Math. & Sci. hum. / Mathematics and Social Sciences (45e année, n° 180, *2007*(4), p. 101–113)

APPRENTISSAGE BOTTOM-UP DES PHONÈMES : UNE ÉTUDE COMPUTATIONNELLE[1],

Rozenn LE CALVEZ[2], Sharon PEPERKAMP[3], Emmanuel DUPOUX[4]

RÉSUMÉ – *Nous présentons une étude computationnelle de l'hypothèse selon laquelle l'information distributionnelle est suffisante pour acquérir les règles allophoniques (et ainsi les phonèmes) de façon* bottom-up. *L'hypothèse a été testée en utilisant une mesure de la théorie de l'information qui compare les distributions. La phase de test a été conduite sur plusieurs corpus de langues artificielles et sur deux corpus de langues naturelles (constitués de transcriptions de parole adressée aux enfants) typologiquement distantes : le français et le japonais. Trois filtres ont été ajoutés à la mesure. L'un concerne la fiabilité statistique due à la taille de l'échantillon, les deux autres prennent en compte les deux propriétés universelles des règles allophoniques suivantes : les éléments d'une règle allophonique doivent être phonétiquement proches, et les règles allophoniques doivent être de nature assimilatoire.*

MOTS CLÉS – Acquisition du langage, Apprentissage statistique, Phonologie

SUMMARY – Bottom-up learning of phonemes: a computational study.
We present a computational evaluation of a hypothesis according to which distributional information is sufficient to acquire allophonic rules (and hence phonemes) in a bottom-up fashion. The hypothesis was tested using a measure based on information theory that compares distributions. The test was conducted on several artificial language corpora and on two natural corpora containing transcriptions of speech directed to infants from two typologically distant languages (French and Japanese). The measure was complemented with three filters, one concerning the statistical reliability due to sample size and two concerning the following universal properties of allophonic rules: constituents of an allophonic rule should be phonetically similar, and allophonic rules should be assimilatory in nature.

KEYWORDS – Language Acquisition, Phonology, Statistical Learning

[1]Cet article est la traduction de la communication "Bottom-Up Learning of Phonemes: A Computational Study", S. Vosniadou, D. Kayser, A. Protopapas (eds.), *Proceedings of the Second European Cognitive Science Conference*, Delphi, Greece, May 23-27 2007, p. 167-172

[2]Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS, DEC-ENS, CNRS) et Université de Paris 6, 29 rue d'Ulm, 75005 PARIS, rozenn.le.calvez@ens.fr

[3]Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS, DEC-ENS, CNRS) et Université de Paris 8, 29 rue d'Ulm, 75005 PARIS, sharon.peperkamp@ens.fr

[4]Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS, DEC-ENS, CNRS), 29 rue d'Ulm, 75005 PARIS, dupoux@lscp.ehess.fr

1. L'ACQUISITION DES RÈGLES ALLOPHONIQUES

Pendant leur première année de vie, les bébés apprennent de nombreux aspects de la phonologie de leur langue maternelle. À la naissance, ils possèdent des capacités universelles de discrimination des segments de parole (unités atomiques correspondant aux consonnes et voyelles) de n'importe quelle langue du monde. Leur perception se spécialise ensuite pour leur langue maternelle : à 6-8 mois [Kuhl *et al.*, 1992], les bébés ont appris les catégories de voyelles de leur langue maternelle et à 10-12 mois les catégories de consonnes [Werker, Tees, 1984]. Les bébés arrivent à ces résultats remarquables avant d'avoir constitué un lexique et avant de commencer à parler. Les mécanismes à l'oeuvre dans cet apprentissage comprennent l'extraction de régularités statistiques présentes dans le signal de parole telles que les distributions de fréquence des segments et les probabilités de transition entre segments [Jusczyk, 1997 ; Maye *et al.*, 2002]. Nous nous intéressons à un aspect de l'acquisition phonologique qui a été peu étudié jusqu'à présent : comment les enfants dépassent-ils la représentation segmentale pour acquérir une représentation plus abstraite des sons, sous forme de phonèmes ?

1.1. PHONÈMES ET RÈGLES ALLOPHONIQUES

Les sons du langage sont représentés à deux niveaux. Au niveau abstrait (sous-jacent), un mot est représenté comme une combinaison d'un ensemble fini de phonèmes. Au niveau de surface, la prononciation d'un mot est spécifiée selon un ensemble plus grand de segments qui dépendent du contexte. Par exemple, en espagnol du Mexique le mot /felis/ (feliz, heureux) se prononce [feliz] lorsqu'il est suivi d'une consonne voisée et [felis] ailleurs.

Les règles allophoniques expriment les réalisations phonétiques d'un phonème en fonction de son contexte. Ainsi, la règle allophonique du voisement en espagnol du Mexique est formulée comme suit :

$$/\mathrm{s}/ \rightarrow \begin{cases} [\mathrm{z}] & \text{avant une consonne voisée} \\ [\mathrm{s}] & \text{ailleurs} \end{cases} \tag{1}$$

Nous appelons [z] l'allophone et [s] le segment par defaut. Ces deux segments n'apparaissent jamais dans les mêmes contextes : on dit qu'ils sont en distribution complémentaire .

Le fait qu'une paire de segments soit impliquée dans une règle allophonique ou non est spécifique à la langue : contrairement à l'espagnol du Mexique, /s/ et /z/ sont deux phonèmes distincts en français si bien que /felis/ et /feliz/ pourraient être deux mots différents en français. Par conséquent, les phonèmes (et les règles allophoniques) doivent être appris au cours de l'acquisition du langage.

1.2. HYPOTHÈSE BOTTOM-UP DE L'ACQUISITION DES PHONÈMES

Quand et comment les phonèmes sont-ils appris ? Cette question reste controversée. Les phonèmes pourraient être appris de façon top-down à l'aide du lexique ou de l'orthographe : à partir de la forme abstraite d'un mot, les enfants apprendraient à lui associer ses réalisations phonétiques [Kazanina *et al.*, 2006]. Une autre hypothèse

est que les phonèmes soient appris très tôt *avant* que les bébés n'aient constitué un lexique, à partir des distributions complémentaires de segments [Peperkamp *et al.*, 2006]. C'est l'hypothèse que nous poursuivons dans cette étude.

1.3. ÉVALUATION COMPUTATIONNELLE

Nous présentons une étude computationnelle de l'apprentissage bottom-up des phonèmes. Nous pensons qu'un nombre significatif de paires allophoniques peut être acquis sans information lexicale. Nous supposons que les bébés sont capables d'extraire les segments à partir de la parole qu'ils entendent, qu'ils peuvent utiliser des mécanismes d'apprentissage statistique, et qu'ils possèdent une métrique de similarité qui leur permet de comparer les segments de leur langue maternelle [Liberman, Mattingly, 1985]. Nous examinons si l'information statistique est suffisante ou si d'autres informations (sous la forme de biais linguistiques) pourraient être nécessaires pour apprendre les règles allophoniques.

1.4. AUTRES TRAVAUX

Des modèles d'induction de règles phonologiques ont été proposés dans différents cadres théoriques. Les linguistes structuralistes ont formulé des procédures permettant de construire à la main les phonèmes à partir d'un ensemble de données [Harris, 1951]. [Johnson, 1984] a présenté une procédure formelle dans un cadre de linguistique générative qui permet de découvrir les règles phonologiques ordonnées. Cependant, aucune de ces approches n'est robuste au bruit.

[Gildea, Jurafsky, 1996] ont proposé un algorithme stochastique qui utilise des tranducteurs à états finis. Ils ont inclus trois biais d'apprentissage qui sont souvent implicites dans les théories linguistiques : la fidélité (les formes de surface sont similaires aux formes sous-jacentes), la communauté (des segments similaires ont tendance à se comporter de façon similaire) et le contexte (l'accès aux règles phonologiques se fait dans leur contexte). Bien que leur approche soit robuste au bruit, le modèle a l'inconvénient d'être supervisé par un professeur virtuel.

Afin de tenter de comprendre comment les enfants pourraient apprendre leur langue, nous avons proposé un algorithme statistique non supervisé. Il recherche les distributions complémentaires à l'aide d'une mesure de la théorie de l'information. Nous avons trouvé que cet algorithme détecte efficacement les paires allophoniques dans un corpus du français à condition que deux filtres linguistiques qui restreignent l'apprentissage aux règles allophoniques universellement possibles soient ajoutés [Peperkamp *et al.*, 2006]. Cette première étude était limitée par certains aspects : la résistance de l'algorithme au nombre de règles et à la complexité n'a pas été étudiée ; les distributions complémentaires parasites dues à la taille limitée du corpus n'étaient pas éliminées ; l'algorithme a été testé sur une seule langue naturelle et les corpus artificiels utilisés (avec des phonèmes équiprobables) étaient très peu réalistes. Dans cette étude, nous ajoutons un filtre de fiabilité statistique qui élimine les paires non fiables en raison de la petite taille de l'échantillon. Les tests ont été effectués sur un panel plus large de langues : des langues artificielles plus réalistes ont été utilisées pour étudier l'influence de la taille du corpus et du nombre de règles ; enfin, deux langues naturelles (français et japonais) ont été étudiées.

Dans la section suivante, nous présentons l'algorithme. Nous l'évaluons ensuite sur plusieurs corpus de langues artificielles. Enfin, l'algorithme est testé sur des corpus de langues naturelles constitués de transcriptions de parole de deux langues typologiquement distantes : le français et le japonais.

2. ALGORITHME

2.1. RECHERCHE DE DISTRIBUTIONS COMPLÉMENTAIRES

L'algorithme recherche les distributions presque complémentaires de segments à l'aide de la divergence de Kullback-Leibler symétrisée (appelée désormais *mesure KL*) qui compare deux distributions de probabilité [Kullback, Leibler, 1951]. Plus spécifiquement nous comparons les distributions de probabilité de deux segments s_1 et s_2 comme suit :

$$m_{KL}(s_1, s_2) = \sum_c \left(P_1 \ \log\left(\frac{P_1}{P_2}\right) + P_2 \ \log\left(\frac{P_2}{P_1}\right)\right) \tag{2}$$

où s_1 et s_2 sont deux segments, c sont les contextes (segment droit, gauche ou les deux) rencontrés dans le corpus, $P_1 = P(c|s_1)$, $P_2 = P(c|s_2)$, $P(c|s) = \frac{n(c,s)+1}{n(s)+N}$ avec $n(c, s)$ le nombre d'occurrences du segment s dans le contexte c, $n(s)$ le nombre d'occurrences du segment s et N le nombre de contextes différents[5].

La mesure est grande pour les paires de segments qui ont des distributions complémentaires. Toutes les paires de segments au-dessus d'un certain seuil (score de $Z > 1$, ce qui correspond à la moyenne des mesures plus un écart-type) sont sélectionnées comme des paires allophoniques potentielles.

2.2. PHONE PAR DÉFAUT OU ALLOPHONE ?

Un critère d'entropie relative détermine les rôles des deux segments de la paire : segment par défaut (qui apparaît globalement plus souvent et dans plus de contextes) ou allophone. Le segment par défaut a la plus petite entropie relative :

$$s_d = \arg\min_s \left[\sum_c P(c|s) \log \frac{P(c|s)}{P(c)} \right] \tag{3}$$

où s sont les deux segments s_1 et s_2 du phonème, et c sont les contextes des segments.

2.3. FILTRE DE FIABILITÉ STATISTIQUE

La fiabilité statistique des estimations de probabilités dépend de la taille du corpus. Nous utilisons un filtre de fiabilité qui élimine les paires non fiables sélectionnées comme des paires allophoniques potentielles par la mesure KL. Le critère Ψ [Jaynes, 2003] compare les fréquences observées à une distribution de probabilité théorique. Il est similaire à un test du Chi^2 mais est valide pour de petits échantillons. Il est défini ainsi :

[5]Nous ajoutons une occurrence de chaque segment au corpus afin d'éviter les probabilités nulles.

$$\Psi(s_1, s_2) = n(s_1) \sum_c f_c(s_1) \ \log\left(\frac{f_c(s_1)}{P(c|s_2)}\right) \tag{4}$$

où $n(s_1)$ est le nombre d'occurrences du segment s_1, $f_c(s) = \frac{n(c,s)}{n(s)}$ avec $n(c, s)$ le nombre d'occurrences du segment s dans le contexte c, $P(c|s_2)$ est l'estimation de la probabilité conditionnelle de c sachant s_2 définie selon l'équation 2.

Le critère évalue donc si les fréquences d'un segment s_1 sont différentes de la distribution de probabilité théorique d'un segment s_2. Si elles ne sont pas considérées comme suffisamment différentes, la paire est éliminée. Nous utilisons un seuil de confiance de 10^{-3} (un apprentissage sur 1 000 échoue) : les paires telles que $\Psi < 3$ sont éliminées comme non fiables.

En calculant ce critère, nous comparons les distributions de segments pris deux à deux. Pour corriger pour le nombre de comparaisons, nous divisons le critère Ψ par le nombre de comparaisons (correction de Bonferroni)[6].

2.3.1. Filtres linguistiques

Dans [Peperkamp *et al.*, 2006], nous avons trouvé que la mesure KL sélectionne des paires allophoniques parasites (fausses alertes) dues aux contraintes phonotactiques (c'est-à-dire distributionnelles) des langues naturelles. Pour les éliminer, deux filtres linguistiques qui prennent en compte les propriétés linguistiques des règles allophoniques ont été ajoutés. D'abord, les paires allophoniques sont constituées de segments phonétiquement proches. En particuler, il ne doit pas y avoir de segment intermédiaire entre eux :

$$\begin{array}{rl} \nexists\ s, & \forall i \in \{1 \dots 6\}, d_i(s_a) \leq d_i(s) \leq d_i(s_d) \\ & \text{or } \forall i \in \{1 \dots 6\}, d_i(s_d) \leq d_i(s) \leq d_i(s_a) \end{array} \tag{5}$$

où s est un segment qui apparaît au moins dans un contexte de l'allophone, s_d est le segment par défaut et s_a l'allophone, $d_i(s)$ est la i^{me} composante de la distance.

Ensuite, les règles allophoniques sont de nature assimilatoire. Ainsi, l'allophone doit être plus proche de ses contextes que le segment par défaut :

$$\forall i, \ |\sum_{C_{s_a}} \Big(d_i(s_a) - d_i(C_{s_a})\Big)| \leq \ |\sum_{C_{s_a}} \Big(d_i(s_d) - d_i(C_{s_a})\Big)| \tag{6}$$

où s, s_d, s_a, d_i sont définis comme ci-dessus, et C_{s_a} sont les contextes de l'allophone.

Pour appliquer les filtres, les segments sont définis selon une échelle numérique d'après leurs propriétés articulatoires. Cette distance est constituée de six dimensions : *place d'articulation* de 1 (bilabiale) à 13 (uvulaire), *sonorité* de 1 (occlusives non voisées) à 12 (voyelles basses), *voisement* (0 ou 1), *nasalité* (0 ou 1), *arrondissement* (0 ou 1) et *longueur* (0 pour les segments simples, 1 pour les géminées et les voyelles longues).

[6]Par exemple, pour 100 ($= 10 \times 10$) comparaisons (environ 10 segments dans la langue), le critère corrigé est de $\Psi = -\log(\frac{10^{-3}}{10^2}) = 5$.

Les filtres linguistiques n'ont pas été utilisés pour les simulations sur les langues artificielles, dans lesquelles les segments sont des symboles arbitraires et pas des segments ayant des propriétés phonétiques. L'algorithme complet est récapitulé en Figure 1.

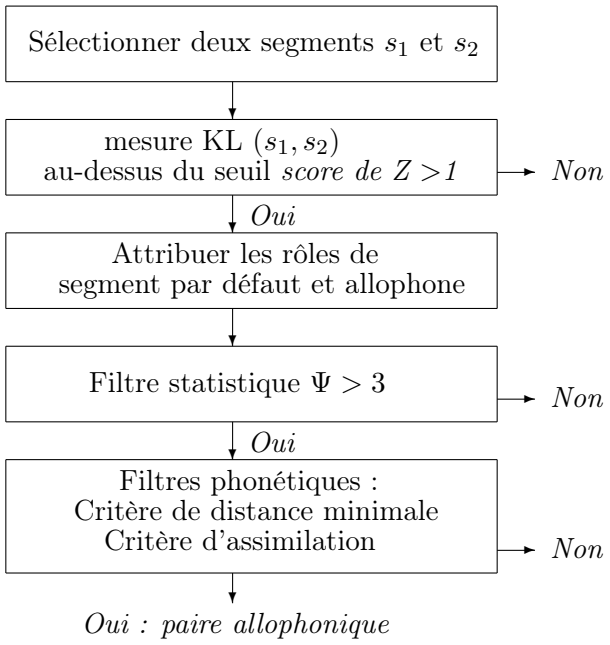

FIGURE 1. Récapitulatif de l'algorithme. La paire sélectionnée est-elle une paire allophonique ?

3. SIMULATIONS AVEC DES LANGUES ARTIFICIELLES

Deux simulations ont été effectuées afin d'évaluer les performances de l'algorithme concernant l'efficacité du filtre statistique et la sensibilité de l'algorithme à deux paramètres importants pour notre problème : la taille du corpus et le nombre de règles allophoniques.

3.1. TAILLE DU CORPUS

3.1.1. Méthode

Des corpus d'une langue artificielle ayant les caractéristiques suivantes, ont été générés : la langue est constituée de 60 segments (ordre de grandeur similaire aux corpus naturels utilisés dans la section suivante) avec un ratio de fréquence de 1 000 entre le segment le plus fréquent et le moins fréquent, et une distribution logarithmique des fréquences. Dans chaque corpus, dix règles allophoniques ont été implémentées. Elles sont déclenchées par des contextes droits déterminés aléatoirement. Six tailles de corpus ont été étudiées de 100 à 10^7 segments ; pour chaque taille, 20 corpus ont été générés.

La performance a été mesurée ainsi : les paires de segments ont été classées suivant leur mesure KL (le rang 1 est assigné à la paire ayant la mesure KL la plus élevée). La performance optimale correspond donc au cas où les paires allophoniques sont classées de 1 à 10. Par conséquent, plus le rang médian des paires allophoniques est élevé, plus la performance est mauvaise.

3.1.2. Résultats

Les résultats sont présentés en Figure 2 : les boîtes à moustaches montrent le rang minimum, les quartiles et le rang maximum.

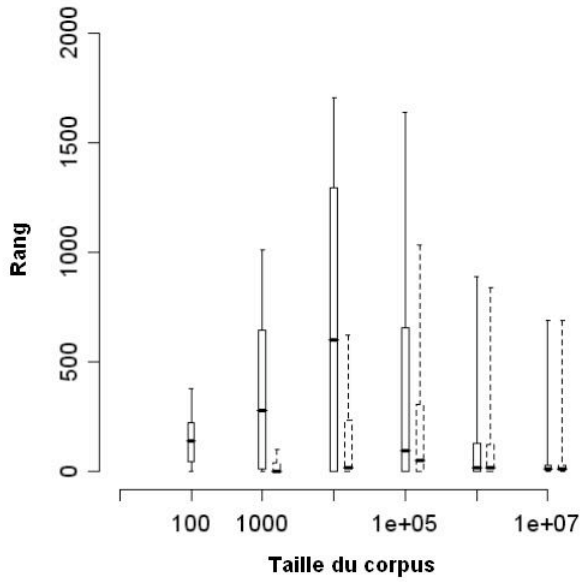

FIGURE 2. Influence de la taille du corpus sur des corpus de 100 à 10^7 segments. Les boîtes à moustaches montrent les résultats obtenus sur 20 corpus aléatoires pour chaque taille. Trait plein (gauche): les rangs des paires allophoniques avant application du filtre statistique. En pointillés (droite) : les rangs des paires allophoniques après application du filtre statistique

Les rangs médians et les quartiles diminuent avec l'application du filtre : le filtre statistique améliore considérablement la performance malgré quelques paires ayant toujours un rang élevé. Le filtre est particulièrement efficace sur les corpus de petite et moyenne taille. Les courbes sont en forme de cloche, que le filtre soit appliqué ou non (avec un maximum autour de 10^5 segments).

Des analyses complémentaires (non présentées) ont montré que la forme de cloche s'explique par deux facteurs : la fréquence des segments et les interactions entre paires allophoniques. En ce qui concerne la fréquence des segments, les paires allophoniques constituées d'un segment rare ne sont pas présentes dans les petits corpus. Dans les corpus de taille moyenne elles sont présentes mais de façon peu fiable, ce qui entraîne une augmentation du rang ; dans les grands corpus, les phonèmes rares sont trouvés de façon fiable et ont un rang bas. En ce qui concerne les interactions,

deux paires allophoniques peuvent par exemple partager le même allophone. L'effet de ces interactions est d'augmenter les rangs des paires allophoniques dans les corpus de taille moyenne. Les deux effets sont réduits après application du filtre statistique.

3.2. NOMBRE DE RÈGLES

3.2.1. Méthodes

La langue artificielle est constituée de 60 segments avec un ratio de fréquence de 1 000 entre le segment le plus fréquent et le moins fréquent, et une distribution logarithmique des fréquences. La taille du corpus a été fixée à une taille fiable de 10^7 segments et le nombre de règles varie dans un intervalle raisonnable pour les corpus de langues naturelles que l'on utilisera ensuite : de 1 à 35 règles sont déclenchées par des contextes droits déterminés aléatoirement. Vingt corpus ont été générés aléatoirement pour chaque nombre de règles. Les simulations ont été effectuées avec application du filtre statistique.

3.2.2. Résultats

Les résultats sont présentés en Figure 3 : les boîtes à moustaches montrent le rang minimum, les quartiles et le rang maximum.

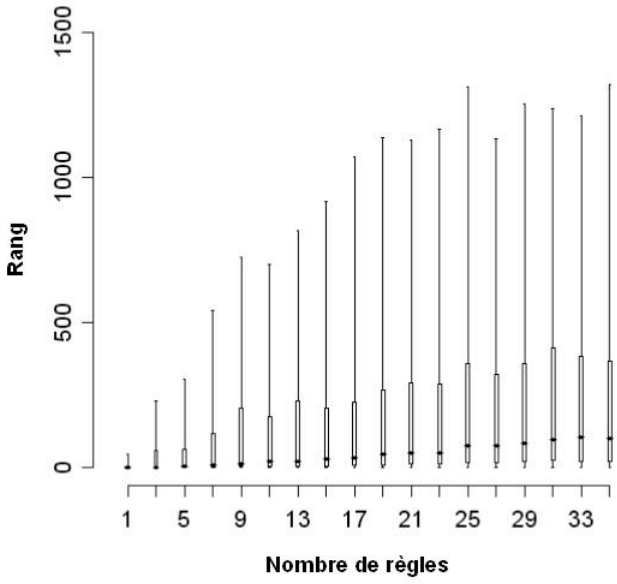

FIGURE 3. Influence du nombre de règles (1 à 35) implémentées dans un corpus aléatoire. Les boîtes à moustaches montrent les résultats obtenus sur 20 corpus aléatoires pour chaque nombre de règles

Les quartiles augmentent avec le nombre de règles. Quelques paires ont toujours un rang élevé. Le plus mauvais rang augmente peu à peu avec le nombre de règles, jusqu'à environ 25 règles. Le rang médian est toujours plus mauvais que sa valeur

optimale, ce qui indique que quelques paires non allophoniques sont toujours mieux classées que les paires allophoniques. Ces paires sont essentiellement constituées d'un allophone et d'un autre segment (allophone ou non) et sont donc le résultat d'un phénomène de confusion allophonique .

Au total, les simulations sur les langues artificielles suggèrent que l'algorithme est assez robuste à la variation de la taille du corpus et du nombre de règles. L'algorithme est particulièrement sensible à trois caractéristiques : la fréquence des segments (les segments fréquents ont tendance à être mieux classés), les interactions entre paires allophoniques dans les corpus de taille moyenne, et la confusion allophonique. Dans les corpus de langues naturelles, les filtres linguistiques réduisent (voire éliminent) les effets négatifs de ces deux dernières caractéristiques.

4. SIMULATIONS AVEC DES LANGUES NATURELLES

Finalement, l'algorithme a été évalué sur des corpus transcrits de parole adressée à des enfants afin d'examiner la performance sur deux langues phonologiquement différentes : le français et le japonais.

4.1. FRANÇAIS

Le corpus est constitué de parole adressée à des enfants issue du corpus CHILDES [Mac Whinney, 2000]. Ce corpus contient des dialogues entre des parents et leurs enfants qui ont été transcrits orthographiquement. Seules les phrases des adultes ont été conservées. Le dictionnaire VoCoLex [Dufour *et al.*, 2002] a été utilisé afin d'obtenir une transcription phonémique du corpus. Onze règles allophoniques ont ensuite été implémentées [Dell, 1973] :

- Les sonantes /r,l,m,n,ɲ,ŋ,ɥ,j,w/ dévoisent lorsqu'elles sont suivies ou précédées d'une consonne non voisée /p,t,k,f,s,ʃ/.
- Les vélaires /k,g/ sont palatalisées quand elles sont suivies de voyelles et de semi-voyelles antérieures /i,y,e,ɛ,ø,œ,j,ɥ,ɛ̃/.

Le corpus semi-phonétique ainsi obtenu est constitué de 45 segments différents dont 11 allophones. Sa longueur est de 43 000 phrases (pour un total d'environ 200 000 segments). L'algorithme a été appliqué au corpus. 432 paires ont été sélectionnées par la mesure KL comme des paires allophoniques, aucune n'a été écartée par le filtre statistique. Parmi les 432, 8 sont des paires allophoniques correctes. Les paires restantes sont des fausses alertes dues à la phonotactique (restrictions dans les distributions des phonèmes de la langue) et à la confusion allophonique.

L'application des filtres linguistiques a éliminé 422 des 424 fausses alertes. Les fausses alertes restantes sont [w̥]-[ɥ] (deux segments appartenant à deux paires allophoniques différentes) et [ə]-[l̥] (dues aux contraintes phonotactiques). L'action des filtres linguistiques est présentée en Figure 4 selon une représentation à deux dimensions de notre distance à six dimensions. Ces deux dimensions montrent grossièrement la place d'articulation sur l'axe horizontal et la sonorité sur l'axe vertical. Sans les filtres, toutes les paires de segments tracées sur la figure ont été sélectionnées

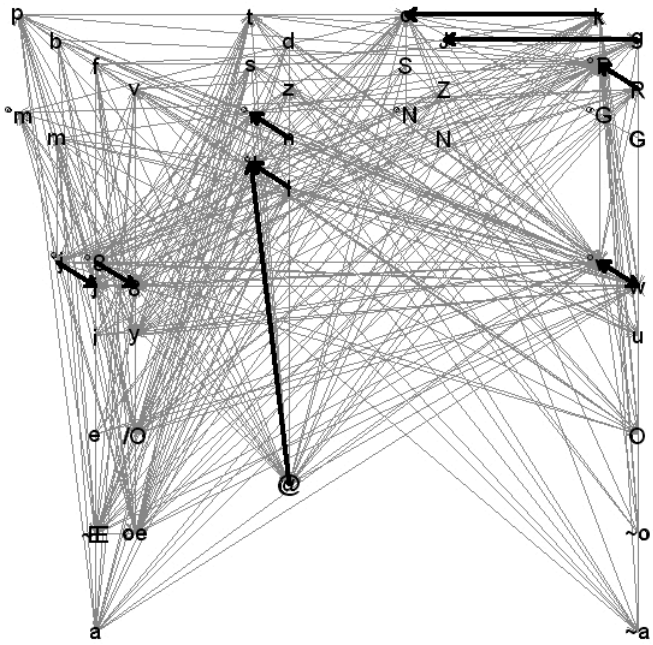

FIGURE 4. Représentation des résultats obtenus avec le corpus CHILDES du français. En noir: paires conservées après application des filtres linguistiques. En gris : fausses alertes éliminées par les filtres

comme des paires allophoniques potentielles. Les filtres ont éliminé toutes les paires tracées en gris et n'ont conservé que les paires tracées en noir. Il est à noter que les segments des fausses alertes (en gris) sont éloignés sur la figure.

Les trois paires allophoniques qui n'ont pas été trouvées par l'algorithme sont [m-m̥], [ŋ-ŋ̊] et [ɲ-ɲ̊]. Les allophones de ces paires sont rares dans le corpus, ont donc une mesure KL faible et n'ont par conséquent pas été sélectionnées.

4.2. JAPONAIS

Le corpus est constitué de parole adressée à des enfants issue du corpus CHILDES du japonais [Mac Whinney, 2000]. Nous avons incorporé dans le corpus un certain nombre de règles phonologiques bien connues (palatalisation, affrication, assimilation nasale). Le corpus ainsi obtenu contient 15 paires allophoniques, dues aux règles allophoniques suivantes :

- /t,d,z/ et leurs géminées deviennent affriqués lorsqu'ils sont suivis de [u].
- /h/ devient [f] lorsqu'il est suivi de [u].
- la nasale moraïque /N/ devient vélaire lorsqu'elle est suivie des consonnes vélaires /k,g/.
- /a,i,u,e,o,aː,iː,uː,eː,oː/ sont nasalisés lorsqu'ils sont suivis de la nasale moraïque /N/.

Le corpus est constitué de 53 segments différents et de 81 000 phrases (environ 800 000 segments au total).

La mesure KL a sélectionné 725 paires allophoniques potentielles, dont 5 ont été éliminées par le filtre statistique. Parmi les 720 paires restantes, 8 sont des paires allophoniques et les autres sont des fausses alertes. Après application des deux filtres linguistiques, seules neuf paires ont été conservées : 8 paires allophoniques et 1 fausse alerte impliquant [h] et [ɴ] (due aux contraintes phonotactiques). L'action des filtres est représentée en Figure 5. Comme pour le français, toutes les paires allophoniques potentielles sélectionnées par la mesure KL sont présentées. Les paires éliminées par les filtres linguistiques sont représentées en gris, les paires passant avec succès les filtres en noir. Les 7 paires allophoniques qui n'ont pas été trouvées (nasalisation des 5 voyelles longues, affrication de /t/ géminé, [h-f]) contiennent des allophones rares.

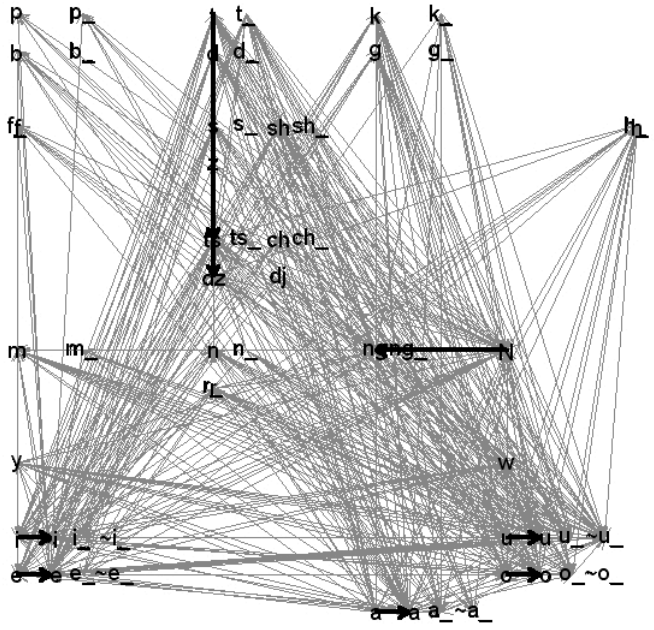

FIGURE 5. Représentation des résultats obtenus avec le corpus du japonais CHILDES. En noir : paires conservées après application des filtres linguistiques. En gris : fausses alertes éliminées par les filtres

4.3. DISCUSSION

L'algorithme complété par les filtres linguistiques obtient de très bonnes performances : il a découvert 8/11 paires allophoniques du français et 7/15 du japonais. Il est à noter que dans les deux langues, certaines règles interagissent. Par exemple en français, plusieurs règles sont appliquées dans les mêmes contextes, ce qui introduit des distributions complémentaires entre segments par défaut et allophones de règles allophoniques différentes (comme [m] et [j̊]). Ces interactions n'ont pas

diminué les performances de l'algorithme car les filtres linguistiques ont éliminé la plupart de ces fausses alertes.

Très peu de fausses alertes ont été conservées : deux pour le français et une pour le japonais. Elles sont dues aux contraintes phonotactiques et à la confusion entre les éléments de paires allophoniques différentes. L'ajout de contraintes sur la participation d'un segment à plusieurs paires allophoniques pourrait aider à les éliminer. Par exemple, on pourrait interdire la sélection de deux paires allophoniques ainsi que d'une troisième paire constituée d'un segment de chacune des deux autres paires. De telles contraintes agiraient sur l'ensemble des paires allophoniques et pas sur les paires allophoniques individuelles. Elles contraindraient ainsi le système phonologique.

En japonais, le filtre statistique élimine plusieurs paires allophoniques, ce qui indique que le corpus est trop petit pour que l'information concernant toutes les paires soit fiable. Il serait nécessaire d'avoir un plus grand corpus pour améliorer les résultats.

5. CONCLUSION

Nous avons présenté un algorithme pour l'apprentissage bottom-up des catégories phonémiques. Les simulations sur les langues artificielles ont étudié l'influence de la taille du corpus et du nombre de règles allophoniques. L'algorithme a été ensuite appliqué à des corpus de deux langues : le français et le japonais. Nous avons obtenu de bonnes performances à condition que trois filtres soient ajoutés à l'algorithme : un filtre statistique qui élimine les paires non fiables en raison d'une taille du corpus trop petite, et deux filtres linguistiques qui contraignent la nature des règles allophoniques. La partie statistique de l'algorithme (recherche de distributions complémentaires) a généré un nombre très important de fausses alertes, dues pour la plupart aux contraintes phonotactiques de la langue. Comme dans [Peperkamp *et al.*, 2006], nous avons montré que ces fausses alertes peuvent être éliminées en utilisant des filtres linguistiques, basés sur une représentation phonétique des segments, qui imposent des contraintes sur les propriétés universelles des règles allophoniques. Cependant, on ne sait pas pour l'instant si une telle représentation phonétique peut être utilisée par les enfants pendant leur première année de vie. Une extension de ce travail sera de remplacer la représentation phonétique conçue à la main par une représentation acoustique. Il pourrait également être intéressant d'acquérir indépendamment les contraintes phonotactiques et d'utiliser ces connaissances pour éliminer les fausses alertes. Finalement, conformément aux théories actuelles sur l'acquisition du langage, cette étude suggère que les bébés peuvent acquérir des connaissances considérables sans lexique à partir de quelques principes computationnels et de biais d'apprentissage appropriés.

Remerciements. Nous tenons à remercier Jean-Pierre Nadal de son aide pour la conception de l'algorithme de recherche de distributions complémentaires. La recherche pour ce travail a été financée par une allocation de recherche du ministère de la Recherche à R. Le Calvez, ainsi qu'un financement de l'Agence Nationale de la Recherche n° ANR-05-BLAN-0065-01.

BIBLIOGRAPHIE

DELL F., *Les règles et les sons*, Paris, Hermann, 1973.

DUFOUR S., PEEREMAN R., PALLIER C., RADEAU M., VoCoLex: une base de données lexicales sur les similarités phonologiques entre les mots français , *L'année Psychologique* 102, 2002, p. 725-746.

GILDEA D., JURAFSKY D., "Learning bias and phonological rule induction", *Computational Linguistics* 22, 1996, p. 497-530.

ZELLIG H., *Methods in structural linguistics*, Chicago, University of Chicago Press, 1951.

JAYNES E.T., *Probability theory: the logic of science*, Cambridge, Cambridge University Press, 2003.

JOHNSON M., "A discovery procedure for certain phonological rules, *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, 1984.

JUSCZYK P., *The discovery of spoken language*, Cambridge (MA), MIT Press, 1997.

KAZANINA N., PHILLIPS C., IDSARDI W., "The influence of meaning on the perception of speech sounds", *PNAS* 103(30), 2006, p. 11381-11386.

KUHL P., WILLIAMS K., LACERDA F., STEVENS K., LINDBLOM B., "Linguistic experience alters phonetic perception in infants by six months of age", *Science* 255, 1992, p. 606-608.

KULLBACK S., LEIBLER R., "Shape description using weighted symmetric axis features", *Annals of Mathematical Statistics* 22, 1951, p. 76-86.

LIBERMAN A.M., MATTINGLY I.G., "The motor theory of speech perception revised", *Cognition* 21, 1985, p. 1-36.

MAC WHINNEY B., *The CHILDES project: tools for analyzing talk*, 3rd edition, Mahwah (NJ), Lawrence Erlbaum Associates, 2000.

MAYE J., WERKER J.F., GERKEN L.A., "Infant sensitivity to distributional information can affect phonetic discrimination", *Cognition* 82(3), 2002, B101–B111.

PEPERKAMP S., DUPOUX E., "Coping with phonological variation in early lexical acquisition", I. Lasser (ed.), *The Process of Language Acquisition*, Frankfurt, Peter Lang, 2002, p. 359-385.

PEPERKAMP S., LE CALVEZ R., NADAL J.-P., DUPOUX E., "The acquisition of allophonic rules: statistical learning with linguistic constraints", *Cognition* 101(3), 2006, B31–B41.

WERKER J., TEES R., "Cross language speech perception: Evidence for perceptual reorganization during the first year of life", *Infant Behavior and Development* 7, 1984, p. 49-63.