# Prédiction et vérification lexicale dans le cadre d'un dialogue oral homme-machine

Laurent Romary, Bernard Mangeol

▶ **To cite this version:**

Laurent Romary, Bernard Mangeol. Prédiction et vérification lexicale dans le cadre d'un dialogue oral homme-machine. XVIIèmes Journées d'Etude sur la Parole, 1988, Nancy, France. hal-00521619

**HAL Id: hal-00521619**
**https://hal.science/hal-00521619**

Submitted on 18 Nov 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

JEP 88.

# Prédiction et vérification lexicale dans le cadre d'un dialogue oral homme-machine

Laurent ROMARY - ESE/CRIN/INRIA
Bernard MANGEOL - CRIN/INRIA

BP 239, 54506 Vandœuvre.
(romary@crin.UUCP)

## Abstract

As a central part of the man-machine oral dialogue system under development at the CRIN at Nancy, we present here our conception of its lexical component. We first show the importance of linguistic knowledge as a guide to word recognition, and particularly, how contextual information can be obtained at any step of a dialogue in a task oriented application. Then, we present some techniques of combining hypotheses thanks to a specific lexical representation and the use of the Dempster-Shafer theory for combining word evidence. At last, we describe the prediction and verification of word presence along the speech signal through the use of macro-classes of phonemes and dynamic programming algorithms.

## 1. Introduction.

Malgré la relative continuité des recherches en intelligence artificielle ces dernières années, de nouvelles façons d'analyser les problèmes qui touchent ce domaine tendent à faire évoluer celui-ci de manière profonde, grâce aux efforts conjoints de nombreuses disciplines regroupées sous le terme général de sciences cognitives. En particulier, un système dit intelligent est de moins en moins vu comme un ensemble clos où se trouvent centralisées toutes les informations et les décisions, mais plutôt comme un univers où plusieurs entités coopèrent pour permettre la réalisation d'une fonction particulière vis à vis du monde extérieur. Ce paradigme permet, par exemple, la considération d'environnements multi-experts, ou la gestion de l'interaction entre plusieurs agents autonomes (des robots) dans un univers physique particulier.

Cette évolution se retrouve dans les travaux touchant la reconnaissance de la parole où l'on parle maintenant de dialogue oral homme-machine en considérant, non plus un système de reconnaissance indépendant du monde extérieur qui tente (parfois vainement) de comprendre un signal qui lui est présenté en entrée, mais un agent intelligent, destiné à converser avec d'autres usagers et pour cela, convié à intégrer sa composante de reconnaissance dans une description plus générale de l'univers qui l'entoure. Cela nécessite, au niveau d'un tel système, la mise en place d'un double mécanisme, à savoir la prise en compte de nouveaux éléments à chaque interaction avec l'extérieur et l'utilisation de ces objets au niveau du module de compréhension pour optimiser son analyse des énoncés à venir.

C'est dans cette optique que nous avons conçu l'architecture du système de dialogue oral homme-machine en cours de développement au CRIN à Nancy [Carbonell 87][Pierrel 87], dont l'organisation interne est elle-même particulièrement modulaire. Nous n'allons détailler ici qu'une certaine vision du lexique, en regardant comment celui-ci échange des informations avec le reste du système, ainsi que les choix de techniques que nous avons été amenés à faire pour intégrer ces informations et réaliser des prédictions sur des mots, à partir de la représentation phonétique d'un énoncé.

## 2. La place d'une composante lexicale dans un système de dialogue oral homme-machine.

Les informations lexicales sont les premiers éléments linguistiques, manipulés par le système, qui ne soient pas propres à la communication orale. En effet, tous les niveaux inférieurs, filtrage, transformées, et décodage acoustico-phonétique travaillent essentiellement sur des informations acoustiques et ceci, d'une manière purement ascendante, du signal jusqu'à une représentation sous forme de phonèmes ou de syllabes. L'expérimentation relative à Hearsay II avait déjà montré ce résultat en adoptant une structure des sources de connaissance dans le Blackboard où le premier grand pôle d'échanges se situait au niveau du mot [Erman 80].

Une entité lexicale peut en effet être vue comme un point de convergence d'informations de types divers, comme un patron phonétique, une classe (ou une structure) grammaticale, et une représentation sémantique. Chacune d'entre elles peut servir de point d'accès au lexique tout entier ou, dans le cas de la reconnaissance d'un énoncé, d'espace de décision particulier en fonction des informations disponibles dans le système.

On considère communément que l'impossibilité d'obtenir un décodage acoustico-phonétique d'excellente qualité, impose aux niveaux supérieurs de limiter progressivement les cas d'ambiguité entre les différentes interprétations possibles du signal. Cependant, nous allons voir que les niveaux linguistiques peuvent posséder un rôle analogue à l'étape de décodage, au sens où eux aussi peuvent proposer certains éléments du lexique qui s'avèrent pertinents à chaque stade de la reconnaissance. Afin de comprendre l'origine de ce processus, il est nécessaire de décrire les connaissances qui vont être utilisées à différents niveaux d'analyse et que nous nommerons de manière générale le *contexte*.

Le contexte n'apparaît pas ici comme une entité externe aux intervenants d'un dialogue, comme parfois il peut être défini en pragmatique [Latraverse 87], entité qui représenterait toutes les conditions préalables dans lesquelles s'insère une suite d'énoncés. Cette vue du contexte n'a aucun sens si l'on désire modéliser le comportement d'un agent intervenant dans le dialogue. En effet, ce qui importe est l'état "mental" d'un des protagonistes et l'influence qui en résulte au niveau de la reconnaissance. Le contexte est donc l'ensemble des connaissances du système à un instant donné de son existence, qui forme son espace de représentation du monde extérieur.

Nous allons raisonner dans le cadre précis de l'application envisagée qui est l'interrogation par un utilisateur d'un centre de renseignements administratifs simulé par la machine (correspondant sensiblement aux informations disponibles dans les pages roses d'un annuaire). L'architecture envisagée (cf fig.1) fait ressortir quatre processeurs indépendants qui interagissent avec le lexique (LEXIQUE), du décodeur acoustico-phonétique (APHON) au module de dialogue (DIALOGUE), en passant par le détecteur d'indices prosodiques (PROSODIE) et les analyseurs syntaxico-sémantiques (ANALYSEURS). A chaque niveau, des informations de plus en plus précises sont disponibles qui toutes entrent dans la définition du contexte.

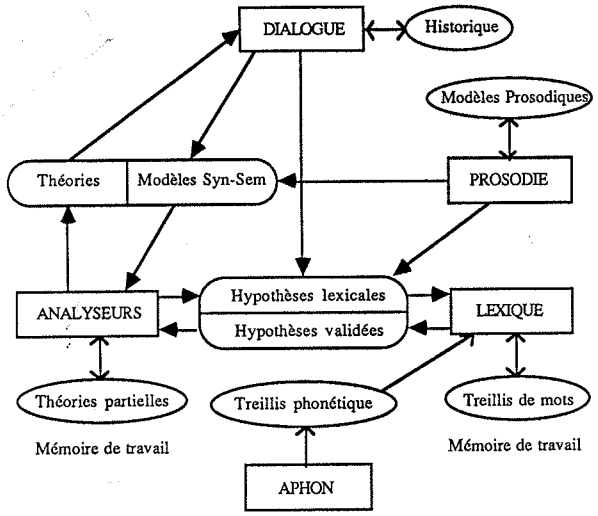



Figure 1.

Le module de dialogue gère en continu les échanges entre le locuteur et la machine. En particulier, il dirige partiellement le dialogue et déclenche les phases d'acquisition d'un nouvel énoncé. En début de dialogue, il va ainsi pouvoir prédire un énoncé de salutation en réponse à une introduction de la part de la machine du type "*Centre de renseignements administratifs, Bonjour!*". Puis un sous-lexique particulier relatif à une demande d'informations va pouvoir être proposé (typiquement : "*Je voudrais...*"), correspondant à l'expression de la requête de l'usager. Enfin, après satisfaction de celui ci, on peut s'attendre à deux types d'énoncés auxquels correspondent deux sous-lexiques particuliers, à savoir un énoncé de relance ou de fin de dialogue ("*Je vous remercie, au revoir.*").

Ces connaissances générales au niveau d'un dialogue peuvent s'affiner à chaque échange puisque progressivement le thème général du discours se définit, comme par exemple une demande d'information relative au renouvellement d'une carte d'identité, ou l'accession à la nationalité française. Localement, seul un sous-ensemble restreint du lexique relatif à ce thème peut être mis en exergue, tout en autorisant éventuellement l'usage de mots moins probables, en cas de rupture de séquence, ou d'énoncés concernant la gestion du canal de communication si l'usager désire préciser une réponse de la machine ("*Pouvez vous répéter, s'il vous plaît*"). Plus précisément, l'usage de certaines structures ou expressions par le locuteur peut être fortement conditionné par un énoncé particulier de la machine comme une question du type :

- *Où habitez vous ?*

où la structure standard de réponse est du type :

- *J'habite à [Nancy]*

ou de façon plus elliptique :

- *à [Nancy]*

Cette grande corrélation entre deux énoncés successifs permet de restreindre de manière importante l'espace d'analyse au niveau du lexique, et bien sûr, de façon plus générale au niveau des structures de la langue.

Les prédictions plus précises au niveau de la structure détaillée d'un énoncé ne concernent plus le module de dialogue, mais plutôt les deux modules d'analyse syntaxico-sémantique et de prosodie. Ce dernier processeur fournit à l'analyseur lexical des frontières possibles de mots, sans être en mesure de préciser la nature exacte de ces mots. Les hypothèses fournies sont donc purement temporelles. Les analyseurs peuvent, quant à eux utiliser des résultats relativement sûrs, obtenus à partir de la reconnaissance d'une partie d'énoncé pour induire des hypothèses sur les éléments restants. Si par exemple le système a reconnu la portion d'énoncé :

*J'habite ...*

Une hypothèse sémantique va pouvoir être générée concernant un sous-lexique de lieu. Ceci peut se faire dans la pratique grâce à des contraintes sémantiques sur les constructions possibles autour du verbe "*habiter*". Ce phénomène de déclenchement sémantique est d'ailleurs connu en psychologie dont les expériences apportent beaucoup à notre approche [Heyer 85].

La phrase suivante montre un cas plus complexe d'analyse faisant intervenir la représentation du monde qu'a la machine en cours d'analyse :
"*Je suis Marocain et je désirerais renouveler ma carte de séjour.*"
L'analyse du début de l'énoncé permet de qualifier l'élément en mémoire qui représente le locuteur avec l'attribut "Marocain", ce qui restreint l'ensemble des demandes possibles de papiers administratifs susceptibles d'être émises par celui-ci. Ce type d'analyse impose que l'énoncé soit interprété dès le début de sa reconnaissance, ce qui n'est pas encore réalisé au stade actuel de développement du système.

Le problème se pose maintenant d'intégrer ces informations dans le cadre d'une analyse particulière. En effet, la relative généralité du domaine donne une taille importante au lexique non-grammatical, aussi est-il nécessaire de posséder un mécanisme de sélection assez efficace pour que le temps et la qualité de la reconnaissance restent raisonnables.

## 3. Intégration d'une hypothèse descendante.

Afin de conserver une certaine souplesse à la reconnaissance, nous avons choisi de préserver la disponibilité de tout le lexique à chaque instant, de sorte qu'un apport d'information relatif à un sous-lexique particulier ne fasse que renforcer celui ci. Nous supposons ici qu'une hypothèse lexicale est par définition précise et sélectionne un sous-ensemble vrai de l'ensemble total des mots. Pour cela le lexique est structuré par avance sous forme d'une arborescence multiple correspondant à une organisation syntaxique ou sémantique particulière. Au niveau sémantique nous avons adopté une grammaire de cas étudiée par Guy Deville et Hans Paulussen [Deville 87], qui structure les mots prédicatifs (verbes, noms ou adjectifs) en fonction de primitives définies à partir de traits sémantiques de base. Pour illustrer ceci, nous pouvons donner une représentation partielle des informations relatives à "*signer*" et de certains de ses voisins.

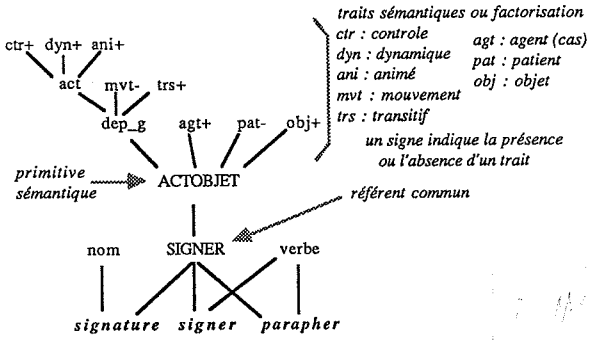



Figure 2.

Au sens défini ci-dessus chaque nœud de cette structure permet de désigner un sous-lexique particulier, et grâce à des opérations ensemblistes élémentaires, il est possible d'exprimer la plupart des contraintes qui nous semblent utiles au niveau du système. Ainsi, une expression du type :

(inter (non verbe) (union ctr+ ben+))

désigne tous les éléments du lexique autres que verbes qui expriment une action contrôlée ou acceptant un bénéficiaire.

Une hypothèse lexicale va donc pouvoir être décrite sous le format suivant :

(*Sous-lexique Score Identifiant Plage-phonétique*)

où *Sous_lexique* est une expression ensembliste signalée ci-dessus, *Score* une évaluation de la qualité de l'hypothèse, *Identifiant* une marque propre à l'émetteur de l'hypothèse et qui permet à celui-ci de la retrouver en cas de validation. Enfin, *Plage-phonétique* est un intervalle au sens large, indiquant la zone temporelle où l'hypothèse est effective.

L'intégration de telles hypothèses se fait relativement simplement grâce à la théorie de Dempster Shafer adaptée ici à un type particulier de distribution de vraisemblance (on peut se référer à [Barnett 83] pour une bonne introduction à cette théorie). Chaque hypothèse accompagnée d'un score va être combinée à la distribution initiale existant au niveau du lexique

en considérant celui-ci comme deux sous-ensembles, à savoir le sous-lexique désigné et son complémentaire.

Avant de détailler le mécanisme de combinaison, il est nécessaire de signaler la provenance de la distribution sur le lexique a priori. Nous sommes partis d'un histogramme de présence des entrées lexicales dans un corpus relatif au même domaine d'application en situation réelle. L'information fréquentielle résultante nous donne une distribution probabiliste vraie au niveau de chaque mot. Cette information peut ensuite être remontée par sommation le long de l'arborescence, pour obtenir ainsi la masse totale d'incertitude que représente un sous-lexique particulier désigné par une hypothèse.

Le problème d'agrégation d'une hypothèse sur la base lexicale peut alors se formaliser ainsi :

Une distribution initiale sur le lexique, qui peut se réduire à une répartition de masse m1 sur les éléments focaux X et $\neg X$ telle que :

$$m1(X) = p \; ; \; m1(\neg X) = 1-p$$

Une hypothèse sur le sous-lexique X avec le score q : (X q), ce qui correspond à une distribution d'incertitude m2 sur les éléments focaux X et $\Theta$ ( le lexique tout entier) telle que :

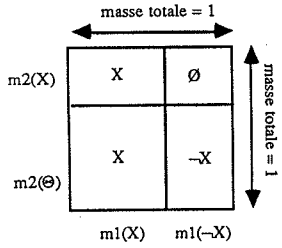
$$m2(X) = q \; ; \; m2(\Theta) = 1-q$$

L'application de la règle de combinaison de Dempster-Shafer appliquée aux distributions m1 et m2 schématisée par la figure 3 donne une nouvelle distribution m dont les éléments focaux sont X et $\neg X$ telle que :

$$m(X) = K*p$$
$$m(\neg X) = K*(1-p)*(1-q)$$

où $K = 1/(1-q+p*q)$ est un facteur de normation tel que :

$$m(X) + m(\neg X) = 1$$




La nouvelle distribution est obtenue en effectuant le bilan des surfaces allouées à chaque sous-ensemble. (ici X et $\neg X$) L'incertitude affectée à l'ensemble vide représente le degré d'incompatibilité entre les deux distributions.

Figure 3.

Au niveau de chaque élément du lexique la distribution finale est obtenue en répartissant m(X) et $m(\neg X)$ sur X et $\neg X$ respectivement, en respectant les proportions correspondant à la distribution initiale m1. L'incertitude obtenue correspond à un certain état de connaissance au niveau du lexique, qui va se traduire de manière effective lors d'une demande de validation de la part des analyseurs syntaxico-sémantiques par exemple. En effet, pour tout sous-lexique dans lequel les analyseurs cherchent à retrouver un mot particulier sur le signal, la distribution relative de certitude pour chaque élément va déterminer l'ordre dans lequel ils seront vérifiés sur le signal de manière à privilégier les plus probables, tout en éliminant plus ou moins les éléments particulièrement improbables, suivant le degré d'avancement de la reconnaissance.

Un tel mode de combinaison peut sembler en première analyse relativement irrévocable, puisqu'il modifie continuellement la base de certitude sur le lexique. Cependant, la règle de Dempster Shafer appliquée à ce type de répartitions bien particulier possède des propriétés intéressantes qui permettent au système de ne pas être contraint par des erreurs antérieures éventuelles.

En dehors des propriétés classiques de commutativité et d'associativité propre au mode de combinaison, qui rendent les effets d'un ensemble d'hypothèses indépendantes de l'ordre de leur application, la structure particulière des distributions de certitude employés ici fait que toute hypothèse (X q) est rétractable, au sens qu'il est possible d'émettre une nouvelle hypothèse ($\neg X$ q) pour la réduire complètement. Ainsi, un module particulier peut revenir sur une de ses analyses et de l'extérieur, intervenir sur le lexique pour redonner une cohérence interne à ses informations.

L'exemple fourni en annexe montre l'effet de l'application d'une hypothèse sémantique sur un ensemble de verbes, dans le cadre d'un lexique réduit pour les besoins de l'exemple. Les mots ainsi prédits doivent alors être vérifiés sur le signal pour être ensuite renvoyé au module émetteur. Nous allons voir maintenant comment une telle vérification peut s'opérer et comment des mots peuvent être prédits de manière ascendante, directement à partir du signal.

## 4. Prédiction et vérification au niveau phonétique.

La phase de prédiction doit être rapide et l'algorithme doit tenir compte des propriétés des résultats du décodage acoustico-phonétique. Le système APHODEX [Fohr 87] détecte très bien les noyaux vocaliques ( environ 95% de noyaux bien détectés, et 5% d'insertions), les fricatives et les plosives (détection de l'ordre de 90%, peu d'insertions pour les fricatives et environ 10% pour les plosives). La détection des liquides (l et R) et des nasales (m,n et $\eta$) est nettement moins bonne, les efforts de recherche s'étant concentrés sur les trois premières classes de phonèmes dans un premier temps. La détection des semi-voyelles (w et $\mu$) n'est pas encore opérationnelle. De plus, si la reconnaissance des trois grandes classes (voyelles, fricatives et plosives) est bonne, l'étiquetage en terme de phonèmes reste perfectible. Un programme de recherche a priori de mots, ne disposant que des seules informations fournies par le décodage, devra donc disposer d'une représentation des mots adaptée à ce niveau de décodage.

Deux solutions principales s'offrent au concepteur du système:

a) Une première approche consiste à regrouper les phonèmes d'un mot de façon contextuelle en blocs ou macro-phonèmes, qui s'avèrent être moins facilement omis par le décodage. Par exemple, le mot "*extension*" sera codé :

| | GV | GP&F | GV | GF | GV |
|---|---|---|---|---|---|
| avec | (ɛ) | (kst) | (â) | (s) | (iô) |
| ou | (ɛ) | (kst) | (â) | (sj) | (ô) |

Aucun des macro-phonèmes retenus ne pourra être élidé complètement, une au moins de ses composantes devra être fournie par le décodage, pour qu'un mot puisse être candidat. Une recherche exacte de cette chaîne de macro-phonèmes sera faite sur le treillis fourni par les niveaux bas, après l'avoir transformé en regroupant la liste des phonèmes fournis en macro-phonèmes. Une deuxième étape avec recherche exhaustive de tous les phonèmes donnera un score de reconnaissance plus fin au mot, ou même le rejettera si ce score devient trop faible.

b) Une deuxième méthode de codage et de recherche très simple fournit des résultats comparables, voire meilleurs. Le patron phonétique d'un mot se réduit à la liste des grandes classes qui le composent, parmi plosives, voyelles ou fricatives: par exemple, le patron de "*chapeau*" sera F V P V.

Un affinage consisterait à prendre en compte des traits supplémentaires tels que le voisement des consonnes ou la nasalisation des voyelles, s'ils sont obtenus avec des degrés de confiance suffisants. Cela permettrait d'augmenter le nombre de classes, et rendrait un patron plus sélectif, mais une seule erreur sur ces indices empêche le repérage d'un mot présent. Dans un premier temps, nous avons donc implanté la version la plus simple, avec trois classes seulement. Le patron d'un mot est alors équivalent à une écriture de ce mot sous une forme numérique, en base n, où n est le nombre de classes retenues. ("*chapeau*" peut ainsi être associé à 1202 en base 3 si l'on associe 0 à une plosive, 1 à une fricative et 2 aux voyelles )

Si ces quatre classes sont correctement trouvées par la phase de prétraitement du décodage acoustico-phonétique, repérer chapeau sur le signal revient à rechercher ce nombre sur quatre positions contiguës. Il suffit de promener un masque de largeur quatre sur la segmentation fournie par le prétraitement. Chaque valeur associée au masque courant se déduit de la précédente en ajoutant la contribution entrante à droite puis en ôtant la contribution sortante à gauche, pour un balayage gauche droite. On peut ainsi examiner tous les mots de longueur L et ( L + 1)

en un seul passage. Bien sûr un même patron correspond à plusieurs mots, mais reste assez sélectif. Par exemple, le patron F V P V ne correspond qu'à une dizaine de mots parmi 1300 entrées lexicales environ, tels que "*chaque*", "*chacun*", "*chapeau*", "*Jean-Paul*", "*Jacques*", "*jusque*", "*jusqu'à*", "*jeudi*"... Une deuxième phase de vérification de chacun de ces mots permet d'en éliminer plus ou moins, selon la dissemblance autorisée.

Ainsi, sur la phrase "*Ils ont de beaux chapeaux tyroliens*", seuls "*chapeaux*" et "*Jean-paul*" sont retenus, ce qui semble très satisfaisant. De récents progrès dans le calcul du fondamental et l'évaluation du critère voisé vs non-voisé d'une plage de signal devrait faciliter notre tâche.

Le nombre de mots significatifs retenus pour un énoncé court tel que celui ci-dessus varie de 0 à 10 maximum, dont environ 50 % de valides. Les mots erronés seront de toute façon retenus en cas de demande sur le signal, leur décodage apparaissant parfois de façon parfaite dans la chaîne phonémique fournie. Le taux de mots détectés en fonction de tous les mots effectivement présents, donc à reconnaître, varie de 30% à plus de 50 % selon les locuteurs. Supposons que la probabilité globale de détecter une classe phonétique parmi les trois que nous avons retenues soit p, et que la probabilité d'insertion d'une de ces classes soit q.

La probabilité de trouver un mot de longueur L est alors $p^{L}(1-q)^{(L-1)}$ ( il faut trouver L classes consécutives, et ne rien insérer dans les L - 1 intervalles), par exemple pour L=6 , p=0.95 et q=0.05 on devrait repérer 56% des mots présents et pour L=6 , p=0.9 et q=0.1 on doit détecter environ 31% de mots.

On remarque que cette méthode est très sensible au niveau de décodage, et que des progrès modestes en segmentation du signal peuvent la rendre très performante. Les résultats pratiques et théoriques sont assez proches, et la différence est liée au niveau de la deuxième phase du décodage, un très mauvais étiquetage pouvant conduire au rejet d'un mot correctement détecté, mais trop mal étiqueté.

Dans une deuxième phase, le niveau lexical ne travaille plus qu'à la demande des niveaux supérieurs. Si la phase de prédiction doit éviter de produire des mots erronés, le but à ce moment de la reconnaissance sera inversé : il faudra éviter d'omettre des mots candidats existants. Nous avons déjà précisé qu'un mot candidat avait le format suivant :

(*Sous-lexique Score Identifiant Plage-phonétique* )

Selon les informations de voisinage déjà prise en compte, la plage phonétique est plus au moins stricte : une hypothèse pourra apparaître au milieu de la plage proposée, sans cadrer ni à gauche ni à droite par exemple, être contrainte d'un côté (le mot attendu par les modules d'analyse est un voisin immédiat d'un autre mot déjà validé ), ou même couvrir toute la plage. Le score sera aussi dépendant du nombre d'informations déjà prises en compte, ainsi que de la confiance que le système place dans ces informations. Il ne tient encore compte d'aucune information phonétique. Il peut servir à fixer un seuil maximum de pénalité avant rejet pour la comparaison, ce seuil étant proportionnel au score a priori du mot.

La vérification est faite sur le signal avec un algorithme de programmation dynamique classique, qui fait trois hypothèses à chaque pas : mise en correspondance des deux phonèmes, élision du phonème attendu, insertion du phonème présent, chacun de ces choix ajoutant les pénalités induites au total déjà trouvé. Si ce total excède le seuil de rejet, on abandonne le chemin correspondant. Si aucun chemin n'aboutit, le mot sera rejeté, sinon la pénalité trouvée sera combinée au score à priori pour établir le score final du mot. Si la plage était floue, elle est remplacée par la plage de signal qui a fourni le meilleur chemin.

Les pénalités à associer à chacun des choix de progression de l'algorithme sont tirées de trois bases de connaissances : La première indique le coût d'une substitution d'un phonème par un autre. Elle n'est pas complètement symétrique, car certains phonèmes ne sont jamais étiquetés par le système, qui n'en a pas la description (w, μ) et certaines confusions sont orientées. Cette base de connaissances est une expertise a posteriori sur les confusions faites réellement par le système de décodage, et prend en compte à la fois les raisons phonétiques (proximité de deux phonèmes), phonologiques (altération d'un phonème par co-articulation) et imperfections du système (erreurs systématiques ou fréquentes). Une deuxième base de connaissances indique la gravité d'une élision. Elle prend surtout en compte les résultats du système, mais aussi des informations contextuelles : si l'élision d'un noyau vocalique est grave, car c'est une erreur rare, l'absence de détection d'un "i" entre s et ʃ l'est beaucoup moins. La troisième base de connaissance traite les insertions. Elle sera associée à la longueur du phonème "coupable" pour produire une pénalité la plus juste possible. Toutefois, un segment plosif très long, et placé en début de mot, pourra être une pause.

Ce module de vérification est déjà opérationnel pour de petits systèmes développés en parallèle avec le système de dialogue oral homme-machine. Le niveau de décodage est suffisant pour un locuteur entraîné pour permettre la reconnaissance de phrases, avec un vocabulaire limité (une centaine de mots). Le nombre de mots à retenir pour ne pas perdre des mots réellement présents est très variable en fonction des locuteurs, les extrêmes vont de 3 à 10 pour 1. La figure donnée en annexe 2 nous montre des copies d'écran d'un test ayant abouti à une reconnaissance parfaite de l'ordre < Copie core dans essai > pour une application "commandes vocales à un système informatique".

## 5. Conclusion et perspectives.

Les éléments de réflexions et de techniques présentés ici s'insèrent en réalité dans une analyse plus complète menée sur la mise en place d'un système de dialogue oral homme-machine sur un domaine relativement complexe. Les derniers éléments dont nous disposons au stade actuel de nos travaux montrent qu'il est difficile d'envisager la composante lexicale de manière totalement indépendante du reste du système. Il est préférable de spécifier son statut exact au regard des phases d'analyse structurale, mais surtout vis à vis de l'espace cognitif de la machine, pour espérer "comprendre" effectivement un dialogue.

Le point important est ici de considérer la machine comme un réel interlocuteur quand il s'agit de mettre en œuvre ce type de dialogue. De récentes expériences [Amalberti 88] ont permis d'ailleurs de montrer l'utilité de l'étude du dialogue homme-homme comme référence en la matière. Cette approche permet d'établir des modèles plus complets d'une organisation de haut niveau , où les techniques de base que nous avons montrées s'insèrent de façon satisfaisante.

## Références.


[Amalberti 88] R. Amalberti, N. Carbonel, P. Falzon, "Dialogue Homme-Homme, Dialogue Homme-Machine : Un même modèle ?", *Actes du 3eme Colloque International de l'ARC*, Toulouse 9-11 mars 1988.

[Barnett 83] J.A. Barnett, "Computational methods for a mathematical theory of evidence". *Proc. IJCAI 83*, pp.868-875.

[Carbonell 87] N.Carbonell and J.M.Pierrel, "Architecture of knowledge sources in a human-computer oral dialogue system". in: M.M.Taylor, F.Neel and D.G.Bouwhuis, eds., *structure of multimodal dialogues*, North-Holland, Amsterdam, 1987 .

[Deville 87] G.Deville, H.Paulussen et J.M. Pierrel, "Une grammaire de cas comme modèle de représentation sémantique d'énoncés de dialogues oraux homme-machine finalisés", *Proc. AFCET-INRIA 6th Cong. RFIA,* Antibes, nov. 1987.

[Erman 80] Lee D. Erman et Victor R. Lesser, "The Hearsay-II Speech Understanding System: A Tutorial". in : W.A. Lea, *Trends in Speech Recognition*, Prentice-Hall,1980

[Fohr 87] D. Fohr, N. Carbonel, J.P.Haton, "APHODEX, an acoustic-phonetic decoding expert system". Proc. IEEE Workshop on Expert Systems and Pattern Analysis. in *International Journal of Pattern Recognition and Artificial Intelligence*. C.H.Chen ed.,V.1 N.2 1987,pp. 207-222.

[ Heyer 85] K. den Heyer, A. Goring et G.L. Dannenbring, "Semantic priming and word repetition , the two effects are additive", *Journal of memory and langage*, v24 1985,pp.699-716.

[Latraverse 87] François Latraverse, *La pragmatique : histoire et critique*, P.Margaga,Bruxelles.

[Pierrel 87] J.M.Pierrel, *Dialogue Oral Homme-Machine*, Hermes, Paris, 1987.

## Annexe I : Effet d'une hypothèse.

**lexique** { la description complète de l'arborescence n'est pas fournie ici}

| NOEUD | NOMBRE | SCORE | ANCETRES | |
|---|---|---|---|---|
| document | 5 | 0,0123 | anime- | nom |
| je | 249 | 0,6118 | anime+ | pronom |
| mesurer | 2 | 0,0049 | mesure1 | verbe |
| compliquer | 1 | 0,0025 | process2 | verbe |
| important | 1 | 0,0025 | statut2 | adjectif |
| importance | 1 | 0,0025 | statut2 | nom |
| importer | 4 | 0,0098 | statut2 | verbe |
| habitation | 1 | 0,0025 | statut1 | nom |
| habiter | 5 | 0,0123 | statut1 | verbe |
| appartenir | 4 | 0,0098 | location1 | verbe |
| remise | 1 | 0,0025 | exchprod | nom |
| remettre | 2 | 0,0049 | exchprod | verbe |
| remercier | 2 | 0,0516 | actanime | verbe |
| changement | 1 | 0,0025 | process1 | nom |
| changer | 8 | 0,0197 | process1 | verbe |
| couter | 1 | 0,0025 | mesure2 | verbe |
| savoir | 29 | 0,0712 | extension | verbe |
| garder | 2 | 0,0049 | location2 | verbe |
| vote | 1 | 0,0025 | atrans | nom |
| obtention | 2 | 0,0049 | echobt | nom |
| obtenir | 24 | 0,0590 | echobt | verbe |
| donner | 21 | 0,0516 | exchprod | verbe |
| signer | 1 | 0,0025 | actobjet | verbe |
| apporter | 2 | 0,0049 | mvmt2 | verbe |
| aller | 18 | 0,0442 | mvmt1 | verbe |

La sélection d'un sous-lexique est seuillée à **niveau = 0,02**

l'hypothèse : **(verbe ())** {est une demande de validation d'un verbe sur le signal.}

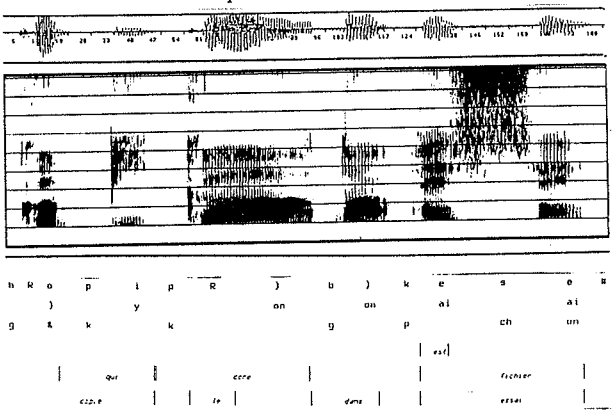
savoir -> score = 0,0712 {premier élément à être vérifié}
obtenir -> score = 0,0590 {deuxième...}
donner -> score = 0,0516
remercier -> score = 0,0516
aller -> score = 0,0442

l'hypothèse : **((inter mvt+ ctr+) 0,9)** {renforce l'ensemble des verbes exprimant un mouvement et un contrôle.}

l'hypothèse : **(verbe ())** {donne alors :}

aller -> score = 0,3066 {"aller" à été mis en évidence
savoir -> score = 0,0494 par l'hypothèse précédente et
obtenir -> score = 0,0409 sera donc vérifié en priorité.}
remercier -> score = 0,0358
donner -> score = 0,0358
apporter -> score = 0,0341

## Annexe II : Exemple de reconnaissance de la phrase "*copie core dans essai*".



- Signal temporel
- Spectrogramme
- Treillis phonétique fourni par APHON.
- Mots reconnus en phase ascendante (les barres verticales indiquent les limites temporelles de ceux-ci).