Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

ENTRAINEMENT LEXICAL SEMI-AUTOMATIQUE D'UN SYSTEME DE RECONNAISSANCE A BASE SYLLABIQUE

LENNIG Matthew
MERMELSTEIN Paul

Recherches Bell-Northern, 3 Place du Commerce
Ile des Soeurs, Québec, Canada H3E 1H6

Il est souhaitable que les systèmes de reconnaissance de la parole s'adaptent facilement aux nouveaux locuteurs. Le système Harpy est bien avancé à cet égard: on peut l'adapter à un nouveau locuteur en lui faisant prononcer seulement vingt phrases (Klatt, 1977; Lowerre 1977). Cependant, l'adaptation que fait Harpy est limité au niveau des gabarits acoustiques. Harpy n'est pas capable de s'adapter facilement aux nouveaux dialectes qui diffèrent au niveau phonologique.

Dans cette communication nous présentons une méthode d'adaptation au locuteur qui, bien que moins automatique que celle de Harpy, est plus générale. Non seulement est-il possible de modifier les gabarits acoustiques par cette méthode, mais aussi de modifier le lexique en même temps. La possibilité de modifier le lexique pendant le procédé d'apprentissage implique qu'on peut adapter les représentations phonémiques des mots au locuteur.

Le système de reconnaissance de la parole continue que nous avons réalisé utilise la syllabe comme unité de segmentation et de reconnaissance. L'avantage d'un système à base syllabique est que ses gabarits acoustiques incorporent déjà une grande partie de la variation allophonique. Cependant, à cause de la variabilité intra-locuteur dans la production de la parole, la segmentation en syllabe n'est pas toujours faite de la même façon. La méthode d'apprentissage que nous proposons pour l'adaptation au locuteur offre aussi une solution au problème de la variabilité de segmentation.

DESCRIPTION GENERALE DU SYSTEME DE RECONNAISSANCE

Notre système de reconnaissance automatique de la parole continue consiste en quatre composants: un composant de prétraitement, qui extrait les paramètres acoustiques du signal de la parole utilisés pour la reconnaissance, un composant de syllabation, qui segmente la parole parametrisée en syllabes (Mermelstein, 1975), un reconnaisseur de phrase, qui dirige l'exploration de l'espace syntaxique afin de déterminer l'identité la plus probable de la phrase inconnue, et un comparateur syllabique, capable de calculer une mesure de distance (ou de dissimilarité) entre une syllabe inconnue et un gabarit de référence. Pour reconnaître une phrase inconnue, la phrase est d'abord prétraitée et segmentée en syllabes. Puis, le reconnaisseur de phrase dirige une exploration parallèle de tous les sentiers syntaxiques possibles en acceptant une syllabe à la fois de la phrase inconnue et en proposant plusieurs gabarits de référence pour lui être comparés.

En se basant sur les distances cumulatives de plusieurs sentiers parallèles d'analyse syntaxique, le reconnaisseur de phrase est capable d'éliminer certains sentiers d'analyse peu probables. Quand le reconnaisseur arrive à la fin de la phrase d'entrée, le sentier ayant la plus petite distance cumulative est choisie comme étant l'analyse la plus probable.

La méthode de spécification syntaxique que nous utilisons est celle de la grammaire sous forme de Réseau de Transition Augmenté (RTA) proposée par Woods (1970). Le RTA consiste en un système de réseaux de transition récursifs dont les arcs sont capables d'exécuter des actions et de tester des conditions arbitraires.

EXEMPLE: LES EXPRESSIONS DATE-HEURE

Le Graphique 1 représente le niveau le plus élevé d'une syntaxe sous forme de RTA qui accepte les expressions 'date-heure' en français. Ce réseau fait appel à deux sortes de sous-réseaux: les sous-réseaux lexicaux qui spécifient la structure des mots, et les sous-réseaux syntagmatiques, qui spécifient quelles suites de mots peuvent constituer des syntagmes. Par exemple, l'arc PUSH LE/ fait appel à un sous-réseau lexical qui spécifie les suites de syllabes possibles pour les différentes réalisations de surface du mot le. Quand l'arc PUSH LE/ est executé, la commande passe au réseau lexical LE/. Après que le réseau lexical accepte le mot le, la commande retourne au réseau du Graphique 1 dans l'état S/LE.

De l'état S/LE, le sentier d'analyse ou l'hypothèse peut prendre cinq arcs differents. Les arcs PUSH PREMIER/, PUSH VINGT/ et PUSH TRENTE/ font référence à des sous-réseaux lexicaux, tandis que les arcs PUSH TEEN/ et PUSH N29/ font référence à des sous-réseaux syntagmatiques. Le sous-réseau syntagmatique TEEN/ est représenté dans le Graphique 2. Il est capable d'accepter tous les nombres entre onze et dix-neuf en faisant appel à différents sous-réseaux lexicaux. Remarquez que l'arc qui porte l'étiquette JUMP permet qu'une hypothèse se déplace de l'état TEEN/DIX à l'état TEEN/END sans accepter de mot. (Cette trajectoire permet au sous-réseau TEEN/ d'accepter le mot dix.) L'arc POP fait retourner la commande au prochain niveau plus haut. Le Graphique 3 montre le sous-réseau syntagmatique N29/ qui accepte tous les nombres entre deux et neuf.

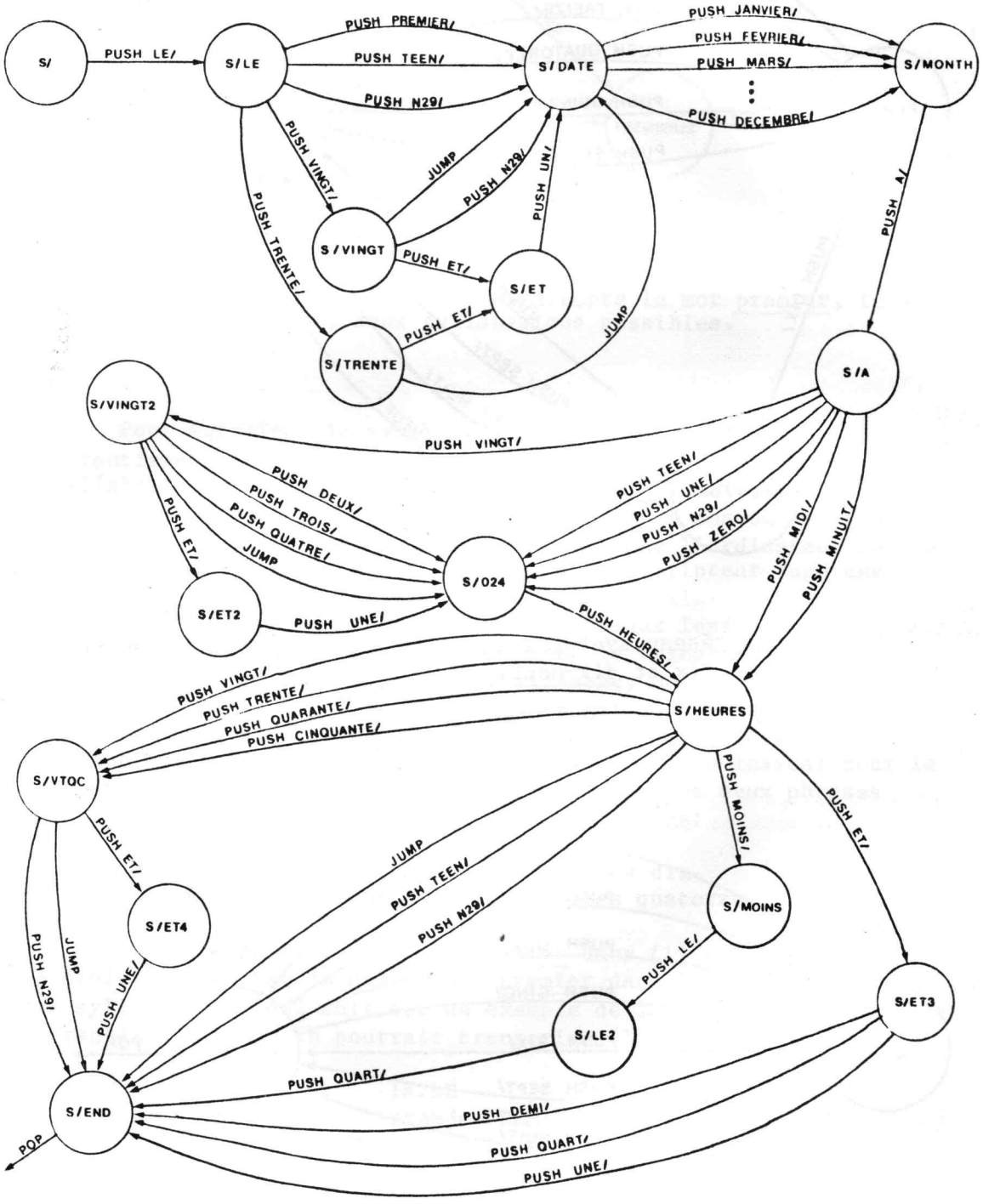
Pour voir comment fonctionnent les sous-réseaux lexicaux, prenons comme exemple celui qui accepte le mot premier, représenté dans le Graphique 4. On voit dans ce graphique que le mot premier peut avoir deux syllabations: il peut être segmenté en deux syllabes, [prə] suivi de [mje], ou bien il peut être segmenté en une seule syllabe: [prəmje]. Ainsi, la variabilité de syllabation peut être prise en main au niveau des sous-réseaux lexicaux.

Chaque fois qu'une hypothèse accepte une syllabe en traversant un arc ACCEPT, le reconnaisseur de phrases fait appel au comparateur syllabique pour calculer la distance acoustique entre la syllabe actuelle d'entrée et le gabarit correspondant à la transcription sur l'arc ACCEPT. Chaque hypothèse garde en mémoire la distance cumulative de toutes les syllabes qu'elle a acceptées.

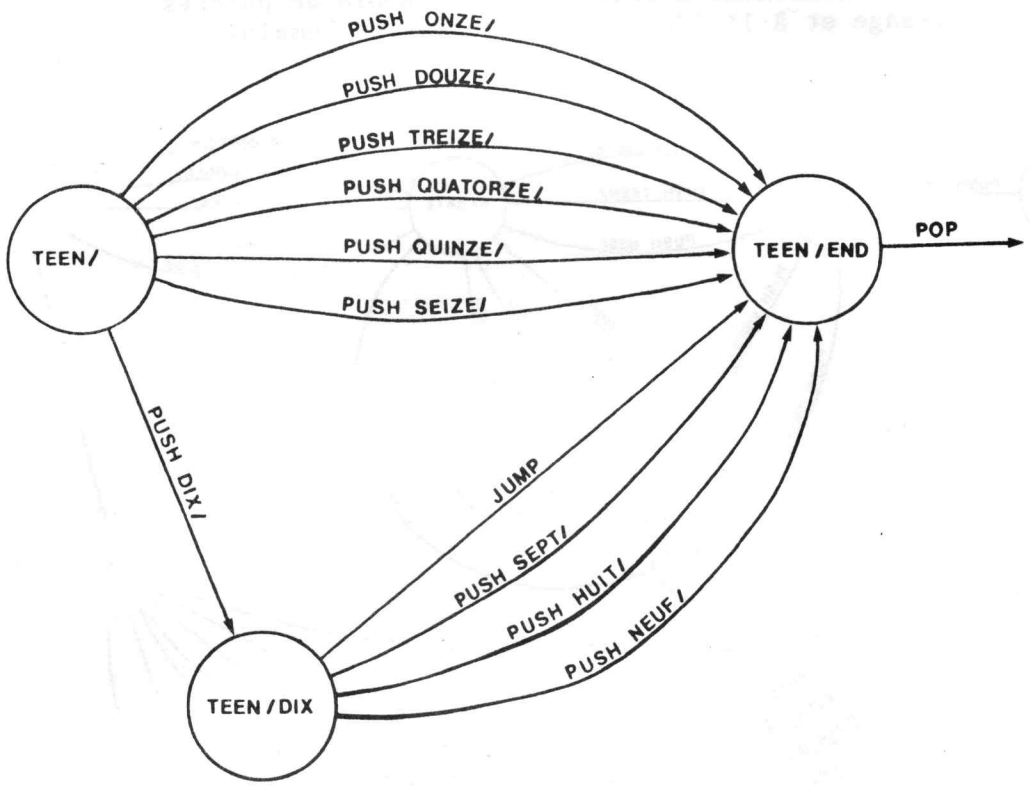
L'ADAPTATION DU LEXIQUE AU LOCUTEUR

L'apprentissage lexical consiste à créer un ensemble de réseaux lexicaux correspondant au vocabulaire du système et un ensemble de

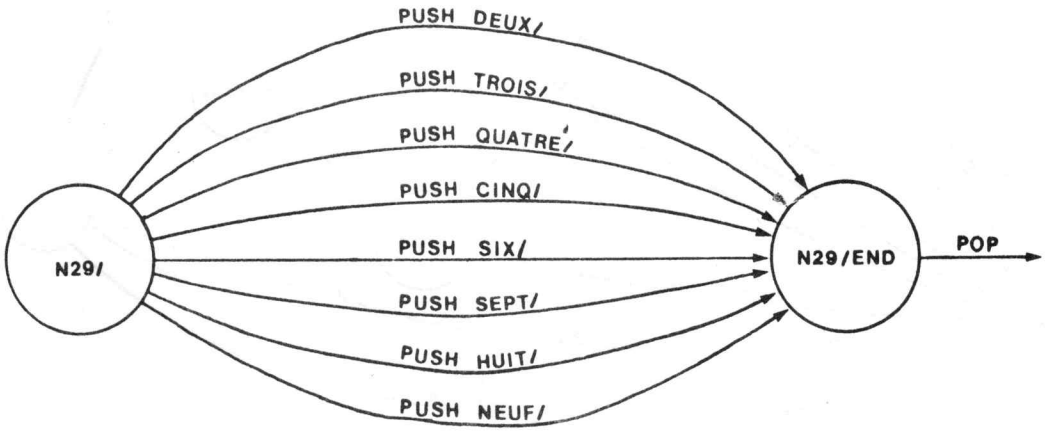
gabarits qui contient un gabarit pour chaque transcription phonétique figurant sur les arcs ACCEPT des sous-réseaux lexicaux. Dans la méthode d'apprentissage semi-automatique, nous générons ces deux ensembles d'informations à partir d'un ensemble de phrases d'apprentissage et à l'aide d'un transcripteur humain.



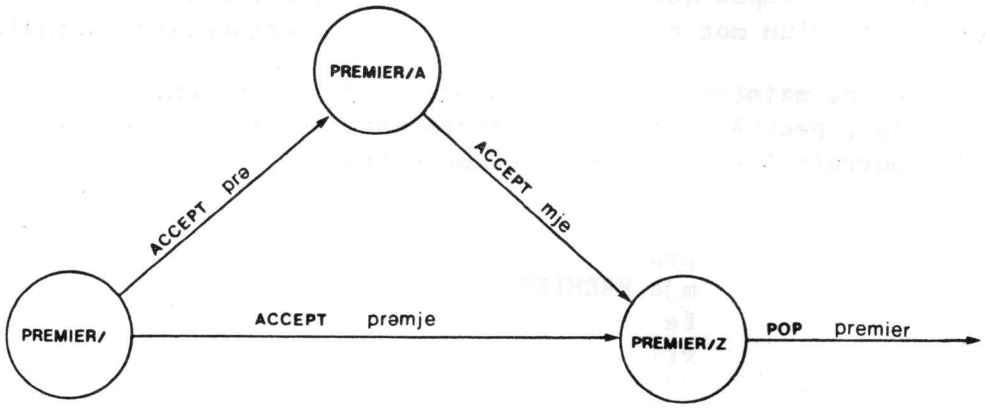
GRAPHIQUE 1. Réseau du niveau de la phrase, faisant partie de la spécification RTA des dates et heures françaises.



GRAPHIQUE 2. Sous-réseau syntagmatique TEEN/ qui accepte les nombres entre dix et dix-neuf.



GRAPHIQUE 3. Sous-réseau syntagmatique N29/ qui accepte les nombres entre deux et neuf.



GRAPHIQUE 4. Sous-réseau lexical qui accepte le mot premier, tenant compte des deux syllabations possibles.

Pour entraîner le système, l'ensemble des phrases d'apprentissage est tout d'abord segmenté en syllabes par le composant de syllabation. Un logiciel d'apprentissage reproduit à travers un haut-parleur le signal acoustique correspondant à chaque syllabe de chaque phrase, une syllabe à la fois. Après que l'ordinateur reproduit chaque syllabe, il pause et attend que le transcripteur tape une transcription phonétique de la syllabe au terminal.

Afin de pouvoir générer les sous-réseaux lexicaux directement, le transcripteur fait entrer non seulement la transcription phonétique de chaque syllabe des phrases d'apprentissage, mais aussi fait-il entrer une indication de la fin de chaque mot aussi bien que son orthographe standard. Nous prenons comme exemple la séquence d'apprentissage nécessaire pour créer le sous-réseau lexical pour le mot premier dans le Graphique 4. Supposons que les deux phrases suivantes figurent dans l'ensemble d'apprentissage:

(1) Le premier décembre à cinq heures dix.
(2) Le premier février à sept heures quatorze.

Supposons aussi que le mot premier dans (1) a été segmenté en une seule syllabe tandis que le mot premier dans (2) a été décomposé en deux syllabes. Ce qui suit est un exemple de la façon selon laquelle le transcripteur humain pourrait transcrire (1):

lə.LE
prəmje.PREMIER
de
sãbr.DECEMBRE
a.A
sɛ̃k.CINQ
œ r.HEURES
dis.DIX

On voit dans cet exemple que le symbole point (.) est utilisé pour signifier la fin d'un mot et qu'il est suivi de l'orthographe normale du mot.

Supposons maintenant que (2) apparaît dans l'ensemble d'apprentissage, peut-être après plusieurs phrases intervenantes. La phrase (2) pourrait être transcrite comme suit:

lə.LE
prə
mje.PREMIER
fe
vri
je.FEVRIER
a.A
sɛt.SEPT
œ r.HEURES
ka
tɔrz.QUATORZE

Nous avons développé un compilateur lexical qui prend comme entrées des données de transcription comme celles-ci et produit comme sortie un lexique qui consiste en un réseau lexical correspondant à chaque mot unique figurant dans l'ensemble d'apprentissage. Le réseau lexical correspondant à chaque mot est capable d'accepter toutes les syllabations de ce mot qui ont apparu dans l'ensemble d'apprentissage.

Un logiciel associé crée un gabarit de référence correspondant à chaque transcription phonétique de syllabe qui peut apparaître sur un arc ACCEPT. Le gabarit consiste en une combinaison des paramètres acoustiques de toutes les syllabes de l'ensemble d'apprentissage qui porte cette transcription.

L'exemple du mot premier démontre une sorte de variation dans la décomposition syllabique: la division ou le manque de division variable à l'intérieur d'un mot. Plusieurs autres exemples de ce phénomène existent et peuvent être pris en compte par le même mécanisme. Maintenant voyons les sortes de variation dans la segmentation syllabique qui présentent le plus de difficultés: celles qui ont lieu à travers une frontière de mot.

MIGRATION PROGRESSIVE DES CONSONNES

Un phénomène fréquent dans la segmentation en syllabes et du français et de l'anglais est le transfert d'une consonne finale ou d'un groupe de consonnes finales au début du mot suivant. Ce phénomène, que nous désignons migration progressive, arrive surtout quand le mot suivant commence par une voyelle. Le Tableau 1 montre des exemples de la migration progressive.

Orthographe normale	Décomposition syllabique			
sept avril	sɛ	ta	vril	
sept avril	sɛt	ta	vril	
quatre heures	kat	trœr		
huit heures	ɥi	tœr		
seize heures	sɛz	zœr		
fifth October	fɪθ	θɔk	to	bə

TABLEAU 1. Cinq exemples français et un exemple anglais de la migration progressive.

Afin de neutraliser ce genre de variabilité de segmentation, nous voudrions, par exemple, que le sous-réseau lexical pour le mot heures accepte la syllabe [trœr] mais seulement si le mot précédent accepté par l'hypothèse se termine en [tr]. Aussi, nous voudrions que le réseau accepte [tœr] comme une réalisation du mot heures, mais seulement au cas où le mot précédent se termine en [t]. De la même façon, le réseau doit accepter [zœr] si le mot précédent se termine en [z]. C'est-à-dire que pour tenir compte de la migration progressive nous sommes obligés d'ajouter de la sensibilité contextuelle à la grammaire. Les actions et les conditions sur les arcs peuvent servir à cette fin.

La façon la plus directe d'arriver à la sensibilité contextuelle nécessaire pour tenir compte de la migration progressive est de faire que chaque hypothèse garde en mémoire le groupe de consonnes finales du dernier mot qu'elle a accepté. Pour ce faire on définit une action qui emmagasine une valeur dans un registre de mémoire que nous appellerons FINAL. Chaque hypothèse aura un registre FINAL qui lui sera associé de la même façon qu'elle a associé avec elle un accumulateur pour garder sa distance cumulative. Nous définissons une action sur l'arc POP du niveau lexical qui met dans le registre FINAL une valeur qui correspond au groupe de consonnes finales du mot. Chaque fois que l'arc POP est executé pour faire retourner la commande au réseau de niveau plus haut, la valeur correspondante au groupe de consonnes finales du mot est emmagasinée dans le registre FINAL. Pour le mot quatre, par exemple, l'arc POP a comme argument le groupe final de consonnes [tr].

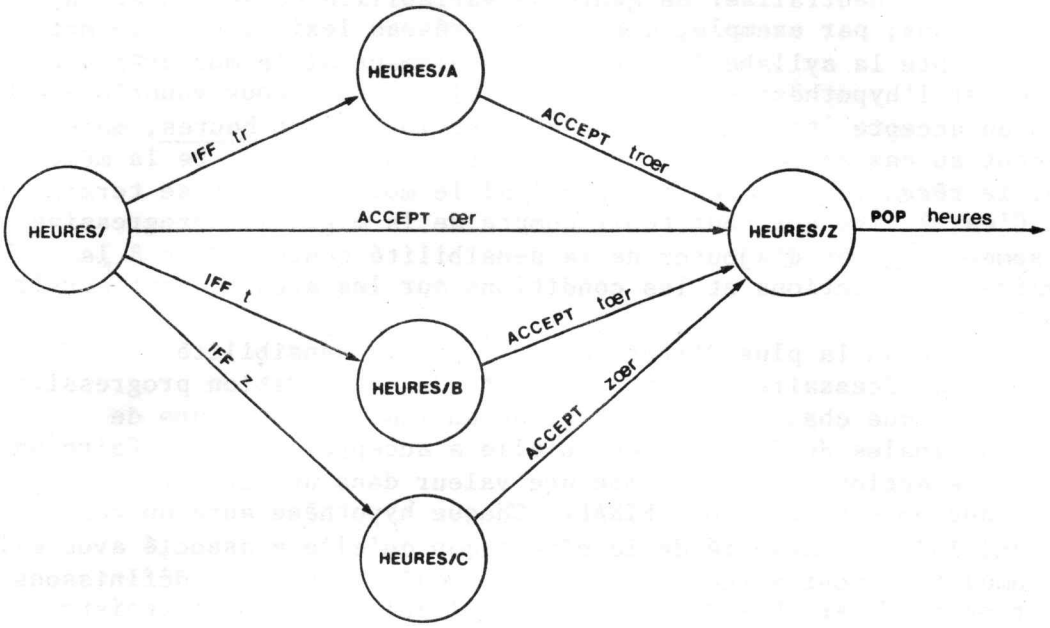
Pour compléter le mécanisme qui tient compte de la migration progressive, nous avons besoin de définir une condition qui teste le contenu du registre FINAL. Nous avons décidé de ce faire en définissant une nouvelle sorte d'arc que nous appelons IF FINAL (abrégé IFF) qui compare la valeur de son argument avec celle du registre FINAL de l'hypothèse qui le traverse. Si FINAL est égal à l'argument de l'arc IFF, l'hypothèse avance à sa destination, exactement comme s'il s'agissait d'un arc JUMP. Si, par contre, le contenu de FINAL n'est pas égal à l'argument de l'arc IFF, l'hypothèse est éliminée. Par exemple, le sous-réseau lexical pour le mot heures pourrait ressembler à celui du Graphique 5. Ce sous-réseau est capable d'accepter quatre formes alternatives du mot heures: [œr], [tœr], [zœr] ou [trœr]. La première de ces transcriptions est accessible à n'importe quelle hypothèse. Les autres exigent que le mot préalablement accepté se termine par un groupe précis de consonnes.

Afin que le compilateur lexical se serve de l'arc IFF et du nouvel argument de l'arc POP pour tenir compte de la migration progressive, nous étions obligés d'ajouter des capacités supplémentaires au procédé de transcription. Ainsi, nous utilisons le symbole virgule (,) pour séparer un groupe de consonnes qui a subi la migration progressive de la première syllabe du mot suivant. Par exemple, la séquence de transcription

.
.
.

kat.QUATRE
tr,œ r.HEURES

.
.
.



GRAPHIQUE 5. Sous-réseau lexical qui accepte quatre variantes du mot heures.

a l'effet de faire apparaître l'argument [tr] sur l'arc POP du mot quatre et de créer un sentier à travers le réseau heures qui accepte la syllabe [trœr] si le mot précédent s'est terminé en [tr] (comme dans le réseau du Graphique 5).

MIGRATION REGRESSIVE DES CONSONNES

Dans certains environnements phonologiques, les frontières syllabiques sont placées de façon qu'elles attachent le groupe initial de consonnes d'un mot à la fin du mot précédent. Cette situation arrive le plus souvent quand le mot précédent se termine par une voyelle. Le problème ressemble à celui décrit ci-haut, sauf que le groupe de consonnes émigre dans la direction inverse. Par exemple, le sept est parfois décomposé comme [ləs] [sɛt].

A cause de son analogie avec la migration progressive, nous avons choisi une notation de transcription similaire pour représenter la migration régressive. Le symbole trait d'union (-) sert à séparer le groupe de consonnes ayant subi la migration régressive du mot précédent. Ainsi, l'exemple le sept serait transcrit comme

.
.
.
lə-s.LE
sɛt.SEPT
.
.
.

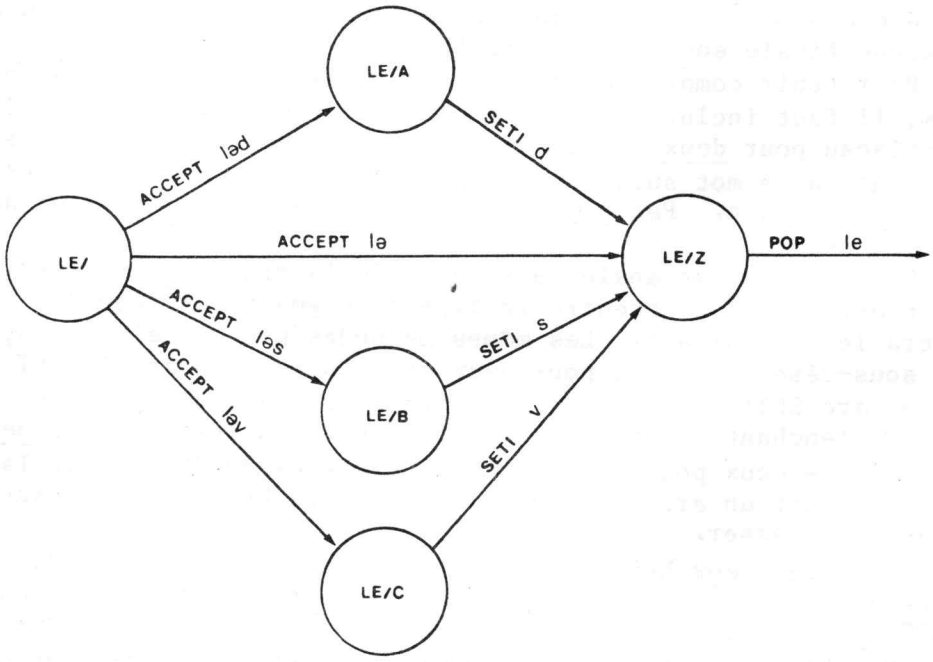
Le trait d'union indique que le [s] fait réellement partie du mot suivant mais qu'il a été attaché au mot le par le composant de

segmentation syllabique. Le compilateur interprète cette notation et se sert de cette information pour construire des sous-réseaux lexicaux pour le et deux qui reflètent la possibilité de cette segmentation. Dans cette partie, nous discutons la représentation de la sensibilité contextuelle nécessaire pour tenir compte de la migration régressive.

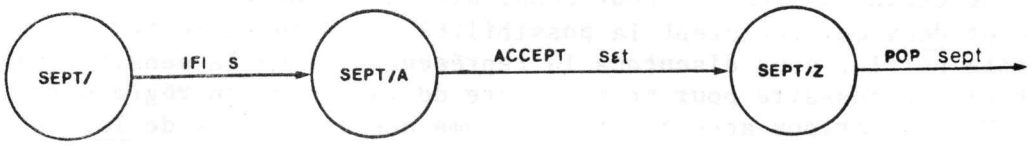
Nous voudrions accepter [ləs] comme une des formes de le, mais seulement si le mot suivant commence par [s]. Au premier abord, ceci paraît paradoxal: comment savoir à l'avance le sentier que suivra une hypothèse? La solution est de permettre que n'importe quelle hypothèse accepte [ləs] et puis de marquer cette hypothèse comme ayant besoin que son prochain mot accepté commence par [s]. Si l'hypothèse essaie d'accepter un mot suivant qui commence par une consonne autre que [s], l'hypothèse est éliminée immédiatement.

Pour marquer une hypothèse comme exigeant que le mot suivant commence par une consonne particulière, on emmagasine une valeur correspondante à la consonne initiale requise dans un registre de mémoire associé avec l'hypothèse. Le nom du registre servant à cette fin est INITIAL. On définit un arc spécial SET INITIAL (abrégé SETI) pour emmagasiner une valeur dans le registre INITIAL. Quand une hypothèse traverse un arc SETI, le reconnaisseur de phrase emmagasine l'argument de l'arc dans le registre SETI de l'hypothèse. Le Graphique 6 montre une configuration possible du réseau lexical du mot le qui utilise des arcs SETI pour forcer l'occurrence de certaines consonnes initiales dans le mot suivant. Si la forme non marquée [lə] est acceptée, le registre INITIAL est implicitement remis à zéro et n'importe quel mot peut suivre.

Afin de tester le contenu du registre INITIAL, nous définissons l'arc IF INITIAL (abrégé IFI). L'arc IFI est traité comme un arc JUMP si son argument est égal à la valeur du registre INITIAL de l'hypothèse qui essaie de la traverser. Sinon, l'arc bloque et l'hypothèse est éliminée. Le Graphique 7 illustre l'emploi de l'arc IFI dans le sous-réseau lexical pour le mot sept.



GRAPHIQUE 6. Sous-réseau lexical qui accepte différentes formes du mot le. Ce réseau illustre l'utilisation de l'arc SETI.



GRAPHIQUE 7. Sous-réseau lexical pour le mot sept illustrant l'emploi de l'arc IFI pour spécifier que la consonne initiale du mot est [s].

En employant des actions et des conditions sur les arcs, nous avons réussi à ajouter à nos descriptions syntaxiques la sensibilité contextuelle nécessaire pour tenir compte de la migration et progressive et régressive. Dans la prochaine partie nous verrons que, sans définir aucun nouveau mécanisme, nous pouvons tenir compte sur un niveau rudimentaire de la liaison française.

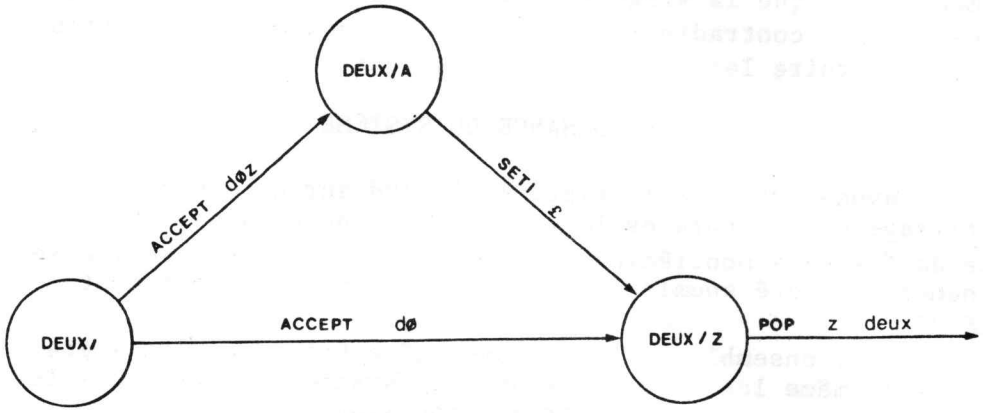
LA LIAISON FRANCAISE

Un exemple de la liaison se produit dans l'expression deux heures. Puisque [z] est la consonne finale sous-jacente du mot deux, il semble raisonnable d'essayer de la spécifier dans l'argument de la consonne finale de l'arc POP du sous-réseau lexical pour le mot deux. Le mécanisme de la spécification des consonnes finales sur les arcs POP existe déjà pour tenir compte de la migration progressive. Comme le mot deux se termine par une voyelle dans sa forme habituelle, l'argument de la consonne finale de l'arc POP reste libre pour emmagasiner la consonne finale sous-jacente. Un des sentiers traversant le sous-réseau pour heures consistera en un arc IFF z suivi d'un arc ACCEPT zœr. Ce sentier sera accessible à toute hypothèse qui a accepté comme mot précédent un mot se terminant en [z] ou bien ayant une consonne finale sous-jacente de [z].

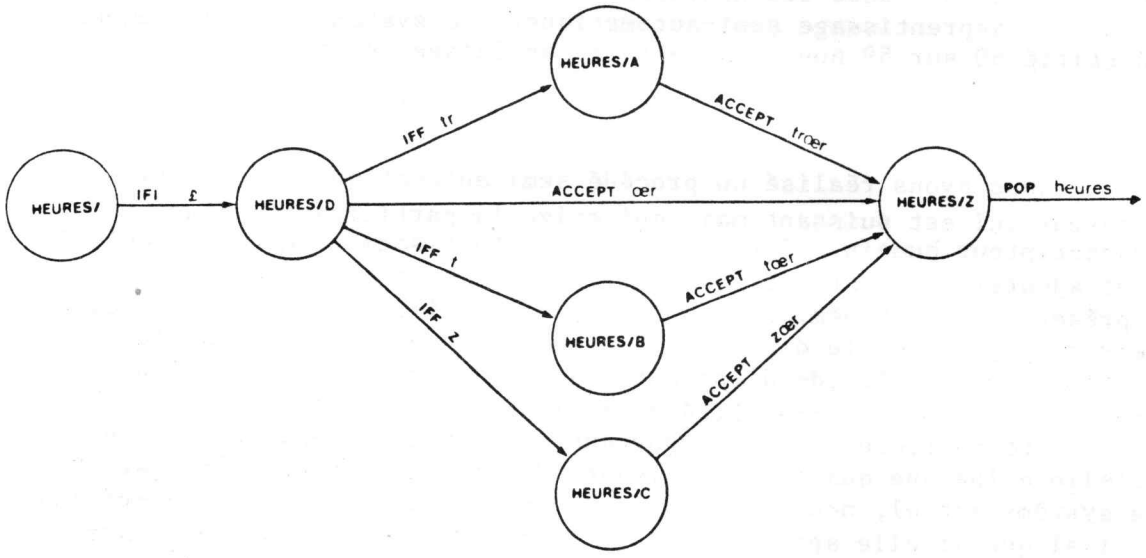
Pour tenir compte des cas où le [z] est prononcé à la fin du mot deux, il faut inclure un sentier qui accepte la syllabe [døz] dans le sous-réseau pour deux. Cependant, nous ne voulons pas accepter [døz] au cas où le mot suivant ne commence pas par un phonème qui déclenche la liaison. Par exemple, la séquence [døz] [me] pour deux mai est impossible.

Ce problème est analogue à celuį de la migration régressive puisqu'on est obligé de prédire le type de segment par lequel commencera le mot suivant. Les mêmes méthodes peuvent s'appliquer. Dans le sous-réseau lexical pour deux on se sert d'un arc ACCEPT døz suivi d'un arc SETI £, où £ est un symbole spécial qui signifie 'segment déclenchant la liaison'. Le sous-réseau lexical pour heures (aussi bien que ceux pour tous les mots capables de déclencher la liaison) contient un arc IFI £ par lequel toute hypothèse entrant dans le réseau doit passer.

L'exemple deux heures est illustré dans le Graphique 8. Le Graphique 8A montre le sous-réseau lexical pour le mot deux, y compris l'arc SETI £ et la spécification de la consonne sous-jacente sur l'arc POP. Le Graphique 8B montre le sous-réseau lexical révisé pour heures avec l'addition de l'arc IFI £ pour indiquer que heures peut déclencher la liaison. Nous voyons qu'en utilisant des mécanismes déjà nécessaire



GRAPHIQUE 8A. Sous-réseau lexical pour le mot deux qui tient compte de sa forme de liaison.



GRAPHIQUE 8B. Sous-réseau lexical pour le mot heures, révisé pour tenir compte de la liaison.

pour tenir compte de la migration des consonnes nous pouvons également tenir compte d'une façon rudimentaire de la liaison.

Nous avons adopté une convention spéciale pour indiquer les occurrences de liaison dans la transcription d'apprentissage. On transcrit la consonne de liaison comme si elle n'appartenait à aucun des deux mots. Par exemple, si les mots deux heures étaient segmentés comme [dØz] [zœ r], alors la transcription d'apprentissage serait

.
.
.

dØ-z.DEUX
z,œ r.HEURES

.
.
.

Le trait d'union indique que le z ne fait pas partie du mot deux en même temps que la virgule indique qu'il ne fait pas partie du mot heures. Cette contradiction de notation signale au compilateur lexical de construire les arcs nécessaires pour la liaison.

LA PERFORMANCE DU SYSTÈME

Nous avons entraîné le système d'abord sur un ensemble d'apprentissage de 100 phrases 'date-heure' prononcées par une locutrice du français montréalais. Quand les 100 phrases de l'ensemble d'entraînement ont été soumises au système, il en a correctement identifié 96.

Un nouvel ensemble de 100 phrases générées au hasard a été prononcé par la même locutrice et soumis au système. Cette fois le système a correctement identifié 76 des 100 nouvelles phrases.

Pour démontré la généralité du procédé d'apprentissage, nous l'avons utilisé pour adapter le système à reconnaître non pas un autre dialecte, mais une autre langue, cette fois choisissant un locuteur de l'anglais britannique. Pour que le système accepte les dates et les heures anglaises, il suffisait de spécifier manuellement les réseaux syntagmatiques: tous les nouveaux sous-réseaux lexicaux étaient créés lors de l'apprentissage semi-automatique. Le système a correctement identifié 50 sur 59 nouvelles phrases anglaises, soit 85%.

CONCLUSIONS

Nous avons réalisé un procédé semi-automatique d'adaptation au locuteur qui est puissant mais qui exige la participation d'un transcripteur humain. Cette méthode est suffisamment générale qu'elle peut ajouter du vocabulaire au lexique ou bien ajouter des représentations phonémiques au vocabulaire existant. Ainsi peut-elle servir de modifier le dialecte ou même le langage reconnu par le système. Cette méthode d'apprentissage incorpore dans le lexique une connaissance de la variabilité de segmentation.

Le problème le plus important de la méthode décrite ici est qu'elle exige une quantité suffisante de données d'apprentissage. Dans le système actuel, nous n'incorporons une variante dans un sous-réseau lexical que si elle apparaît dans l'ensemble d'apprentissage. Parce qu'on voudrait pouvoir entraîner un système avec un minimum de données d'apprentissage, il serait profitable de chercher des algorithmes qui pourront se généraliser à d'autres segmentations possibles à partir d'un ensemble limité d'apprentissage.

BIBLIOGRAPHIE

KLATT (D.H.), 1977, "Review of the ARPA speech understanding project". J. Acoust. Soc. Am. 62, pp. 1345-1364.

LOWERRE (B.T.), 1977, "Dynamic speaker adaptation in the Harpy speech recognition system". Conference Record of the 1977 IEEE International Conference on Acoustics, Speech and Signal Processing, Hartford, 9-11 May, pp. 788-790.

MERMELSTEIN (P.), 1975, "Acoustic segmentation of speech into syllabic units". J. Acoust. Soc. Am. 58, pp. 880-883.

WOODS (W.A.), 1970, "Transition network grammars for natural language analysis". Comm. Assoc. Computing Machinery 13, pp. 591-606.