Approches segmentales multilingues pour l'identification automatique de la langue : phones et syllabes

Fabien Antoine, Dong Zhu, Philippe Boula de Mareüil, Martine Adda-Decker*

LIMSI-CNRS, Université de Paris-Sud, BP 133, 91403 Orsay CÉDEX, France
* DGA-CTA, 16 bis av. Prieur de la Côte d'Or, 94114 Arcueil CÉDEX, France
Mél : Fabien.Antoine@etca.fr, Dong.Zhu@limsi.fr, Philippe.Boula.de.Mareuil@limsi.fr, Martine.Adda@limsi.fr

ABSTRACT

This study focuses on unit modelling within the framework of the phonotactic approach to automatic language identification. The unit set is extended from a multilingual phone set to a multilingual syllable inventory. Various motivations and questions have originated this work : in particular, do syllabotactic constraints provide more useful information for language identification than phonotactic knowledge ? The development of a common syllable base enables the description and analysis of large multilingual corpora. The present study is a first step in this direction. It makes use of radio and TV broadcast shows in 8 languages (Arabic, Chinese, English, French, German, Italian, Portuguese and Spanish) for which orthographic transcriptions are available. Using a shared multilingual phone set of 74 symbols, phonemic transcriptions were generated. A general syllabification algorithm was applied on this corpus, leading to a 5,380 syllable set. Two systems, based on phonotactic and syllabotactic approaches were developed, and a preliminary comparison is provided.

1. INTRODUCTION

Les travaux présentés ici s'inscrivent dans le cadre du projet MIDL (Modélisations pour l'IDentification des Langues) du programme interdisciplinaire STIC-SHS « Société de l'Information ». L'objectif de ce projet est de combiner connaissances linguistiques et compétences technologiques autour de la problématique de l'identification des langues par les humains et les machines [1]. Si l'identification automatique des langues (IAL) est un domaine de recherche actif depuis 30 ans, les recherches ont connu un véritable essor dans les année 90 avec principalement les méthodes à base de modèles de Markov cachés (HMM), pour modéliser les niveaux acoustiques, et la mise en place de corpus multilingues. Le nombre de langues étudiées et la quantité de données pour l'apprentissage des modèles d'identification croissent sans cesse ; toutefois, on reste très loin des quelques milliers de langues répertoriées dans le monde.

Plusieurs niveaux d'information contribuent à l'identification humaine de la langue, qu'il est usuel de séparer en acoustique, phonétique, phonotactique, prosodique, morphologique et lexical. Ces niveaux ne présentent pas une difficulté équivalente de modélisation, et les plus explorés sont l'acoustique, le phonétique et le phonotactique [10, 4]. Le succès de ces modélisations est dû d'abord au fait que ces niveaux portent une quantité d'information souvent décisive, mais aussi à leur relative facilité de modélisation et de mise en œuvre, grâce à des techniques partagées avec la reconnaissance de la parole. Le niveau prosodique porte certainement une information utilisée par les humains [6] : sa modélisation et sa mise en application dans un système d'IAL relèvent cependant toujours un peu du défi [7]. La prise en compte des niveaux morphologique et lexical est possible pour les langues à tradition écrite (soit quelques centaines dans le monde). Ici, la complexité augmente très vite : on se rapproche de la mise au point d'une batterie de systèmes de transcription automatique [9] en parallèle. Afin d'optimiser les taux d'identification automatique des langues, plusieurs systèmes, avec des modèles et des approches différents, peuvent être mis en parallèle et combinés pour la décision quant à l'identité de la langue parlée. Dans ce cadre, toute approche nouvelle peut être utilement intégrée.

Dans le travail décrit par la suite, nous proposons une extension de l'approche phonotactique classique à une unité plus longue que le phonème. On exploite d'habitude les fréquences d'observation de séquences de n phones décodés automatiquement (typiquement n=3). Or les taux de reconnaissance d'unités courtes comme les phones sont en général significativement plus faibles que pour les unités plus longues. Une autre motivation en faveur d'une unité telle que la syllabe provient du fait qu'un modèle phonotactique permet de générer des séquences de phones impronon̂çables, alors qu'un modèle qui génère des séquences de syllabes atténue ce défaut. Enfin, le choix d'une unité de type syllabique permet également d'envisager des études comparatives des langues [3], et d'intégrer des paramètres prosodiques dans une étape ultérieure. La syllabe n'admet cependant pas une définition unique : variable d'une langue à l'autre, elle peut même être controversée à l'intérieur d'une même langue. En dépit de ces variations, la syllabe nous semble intéressante à explorer tant au niveau linguistique que dans la perspective de développement de systèmes peu supervisés pour l'identification d'un grand nombre de langues.

La section suivante présente l'approche générale et le corpus. La notion de (pseudo-)syllabe est développée dans la section 3, tandis que la section 4 décrit les premières mises en œuvre et compare avec l'approche phonotactique classique.

2. APPROCHE GÉNÉRALE ET CORPUS

2.1. Description du corpus

Les corpus d'émissions de radio et télévision offrent une variété linguistique plus large que les corpus téléphoniques généralement utilisés pour l'identification des

langues, et offrent aussi une meilleure qualité acoustique pour la modélisation phonétique et syllabique. Un corpus multilingue a ainsi été constitué pour les langues suivantes : arabe standard, chinois mandarin, anglais américain, espagnol latino-américain, allemand, italien, français et portugais européen. Les corpus français et arabe sont des ressources fournies par la DGA. L'anglais, l'espagnol et le chinois sont extraits de corpus Hub4 du LDC. Les corpus allemand, portugais et italien sont issus de divers projets européens FP5 LE (OLIVE, ALERT) ou obtenus auprès d'ELDA – le corpus portugais étant le plus petit. Nous avons limité la plupart des expériences à un corpus d'apprentissage de trois heures. Pour toutes ces langues, des transcriptions orthographiques sont disponibles, et des lexiques de prononciation ont été construits.

2.2. *Systèmes d'IAL explorés*

La figure 1 illustre l'approche utilisée pour l'identification des langues à base phonotactique ou syllabotactique. Dans les deux cas, le corpus d'apprentissage sert à estimer dans un premier temps le modèle acoustique multilingue (de type HMM), et dans un second temps les modèles de langage (de type n-gramme), dont le vocabulaire est respectivement constitué par le jeu de phonème et par le jeu de syllabes.

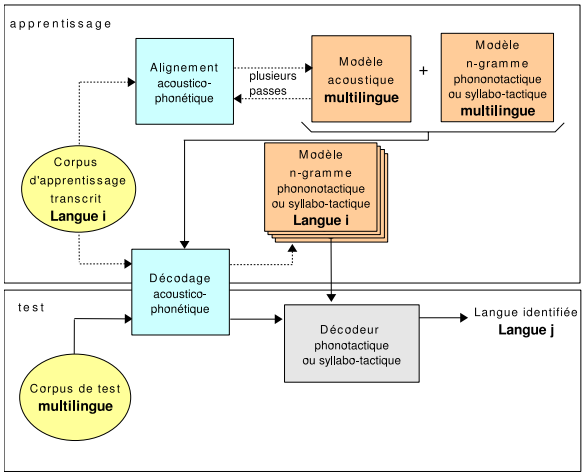

FIG. 1: Identification automatique de la langue par une approche phonotactique ou syllabotactique. L'apprentissage se décompose en un apprentissage du modèle acoustique multilingue et un apprentissage du modèle de langage (phonotactique ou syllabotactique). L'identification elle-même d'un échantillon consiste à en faire un décodage phonétique ou syllabique, dont la vraisemblance est estimée à partir du modèle de langage correspondant.

La première étape pour un tel système consiste à définir un jeu de phonèmes commun pour les langues, dont le choix a été fait dans [1]. Nous donnons une partie de ce jeu de 74 phonèmes dans la table 1.

3. SYLLABATION

3.1. *La syllabe*

La syllabe est traditionnellement rapportée à l'activité des muscles liés à la respiration, les muscles intercostaux : selon l'intensité naturelle des phonèmes émis, l'activité en est plus ou moins importante. L'émission de parole peut alors être considérée comme un flux d'air d'intensité variable, où les pics correspondent de façon idéale aux sommets de syllabes. La plupart du temps, les pics d'intensité sont représentés par des voyelles, mais peuvent être aussi représentés par d'autres phonèmes.

Pour de nombreuses langues, la syllabe est reconnue comme unité structurante du langage. On peut la schématiser par une voyelle (ou une diphtongue) facultativement entourée de consonnes ; la structure interne de la syllabe est alors constituée de trois parties : l'attaque, le noyau et la coda – ces ceux dernières parties regroupées constituant la rime. L'attaque, quand elle n'est pas vide, est constituée par une ou plusieurs consonne(s) précédant le sommet ; le noyau représente le sommet de la syllabe, et la coda est optionnellement constituée de consonnes suivant le noyau. Pour le mot français *piste* [pist], par exemple, composé d'une seule syllabe, l'attaque est [p], le noyau [i] et la coda [st].

3.2. *Méthode générale de syllabation*

Depuis Saussure [8], différentes théories de la syllabation ont été proposées afin de découper la parole en syllabes. Le *Principe de Sonorité* propose d'ordonner les phonèmes sur une échelle de sonorité (voir table 1), qui correspond grosso modo à leur degré d'intensité perçue (« loudness »). Suivant cette échelle, les voyelles sont les plus sonores, suivies par les glides (semi-voyelles ou semi-consonnes), les liquides, les fricatives et les plosives. Le *Principe de Sonorité* stipule alors que les consonnes en début de syllabe doivent se succéder selon une intensité croissante.

TAB. 1: Exemples de phonèmes, par classe de sonorité dans un ordre décroissant.

voyelles ouvertes	a ɑː ɒ æ ɑ̃
voyelles semi-ouvertes	ɛ œ ʌ ɔ ɛ̃ ə
voyelles semi-fermées	e ø o õ
voyelles fermées	i iː ɪ y yː ʏ u uː ʊ ũ
glides	j w ɥ
rhotiques	r ɾ ʁ ʁ
latérales	l ʎ
nasales	n m ŋ ɲ
fricatives sonores	v z ʒ ð ðˤ
fricatives sourdes	f s sˤ ʃ ç θ ħh x ɕ
occlusives sonores	b d dˤ ɡ
occlusives sourdes	p t tˤ k ʔ

Un autre principe, celui de l'*Attaque Maximale* (« Maximum Onset Principle », [5]) stipule que la frontière syllabique entre deux voyelles séparées par des consonnes est placée de façon à maximiser le nombre de consonnes en attaque de la deuxième syllabe. Ces consonnes, cependant, doivent constituer des groupes « légaux », c'est-à-dire des clusters qui peuvent apparaître en début de mot dans la langue.

Globalement, les règles retenues pour la syllabation dans notre système d'IAL sont un compromis entre le *Principe de Sonorité* et celui d'*Attaque Maximale*. Ces règles sont difficiles à mettre en œuvre indépendemment de la langue, car il faut prendre en compte les particularités de certaines

langues comme le français, où l'on considère traditionnellement que la syllabation se fait indépendamment de la division grammaticale et orthographique des mots. L'algorithme de syllabation pour le français est décrit plus en détail dans la sous-section suivante.

3.3. *Algorithme de syllabation pour le français*

Le principe des règles de syllabation pour le français est le suivant [2] : on repère certains motifs de phonèmes à découper parmi les groupes V, L, G, O, C et x (voir ci-dessous). On découpe en priorité le motif de la première ligne de la table 2, puis en cas d'impossibilité le motif de la seconde, et ceci jusqu'au dernier.

TAB. 2: Règles de découpage syllabique en français. V={voyelle} ; C{0,4}={de 0 à 4 consonnes} ; G={glide}, O={occlusive, fricative ou nasale} ; L={liquide} ; x={phonème quelconque}.

séquence	*découpage*	*exemple*	*prononcé*
əC{0,4}V	ə.C{0,4}V	re-structurer	[ʁə.stʁykt...]
VV	V.V	co-opérer	[ko.ɔpeʁe]
VxV	V.xV	i-miter	[i.mite]
VxGV	V.xGV	stu-dio	[sty.djo]
VOLV	V.OLV	pu-blic	[py.blik]
VxxV	Vx.xV	sor-tir	[sɔʁ.tiʁ]
VOLGV	V.OLGV	em-ploi	[ɑ̃.plwa]
VxxGV	Vx.xGV	vic-toire	[vik.twaʁ]
VxOLV	Vx.OLV	es-prit	[ɛs.pʁi]
VxxxV	Vxx.xV	ex-pert	[ɛks.pɛʁ]
VxOLGV	Vx.OLGV	al-truiste	[al.tʁɥist]
VxxxxV	Vxx.xxV	ex-pier	[ɛks.pje]
VxxxxGV	Vxx.xxGV	ex-ploit	[ɛks.plwa]

3.4. *Recensement des syllabes dans les 8 langues*

Des particularités sont à prendre en compte en marge de l'approche générale que nous avons suivie. La morphologie, en effet, est très différente d'une langue à l'autre. En chinois par exemple (langue largement monosyllabique), la syllabe joue un rôle morphologique fondamental. Phonétiquement, il existe seulement deux formes de syllabes : vocalique simple (V, où V peut être une diphtongue ou bien une voyelle suivie d'une nasale), ou bien consonantique (CV, où C peut être une consonne complexe, alors décomposée dans notre jeu de phonèmes : [tɕ^{h}] devient par exemple [tɕh]. Dans les langues romanes (espagnol, italien, portugais, français), la structure CV est dominante, le sommet de syllabe V peut également être constitué d'une diphtongue (ex. *rei* [rei] en italien) ; mais les syllabes peuvent de plus avoir le type VC. L'italien a par ailleurs la particularité de présenter un grand nombre de géminées, qui sont dédoublées de chaque côté de la frontière de syllabes qu'elles délimitent. Les langues romanes considérées ne présentent pas dans leur vocabulaire natif de syllabe du type CCCVCC : les seules exceptions sont des mots comme *script* ([skʁipt] en français). En anglais et en allemand, la proportion de syllabes fermées est nettement plus élevée et avec des structures plus complexes. L'allemand représente la langue la plus riche en types de syllabes : la syllabe la plus longue est de type CCCVCCCC (*strolchst* [ʃtʁɔlçst]). En arabe, la syllabe débute toujours par une consonne, et peut être de 5 types : court et ouvert (CV), long et ouvert (CV:), long et fermé (CVC, CV:C et CVCC, qui se trouve uniquement en fin de mot). En termes de phonèmes codés avec les 74 symboles retenus, les syllabes les plus fréquentes dans nos corpus sont indiquées dans la table 3.

TAB. 3: Exemples des syllabes les plus fréquentes obtenues par alignement sur le corpus d'apprentissage, pour chaque langue.

all.	[gə] , [tən] , [tə] , [fəʁ] , [bə] , [ən] , [ʁ] , [ti] , [a] , [li]
ang.	[liː] , [ə] , [tiː] ,[ɹiː] , [dɹ] , [tɚ] , [ʃən] , [təd] , [m] , [ɹɪ]
ara.	[tæ] , [tɹ] , [wæ] , [lɪ] , [bɪ] , [ʔæ] , [mæ] , [muː] , [jæ] , [hæː]
esp.	[a] , [do] , [ta] , [te] , [ka] , [ti] , [ra] , [deˌ] , [naˌ] , [da]
fra.	[a] , [de] , [te] , [e] , [li] , [sjõ] , [ti] , [ʁe] , [ʁa] , [mɑ̃]
ita.	[ti] , [ta] , [tɔ] , [te] , [re] , [ri] , [nɔ] , [ka] , [ne] , [ra]
man.	[si] , [tji] , [i] , [tʃi] , [li] , [tjhi] , [tsɪ] , [wu] , [fu] , [pu]
por.	[ɐ] , [du] , [tɐ] , [ti] , [kɐ] , [də] , [tə] , [dɐ] , [rə] , [zə]

3.5. *Unification des syllabes*

La syllabation des 8 langues est fondée sur les lexiques de prononciation propres à chaque langue. On constate dans notre corpus que le nombre de syllabes varie de 400 pour le chinois (les tons sont ignorés) à 3666 pour l'anglais. Le taux de couverture en fonction du nombre de syllabes pour chaque système de syllabe est reporté dans la figure 2. On

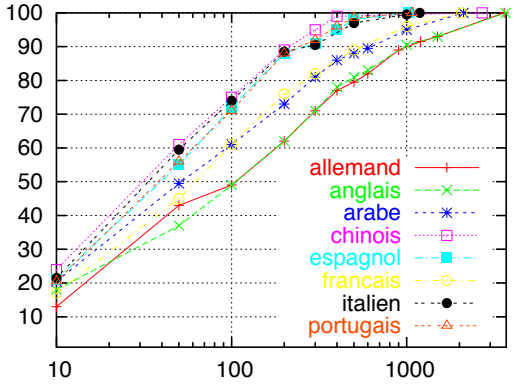

FIG. 2: Taux de couverture syllabique sur le corpus pour chaque langue en fonction du nombre de syllabes.

remarque que les langues romanes méridionales sont très proches dans leur comportement les unes des autres : il y a peu de syllabes différentes, et le taux élevé de couverture est très vite atteint (98 % avec 500 syllabes). Le chinois a un comportement similaire sur le graphique, puisque le taux de couverture est de 99,5 %. L'arabe et le français ont un comportement très proche en matière de couverture syllabique, ainsi que l'anglais et l'allemand.

Afin d'établir un inventaire syllabique multilingue, on choisit alors dans chaque langue un nombre de syllabes de sorte à recouvrir au moins 95 % du corpus pour chaque langue : 400 pour le chinois, 1000 pour l'espagnol, l'italien et le portugais, 1200 pour l'arabe et le français, et 1500 pour l'allemand et l'anglais. Au total pour le lexique de reconnaissance, cela représente 5380 syllabes distinctes, dont certaines sont partagées.

3.6. *Reconnaissance des syllabes*

L'inventaire de « syllabes multilingues » sert de lexique dans le système de reconnaissance syllabique, conformément à la figure 1. Idéalement, un signal de parole de langue $\mathcal{L}$ devrait être transcrit avec des syllabes issues de $\mathcal{L}$. Dans le flux de syllabes décodées automatiquement, la proportion de syllabes étrangères (xéno-syllabes) peut être importante. Nous utilisons comme mesure diagnostique un taux de syllabes propres à la langue décodée. En terme de couverture, la proportion de syllabes propres après reconnaissance automatique des syllabes varie de 84 % pour le français à 30 % pour le chinois, avec pour les langues romanes méridionales et l'arabe environ 50 % ; pour l'anglais et l'allemand environ 65 % – le score moyen étant d'environ 40 %. La faiblesse des scores de reconnaissance automatique des syllabes doit être reliée à la modélisation actuelle des phonèmes complexes (diphtongues, consonnes géminées ou affriquées).

4. EXPÉRIENCES

Le corpus utilisé dans l'expérience suivante comporte pour chaque langue 3 heures de données audio avec leurs transcriptions orthographiques, dont 90 % ont été utilisés pour l'apprentissage. Le test comprend le complémentaire (20 minutes par langue, soit au total 2 heures).

4.1. *Résultats préliminaires*

Des résultats d'identification de la langue par des modèles trigrammes, avec l'approche syllabotactique sont donnés dans la table 4. En moyenne, le taux d'identification sur 7 langues (sans l'espagnol pour lequel le prétraitement a été paramétré différemment par erreur) est de 79 % avec des échantillons de test de 10 secondes. Ce taux monte à 86% pour des séquences de test de 20 secondes. Bien que inférieurs aux performances mesurées avec une approche phonotactique [1], ces premiers résultats avec l'approche syllabotactique sont très prometteurs pour les développements futurs.

TAB. 4: Taux d'IAL avec l'approche syllabotactique pour 7 langues sur des échantillons de 10 et 20 secondes.

config.	*all*	*ang*	*ara*	*fra*	*ita*	*man*	*por*	**moy.**
syll-10s	79%	91%	83%	87%	72%	83%	55%	**79%**
syll-20s	90%	96%	90%	91%	91%	95%	91%	**86%**

4.2. *Apprentissage et quantité de données*

Des expériences ont été menées sur la quantité de données nécessaires à l'apprentissage des modèles phonotactiques ; ils montrent que la quantité de données de 3 heures fixée, correspondant à environ 100 000 phonèmes par langue, est en-deçà des limites théoriques des modèles phonotactique. Le corpus sera étendu à 10 heures par langue, sur 6 langues, en retirant le portugais et le chinois qui représentent moins de données transcrites. La saturation des modèles syllabotactiques n'est clairement pas atteinte, puisque le vocabulaire des modèles contient 5380 syllabes pour des textes d'apprentissage ne comportant pas plus de 50 000 syllabes par langue. D'autre part, le choix du jeu de phonèmes a été optimisé pour l'approche phonotactique. Favorisant l'apparition de xéno-syllabes lors de la phase de reconnaissance, il semble trop détaillé. Les problèmes rencontrés en particulier en chinois soulèvent l'intérêt qu'aurait une approche avec un jeu de phonèmes plus réduit. Nous travaillons actuellement sur un jeu de phones limité, représentant seulement 26 groupes de phonèmes. Les travaux menés dans [1] montrent que la réduction à une vingtaine de classes de phonèmes n'est pas limitative pour l'approche phonotactique.

5. CONCLUSIONS ET PERSPECTIVES

La définition d'un jeu de phones multilingues, par opposition aux jeux de phonèmes monolingues, nous a permis de construire deux systèmes fondés sur des approches segmentales : une approche purement phonotactique avec une base de 74 symboles de phones, et une approche syllabotactique avec une base de 5380 syllabes. Nous avons présenté un algorithme de syllabation et une première comparaison des deux approches. Les résultats encore préliminaires obtenus avec cette première version syllabotactique sont très encourageants. De nombreux paramètres restent cependant à explorer et à optimiser : les dictionnaires de prononciation avec plus ou moins de variantes, différentes conventions pour les diphtongues, affriquées et géminées, différents jeux de phones plus ou moins détaillés, différentes options de syllabation, et bien sûr plus de données et de langues.

RÉFÉRENCES

[1] M. Adda-Decker *et al*. Phonetic knowledge, phonotactics and perceptual validation for automatic language identification. In *Proc. of ICPhS*, Barcelone, 3-9 août 2003.

[2] P. Boula de Mareüil. *Étude linguistique appliquée à la synthèse de la parole à partir du texte*, Thèse de doctorat de l'Université Paris XI, Orsay, 1997.

[3] P. Delattre. *Comparing the phonetic features of English, German, Spanish and French*. Julius Gross Verlag, Heidelberg, 1965.

[4] J.-L. Gauvain, L. F. Lamel. Identification of non-linguistic speech features. In *Proc. of ARPA Workshop on Human Language Technology*, pages 96-101, mars 1993.

[5] D. Kahn. *Syllable-based generalisations in English phonology*. PhD thesis, MIT, 1976.

[6] F. Ramus, J. Mehler. Language identification with suprasegmental cues : a study based on speech resynthesis. *JASA* 105 :512–521, 1998.

[7] J.L. Rouas, Jérôme Farinas, François Pellegrino, Régine André-Obrecht. *Modeling prosody for language identification on read and spontaneous speech*. In *Proc. of IEEE ICASSP*, Hong Kong, 6-10 avril 2003.

[8] F. de Saussure. *Cours de linguistique générale*. Payot, Paris, 1915.

[9] T. Schultz, A. Waibel. Experiment on cross-language acoustic modeling. In *Proc. of Eurospeech*, Alborg, 1-4 septembre 2001.

[10] M.A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. on Speech and Audio Processing*, 4(1), 1996.