer les probabilités des n-grammes de fréquence nulle par une fraction de la probabilité des n-grammes de taille inférieure (modèle de Backing-Off (Katz 1987)).

Nous avons testé, dans le cadre de cette tâche, plusieurs schèmes d'interpolation différents, qui ont conduit à des résultats dans l'ensemble très comparables. Les valeurs numériques que nous donnons dans la partie évaluation correspondent à une interpolation linéaire-empirique, qui consiste à calculer récursivement la probabilité n-gramme de la manière suivante (on note $h^k(x_i)$ l'histoire de taille k du symbole x_i, i.e. la séquence $x_{i-k}\ldots x_{i-1}$, $c(x)$ la fréquence observée de la séquence x et $d(x*)$ le nombre de séquences différentes commençant par x) :

$P(x_i|h^{n-1}(x_i)) =$

- $\frac{c(h^{n-1}(x_i)x_i)}{c(h^{n-1}(x_i))+d(h^{n-1}(x_i)*)} + \lambda(h^{n-1}(x_i))P(x_i|h^{n-2}(x_i))$

  avec $\lambda(h^{n-1}(x_i)) = \frac{d(h^{n-1}(x_i)*)}{c(h^{n-1}(x_i))+d(h^{n-1}(x_i)*)}$,

  si $c(x_{i-n+1}\ldots x_{i-1}) > 0$
- $P(x_i|h^{n-2}(x_i))$ sinon

2.3. Choix de la meilleure langue

Si l'on dispose d'un tel modèle pour chacune des langues à reconnaître, et que l'on est assuré que tous les mots à reconnaître appartiennent une ces langues, la manière la plus directe de choisir la langue d'appartenance d'un nom inconnu x est de calculer sa probabilité d'occurence dans un modèle de chacune des langues, et de choisir celle dans laquelle le nom propre est le plus probable :

$$\begin{aligned} L^\star &= \operatorname*{argmax}_{L_i} P(x|L_i)P(L_i) \\ &= \operatorname*{argmax}_{L_i} P(x|L_i) \end{aligned} \qquad (3)$$

sous l'hypothèse d'équiprobabilité des langues.

Compte-tenu de la multitude des origines différentes attestées dans les corpus de noms propres, il est en général difficile de modéliser et d'identifier toutes les langues. Dans ce cas, il est nécessaire de faire l'hypothèse supplémentaire que le nom inconnu x n'appartient pas nécessairement à l'une des langues que l'on sait reconnaître, mais qu'il peut aussi appartenir à l'une des langues non-modélisées, regroupées dans L_{N+1}.

Dans ce contexte, la décision d'affectation d'un nom à une langue est prise en testant, pour chacun des langues, l'hypothèse que le nom appartient ou pas à la langue, à l'aide d'un test du rapport de vraisemblance. Pour chaque langue L_i, étant donnée une modélisation du complémentaire $\overline{L_i}$, on teste le rapport : $r_i(x) = \frac{P(x|L_i)}{P(x|\overline{L_i})}$. Si ce rapport excède un certain seuil, on rejette l'hypothèse que x appartient à L_i, sinon on la conserve.

De manière à pouvoir prendre une décision y-compris dans les cas où ce test ne permet pas de rejeter toutes les langues sauf une; de manière également à nous affranchir de l'arbitraire du choix d'un seuil de décision, nous avons estimé, pour chacune des langues L_i, la distribution empirique de ce rapport de vraisemblance en retirant (10 000 fois) avec remise dans les exemples positifs disponibles. Étant donné un nom inconnu x, et la valeur $r_i(x)$ de ce rapport pour la langue L_i, nous pouvons approximer la valeur $\alpha_i(x)$ correspondant à $\alpha_i = P(r > r_i(x)))$. La langue choisie est alors L^* telle que :

$$L^* = \operatorname*{argmax}_{L_i} P(r > \frac{P(x|L_i)}{P(x|\overline{L_i})}) \qquad (4)$$

Dans le cas où toutefois aucune langue ne permet d'obtenir une valeur α_i supérieure à un seuil arbitraire α_{min}, alors le nom propre est affecté à la langue par défaut L_{N+1}. L'intérêt principal de cette technique par rapport aux autres méthodes de rejet, est, du moins dans la manière dont nous l'avons mise en œuvre, d'opérer une sorte de normalisation des vraisemblances.

2.4. Différentes modélisation de la non-langue

La mise en œuvre de tests du rapport de vraisemblance présuppose que l'on soit capable de modéliser $P(x|\overline{L_i})$, c'est à dire la vraisemblance de x dans l'ensemble complémentaire de L_i. Dans les expériences décrites ci-dessous, étant donnés des modèles pour les langues $L_i, i = 1\ldots N$, nous avons utilisé les modes de calcul suivants :

$$P(x|\overline{L_i}) = \sum_{j\neq i} P(x|L_j) \qquad (5)$$

$$P(x|\overline{L_i}) = \sum_{j\in C(i)} P(x|L_j) \qquad (6)$$

$$P(x|\overline{L_i}) = P(x|\bigcup_j L_j) \qquad (7)$$

L'équation (5) correspond à l'hypothèse que les langues modélisées recouvrent complètement l'ensemble des langues possibles. Dans ce cas, $\overline{L_i}$ est variable d'une langue à l'autre. Dans le cas de l'équation (6), la sommation se fait seulement sur des langues choisies *a priori* comme ressemblant le plus à L_i (modèle de cohorte), ce qui revient à supposer négligeable dans (5) l'influence des langues trop éloignées de L_i. Enfin, utiliser l'équation (7) dans un test du rapport de vraisemblance revient à comparer la vraisemblance de x dans le modèle de L_i à une vraisemblance moyenne, estimée sur un échantillon qui regroupe tous les échantillons de langues disponibles.

3. Expériences et Résultats

3.1. Les noms de villes

Nous présentons dans cette section les résultats des divers modèles d'identification présentés ci-dessus, dans un contexte de choix parmi un ensemble fermé de langues.

3.1.1. Les données La première tâche sur laquelle nous avons testé nos systèmes de reconnaissance concerne l'identification de l'origine d'un petit ensemble (287) de noms de localités suisses, celles pour lesquelles plusieurs orthographes coexistent, correspondant aux 3 langues parlées en Suisse, à savoir l'allemand, le français et l'italien. Ainsi, trois entrées de cette liste réfèrent la ville de Genève, *Genève* (version française), *Ginovra* (version italienne) et *Genf* (version allemande). L'intérêt de cette tâche est que les données d'Onomastica concernant les noms de localités sont nettement plus homogènes que celles qui concernent les noms de personnes, qui contiennent une proportion importante d'emprunts. Nous sommes donc ici dans la situation quasi-idéale où les données d'apprentissage comme les données de tests sont très peu bruitées, et où l'ensemble des langues à reconnaître est connu *a priori* et très restreint. Ceci nous permet d'avoir une idée de la borne maximale des performances atteignables avec nos modèles.
Nous avons utilisé, pour apprendre à discriminer ces trois langues, les bases de données Onomastica, qui contiennent un nombre important de noms de localités (de l'ordre respectivement de 42 000 pour l'Allemagne, de 38 000 pour la France (hors localités corses et alsaciennes), et 32 000 pour l'Italie).

3.1.2. Résultats Les résultats obtenus pour cette tâche sont résumés dans le tableau 1. Dans ce tableau, les valeurs correspondant aux résultats du modèle de Vitale ont été obtenues en utilisant comme séquences typiques les 3-grammes. Le calcul direct des probabilités comme un rapport de fréquences d'occurence que cette approche met en œuvre rend en effet inefficace l'utilisation de séquences plus grandes, à cause du nombre trop important de n-grammes de probabilité nulle dans le corpus de test. Les valeurs correspondant aux lignes Modèle (I) et Modèle (V) ont été obtenues en utilisant des modèles de 4-grammes; dans le premier cas le choix de la langue optimale se fait selon (3), dans le second, elle se fait selon (4) (avec $\alpha_{min} = 0.0$); dans les deux cas les meilleurs résulats ont été obtenus en modélisant $\overline{L_i}$ de manière variable selon (5).

	All.	Fran.	Ital.	Total
Vitale	70.7%	79.0%	85.7%	74.2%
Modèle (I)	68.9%	87.0%	95.2%	76.6%
Modèle (V)	72.6%	85.0%	95.2%	78.0%

FIG. 1 – *Taux de reconnaissance de l'origine des noms de localité suisses*

Dans l'ensemble, les différents modèles testés permettent d'obtenir des taux de reconnaissance relativement satisfaisants, d'autant plus que, selon notre jugement, environ la moitié des erreurs de classement qui restent paraissent difficile à corriger (ainsi, par exemple, on voit mal comment nos algorithmes pourraient prédire que *Celerina*, *Valcava* et *Villa* sont des orthographes allemandes; ou encore que *Schaffenhouse*, *Schwytz*, et *Zurich* sont en fait des noms français).

L'utilisation de modèles de langage permet d'améliorer à la marge les résultats obtenus par la méthode de Vitale, le gain le plus sensible se situant dans l'amélioration de la discrimination entre français et italien. Le principal reliquat d'erreurs, pour les deux types de modèles, consiste en des détections incorrectes de noms allemands : environ 30% de ces noms sont mal classés, approximativement équirépartis entre français et italien. Par contre, des différences plus sensibles entre les deux types de modélisation apparaissent lorsque l'on rentre dans le détail des erreurs : près de 40% des erreurs commises avec le modèle de Vitale sont corrigées lorsque l'on a recourt à des modèles de langage; près de 30% des noms mal identifiés par des modèles de langage sont correctement classés par l'algorithme de Vitale.
Les deux résultats obtenus avec des modèles de langage sont très similaires, y-compris dans le détail des noms erronés. Le meilleur score obtenu pour la règle de décision (4) par rapport à la règle de décision est, selon nous, anecdotique, et provient, pour l'essentiel, de la normalisation des vraisemblances qui est réalisée dans la procédure de décision correspondant au modèle (V).
Il est intéressant toutefois de noter que les résultats du modèle (V) ont été obtenus en modélisant, pour les trois langues à identifier, $P(x|\overline{L_i})$ comme $\sum_{j \neq i} P(x|L_j)$, où j varie *sur l'ensemble des langues que pouvions modéliser*, c'est-à-dire y-compris sur des langues que nous ne cherchions pas à reconnaître (l'anglais, le hollandais, etc). Différentes modélisations, moins riches, de $P(x|\overline{L_i})$, par exemple celle qui ne fait intervenir que les langues à identifier, se sont révélées moins efficaces, probablement parce qu'elles permettent de modéliser moins finement les distributions de $\frac{P(x|L_i)}{P(x|\overline{L_i})}$.

3.2. Les noms de personnes

Nous testons dans cette section les capacités des différents modèles à identifier correctement l'origine de patronymes européens, dans un contexte où les noms à identifier sont mêlés à des patronymes choisis dans d'autres langues. Le but est donc ici également d'éprouver les différents mécanismes de rejet.

3.2.1. Les données

Nous avons, pour cette série d'expériences, délibérément choisi une tâche sensiblement plus difficile qu'à la section précédente, puisqu'il s'agit d' identifier 7 langues (ou groupes de langues), à savoir : l'allemand, l'italien, le français, l'anglais, les langues ibères (espagnol, portuguais), les langues nordiques (danois, norvégien, suédois) et les langues slaves (bulgare, polonais, serbo-croate russe, tchèque).
Le jeu de données que l'on utilise pour tester les capacités de discrimination de nos algorithmes est très bruité, puisqu'il inclut, en quantité importante, des noms originaires d'autres langues européennes (grec, turc, finlandais, arménien, albanais), ou encore d'autres zones linguistiques (Afrique du nord, Afrique Noire, Asie du Sud-Est). De surcroît, ces noms ne contiennent aucun signe diacritique, et constituent donc un ensemble de test spécialement difficile, puisqu'un certain nombre de graphèmes permettent d'identifier avec une quasi-certitude des langues ou familles de

langues : ainsi le ø pour les langues nordiques, le ß pour l'allemand, etc. La répartition des origines linguistiques dans cet ensemble de (10 000) noms propres est donnée dans la table 2, dans lequel l'on a également fait figurer l'étiquette correspondant à chacune des langues à reconnaître.

Allemand (A)	550	Afr. du Nord	1437
Français (F)	896	Afr. Noire	681
Italien (I)	721	Asie Sud-Est	644
Anglais (R)	868	Grec, Turc	830
Ibèrique (E)	411	Hollandais	502
Nordique (N)	332	Autres	792
Slave (S)	1336		
Total	5114	Total	4886

FIG. 2 – *Distribution des noms à reconnaître par origine linguistique.*

Pour l'apprentissage des modèles, nous avons utilisé les patronymes disponibles dans les bases Onomastica. Dans la mesure toutefois où ces bases contiennent, pays par pays, un volume important d'emprunts, nous avons jugé préférable de les rendre plus homogène. À cet effet, nous avons utilisé un filtre à base de règles (environ 600), qui nous a permis de supprimer, pour chacun des fichiers d'apprentissage, les mots reconnus comme appartenant à une autre langue. Notons toutefois que pour le français et surtout l'allemand, compte-tenu de la proportion trop importante d'emprunts après filtrage, compte-tenu également des volumes de données très important disponibles pour ces langues, nous avons utilisé non pas les bases filtrées, mais *directement les listes de noms identifiées par les règles*. Ces règles nous ont également permis d'extraire une base d'apprentissage importante pour les langues slaves (absentes d'Onomastica), ainsi que pour un certain nombre d'autres langues ou familles de langues (quoiqu'avec des volumes bien plus réduits). Ces noms propres ont été regroupés dans un unique ensemble (étiqueté M), qui a été utilisé pour construire une caractérisation négative des langues à reconnaître.

3.2.2. Résultats Pour cette série d'expériences, nous avons utilisé pour le modèle de Vitale une caractérisation par les séquences de taille 3. Le mécanisme de rejet associé à ce modèle est la caractérisation négative : la langue choisie est donc celle qui satisfait (2), où le maximum est calculé sur l'ensemble des langues à reconnaître plus la « langue » correspondant à tous les exemples négatifs disponibles. Les résultats détaillés de cette expérience sont donnés dans la table 3.

	E	F	I	N	R	A	S	M	Tot.
E	63.87 2.51 22.19	4.58 0.18 1.78	21.37 0.84 5.03	1.78 0.07 1.23	3.31 0.13 1.37	1.53 0.06 0.59	0.25 0.01 0.07	3.31 0.13 0.59	100.00 3.93 3.93
F	10.83 0.97 8.58	58.37 5.23 51.73	11.50 1.03 6.17	3.91 0.35 6.14	5.58 0.50 5.27	3.79 0.34 3.35	0.56 0.05 0.34	5.47 0.49 2.24	100.00 8.96 8.96
I	9.71 0.70 6.19	1.66 0.12 1.19	84.33 6.08 36.43	0.55 0.04 0.70	0.14 0.01 0.11	1.94 0.14 1.38	1.11 0.08 0.54	0.55 0.04 0.18	100.00 7.21 7.21
N	0.60 0.02 0.18	2.41 0.08 0.79	1.51 0.05 0.30	51.81 1.72 30.18	9.94 0.33 3.48	23.80 0.79 7.79	6.33 0.21 1.43	3.61 0.12 0.55	100.00 3.32 3.32
R	3.00 0.26 2.30	8.76 0.76 7.52	4.84 0.42 2.52	5.07 0.44 7.72	59.22 5.14 54.16	11.52 1.00 9.86	4.49 0.39 2.65	3.11 0.27 1.24	100.00 8.68 8.68
A	1.27 0.07 0.62	2.36 0.13 1.29	2.00 0.11 0.66	3.45 0.19 3.33	3.09 0.17 1.79	77.45 4.26 42.01	7.27 0.40 2.72	3.09 0.17 0.78	100.00 5.50 5.50
S	5.69 0.76 6.72	1.65 0.22 2.18	5.46 0.73 4.37	2.40 0.32 5.61	1.95 0.26 2.74	5.39 0.72 7.10	72.08 9.63 65.51	5.39 0.72 3.29	100.00 13.36 13.36
M	12.12 5.92 52.34	6.94 3.39 33.53	15.21 7.43 44.52	5.24 2.56 44.91	5.91 2.89 30.45	5.77 2.82 27.81	8.04 3.93 26.73	40.77 19.92 91.13	100.00 48.86 48.86
Total	11.31 11.31 100.00	10.11 10.11 100.00	16.69 16.69 100.00	5.70 5.70 100.00	9.49 9.49 100.00	10.14 10.14 100.00	14.70 14.70 100.00	21.86 21.86 100.00	100.00 100.00 100.00

FIG. 3 – *Matrice de confusion entre langues obtenue en utilisant le modèle de Vitale*

La case correspondant à l'intersection de la ligne A et de la colonne A donne successivement le pourcentage de noms allemands classés comme allemands, la fraction (par rapport au total) de ces noms, et le pourcentage de noms allemands parmi ceux qui ont été classés comme allemand.

Le taux de reconnaissance global obtenu avec cette méthode est de 54.49%, la large majorité des erreurs (près de 30%) provenant de « fausses alarmes », c'est-dire de noms étrangers incorrectement reconnus comme appartenant à une des 7 langues apprises. Par contre, les noms appartenant aux langues effectivement apprises sont dans l'ensemble assez bien reconnus. En effet, si l'on se restreint à ces noms, le pourcentage d'identifications correctes est de 67.6%, avec des disparités importantes, puisque l'italien est correctement identifié à plus de 80%, alors que pour le français, le taux de reconnaissance n'est que de 58%. De surcroît, la majorité des erreurs commises sur ces noms correspond à un mauvais classement dans une autre langue à reconnaître, puisque les faux rejets sont très rares. L'examen de la table 3 permet enfin de constater que l'on retrouve bien des taux de confusion élevés entre les langues de même famille (italien et espagnol, anglais et allemand, etc).

Les modèles de langage utilisés dans ces expériences sont également des modèles de 3-gramme. Nous avons comparé plusieurs méthodes différentes pour identifier les noms à rejeter (avec ou sans caractérisation négative, avec plusieurs modèles de $\overline{L_i}$, plusieurs valeurs de α_{min} pour les rejets par test LR, etc). Ces expériences ont permis de montrer que le choix de l'une ou l'autre stratégie de rejet pouvait avoir des conséquences très marquées, tant en ce qui concerne les taux de reconnaissance globaux que les taux d'identifications correctes (pourcentage calculé sur les seuls noms appartenant effectivement aux 7 langues à discriminer).

En particulier, il est apparu qu'ajouter une caractérisation négative en apprenant un modèle des exemples négatifs, aussi disparates soient ces exemples, permettait de réaliser des gains très sensibles pour les ces deux scores : ainsi, apprendre un modèle des exemples négatifs, permet, pour $alpha_{min} = 0.05$, et pour une modélisation uniforme de $\overline{L_i}$ selon (7) permet d'augmenter de près de 20 points le taux de reconnaissance global, tout en améliorant de 2 points de taux d'identifications correctes.

Dans notre contexte, il est également apparu que l'utilisation de modèles de cohortes, dans lesquels les cohortes sont définies sur la base de familles linguistiques (langues romanes vs. langues germaniques) ne permettait pas d'améliorer les résultats.

Enfin, nous avons constaté, à modèle constant, que l'augmentation progressive du seuil $alpha_{min}$ avait des effets contrastés : passer de 0.01 à 0.05 permet d'améliorer simultanément le taux de reconnaissance global et le taux d'identification correcte; pour des valeurs supérieures toutefois, si le taux de reconnaissance global continue mécaniquement d'augmenter (nos modèles pêchant tous par manque de rejet, augmenter $alpha_{min}$ revient à augmenter le taux de rejet, et donc à améliorer les scores), le taux d'identifications correctes lui diminue.

Le meilleur compromis global a été réalisé en combinant une caractérisation négative, un seuil $alpha_{min}$ égal à 0.05, et une modélisation de $\overline{L_i}$ uniforme. Dans cette configuration, nous avons obtenu un taux de reconnaissance globale de 55.65%, avec un pourcentage d'identifications correctes de 70.16%, soit une amélioration sur les deux tableaux par rapport au modèle de Vitale. Les résultats complets correspondant à cette expérience sont donnés dans la table 1. Il est à noter que le meilleur résultat en reconnaissance global obtenu sur toutes nos expériences a été de 59.17%, et que le meilleur taux d'identifications correctes a été de 72.08%.

4. Perspectives

4.1. Quelques leçons

Ces premières tentatives dans le domaine de discrimination des langues sur la base de chaînes graphiques isolées se sont avérées riches en enseignement. Les différents modèles statistiques utilisés permettent d'obtenir des scores de reconnaissance relativement satisfaisants, y compris où l'ensemble des noms à identifier est très bruité. Utiliser des modèles de langage permet d'améliorer les performances obtenues en utilisant le modèle proposé par Vitale, qui est toutefois, malgré sa simplicité, remarquablement robuste. Dans tous les cas, il s'avère enfin nécessaire de disposer d'exemples négatifs si l'on souhaite traiter des corpus « réels », qui contiennent une proportion très importante d'emprunts à des langues non-modélisées.

Dans le cadre de nos applications de transcription, nous voyons ces modèles comme des auxiliaires tout à fait compétents de modèles à base de règles, et ce à double titre. Les règles dont nous disposons sont en effet relativement fiables, mais elles ne permettent de traiter qu'une fraction des noms propres (environ 40% seulement du corpus de patronymes français d'Onomastica). Disposer dans ce contexte de modèles de langages permet d'une part de classer la fraction de noms résiduelle; permet d'également d'enrichir le formalisme de règles que nous utilisons. En effet, nos règles ne permettent pour l'instant de que de prendre des décisions positives, du type : la séquence x identifie la langue Y. Disposer additionnellement de mécanismes statistiques permet d'utiliser les règles comme des mécanismes d'affectation des probabilités *a priori* des langues. Une règle positive associe une probabilité *a priori* nulle à toutes les langues sauf une, mais d'autres types de règles peuvent également être utilisées, qui restreignent les langues à discriminer à un petit sous-ensemble des langues possibles. Ce type de règles s'avèrerait certainement très productif dans le cas précis des noms propres, pour lesquels il existe un recouvrement non-négligeable entre les suffixes caractéristiques. Ainsi *-ger* est simultanément caractéristique de l'allemand (*Wenger*) comme du français (*Boulanger*), *-ova* est typique du tchèque (e.g. dans *Ivanova*), mais également fréquent en italien (*casanova*), etc.

4.2. Voies d'amélioration

La principale faiblesse des techniques statistiques que nous avons mis en œuvre, est, à nos yeux, de traiter de manière uniforme tous les n-grammes figurant dans un mot à classer. Les suffixes étant en effet, pour diverses raisons (morphologique pour certaines langues, historiques pour d'autres) des indices particulièrement saillants de l'origine d'un patronyme, l'information apportée par ces séquences gagneraient donc à être sur-pondérés.

	A	E	F	I	N	M	R	S	Tot.
A	76.91 4.23 46.74	2.55 0.14 0.92	1.82 0.10 0.85	0.36 0.02 0.16	2.91 0.16 3.17	6.18 0.34 1.52	4.91 0.27 2.16	4.36 0.24 2.01	100.00 5.50 5.50
E	0.73 0.03 0.33	76.40 3.14 20.63	5.60 0.23 1.96	8.76 0.36 2.97	1.95 0.08 1.59	1.70 0.07 0.31	4.87 0.20 1.60	0.00 0.00 0.00	100.00 4.11 4.11
F	2.01 0.18 1.99	12.72 1.14 7.49	66.85 5.99 50.98	3.68 0.33 2.72	2.46 0.22 4.37	3.01 0.27 1.21	9.04 0.81 6.47	0.22 0.02 0.17	100.00 8.96 8.96
I	1.66 0.12 1.33	16.09 1.16 7.62	0.97 0.07 0.60	78.50 5.66 46.66	0.42 0.03 0.60	0.83 0.06 0.27	0.69 0.05 0.40	0.83 0.06 0.50	100.00 7.21 7.21
N	23.49 0.78 8.62	0.90 0.03 0.20	0.90 0.03 0.26	0.90 0.03 0.25	53.31 1.77 35.12	4.52 0.15 0.67	12.05 0.40 3.20	3.92 0.13 1.09	100.00 3.32 3.32
R	8.76 0.76 8.40	2.65 0.23 1.51	6.45 0.56 4.77	0.92 0.08 0.66	2.30 0.20 3.97	3.92 0.34 1.52	72.24 6.27 50.12	2.76 0.24 2.01	100.00 8.68 8.68
S	5.46 0.73 8.07	9.43 1.26 8.28	2.69 0.36 3.06	1.95 0.26 2.14	1.80 0.24 4.76	9.96 1.33 5.96	2.69 0.36 2.88	66.02 8.82 73.68	100.00 13.36 13.36
M	4.54 2.22 24.53	16.62 8.12 53.35	9.03 4.41 37.53	11.03 5.39 44.44	4.79 2.34 46.43	40.46 19.77 88.54	8.49 4.15 33.17	5.03 2.46 20.55	100.00 48.86 48.86
Total	9.05 9.05 100.00	15.22 15.22 100.00	11.75 11.75 100.00	12.13 12.13 100.00	5.04 5.04 100.00	22.33 22.33 100.00	12.51 12.51 100.00	11.97 11.97 100.00	100.00 100.00 100.00

TAB. 1 – *Matrice de confusion entre les langues obtenue en utilisant des modèles de langage*

En second lieu, il nous semble que la manière dont les différentes probabilités n-gramme sont estimées sont finalement assez mal adaptées à la tâche de classification pour laquelle nous les utilisons. Ce qui importe réellement, c'est non pas la probabilité d'occurence d'une séquence, mais sa capacité à discriminer entre les langues. Dans ce cadre, on pourrait envisager de remplacer, dans le modèle de Vitale, les scores des n-grammes par des indices de typicité, tels que l'information mutuelle entre le n-gramme et la langue, ou d'autres indices statistiques, obtenus en centrant les fréquences d'occurence par rapport la fréquence d'occurence moyenne (toutes langues confondues).

Dans un tel contexte, il probablement possible d'améliorer les scores obtenus dans nos expériences, que se soit en utilisant d'autres modèles statistiques plus orientés vers sur la discrimination (nous pensons en particulier à des méthodes neuronales (**?**) ou encore à des méthodes de discrimination factorielles), ou des techniques symboliques tels que l'apprentisage d'arbres de décision, ou encore à méthodes fondées sur l'extraction d'analogies.

Finalement, il convient de revenir sur la finalité de ces travaux, à savoir l'amélipration de la transcription orthographique-phonétique des noms propres pour la synthèse de parole. Améliorer les performances absolues d'identification est certes souhaitable, mais les résultats qui en dernière instance importent concernent non pas le nombre d'erreurs d'identification, mais le nombre d'erreurs de prononciation : les deux ne vont pas de pair, d'une part parce que certaines langues (en particulier celles dont les noms sont translités d'un autre alphabet graphique que l'alphabet latin sont en général plus « faciles » à prononcer que les noms européens), d'autre part parce que la familiarité est plus grande avec les langues voisines (du français), qui sont à la fois mieux identifiées (par l'humain) et mieux prononcées. Il y a ici un important travail empirique à conduire, pour à la fois mesurer correctement l'apport de systèmes d'identification, pour également obtenir une meilleure idée du nombre de langues (ou groupes de langues) à identifier.

Remerciements

Les auteurs tiennent à remercier vivement Frédéric Bimbot pour quelques discussions fructueuses concernant l'identification et la reconnaissance du locuteur, ainsi que Karim Nedir (Swiss Telekom) qui a fourni les données concernant les localités suisses.

Références

BELHOULA, K. 1993. Rule-based grapheme-to-phoneme conversion of names. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 881–884.

CARLSON, R.; GRANSTRÖM, B. et LINDSTRÖM, A. 1989. Predicting name pronunciation for a reverse directory service. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 113–116.

FEDERICO, M.; CETTOLO, M.; BRUGNARA, F. et ANTONIOL, G. 1995. Language modelling for efficient baum-search. *Computer Speech and Language* 9:353–379.

GOLDING, A. R. 1991. *Pronouncing names by a combination of rule-based and case-based reasoning*. Ph.D. Dissertation, Stanford University, Stanford, CA.

GUSTAFSON, J. 1996. A Swedish name pronunciation system. Licentiate Thesis, Royal Institute of Technology, Stockolm.

HOCHBERG, J.; MNISZEWSKI, S.; CALLEJA, T. et PAPCUN, G. J. 1990. What's in a name ? Last names as a computational problem. Unpublished paper, Los Alamos National Laboratory, Los Alamos, NM.

JACK, M.; SCHMIDT, M.; SUTHERLAND, A.; DALSGAARD, P.; GRANDSTRÖM, B.; VAN DEN HEUVEL, H.; LOGOTHETHIS, M.; MENGEL, A.; ; OTTESEN, G.; PARDO, J.; PIRRELLI, V.; TRANCOSO, I. et YVON, F. 1995. Multi-language pronunciation dictionary of proper names and place names. Final Report, LRE Project # LRE-6104.

JELINEK, F. 1990. Self-organized language modeling for speech recognition. In WAIBEL, A. et LEE, K.-F., eds., *Readings in Speech Recognition*. San Mateo, CA: Morgan-Kaufman. 450–506.

KATZ, S. M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-35(3):400–401.

MÉTAYER, I. et VINCENT, M. 91. La synthèse vocale des noms propres : problèmes et propositions de solutions. Technical report, CNET.

SCHMIDT, M.; FITT, S.; SCOTT, T. et JACK, M. 1993. Phonetic transcription standards for European names (ONOMASTICA). In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, 279–282.

VITALE, T. 1991. An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computationnal Linguistics* 17(3):257–276.

YVON, F. 1996. *Prononcer par analogie : motivations, formalisations et évaluations*. Ph.D. Dissertation, École Nationale Supérieure des Télécommunications, Paris.