Zeitschrift: Études de Lettres : revue de la Faculté des lettres de l'Université de Lausanne

Herausgeber: Université de Lausanne, Faculté des lettres

Band: - (1997)

Heft: 3

Artikel: La fluidité en synthèse de la parole

Autor: Zellner, Brigitte

DOI: https://doi.org/10.5169/seals-870415

Nutzungsbedingungen

Die ETH-Bibliothek ist die Anbieterin der digitalisierten Zeitschriften auf E-Periodica. Sie besitzt keine Urheberrechte an den Zeitschriften und ist nicht verantwortlich für deren Inhalte. Die Rechte liegen in der Regel bei den Herausgebern beziehungsweise den externen Rechteinhabern. Das Veröffentlichen von Bildern in Print- und Online-Publikationen sowie auf Social Media-Kanälen oder Webseiten ist nur mit vorheriger Genehmigung der Rechteinhaber erlaubt. Mehr erfahren

Conditions d'utilisation

L'ETH Library est le fournisseur des revues numérisées. Elle ne détient aucun droit d'auteur sur les revues et n'est pas responsable de leur contenu. En règle générale, les droits sont détenus par les éditeurs ou les détenteurs de droits externes. La reproduction d'images dans des publications imprimées ou en ligne ainsi que sur des canaux de médias sociaux ou des sites web n'est autorisée qu'avec l'accord préalable des détenteurs des droits. En savoir plus

Terms of use

The ETH Library is the provider of the digitised journals. It does not own any copyrights to the journals and is not responsible for their content. The rights usually lie with the publishers or the external rights holders. Publishing images in print and online publications, as well as on social media channels or websites, is only permitted with the prior consent of the rights holders. Find out more

Download PDF: 06.07.2025

ETH-Bibliothek Zürich, E-Periodica, https://www.e-periodica.ch

LA FLUIDITÉ EN SYNTHÈSE DE LA PAROLE

En français, peu de travaux ont jusqu'à présent été consacrés à la dynamique temporelle de la parole. Il sera montré dans cet article en quoi cette dimension appartient à la fluence verbale et doit être prise en compte dans la modélisation de la prosodie. La démonstration empirique de l'importance de cette dimension peut être faite en synthèse de la parole. Différentes approches seront présentées dans leur capacité à bien prédire la structure temporelle des énoncés pour un synthétiseur. Les résultats seront ensuite discutés et permettront de reconsidérer la relation entre structure temporelle et structure mélodique pour le français.

1. Introduction

Une des difficultés rencontrées dans ce domaine de la communication humaine provient du fait que la parole est un processus *dynamique* et rapide. Une façon d'approcher cette dynamique est de considérer la dimension temporelle. En effet, une parole agréable à écouter n'est pas stable dans le temps : elle s'accélère, se ralentit, se prolonge, s'arrête, reprend. On peut donc se poser la question de savoir ce qui conditionne ce rythme ou ces interruptions.

La composante temporelle a souvent été le « parent pauvre » dans les modèles utilisés en synthèse de la parole. L'objet du présent article est de montrer comment la dynamique temporelle de la parole peut être appréhendée et modélisée. Plus précisément, il s'agit d'évaluer en quoi un modèle syntaxique ou phonologique, c'est-à-dire un modèle de la structure de la langue, peut ou non rendre compte du fonctionnement de la dynamique temporelle. La discussion sera basée sur une évaluation statistique et auditive de différents modèles.

2. *Parler : un ensemble d'actions dans un espace multi-temporel*

2.a. *La temporalité inhérente à l'activité de langage*

Ce que parler veut dire. « Panta Rhéi ! » Tout s'écoule, selon Héraclite. Ainsi en est-il de l'homme ayant le don de parole et produisant un mélodieux continuum sonore. De couler ou – en latin « fluere » – vient le « flux de parole » et la « fluence »: cette qualité verbale qui permet à l'homme de s'exprimer avec aisance, rapidité et sans heurt.

Comme pour toute activité humaine, l'activité de parole ne peut s'exercer en continu. Ni notre appareil de phonation, ni notre appareil auditif ne nous permettraient de produire ou de percevoir des sons en permanence. Du point de vue de la production de parole, la coordination neuro-musculaire des différents organes phonatoires, en particulier le larynx, le pharynx, et la langue, imposent des limites temporelles. Même en parlant très vite, il existe un seuil de vitesse au delà duquel un geste phonatoire ne peut plus être accompli. Selon les résultats des analyses effectuées dans notre laboratoire, 30 à 40 ms correspondent à cette durée minimale pour produire un son du français — ou segment. Également, du point de vue de la perception, ni nos oreilles, ni notre cerveau auditif ne sont capables de traiter un flot sonore linguistique en continu. Le message oral serait impossible à comprendre car nous n'aurions pas le temps de capter l'information, de l'analyser et de la comprendre. À cet égard, les travaux des psychoacousticiens ont montré que la fonction auditive est une fonction hautement complexe (Canévet, 1996). Selon ces spécialistes, en situation d'écoute attentive, 15 à 30 sons du langage peuvent être identifiés à la seconde (Zwicker and Feldkeller, 1981; Botte *et al.*, 1989). C'est en soi une performance prodigieuse car elle signifie que 7 à 15 syllabes par seconde, soit trois à huit mots[1] par seconde peuvent être traités par notre cerveau auditif ! Au delà de ce seuil, il devient difficile « d'écouter » de la parole. En effet, succinctement, les principales opérations effectuées en perception de la parole sont les suivantes :

1. *Capter* le signal sonore pertinent dans un milieu plus ou moins bruyant parmi d'autres conversations, par ex., dans une cafétéria,

---

1. Cependant, généralement, du fait de la redondance du langage et de la parole, le cerveau auditif semble ne traiter que partiellement l'information sonore qui lui parvient.

lorsqu'on appelle un ami qui vient d'arriver : "houhou, je suis là".
2. *Segmenter* ce signal en unités signifiantes, — [houhou/ je/suis/là] et non pas : *[hou /houjesuis/là] ou *[houhouje/ suislà].
3. *Identifier* les unités— [je], mot grammatical, référence à la 1ère personne, etc.
4. *Comprendre* le message — "houhou, je suis là".

Ces traitements complexes en perception et en production ne sont possibles que dans la mesure où le flux de parole est marqué sur le plan prosodique. Ce flux est jalonné de silences plus ou moins longs, et peut être modifié par des accélérations et des ralentissements. Il s'en suit que ces modifications de la rapidité ou de la lenteur du tempo, « l'agogique[2]», caractérisent la gestuelle verbale. La *parole fluide est donc en fait discontinue*. Une modélisation de la structure temporelle de la parole en vue de la lecture automatique par ordinateur devra tenir compte de cette caractéristique essentielle de la dynamique langagière.

Dimension cognitive de la temporalité de la parole. Pour autant, ces discontinuités n'interviennent pas n'importe où, lors de la production d'un énoncé. Par exemple, même en situation d'hésitation, les "euh", les pauses, et autres marques d'hésitation se produisent à certains moments privilégiés et à certains endroits dans l'énoncé. Chez un locuteur ne souffrant d'aucune pathologie de la parole, la production d'une hésitation au milieu d'une syllabe ou d'un mot est par exemple extrêmement rare. En revanche, il est tout à fait usuel de placer une marque d'hésitation entre deux groupes de sens (Zellner, 1992). Une abondante littérature motive ce phénomène du positionnement des discontinuités dans le flux de parole comme un reflet de l'activité mentale du locuteur (Goldman-Eisler, 1968, 1972 ; Boomer et Ditman, 1962 ; Cook, Smith et Lalljee, 1974 ; Chafe, 1980). Ainsi, les pauses et les hésitations interviennent comme des « balises » qui aident à structurer l'énoncé et à le rendre plus intelligible.

Dans le cas du français, les travaux de Grosjean et Deschamps (1975) appuient tout à fait cette thèse. Ils ont montré que plus un locuteur est soumis à une tâche complexe, plus il a besoin de « balises » pour structurer son discours, c'est-à-dire qu'il produit des pauses plus longues et/ou plus nombreuses (Grosjean et

---

2. Agogique: en musique, correspond aux modifications passagères du mouvement.

Deschamps, 1975). Le fait que ces discontinuités surgissent plutôt entre deux groupes de sens, montre que l'activité de parole s'organise en fonction d'un certain degré de cohérence entre les mots.

Cette cohérence a été mise en évidence, en particulier, par les psycholinguistes qui ont bien exploré certains aspects cognitifs de cette structuration orale du discours (Grosjean, 1980 ; Grosjean *et al.*, 1979, 1983.) Ils ont montré qu'en lecture, un énoncé est généralement dit en différents fragments, chaque fragment présentant une cohérence interne du point de vue du sens, de la prosodie — mélodie, durées, tempo, et intensité — et de la syntaxe. Eu égard à cette cohérence, il s'avère que d'une lecture à une autre, et d'un locuteur à un autre, un même énoncé est segmenté de façon assez similaire, c'est-à-dire que les découpages en groupes de mots, ou structures psycholinguistiques, obtenus d'un locuteur à un autre sont très proches. Dans la perspective d'une modélisation pour la synthèse de la parole, ces structures stables sont évidemment très intéressantes.

La lecture automatique à haute voix suppose qu'à partir du texte, un certain profil temporel se dégage, créant ainsi une parole de synthèse fluide agréable à écouter. Pour cela, il faut donc *prédire* correctement la cohérence temporelle (comment regrouper les mots, où placer des pauses et quelle longueur leur attribuer, où introduire des ralentissements et des accélérations de la vitesse de parole, etc. ?). Il sera montré que la connaissance des contraintes psycholinguistiques constitue un élément majeur dans la modélisation de la temporalité.

2.b. Le phénomène social de la temporalité de la parole

La langue et ses spécificités prosodiques. L'acte de parole n'est pas seulement contraint par une / des dimension(s) cognitive(s) liée(s) à la communication, mais également par tout un ensemble de règles sociales qui permettent qu'un locuteur puisse se faire comprendre d'un autre locuteur. Ainsi, chaque langue se sonorise avec des caractéristiques mélodiques et rythmiques qui lui sont particulières. L'étude de la phonologie d'une langue — système des sons — et de la prosodie vise à mettre en évidence à la fois les spécificités sonores de chaque langue et les invariants entre différentes langues.

La réalité de ces spécificités vernaculaires apparaît par exemple dans les travaux récents effectués dans le domaine de l'acquisition du langage. Autrement dit, ce que l'ordinateur doit

apprendre, le jeune enfant l'a acquis dès sa prime enfance. Ainsi, par rapport à des nouveaux-nés, les bébés de quatre jours sont déjà capables de différencier sur le plan auditif les sons qui appartiennent à leur langue maternelle de ceux d'une autre langue (Boisson-Bardies, 1996). Plus tard, vers neuf ou dix mois, le babillage des bébés devient également marqué, c'est-à-dire qu'un bébé anglais ne babille plus comme un bébé allemand ou français (Boisson-Bardies, 1996). En ce qui concerne les structures temporelles du français, c'est entre 8 et 24 mois que l'enfant évolue peu à peu vers la maîtrise de la structure rythmique de la langue (Konopczynski, 1986). L'enfant se montre peu à peu capable de produire l'allongement de la syllabe en fin de groupe prosodique, et d'adapter la durée syllabique aux contraintes imposées par la structure de la syllabe.

Ainsi, pour le jeune enfant apprenant, comme pour l'ordinateur qui doit parler, il est crucial de savoir capter les structures prosodiques qui appartiennent à toute une communauté linguistique et qui font que même en produisant des lallations — émission de sons dépourvus de sens —, on puisse reconnaître la langue et souvent l'intention qui y est exprimée.

Le problème de la représentation de ces structures a donné lieu à de nombreuses études, y compris pour le français. Le Laboratoire d'Analyse Informatique de la Parole (LAIP) a retenu et développé les modèles qui lui semblaient le plus opératoires en vue de la synthèse de la parole (Werner, ce volume ; Keller, ce volume ; Keller et Zellner, 1995, 1996 ; Zellner, 1994, 1996).

La norme socio-culturelle. Dans les formes sociales qui contraignent la temporalité de la parole, les aspects socio-culturels sont également très importants. Street et Brady (1982) montrent ainsi que les auditeurs acceptent des variations du débit de parole sur une certaine étendue, et qu'au-delà ou en deçà de cette étendue, l'évaluation sociale du débit devient négative. Dans notre société, cette zone d'acceptabilité de débit s'étend d'une vitesse modérée à une vitesse assez rapide. Dans cette zone de vitesse de parole, il se produit en effet selon Street et Brady, un effet d'attraction sociale. Quelqu'un qui parle avec fluidité et aisance (c'est-à-dire avec un débit assez rapide) est perçu comme une personne compétente et sûre d'elle-même. Par exemple dans une conversation, lorsqu'un locuteur parle rapidement et produit peu de pauses, il donne peu d'occasions à son interlocuteur de lui répondre. C'est ainsi que le locuteur paraît convaincant (Miller *et*

al., 1976). En revanche, une personne qui s'exprime très lentement peut être rejetée de son corps social, la lenteur de parole étant souvent associée à l'idiotie. Or, cette lenteur d'élocution peut être causée par une pathologie particulière, ou par un traitement thérapeutique qui modifie temporairement la dynamique vocale. Scherer *et al.* (1989) ont rappelé les nombreuses études qui décrivent comment un locuteur, dépressif ou névrosé, présente dans la voix des caractéristiques prosodiques très particulières.

La parole de synthèse doit bien sûr respecter ces normes sociales d'acceptabilité en matière de temporalité, d'autant que les exigences des auditeurs sont sévères vis-à-vis d'un synthétiseur de parole. Les modèles de représentation de la parole retenus pour la synthèse doivent intégrer ces contraintes. Une façon d'évaluer le niveau d'acceptabilité est de recourir aux tests auditifs afin de mesurer le degré de proximité entre une vitesse de parole artificielle et un débit de parole dit "naturel".

La composante pragmatique et affective de la temporalité. Enfin, chaque événement de communication est unique eu égard aux circonstances dans lesquelles il s'insère. Ceci signifie que le degré de cohérence entre les mots est aussi fonction des contenus du texte, des intentions du locuteur, de sa perception de la situation et des besoins du ou des auditeurs (Caelen-Haumont, 1994). Par exemple, dans certaines circonstances, un locuteur devra s'exprimer de manière très explicite, se mettant alors en état d'hyperarticulation. Dans d'autres situations, la rapidité de la communication constituera la contrainte majeure. Les stratégies de lecture s'inscrivent donc dans une fonction pragmatique de la communication, stratégies que nous ne savons pas encore modéliser mais qui sont actuellement l'objet d'études approfondies pour le français, en particulier au CLIPS, à Grenoble (Caelen-Haumont et Keller, ce volume).

Une situation de communication se caractérise aussi par l'état émotif des locuteurs. La force illocutoire d'un message — message sous-jacent — tout comme sa force perlocutoire — conséquence du message — peuvent altérer l'acte de communication. En effet, les changements d'états émotifs perturbent en particulier l'équilibre endocrinien, la température du corps, la tension musculaire. Ce sont autant de bouleversements physiologiques qui « s'entendent » dans la voix, car l'état de tension et de constriction du conduit vocal s'en trouve alors modifiés (Scherer, 1984, 1986). Le timbre devient plus ou moins rauque, plus ou moins profond,

plus ou moins coloré, plus ou moins tonique. Ces différentes configurations vocales ont des implications acoustiques perceptibles, en particulier sur le plan temporel. Mais elles sont encore très difficiles à modéliser car la diversité des modes d'agir est prodigieuse. Dans ce domaine, on commence à peine à savoir catégoriser des échantillons de parole en termes de joie, tristesse, colère, etc.

2.c. Conclusion : une définition de la fluence

En résumé, la temporalité de la parole se définit d'une part par les limites « biopsychologiques » de l'activité de parole, c'est-à-dire par des limites de perception, de production et d'activité cognitive. Et d'autre part, elle se définit par des limites sociales, c'est-à-dire par des limites liées à la langue, à la communauté socio-linguistique, à l'expression des affects et aux situations de communication. Autrement dit, « tout n'est pas possible ». Comme pour la plupart des gestuelles humaines, la parole est une gestuelle hautement complexe du fait qu'elle évolue dans un *espace multi-temporel*, chaque niveau de traitement ayant ses propres exigences temporelles.

La modélisation d'un tel ensemble d'événements multi-temporels nécessite la notion de « tamponnage » ou « boucle interne » comme suggéré par Levelt (1991). La fluence désigne cette capacité à *co-ordonner* tous ces événements au sein d'un système dont les composantes opèrent à différentes vitesses. En ce sens, Berthoz (1997) mentionne la proposition originale du biologiste Llinas selon laquelle « nous pensons à 40 Hz et nous bougeons à 10 Hz ». Compte tenu que le temps de base pour le traitement cognitif est d'environ 25 ms, et que le temps de base pour contrôler un mouvement est d'environ 100 ms, il est proposé que le cerveau utilise des boucles internes oscillant à 40 Hz pour permettre un traitement perceptif multisensoriel (Berthoz, 1997). Nous pensons que dans le domaine de la parole, une telle coordination multi-temporelle est reliée à la dimension temporelle et qu'elle se caractérise à la fois par une aisance dans la production du geste, — chaque événement succède à l'autre assez régulièrement —, par une tonicité adaptée qui permet le moindre effort, et par un débit assez rapide (Pfauwadel, 1986 ; Zellner, 1992, 1994).

Lorsqu'une gestuelle présente des caractères d'harmonie, d'équilibre et de cohérence, autrement dit, lorsque la gestuelle est *fluide*, il peut être supposé que cet ensemble de mouvements est

prédictible. Chaque séquence de mouvements succède à une autre en eurythmie[3], et les différents mouvements sont exécutés sans heurt, selon cette forme de cohérence temporelle qu'est la « synchronisation ». La parole peut par conséquent être décrite comme un système dynamique qui évolue selon un *pattern* assez régulier *(fluent)*. Lorsque cette régularité est soudainement perturbée, du fait soit de conditions inhérentes au locuteur — comme une émotion —, soit de conditions extérieures — comme un bruit très intense — le système peut alors devenir disfluent (chaotique) plus ou moins longtemps.

Des évidences supplémentaires en appui de cette hypothèse proviennent des erreurs de parole des locuteurs ne souffrant d'aucune pathologie. Beaucoup d'erreurs d'anticipation (comme les lapsus, les faux-départs, etc.) suggèrent que certains événements de parole sont planifiés simultanément ou légèrement en avance d'autres événements, et qu'ils doivent être maintenus en attente jusqu'à ce que le moment d'articulation de l'événement en question arrive (Fromkin, 1980; Dechert et Raupach, 1980). Ce type de traitement de l'information ne peut être modélisé qu'en proposant un modèle dont les composantes opèrent à différentes vitesses et communiquent via un système de boucles internes (Lapicque, 1943 ; Keller, 1984; Starkweather, 1987).

Dans cette perspective, ma proposition ne consiste évidemment pas à tenter de modéliser le comportement humain dans toute cette complexité avec ces systèmes de tamponnages. Mais on comprend qu'une production de parole sera fluente si on peut générer en particulier une organisation temporelle adéquate, où les périodes de parole alternent avec les périodes de silence et de ralentissement du débit. Le point crucial est que ces périodes ne se produisent pas de manière aléatoire et que ceci doit être modélisé dans les systèmes de synthèse de parole si ces systèmes doivent générer une parole dite « naturelle ».

3. La modélisation du cadre temporel

En sciences humaines, il est traditionnel de chercher à construire des modèles pour tenter d'expliquer des phénomènes

---

3. L'eurythmie est cette qualité qui désigne un mouvement harmonieux et équilibré. En parole, ce principe repose sur l'équilibre syllabique.

humains observés, qu'ils soient génotypiques — innés — ou phénotypiques — acquis —, individuels ou collectifs. Il est aussi généralement très difficile d'*évaluer* ces modèles. Cela impliquerait en effet de pouvoir soumettre l'homme à des conditions de laboratoire. Or, l'humain est si complexe qu'il est souvent hasardeux de prétendre que *tout* puisse être contrôlé lors d'une expérimentation. Ainsi en est-il dans le domaine des sciences de la parole car le fait de parler est éminemment complexe. Dans cette perspective, la possibilité de créer une parole artificielle offre des voies d'exploration intéressantes. En effet, la mise à disposition d'un nouvel outil permet d'éclairer le domaine scientifique sous des angles restés obscurs jusqu'alors. La parole artificielle offre désormais la possibilité de tester des hypothèses qu'il était impossible d'éprouver auparavant. Une démonstration limitée en est faite ci-après sur la question très spécifique de l'organisation temporelle de la parole.

Pour générer une parole de synthèse fluide, il importe de prédire correctement la structuration d'un énoncé en groupes de mots cohérents, et en particulier de déterminer où placer des pauses, et où introduire des ralentissements et des accélérations de la vitesse de parole — c'est-à-dire, aux frontières temporelles. De plus, ceci implique *aussi* de respecter les formes prosodiques spécifiques à la langue traitée. En effet, la cohérence temporelle est un problème qui se pose tout au long de la chaîne parlée, c'est à dire *aussi* entre deux frontières temporelles. Pour ce faire, deux types d'approches sont envisageables, les modèles basés sur les hypothèses linguistiques et ceux basés sur les hypothèses psycholinguistiques.

Poser la question de l'organisation temporelle de la parole en ces termes est une question nouvelle et c'est une question qui promet d'être utile par exemple pour des chercheurs en prosodie, pour des thérapeutes dans leurs efforts de compréhension des dysfonctionnements et leurs recherches de thérapies efficaces, ou encore pour l'amélioration des méthodes d'enseignement du français langue étrangère.

3.b. Les approches linguistiques

En synthèse de la parole, les modèles linguistiques, c'est-à-dire ceux fondés sur un modèle de la langue, sont à la base de nombreux algorithmes de regroupement des mots, regroupements qui permettront ensuite le calcul des paramètres prosodiques. Dans ce type d'approches, les structures syntaxiques et/ou pho-

nologiques permettent de rendre compte de cette logique de regroupement. Le trait commun à tous les algorithmes construits dans cette perspective est qu'ils conçoivent le plus souvent l'accent[4] comme la clé de voûte de l'organisation temporelle de la parole.

Les modèles prosodiques dérivés des structures syntaxiques. Historiquement, l'héritage de l'école américaine chomskyenne a beaucoup influencé les premiers travaux en synthèse de la parole comme par exemple, ceux du « MITalk System » (cf. Chomsky et Halle, 1968 ; Chomsky, 1970 ; Allen *et al.*, 1987). Selon ces modèles, on admet généralement une dépendance entre syntaxe et prosodie. Les frontières prosodiques épousent les frontières syntaxiques (Di Cristo, 1975). L'accent de fin de groupe de mots signale une frontière syntaxique et se caractérise par des variations de mélodie et / ou d'intensité, ainsi que par un allongement de la durée de la dernière syllabe. La modélisation de ces principes en synthèse de la parole aboutit à ce que la durée des unités syllabiques ou segmentales soit calculée en fonction de la proximité de ces unités avec les frontières prosodiques, qui elles-mêmes correspondent à des cibles accentuelles plus ou moins saillantes (Bailly, 1983).

Les modèles prosodiques dérivés des structures phonologiques. Les approches phonologiques[5] proposent un autre type de hiérarchie linguistique pour mettre en évidence les relations de proéminence entre les unités (par exemple, Martin 1975 ; Liberman et Prince, 1977 ; Selkirk, 1984 ; Dell, 1984 ; Rossi, 1985). Divers modèles sont proposés pour exprimer ces proéminences et un certain consensus s'est établi autour du concept de constituant prosodique pour décrire ces structures (Wightman, 1992). Les frontières de ces constituants sont supposées être plus ou moins marquées grâce à l'assignation des accents intonatifs (voir par exemple l'algorithme de Bachenko et Fitzpatrick, 1990). L'implantation de ces approches dans un synthétiseur implique que la durée des unités syllabiques ou segmentales s'allonge plus ou

---

4. «Accent» se comprend ici en tant que phénomène prosodique se traduisant au plan acoustique par des variations spectro-fréquentielles perceptivement saillantes.

5. Sous ce même «chapeau», on me pardonnera de mettre ensemble entres autres la phonologie générative (par ex. Chomsky et Halle), la phonologie autosegmentale (par ex. Goldsmith) et la phonologie prosodique (par ex Selkirk).

moins, en fonction de la proximité d'un accent et de sa hauteur dans la hiérarchie phonologique (Barbosa, 1993 ; Delais, 1995).

3.c. Evaluation de ces deux approches

Matériel. Quinze phrases ont été extraites d'un corpus lu par un locuteur à un débit assez rapide (cf. Keller et Zellner, 1995 ; Zellner, 1996). La place du verbe dans la phrase a été prise en compte car la frontière entre le groupe nominal sujet et le groupe verbal constitue une frontière majeure pour la majorité des modèles syntaxiques[6] utilisés en synthèse de la parole et en traitement automatique de la langue, bien que quelques linguistes, comme Tesnière (1959), aient défendu une mise au même niveau du sujet, du verbe et de ses compléments.

Le but de cette expérimentation est de tester si les structures syntaxiques et phonologiques d'un texte peuvent prédire les structures temporelles de la parole. Si l'organisation temporelle d'un énoncé oral est directement reliée aux structures syntaxiques et/ou phonologiques, cette frontière majeure devrait être fortement marquée sur le plan temporel, quel que soit son emplacement dans la phrase.

Pour de nombreux auteurs, il existe une correspondance assez forte entre ces frontières syntaxiques majeures et les structures temporelles. Cependant, il apparaît aussi chez ces mêmes auteurs, aussi bien dans les études portant sur le français (par exemple, Delais, 1995) que dans d'autre langues (par exemple, Terken, 1992), que la plupart des phrases analysées sont structurées de manière telle que la coupure syntaxique majeure est située vers le milieu de la phrase. Il se trouve que des psycholinguistes comme Grosjean ont de leur côté observé une tendance à produire une coupure majeure en milieu de phrase, et ils l'expliquent davantage à l'aide de critères psycho-rythmiques. Il n'apparaît donc pas clairement si le locuteur fait une pause du fait de la frontière majeure entre le syntagme nominal (SN) et le syntagme verbal (SV), ou du fait de la position centrale de cette frontière.

Par conséquent, parmi ces quinze phrases, cinq contenaient le verbe en début d'énoncé, cinq autres en milieu d'énoncé et cinq autres vers la fin de l'énoncé. Dans chacun des trois groupes, la

---

6. Jean Véronis, Université d'Aix-en-Provence, confirme que ce principe de la division majeure est opératoire dans de nombreux systèmes en traitement automatique de la langue.

longueur des phrases variait en moyenne entre 8 et 14 mots (voir en annexe).

Mise en évidence de l'organisation temporelle des phrases. M'inspirant des principes méthodologiques développés par Grosjean (Monnin et Grojean, 1993), le profil temporel des phrases est révélé grâce à l'analyse des mesures des durées syllabiques produites par le locuteur. Les mesures des durées des syllabes sont tout d'abord standardisées au moyen d'une transformation logarithmique[7]. Puis, l'aire sous la courbe gaussienne est subdivisée en 5 zones : la zone centrale correspond aux durées syllabiques proches de la moyenne statistique. De part et d'autre de ces durées proches de la moyenne viennent ensuite les zones comprenant les syllabes avec un petit écart par rapport à la moyenne (entre un demi et un écart-type), et enfin aux deux extrêmes de la courbe, viennent les zones qui comprennent les syllabes s'écartant de plus d'un écart-type (cf. figure 1).

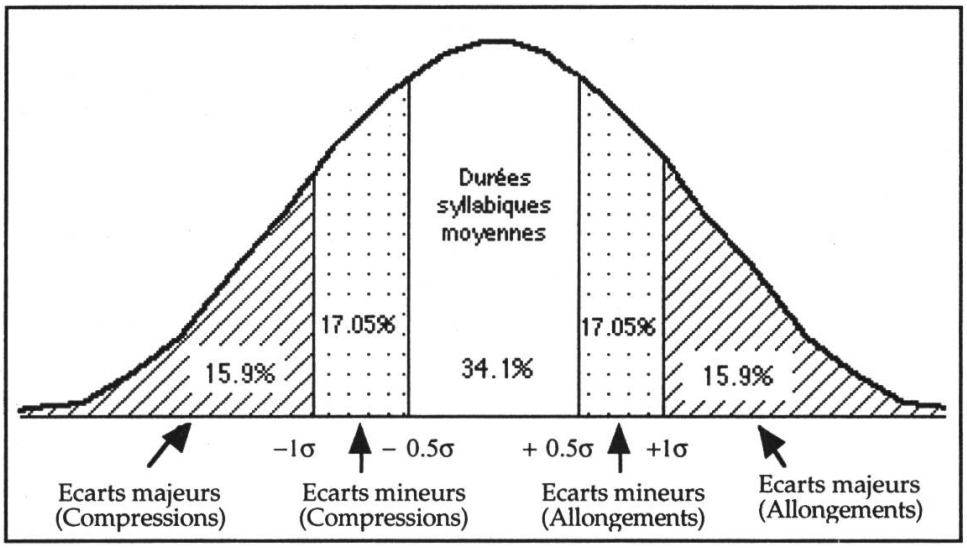

Figure 1. Subdivision des durées syllabiques après normalisation.

Cette répartition des durées est interprétée comme suit. La durée syllabique peut osciller autour d'une durée moyenne, ou au

7. Ce type de normalisation statistique a été retenu au LAIP après de nombreuses investigations, sous la direction du Professeur de statistiques, M. Gualtierotti, HEC, Université de Lausanne. La distribution des durées s'approche à un degré satisfaisant d'une distribution normale après une transformation log. D'autres transformations plus sophistiquées ne fournissaient pas d'approximations significativement meilleures. L'intérêt d'une normalisation des données est qu'elle autorise ensuite l'emploi de tests statistiques standard.

contraire subir des compressions ou des allongements plus ou moins importants. Un indice élevé (classe 3 ou 4) correspond à la réalisation d'un allongement de la durée syllabique tandis qu'un indice bas (classe 0 ou 1) correspond à la compression de la durée d'une syllabe.

Sachant que de nombreux paramètres interviennent dans l'explication de la variance des durées syllabiques (cf. figure 2), cette catégorisation permet d'expliquer dans un modèle de prédiction statistique environ 37% de la variance totale des durées syllabiques (Zellner, 1996).

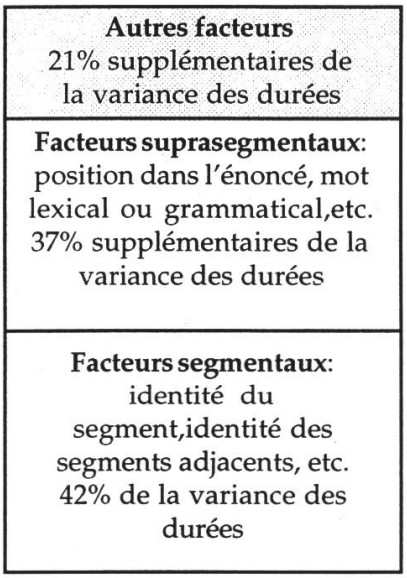

Figure 2. Selon notre modèle statistique des durées (Keller et Zellner, 1996), les facteurs suprasegmentaux permettent d'expliquer 37% de la variance totale des durées syllabiques (Zellner, 1996).

Voici un exemple d'indexation des syllabes, en fonction de l'organisation temporelle produite par un locuteur pour la phrase : "Ce fin cordon-bleu nous cuisinait parfois un canard eurasien émincé à l'armagnac":

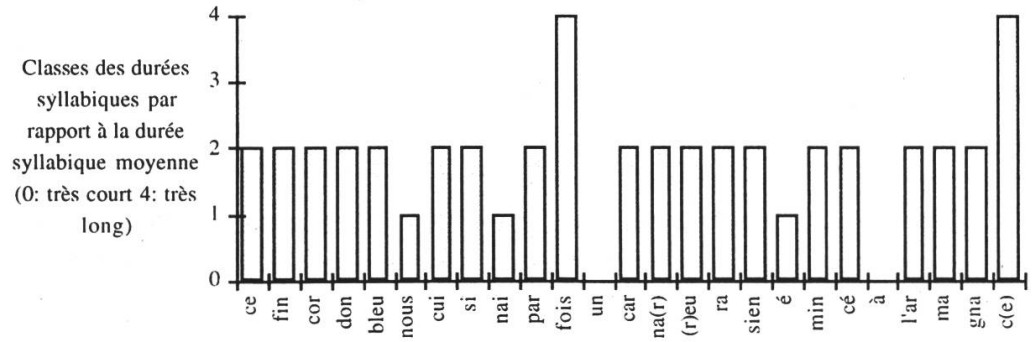

Figure 3. Représentation du profil temporel produit par un locuteur, selon la méthode d'indexation des syllabes.

Les comparaisons qui suivent, entre le profil temporel réalisé par le locuteur et ceux prédits par les différents modèles exploitent la logique suivante. Si la grammaire testée — qu'elle soit syntaxique, phonologique ou autre — est capable de prédire le degré de cohérence entre les mots appartenant à un syntagme et entre les syntagmes d'un énoncé, et si ce degré de cohérence interlexicale est en mesure de prédire les modulations temporelles des syllabes interlexicales d'un énoncé, alors, une corrélation étroite devrait émerger entre le type de durée syllabique produit en un point de l'énoncé et l'importance de la frontière prédite en ce point par cette grammaire. La comparaison statistique porte donc sur les classes de durées prédites à chaque fin de mot dans l'énoncé.

Comparaison avec les profils prédits selon les approches linguistiques : profils temporels prédits à partir de l'analyse syntaxique. L'analyse syntaxique qui a été ici testée pour prédire le profil temporel est fondée sur la grammaire normative *Le Bon Usage*, de Grevisse[8], fournie par l'application *Le Correcteur 101* (Machina Sapiens, Montréal). Ce logiciel propose une analyse syntagmatique hiérarchisée. Il suffit ensuite de compter le nombre de nœuds[9]. Puis, cette hiérarchisation syntaxique est traduite sur une échelle de 0 à 4 pour pouvoir être comparée avec les observations.

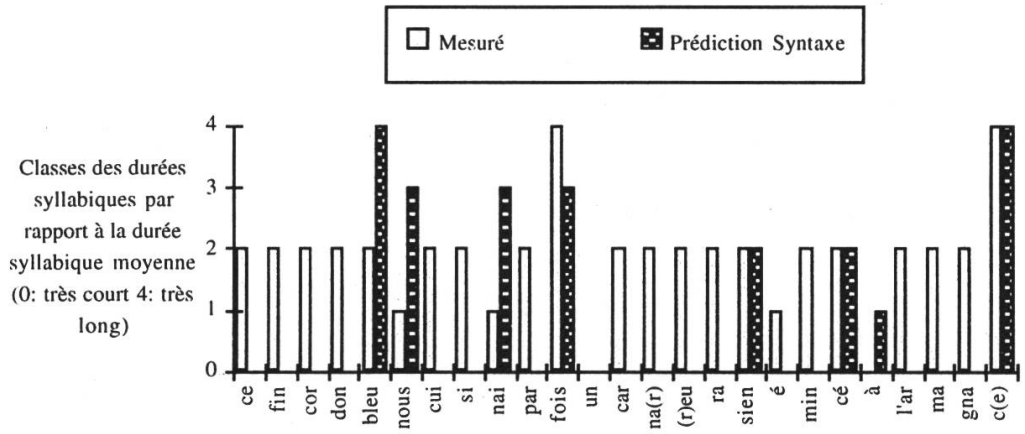

Figure 4. Comparaison entre le profil temporel mesuré et le profil temporel prédit selon l'analyseur syntaxique.

---

8. Parmi les raisons qui ont motivé ce choix, retenons que cette grammaire normative constitue souvent une référence pour les enseignants du français. Par conséquent, même si cette grammaire fournit des analyses syntaxiques critiquables pour nombre de linguistes, elle est «validée» par une certaine communauté. Une autre raison de ce choix est que l'utilisation du logiciel permettait d'éviter tout biais interprétatif de ma part.

9. Compte tenu que l'analyseur ne fait pas de prédiction à l'intérieur du mot,

Ainsi, selon cette analyse, la prédiction d'un allongement sur la syllabe "bleu" indique une coupure syntaxique majeure — qui correspond à la distinction entre groupe nominal et groupe verbal. Si cette analyse prédit correctement le profil temporel, elle prédit donc que la rupture temporelle sera élevée en ce point de l'énoncé (4). Cependant, le graphique montre que ce n'est pas ce que le locuteur a réalisé.

Globalement, le degré de similarité des syllabes entre la structure syntaxique et l'organisation temporelle effective est de 63.02%. Le taux des frontières prédites par l'analyse syntaxique mais non produites par le locuteur est de 40.7%.

Comparaison avec les profils prédits selon les approches linguistiques : profils temporels prédits à partir de l'analyse phonosyntaxique. L'algorithme sommairement décrit ici est celui décrit par Delais (1995). Les groupes prosodiques sont dégagés à partir de l'interaction entre deux types de structures : la structure syntaxique des phrases — selon la théorie X-barre — et la structure phono-rythmique basée sur le repérage de l'accent (Delais, 1995). La structure prosodique qui est finalement retenue est celle qui présente une compatibilité optimale avec ces deux structures linguistiques, grâce à une évaluation et une hiérarchisation combinée des contraintes syntaxiques et phonologiques (Delais, 1995). Toutes les configurations d'accents possibles sont ainsi examinées. La ou les configurations finalement retenues sont celles qui contiennent à la fois le moins de contraintes violées dans les deux domaines et le moins haut possible dans la hiérarchie des contraintes. Puis, le niveau de frontière retenu est indexé selon la même échelle que celle appliquée plus haut, à fins de comparaisons.

Globalement, le degré de similarité des syllabes entre le profil prédit par l'algorithme phonosyntaxique et l'organisation temporelle effective est de 81.3%. L'apparent bon score du degré de similarité entre les deux profils s'explique en fait par une sous-estimation du nombre de frontières. En effet, du fait que le nombre de syllabes non proéminentes est supérieur au nombre de syllabes positionnées aux frontières, l'algorithme a fait moins d'erreurs que le précédent par « abstention » de prédiction. Par ailleurs, parmi les syllabes proéminentes prédites, le taux des

---

toutes les syllabes intralexicales ont été mises à zéro dans le graphique pour les besoins d'illustration.

frontières prédites par l'analyse phonosyntaxique mais non produites par le locuteur est de 38.9%, ce qui révèle finalement un score assez faible de prédictions effectivement correctes. Cela est d'ailleurs confirmé sur le plan sonore. Le synthétiseur LAIPTTS muni de cette logique produit une parole très bizarre du point de vue de l'organisation temporelle, suscitant un rejet immédiat des auditeurs.

3.d. Discussion

Les résultats obtenus concordent avec ceux obtenus dans d'autres langues (Ostendorf, 1994). Globalement, les structures syntaxiques et phonosyntaxiques n'offrent pas une bonne prédiction des frontières temporelles mineures, or ce sont les plus nombreuses. Enfin, le déplacement du verbe dans la phrase induit de 50% à 60% de prédictions erronées pour ces deux parseurs. La coupure entre SN et SV n'est donc pas automatiquement marquée par une coupure temporelle majeure si cette coupure est éloignée du centre de la phrase. Cela renforce la proposition de Caelen-Haumont (1991) selon laquelle en lecture, les groupes de mots ne sont pas uniquement formés selon une stratégie syntaxique. Ces résultats remettent en cause le traditionnel triplet {frontière syntaxique ou phonologique —> accent de fin de groupe —>frontière de groupe temporel}.

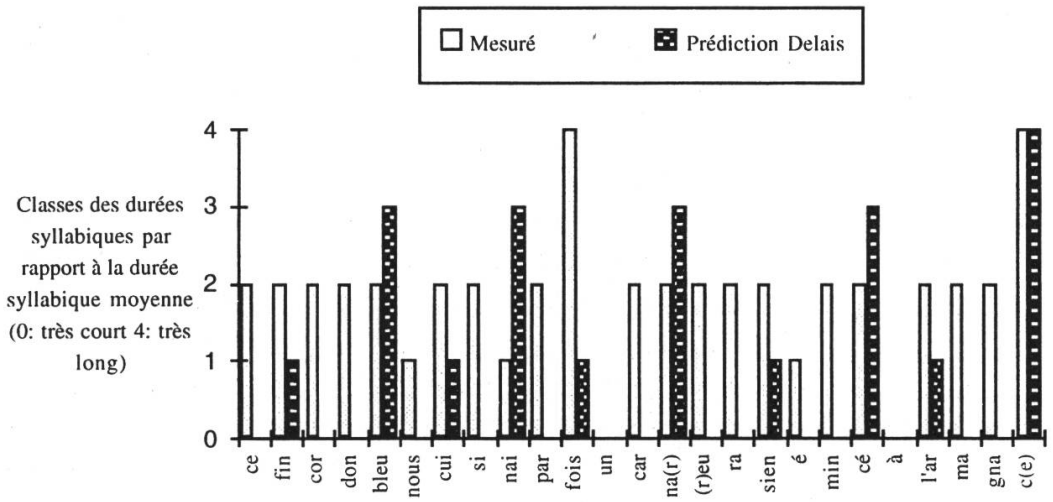

Figure 5. Comparaison entre le profil temporel mesuré et le profil temporel prédit selon l'analyseur phono-syntaxique.

Ces observations montrent que le lien entre frontière syntaxique et/ou phonologique et frontière temporelle n'est pas immédiat[10]. En effet, ces approches ne peuvent satisfaire l'objectif de prédiction du profil temporel de la parole. Cela s'explique par le fait que la temporalité de la parole est à l'interface de nombreux processus autres que les processus linguistiques (voir la première partie de cet article). En outre, il est à remarquer qu'en héritage de l'école anglo-américaine, la prédiction des modèles prosodiques, et en particulier temporels, pour le français, se fait souvent à partir de la localisation de l'accent (Ex : Bailly, 1983 ; Barbosa, 1993 ; Beaugendre, 1994 ; Delais, 1994, 1995 ; Di Cristo et Hirst, 1994 ; Jun et Fougeron, 1995 ; Martin, 1987 ; Mertens, 1987, 1993 ; Pasdeloup, 1992.). Cependant, le français n'est pas une langue accentuelle comme l'anglais ou l'allemand. Et c'est sans doute pourquoi dans la littérature il n'existe de consensus ni sur le placement des accents en français, ni sur les différents types d'accents. Guaïtella, (1996), Young-Lim (1996) et Zellner (1996) parviennent, par des analyses différentes, à des conclusions semblables quant à l'incertitude du repérage de l'accent en français. Par exemple, Behne distingue deux accents de syntagme : l'accent final et l'accent de focus basé sur les corrélats acoustiques qu'elle a pu mettre en évidence. Pasdeloup (1992) repère également deux accents : l'accent primaire et l'accent secondaire mais qui se traduisent au plan acoustique par 11 types de proéminences différenciées par le contour de F0 — variations de hauteur — et le type de durée. Astesano *et al.* distinguent quant à eux 6 types d'accents en fonction de leur emplacement dans l'unité intonative et de la présence ou non d'une emphase. S'il s'avère difficile d'aboutir à un consensus sur le concept d'accent en *production* de parole (voir Mertens, 1987), remarquons que la *perception* de l'accent pose également problème. Ko Young-Lim (1996) a montré dans des tests de perception avec huit auditeurs francophones, qu'il existait des divergences significatives sur la localisation de l'accent, et elle interprète ces résultats selon le fait que tous les auditeurs n'attribuent pas les mêmes valeurs aux mêmes indices. Dans ces

---

10. Les récentes observations neuro-linguistiques confirment d'ailleurs que l'encodage prosodique est distinct du traitement lexico-syntaxique. C'est par l'intervention de nombreuses structures médiatrices que la cohérence est assurée entre les différentes aires spécialisées pour le langage et la parole (Damasio *et al.*, 1997).

circonstances, il semble légitime de questionner la validité de l'indice accentuel pour la prédiction des modèles temporels.

Enfin, au niveau algorithmique, cette logique implique que l'organisation temporelle de la parole soit subordonnée aux phénomènes accentuels puisque tout le calcul des durées syllabiques est fonction de ces derniers. Or s'il est clairement montré dans la littérature qui vient d'être citée que ces deux paramètres sont en inter-relations, il n'est paru à ce jour aucune étude qui ferait la preuve d'une telle subordination.

4. L'approche psycholinguistique

Les modèles théoriques de la langue ne permettent pas actuellement de rendre compte de façon satisfaisante de l'organisation temporelle de la parole. D'autres chercheurs avaient déjà mis en lumière ce problème (Gee et Grosjean, 1983 ; Pasdeloup, 1992 ; Delais, 1995). Par conséquent, une autre approche doit être envisagée et elle vient de la psycholinguistique, c'est-à-dire d'un modèle du *fonctionnement langagier en temps réel (real-time processing model)*. Dans cette approche, l'agencement des mots qui se fait au cours de l'élocution est fonction de paramètres linguistiques et paralinguistiques.

4.a. L'apport de la psycholinguistique

Les structures psycholinguistiques, ou « structures de performance[11] », désignent ces groupements de mots produits spontanément par un lecteur. Elles sont qualifiées de « psycholinguistique » au sens où elles ont été mises en évidence à partir d'observations empiriques du fonctionnement langagier en temps réel, selon différentes contraintes cognitives. Ces structures ont été dégagées dans différentes langues, en français, en anglais (Gee et Grosjean, 1983), ainsi que dans la langue américaine des signes (ASL) utilisée par les sourds (Grosjean, 1980). En ce sens, ces groupements spontanés ou structures psycholinguistiques sont caractéristiques de l'expression langagière dans sa dimension dynamique, indépendamment de la modalité d'expression — orale ou visuelle (Grosjean, 1980).

---

11. Ici, le terme «structure de performance» renvoie simplement au concept d'une structure générée en temps réel.

La réalité psycholinguistique de ces structures est apparue lors de diverses expériences (Grosjean, 1980 ; Gee et Grosjean, 1983 ; Grosjean et Dommergues, 1983 ; Grosjean et Lane, 1981). Ainsi, si des phrases sont lues auprès de différents auditeurs et qu'il leur est demandé de redire cette phrase, les groupes de mots obtenus sont semblables. Si ces mêmes phrases sont fournies sous forme écrite à des non-linguistes et qu'il leur est demandé d'indiquer intuitivement au crayon les groupes de mots, les segmentations obtenues sont semblables. Si ces phrases sont lues à haute voix par plusieurs locuteurs, l'analyse des pauses interlexicales révélera des structures temporelles similaires. Et si on compare les structures obtenues entre ces trois types d'expériences, elles sont également semblables.

Ces structures psycholinguistiques se caractérisent par quelques traits majeurs. Les groupements de mots de base sont tous à peu près de même longueur et ils s'intègrent dans des groupes supérieurs. Par exemple : /la petite fille de mon amie/ peut être structuré en deux sous-groupes : /(la petite fille) (de mon amie)/. Cette structure a tendance à être symétrique autour d'un point central dans la phrase, ce qui lui confère un certain équilibre rythmique (Grosjean et Dommergues, 1983 ; Monnin et Grosjean, 1993). Le caractère « psycholinguistique » de ces structures apparaît dans diverses études empiriques. Par exemple, le problème de la longueur de groupes de mots est appuyé par les études d'empan de coordination œil-voix (Kondo *et al.*, 1996) qui montrent que le lecteur ne regarde que très peu de caractères en amont, avant de lire à voix haute, et ceci suggère que la génération de la prosodie ne peut pas être basée sur l'analyse de toute la phrase. De plus, la réalité psychologique de ces petits groupements de mots est beaucoup plus appuyée que celle de la phrase, au regard de nombreux traitements cognitifs qui sont faits à ce niveau (pour une revue de la question, voir Caelen-Haumont, 1997, à paraître).

Également, la distinction entre mots *grammaticaux* (mots outils : prépositions, articles, etc.) et mots *lexicaux* (autres mots : verbes, noms, adjectifs, etc.) reposent sur des observations empiriques. Il a été abondamment rapporté dans la littérature qu'une même syllabe produite en tant que mot fonction ou mot de contenu recevait une durée différente, le mot de contenu étant plus long du fait de son poids sémantique. Cette distinction a été dans un premier temps attribuée à un critère de catégorisation discrète, suite à des observations de patients aphasiques. Mais de

récentes analyses neurobiologiques suggèrent plutôt que cette distinction serait due à des conditions d'acquisition différentes et résulterait en des représentations neuronales différentes (Mueller, 1997, à paraître).

Compte tenu de leurs caractéristiques, ces structures peuvent se prédire aisément comme proposé par Monnin et Grosjean (1993). De plus, nous avons montré qu'il était possible de prédire ces structures tout en simplifiant considérablement leur algorithme (Keller *et al.*, 1993), c'est-à-dire en appliquant une analyse de proximité du matériel textuel plutôt qu'une analyse exhaustive de la structure phrastique, et ceci sans altérer de manière significative la qualité de la prédiction.

En lecture neutre, ces groupements de mots sont hautement corrélés au profil prosodique puisqu'ils peuvent être également révélés par l'analyse des pauses interlexicales et des durées des syllabes finales (Monnin et Grosjean, 1993). Dans ces circonstances, la modélisation prosodique — et en particulier temporelle — de l'énoncé consiste simplement à prédire ces structures psycholinguistiques.

En lecture neutre, la modélisation temporelle peut être confondue avec la prédiction des structures psycholinguistiques.

En l'absence de charge émotive, sémantique ou pragmatique particulière, le respect des contraintes psycholinguistiques pourrait agir comme une sorte de modèle par défaut pour la production d'une organisation temporelle correcte.

4.b. Un nouveau parseur basé sur ces structures

Dans ce nouveau parseur, les groupes de mots sont conçus comme résultant de certaines contraintes psycholinguistiques. Ces contraintes peuvent, en lecture neutre, se traduire directement en modélisation temporelle. Ainsi, la composante temporelle apparaît ici non plus comme une composante dérivée de la composante mélodique, mais comme une composante indépendante qui contribue pleinement à la modélisation prosodique (cf. figure 6).

Selon les activités de parole, il est supposé que les trois composantes (traitements temporel, mélodique et énergie) sont plus ou moins prégnantes, c'est-à-dire qu'elles peuvent intervenir avec des « poids différents » dans la génération de la prosodie.

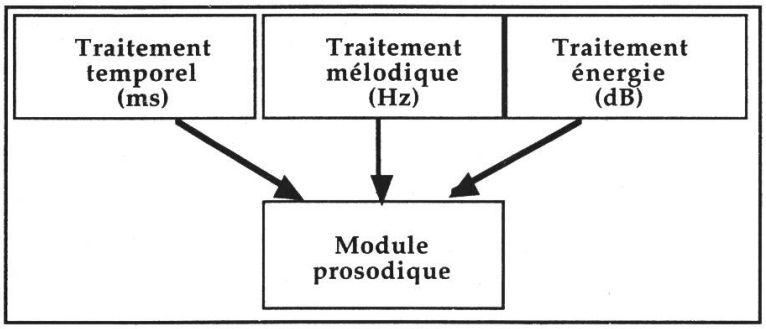

Figure 6. Le modèle prosodique résulte de la combinaison des composantes temporelle, mélodique et énergétique.

Ainsi, en activité de lecture neutre en chambre calme —situation de parole assez exceptionnelle en soi —, il est supposé que la composante temporelle peut suffire à donner le profil prosodique général — car proche des contraintes psycholinguistiques (petits groupes de mots fondés sur la distinction entre mots fonction et mots grammaticaux, équilibrage de la longueur des groupes autour d'une coupure centrale), les traitements mélodique et d'énergie « épousant » le profil temporel.

Par contre, en parole spontanée, on peut s'attendre à ce que le traitement mélodique et le traitement d'intensité prenne plus d'importance — ce qui reste à démontrer — en ce sens qu'il y aurait davantage d'informations *(meaning)* à transmettre.

Autrement dit, selon le type de parole, je suppose que les relations entre les composantes prosodiques *varient*. En ce sens, je rejoins la notion d'*inter-relations* fort bien développée par Local (dans ce volume) pour rendre compte de la complexité dynamique de la prosodie de la parole.

Mise en évidence des profils temporels. Selon cette approche, l'organisation temporelle d'un énoncé se subdivise en groupes majeurs (GM) qui eux-mêmes se subdivisent en groupes mineurs (gm). Ces groupes temporels sont signalés par les durées aux frontières et entre les frontières. Les frontières temporelles mineures et majeures sont marquées par les catégories de durées. Ainsi, les frontières mineures sont signalées par des durées situées dans la zone des écarts mineurs, et les frontières majeures par des durées situées dans la zone des écarts majeurs[12].

12. Dans nos données, cette catégorisation est robuste. En effet, seul un petit nombre de syllabes y échappe: les *syllabes unisegmentales* ne sont distribuées

Modélisation de la temporalité de la parole à partir du texte. La première étape pour la modélisation de la temporalité consiste à poser les frontières de la manière suivante :

• Distinguer les mots lexicaux (noms, verbes, adjectifs, adverbes) des mots grammaticaux dans une situation de lecture neutre.

• Poser une frontière de groupe temporel mineur après tout mot lexical suivi d'un mot grammatical. Quelques règles supplémentaires sont appliquées si le nombre de mots lexicaux est trop important dans un groupe mineur (cf. Zellner, 1996).

• Poser une frontière de groupe majeur

a. après tout signe de ponctuation (marque textuelle),

b. ainsi que sur la frontière de groupe mineur la plus proche du milieu de la phrase. (Le centre de la phrase est compris ici en termes de nombre de syllabes. Cette frontière intervient si la phrase comprend au moins douze syllabes.)

• Enfin, une pause — un silence — est insérée après une frontière majeure au milieu de la phrase et à la fin de la phrase.

La seconde étape pour prédire la structure temporelle consiste à attribuer une catégorie de durée syllabique en fonction du type de frontière prédit. Par exemple :

début de groupe mineur -> catégorie 1 (compression majeure)
fin de groupe mineur -> catégorie 3 (allongement mineur)
fin de groupe majeur -> catégorie 4 (allongement majeur)
dans un groupe mineur -> catégorie 2 (durée proche de la durée moyenne)

Résultats. Ces règles ont été appliquées aux mêmes 15 phrases soumises aux précédents algorithmes. Globalement, le degré de similarité entre les prédictions du modèle temporel et l'organisation temporelle effective est de 98.4%. Le taux des frontières prédites par cet algorithme mais non produites par le locuteur est de 9.6%.

4.c. Evaluation auditive

Finalement, un test de perception a été proposé à dix-huit étudiants en informatique, ne souffrant d'aucune déficience auditive.

---

que sur 3 zones (compression majeure, compression mineure, durée moyenne) mais seulement 0.87% de ces syllabes sont situées à la fin d'un groupe et requièrent un allongement. *Les syllabes de quatre segments* ne sont distribuées que sur 3 zones (durée moyenne, allongement mineur, allongement majeur) mais seulement 0.2% de ces syllabes sont situées au début d'un groupe et requièrent une compression.

Le test portait sur 16 phrases du corpus LAIP[13] (à fins de comparaisons avec un locuteur naturel) et 16 autres phrases prises dans des journaux ou magazines (cf. annexe l'ensemble du corpus). Pour mieux contrôler la complexité de la segmentation syntagmatique, le choix des phrases était fonction de la place du verbe dans la phrase : un tiers des phrases contenait le verbe dans le premier groupe prosodique, un tiers des phrases contenait le verbe dans le second groupe prosodique, un tiers des phrases contenait le verbe dans le troisième groupe ou au-delà. Chacune de ces phrases a ensuite été synthétisée selon deux modes. Le premier mode intègre le profil temporel prédit par l'analyse syntaxique automatique présentée au point 3.b., le second mode intègre le profil temporel prédit par le modèle temporel présenté au point 4.b. Chaque paire était présentée deux fois de suite dans un ordre aléatoire. Le test d'écoute à choix forcé consistait à choisir, pour ces 32 phrases lues par le synthétiseur, laquelle des deux versions semblait la plus naturelle.

Résultats. Le modèle temporel basé sur une approche psycholinguistique, a été préféré dans plus de 96% des cas. Le cas où l'approche syntaxique obtient son plus haut score (33.3% des choix) portait sur la phrase infinitive, qui est sans doute une structure phrastique moins habituelle.

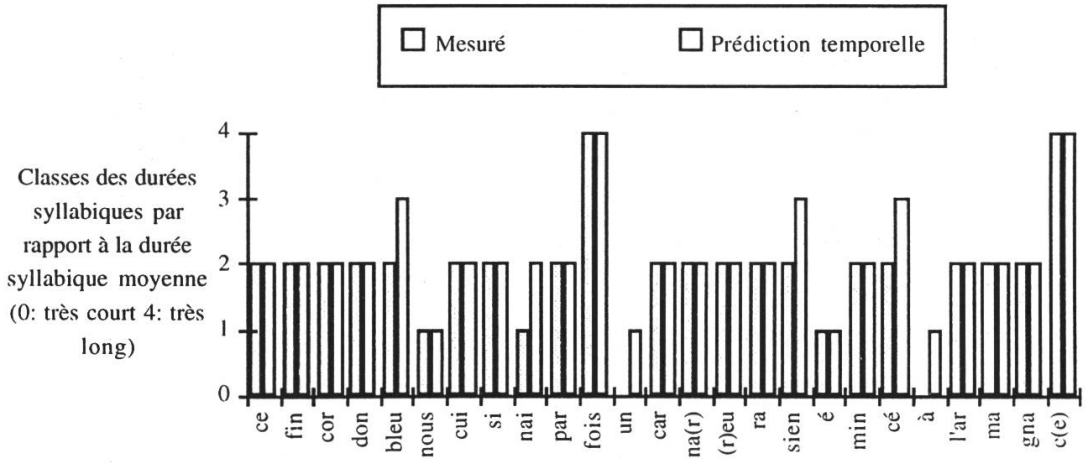

Figure 7. Comparaison entre le profil temporel mesuré et le profil temporel prédit selon le modèle temporel.

13. LAIP: Laboratoire d'Analyse Informatique de la Parole.

4.d. Discussion

Le tableau résumant les prédictions selon les différentes approches montre que pour les phrases testées, l'approche basée sur les hypothèses psycholinguistiques est celle qui permet d'obtenir le profil temporel le plus proche de celui produit par le locuteur, tant du point de vue général que du point de vue des types de frontières (mineure ou majeure).

Prédiction du profil temporel	Syntaxe	Phonosyntaxe	Psycholinguistique
Similarité avec les structures produites	63. 02%	81. 3%	98. 4%
«Fausses» frontières	40. 7%	38. 9%	9. 6%

Tableau 1 : Récapitulation des différentes approches

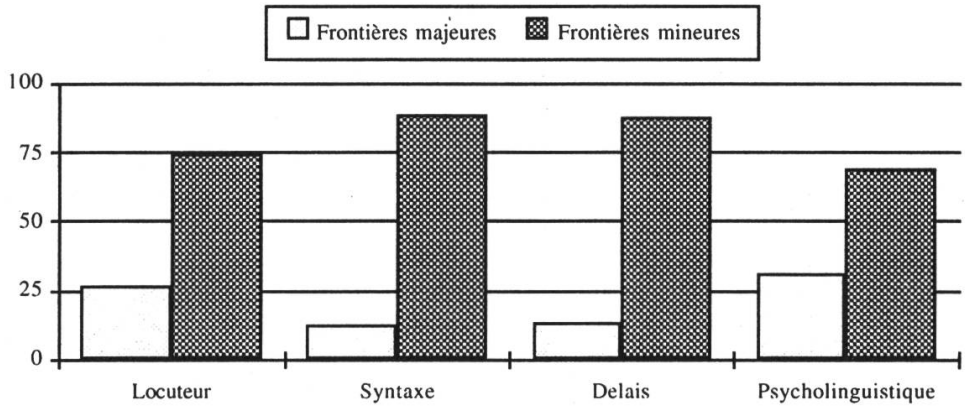

Figure 8. Comparaison de la répartition des frontières temporelles mineures et majeures pour les 50 phrases.

Il est clair que seules quelques approches syntaxiques et phonologiques ont été retenues. Il s'agissait d'approches respectant la linéarité du texte, ce qui se prête bien à la formalisation. Il n'est pas de notre propos de spéculer ici quant à la réalité psychologique de cette linéarité. Les processus liés à la production de la parole sont complexes (représentation du message, accès au lexique, encodage grammatical, encodage sémantique, encodage phonologique, etc.). La linéarité s'impose au moment de la lec-

ture elle-même lorsqu'il s'agit de produire sur le plan moteur les différents éléments du texte, l'un après l'autre. C'est à ce niveau de linéarité que nous nous situons.

D'autre part, cette évaluation est à considérer avec prudence dans la mesure par exemple où les interactions avec la mélodie ne sont pas entièrement connues. Une fréquence fondamentale très plate peut par exemple biaiser la perception et donner l'impression que la parole est plus rapide qu'elle ne l'est véritablement (Rietveld *et al.*, 1987).

Cela étant, cette étude avec évaluation statistique et auditive, indique que l'approche psycholinguistique, telle que formulée ci-dessus, semble appropriée pour prédire des structures temporelles proches de celle produites naturellement en situation de lecture. L'originalité de cette modélisation temporelle est qu'elle est indépendante des structures mélodiques, alors que généralement, les modules prosodiques pour la synthèse de la parole font entièrement dépendre le module temporel du module mélodique. Par là, il est démontré que les structures temporelles ne sont pas systématiquement sous la dépendance des structures mélodiques.

Notre revue des contraintes temporelles dans la parole, ci-dessus, a montré que la dynamique temporelle est multidimensionnelle, et non pas uniquement linguistique. Nous supposons donc que les structures linguistiques sont « pré-formées » par les structures langagières. Cette « pré-empreinte » psycholinguistique serait particulièrement évidente dans une situation de communication neutre.

5. Conclusion

En français, les algorithmes de prédiction des structures temporelles sont généralement basés sur la localisation des phénomènes accentuels. Or, les voix de synthèse issues de ces algorithmes ne présentent pas un bon niveau de fluence verbale (Zellner, 1996). Il apparaît que pour le français, l'accent ne semble pas être le facteur premier qui affecte la structure temporelle de la parole en situation de lecture neutre. Une étude pilote montre qu'un algorithme temporel, fondé sur une approche psycholinguistique, peut prédire des structures temporelles proches de celles produites par un locuteur, c'est-à-dire, sans « erreurs de performance ». Outre

l'intérêt de cette démarche dans la compréhension du phénomène de la fluence, cette approche permet également de reconsidérer la relation entre structure mélodique et structure temporelle.

Brigitte ZELLNER
Laboratoire d'analyse informatique de la parole (LAIP)
Section d'informatique et de méthodes mathématiques
Université de Lausanne
brigitte.zellner@imm.unil.ch

Références

ALLEN, J., HUNNICUTT, M. S., & KLATT, D. (1987). *From text to speech. The MITalk system.* Cambridge, England: Cambridge University Press.

ASTESANO, C., DI CRISTO, A. & HIRST, D. J. (1995). Discourse based empirical evidence for a multi-class accent system in French. *Proceedings of XIIIème Congrès International des Sciences Phonétiques, 4,* 630-633.

BACHENKO J., FITZPATRICK E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computationnal Linguistics, 16,* 155 -170.

BAILLY, G. (1983). *Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Établissement d'un modèle de génération.* Thèse d'ingénieur, Institut National Polytechnique de Grenoble.

BARBOSA, P. A. (1994). *Caractérisation et génération automatique de la structuration rythmique du français.* Thèse de Doctorat. U. R. A. CNRS n°368 – INPG/ENSERG, Université Stendhal, Grenoble.

BEAUGENDRE, F. (1994). *Une étude perceptive de l'intonation du français.* Thèse de Doctorat en Sciences de l'Université Paris XI. LIMSI n° 94-25.

BERTHOZ, A. (1997). *Le sens du mouvement.* Paris: Odile Jacob.

BOOMER, D. S., & Ditmann, A. T. (1962). Hesitation pauses and juncture pauses in speech. *Language and Speech, 5,* 215.

BOTTE, M. C., CANÉVET, G., DEMANY, L., & SORIN, C. (1989). *Psychoacoustique et perception auditive.* Série audition. Paris: Inserm/Sfa/CNET.

BOISSON-BARDIES, B. (1996). Que nous apprennent les enfants en babillant? *Revue Française de Linguistique Appliquée: La communication parlée, 1,* 55-65.

CAELEN-HAUMONT, G. (1991). *Stratégies des locuteurs et consignes de lecture d'un texte: Analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques.* Thèse d'Etat, Aix-en-Provence.

CAELEN-HAUMONT, G. (1994). Synthesis: Semantic and Pragmatic Predictions of Prosodic Structure, in E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition*. Chichester: J. Wiley & Sons, Ltd, 271-293.

CAELEN-HAUMONT, G. (1997, à paraitre). *Prosodie et sens, une approche expérimentale*. Ed. du CNRS.

CAELEN, J. (1996). *Le Dialogue Homme-Machine: vers une logique dialogique*. Lausanne: Séminaire en Intelligence Artificielle.

CANEVET, G. (1996). La psychoacoustique. Aspects fondamentaux et exemples d'applications. *Acoustique et Technique, 5,* 7-13.

CHAFE, W. (1980). Some reasons for hesitating. In W. Dechert & M. Raupach, (eds), *Temporal variables in speech*. The Hague: Mouton. 169-180.

CHOMSKY, N., & HALLE, M. (1968). *The sound pattern of English*. New-York: Harper and Row.

CHOMSKY, N. (1970). Deep structures, surface structures and semantic interpretation. In Jakobson and Kawamoto (eds), *Studies in General and Oriental Linguistics*. Tokyo: TEC, 52-91.

COOK, M., SMITH, J., & LALLJEE, M. (1974). Filled pauses and syntactic complexity. *Language and Speech, 17*, 11-16.

DAMASIO, A. & DAMASIO, H. (1997). Le cerveau et le langage. *Pour la Science: Les langues du monde*, 8-15.

DECHERT, W., & RAUPACH, M. (1980). *Temporal variables in speech*. The Hague: Mouton.

DELAIS, E. (1994). Prédiction de la variabilité dans la distribution des accents et les découpages prosodiques en français. *Actes des 20èmes Journées d'Etude sur la Parole*, 379-384.

DELAIS-ROUSSARIE, E. (1995). *Pour une approche parallèle de la structure prosodique: étude de l'organisation prosodique et rythmique de la phrase française*. Thèse de Doctorat, Université de Toulouse-Le Mirail.

DELL, F. (1984). L'accentuation dans les phrases en français. In F. Dell, D. Hirst, & J.-R. Vergnaud (eds), *Forme sonore du langage*. Paris: Hermann, 65-122.

DI CRISTO, A. (1975). Recherches sur la structuration prosodique de la phrase en français. *Actes des 6èmes Journées d'Études sur la Parole*, GALF-CNRS, 95-116.

DI CRISTO, A. & HIRST, D. (1994). Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français. *Travaux de l'Institut de Phonétique d'Aix, 15,* 13-24.

FROMKIN, V. (1980). *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*. New York: Academic Press.

GEE, J. P., GROSJEAN, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15*, 411-458.

GOLDMAN–EISLER, F. (1968). *Psycholinguistics: Experiments in spontaneous speech.* New York: Academic Press.

GOLDMAN–EISLER, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15,* 103-113.

GROSJEAN, F. (1980). Comparative studies of temporal variables in spoken and sign languages: A short review. In W. Dechert & M. Raupach (eds.). *Temporal variables in speech.* The Hague: Mouton. 307-312.

GROSJEAN, F., & DESCHAMPS, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français. *Phonetica, 31,* 144-184.

GROSJEAN, F., & DOMMERGUES, J. Y. (1983). Les structures de performance en psycholinguistique. *L'Année psychologique, 83,* 513-536.

GROSJEAN, F., & LANE, H. (1981). Temporal variables in the perception and production of spoken and sign languages. In P. D. Eimas & J. L. Miller (eds), *Perspectives on the Study of Speech.* Hillsdale: Lawrence Erlbaum Associates, 207-236.

JUN, S-A. & FOUGERON, C. (1995). The accentual phrase and the prosodic structure of French. *Proceedings of XIIIth. International Congres of Phonetic Sciences, 2,* 722-725.

KELLER, E. (1984). *Introduction aux systèmes psycholinguistiques.* Boucherville, Québec: Gaëtan Morin.

KELLER, E., ZELLNER, B., WERNER, S., & BLANCHOUD, N. (1993). The prediction of prosodic timing: Rules for final syllable lenthening in French. *Proceedings ESCA Workshop on Prosody,* 212-215.

KELLER, E., & ZELLNER, B. (1995). A statistical timing model for French. *Proceedings of XIIIth. International Congres of Phonetic Sciences, 3,* 302-305.

KELLER, E., & ZELLNER, B. (1996). A timing model for fast French. *York Papers in Linguistics, 17,* 53-75.

KO YOUNG-LIM (1996). *Étude prosodique du discours oral en français: variables temporelles et variables mélodiques dans l'interview radiophonique.* Thèse de Doctorat. Université de Strasbourg.

KONOPCZYNSKI, G. (1986). Vers un modèle développemental du rythme français: Problèmes d'isochronie reconsidérés à la lumière des données de l'acquisition du langage. *Bulletin de l'Institut de Phonétique de Grenoble, 15,* 157-190.

LAPICQUE, L. (1943). *La Machine nerveuse.* Paris: Flammarion.

LEVELT, W. J. M. (1991). *Speaking. From intention to articulation.* Cambridge: MIT Press.

LIBERMAN, M., & PRINCE A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8,* 249-336.

MARTIN, P. (1975). Intonation et reconnaissance automatique de la structure syntaxique. *Actes des 6èmes JEP, Galf-CNRS,* 52-62.

MARTIN, P. (1987). Structure rythmique de la phrase française. Statut théorique et données expérimentales. *Proceedings des 16èmes Journées d'Études sur la Parole.* Hammamet, 255-257.

MERTENS, P. (1987). *L'Intonation du français. De la description linguistique à la reconnaissance automatique.* Thèse doctorale, Katholieke Universiteit Leuven.

MERTENS, P. (1993). Intonational grouping, boundaries and syntactic structure in French. *Proceedings ESCA Workshop on Prosody,* 156-159.

MILLER, N., MARUYANA, G., BEABER, R. J. & VALONE, K. (1976). Speed of speech and persuasion. *Journal of Personality and Social Psychology, 4,* 615-624

MONNIN, P, & GROSJEAN, F. (1993). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique, 93,* 9-30.

OSTENDORF, M. & VEILLEUX, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics, 20,* 1-27.

PASDELOUP, V. (1992). A prosodic model for French text-to-speech synthesis: A psycholinguistic approach. In G. Bailly & C. Benoît (eds), *Talking Machines. Theories, Models, and Designs.* Amsterdam: Elsevier Science Publishers, 335-348.

PFAUWADEL, M. -C. (1986). *Être bègue.* Paris: Retz.

RIETVELD, A. C. M., & GUSSENHOVEN, C. (1987). Perceived speech rate and intonation. *Journal of Phonetics, 15,* 273-285.

ROSSI, M. (1985). L'intonation et l'organisation de l'énoncé. *Phonetica, 42,* 135-153.

SCHERER, K. (1984). Les émotions: fonctions et composantes. *Cahiers de Psychologie Cognitive, 4,* 9-39.

SCHERER, K. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143-165.

SCHERER, K. & ZEI, B. (1989). La voix comme indice affectif. *Revue médicale de la Suisse Romande, 109,* 61-66.

SELKIRK, E. O. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge: MIT Press.

STARKWEATHER, W. C. (1987). *Fluency and Stuttering,* New York: Prentice Hall.

STREET, R. & BRADY, R. (1982). Speech rate acceptance ranges as a function of evaluative domain, listener speech rate, and communication context. *Communication Monographs, 49,* 290-308.

TERKEN, J. & COLLIER, R. (1992). Syntactic influences on Prosody. In Y. Tohkura *et al.* (eds), *Speech Perception, Production and Linguistic Structure.* Amsterdam: IOS Press, 427-438.

TESNIERE, L. (1959). *Elements de syntaxe structurale.* Paris: Klincksieck.

WIGHTMAN, C. W. (1992). *Automatic detection of prosodic constituents for parsing.* Doctoral dissertation. Boston University Graduate School.

ZELLNER, B. (1992). Le bé bégayage et euh... l'hésitation en français spontané. *Actes des 19èmes Journées d'Études sur la Parole (J. E. P),* 481-487.

ZELLNER, B. (1994). Pauses and the temporal structure of speech. In E. Keller (ed.), *Fundamentals of speech synthesis and speech recognition.* Chichester: John Wiley, 41-62.

ZELLNER, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée, 1,* 7-23.

ZWICKER, E., & FELDKELLER, R. (1981). *Psychoacoustique: l'oreille récepteur d'information.* Collection technique et scientifique des télécommunications. Paris: Masson.

Annexe

Phrases testées. Elles sont marquées d'une, deux ou trois étoiles selon l'emplacement du verbe dans le premier, deuxième ou troisième groupe temporel. Les parenthèses indiquent que la phrase a été uniquement utilisée pour le test d'écoute.

C'est à l'autre service des achats de la rue des Cardinaux. *
Il a un ami chaleureux d'un âge honorable. *
Certains analystes-programmeurs prétendent éhontément que la mini-informatique ne résoudra jamais leurs problèmes. **
Les articles des femmes scientifiques sont plus souvent cités que ceux des hommes. (***)
C'est au sujet de la référence complète d'un ballon de rugby qui pose problème. *
Le tout fut servi avec un bon beaujolais. **
On boit volontiers à l'ombre des peupliers en regardant déambuler les touristes. *
Ce fin cordon-bleu nous cuisinait parfois un canard eurasien émincé à l'armagnac. **
Elle a chuchoté pendant le cours magistral qu'elle avait le fameux livre. (*)
Dans un contexte de concurrence internationale, les entreprises doivent renforcer leur compétitivité. (***)
Selon le ministre des finances cette crise boursière ne doit pas inquiéter outre mesure les citoyens. ***
Ils étaient déçus et ils attendaient avec impatience la fin de cet ennuyeux spectacle. (*)
Cet homme discret recueillit les dernières confidences de son très grand ami d'enfance. (**)
Les écrans d'ordinateurs et de téléviseurs émettent un champ électro-magnétique. (***)
Chez de nombreuses espèces de fourmis se sont développées des castes de soldats destinées à défendre la colonie. (***)
Il se hâte de fuir les immeubles subventionnés alors qu'il pourrait tellement épargner. *
L'américain se nourrissait de sandwichs tout en racontant de nouveaux gags. (**)
Henri bredouille affreusement d'après le docteur. **
Ils ont apporté des piles de livres anciens en prenant le plus grand soin pour les transporter. (*)
Il régnait une moiteur esquintante dans ce parc à l'abandon. *
Le mot suivant est "flouc". (**)
Quelques moutons broutaient derrière le mur qui s'assombrissait. (**)
L'élaboration des nids de pigeons guyanais est restée inchangée depuis des siècles. ***

Une reine terrible au cou vermeil se baigna sans enlever sa robe impressionnante. ***

La réussite et le développement des entreprises technologiques réclament en particulier la construction de liens durables avec des laboratoires de recherche académique. (***)

Le ridicule est donc la perception immédiate d'une perturbation dans l'ordre de la durée humaine. (**)

Vouloir annihiler la pollution au cyclohexane graisseux avec de la soude est totalement illusoire. ***

Ce reportage analyse très bien l'extrême complexité de la situation conflictuelle au Sri Lanka. (*)

Mais une telle tentative devait malheureusement se heurter à des difficultés considérables. (**)

Le garnement vaniteux effrayait son cousin provincial en déambulant comme une troupe de soldats. (**)

Quel bien-être de vivre dans un deux-pièces et de disposer d'une buanderie.